

Decoding Fake Narratives in Spreading Hateful Stories: A Dual-Head RoBERTa Model with Multi-Task Learning

Yash Bhaskar¹

IIIT Hyderabad

yash.bhaskar@research.iiit.ac.in

Sankalp Bahad¹

IIIT Hyderabad

sankalp.bahad@research.iiit.ac.in

Parameswari Krishnamurthy²

IIIT Hyderabad

param.krishna@iiit.ac.in

Abstract

Social media platforms, while enabling global connectivity, have become hubs for the rapid spread of harmful content, including hate speech and fake narratives (Davidson et al., 2017; Shu et al., 2017). The Faux-Hate shared task focuses on detecting a specific phenomenon: the generation of hate speech driven by fake narratives, termed Faux-Hate. Participants are challenged to identify such instances in code-mixed Hindi-English social media text. This paper describes our system developed for the shared task, addressing two primary sub-tasks: (a) Binary Faux-Hate detection, involving fake and hate speech classification, and (b) Target and Severity prediction, categorizing the intended target and severity of hateful content. Our approach combines advanced natural language processing techniques with domain-specific pretraining to enhance performance across both tasks. The system achieved competitive results, demonstrating the efficacy of leveraging multi-task learning for this complex problem.

1 Introduction

Social media has revolutionized communication, providing unprecedented connectivity across the globe. This increased connectivity, however, has also inadvertently fostered the rapid dissemination of harmful content, including the troubling combination of hate speech and fabricated narratives. Hate speech, particularly when intertwined with falsehoods, exacerbates its detrimental impact, fueling discrimination, violence, and societal unrest.

In response to this growing concern, the Faux-Hate shared task (Biradar et al., 2024a), based on a phenomenon recently characterized and dataset curated by Biradar et al. (Biradar et al., 2024b), introduces a unique challenge: identifying and categorizing instances of hate speech generated through fake narratives in code-mixed Hindi-English text. This task emphasizes the importance of detecting

and analyzing content that misleads and provokes through a combination of misinformation and hateful language. Researchers and practitioners have increasingly turned their attention to understanding and combating these complex phenomena.

The shared task comprises two sub-tasks: Task A focuses on binary classification of fake and hate labels, while Task B involves predicting the target and severity of hateful content. This paper describes our system, methodologies, and experimental results for both sub-tasks, contributing to the broader effort to address hate speech and fake narratives in multilingual, code-mixed contexts.

2 Related Work

The detection of hate speech and misinformation on social media has been a prominent area of research within natural language processing (NLP). Studies have extensively explored techniques for identifying hate speech across various languages and platforms (Warner and Hirschberg, 2012), often leveraging machine learning and deep learning approaches. Recent advancements include transformer-based models like BERT (Devlin et al., 2019), RoBERTa, and multilingual BERT (mBERT), which have shown significant success in text classification tasks, including hate speech detection.

Fake news and misinformation detection have similarly gained attention (Zubiaga et al., 2018), with methods ranging from linguistic feature analysis to neural network-based classification. The intersection of hate speech and fake narratives, however, remains a relatively unexplored domain, particularly in code-mixed languages like Hindi-English. Prior work in code-mixed text processing has highlighted the challenges posed by non-standard grammar, orthographic variations, and the lack of annotated datasets.

This shared task builds on these research threads, offering a novel opportunity to investigate Faux-

Hate in a multilingual and culturally nuanced context. Our approach draws inspiration from prior work in hate speech and fake news detection while tailoring solutions to the unique challenges of the code-mixed Hindi-English dataset provided in this task.

3 Methodology

This section outlines the architecture, components, and training methodology of the dual-head RoBERTa model developed for the Faux-Hate shared task. Our system leverages RoBERTa-base (Liu et al., 2019) as the backbone encoder and extends it with a dual-head classification mechanism for simultaneous hate speech and fake news detection. The architecture adopts a multi-task learning approach (Caruana, 1997), enabling the model to effectively share information across tasks while maintaining task-specific parameterization through dedicated classification heads. The code for our system, including implementation details and pre-trained models, is available at our GitHub repository: <https://github.com/yash9439/ICON-Faux-Hate-Shared-Task>.

