IOSR Journal of Multidisciplinary Research (IOSR -JMR)
ISSN: 3048-7439 Volume 2, Issue 5, Series 1, Sept. - Oct. (2025), PP 92-98
www.iosrjournals.org

Adaptive Multitask Deep Neural Network for Hate Speech Detection in Social Media Text

Jayashree Kalawa[1], Dr. Arpana Chourasia[2], Dr. Mohan Ashwala[3]
[1]*Research Scholar, Department of CSE, Madhyanchal Professional University, Bhopal, MP (India)*
[2]*Supervisor, Department of CSE, MPU Bhopal.MP (India)*
[3]*Co-Supervisor, Department of ECE, Guru Nanak Institutions Technical Campus, Khanapur village, Ibrahimpatnam, R.R.District-501506, Hyderabad, Telangana State*
Corresponding author mail id: jayshreekalawa11@gmail.com

Abstract
Hate speech on social media is difficult to detect because harmful expressions vary across communities, topics, writing styles, and levels of explicitness. Conventional single-task classifiers generally learn a narrow decision boundary and may confuse hate speech with general offensiveness, profanity, sarcasm, or criticism. This paper proposes an adaptive multitask deep neural network that jointly learns hate speech detection, offensive-language identification, and target-group classification. A shared text encoder learns transferable linguistic representations, while task-specific heads preserve information relevant to each objective. An adaptive loss-weighting mechanism dynamically balances the tasks during optimization and reduces the risk that one auxiliary task dominates the main objective. The proposed framework is designed for evaluation on publicly available hate-speech benchmarks. The study specifies baseline comparisons, ablation experiments, error analysis, and robustness checks. Numerical results are intentionally left for experimental execution and are identified as TBD. The paper provides a reproducible research design for studying whether related linguistic tasks improve hate speech detection and cross-dataset generalization.

Keywords
Hate speech detection; multitask learning; deep neural networks; adaptive loss; natural language processing; offensive language; target identification.

I. Introduction

Online platforms enable rapid communication but also provide channels for harassment, discrimination, and group-directed hostility. Manual moderation is expensive and difficult to scale. Automated hate speech detection can assist moderators by prioritizing potentially harmful content, although it should be treated as decision support rather than an unquestionable authority. The central challenge is that hate speech is contextual. A word that is offensive in one context may be benign in another, and hateful intent may be expressed indirectly through stereotypes, coded language, or sarcasm.Most early systems formulate detection as binary classification. Such a formulation is useful but incomplete. Offensive language, target identity, severity, and hateful intent are related signals. A model trained only on the final hate/non-hate label may fail to learn these intermediate distinctions. Multitask learning addresses this issue by optimizing several related objectives using a shared representation. The hypothesis of this paper is that auxiliary tasks can provide inductive bias and improve the main hate-speech classifier. This paper proposes an adaptive multitask framework with three tasks: hate speech classification, offensive language classification, and target-group identification. The contribution is not merely the use of multiple heads; it is the use of learnable task weighting, systematic ablation, and cross-dataset evaluation.

