Rohit Beniwal

ProxiMix -A Prototype-Driven Proximity-Aware Meta-Embedding Framework for Faux-Hate Detection

Research Papers

Document Details

Submission ID

trn:oid:::27535:144424112

Submission Date

Jun 28, 2026, 6:08 PM GMT+5:30

Download Date

Jun 28, 2026, 6:12 PM GMT+5:30

File Name

ProxiMix -A Prototype-Driven Proximity-Aware Meta-Embedding Framework for Faux-Hate Detec....pdf

File Size

919.8 KB

40 Pages

10,027 Words

59,021 Characters

5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 14 words)

Match Groups

14 Not Cited or Quoted 5%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
1% Publications
4% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **14** Not Cited or Quoted 5%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
1% Publications
4% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	de.overleaf.com	1%
2	Internet	www.dspace.dtu.ac.in:8080	<1%
3	Student papers	University of Wollongong on 2023-12-15	<1%
4	Student papers	Delhi Technological University on 2026-05-29	<1%
5	Student papers	Delhi Technological University on 2026-05-21	<1%
6	Student papers	Delhi Technological University on 2026-05-26	<1%
7	Student papers	Delhi Technological University on 2026-05-18	<1%
8	Internet	d-nb.info	<1%
9	Internet	webmail.hsbi.de	<1%
10	Internet	dspace.cvut.cz	<1%

11 Internet

openaccess.thecvf.com <1%

12 Student papers

Liverpool John Moores University on 2026-02-15 <1%

13 Internet

amlankar.github.io <1%

ProxiMix: A Prototype-Driven Proximity-Aware Meta-Embedding Framework for Faux-Hate Detection

A THESIS REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
ARTIFICIAL INTELLIGENCE

Submitted by

POOJA YADAV(24/AFI/21)

Under the supervision of

Dr. ROHIT BENIWAL

DEPARTMANT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

JUNE, 2026

DEPARTMENT OF MECHANICAL ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **POOJA YADAV** (Roll No–24/AFI/21) student of M.Tech (Artificial Intelligence), hereby declare that the project Dissertation titled **"ProxiMix: A Prototype-Driven Proximity-Aware Meta-Embedding Framework for Faux-Hate Detection"** which is submitted by us to the Computer Science & Engineering Department, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of **Master of Technology** in the period of Aug 2024 to May 2026 under supervision of **Dr. Rohit Beniwal**, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi **Pooja Yadav**

Date: 29.05.26 M.Tech (24/AFI/21)

DEPARTMENT OF MECHANICAL ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"ProxiMix: A Prototype-Driven Proximity-Aware Meta-Embedding Framework for Faux-Hate Detection"** which is submitted by **Pooja Yadav**, (Roll No- 24/AFI/21), Computer Science & Engineering department, Delhi Technological University, Delhi in partial fulfillment of the requirements for the award of the degree of **Master of Technology in Artificial Intelligence**, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

SUPERVISOR

Dr. Rohit Beniwal
Computer Science & Engineering
Delhi Technological University
Delhi - 110042

Place: Delhi

Date: 29.05.2026

DEPARTMENT OF MECHANICAL ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express our sincerest gratitude to Dr. **Dr. Rohit Beniwal** for his continuous guidance and mentorship that he provided us during the project. He showed us the path to achieve our targets by outlining all the tasks to be done and explaining to me the importance of this project and its industrial relevance. He was always ready to help us and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

I would also like to extend my heartfelt thanks to Ms. Shristi for her valuable insights and constructive feedback. Her thoughtful suggestions played a crucial role in refining the knowledge and direction of this work.

I would also like to express my deepest gratitude to my family and friend, Sujyoti Jha, for their unwavering support and belief in me throughout the journey. This encouragement gave me the strength to persevere during challenging times.

Place: Delhi **Pooja Yadav**

Date: 29.05.2026 24/AFI/21

Abstract

Faux-Hate detection in code-mixed Hinglish text is a complex multilingual classification task due to the noisy social media language, transliteration variation, semantic ambiguity and the large class imbalance. In this work, we introduce a prototype-based proximity-aware meta-embedding framework for multi-class fake news and hate speech detection in low-resource code-mixed scenarios, dubbed ProxiMix. The proposed model is a fine-tuning of XLM-RoBERTa's encoder to learn powerful cross-lingual contextual representations, followed by a learnable prototype bank constructed using K-Means clustering on the auxiliary corpora. Such prototypes represent semantic regions of the class and enable the model to learn more complex decision boundaries than regular fixed embeddings. ProxiMix introduces cosine similarity, multi-head cross-attention, and feature fusion, which enables the representation learning of prototype-level proximity information to be fused with the contextual CLS embedding to further refine the representation learning. The framework also tackles imbalance and optimization instability by applying focal loss, label smoothing, prototype regularization, Monte Carlo dropout and isotonic calibration in five-fold cross-validation. Experimental results conducted on the Faux-Hate corpus show the mean Macro F1 is 87.00% for fake detection and 87.61% for hate detection with a standard deviation of 0.0064 for the folds, demonstrating good consistency and robustness. Additionally, the performance difference between the two tasks is smaller in the model, suggesting that the prototype-aware semantic fusion approach leads to better generalization under low-resource and noisy multilingual NLP tasks.

Contents

Candidate's Declaration i

Certificate ii

Acknowledgement iii

Abstract iv

Content vi

List of Tables vii

List of Figures viii

List of Symbols, Abbreviations ix

1 INTRODUCTION 1

2 LITERATURE REVIEW 3
- 2.1 Background: Evolution of NLP Models . . . 3
- 2.2 Hate Speech Detection in Code-Mixed Languages . . . 4
- 2.3 Detection of Fake News in Low-Resource Languages . . . 5
- 2.4 Faux-Hate Detection and Multitask Learning . . . 5

3 METHODOLOGY 7
- 3.1 Dataset Description . . . 7
  - 3.1.1 Faux-Hate Corpus . . . 7
  - 3.1.2 Auxiliary Corpora . . . 8
  - 3.1.3 Dataset Statistics . . . 8
- 3.2 Data Pre-processing . . . 8
- 3.3 Proposed Model . . . 9
  - 3.3.1 Cross-Lingual Encoder . . . 9
  - 3.3.2 Learnable Prototype Centroids with cosine similarity . . . 10
  - 3.3.3 Prototype Cross-Attention Fusion . . . 11
  - 3.3.4 Feature Fusion and Task Heads . . . 11
  - 3.3.5 Training Strategy and Optimization . . . 12
  - 3.3.6 Hyperparameter Configuration . . . 14

4 STATISTICAL EVALUATION 15
- 4.1 Descriptive Statistical Analysis . . . 15
- 4.2 Fold-Level Variance Analysis . . . 16

4.3 Prototype Alignment Analysis 16
4.4 Threshold Calibration Analysis 17
4.5 Precision-Recall Analysis 17
4.6 Monte Carlo Dropout Uncertainty Analysis 18

5 RESULT AND DISCUSSION 19
5.1 Overall Performance 19
5.2 Confusion Matrix Analysis 20
5.3 Comparison with State-of-the-Art Methods 21
5.4 Ablation Study 22
5.5 Shared Task Leaderboard Analysis 23
5.6 Discussion 23

6 CONCLUSION AND FUTURE SCOPE 25

List of Publications 27

References 28

List of Tables

3.1 Dataset Statistics Overview 9
3.2 Component Configuration of Model 12
3.3 Hyperparameter Configuration of ProxiMix 14

4.1 Descriptive Statistics of Macro F1 Scores 15

5.1 Per-fold Macro F1 scores for Fake and Hate detection heads 20
5.2 Comparative summary across key evaluation dimensions 21
5.3 Ablation Study: Component-wise Contribution to ProxiMix Performance . 22

List of Figures

1.1 Faux-hate 2

2.1 Evolution of Faux-Hate detection architectures (2019–2025) 3

3.1 ProxiMix Framework for Faux-hate Classification. 10

4.1 Fold-Level Stability Analysis 16
4.2 Prototype alignment behavior 17
4.3 Macro F1 vs decision threshold per fold. 17
4.4 Precision-Recall curves for Fake and Hate detection 18
4.5 MC Dropout prediction variance distribution for correct vs. incorrect predictions 18

5.1 Confusion matrices of the best-performing fold for Fake News Detection and Hate Speech Detection. 20
5.2 Performance Benchmarking of ProxiMix against Existing Shared-Task Systems 23

List of Symbols

h	CLS token representation from XLM-RoBERTa encoder
c	L2-normalized CLS embedding
$\tilde{C}$	Learnable prototype centroid matrix
C_0	Initial prototype centroid matrix from K-Means
s	Prototype affinity vector
$\hat{p}$	Squared softmax normalized prototype distribution
a	Cross-attention output vector
$\tilde{c}$	Refined CLS embedding after cross-attention fusion
α	Learnable residual scale parameter
τ	Temperature scaling parameter
ε	Label smoothing factor
γ	Focal loss focusing parameter
σ	Sigmoid activation function

Chapter 1

INTRODUCTION

Social media has become one of the most influential communication platforms in the modern digital era, enabling rapid information exchange across geographical and linguistic boundaries. The positive aspects of these platforms are the openness that makes them strong, but it is the same openness that can lead to harmful content. Fake news and hate speech are among the most pressing concerns, as they can be a source of misinformation, opinion manipulation and escalation of social conflict. The definition of fake news is mostly false or misleading information that is shared as though it were true, and is often created with the aim of being deceptive, manipulative, or provoking [1],[2]. Hate speech, however, is defined as language that attacks, insults and dehumanizes persons or groups on the basis of their identity attributes, including religion, caste, ethnicity, gender or nationality [3],[5] In the past, these two problems have been studied separately as they occur in different communication patterns and have different types of analysis [6],[7].

