Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)

Shankar Biradar[1], **Kasu Sai Kartheek Reddy**[2], **Sunil Saumya**[2], **Md. Shad Akhtar**[3]
[1] KLE Tech DR. M.S.S College Belagavi , [2] IIIT Dharwad, [3] IIIT Delhi.
{shankar,sunil.saumya}@iiitdwd.ac.in
saikartheekreddykasu@gmail.com
shad.akhtar@iiitd.ac.in

Abstract

The rapid expansion of social media has led to an increase in code-mixed content, presenting significant challenges in the effective detection of hate speech and fake narratives. To advance research in this area, a shared task titled "Decoding Fake Narratives in Spreading Hateful Stories" (Faux-Hate) was organized as part of ICON 2024. This paper introduces a multi-task learning model designed to classify Hindi-English code-mixed tweets into two distinct categories: hate speech and false content. The proposed framework utilizes fastText embeddings to create a shared feature space that adeptly captures the semantic and syntactic intricacies of code-mixed text, including transliterated terms and out-of-vocabulary words. These shared embeddings are then processed through two independent Support Vector Machine (SVM) classifiers, each specifically tailored for one of the classification tasks. Our team, secured 10th place among the participating teams, as evaluated by the organizers based on Macro F1 scores.

1 Introduction

The rise of social media platforms has revolutionized the way people communicate, but it has also introduced significant challenges, particularly in terms of the spread of harmful and toxic content. One of the most concerning issues emerging in recent times is Faux Hate, a novel form of hate speech that arises from the intersection of fake narratives and hate speech. In many instances, individuals unknowingly express hateful opinions based on fabricated claims that have gained traction online. This dangerous combination not only fuels further divisiveness but also perpetuates the spread of false information, creating an environment where harmful rhetoric can thrive unchecked. Faux Hate is a growing issue, particularly in the context of online communities where misinformation circulates rapidly, exacerbating the spread of hatred.

The challenge with Faux Hate lies in its complexity and the difficulty of identifying it in many cases. While traditional hate speech detection models have made significant strides[cite], Faux Hate presents an additional layer of complexity. Hateful comments fueled by false claims can often appear indistinguishable from genuine hate speech, especially if the underlying fake narrative is not widely recognized or acknowledged. This makes detection particularly difficult, as it requires not only identifying the harmful speech but also understanding the falsehood that triggers it. In cases where the fake claim is not obvious or well-known, identifying Faux Hate can be a formidable task, even for sophisticated automated systems.

What sets Faux Hate apart from conventional hate speech is the need for a deeper understanding of the context in which the hate is generated. Hate speech that stems from misinformation or fake claims does not exist in isolation—it is intricately linked to the narratives that are propagated alongside it. Recognizing faux hate, therefore, requires both an awareness of the falsity of the claim and an understanding of how this claim may fuel or amplify hate speech. This intersection of fake narratives and harmful rhetoric represents a novel challenge in the field of online content moderation and demands a tailored approach to detection that goes beyond the capabilities of traditional hate speech classification systems.

To address this emerging problem, the shared task on Faux Hate detection was introduced as part of ICON 2024, providing a unique opportunity for the research community to develop and evaluate methodologies specifically aimed at identifying Faux Hate. As a part of shared task we have released FEUD[1] Data set comprising of 8000 fake-instigated hateful comments in Hindi-English code-mixed text includes real-world examples of online discourse, where participants were challenged to

[1] **F**aux hat**E** m**U**lti-label **D**ataset

distinguish between genuine hate speech and Faux Hate generated by fake narratives using multi-task learning. This task encourages researchers to explore innovative approaches that not only identify harmful speech but also detect the underlying false claims that give rise to it.

The significance of this task lies in its potential to advance the state of the art in hate speech detection. By focusing on Faux Hate, the shared task draws attention to the need for more nuanced models that can understand the complexities of online discourse. In this paper, we present the results of the competition, analyzing the performance of different methodologies submitted by participants.