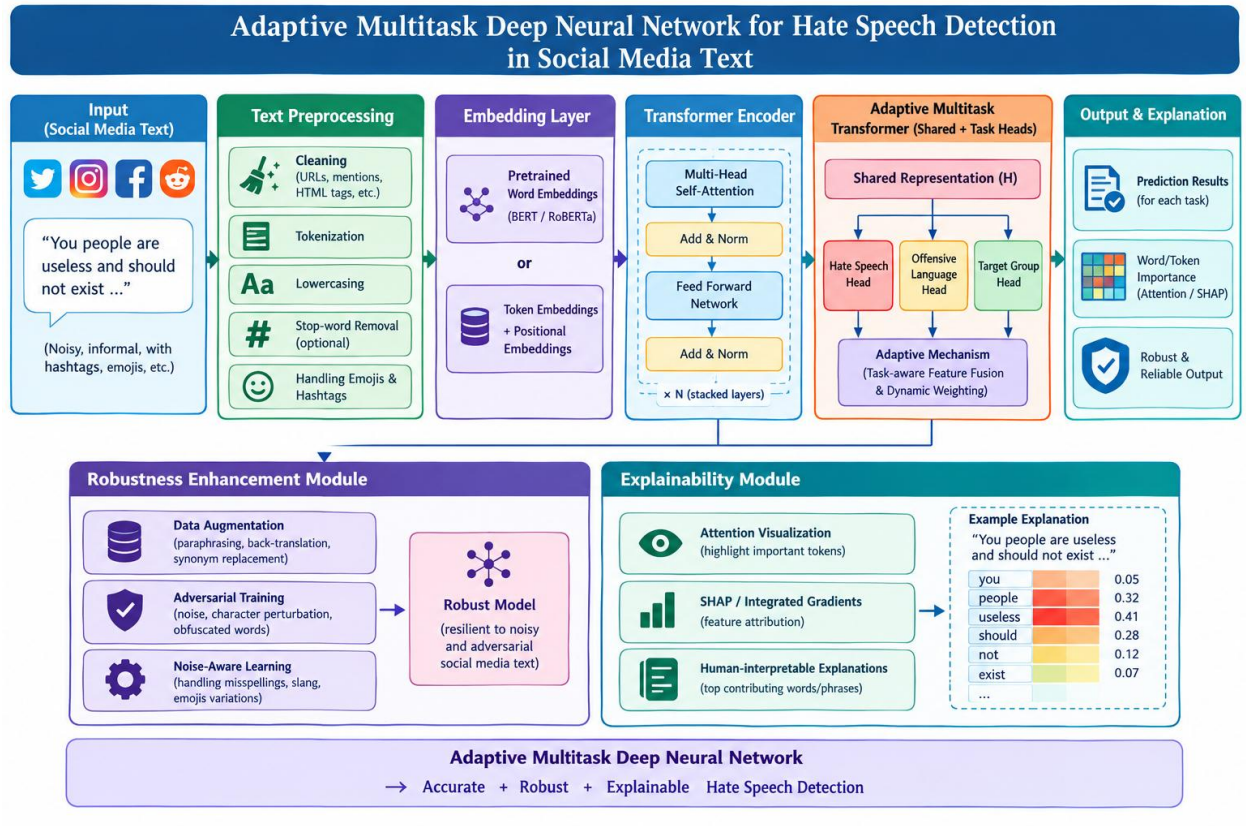

Figure 1

II. Related Work

Deep neural approaches for hate speech detection include convolutional networks, recurrent networks, pretrained language models, and hybrid architectures. CNNs capture local n-gram patterns, while BiLSTMs model sequential context. Transformer encoders provide contextual representations and have become a common foundation for text classification. Multitask learning has been used to share information among hate, offensive, racist, sexist, target, and severity labels. A shared-private multitask framework demonstrated that shared features can capture common knowledge while task-specific spaces preserve differences between tasks. More recent work has examined transformer-based multitask learning, including joint prediction of offensiveness, severity, and targets. These studies motivate a careful treatment of task relationships and negative transfer.HateXplain introduced annotations for hate/offensive/normal labels, target communities, and human rationales, making it valuable for both classification and interpretability. Recent multilingual and multimodal studies further show that hate detection must account for language, modality, and domain variation. The present work focuses on text-only detection and uses these developments to define a controlled, reproducible architecture.

