Shri Dharmasthala Manjunatheshwara College of Engineering & Technology, Dharwad

&

Indian Institute of Information Technology, Dharwad

A Machine Learning Framework for Detecting Hate Speech and Fake Narratives in Hindi-English Tweets

Dr. R. N. Yadawad, Dr. Sunil Saumya,
K. N. Nivedh, Siddhaling S. Padanur, Sudev Basti

Agenda

- **The Problem -** Harmful Content Online
- **Our Approach -**Tackling the Challenge
- **Dataset -** Building the Foundation
- **Methodology -** Unpacking the Process
- **Models & Techniques -** Exploring Options
- **Results -** Key Findings
- **Leaderboard Ranking -** Performance
- **Applications -** Real-World Impact
- **Ethics -** Responsible Deployment

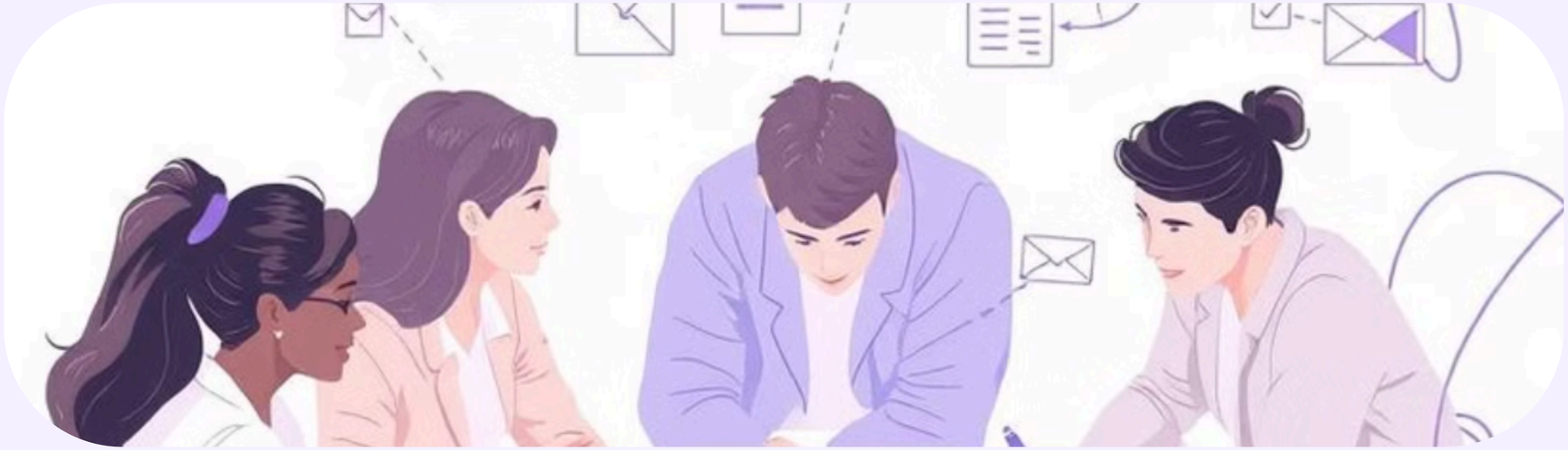
The Problem: Harmful Content Online

Rising Hate Speech

Social media platforms are increasingly plagued by harmful content.

Code-Mixed Challenges

Analyzing Hindi-English mixed tweets poses several difficulties.



Tackling the Challenge: Our Approach

1 Detection System

Develop a robust system to detect hate speech and fake narratives.

2 Target Identification

Identify the targets of hate speech and classify its severity.

3 Code-Mixed Data

Focus on analyzing Hindi-English code-mixed social media data.

Dataset: Building the Foundation

Faux-Hate Shared Task Dataset

6,397 tweets for training,
801 for validation &
801 for testing.

Task Breakdown

Task A: Binary detection (hate speech or not).
Task B: Target and severity prediction.