3.1 Base Architecture

The proposed model is built upon RoBERTa-base, a transformer-based pre-trained language model renowned for its effectiveness in natural language understanding tasks. RoBERTa-base serves as the backbone encoder, processing input text and providing contextualized representations.

3.1.1 Base Encoder

The RoBERTa-base encoder processes input sequences using the following steps:

- **Input Representation:** The input text is tokenized using the RoBERTa tokenizer, and positional embeddings are added. The resulting embeddings are passed through the transformer layers.
- **Hidden States:** The encoder maintains the original configuration of hidden states and extracts the representation of the [CLS] token for downstream classification tasks.
- **Dropout Regularization:** Dropout is applied to the pooled [CLS] representation, with the probability inherited from the RoBERTa-base configuration to reduce overfitting.

3.2 Dual-Head Classification System

The model implements two parallel classification heads, one for hate speech detection and the other for fake news detection. Each classification head adopts a sophisticated multi-layer architecture designed to handle the complexity of the respective tasks.

3.2.1 Classification Head Architecture

The classification heads share the same architecture but maintain independent parameters to allow task-specific learning. Each head is composed of the following layers:

- **Input Layer:** A linear transformation maps the RoBERTa hidden size (768) to a custom hidden size (768) for further processing.
- **Intermediate Layers:**
 - *Layer Normalization:* Applied after the input layer to improve training stability and convergence.
 - *GELU Activation:* Ensures smooth activation and enhances non-linearity in the model.
 - *Dropout Regularization:* A dropout layer with a probability of 0.2 is used to mitigate overfitting.
 - *Dimensionality Reduction:* A linear layer reduces the feature dimensions from 768 to 384, followed by a second layer normalization and GELU activation.
 - *Reduced Dropout:* Another dropout layer with a lower probability (0.1) is applied for regularization.
- **Output Layer:** A final linear transformation maps the 384-dimensional features to the number of output classes (binary classification for task A and 4 class classification for task B).

3.2.2 Additional Features

The classification heads incorporate the following additional features:

- **Residual Connections:** While not enabled in the current configuration, residual connections can be added to facilitate gradient flow and improve training.
- **Shared Dropout Layer:** A shared dropout layer is applied to the pooled RoBERTa output before feeding it into the classification heads.

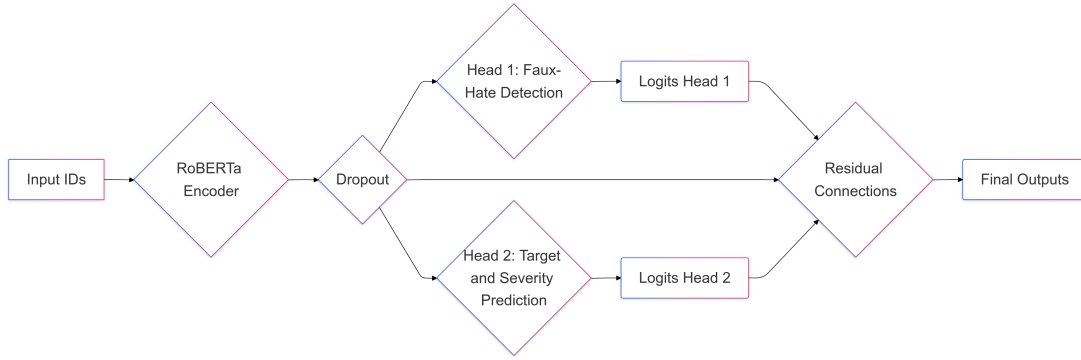


Figure 1: Model architecture

- **Independent Loss Computation:** Each head computes its own task-specific loss, which are combined through averaging for multi-task learning.

3.3 Key Innovations

The proposed architecture incorporates several innovations to enhance performance:

- **Multi-Layer Classification Heads:** The use of progressive dimensionality reduction in the classification heads enables efficient feature extraction and task-specific learning.
- **Dual Regularization Strategy:** The combination of dropout layers and layer normalization reduces overfitting and stabilizes training.
- **Residual Connections:** While disabled in the current configuration, these connections provide the potential to improve gradient flow in future experiments.
- **Balanced Loss Computation:** Independent loss computation and balanced averaging ensure that both tasks are treated equally during training.