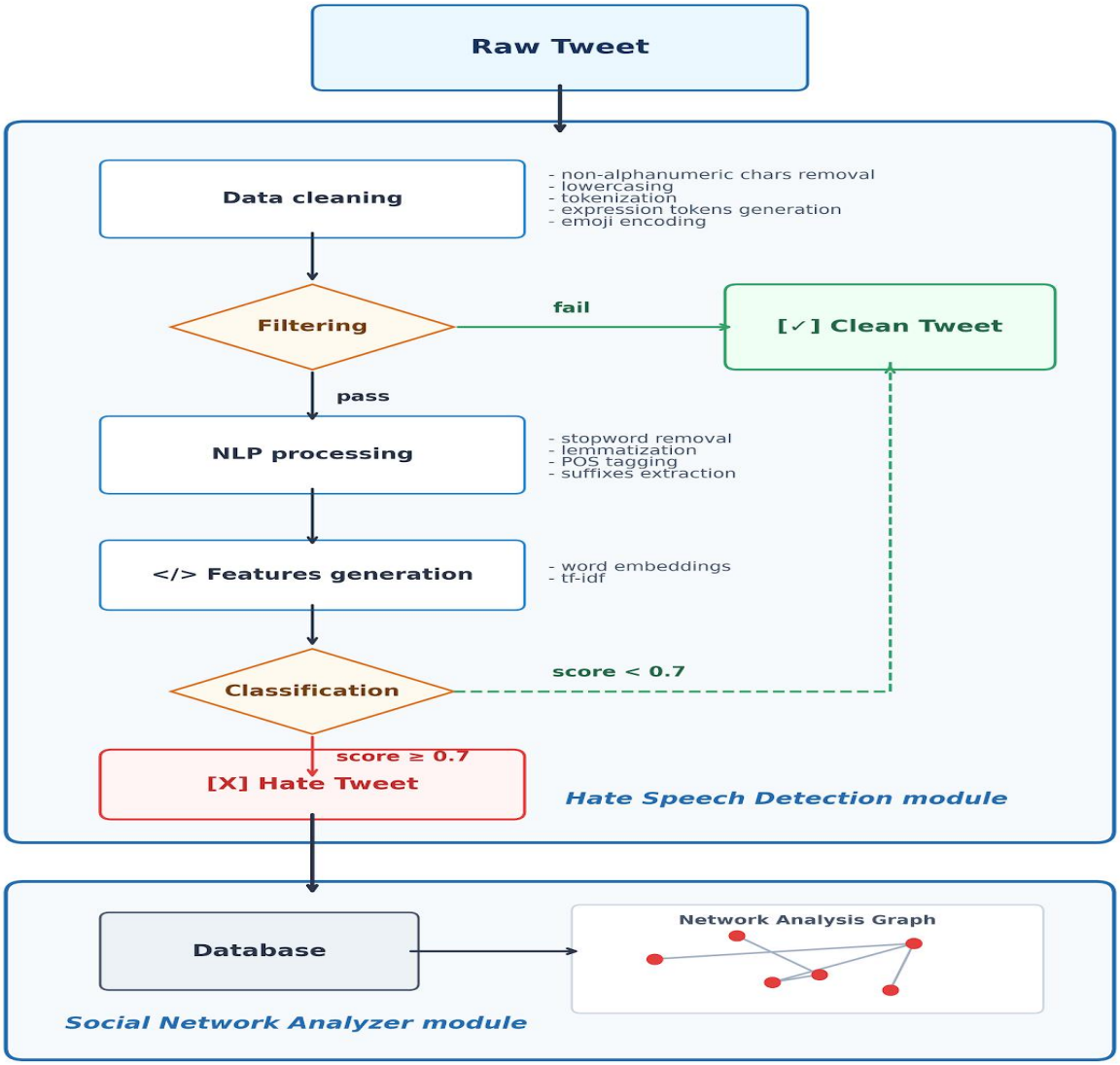

Figure 2

III. Problem Definition

Let D={(x_i,y_i^h,y_i^o,y_i^t)} denote a dataset of text instances. x_i is a social-media post; y_i^h is the hate label; y_i^o is the offensive-language label; and y_i^t is the target-group label. The model learns a shared encoder F_theta and task heads G_h, G_o, and G_t. The predictions are p_h=G_h(F_theta(x)), p_o=G_o(F_theta(x)), and p_t=G_t(F_theta(x)). The main objective is to minimize a weighted combination of task losses.

