ICON 2024

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

Proceedings of the Share task

December 19-22, 2024

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