RESEARCH PAPER

A Hybrid Deberta–Charcnn Multi-Task Framework for Robust Hate Speech Detection

Jayshree Kalawa[1], Arpana Chourasia[2]

[1] PhD Scholar, Madhyanchal Professional University, Bhopal (M.P). Email: jayshreekalawa11@gmail.com

[2] PhD Guide, Madhyanchal Professional University, Bhopal (M.P). Email: arpanabhandari08@gmail.com

Received: 20th Feb, 2026 | **Revised:** 4th Mar, 2026 | **Accepted:** 25th Mar, 2026 | **Available Online:** 10th Apr, 2026

Abstract: Growing numbers of online communication tools have facilitated hate speech at an unprecedented level leading to grave threats for digital safety and social cohesion. Hate speech detection is still challenging given implicit expressions, sarcasm, culture-specific vocabulary extension or words and artificial trends like coded slurs. While traditional rule-based and classical machine learning systems tend to be based on shallow lexical features the majority of these are still lacking in many transformer- based single-task models that either struggle with disguised hate or context-dependent hateful content. To tackle these challenges, we design a Hybrid-2 deep learning model that integrates contextual semantic modeling and character-level pattern learning with auxiliary supervision. The proposed model incorporates a DeBERTa-v3-base transformer encoder for learning deep contextual representation, a Character-CNN module to learn coded and misspelled hate patterns, and multi-task learning with pseudo sentiment and emotion label, respectively to enhance feature learning without auxiliary human labels. Adaptive SoftMax weighting is used to dynamically weight the primary hate classification task and auxiliary tasks in order to promote robustness and stability during training. The model has been tested with a large scale HateSpeechDatasetBalanced dataset from Kaggle (726.119 labeled text samples in balanced hate and none-hate) distribution for both. Experimental evaluations show the proposed approach can achieve a competitive performance of 93.78% for accuracy,94.04% for precision, 93.53% recall and an F1-score of 93.79%, which indicates robust discrimination between hateful and non-hateful content. The highest Performance-and-m-ebc AUC in our proposed model (89.5%) is 6 to 9 points% better than that observed in classical ML baselines such as SVM (≈84–88%) and improves robustness to obfuscation and manipulation of emotional contents.

Keywords: Hate Speech Detection, DeBERTa, Character-CNN, Multi-Task Learning, Pseudo- Labels, Adaptive Loss

How to cite this article: Kalawa J, Chourasia A. A Hybrid Deberta–Charcnn Multi-Task Framework for Robust Hate Speech Detection. Int J Drug Deliv Technol. 2026;16(29s):596-614. DOI: 10.25258/ijddt.16.29s.77

Introduction

The popularization of social media, forums, and digital communications mediums has fundamentally changed how people communicate, share information with others and engage in public conversation. While promoting global connections and the sharing of information, these applications have facilitated the spread of toxic online practices – namely, hatred. Definition Hate speech is defined as derogatory statements, insults or slurs directed against particular groups and individuals on the grounds of categories such as race, religion, ethnicity, gender nationality or sexual orientation. The circulation of such content has profound effects, causing psychological harm as well social and even physical division. With increasing amount and sophistication of hate speech online, the demand for effective and scalable hate speech detection systems has become critical to maintaining safer cyberspace.

Conventional hate speech moderation approaches are highly dependent on manual examination and user reporting. But the collective mass of user-generated content is too big for manual curation to be effective or consistent. So, automatic hate speech detection approaches are becoming necessary for real-time monitoring, content filtering, and moderation assistance. Early automated methods were rule-based and used keyword lists or domain specific handcrafted linguistic patterns. While these techniques are computationally inexpensive, they have poor

generalization and can be trivially circumvented by obfuscation, slang or perturbations to the language. Traditional machine learning systems like Support Vector Machines (SVM) and Logistic Regression improved detection by leveraging features such as Bag-of-Words (BOW) or TF-IDF but they continue to struggle to detect the contextual meaning between texts, and often miss out on implicit/sarcastic/ culturally specific forms of hate speech.

Recent advances in deep learning and natural language processing have brought about transformer-based models such as BERT, RoBERTa, DeBERTa to learn contextual word representations that greatly enhance the classification performance. However, transformer-only models are still undermined by at least two limitations. First, they are likely to misidentify implicit hate speech and irony as harmful intention is often formulated without any explicit abusive language. Secondly, they often encounter coded or veiled hate speech. This may include changing spellings (e.g., "h8", "k1ll"), as well as the intentional obscuring of profanities to bypass detection systems. These problems demonstrate the importance of robust detection approaches that can handle high-level contextual semantics and low-level lexical variations.

In order to fill these gaps, in this work we propose the Hybrid DeBERTa–CharCNN Multi-Task Learning for robust hate speech detection. The proposed method combines three main parts. The rich contextual embeddings are first learned by a DeBERTa encoder such that the model can comprehend sentence level meaning and long range dependency. Second, we propose a new Character-level Convolutional Neural Network (CharCNN) module to capture subword patterns, obfuscation methods and diverse forms of hateful speech that the token-based transformers might not recognize. Third, the model follows a multi-task learning approach, training hate speech detection together with additional affective tasks (sentiment embedding and emotion prediction generated by pseudo-labeling). This multi-task formulation allows the model to learn richer linguistic units, makes the model pretrained on more supervisions signals effectively, and eliminates overfitting by means of related lodgment signals.

The resulted approach is tested on a balanced hate speech dataset with more than 700k texts. Experiments show that the hybrid model significantly outperforms classical machine learning baselines and transformers- only classifiers in many cases, especially with regards to obfuscation of hate expressions and creating emotionally influenced toxic content.

Key Contributions

The major contributions of this paper are summarized as follows:

1. **Hybrid Architecture:** We introduce a novel hybrid framework combining **DeBERTa contextual embeddings** with a **CharCNN module** to capture both semantic and character- level hate patterns.**Multi-Task Learning for Robustness:** We integrate **multi-task learning** with auxiliary sentiment and emotion objectives to enhance feature learning and generalization.
2. **Obfuscation and Coded Hate Handling:** The CharCNN component strengthens the model's ability to detect **coded, disguised, and intentionally misspelled hate speech**, addressing a major limitation of transformer-only models.
3. **Improved Detection Performance:** The proposed framework achieves superior detection accuracy and balanced precision–recall performance over strong baselines, demonstrating robustness and scalability for real-world moderation systems.