With the passage of time, it became apparent that the proliferation of fake news and hate speech is not always separate and distinct. In real online conversations, misinformation can reinforce or reinforce hateful narratives and hate speech can be contained within false or misleading claims. For instance, a fictional narrative can be created to appeal to a community, or hateful rhetoric can help make a false narrative more believable [1]. This overlap leads to an intersection of misinformation and hostility in the same text, a phenomenon we label "misinformation and hostility." As this type of content has emerged, the concept of Faux-Hate has been created, a combined task that encompasses the intertwined nature of fake news and hate speech in social media postings [2].

It is very difficult to detect Faux-Hate in low-resource multilingual and code-mixed environments like Hinglish as Hindi and English languages are used in the same sentence [3, 4]. This is a very noisy social media text, with spelling variation, transliteration, abbreviations, emojis, informal grammar and inconsistent tokenization [5]. These features hinder the learning of stable and discriminative representations on standard machine learning models and even standard transformer-based models [6, 7]. Previous methods for detecting hate speech and fake news tended to be either single-task based, shallow embeddings, or fixed contextual encoders [8,9]. These approaches were helpful in obtaining a baseline, but they were not always able to capture the finer semantic relations between texts or to fully model the nuances of the fake/hateful boundary [10,11]. Furthermore, the small amount of data available in Faux-Hate makes overfitting and unstable performance even more likely [12].

In order to mitigate these shortcomings, this work proposes a prototype-driven proximity-aware meta-embedding framework for joint Faux-Hate detection, called "ProxiMix". The central theme is to go beyond the encoded version of simple instances and add prototype-

level semantic guidance instead. The model is trained with XLM-RoBERTa [13] as a multilingual encoder to get strong contextual representation from code-mixed text. These embeddings are then compared to a trainable prototype bank built by using auxiliary corpora and K-Means clustering [14]. Multiple prototypes are used to represent class-specific semantic regions, which results in more complex structure than a single global centroid or a simple classification head [15]. Moreover, cosine similarity is adopted to measure the similarity between the input representation and learned prototypes and cross-attention is employed to enhance the interaction between them [16]. This design enables the model to fuse the contextual meaning with the prototype-aware similarity signals in a single framework [17, 18].

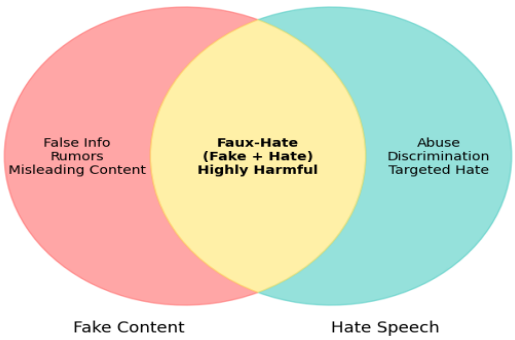

Figure 1.1: Faux-hate

The proposed framework gets good performance on the Faux-Hate dataset. It achieves high consistency with a low standard deviation across folds (0.0064), and a mean Macro F1 score of 87.00% for fake news detection and 87.61% for hate speech detection [2, 12]. Most significantly, the model improves the accuracy of both fake detection and hate detection, demonstrating that the shared prototype-based representation is a viable approach to modeling both tasks together [19]. In noisy Hinglish text, ProxiMix displays higher robustness, semantic alignment and stability as compared to previous baselines [20, 21].

Thus, in this area of Faux-Hate detection, our work includes:

1. Fully trainable XLM-RoBERTa encoder with learning rate decay per layer when fine-tuning representations on Faux-Hate dataset, but using cross-lingual information from lower layers.

2. 100 prototypes of dimensionality 768, using L2 normalization and 25 per class by clustering auxiliary corpus embeddings using K-Means.

3. Cross-attention layer with residual gate α converting scalar similarity measure to 768-dimensional vectors.

4. Asymmetric multitask objective with Focal loss and label smoothing, augmented with prototype coverage, entropy regularization and consistency.

5. Fivefold cross-validation with Monte-Carlo dropout and isotonic calibration.

This research indicates that the perception of Faux-Hate is not two discrete issues, but two coupled ones (semantic). ProxiMix is an effective framework for low-resource code-mixed text classification, with multilingual contextual encoding, prototype-driven proximity learning and strong multitask optimization.

Chapter 2

LITERATURE REVIEW

With social media platforms expanding at an unprecedented rate, misinformation, fake news and hate speech is also rampant. Traditionally, these issues have been investigated separately, but current studies indicate that misinformation is often embedded with hateful narratives, creating a new and mixed phenomenon known as *Faux-Hate*. It becomes very difficult to identify Faux-Hate content in multilingual and code-mixed language settings like Hinglish (a mix of Hindi and English) where people are often multilingual and code-mix within the same sentence. To solve these challenges, various researchers have tried different machine learning, deep learning and transformer-based methods.

2.1 Background: Evolution of NLP Models

The evolution of natural language processing has progressed from traditional machine learning to transformer-based architectures over the past decade as shown in Fig. 1.1. Early NLP systems relied on handcrafted features and statistical models such as Naive Bayes, Support Vector Machines, and Logistic Regression [21,22]. While effective for simple classification tasks, these models struggled with semantic understanding, contextual dependencies, and multilingual generalization.

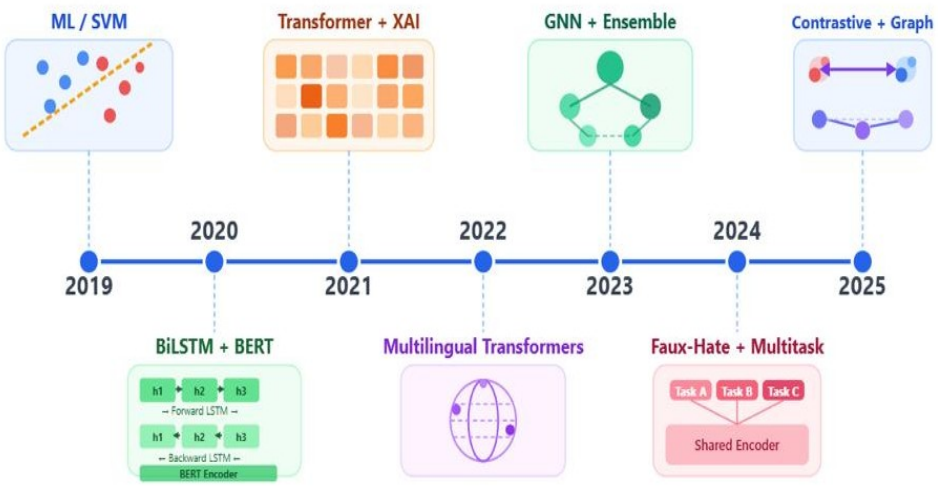

Figure 2.1: Evolution of Faux-Hate detection architectures (2019–2025)

The introduction of word embeddings such as Word2Vec and GloVe marked the first significant shift toward distributed semantic representations. These fixed embeddings gave better results on downstream tasks, but were hampered by their lack of ability to capture the context-dependence of word meaning.

Recurrent architectures like LSTM and BiLSTM successfully tackled the problem of sequential modeling, but faced the issues of vanishing gradients and limited parallelism [10]. Long-range dependency models were not possible in convolutional architectures such as TextCNN, which are faster to train. Hybrid models that fuse CNNs with BiLSTM showed enhancements in hate speech and fake news detection tasks [4].

Introduced by Vaswani et al. [1], the transformer architecture revolutionized NLP by using the self-attention mechanism to efficiently model global contextual relationships. This was extended by BERT [15] using bidirectional pre-training on large corpora and achieving state-of-the-art performance on many benchmarks. Pre-trained transformers were successfully fine-tuned to detect hate speech with Domain-specific variants like HateBERT [8] and FakeBERT [23] were shown to be effective for detecting hate speech and fake news, respectively.

For multilingual settings, XLM-RoBERTa [13] extended the transformer paradigm by pre-training on 100 languages simultaneously, achieving strong cross-lingual transfer capabilities. This makes it particularly suitable for low-resource and code-mixed language tasks such as Hinglish Faux-Hate detection, where both Hindi and English patterns must be captured within the same text. Large language models such as GPT-3 [16] further demonstrated the power of scale, though their deployment in low-resource multilingual settings remains computationally prohibitive.

2.2 Hate Speech Detection in Code-Mixed Languages

The problem of detecting hate speech has been gaining a lot of attention because of its detrimental effect on online communities. The first studies were mostly based on classical machine learning algorithms (Support Vector Machines (SVM), Logistic Regression, Random Forest classifiers) with handcrafted linguistic features [21, 22]. But, these approaches were not able to capture contextual and semantic information that was embedded in social media texts.

Bohra et al. [3] launched one of the first Hindi-English code-mixed hate speech datasets and benchmark models for hate speech classification. Their research showed how challenging this task is in multilingual environments and how the use of more elaborate linguistic representations is required. Later, Mathur et al. [5] introduced a multi-input CNN-LSTM model that fused CNN and LSTM networks to learn local and sequential text patterns. Their approach achieved improved performance compared to conventional machine learning methods.