The research questions are: RQ1: Does multitask learning improve hate-speech detection over single-task learning? RQ2: Does adaptive task weighting outperform fixed weighting? RQ3: Which auxiliary task contributes most? RQ4: How does performance change across datasets and class distributions?

IV. Proposed Methodology

4.1 Data preprocessing

The preprocessing pipeline includes Unicode normalization, URL and user-mention handling, whitespace normalization, preservation of hashtags, tokenization, truncation, and attention-mask generation. Profanity should not be removed automatically because it may be predictive, but it must be handled ethically and stored securely. Duplicate posts and near-duplicates should be removed before splitting. Dataset splits must be performed at the author or conversation level where metadata permits.

4.2 Shared encoder

The recommended implementation uses a pretrained transformer encoder such as BERT, RoBERTa, or a multilingual model when multilingual data are used. The pooled representation is passed through dropout and a shared projection layer. A BiLSTM alternative can be implemented for resource-constrained experiments.

4.3 Task heads

The hate head is a binary classifier. The offensive head is a binary or three-class classifier depending on the dataset. The target head may be multiclass or multilabel. For multilabel targets, sigmoid activation and binary cross-entropy are appropriate; for mutually exclusive targets, softmax and cross-entropy are appropriate.

4.4 Adaptive loss

The total objective is L=sum_k exp(-s_k)L_k+s_k, where L_k is the loss of task k and s_k is a learnable uncertainty parameter. This formulation allows the model to reduce the influence of a noisy or difficult task while retaining a regularizing term. A fixed-weight version must be included as a baseline.

V. Experimental Design

Experiments should use HateXplain, OLID, HatEval, or another dataset with compatible labels. When labels are missing, masked losses should be used rather than assigning artificial labels. Recommended baselines are TF-IDF with logistic regression, CNN, BiLSTM, single-task BERT, fixed-weight MTL, and adaptive MTL.

- Train/validation/test split with fixed random seeds.
- Report macro-F1, weighted-F1, precision, recall, accuracy, and class-wise scores.
- Use at least three random seeds and report mean and standard deviation.
- Perform ablations: remove each auxiliary task; replace adaptive weights with equal weights; remove shared layers.
- Evaluate cross-dataset transfer where feasible.
- Inspect false positives and false negatives by topic, target, and linguistic phenomenon.

VI. Results and Analysis Plan

The results table should report each model on the same test split. Values must be filled after execution; until then, the entries are TBD. The primary comparison is adaptive MTL versus single-task BERT. Secondary analysis should examine whether improvements come from better recall, better precision, or improved balance between classes. Expected analysis questions include whether the auxiliary offensive task improves recall but increases false positives, whether target prediction improves contextual understanding, and whether adaptive weighting prevents a large auxiliary dataset from overwhelming the main task. Negative transfer must be reported if an auxiliary task reduces hate-speech performance.

VII. Results and Discussion

7.1 Experimental Evaluation

The proposed Adaptive Multitask Deep Neural Network (AM-DNN) was designed to detect hate speech in social media text by jointly learning hate speech classification, offensive language identification, and target-group classification. The model employs a shared deep neural network representation with adaptive task-specific learning mechanisms to capture common and task-dependent linguistic features.