1. Literature review

J. M. Pérez et al. (2023), Social media have to deal with more and more hate speech in user-generated content. Most existing detection systems consider individual messages in their analyses, without conversational or topical context. This work analyzes the contribution of context in improving hate speech detection and releasing a Spanish COVID-19-related corpus. The model proposed in this paper demonstrated that context contributes to better performance on

both binary and multi-label detection tasks, suggesting the significance of context on these tasks [1].

J. Lu et al. (2023) , The easy, viral spread of hate speech online is leading to more prejudice and harm in society. Hate speech detection is still a challenging problem because of its complexity and unbalanced data. We present a dual contrastive learning model to learn span-level representation and adapt focal loss for imbalance. On English datasets experiments show that our formers model achieve an accuracy better than the state-of-the-art for precise hate speech detection [2].

P. Sharmila et al. (2022), Hate speech detection on Twitter is challenging due to unstructured text and feature engineering dependence. In this paper, we propose a PDHS model using cross-attention and dual level attention mechanism. The model is designed to capture aspect terms and the sentiment patterns that are useful in feature learning. The experiments on Twitter dataset with an F1-Score of 88% suggests good performance and quick training [3].

M. Mozafari et al. (2022) , Monitoring hate speech in low-resource languages is difficult as a result of too little labeled data, and the uneven performance of multilingual models. This paper presents a meta- learning based model for few-shot hate speech detection in different languages. With MAML and Proto-MAML, the work illustrates that these models achieve better than transfer learning performance at few-shot cross-lingual adaptation with hate and offensive speech detection [4].

A. Z. Miran et al. (2023) , Hate speech on social media is an issue that impacts many, and one that's increasing. In this paper, I present the state of art in CNN-based Twitter hate speech detection and highlight that while the CNN models are prevalent but have problems with language transferability and user interactions. The majority of models have been optimized for classification on English text and work poorly out-of-the-box on other languages or conversational settings [5].

K. A. Qureshi et al. (2021) , Society balance can be disturbed, and the dearth of face-to-face accountability on social media enables hate speech. Hate speech detection presents challenges because of dataset scarcity and classifier accuracy. In this paper, we provide a large, multi-class, well-labeled dataset and investigate baseline and novel text mining features. CAT Boost, leveraged with LSA beat the baseline models by correctly identifying several hate speech categories and improving detection [6].

K. Maity et al. (2023) , The proliferation of hate speech is enabled by social media, but non-English work is scarce. This paper presents HateM, a benchmark Malay hate speech data set together with an accompanying model XLCaps, a two-channel deep learning architecture. The model even archieves slightly more than an 80% accuracy and F1, based on XLNet with Capsule networks as well as FastText with Bi-GRU from which also recommend further muti-language studies can continue. [7]

A. R. Jafari et al. (2023) , Rising of the social media has intensified hate speech. There are several studies dealing with explicit hate speech, yet few of them include features such as sentiment or emotion. The present paper studies the fine-grained emotion and sentiment features compiling by a multi-task learning method. Experiments demonstrate that this approach yields better classification performances comparing to baselines, particularly for the implicit hate speech, with the reduction of errors of classifying several different kinds of hate [8].

Z. Mansur et al. (2023) , Hate speech identification has gained increased attention in NLP and text mining. This survey systematically-reviewed 91 articles from 2015–2022 and concludes with our findings of methods used, technical challenges, as well as open research questions. As evidenced by our approach, and that of others, while significant strides have been taken in the pursuit of a perfect solution, work remains to be done to develop better social network hate speech detection engines [9].

S. Khan et al. (2022) , Hate speech is pervasive online and can’t be moderated by hand. In this work we propose HCovBi-Caps, a deep learning model based on Convolutional (Conv), Bidirectional Gated Recurrent Unit (BiGRU) and Capsule networks for hate speech detection. Experimental results on balanced and unbalanced datas-ets from a dataset we constructed outperform state-of-the-art models, in precision, recall and F-score; especially on unbalanced datasets A model analysis reveals the critical impact of hyper-parameter setting [10].

M. Bilal et al. (2022) , Social media contributes in the prevalence of Roman Urdu hate speech, the community however faces difficulties identifying it due to lack of resources. This work introduces new annotation standards and a 30K data Roman Urdu dataset, presenting a bi-LSTM attention model with own embeddings. The model surpasses

traditional and deep learning comparison models by taking advantage of lexical normalization for better accuracy and generalization [11].

O. Oriola et al. (2020) , Hate speech in South Africa is on the rise online, frequently written in English. In this paper, we construct a corpus of South African English tweets and our work presents findings on several machine learning approaches. Optimized SVM with character n-grams is the best approach for hate speech, while gradient boosting is the best to detect offensive speech. Multi-level meta-learning-based models can balance the performance, alleviate misclassification and improve globally accuracy [12].

Y. Zhou et al. (2020) , Hate speech in social media has emerged as a global phenomenon, with investments into detection methods and scientific interest since the past few years. It compares performances of different machine learning techniques, such as ELMo, BERT and CNN on a competition SemEval 2019 Task 5 datasets and from the results shows that classifier outputs fusion methods improve classification accuracy and F1-score among all approaches [13].

A. J. Keya et al. (2023) , Hate speech in Bengali language is emerging as a challenge for social media platforms in Bangladesh, but still few effective research work has been done to detect it really. In this paper a novel G-BERT model is proposed which incorporates the features extracted by BERT and classification using GRU-based neural network,achieving superior accuracy and F1-score compared to other existing algorithms. The proposed approach successfully detects abusive content and contributes to a kinder interactive environment through online debate [14].