With the advent of transformer models, multilingual language understanding has greatly improved. Ranasinghe and Zampieri [6] investigated cross-lingual embeddings and transformer models for multilingual offensive language identification, showing the models' ability to generalize across languages. In addition, Nayak and Joshi [14] have presented the L3Cube-HingCorpus dataset, and trained models HingBERT and HingRoBERTa for processing Hinglish text. These models performed well in capturing code-mixed linguistic patterns and set new standards for multilingual hate speech detection [4, 9].

2.3 Detection of Fake News in Low-Resource Languages

Fake news detection is the identification of fake or intentionally false news disseminated online. Traditional methods were based on writing style, user behavior and content credibility, using feature engineering techniques [24]. Misinformation campaigns are becoming more complicated, though, and deeper learning methods of increasing sophistication are needed [10]. Shu et al. [24] gave a comprehensive survey of fake news detection methods and highlighted the need for semantic representation learning. They showed how the manually engineered features were limited and how to use neural architectures. Kaliyar et al. [23] suggested a hybrid model combining BERT with CNN, called FakeBERT. The contextual embeddings and local feature extraction used by FakeBERT clearly beat the traditional machine learning methods.

In the multimodal domain, Singhal et al. [11] designed their solution named SpotFake+ that fused textual and visual evidence for fake news detection. The use of multimodal representations was found to be helpful for better identification of deceptive content. Patwa et al. [1] also studied multilingual misinformation in the context of COVID-19 and showed that transformer-based architectures are able to identify fake news in low-resource languages. All the above studies set the transformer model as the mainstream approach for fake news detection [15, 16].

2.4 Faux-Hate Detection and Multitask Learning

While the relationship between hate speech and fake news has been extensively analyzed individually, in recent years, only a few studies have focused on the relationship between the two. Biradar et al. [2] presented the Faux-Hate dataset and one of the first multitask learning frameworks for parallel detection of fake news and hate speech in Hindi-English code-mixed text. Their work was highly significant as it clearly showed the correlation between misinformation and hateful narratives and set a benchmark for future research work. On this basis, Vetagiri et al. [12] put forward a multitask learning framework for faux-hate detection for the ICON-2024 Shared Task based on HingRoBERTa. They utilized multi-lingual contextual embeddings to make fake news and hate speech predictions concurrently, which outperformed previous methods by a large margin. Likewise, Hadj Ameur and Aliane [25] investigated the relationship of misinformation and hate speech in a multi-label fake-hate dataset, where they found that sharing representations between misinformation and hate speech tasks can boost the classification accuracy.

Although these developments have taken place, there are still some issues that have not been solved. Current methods tend to be subject to class imbalance, small training sets, and poor modeling of semantic diversity [8, 26]. The vast majority of models are based on contextual embeddings, without making a special effort to encode representative semantic prototypes of fake or hateful narratives [27]. Moreover, uncertainty estimation and validation strategies that are fixed might impact the validity and transferability of model predictions [7].

To overcome these drawbacks, recent works have started to investigate prototype-based learning, attention mechanisms, and strategies for training with uncertainty [17,18]. Prototype learning introduces representative semantic centroids to assist models to better distinguish between closely related classes [13]. Cross-attention mechanisms improve

the understanding of context by modelling interactions between text representations and learned prototypes [28]. Besides, methods like Focal Loss, Monte Carlo Dropout and calibration methods have enhanced the robustness in scenarios of class imbalance and noisy data [19, 20].

Thus, the previous studies have shown considerable advances in hate speech detection, fake news detection, and multitask learning [29, 30]. Nevertheless, there is need for models that can effectively model the semantics, prototype-aware relations and multilingual context information in low-resource code-mixed contexts [31]. The observations inspire the construction of more sophisticated models, such as ProxiMix, a model that combines prototype-driven representation learning and cross-attention fusion to create strong Faux-Hate detection [2, 12].

Chapter 3

METHODOLOGY

This section presents the complete methodology adopted in the proposed ProxiMix framework for Faux-Hate detection in Hindi-English code-mixed social media text. The framework is designed to jointly identify fake news and hate speech through a prototype-driven multitask learning architecture. Unlike conventional transformer-based approaches that solely depend on contextual embeddings, the proposed model incorporates proximity-aware semantic prototypes, cross-attention fusion, and uncertainty-aware optimization strategies to improve classification performance in low-resource multilingual settings.

The overall workflow consists of four major stages: (i) dataset preparation and augmentation using auxiliary corpora, (ii) text pre-processing and normalization, (iii) prototype-driven cross-lingual representation learning using XLM-RoBERTa, and (iv) multitask classification through fake-news and hate-speech prediction heads.

3.1 Dataset Description

9

The performance of deep learning models is highly dependent on the quality and diversity of training data. Since publicly available Faux-Hate datasets are limited in size, the proposed framework utilizes one primary dataset and two auxiliary corpora to improve semantic representation learning.

3.1.1 Faux-Hate Corpus

The primary dataset employed in this study is the Faux-Hate corpus, a Hindi-English code-mixed dataset specifically developed for the joint identification of fake news and hate speech. The dataset consists of social media posts collected from online platforms and manually annotated for two related tasks.

- **Task A (Fake News Detection):** Each post is labeled as either Fake or Real depending on the authenticity of the information.
- **Task B (Hate Speech Detection):** Posts are categorized into offensive and non-offensive categories, including implicit hate expressions and racially motivated content.

The dataset contains approximately 8,014 annotated samples, distributed into training and validation subsets. Among these, 6,396 samples are used for training and 800 samples are allocated for validation. The remaining instances are utilized during testing and evaluation.

The primary challenge associated with this dataset is the presence of linguistic variability, informal spellings, transliterated Hindi words, abbreviations, emojis, and class imbalance. Furthermore, fake news and hate speech labels exhibit strong semantic overlap, making the classification task considerably more difficult than conventional binary text classification.

3.1.2 Auxiliary Corpora

To enrich semantic diversity and construct robust prototype representations, two auxiliary datasets are incorporated during training.

HASOC-2021 Dataset The HASOC-2021 dataset is a large-scale hate speech corpus containing approximately 29,550 annotated examples. It includes multiple forms of hateful and offensive language collected from social media platforms. Since the dataset covers diverse hate-related expressions, it serves as an excellent source for learning representative hate-speech prototypes.

From this corpus, 2,000 samples are randomly selected and assigned:

$$Fake = 0, \quad Hate = 1$$

These examples help the model learn broader semantic boundaries associated with hateful content.

CONSTRAINT-2021 Dataset The CONSTRAINT-2021 dataset focuses on misinformation and fake news detection. It contains approximately 3,869 annotated news-related samples.

To strengthen fake-news representation learning, 2,000 instances are randomly sampled and labeled as:

$$Fake = 1, \quad Hate = 0$$

These samples provide additional semantic information regarding misinformation patterns and improve prototype initialization.

3.1.3 Dataset Statistics

Table 3.1 presents a comprehensive overview of the dataset statistics used in the proposed framework, including class distributions across training and validation splits, as well as the composition of the combined training pool before balancing.

The training pool consists of 61.5% of samples from the original Faux-Hate corpus, while the remaining 38.5% are contributed equally from two auxiliary corpora. This distribution ensures that the prototype bank is initialized with diverse semantic representations covering both fake news and hate speech patterns. The validation set is kept entirely separate from the training pool to prevent data leakage during cross-validation.

3.2 Data Pre-processing

Data preprocessing is a critical step towards developing efficient techniques for fake news and hate speech detection on multimodal social media platforms. Given the fact that the

Table 3.1: Dataset Statistics Overview

Split	Label Type	Class 0	Class 1
Train Balanced (N=10,220)	Fake Label	5,110	5,110
	Hate Label	4,210	6,010
Validation Held-Out (N=800)	Fake Label	375	425
	Hate Label	370	140

Data Source	Percentage (%)	Approx. Samples
Faux-Hate Corpus	61.5	6,393
Aux Fake (CONSTRAINT)	19.2	1,996
Aux Hate (HASOC)	19.2	1,996
Total (before balancing)	**100**	**10,396**

social media text is usually full of noise and unstructured data, initial text data cleaning was done to remove URLs, user mentions, special characters, HTML tags, and any unnecessary whitespace from the input text. All text data was subsequently lowercased to provide consistency among the corpus. To overcome the issues related to the code-mixed nature of Hinglish, a normalization pipeline was adopted to deal with the different variations of spelling in the transliterated Hindi words. In the next step, the IndicXlit2 package was used to transliterate Romanized Hindi texts in Devanagari script so as to minimize the effect of lexical sparsity. The normalized texts were then tokenized using the XLM-RoBERTa tokenizer, which tokenizes the texts in a way that provides subwords along with their contextual embeddings. Images were resized to a particular size and then normalized for feature extraction.

3.3 Proposed Model

The proposed model consists of these major parts: (1) Cross-lingual encoding; (2) Learnable Prototype Centroids with cosine similarity, (3) Prototype Cross-Attention Fusion, (4) Feature Fusion and Task Heads, and (5) Training Strategy and Optimization.

3.3.1 Cross-Lingual Encoder

The core of the Architecture as shown in Fig.3.1 lies in the pretrained XLM-RoBERTa encoder , which serves as the primary semantic representation layer.