The structure of the paper is as follows: Section 2 provides an overview of the label Taxonomy. Sections 4 and 5 describe the dataset used in the shared task, outlining the task setup and the evaluation metrics employed. Section 6 presents a comprehensive analysis of the methodologies used by the participants. Section 7 discusses the findings of the shared task.

2 Label Taxonomy

The FEUD data set is annotated using a two-level hierarchical structure. The annotation process begins by categorising social media comments into fake and non-fake labels. Subsequently, the fake and non-fake content undergo a further labelling process to determine the presence of hate content, forming the binary class labelling stage. In the second phase, the focus shifts exclusively to hate labels, enabling more fine-grained multi-class labelling for target and severity annotation. These designated labels are crucial in understanding hateful content's complex nature, facilitating a more comprehensive and insightful analysis.

The employed taxonomy for annotating the proposed data set is as follows:

Fake: Fake label denotes comments deliberately crafted to spread misinformation with the potential to mislead readers (Allcott and Gentzkow, 2017). Comments bearing false content are assigned the fake label '1', while others receive '0'.

Hate: Hate label refers to comments that target and marginalize individuals or communities based on attributes such as religion, physical appearance, skin colour, ethnicity, and political opinion (Chowdhury et al.). Authors label comments with hateful content as '1' and those without as '0'.

Target Target label represents the subject of interest in hate posts. Based on the target audience, hate speech is further categorized into three sub-classes.

- **Individual:** Involves hate directed at specific individuals, including politicians, celebrities, ordinary individuals, or industrialists. The presence of hate toward an individual is labelled as 'I'.
- **Organisation :** Encompasses hate aimed at groups of people united by a common goal. In this context, "groups" refers to collectives such as organizations, companies, or any body of individuals who are associated by a shared purpose or affiliation. For example, hate speech directed at a company often targets not just the entity itself but also the people working within it. Organisational hate is annotated with an 'O'.
- **Religion:** Encompasses derogatory references to religions like Hinduism, Islam, and Christianity. Such comments are designated with 'R'.

Severity Determining the severity level is essential to ensure freedom of expression while appropriately addressing hate speech. In the proposed data set, each hateful comment is classified into one of three severity sub-classes.

- **Low Severity:** Includes comments expressing disagreement with ideas, responding to challenging claims, or attempting to alter the target's viewpoint using non-violent means. These comments are tagged with 'L'[2].
- **Medium Severity:** Encompasses comments targeting individuals or groups with insulting language like 'thief' or 'stupid' without provoking violence. Such comments are labelled as 'M'[2].
- **High Severity:** Relates to comments advocating violence, incitement, or targeting specific religious beliefs. These are identified as 'H'[2] .

3 Tasks

We proposed to include the following sub-tasks as part of this shared task.

- **Task A – Binary Faux-Hate Detection** Participants will receive a dataset containing text samples, each labeled with:

[2]https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/

Attribute	Train Set	Validation Set	Test Set
Number of Records			
Total Records	6398	800	800
Hate Labels			
Non-Hate (0)	2277	275	309
Hate (1)	4090	522	488
Fake Labels			
Non-Fake (0)	3063	390	391
Fake (1)	3305	407	406
Target Classes			
I (Individual)	1080	137	137
O (Organization)	2271	297	259
R (Religion)	745	89	93
Severity Classes			
L (Low)	1964	254	253
M (Medium)	1555	199	165
H (High)	578	70	71

Table 1: Dataset Analysis for Train, Validation, and Test Sets

  - **Fake:** Binary label indicating if the content is fake (1) or real (0).
  - **Hate:** Binary label indicating if the content is hate speech (1) or not (0).

  The objective of this sub-task is to develop a single model that outputs both the fake and hate labels for each text sample.

- **Task B - Target and Severity prediction:** Participants will receive a dataset containing text samples, each labeled with:
  - **Target:** Categorical label indicating the target of the content (Individual(I), Organization(O), and Religion(R)).
  - **Severity:** Categorical label indicating the Severity of the content (Low(L), Medium(M), and High(H)).