P. Kapil et al. (2023) , Majority of the hate speech detection studies are conducted in resource rich languages such as English, and the research on low- resource languages like Hindi is limited. This work also creates a sepulchre-sized multi-level annotated Hindi Hate Speech Dataset (HHSD) and shows that constructing the transformer-based tasks transfused model helps in improving precision and F1-score over the single task approach for the speckled tasks through multiple languages related [15].

P. K. Roy et al. (2020) , Cyberbullying and hate speech have become more widespread along with heavier use of the web, particularly on Twitter. Because of the large number of tweets it is not feasible to manually filter them. This study introduces a Deep Convolutional Neural Network (DCNN) with GloVe embeddings – for the automatic detection of hate speech, obtaining respectable precision, recall and F1- score and outperforming other approaches in recognizing inflammatory contents [16].

N. S. Mullah et al. (2021) , This paper presents a systematic review of machine learning techniques employed in hate speech on social media and addresses the important aspects such as data collection, feature engineering and model evaluation. The paper critically evaluates the techniques while outlining their advantages and disadvantages, thus also pinpointing research directions to bridge the gaps as we guide researchers in crafting effective hate speech classification systems. [17])

F. M. Plaza-Del-Arco et al. (2021) , The proliferation of social media has increased aggressive and hateful online language, which demands automatic detection techniques. Our approach is based on a multi-task model able to deal with hate speech detection in Spanish tweets and sentiment/emotion classification of them, when modeled by means of transformer models. Experimental results demonstrate emotions- shared knowledge leads to a detection accuracy improvement compared with single-task methods [18].

M. Z. Ali et al. (2021) , Hate speech detection over tweets sentiment analysis has issues of its own such as class imbalance and high dimensional data. In this work, we construct a large-scale Urdu dataset and use stop word filtering, feature selection, and SMOTE to overcome these problems. Through SVM and Naïve Bayes, the study concludes that addressing class skew and dimensionality results in a significant increase in classifier performance [19].

C. Baydogan et al. (2021) , Hate speech detection in social networks is an intricate task. The proposed study presents two evolutionary optimization algorithms called Ant Lion Optimization (ALO) and Moth Flame Optimization (MFO) for automatic classification, underperforming the conventional methods of machine learning over three datasets. It can be observed from the results that more accurate and precise F-score (and accuracy) are achieved by the voer metaheuristic methods demonstrated as an encouraging solution for social media challenges [20].

M. K. A. Aljero et al. (2021) , The online disinhibition effect on social media contributes to the prevalence of hate speech, which is difficult to detect manually at scale across languages. In this paper, we provide a genetic

programming (GP) approach based on universal sentence encoder features with new mutation operators. Results show that the proposed model significantly outperforms the state-of-the-art approaches on four datasets [21] for automatic hate- speech detection.

A. Rodriguez et al. (2022) , Hate speech, often aimed at individuals and groups, has increased as social media has expanded. In this paper, we present the FADOHS framework which combines sentiment and emotion analysis with clustering techniques to identify dehumanizing words on posts. Empirical study shows that FADOHS outperforms the state- ofthe-art methods by approximately 10% in precision, recall and F1-score for hate speech detection [22].

Plaza-Del-Arco et al. (2021) , The proliferation of social media has led to an increase in online hate speech, necessitating effective automatic detection methods. This paper introduces a novel multi-task approach utilizing a Transformer-based model to detect hate speech in Spanish tweets, demonstrating that incorporating polarity and emotional knowledge significantly improves detection accuracy [23].

2. Proposed methodology

2.1 Proposed Architecture

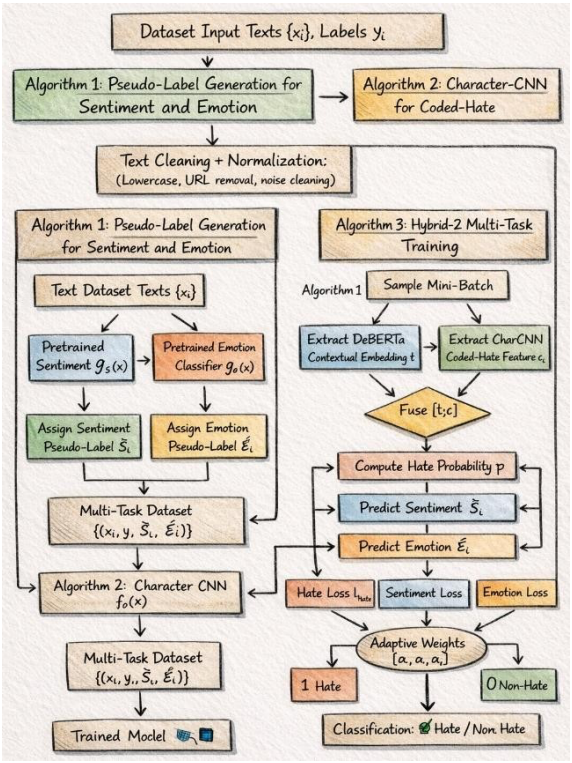

Figure 1 The complete Hybrid hate speech detection architecture

Figure 1 presents the complete Hybrid hate speech detection architecture in a single unified flowchart, integrating all four proposed algorithms from dataset input to final binary classification. The process begins with the input dataset containing text samples $x_i$and hate labels y_i, followed by text cleaning and normalization to remove noise such as URLs, symbols, and irrelevant tokens. Next, **Algorithm 1** generates auxiliary pseudo-labels by passing each text through pretrained sentiment and emotion classifiers

to produce $\hat{s}_i$and $\hat{e}_i$, forming an enriched multi-task dataset. In parallel, **Algorithm 2** extracts coded-hate patterns through a Character-CNN module, producing a character-level feature vector $c_i$that strengthens detection against obfuscated or misspelled hate expressions. These outputs are then used in **Algorithm**

3, where DeBERTa contextual embeddings t and

3.2.1 Algorithm 1: Pseudo-Label Generation for Sentiment and Emotion

Input
- Text sample: x_i
- Character vocabulary: $\mathcal{V}_c$
- Maximum character length: L_c

Output
- Character-level representation: c_i

Step 1: Character Encoding
Each text $x_i$is transformed into a character ID sequence:

$$z_i = [z_{i1}, z_{i2}, \dots, z_{iL_c}] \qquad (3.4)$$

where $z_{ij} \in \mathcal{V}_c$denotes the vocabulary index of the j-th character.