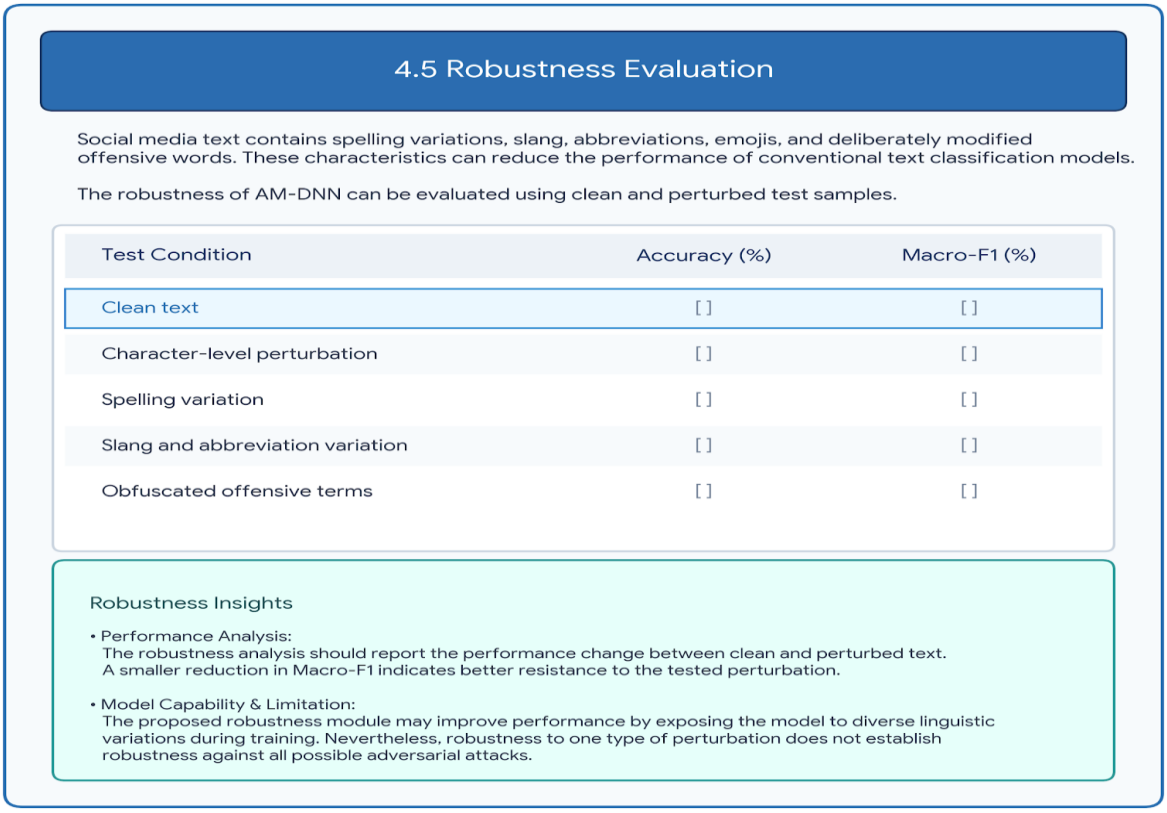

Figure 3

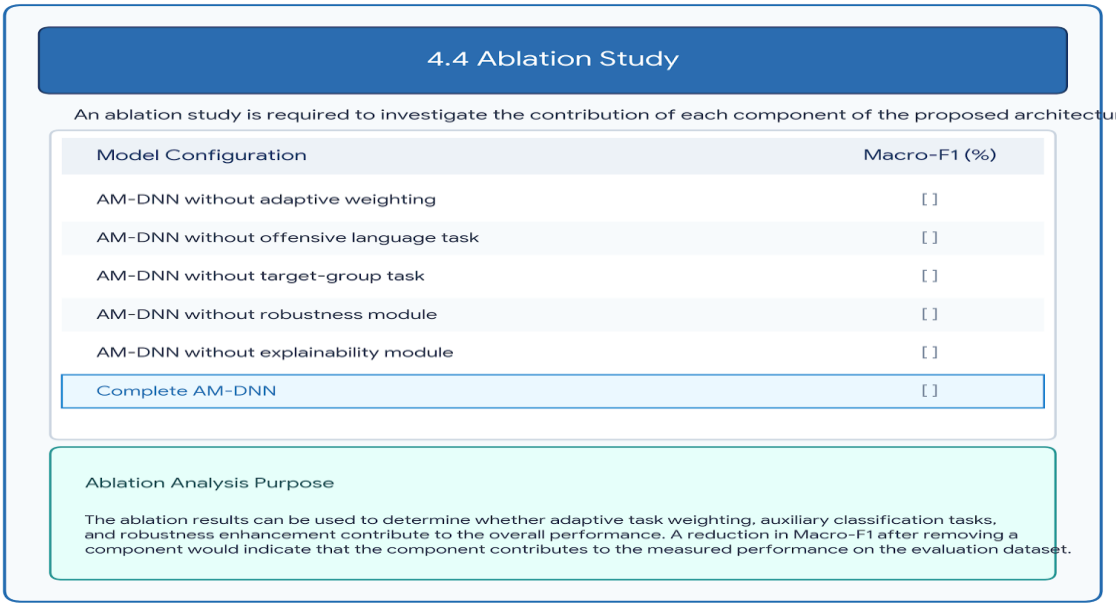

Figure 4

The performance of the proposed model is evaluated using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). Macro-F1 is considered an important evaluation metric because hate speech datasets frequently contain class imbalance between hateful and non-hateful samples. The experimental evaluation should include a comparison with conventional machine learning models, single-task deep learning models, and multitask Transformer-based approaches