  The objective of this sub-task is to develop a single model that generate both the Target and Severity labels for given text sample.

4 Data & Resources

We gathered data from multiple social media platforms, including Twitter and YouTube, covering various topics such as Religion, Sports, Health, Politics, Finance, and Entertainment to ensure data diversity. Initially, fake narratives were sourced from reputable fact-checking websites like Alt-News, Boomlive, and Factly. These fake narratives were then used as search queries to extract hateful reactions. The data was annotated using crowdsourcing. Further details on data annotation and the annotation guidelines are provided in (Biradar et al., 2024).

The detailed description of the dataset can be found in Table 1

Rank	TEAM	Macro F 1 Score for Task A
1	**DCST_unigoa**	**0.79**
2	**Radicaldecoders run1**	**0.7761**
3	**chakravyuh coders run1**	**0.7721**
4	Tensor_Text	0.772
5	Keyboardwarriors run1	0.76
6	MUCS run1	0.7589
7	Rejected_cookies	0.7557
8	Radicaldecoders run2	0.7522
9	NOVA-RMK-ADS	0.7479
10	VTU_BGM	0.7445
11	Keyboardwarriors run2	0.73
12	MUCS run2	0.7026
13	Vector_Visionaries	0.6803
14	CNLP-NITS-PP run1	0.65
15	RMK_Mithra	0.5232
16	CNLP-NITS-PP run1	0.51

Table 2: Rank list based on Macro F1 score for Task A

5 Evaluation Parameter

The proposed FEUD corpus is highly imbalanced, especially for Task B. To address this, we used the Macro F1 score as an evaluation parameter, which gives equal importance to all class labels. This approach is beneficial for assessing the performance of minority labels without prioritizing the majority label. By opting for the Macro F1 score, we ensure that all class labels are given equal weight in the evaluation.

6 System Description

This section describes the systems submitted for the shared task. A total of 22 teams from various institutes registered for the shared task. Out of these, 12 teams participated in Task A, which focuses on binary classification, and 10 teams participated in Task B, which involves more fine-grained classification with multi-task learning. Several teams submitted more than one run during the competition. This section briefly discusses the methodologies used in each submitted model.

The team DCST_unigoa (Fondekar et al., 2024) participated in both Task A and Task B. They used HingRoBERTa, a pre-trained transformer fine-tuned on Hindi-English code-mixed text with a class weightage technique, securing first place in both tasks. Team RMK_Mithra (Murugan et al., 2024; Yadawad et al., 2024) used TF-IDF-based features with AutoML models in their proposed

RANK	TEAM	Macro F 1 Score for Task B
1	**DCST_unigoa**	**0.6155**
2	**NOVA-RMK-ADS**	**0.6048**
3	**Radicaldecoders run1**	**0.5947**
4	Rejected_cookies	0.5926
5	Tensor_Text	0.5887
6	MUCS run1	0.5746
7	CNLP-NITS-PP run1	0.57
8	Keyboardwarriors run1	0.56
9	Radicaldecoders run2	0.5416
10	Keyboardwarriors run2	0.54
11	RMK_Mithra	0.4818
12	MUCS run2	0.4359
13	chakravyuh coders run1	0.13

Table 3: Rank list based on Macro F1 score for Task B

work, achieving 15^{th} place in Task A and 11^{th} place in Task B.

The team KeyboardWarriors (Bhaskar and Bahad, 2024) employed a dual head attention weight with a Roberta model, securing 5th and 8th place in Task A and Task B, respectively. Additionally, the team CNLP-NITS-PP (Vetagiri and Pakray, 2024) developed a Conv-LSTM network with single-task learning to address the Faux-Hate issue, securing 14th place in Task A and 7th place in Task B.

The team MUCS (Hegde and Shashirekha, 2024) used two distinct models for Task A and Task B. They proposed the Hing_MTL model for Task A and the Ensemble_MTL model for Task B, securing 6th place in both tasks. Team Rejected Cookies (Joel and Srivastava, 2024) used a BERT and Hybrid Quantum Neural Network-based method for Faux-Hate detection, achieving 7th place in Task A and 4th place in Task B.