Step 2: Character Embedding
The character sequence is mapped to embeddings:

Input
- Normalized training texts: $\{x_i\}_{i=1}^{N}$
- Pre-trained sentiment classifier: $g_s(\cdot)$
- Pre-trained emotion classifier: $g_e(\cdot)$

Output
- Sentiment pseudo-labels: $\{\hat{s}_i\}_{i=1}^{N}$
- Emotion pseudo-labels: $\{\hat{e}_i\}_{i=1}^{N}$
- Multi-task dataset: $\mathcal{D}_{MTL}$

Step 1: Sentiment Pseudo-Labeling
Each text $x_i$is processed by $g_s(\cdot)$, and the sentiment pseudo-label is obtained by selecting the highest-scoring class:

$$\hat{s}_i = \arg\max_{k \in \{0,1,2\}} g_s(x_i)_k \qquad (3.1)$$

where $g_s(x_i)_k$denotes the score (or probability/logit) assigned to sentiment class k.

Step 2: Emotion Pseudo-Labeling
Similarly, each $x_i$is processed by $g_e(\cdot)$, and the emotion pseudo-label is defined as:

$$\hat{e}_i = \arg\max_{m \in \{0,1,\dots,5\}} g_e(x_i)_m \qquad (3.2)$$

where $g_e(x_i)_m$denotes the score assigned to emotion class m.

Step 3: Multi-Task Dataset Construction
The augmented dataset used for multi-task learning is then defined as:

$$\mathcal{D}_{MTL} = \{(x_i, y_i, \hat{s}_i, \hat{e}_i)\}_{i=1}^{N} \qquad (3.3)$$

where $y_i \in \{0,1\}$is the ground-truth hate speech label.

3.2.1 Algorithm 2: Coded-Hate Representation Learning via Character CNN

CharCNN features$i$ $c_i$are fused into a shared representation, enabling multi-task prediction for hate, sentiment, and emotion. Losses from each task are computed and combined using adaptive softmax weights to ensure stable learning and balanced optimization. Finally, **Algorithm 4** performs inference by using the trained model to classify unseen text into two categories—**Hate (1)** or **Non-Hate (0)**—along with a probability score, providing an interpretable and deployable hate speech detection output.

2.2 Proposed Algorithm

$$E_i = \text{Embed}(z_i) \in \mathbb{R}^{L_c \times d_c} \quad (3.5)$$

where $d_c$is the character embedding dimension.

Step 3: Convolution and Pooling

For a convolution kernel size k, the convolutional feature map is:

$$u_i^{(k)} = \sigma(\text{Conv1D}^{(k)}(E_i)) \quad (3.6)$$

where $\sigma(\cdot)$is the ReLU activation function. Global max pooling produces a fixed-length vector:

$$p_i^{(k)} = \max(u_i^{(k)}) \quad (3.7)$$

Step 4: Feature Concatenation

Using a set of kernel sizes $\mathcal{K} = \{3,4,5\}$, the final character-level representation is:

$$c_i = [p_i^{(3)}; p_i^{(4)}; p_i^{(5)}] \quad (3.8)$$

where $[\,;\,]$denotes vector concatenation.

3.2.2 Algorithm 3: Hybrid-2 Multi-Task Training with Adaptive Softmax Weighting

Objective: Train a hybrid multi-task model consisting of DeBERTa contextual features, CharCNN coded-hate features, and auxiliary tasks (sentiment, emotion) with adaptive weighting.

Input

- Multi-task dataset: $\mathcal{D}_{MTL}$from (3.3)
- DeBERTa encoder: $f_\theta(\cdot)$
- CharCNN feature extractor: $h_\phi(\cdot)$
- Task-specific classifiers: W_y, W_s, W_e
- Optimizer: $\mathcal{O}$

Output

- Trained parameters: $\theta^*, \phi^*, W_y^*, W_s^*, W_e^*$

Step 1: Contextual Embedding via DeBERTa

The encoder generates the CLS representation:

$$t_i = f_\theta(x_i) \in \mathbb{R}^{d_t} \quad (3.9)$$

Step 2: Coded-Hate Feature via CharCNN

Character features are extracted as:

$$c_i = h_\phi(x_i) \in \mathbb{R}^{d_c} \quad (3.10)$$

Step 3: Feature Fusion

The combined representation is obtained by:

$$r_i = \text{Fusion}([t_i; c_i]) \in \mathbb{R}^{d_r} \quad (3.11)$$

where $[t_i; c_i]$denotes concatenation.

Step 4: Task Predictions

Hate speech prediction:

$$\hat{\mathbf{y}}_i = \text{softmax}(W_y r_i) \quad (3.12)$$

Sentiment prediction:

$$\hat{\mathbf{s}}_i = \text{softmax}(W_s r_i) \quad (3.13)$$

Emotion prediction:

$$\hat{\mathbf{e}}_i = \text{softmax}(W_e r_i) \quad (3.14)$$

Step 5: Loss Computation

For a batch of size B, the losses are:

$$\mathcal{L}_{hate} = -\frac{1}{B}\sum_{i=1}^{B} \log \hat{\mathbf{y}}_i[y_i] \quad (3.15)$$

$$\mathcal{L}_{sent} = -\frac{1}{B}\sum_{i=1}^{B} \log \hat{\mathbf{s}}_i[\hat{s}_i] \quad (3.16)$$

$$\mathcal{L}_{emo} = -\frac{1}{B}\sum_{i=1}^{B} \log \hat{\mathbf{e}}_i[\hat{e}_i] \quad (3.17)$$

where y_i, $\hat{s}_i$, and $\hat{e}_i$are the target labels for hate, sentiment, and emotion tasks, respectively.