Methodology: Unpacking the Process

1

Preprocessing

Remove URLs, hashtags, mentions, and special characters. Tokenize and remove stop words.

2

Feature Extraction

TF-IDF vectorization with unigrams and bigrams. Additional feature: Text length.

3

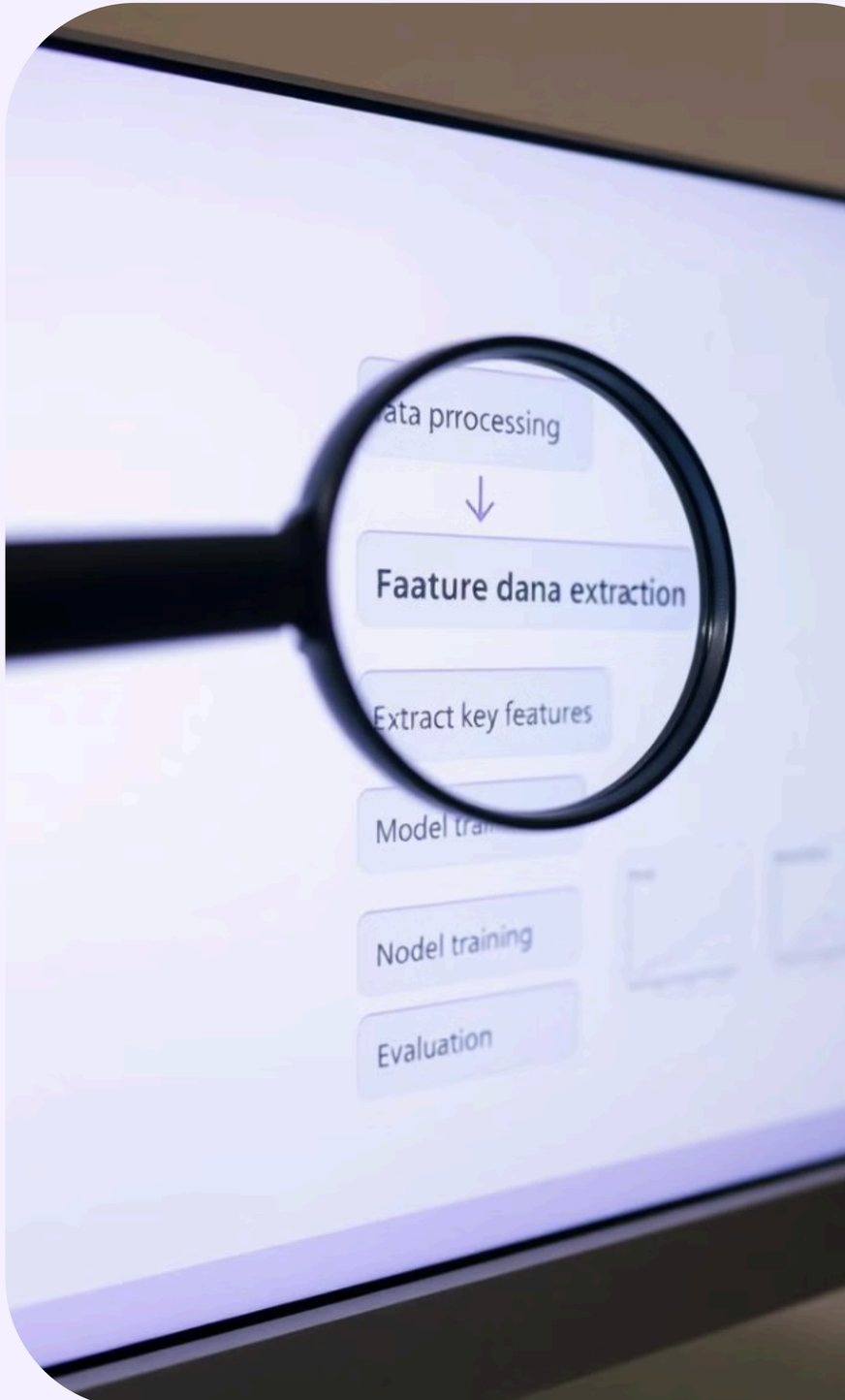
Model Training

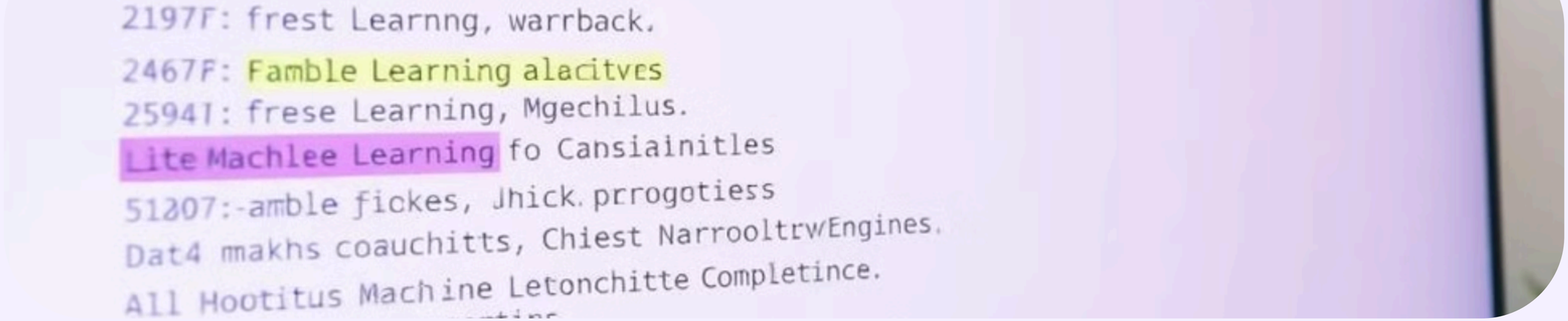
Train an FFNN for classification using SMOTE to handle class imbalance.

4

Evaluation

Assess performance metrics like accuracy and F1-score.





Models and Techniques: Exploring Options

Random Forest

Our primary classifier for this task.

Baselines

Logistic Regression and SVM for comparison.

Advanced Models

Gradient Boosting,XGBoost,LightGBM Feedforward Neural Networks networks.

Results: Unveiling the Findings

Untiret	Lient	Midulne	Sunary	Becore	Faccoce
Pacliast.	303	12.84	7.21	11.03	7.32S
11 Lntringes	345	12.76		12.91	3.991
51. Ercstestly	490	17.35		11.21	1.594
32 Arrcloady	360	38.54		14.59	7.3%
91. Ponersily F2 Pecterstly	215 450	34.24 37.76		13.33 1.71	2.374 1.27%
57. Seceroley	500	30.06		13.31	1.5%