Team Tensortext (Gopal et al., 2024) used an XLM-RoBERTa-based ensemble network with multi-task learning, securing 4th place in Task A and 5th place in Task B. Team Radicaldecoders (Hitesh et al., 2024) employed a hard parameter shared XLM-RoBERTa and HateBERT-based multi-task learning model, securing 2nd place in Task A and 3rd place in Task B. Team Vector_Visionaries (Onajol et al., 2024) developed an attention-weight-tuned LoRA adopter-based model for building task-specific classification heads, securing 13th place in Task A. Lastly, team VTU_BGM (Kavatagi et al., 2024) used a FastText-based multi-task learning approach, securing 10th place in Task A.

7 Discussion

Among the submitted teams, DCST_unigoa secured the first rank in both Task A and Task B, achieving a macro F1 score of 0.6155. They used a pre-trained HingRoBERTa model, trained on Hindi-English text, to address the Faux-hate issue. Key observations from the shared task include that domain-specific and language-agnostic pre-trained models achieved better results than other models. Additionally, most models failed to achieve better results for the more fine-grained classification in Task B. Among the submitted teams, the popular choice was a hard parameter shared encoder model, with a task-specific classification layer on top to solve this multi-task learning problem. The leaderboard for both Task A and Task B is illustrated in Table 2 and 3.

8 Conclusion

This paper provides an overview of the models submitted to our first shared task on Faux-hate identification, part of the ICON 2024 conference. The shared task addresses the critical issue of Faux-hate content detection in Hindi-English code-mixed text through multi-task learning. A total of 15 teams participated in Task A, and 13 teams participated in Task B. This work aims to promote research in this critical category of hate speech, which is particularly harmful.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31.

Yash Bhaskar and Sankalp Bahad, editors. 2024. *Decoding Fake Narratives in Spreading Hateful Stories: A Dual-Head RoBERTa Model with Multi-Task Learning*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2024. 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.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Ashweta Fondekar, Milind Shivolkar, and Jyoti Pawar, editors. 2024. *Unpacking Faux-Hate: Addressing Faux-Hate Detection and Severity Prediction in Code-Mixed Hinglish Text with HingRoBERTa and Class Weighting Techniques*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Sunil Gopal, Shreyas Srinivas, R Sushanth, and C B Abhilash, editors. 2024. *Faux-Hate Multitask Framework for Misinformation and Hate Speech Detection in Code-Mixed Languages*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Asha Hegde and H L Shashirekha, editors. 2024. *Transformer-driven Multi-task Learning for Fake and Hateful Content Detection*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

N P Hitesh, D Ankith, A N Poornachandra, and C B Abhilash, editors. 2024. *Multi-Task Learning for Faux-Hate Detection in Hindi-English Code-Mixed Text*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Joy Joel and Naman Srivastava, editors. 2024. *Rejected Cookies @ Decoding Faux-Hate: Predicting Fake Narratives and Hateful Content*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Sanjana Kavatagi, Rashmi Rachh, and Prakul Hiremath, editors. 2024. *Shared Feature-Based Multitask Model for Faux-Hate Classification in Code-Mixed Text*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Shanthi Murugan, R Arthi, E Boomika, S Jeyanth, and S Kaviyarasu, editors. 2024. *Challenges and Insights in Identifying Hate Speech and Fake News on Social Media*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Abhinandan Onajol, Varun Gani, Praneeta Marakatti, Bhakti Malwankar, and Shankar Biradar, editors. 2024. *LoRA adapter weight tuning with multi-task learning for Faux-Hate detection*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Advaitha Vetagiri and Partha Pakray, editors. 2024. *Detecting Hate Speech and Fake Narratives in Code-Mixed Hinglish Social Media Text*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

R Yadawad, Sunil Saumya, K Nivedh, Siddhaling Padanur, and Sudev Basti, editors. 2024. *A Machine Learning Framework for Detecting Hate Speech and Fake Narratives in Hindi-English Tweets*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.