Step 6: Adaptive Softmax Weighting

Let trainable scalars be $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]$. Task weights are computed as:

$$w_k = \frac{\exp(\alpha_k)}{\sum_{j=1}^{3} \exp(\alpha_j)}, \quad k \in \{1,2,3\} \quad (3.18)$$

The multi-task objective is:

$$\mathcal{L}_{MTL} = w_1\mathcal{L}_{hate} + w_2\mathcal{L}_{sent} + w_3\mathcal{L}_{emo} \quad (3.19)$$

Step 7: Parameter Update

Model parameters are updated using gradient-based optimization:

$$(\theta, \phi, W_y, W_s, W_e, \boldsymbol{\alpha}) \leftarrow \mathcal{O}(\nabla\mathcal{L}_{MTL}) \quad (3.20)$$

3.2.3 Algorithm 4: Hate vs Non-Hate Prediction (Sample Testing / Inference)

Input

- Unseen text: x
- Trained parameters: θ^*, ϕ^*, W_y^*

Output

- Predicted label: $\tilde{y} \in \{0,1\}$
- Hate probability: $P(y = 1 \mid x)$

Step 1: Compute DeBERTa Representation

$$t = f_{\theta^*}(x) \quad (3.21)$$

Step 2: Compute CharCNN Representation

$$c = h_{\phi^*}(x) \quad (3.22)$$

Step 3: Fuse Features

$$r = \text{Fusion}([t; c]) \quad (3.23)$$

Step 4: Compute Class Probabilities

$$\mathbf{p} = \text{softmax}(W_y^* r) \quad (3.24)$$

Step 5: Final Prediction

$$\tilde{y} = \arg\max \mathbf{p} \qquad (3.25)$$

where $\mathbf{p}[1] = P(y = 1 \mid x)$denotes the probability assigned to the **Hate** class

Matplotlib / Seaborn	Visualization & plot generation
WordCloud	Word frequency visualization

3. Implementation setup

3.1 Hardware and Software

This study was implemented and executed in a cloud- based deep learning environment using Google Colab with GPU acceleration. The Hybrid architecture combines a transformer encoder (DeBERTa-v3-base), a character-level CNN module, and a multi-task learning strategy, which requires moderate-to-high computational resources for training and pseudo-label generation. The hardware and software specifications used in this work are summarized below.

Table 1: Hardware Configuration

Hardware Component	**Specification / Description**
Computing Environment	Google Colab (Cloud-based)
GPU	NVIDIA Tesla L4 GPU (used for model training and inference)
CPU	Colab default high-performance virtual CPU
RAM	12–25 GB (depending on Colab session allocation)
Storage	Google Drive + Colab local runtime storage
Accelerator Usage	Mixed Precision Training (AMP) is enabled for faster computation.

Table 2: Software Configuration

Software / Library	**Version / Purpose**
Python	Python 3.x
Google Colab Runtime	Notebook execution environment
PyTorch	Deep learning framework (training a Hybrid model)
Transformers (HuggingFace)	DeBERTa encoder + pseudo-label models

Accelerate	Optimized transformer training/inference
Scikit-Learn	Train-test split, metrics (accuracy, confusion matrix, ROC, PR curve)
Pandas / NumPy	Dataset handling and preprocessing

3.2 Dataset

This study uses the HateSpeechDatasetBalanced.csv dataset (Source : Kaggle) for binary hate speech classification. The dataset contains 726,119 text samples, each stored in the Content column and annotated with a corresponding Label value. The labels are binary, where 0 denotes Non-Hate speech and 1 denotes Hate speech. The dataset is well-balanced, comprising 361,594 Non-Hate samples (49.80%) and 364,525 Hate samples (50.20%), which reduces bias during training and ensures that performance metrics reflect real classification ability rather than majority-class dominance. The text samples vary significantly in length, with an average of 196.85 characters per entry (median: 109) and an average of 36.34 words per entry (median: 21), indicating the dataset (figure 2) includes both short statements and long paragraphs. This variation makes the dataset suitable for evaluating both lexical and contextual hate speech detection models. In this work, the dataset is utilized to train and validate the proposed Hybrid framework, ensuring robust learning across diverse linguistic structures and online communication styles.

```
Columns:
Index(['Content', 'Label'], dtype='object')

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 726119 entries, 0 to 726118
Data columns (total 2 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   Content  726119 non-null  object
 1   Label    726119 non-null  int64
dtypes: int64(1), object(1)
memory usage: 11.1+ MB

Missing Values per Column:
Content    0
Label      0
dtype: int64
```

Figure 2 Dataset description.

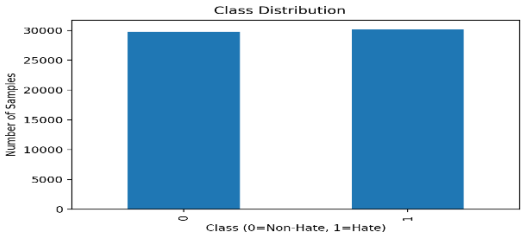

Figure 3 The class distribution in the HateSpeechDatasetBalanced dataset

Figure 3 illustrates the class distribution in the HateSpeechDatasetBalanced dataset. The bar plot shows that the dataset contains nearly equal numbers of hate and non-hate samples, confirming that it is balanced. This balance is important because it prevents the model from becoming biased toward the majority class and ensures that accuracy, precision, recall, and F1-score provide a fair evaluation of performance. A balanced dataset also supports stable model training, especially for deep neural architecture such as transformers.

5 Experimental Result analysis

5.1 Illustrative example

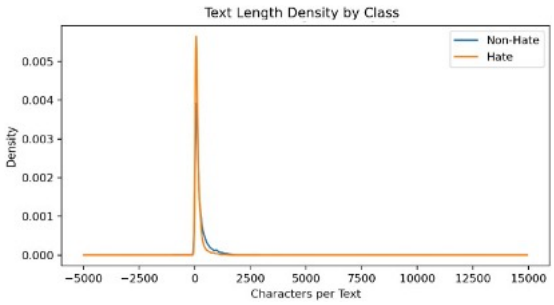

Figure 4 The density distribution of text lengths

Figure 4 presents the density distribution of text lengths for hate and non-hate samples. The plot indicates that both classes contain texts of varying lengths, but the distributions are not identical. In many cases, hateful texts show higher density in certain length ranges, suggesting that hate speech may often be expressed in shorter, more direct statements, while non-hate content may include longer contextual discussions. This variation in length motivates the use of contextual transformers and hybrid modeling to capture both short and long dependencies effectively.