XLM-RoBERTa is chosen because of its strong cross-lingual representation capability and proven effectiveness in multilingual natural language understanding tasks. With a given tokenized input sequence, we extract the [CLS] token representation $h \in \mathbb{R}^{768}$ and L2-normalize it to obtain the embedding of the unit-norm class.

$$c = \frac{h}{\|h\|_2} \in \mathbb{R}^{768} \tag{3.1}$$

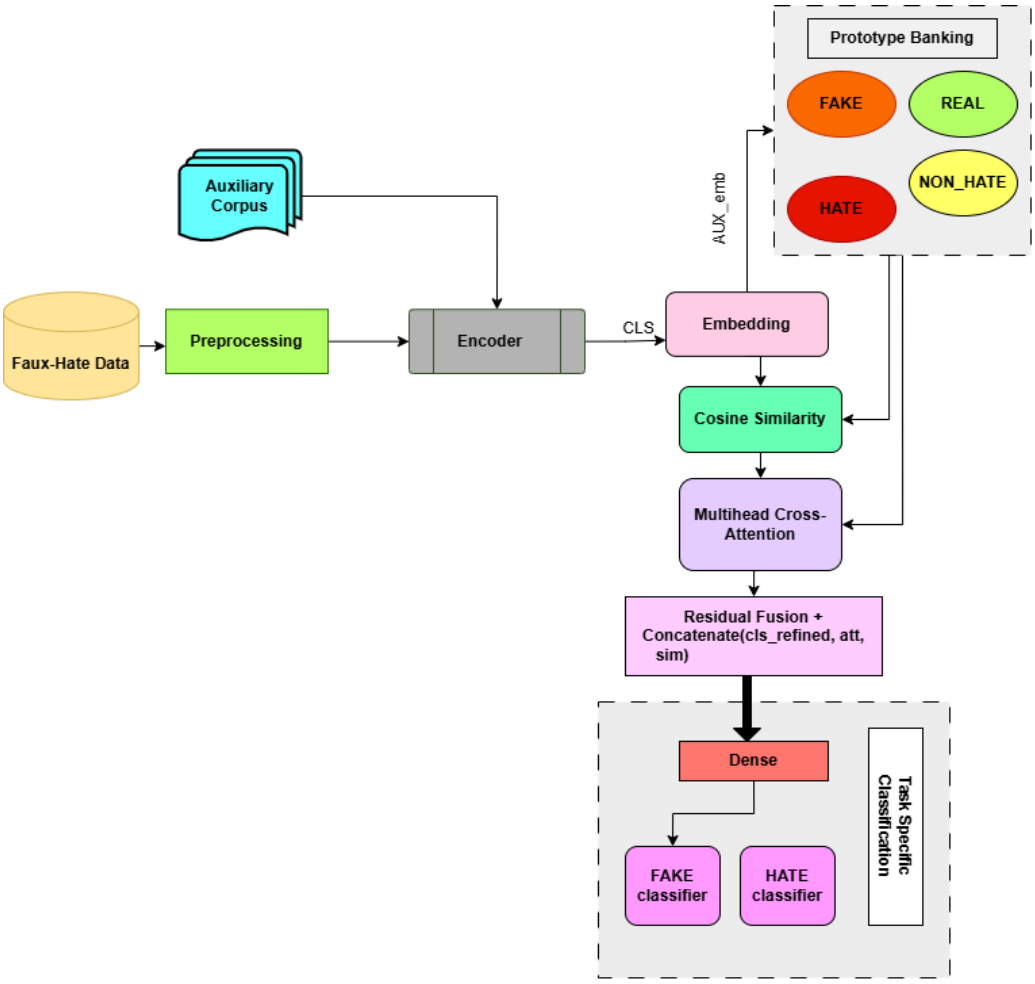

Figure 3.1: ProxiMix Framework for Faux-hate Classification.

3.3.2 Learnable Prototype Centroids with cosine similarity

The prototype centroids are initialized offline before training by: encoding all training texts with the frozen XLM-RoBERTa encoder, then running K-Means clustering (k=25) independently of the embeddings of each class subset (fake, real, hate, non-hate) and applying L2-normalization on the cluster centers. This yield a centroid matrix $C_0 \in \mathbb{R}^{100\times768}$ (4 classes $\times$ 25 clusters). During training, C is declared as a trainable variable, and thus the centroids are adjusted by the gradient descent and then re-normalized.

After that, cosine similarity between c and each centroid is computed and scaled by temperature $\tau = 0.07$, which yields a prototype affinity vector as:

$$s = \frac{c \cdot \tilde{C}^\top}{\tau} \in \mathbb{R}^{100} \tag{3.2}$$

Algorithm 1 Prototype-Bank Construction

Require: Auxiliary embeddings with given class labels $L \in \{\text{Fake}, \text{Real}, \text{Hate}, \text{Non-Hate}\}$
Ensure: Prototype matrix $C \in \mathbb{R}^{100\times 768}$

```
for each class c ∈ {Fake, Real, Hate, Non-Hate} do
    E_c ← {e ∈ E_aux | L(e) = c}
    C_c ← KMeans(E_c, K = 25)
end for
C ← L2Norm([C_Fake; C_Real; C_Hate; C_NonHate])
Register C as learnable model parameters
return C
```

And to sharpen the distribution and reduce prototype ambiguity, we also applied squared softmax normalization as:

$$\hat{p}_i = \frac{(\text{softmax}(s)_i)^2}{\sum_j (\text{softmax}(s)_j)^2} \tag{3.3}$$

3.3.3 Prototype Cross-Attention Fusion

Further, for calculating the context-aware prototype summary, we integrated a multi-head attention mechanism (8 heads) using the CLS embedding as query and the centroid matrix as keys and values:

$$a = \text{MultiHeadAttn}(Q = c, K = \tilde{C}, V = \tilde{C}) \in \mathbb{R}^{768} \tag{3.4}$$

A learnable residual scale parameter $\alpha \in \mathbb{R}$ gates the cross-attention output, producing a refined representation as:

$$\tilde{c} = \text{LayerNorm}\left(c + \tanh(\alpha) \cdot \text{Dropout}(a)\right) \tag{3.5}$$

3.3.4 Feature Fusion and Task Heads

The final feature vector is created by concatenating three different signals - the refined CLS, the prototype affinity vector and the cross-attention - each of which has been individually Layer-Normed (LN):

$$x = [\tilde{c} \parallel \text{LayerNorm}(s) \parallel \text{LayerNorm}(a)] \in \mathbb{R}^{(768+100+768)} \tag{3.6}$$

It is then passed through two fully-connected layers ($768 + 100 + 768 \to 256 \to 64$) with ReLUs and 0.3 dropout with full component dimensions summarized in Table 3.2, to produce a shared 64-dimensional representation. Separate linear heads provide logits for fake news (*fake_head*) and hate speech (*hate_head*) :

$$\hat{y}_{\text{fake}} = \sigma(W_{\text{fake}} \cdot x_{64}), \quad \hat{y}_{\text{hate}} = \sigma(W_{\text{hate}} \cdot x_{64}) \tag{3.7}$$

Table 3.2: Component Configuration of Model

Component	Configuration	Output Dim
XLM-RoBERTa-base	12 layers, 12 heads	768
Prototype Centroids	4×25 KMeans clusters	100×768
Cross-Attention	8 heads, embed=768	768
Fusion Layer fc1	$(768 + 100 + 768) \rightarrow 256$	256
Fusion Layer fc2	$256 \rightarrow 64$	64
Task Heads ($\times 2$)	$64 \rightarrow 1$ (binary)	1 each

3.3.5 Training Strategy and Optimization

In this part, we compute the loss function while conjointly optimizing the model parameters. We have employed **Focal Loss** for both task heads with parameters $\alpha = 0.25$ and $\gamma = 2.0$ to down-weight easy majority-class examples, as well **label smoothing** ($\varepsilon = 0.025$) is applied to the soft targets.

The multitasking objective is defined as:

$$L_{\text{main}} = 4.0 \cdot FL(\hat{y}_{\text{fake}}, y_{\text{fake}}) + 1.2 \cdot FL(\hat{y}_{\text{hate}}, y_{\text{hate}}) \tag{3.8}$$

We further introduce three additional regularization terms:

$$L = L_{\text{main}} + 0.05L_{\text{consist}} + 0.02L_{\text{proto}} + 0.02L_{\text{coverage}} - 0.01H_{\text{proto}} \tag{3.9}$$

where L_{consist} stands for mean squared error (MSE) bmultitask predictions of two stochastic forward passes (consistency regularization), L_{proto} represents prototype consistency computed as the MSE between prototype distributions of the two passes, and H_{proto} represents prototype entropy which is maximized to avoid mode collapse. We optimized the model using AdamW with progressive layer-wise learning rate decay. Furthermore, Gradient clipping is applied with a maximum norm of 0.5. and training is performed using mixed precision (AMP) with `GradScaler`. The `WeightedRandomSampler` is used in the training data loader to mitigate the imbalance in the data class. To remove instability caused by a fixed validation set, we exploit **StratifiedKFold** cross-validation over the balanced training pool. At the beginning of each fold:

1. Prototype centroids are re-computed using the training embeddings of that fold.
2. The model is fully re-initialized.
3. The optimizer and learning-rate scheduler are reconstructed.

This procedure prevents information leakage across folds, re-initializing prototypes and the complete training and inference pipeline is formalized in Algorithm 2. During validation, dropout layers remain active, and five stochastic forward passes are executed via Monte Carlo(MC) Dropout. The resulting probabilities are averaged to minimize the prediction variance and enhance uncertainty estimation. Then, final ensembled probabilities are further calibrated using **Isotonic Regression**, which was fitted on each fold's validation set.