75.75%

Accuracy

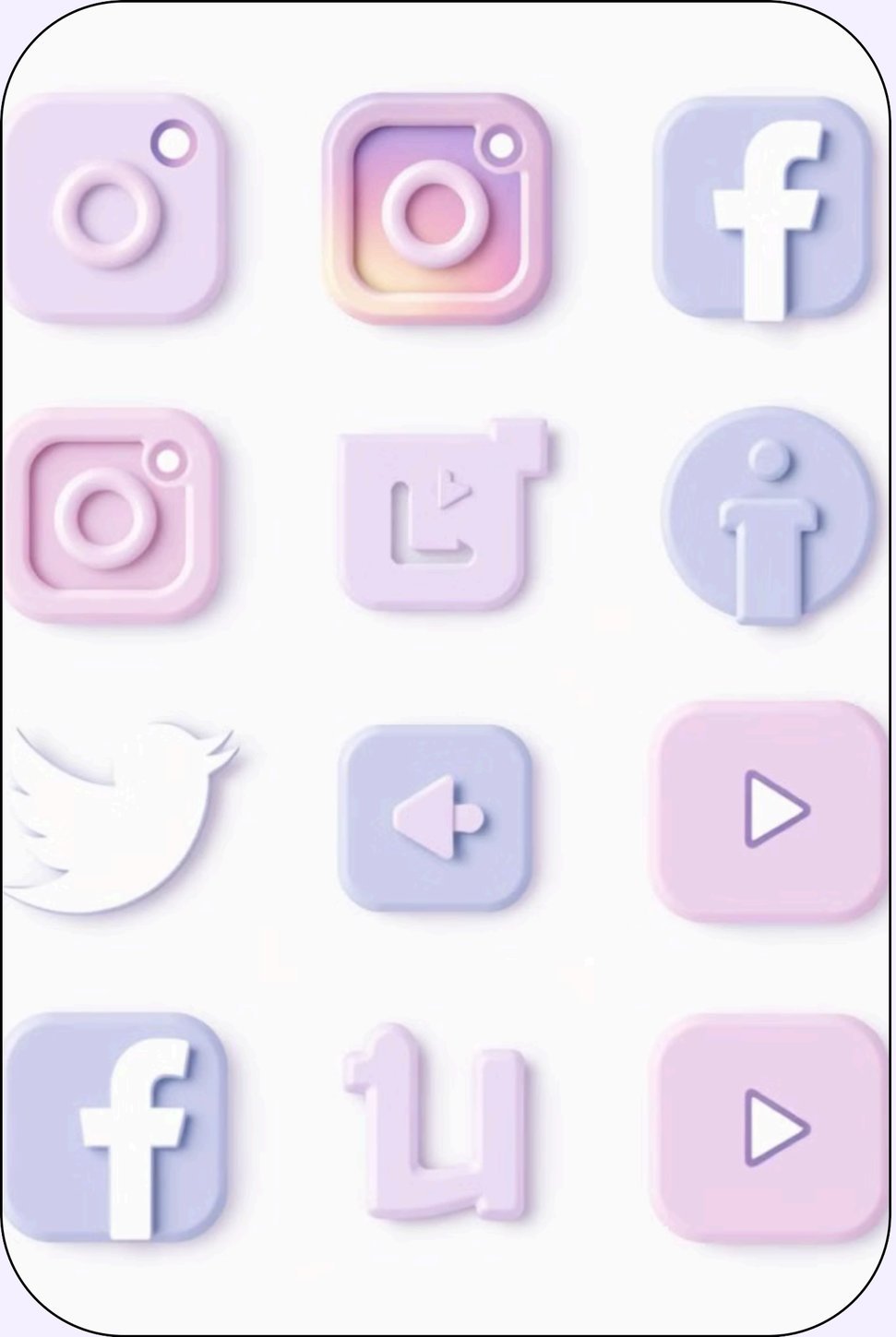
77.21%

Macro F1-Score

Leader Board Ranking for Task A

Team	F1-Score	Rank
DCST Unigoa	0.79	1
Radicaldecoders run1	0.7761	2
Chakravyuh coders run1	**0.7721**	**3**

TASK A OF THE **ICON 2024** SHARED TASK COMPETITION EARNED US A **COMMENDABLE 3RD PLACE.**



Real-World Applications: Impact and Potential

Online Moderation

Moderating online platforms to remove harmful content.