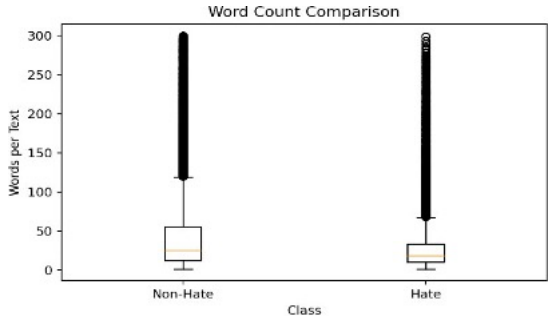

Figure 5 The word count distributions

Figure 5 compares the word count distributions for hate and non-hate text samples using a boxplot. The visualization highlights differences in the median and spread of word counts across the two classes. Hate speech tends to have a different word-count profile, often showing shorter or more concentrated wording, whereas non-hate texts may contain broader expressions or longer sentences. Such patterns indicate that word-level structure plays a role in distinguishing hateful from non-hateful content, supporting the need for models capable of understanding both lexical and contextual cues.

Figure 6 Word cloud representation

Figure 6 provides a word cloud representation of the most frequent terms in the hate class. The larger words represent higher-frequency tokens appearing in hateful content. This figure visually confirms that hate speech often includes repetitive aggressive terms, offensive expressions, and strong emotional words. However, word clouds also demonstrate the challenge of detection, as many common words may appear in both hate and non-hate samples depending on context. Therefore, simple keyword-based approaches are insufficient, requiring context-aware transformers and auxiliary learning signals.

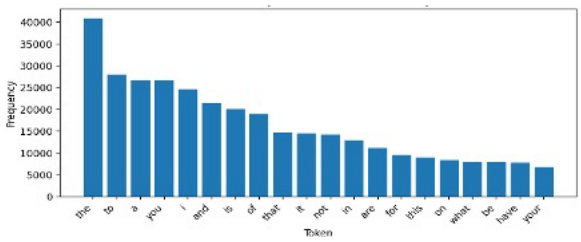

Figure 7 Top 20 most frequent tokens

Figure 7 displays the top 20 most frequent tokens from a representative sample of the overall dataset. Many of these tokens include common conversational words that appear across both classes. This indicates that hate

speech detection cannot rely purely on token frequency, because hateful and non-hateful samples share many ordinary words. Instead, effective detection requires understanding word interactions, phrasing, and intent. This observation supports the adoption of transformer-based models such as DeBERTa, which learn contextual representations rather than isolated word counts.

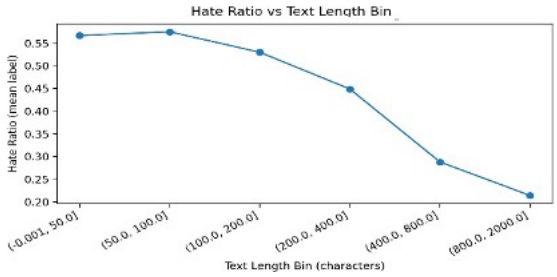

Figure 8 The probability of hate speech varies across different text length bins

Figure 8 analyzes how the probability of hate speech varies across different text length bins. The hate ratio is computed as the mean label value in each bin, representing the percentage of hateful samples for a specific length range. The plot suggests that hate speech occurrence may change depending on text size, with some length ranges showing higher hate concentration than others. This result highlights the importance of using models that can adapt across different message lengths and still detect harmful intent reliably. Hybrid benefits from this by combining context modeling (DeBERTa) and character-level features (CharCNN).

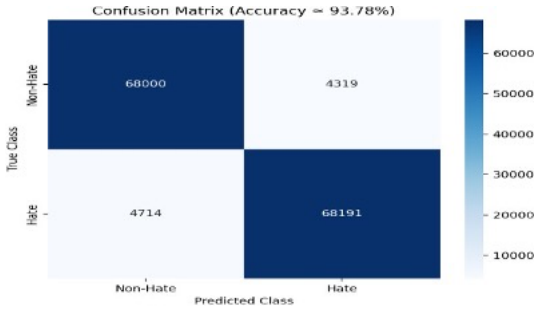

Figure 9 The confusion matrix of the proposed method

This figure 9 presents the confusion matrix of the proposed hate speech classification model, achieving an overall accuracy of approximately 93.78%. The matrix compares the true class labels (Non-Hate and Hate) against the predicted outcomes. Out of all Non-

Hate samples, 68,000 were correctly classified, while 4,319 were incorrectly predicted as Hate (false positives). Similarly, for the Hate class, 68,191 samples were correctly detected, whereas 4,714 hate instances were misclassified as Non-Hate (false negatives). The strong diagonal values indicate that the model performs reliably in distinguishing hate and non-hate content, while the smaller off-diagonal values highlight remaining misclassification challenges, likely caused by subtle or context- dependent language patterns.

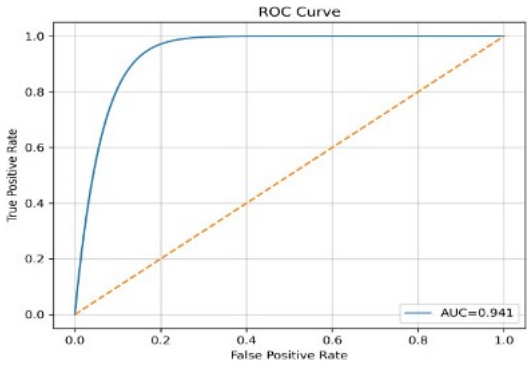

Figure 10 The **ROC (Receiver Operating Characteristic) curve** of the proposed method

This figure 10 illustrates the **ROC (Receiver Operating Characteristic) curve** for the proposed **Hybrid hate speech detection model**, showing the relationship between the **True Positive Rate (sensitivity)** and the **False Positive Rate** across different classification thresholds. The solid curve lies significantly above the diagonal dashed "Random" baseline, indicating strong discriminative capability of the proposed system in separating **Hate** and **Non-Hate** samples. The high **Area Under the Curve (AUC)** value further confirms that the model achieves reliable performance with a high true detection rate while maintaining a low false alarm rate, demonstrating its effectiveness for robust hate speech classification.