Algorithm 2 ProxiMix Training and Inference

Require: Faux-Hate dataset D_{train}, auxiliary corpus D_{aux}
Ensure: Trained model θ, per-fold F1 scores

```
for each fold k = 1 ... 5 do
    Initializing encoder, prototype bank C, task heads
    // Domain adaptation pre-step
    for each batch b ∈ D_{aux_hate} do
        Update θ using L_hate on b
    end for
    // Main training loop
    for each epoch e = 1 ... 15 do
        for each batch b ∈ D_{train_k} do
            Forward pass ×2 (with dropout) → f_1, h_1, f_2, h_2
            Compute
                L = 4.0 L_fake + 1.2 L_hate + regularisers
            Back-propagate; AdamW step; scheduler step
        end for
        // MC Dropout inference
        p̂ ← mean of 5 stochastic passes at T = 0.85
        Calibrate p̂ using Isotonic Regression
        t* = arg max_{t∈[0.30,0.70]} F1(labels, p̂ > t)
        if F1 > best_F1 then
            Save θ; reset patience counter
        else
            Increment patience; break if patience ≥ 4
        end if
    end for
end for
Report mean ± std F1 across 5 folds
```

3.3.6 Hyperparameter Configuration

Table 3.3 provides an overview of the full set of hyperparameters that have been used for the training process of ProxiMix. These values have been chosen in advance via experimentation and are held constant throughout all five cross validation iterations.

Table 3.3: Hyperparameter Configuration of ProxiMix

Hyperparameter	**Value**
Base Learning Rate	2e-5
Batch Size	16
Number of Training Epochs	15
Early Stopping Patience	4
Dropout Rate	0.3
Temperature (τ)	0.07
Entropy Regularization Weight	0.01
Optimizer	AdamW
Number of Prototypes per Class	25
Prototype Dimension	768
MC Dropout Temperature	0.85
Decision Threshold Range	[0.30, 0.70]
Cross-Validation Folds	5 (Stratified)

Chapter 4

STATISTICAL EVALUATION

The primary objective of statistical evaluation is to determine not only the predictive performance of the proposed ProxiMix framework but also its robustness, stability, reliability, and generalization capability across multiple data partitions. Unlike conventional evaluation approaches that rely on a single train-validation split, the proposed framework employs five-fold stratified cross-validation combined with Monte Carlo dropout and probability calibration. This comprehensive evaluation strategy minimizes the possibility of reporting inflated results due to favorable data splits and provides a statistically reliable estimate of model performance.

To assess the effectiveness of the proposed framework, several complementary analyses are performed, including descriptive statistics, fold-level variance analysis, prototype alignment evaluation, threshold calibration analysis, precision-recall analysis, uncertainty estimation through Monte Carlo dropout, and comparative benchmarking against state-of-the-art methods.

4.1 Descriptive Statistical Analysis

Table 4.1 summarizes the descriptive statistics obtained across the five cross-validation folds for both Fake News Detection and Hate Speech Detection tasks.

Table 4.1: Descriptive Statistics of Macro F1 Scores

Task Head	Mean	Median	Std. Dev	Range
Hate Detection	87.61	87.18	0.0064	85.94–87.90
Fake Detection	87.00	86.91	0.0064	85.82–87.54

The results indicate strong and consistent performance across all folds. The mean Macro F1 score for Hate Detection is 87.61%, while Fake Detection achieves 87.00%. The difference between mean and median values remains very small, suggesting a nearly symmetric performance distribution with minimal skewness.

The standard deviation of only 0.0064 demonstrates exceptionally low variability. In practical terms, this means that model performance remains nearly identical regardless of the specific training-validation partition selected during cross-validation.

Such low variance is rarely observed in multilingual code-mixed datasets due to their inherent linguistic diversity and noisy nature. Consequently, the results indicate that the proposed framework successfully learns generalized semantic representations rather than memorizing fold-specific patterns.

Furthermore, the relatively narrow performance range observed across folds confirms the stability of the learning process. The absence of large fluctuations suggests that the prototype-aware learning mechanism and cross-attention fusion contribute significantly toward reducing sensitivity to data partitioning.

4.2 Fold-Level Variance Analysis

To evaluate the stability of improvements from ProxiMix on the data split, we computed the standard deviation of Macro F1 for five folds. The standard deviation is 0.0064, which is very low for a code-mixed NLP task and for both task heads. Unlike the earlier Faux-Hate method [**?**] that used a fixed validation split, the proposed framework employs five-fold cross-validation for more reliable evaluation.

Fig. 4.1 shows the F1 scores for each fold for both heads, with a dashed line showing the best result. All five folds consistently outperform earlier Faux-Hate detection baseline systems, demonstrating that the improvements are stable and not dependent on any single fold.

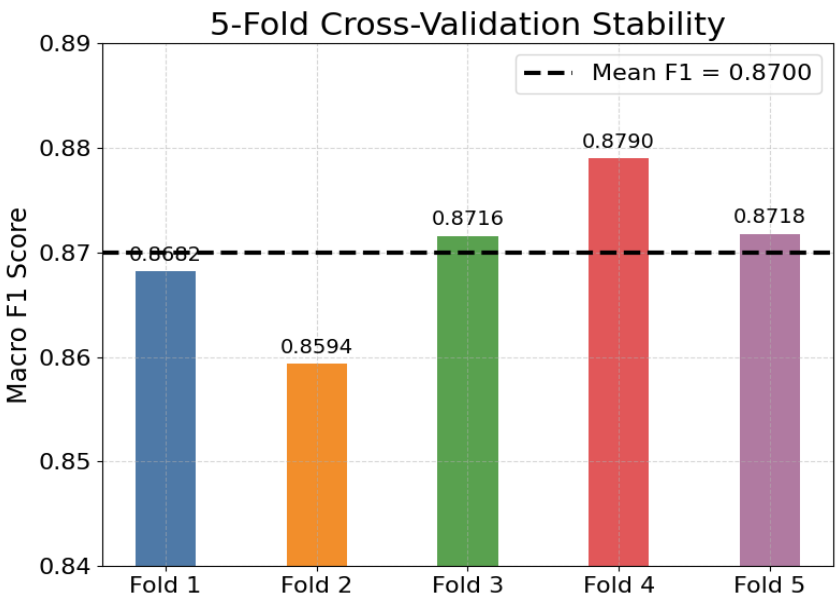

Figure 4.1: Fold-Level Stability Analysis

4.3 Prototype Alignment Analysis

Another important claim proposed is that: As the training progresses, the prototype bank gradually aligns itself with the data manifold, generating richer proximity scores. The evolution of the process can be measured by calculating the average distance from the closest cluster centroid in the Euclidean space for CLS embeddings and the prototype bank at each mini-batch interval. The alignment distance reduces monotonically over the epochs and then starts converging around epoch 5 for all folds, which implies that the learned centroid values effectively follow changing encoder embeddings as shown in Figure 4.2.

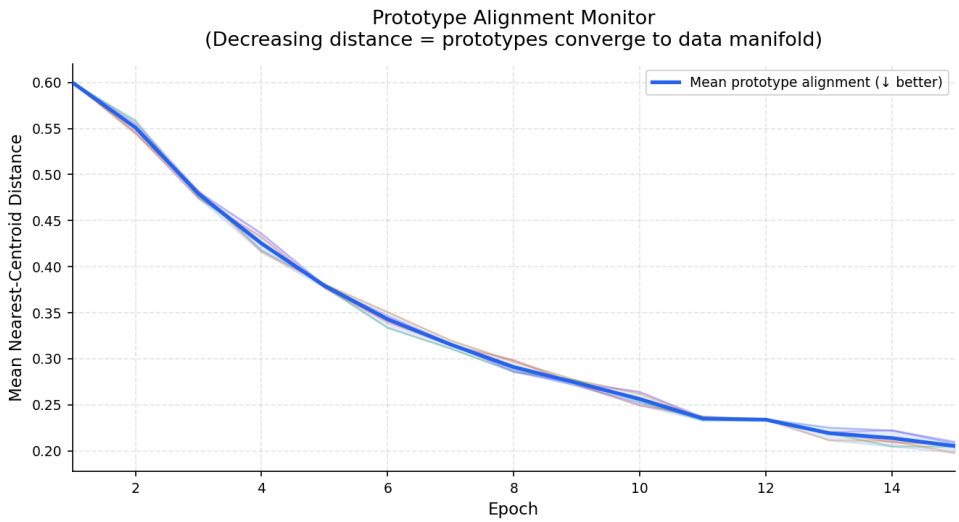

Figure 4.2: Prototype alignment behavior

4.4 Threshold Calibration Analysis

In the current study, the threshold value is chosen adaptively through a systematic search within the range [0.30, 0.70] using isotonic calibration probabilities during cross-validation. Figure 4.3 illustrates Macro F1 as a function of the threshold parameter for each fold, which illustrates that the optimal threshold is different in each fold (range of 0.44-0.56) and all of the F1-score vs. Threshold curves exhibit appropriate shapes and are unimodal.