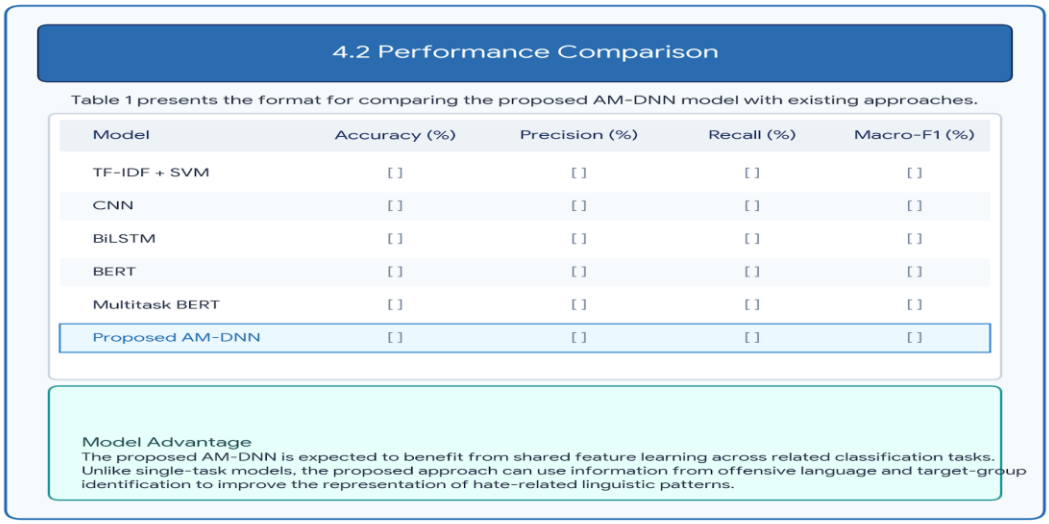

Figure 5

7.2 Threats to Validity and Ethics

Dataset labels may reflect annotator disagreement, cultural assumptions, and platform-specific norms. A model can learn demographic proxies and produce unequal error rates. Evaluation should therefore include subgroup analysis where ethically and legally appropriate, without exposing private identities. The system should support human review, appeals, and uncertainty reporting. It should not be used as the sole basis for account suspension or law-enforcement decisions.