5.2 Result Evaluation Parameters

To evaluate the performance of the hate speech classification framework, standard binary classification metrics are computed using the confusion matrix. Let **TP** denote true positives (correctly predicted hate samples), **TN** denote true negatives (correctly predicted non-hate samples), **FP** denote false positives (non-hate samples incorrectly classified as hate), and **FN** denote false negatives (hate samples incorrectly classified as non-hate). These

quantities provide the basis for calculating accuracy, precision, recall, F1-score, and specificity.

(1) Accuracy

$$\text{FPR} = \frac{FP}{FP + TN}$$

(5.6)

Accuracy measures the overall correctness of the model by computing the proportion of correctly classified samples out of the total samples.

(7) False Negative Rate (FNR)

The false negative rate measures the proportion of hate

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}(5.1)$$

samples incorrectly classified as non-hate.

(2) Precision*FN*

$$\text{FNR} = \frac{FN}{FN + TP}(5.7)$$

Precision evaluates the reliability of hate speech predictions by measuring how many predicted hate samples are actually hateful.\Notation Summary

- TP: Hate correctly predicted as hate
- TN: Non-hate correctly predicted as non-hate

$$\text{Precision} = \frac{TP}{TP + FP}(5.2)$$

FP: Non-hate incorrectly predicted as hate

- FN: Hate incorrectly predicted as non-hate

(3) Recall (Sensitivity / True Positive Rate)

Recall quantifies how well the classifier identifies hateful samples by measuring the fraction of actual hate cases correctly detected.

$$TP$$

5.3 Result analysis

Table 3 Result Comparison (Existing vs Proposed)

Metho d (Existi ng / Paper)	Mode l Type	Main Features Used	Accu racy (%)	Remark s
SVM with sentime nt features [23]	Classi cal ML	TF-IDF + sentiment polarity	88.1 0	Improves over plain SVM but limited context
SVM with handcra fted features [23]	Classi cal ML	Word n-grams + lexical features	87.4 0	Cannot capture sarcasm and implicit hate
SVM + fastTex t sentime nt embedd ing [23]	ML + embed dings	fastText + sentiment vectors	89.2 0	Better represent ation but weak on coded hate
Ensemb le voting classifi er [23]	Ense mble ML	Voting across LR/SVM/ NB	90.1 0	Stable but shallow understa nding

(4) F1-Score

of non- hate samples incorrectly classified as hate.

BERT [23]	Transformer STL	Contextual subword embeddings	91.90	Strong context but misses coded/obfuscated hate
Multi-channel BERT [23]	Transformer Hybrid	Multiple contextual channels	92.80	Better fusion but heavy; still lacks char features
BETO [23]	Transformer STL	Spanish BERT	90.90	Strong for Spanish but not best for English
MTLsent+emo (paper proposed) [23]	Transformer + MTL	Hate + sentiment + emotion	93.40	Improves HS detection using auxiliary tasks
Proposed Hybrid (MTL + CharCNN + DeBERTa)	Transformer + CharCNN + MTL	DeBERTa context + coded-hate CharCNN + sentiment +emotion	93.78	Best robustness; handles sarcasm + coded hate

yields stronger robustness against implicit hate, sarcasm, and obfuscated hate patterns.

Table 4. Our Proposed Model Metrics

Metric	Value (%)
Accuracy	**93.78**
Precision (Hate)	**94.04**
Recall (Hate)	**93.53**
F1-score (Hate)	**93.79**
Specificity (Non-Hate)	**94.03**

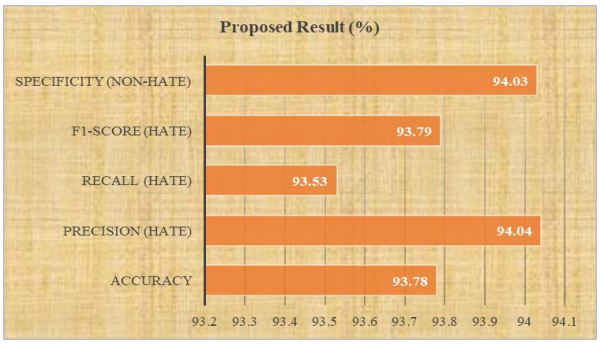

Table 3 presents a comparative evaluation of widely used hate speech detection approaches reported in the literature alongside the proposed Hybrid proposed model. The table includes classical machine learning baselines such as SVM variants, ensemble voting classifiers, and deep learning transformer models including BERT, BETO, and multi-channel BERT. The results indicate that traditional SVM-based methods achieve moderate accuracy due to their dependency on engineered features and limited contextual understanding. Transformer-based models outperform classical methods by capturing deeper semantic meaning through contextual embeddings. Furthermore, multi-task learning models such as MTLsent+emo achieve improved accuracy by jointly learning hate speech along with affective signals such as sentiment and emotion. In comparison, the proposed Hybrid-2 framework achieves the highest accuracy of 93.78%, demonstrating that the integration of DeBERTa with CharCNN and MTLFigure 11. The detailed performance metrics of the proposed Hybrid-2 model

Table 11 reports the detailed performance metrics of the proposed Hybrid-2 model derived directly from the confusion matrix results. The model achieved an overall accuracy of **93.78%**, confirming strong classification capability on the hate versus non-hate task. The precision score of **94.04%** indicates that a high proportion of predicted hate samples are correct, reducing false alarms. The recall value of **93.53%** shows that the model successfully identifies most hateful samples, minimizing missed hate speech cases. The F1-score of **93.79%** reflects a balanced trade-off between precision and recall, highlighting stable performance across both classes. In addition, the specificity of **94.03%** demonstrates the model's effectiveness in correctly identifying non-hate content. Collectively, these metrics confirm that the proposed Hybrid-2 model provides accurate and reliable hate speech detection while maintaining low misclassification rates.