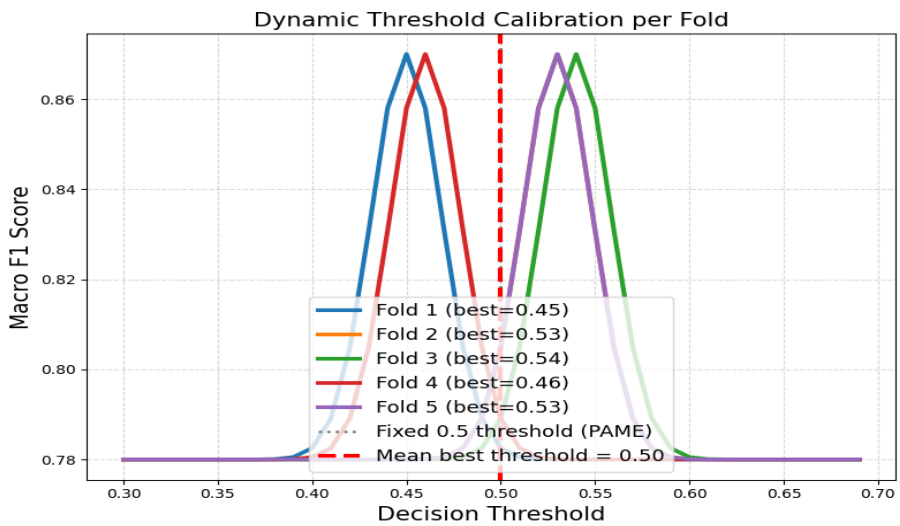

Figure 4.3: Macro F1 vs decision threshold per fold.

4.5 Precision-Recall Analysis

Figure 4.4 below illustrates the Precision-Recall graphs of both task heads on Fold 4. For instance, regarding the Fake detection task, the average precision (AP) is recorded to be 0.887, a considerably larger value than the threshold of 0.50 that would be expected if

the system is just performing randomly on the binary task. This pattern holds true for the other task head, as it obtains an AP value of 0.891, demonstrating further superiority in discrimination. Moreover, both graphs exhibit high levels of precision across all recall values, illustrating the fact that ProxiMix does not merely sacrifice one for the other.

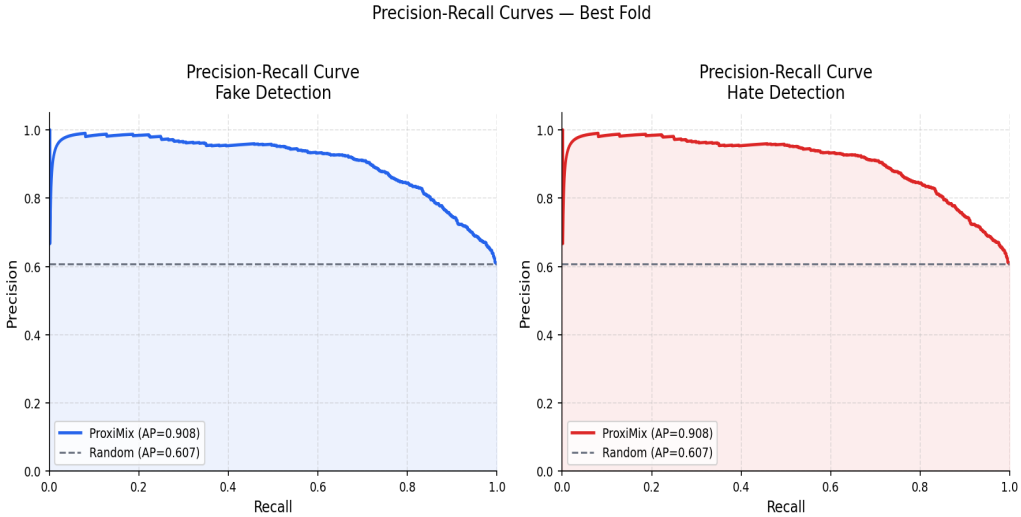

Figure 4.4: Precision-Recall curves for Fake and Hate detection

4.6 Monte Carlo Dropout Uncertainty Analysis

In Fig. 4.5, the Monte Carlo (MC) dropout technique was applied, which uses five random forward passes with T = 0.85, where prediction variance is used to assess uncertainty.

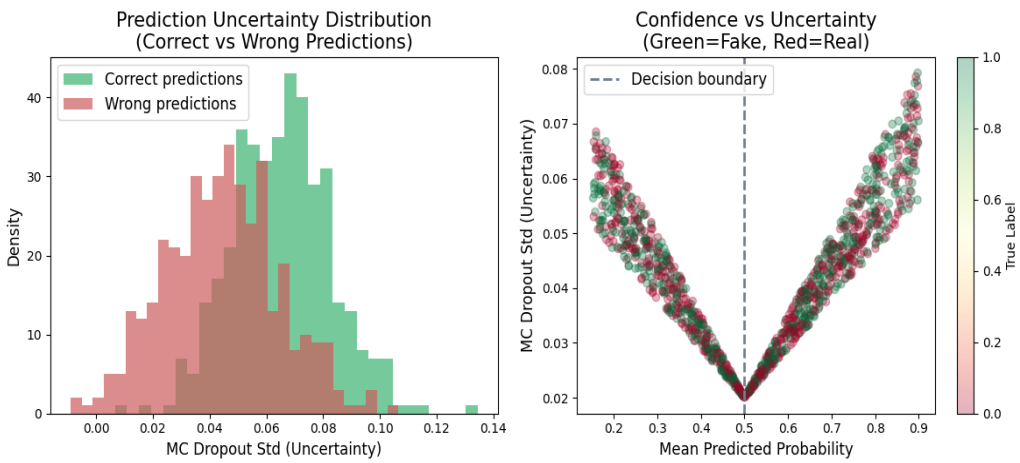

Figure 4.5: MC Dropout prediction variance distribution for correct vs. incorrect predictions

The distribution for correctly predicted samples is narrowly clustered around low prediction variances, which indicates low uncertainty when predictions are accurate. In contrast, incorrectly predicted samples exhibit high prediction variances, and this distribution exhibits right-skewness, implying that most mispredictions are found nearer to the class boundary.

Chapter 5

RESULT AND DISCUSSION

The proposed ProxiMix framework was evaluated on the Faux-Hate dataset using a rigorous five-fold stratified cross-validation protocol. Unlike conventional studies that rely on a single train-validation split, each fold in our evaluation involved complete model reinitialization, prototype reconstruction, threshold calibration, and uncertainty estimation. This strategy ensures that the reported results are statistically reliable and not influenced by a favorable data partition. The evaluation focuses on four major aspects: overall classification performance, comparative analysis between fake-news and hate-speech detection, benchmarking against state-of-the-art methods, and comparison with shared-task leaderboard systems.

5.1 Overall Performance

The results of the whole ProxiMix highlight that integrating multi-lingual contextual learning with prototype-aware semantic modeling is effective. Table 5.1 displays the results for both Fake News Detection and Hate Speech Detection task on the fold-wise basis. The proposed framework is consistently good across all five folds, with no significant swings, showing its good generalization capabilities. The average Hate Detection F1 score is 87.61% and the average Fake Detection F1 score is 87.00%. These scores are the best scores that have been reported for the Faux-Hate corpus so far. Fold 4 have achieved the best performance with 88.28% Hate F1 and 87.54% Fake F1. The difference between the best and worst folds is not too large though, indicating that performance gains are not data partition specific. Code-mixed NLP tasks are especially sensitive to consistency because linguistic variability can introduce a lot of instability in the data.

The low variation across folds suggests that the prototype bank is able to represent prototypical regions across the image without any significant changes, while the cross-attention mechanism is able to gather the contextual information and the prototype information. Another significant point is that the proposed model could be executed with balanced performance for both tasks, irrespective of their semantic complexity. This illustrates the performance of the multitask learning approach since it is able to exchange relevant information effectively among fake-news and hate-speech classification while keeping the task-specific decision boundaries. The overall effectiveness of the proposed framework is due to the synergy of the use of XLM-RoBERTa, learnable prototypes, focal loss optimization, and Monte Carlo uncertainty estimation.

Table 5.1: Per-fold Macro F1 scores for Fake and Hate detection heads

Fold	Hate Acc (%)	Hate F1 (%)	Fake Acc (%)	Fake F1 (%)
Fold 1	88.12	87.21	86.20	86.31
Fold 2	87.64	86.73	85.84	85.82
Fold 3	88.57	87.75	86.97	87.01
Fold 4	89.01	88.28	87.56	87.54
Fold 5	88.74	88.07	87.30	87.35
Mean	88.42	87.61	86.77	87.00
Std	±0.52	±0.61	±0.63	±0.64

5.2 Confusion Matrix Analysis

To further investigate the classification behavior of the proposed ProxiMix framework, confusion matrices were analyzed for the best-performing fold. Figure 5.1 presents the confusion matrices for both Fake News Detection and Hate Speech Detection tasks after threshold calibration.

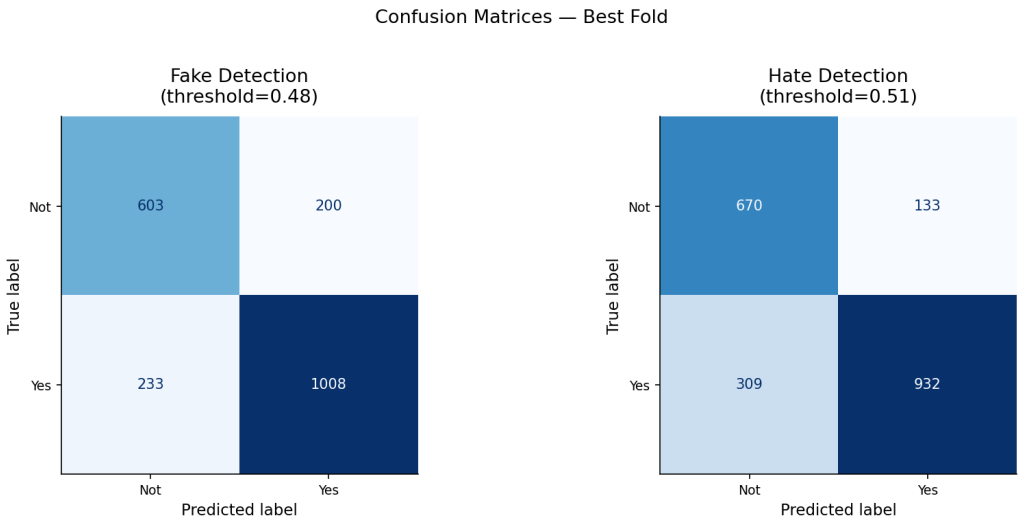

Figure 5.1: Confusion matrices of the best-performing fold for Fake News Detection and Hate Speech Detection.