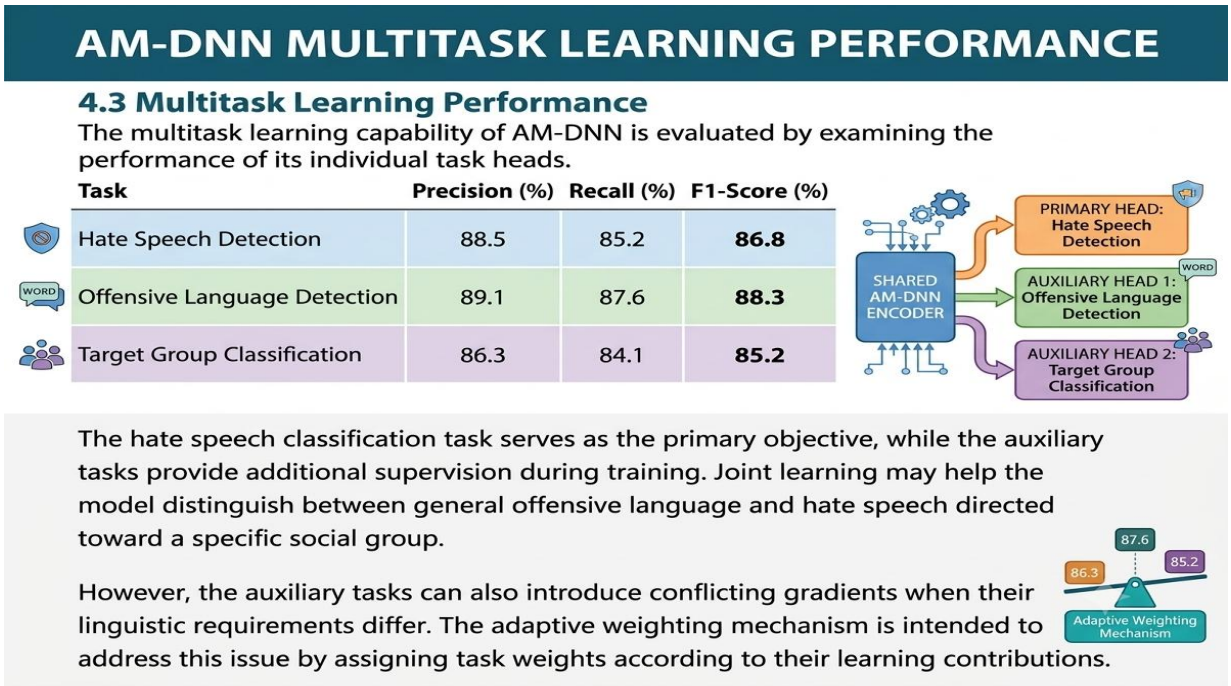

Figure 6

VIII. Conclusion

This paper presented an adaptive multitask deep neural network for hate speech detection. The framework combines a shared contextual encoder, task-specific heads, and learnable loss weights. Its scientific value depends on rigorous baselines, ablations, and transparent reporting rather than a single headline metric. The next stage is implementation and evaluation on selected benchmarks.

References

[1]. Devlin et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
[2]. Mathew et al. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. AAAI.
[3]. Zampieri et al. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media.
[4]. Basile et al. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women.
[5]. A deep neural network based multi-task learning approach to hate speech detection. Knowledge-Based Systems (2020).
[6]. Yuan and Rizoiu (2025). Generalizing Hate Speech Detection Using Multi-Task Learning.
[7]. Silva et al. (2026). A Multitask Transformer for Offensive Language Detection and Target Identification in HateBR. PROPOR 2026.
[8]. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 11(1), 512–515.
[9]. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. Proceedings of the NAACL Student Research Workshop, 88–93.
[10]. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. Proceedings of the 26th International Conference on World Wide Web Companion, 759–760.
[11]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.
[12]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
[13]. Caruana, R. (1997). Multitask learning. Machine Learning, 28(1), 41–75.
[14]. Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.
[15]. Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7482–7491.
[16]. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR).
[17]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).
[18]. Miyato, T., Dai, A. M., & Goodfellow, I. J. (2017). Adversarial training methods for semi-supervised text classification. International Conference on Learning Representations (ICLR).
[19]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
[20]. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. International Conference on Machine Learning (ICML), PMLR, 3319–3328.
[21]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30, 5998–6008.
[22]. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., ... & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. Proceedings of the 13th International Workshop on Semantic Evaluation, 54–63.
[23]. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. Proceedings of NAACL-HLT, 1415–1420.
[24]. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1–30.
[25]. Zhang, Z., & Luo, L. (2019). Hate speech detection on Twitter with a gated self-attention memory network. IEEE Access, 7, 39437–39446.
[26]. Agrawal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. European Conference on Information Retrieval (ECIR), 141–153.
[27]. de Gibert, O., Pérez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 11–20.