4 Experiments and Results

This section outlines the experimental setup, training procedure, evaluation metrics, and the results obtained for both tasks in the Faux-Hate shared task. Additionally, we analyze the impact of architectural variations, specifically the inclusion and exclusion of residual connections in the classification heads.

4.1 Experimental Setup

4.1.1 Training and Evaluation

The experiments were conducted on the provided training and validation datasets for both Task A (Bi-

nary Faux-Hate Detection) and Task B (Target and Severity Prediction). The training process for each task spanned six epochs, with the model evaluated after every epoch.

We implemented two variants of the dual-head RoBERTa model:

- **Run 1:** Model with residual connections in the classification heads.
- **Run 2:** Model without residual connections.

4.1.2 Evaluation Metrics

The models were evaluated using the following:

- **Accuracy:** For each classification head, measuring performance in binary and categorical classification tasks.
- **Loss:** Validation loss for both tasks to monitor overfitting and convergence.
- **Overall Accuracy:** Average of the two heads for task A and task B.

4.2 Results

Table 1 presents the results for both Task A and Task B.

Variant	Task	Test Set F1 Score
With Residual Connection	Task A	0.76
	Task B	0.56
Without Residual Connection	Task A	0.73
	Task B	0.54

Table 1: Comparison of Task A and Task B results with and without residual connections.

5 Analysis

For Task A, which involved binary classification of Faux-Hate instances into hate speech or fake content, we evaluated our model’s performance based on standard classification metrics. The results, as shown in Table 1, demonstrate that the model effectively learned the underlying patterns that distinguish between fake and hate speech. Specifically, the variant with residual connections achieved a test set F1 score of 0.76, outperforming the variant without residual connections (0.73). This indicates that residual connections helped the model better capture subtle distinctions in the data. However, we observed slightly higher false positives in some cases, which could be attributed to the inherent challenges of the dataset, such as the overlapping features of fake narratives and hate speech.

For Task B, which focused on a multiclass classification task involving the prediction of the target and severity of hateful content, the model performed admirably despite the complexity of the task. As summarized in Table 1, the F1 scores for the test set were 0.56 and 0.54 for the variants with and without residual connections, respectively. The slight improvement with residual connections highlights their role in enhancing the model’s ability to generalize across the greater diversity of content within each class. This task posed additional challenges due to the varied nature of the input, but the multitask learning approach enabled the model to leverage shared knowledge between the two tasks, which likely contributed to its strong performance.

The analysis of these results underscores the importance of task-specific fine-tuning, particularly for the classification heads, in achieving high performance. The shared model architecture also proved beneficial in efficiently utilizing training data, thereby improving outcomes for both tasks. Future work could focus on refining fine-tuning strategies and incorporating more diverse datasets to enhance the model’s robustness in real-world applications.

6 Conclusion

We presented a dual-head RoBERTa model for the Faux-Hate shared task, addressing both binary classification of fake and hate speech (Task A) and multiclass classification of target and severity (Task B). Our system achieved competitive results, demonstrating the effectiveness of multitask learning in handling the complexities of code-mixed Hindi-

English text. The model showed strong performance on both tasks, and future work will focus on refining the model with additional data and fine-tuning techniques to further improve its accuracy.

References

- Shankar Biradar, Sai Kartheek Reddy Kasu, Sunil Saumya, and Md. Shad Akhtar, editors. 2024a. *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.
- Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2024b. Faux hate: unravelling the web of fake narratives in spreading hateful stories: a multi-label and multi-class dataset in cross-lingual hindi-english code-mixed text. *Language Resources and Evaluation*, pages 1–32.
- Rich Caruana. 1997. Multitask learning. Technical report, Technical report, Carnegie Mellon University.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.
- William Warner and Julia Hirschberg. 2012. Challenges in detecting hate speech on the world wide web. In *Proceedings of the NAACL-HLT 2012 Workshop on Language in Social Media*, pages 11–20.
- Arkaitz Zubiaga, Maria Liakata, Rob Procter, Guo Wei Hoi, and Peter Tolmie. 2018. Rumour detection on social media: A critical review. *ACM Computing Surveys (CSUR)*, 51(2):1–36.