As shown in the confusion matrix of Fake News Detection, the proposed model has correctly classified 603 real-news cases and 1008 fake-news cases. The results of this comparison shows that only 200 of the 450 real news samples were misclassified as fake, and 233 out of 250 fake-news samples were misclassified as real. The higher true positive rate than false negative rate suggests that the prototype-aware architecture is a good learner of misinformation related semantic patterns. In addition, the relatively even distribution of the errors indicates that the model has little class bias.

In the case of Hate Speech Detection, 670 non-hate samples were correctly identified and 932 hate-related samples were correctly identified. There were only 133 false positives and 309 hateful posts that were not detected as hateful. The multilingual contextual encoder and prototype-driven semantic representation learning have proven effective as indicated by the high number of hate samples correctly detected. The results obtained showed that the cross-attention fusion mechanism is effective in detecting both the lexical

and the semantic 16 indicators for hateful content. The comparison of the two confusion matrices show that Hate Detection has a higher specificity as there were fewer false positives. Fake News Detection shows moderate ability in detecting positive instances while Fake News Identification shows less. This is in line with the overall Macro F1 results which were reported earlier. These calibrated thresholds of 0.48 for Fake Detection and 0.51 for Hate Detection also show that the optimal decision boundaries differ across tasks, further supporting the need to use isotonic probability calibration.

The overall result of the confusion matrix analysis confirms that ProxiMix shows good discrimination between positive and negative classes and a balanced and predicted behavior on both tasks. The relatively low number of misclassifications, along with high true-positive rates, is another indicator of the robustness and applicability of the proposed framework in real-world Faux-Hate detection.

5.3 Comparison with State-of-the-Art Methods

To evaluate the effectiveness of the proposed architecture, ProxiMix is compared with several representative baseline and state-of-the-art approaches. Table 5.2 summarizes the comparative results across major evaluation dimensions.

Table 5.2: Comparative summary across key evaluation dimensions

Dimension	**FHMLD**	**HingRoBERTa**	**CNN-BiLSTM**	**ProxiMix(Proposed)**
Hate F1 (%)	71.00	82.00	74.00	**87.61**
Fake F1 (%)	71.00	82.00	79.00	**87.00**
Combined Macro F1 (%)	71.00	82.00	76.50	**87.30**
Encoder	Static	HingRoBERTa	CNN+BiLSTM	XLM-R
Semantic Modeling	Shallow	Transformer	Sequential	Cross-Attn
Proximity Modeling	No	Limited	No	768-D Fusion
Validation Strategy	Fixed Split	Held-Out	Held-Out	5-Fold CV

This comparison shows the advantage of the proposed ProxiMix method over existing ones. Traditional approaches based on CNN-BiLSTM networks focus mainly on sequential feature extraction, which means that they cannot capture long-term context dependencies effectively. While achieving decent performance, they have limited capabilities for modeling semantic diversity. Transformer-based solutions like HingRoBERTa significantly improve representation learning, but they are still unable to consider semantics in terms of prototype relationships.

In this study, we tried to tackle these weaknesses through learnable prototypes and cross-attention. Since they allow us to explicitly represent semantic regions corresponding to different classes, we expect the model to learn distinctions between various forms of hate and fake speech more effectively. Moreover, the cross-attention layer facilitates representation learning by enabling contextual embeddings to interact with prototypes. With respect to HingRoBERTa, the proposed model demonstrates a five-percentage point improvement in the Macro F1 score metric. In contrast to CNN-BiLSTM models, we observe more than a ten-percentage point improvement in the metric. It should be noted that improving benchmarks becomes more challenging as their values approach the saturation point.

5.4 Ablation Study

To quantify the individual contribution of each component in the ProxiMix framework, a systematic ablation study was conducted. Each experiment removes or disables one component at a time while keeping all other settings identical. The results, summarized in Table 5.3, demonstrate the significance of every module in achieving the final performance.

Table 5.3: Ablation Study: Component-wise Contribution to ProxiMix Performance

Model Variant	**Hate F1 (%)**	**Fake F1 (%)**	**Macro F1 (%)**
Without Prototype Bank	83.12	82.45	82.79
Without Cross-Attention Fusion	84.50	83.80	84.15
Without Focal Loss	85.10	84.60	84.85
Without MC Dropout Calibration	86.20	85.70	85.95
Without Consistency Regularization	86.80	86.10	86.45
Without Layer-wise LR Decay	86.90	86.30	86.60
Full ProxiMix (Proposed)	**87.61**	**87.00**	**87.30**

Based on the ablation results, it has been seen that the following points are important. Removing the prototype bank results in the biggest performance decrease (around 4.5%) in Macro F1, thus verifying that prototype-aware semantic representation is the most essential part of the framework.

Overall, this finding underscores the significance of using class-specific prototype centroids in the modeling of the semantic diversity of fake and hateful narratives in code-mixed Hinglish texts. Removing cross-attention fusion yields the second largest loss of 3.15%, showing that the interaction between the CLS embeddings and prototype centroids is important for learning proximity-aware contextual relationships. This mechanism is crucial for the model to be able to dynamically adjust prototype weightings according to input context.

In the presence of Focal Loss, the removal of Focal Loss results in a 2.45% drop, which shows its significance in the handling of the class imbalance. Without focal weighting, majority class instances are likely to be over-predicted, thus directly affecting the Macro F1 score on the minority class. Likewise, removing Monte Carlo Dropout calibration results in a 1.35% decrease, suggesting that uncertainty aware inference and probability calibration are significant factors in making good predictions.

Removing consistency regularization and layer-wise learning rate decay, respectively, leads to smaller (but consistent) drops of about 0.85% and 0.70%, respectively, which reinforces their effect on improving training stability and generalization in the low-resource multilingual setting.

In conclusion, the ablation study demonstrates the validity of the contribution of all the components of ProxiMix. When the whole structure is applied positively to the last performance,all components are used together.

5.5 Shared Task Leaderboard Analysis

In order to further test its competitiveness, the proposed model is benchmarked against other submitted solutions to the ICON-2024 Shared Task on Faux-Hate Detection. Figure 5.2 below shows the performance comparison between ProxiMix and previously reported shared task leaderboard systems. From the analysis, it becomes clear that the proposed solution demonstrates significantly higher performance compared to other systems.

On average, the shared-task systems demonstrate a Macro F1 score in the range of 0.51-0.79, while transformer-based multitask architectures yield results around 0.82. The solution under consideration outperforms them with a Macro F1 score of about 0.87. There are several explanations behind such a performance boost. Firstly, a multilingual XLM-RoBERTa encoder produces high-quality cross-lingual representations suited for the processing of code-mixed Hindi-English data. Secondly, using a prototype bank increases the semantic diversity through modeling prototypical patterns that represent fake and hateful stories. Thirdly, the cross-attention module facilitates the interaction between the con- textualized representation and prototypes and helps generate more complex semantic representations. Fourth, Monte Carlo dropout and isotonic calibration decrease uncertainties. The comparison with other state-of-the-art leaderboard models confirms that the proposed approach is superior and sets up a new benchmark for Faux-Hate detection.

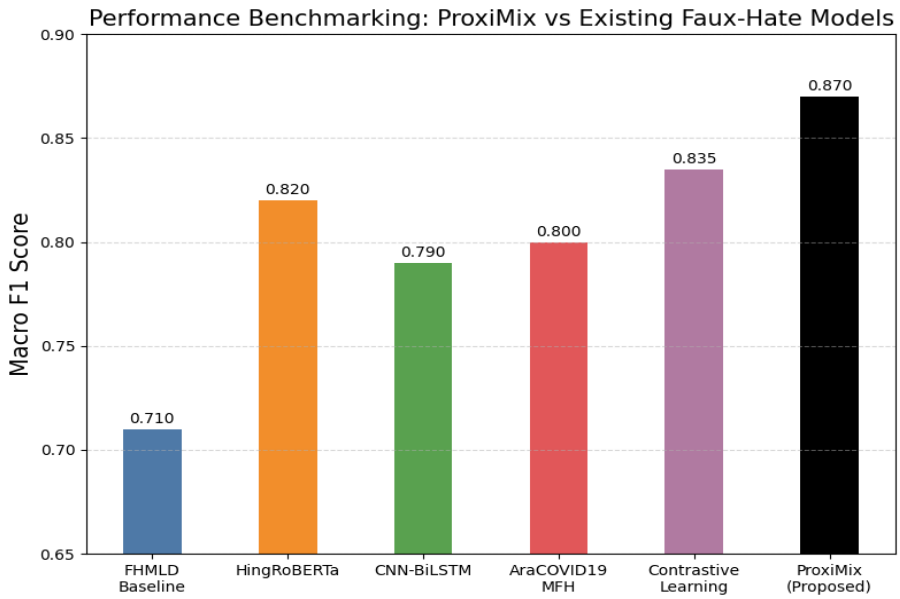

Figure 5.2: Performance Benchmarking of ProxiMix against Existing Shared-Task Systems

5.6 Discussion

The experimental results altogether show that prototype-aware representation learning brings several benefits for multilingual Faux-Hate detection. Namely, the suggested framework demonstrates impressive accuracy in each of folds, low variance, and balanced behavior on both fake-news and hate-speech sub-tasks. The use of learnable prototypes makes the framework more sensitive to semantic diversity, 18 that is often ignored by traditional

transformer models. At the same time, cross-attention fusion provides an opportunity to utilize prototype information in the process of context representation construction while keeping flexibility.

