



ICON 2024

**Shared Task on Decoding Fake Narratives in Spreading
Hateful Stories (Faux-Hate)**

Proceedings of the Share task

December 19-22, 2024

©2024 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN

INTRODUCTION

Social media has undeniably transformed how people communicate; however, it also brings significant drawbacks, particularly the proliferation of fake and hateful comments. Recent observations indicate that these issues frequently coexist, with discussions on hate topics often dominated by fake narratives. Therefore, it is imperative to explore the role of fake narratives in disseminating hate in contemporary times.

In this context, we propose the task of detecting the role of fake narratives in spreading hateful stories, termed "Faux-Hate." We introduce a novel dataset, **FEUD**, comprising 8,000 fake-instigated hateful comments in Hindi-English code-mixed text. Existing datasets face two primary challenges. Firstly, they have not established a link between fake and hate content within the context of Hindi-English code-mixed text. Secondly, many datasets focus solely on binary hate or non-hate classification, neglecting the more intricate aspects of identifying the target audience and the societal impact. Our proposed dataset addresses these limitations by constructing a multiclass and multi-label dataset, incorporating target and severity level detection for hate content within the FEUD corpus. The data is sourced from both YouTube and Twitter, reducing inherent biases from a single platform.

Moreover, the proposed corpus is built from instances of fake narratives that emerged during emergencies, triggering waves of hateful responses. These narratives were sourced from reputable fact-checking websites such as AltNews, Boomlive, FactChecker, and FACTLY, covering various topics including religion, sports, health, politics, finance, and entertainment. Subsequently, these fake narratives were used as search queries to scrape hateful reactions from Twitter and YouTube, resulting in the comprehensive FEUD corpus.

Tasks

We propose 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:

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).

This sub-task objective aims 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)).

This sub-task aims to develop a single model that generates both the Target and Severity labels for a given text sample.

Participation

A total of 22 teams from various institutes registered for the shared task. Twelve teams participated in Task A, which focuses on binary classification, and ten 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.

Organizing Committee

Shankar Biradar

Kasu Sai Kartheek Reddy

Sunil Saumya

Md. Shad Akhtar

Table of Contents

<i>Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)</i> Shankar Biradar, Kasu Sai Kartheek Reddy, Sunil Saumya and Md. Shad Akhtar	1
<i>Unpacking Faux-Hate: Addressing Faux-Hate Detection and Severity Prediction in Code-Mixed Hinglish Text with HingRoBERTa and Class Weighting Techniques</i> Ashweta A. Fondekar, Milind M. Shivolkar and Dr. Jyoti D. Pawar.....	6
<i>Decoding Fake Narratives in Spreading Hateful Stories: A Dual-Head RoBERTa Model with Multi-Task Learning</i> Yash Bhaskar and Sankalp Bahad	12
<i>Challenges and Insights in Identifying Hate Speech and Fake News on Social Media</i> Shanthi Murugan, Arthi R, Boomika E, Jeyanth S and Kaviyarasu S.....	16
<i>Detecting Hate Speech and Fake Narratives in Code-Mixed Hinglish Social Media Text</i> Advaita Vetagiri and Partha Pakray	22
<i>Transformer-driven Multi-task Learning for Fake and Hateful Content Detection</i> Asha Hegdea and H L Shashirekhab	29
<i>Rejected Cookies @ Decoding Faux-Hate: Predicting Fake Narratives and Hateful Content</i> Joel D Joy and Naman Srivastava	36
<i>A Machine Learning Framework for Detecting Hate Speech and Fake Narratives in Hindi-English Tweets</i> R.N.Yadawad, Sunil Saumya, K.N.Nivedh, Siddhaling S. Padanur and Sudev Basti	40
<i>Faux-Hate Multitask Framework for Misinformation and Hate Speech Detection in Code-Mixed Languages</i> Sunil Gopal C V, Sudhan S, Shreyas Gutti Srinivas, Sushanth R and Abhilash C B	45
<i>Multi-Task Learning for Faux-Hate Detection in Hindi-English Code-Mixed Text</i> Hitesh N P, D Ankith, Poornachandra A N and Abhilash C B	50
<i>LoRA adapter weight tuning with multi-task learning for Faux-Hate detection</i> Abhinandan Onajol, Varun Gani, Praneeta Marakatti, Bhakti Malwankar, and Shankar Biradar ..	56
<i>Shared Feature-Based Multitask Model for Faux-Hate Classification in Code-Mixed Text</i> Sanjana Kavatagi, Rashmi Rachh and Prakul Hiremath	61