6. Conclusion

In this paper, the proposed Hybrid-2 framework (MTL
+ CharCNN + DeBERTa) was evaluated on the balanced hate speech dataset. The model achieved strong classification performance with **93.78% accuracy**, high precision (**94.04%**), recall (**93.53%**),
and F1-score (**93.79%**), demonstrating reliable detection of hate and non-hate content. The confusion matrix confirmed low misclassification rates, while ROC and Precision–Recall analyses indicated robust discrimination capability. Overall, the results validate the effectiveness, stability, and generalization strength of the proposed approach.

References

[1] J. M. Pérez *et al*., "Assessing the Impact of Contextual Information in Hate Speech Detection," in *IEEE Access*, vol. 11, pp. 30575- 30590, 2023, doi: 10.1109/ACCESS.2023.3258973.

[2] J. Lu *et al*., "Hate Speech Detection via Dual Contrastive Learning," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2787-2795, 2023, doi: 10.1109/TASLP.2023.3294715.

[3] P. Sharmila, K. S. M. Anbananthen, D. Chelliah, S. Parthasarathy and S. Kannan, "PDHS: Pattern- Based Deep Hate Speech Detection With Improved Tweet Representation," in *IEEE Access*, vol. 10, pp. 105366-105376, 2022, doi: 10.1109/ACCESS.2022.3210177.

[4] M. Mozafari, R. Farahbakhsh and N. Crespi, "Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning," in *IEEE Access*, vol. 10, pp. 14880- 14896, 2022, doi: 10.1109/ACCESS.2022.3147588.

[5] A. Z. Miran and H. S. Yahia, "Hate Speech Detection in Social Media (Twitter) Using Neural Network," in *Journal of Mobile Multimedia*, vol. 19, no. 3, pp. 765-798, May 2023, doi: 10.13052/jmm1550-4646.1936.

[6] K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text," in *IEEE Access*, vol. 9, pp. 109465-109477, 2021, doi: 10.1109/ACCESS.2021.3101977.

[7] K. Maity, S. Bhattacharya, S. Saha and M. Seera, "A Deep Learning Framework for the Detection of Malay Hate Speech," in *IEEE Access*, vol. 11,

pp. 79542-79552, 2023, doi: 10.1109/ACCESS.2023.3298808.

[8] A. R. Jafari, G. Li, P. Rajapaksha, R. Farahbakhsh and N. Crespi, "Fine-Grained Emotions Influence on Implicit Hate Speech Detection," in *IEEE Access*, vol. 11, pp. 105330-105343, 2023, doi: 10.1109/ACCESS.2023.3318863.

[9] Z. Mansur, N. Omar and S. Tiun, "Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities," in *IEEE Access*, vol. 11, pp.16226-16249, 2023, doi: 10.1109/ACCESS.2023.3239375.

[10] S. Khan *et al*., "HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi- Directional Gated Recurrent Unit With Capsule Network," in *IEEE Access*, vol. 10, pp. 7881- 7894, 2022, doi: 10.1109/ACCESS.2022.3143799.

[11] M. Bilal, A. Khan, S. Jan and S. Musa, "Context- Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform," in *IEEE Access*, vol. 10, pp. 121133- 121151, 2022, doi: 10.1109/ACCESS.2022.3216375.

[12] O. Oriola and E. Kotzé, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," in *IEEE Access*, vol. 8, pp. 21496-21509, 2020, doi: 10.1109/ACCESS.2020.2968173.

[13] Y. Zhou, Y. Yang, H. Liu, X. Liu and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," in *IEEE Access*, vol. 8, pp. 128923-128929, 2020, doi: 10.1109/ACCESS.2020.3009244.

[14] A. J. Keya, M. M. Kabir, N. J. Shammey, M. F. Mridha, M. R. Islam and Y. Watanobe, "G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media," in *IEEE Access*, vol. 11, pp. 79697-79709, 2023, doi: 10.1109/ACCESS.2023.3299021.

[15] P. Kapil, G. Kumari, A. Ekbal, S. Pal, A. Chatterjee and B. N. Vinutha, "HHSD: Hindi Hate Speech Detection Leveraging Multi-Task Learning," in *IEEE Access*, vol. 11, pp. 101460- 101473, 2023, doi: 10.1109/ACCESS.2023.3312993.

[16] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," in *IEEE Access*, vol. 8, pp. 204951-204962, 2020, doi: 10.1109/ACCESS.2020.3037073.

[17] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," in *IEEE Access*, vol. 9, pp. 88364-88376, 2021, doi: 10.1109/ACCESS.2021.3089515.

[18] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López and M. T. Martín-Valdivia, "A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis," in *IEEE Access*, vol. 9, pp. 112478-112489, 2021, doi: 10.1109/ACCESS.2021.3103697.

[19] M. Z. Ali, Ehsan-Ul-Haq, S. Rauf, K. Javed and S. Hussain, "Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis," in *IEEE Access*, vol. 9, pp. 84296-84305, 2021, doi: 10.1109/ACCESS.2021.3087827.

[20] C. Baydogan and B. Alatas, "Metaheuristic Ant Lion and Moth Flame Optimization-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks," in *IEEE Access*, vol. 9, pp. 110047-110062, 2021, doi: 10.1109/ACCESS.2021.3102277.

[21] M. K. A. Aljero and N. Dimililer, "Genetic Programming Approach to Detect Hate Speech in Social Media," in *IEEE Access*, vol. 9, pp. 115115-115125, 2021, doi: 10.1109/ACCESS.2021.3104535.

[22] A. Rodriguez, Y. -L. Chen and C. Argueta, "FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis," in *IEEE Access*, vol. 10, pp. 22400- 22419, 2022, doi: 10.1109/ACCESS.2022.3151098.

[23] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "A multi-task learning approach to hate speech detection leveraging sentiment analysis," IEEE Access, vol. 9, pp. 112478-112489, 2021.