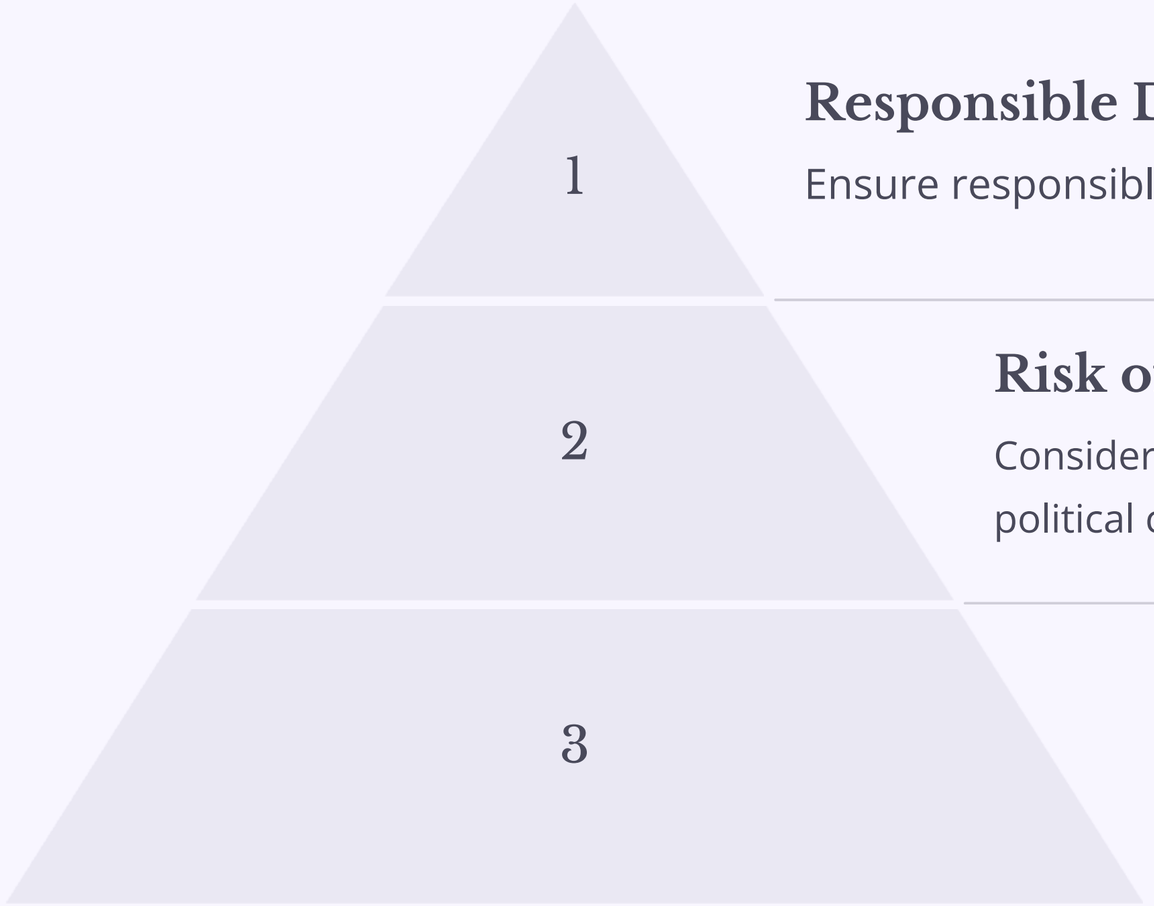
Social Media Monitoring

Real-time monitoring of social media for hate speech detection.

Automated Detection

Automated detection of harmful narratives in code-mixed data.

Ethical Considerations: Responsible Deployment



Responsible Deployment

Ensure responsible use of our system, preventing bias and misuse.

Risk of Misuse

Consider potential misuse in sensitive domains, like elections or political campaigns.

Ethical Use

Promote guidelines for ethical use of our system in the future.

CONCLUSION & FUTURE WORK

- Our approach provided the solution to the TASK A very efficiently.
- We used TF-IDF, SVM, Random Forest and Neural Networks to propose the solution to the TASK A.
- However, we could not succeed in providing efficient solution to TASK B, which we want to carry out further.

REFERENCES

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2024b. Faux hate: unravelling the web of fake nar- ratives 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.

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of ma- chine learning research, 6(11).

Galen Andrew and Jianfeng Gao. 2007. Scalable training of 1 1-regularized log-linear models. In Proceed- ings of the 24th international conference on Machine learning, pages 33-40.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collabora- tion and Computing for Under-Resourced Languages (CCURL), pages 202-210, Marseille, France. Euro- pean Language Resources association

Thank you