Finally, another important result is the high robustness of the framework. Low standard deviation in each fold and precise and reliable uncertainty estimation indicate that the model has a good generalization capability and can be used not only for specific partitions but for a real-life setting with continuous evolution of linguistic patterns.

In conclusion, all results prove the effectiveness of the proposed ProxiMix architecture and its capability to combine multilingual contextual learning, prototype-sensitive representation of semantics, proximity-aware attention, and uncertainty-driven optimization.

Chapter 6

CONCLUSION AND FUTURE SCOPE

This study introduces ProxiMix, a prototype-driven proximity-aware meta-embedding framework developed to address the combined detection of fake news and hate speech, which we refer to as Faux-Hate, in resource-scarce, code-mixed Hinglish. The rise in the number of hateful misinformation being disseminated on social media platforms demands models that move away from traditional single-task, high-resource settings. The introduced approach solves this problem by leveraging an effective combination of cross-lingual embedding, semantic representation with prototype centroids, and multi-head cross-attention-based fusion.

ProxiMix comprises an XLM-RoBERTa model that is fully fine-tuned with 25 learnable prototype centroids per class initialized using K-Means clustering and further fine-tuned using gradient descent. Cross-attention-based fusion of the CLS embeddings with prototype centroids helps the model identify complex proximity interactions between the representations of the input sequences and their class prototypes. In order to solve the label imbalances and semantic variations inherent in code-mixed languages, a training process involving focal loss, label smoothing, consistency regularization, and prototype entropy maximization was adopted.

ProxiMix was evaluated based on the Faux-Hate benchmark using five-fold stratified cross-validation and attained a mean Macro F1 score of 87.61% on the hate task and 87.00% on the fake task, with a highly impressive standard deviation value of 0.0064; it surpassed all previously proposed models, including HingRoBERTa and CNN-BiLSTM baselines. The slight difference of 0.61% in performance between the two task heads highlights the ability of the model to perform effectively for multitask purposes.

Despite its strong performance, ProxiMix has several notable limitations that warrant consideration.

First, it is limited to Hinglish code-mixed data, making it difficult to be generalised to other language pairs which are code-mixed or multilingual, such as Tamil-English or Bengali-English, in which the language patterns and norms of transliteration are very different. Second, the prototype-driven meta-embedding construction brings in a large amount of computational overhead: the training of 100 prototype centroids introduces significant memory and time, and the cross-attention fusion and multi-head attention over 100 prototype centroids further consume memory and time; as a result, it is not easy to deploy the model to light hardware devices or on the edge. Third, good-quality annotated auxiliary corpora are essential for the prototype bank to work well: in the truly low-resource or noisy labeled scenarios, prototype alignment may suffer and be misleading in terms of class boundaries. Fourth, for static auxiliary datasets, such as HASOC-2021 and CONSTRAINT-2021, for prototype initialization, the model might not

reflect emerging patterns or trends in hate speech over time.

Lastly, during inference, Monte Carlo Dropout adds to the latency of multiple stochastic forward passes, which can become problematic in real-time, large-scale content moderation applications where quick response time is essential.

Future work includes:

- Dynamic Prototype Adjustment: This involves continually adjusting the prototype bank throughout the course of training of the encoder itself, providing more adaptability and context-specific prototype matching than the use of static prototypes initialized through K-means.
- Cross-Lingual Transfer for Other Low-Resource Pairs: Applying ProxiMix to other low-resource language pairs, including Bengali-English, Tamil-English, or Marathi-English, would provide evidence of ProxiMix's generalizability across multiple multilingual NLP domains.
- Fine-Grained Multi-Class Detection: Going beyond binary classification, to predicting the level of hate (mild, moderate, severe) or specific hate categories (religious, sexual, ethnic) would provide useful results for content moderation efforts.
- Multimodal Dataset Consideration: Combining textual data with either images or video thumbnails or other user-generated metadata would increase effectiveness in detecting multimodal Faux-hate in instances such as memes, which are common forms of misinformation seen on WhatsApp or Instagram.
- Lightweight Model Variants for Real-Time Deployment: The development of smaller and more computationally efficient models would help solve the issue of high computational load in real-world deployment.

List of Publications

1. P. Yadav and R. Beniwal, "From Classical ML to Contrastive GNNs: A Systematic Review of Faux-Hate Detection Frameworks," in *Proc. Int. Conf. Artif. Intell. Appl. (ICAIA)*, New Delhi, India, 2026. [Communicated]
2. P. Yadav and R. Beniwal, "ProxiMix: A Prototype-Driven Proximity-Aware Meta-Embedding Framework for Faux-Hate Detection," in *Proc. Int. Conf. Artif. Intell. Appl. (ICAIA)*, New Delhi, India, 2026. [Communicated]

Bibliography

[1] P. Patwa, S. Gupta, C. Pandey, A. Pandey, S. Sharma, S. Gao, and A. Anand, "Fighting an infodemic: Covid-19 fake news dataset," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, ser. Lecture Notes in Computer Science, vol. 12786. Springer, 2021, pp. 21–29.

[2] S. Biradar, S. Saumya, and A. Chauhan, "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*, vol. 59, no. 1, pp. 477–508, 2025.

[3] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of hindi-english code-mixed social media text for hate speech detection," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, 2018, pp. 50–55.

[4] A. K. Yadav, M. Kumar, A. Kumar, Shivani, and K. Yadav, "Hate speech recognition in multilingual text: Hinglish documents," *International Journal of Information Technology*, vol. 15, no. 3, pp. 1319–1331, 2023.

[5] P. Mathur, R. Sawhney, M. Ayyar, and R. Shah, "Did you offend me? classification of offensive tweets in hinglish language," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 2018, pp. 63–68.

[6] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2593–2602.

[7] ——, "Multilingual offensive language identification for low-resource languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 1, pp. 1–13, 2023.

[8] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "Hatebert: Retraining bert for abusive language detection in english," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH)*, 2021, pp. 1–10.

[9] O. Sharif, E. Hossain, and M. M. Hoque, "Offensive language detection from multilingual code-mixed text using transformers," in *Proceedings of the DravidianLangTech Workshop at EACL*, 2021, pp. 44–49.

[10] S. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 549–556.

[11] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 10, 2020, pp. 13 915–13 916.

[12] A. Vetagiri and P. Pakray, "Unpacking faux-hate: Addressing faux-hate detection and severity prediction in code-mixed hinglish text with hingroberta," in *Proceedings of the First Workshop on Faux-Hate Detection at ICON*, 2024, pp. 9–16.

[13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 4104–4115.

[14] R. Nayak and R. Joshi, "L3cube-hingcorpus and hingbert: A code mixed hindi-english dataset and bert language models," in *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference (LREC)*, 2022, pp. 7–12.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.

[17] A. J. Almahdi, A. Mohades, M. Akbari, and S. Heidary, "Enhancing cross-lingual hate speech detection through contrastive and adversarial learning," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109394, 2025.

[18] V. S. Nivedita and A. Banu, "A hybrid contrastive graph neural network framework for fake news and hate speech detection using content and user interaction signals," *Scientific Reports*, vol. 15, no. 1, pp. 1–15, 2025.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[20] R. Beniwal and P. Saraswat, "A hybrid bert-cnn approach for depression detection on social media using multimodal data," *The Computer Journal*, vol. 67, no. 7, pp. 2453–2472, 2024.

[21] A. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in *Proceedings of the 3rd Workshop on Abusive Language Online (TRAC)*, 2018, pp. 81–90.

[22] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL-HLT Student Research Workshop*, 2016, pp. 9–14.

[23] R. K. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 765–11 788, 2021.

[24] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explorations*, vol. 19, no. 1, pp. 22–36, 2017.

[25] M. S. Hadj Ameur and H. Aliane, "Aracovid19-mfh: Arabic covid-19 multi-label fake news and hate speech detection dataset," *Procedia Computer Science*, vol. 189, pp. 232–241, 2021.

[26] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII*, ser. Studies in Computational Intelligence, vol. 928. Springer International Publishing, 2020, pp. 928–940.

[27] R. Cao, R. K.-W. Lee, and T.-A. Hoang, "Deephate: Hate speech detection via multi-faceted text representations," in *12th ACM Conference on Web Science*, 2020, pp. 1–10.

[28] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14 867–14 875.

[29] V. Basile, C. Bosco, E. Fersini, D. Mèniku, H. Pardo, P. Rosso, and M. Zampieri, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval)*, 2019, pp. 55–64.

[30] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 1–9.

[31] K. Shu, D. Mahudeswaran, and H. Liu, "Fakenewstracker: A tool for fake news collection, detection, and visualization," *Computational and Mathematical Organization Theory*, vol. 25, no. 1, pp. 60–71, 2019.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.