**Title:** Linking Memes to Knowledge Graphs for Hateful Meme Detection in Kannada

**Abstract**—Memes serve as a powerful and viral form of digital communication, often blending humor with socio-political commentary. However, they also pose a threat when used to propagate hate speech and offensive content. This study proposes a multimodal approach to detect hateful content in Kannada memes using a combination of optical character recognition (OCR), keyword-based hate identification, semantic topic mapping, knowledge graph construction, and classification using both traditional and deep learning models. Our dataset comprises 140 Kannada memes sourced from Instagram. Text was extracted via PyTesseract OCR, with manual correction for failed cases. We manually annotated the memes as hateful or non-hateful, created a hate lexicon in Kannada, and built a knowledge graph mapping keywords to semantic categories. Using this structured representation, we applied and evaluated multiple machine learning (Logistic Regression, SVM, Random Forest) and deep learning models (IndicBERT, mBERT, CNN), along with fusion strategies. The best results were achieved using IndicBERT + CNN with 71% accuracy and 0.81 F1-score.

**I. INTRODUCTION**

Memes have emerged as a dominant medium of expression on the internet. Despite their humorous appeal, memes can be potent carriers of hate, sexism, casteism, and other forms of online abuse. While hateful meme detection in English has been extensively studied [1][2], the Kannada language—a major Indian language with over 50 million speakers—remains underexplored due to limited NLP resources. We address this gap through a multimodal pipeline that combines text and image understanding, enriched with knowledge graph reasoning.

A person on a motorcycle

AI-generated content may be incorrect.

To develop a curated list of Kannada hate speech keywords, and categorize them into semantic topics (e.g., gender abuse, politics, caste-based hate).

To construct a knowledge graph linking meme texts, hate keywords, semantic topics, and labels to provide interpretable hate reasoning.

To implement and evaluate machine learning and deep learning models (SVM, Random Forest, Logistic Regression, IndicBERT, mBERT, CNN) for text-only and image-only meme classification.

To analyze the performance of all models using standard metrics (accuracy, precision, recall, F1-score) and identify the best-performing setup.

To contribute toward resource development for regional hate speech detection in Kannada and encourage further work in low-resource Indian languages.

**II. RELATED WORK**

The Hateful Memes Challenge [1] and the MultiOFF dataset [2] emphasized multimodal hate detection combining vision and language. Chakravarthi et al. [3] and Hegde et al. [4] developed Kannada offensive language datasets. IndicBERT [5] and Krishiq-BERT [6] offer transformer models pretrained for Indian languages. MemeFier [7], MEMEX [8], and KnowMeme [9] proposed dual-stage fusion and graph reasoning techniques. MemeGraphs [10] and IMKG [11] explored integrating memes with external knowledge bases like Wikidata. Sharma et al. [14] presented the HASTIKA Kannada-English hate corpus. Dandapat et al. [13] surveyed hate detection across Indian languages.

**III. DATA COLLECTION AND PREPROCESSING**

**A. Meme Collection**  
We collected 140 Kannada memes from Instagram via screenshots, saved in .jpg format. All metadata and labels were stored in meme\_extracted.xlsx.

**B. Text Extraction (OCR)**  
We used PyTesseract with lang='kan' to extract text. Roughly 15% of the memes had OCR failures, for which Google Lens was used manually.

**C. Manual Annotation**  
Each meme was annotated as either **Hateful (1)** or **Non-Hateful (0)** based on text and image context.

**D. Modelling**  
Non-Kannada characters were removed. Tokenization, stopword removal, and lowercasing were applied.

**Text-Only Models**

We used two pretrained transformer-based models: **IndicBERT** and **mBERT**, which are designed to handle Indian languages and multilingual content respectively.

**Embedding Extraction**: The clean meme texts were tokenized and passed through IndicBERT/mBERT to extract 768-dimensional contextual embeddings.

**Classifiers Used**:

**Support Vector Machine (SVM)**

**Logistic Regression**

**Random Forest**

These classifiers were trained on the text embeddings to predict whether a meme was hateful or non-hateful. Performance was evaluated using standard classification metrics.

**Image-Only Model**

To classify memes using visual features alone, we built a **Convolutional Neural Network (CNN)** model using the image\_CNN.ipynb notebook:

* + Two Conv2D layers with ReLU activation followed by MaxPooling
  + A Flatten layer
  + A dense fully connected layer with Dropout
  + Final Dense(1, activation='sigmoid') layer for binary classification
* **Training Setup**:
  + Image size: 128x128 (resized)
  + Batch size: 16
  + Optimizer: Adam
  + Loss: Binary Crossentropy

**Multimodal Fusion Models**

To leverage both text and image cues, we developed **early fusion** models using a combination of CNN (for image) and IndicBERT/mBERT (for text):

* **Fusion Approach**:
  + The meme image is passed through the CNN branch to obtain a 512-dimensional feature vector.
  + The meme text is passed through the transformer (IndicBERT or mBERT) to extract a 768-dimensional embedding.
  + These two vectors are **concatenated** and passed through a dense neural network with dropout regularization for final binary classification.
* **Optimization**:
  + Used Adam optimizer with a learning rate of 1e-4
  + Batch size: 8
  + Early stopping on validation loss

**Performance**:

* **IndicBERT + CNN fusion** model achieved the best performance:
  + **Accuracy**: 71%
  + **F1-score**: 0.81
* **mBERT + CNN fusion** followed closely with 64% accuracy and 0.78 F1-score.

These results highlight the strength of multimodal representation in capturing both subtle visual innuendo and explicit textual hate.

| **Embedding** | **Classifier** | **Class** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- | --- | --- |
| IndicBERT | SVM | Non-Hateful | 54% | 0.69 | 0.5 | 0.58 |
|  |  | Hateful |  | 0.4 | 0.6 | 0.48 |
|  | Random Forest | Non-Hateful | 61% | 0.65 | 0.83 | 0.73 |
|  |  | Hateful |  | 0.4 | 0.2 | 0.27 |
|  | Logistic Regression | Non-Hateful | 54% | 0.65 | 0.61 | 0.63 |
|  |  | Hateful |  | 0.36 | 0.4 | 0.38 |
| mBERT | SVM | Non-Hateful | 64% | 0.7 | 0.78 | 0.74 |
|  |  | Hateful |  | 0.5 | 0.4 | 0.44 |
|  | Random Forest | Non-Hateful | 68% | 0.67 | 1.00 | 0.18 |
|  |  | Hateful |  | 1.00 | 0.1 | 0.8 |
|  | Logistic Regression | Non-Hateful | 61% | 0.71 | 0.67 | 0.69 |
|  |  | Hateful |  | 0.45 | 0.5 | 0.48 |

We applied Logistic Regression, SVM, and Random Forest on embeddings from mBERT [5] and IndicBERT [6].

| **Class** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Non-Hateful | 43% | 0.64 | 0.37 | 0.47 |
| Hateful |  | 0.29 | 0.56 | 0.38 |

A simple CNN was trained using TensorFlow. It consisted of two Conv2D+MaxPool layers, followed by Dense layers using only image features. The model achieved **43% accuracy**.

| **Class** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Non-Hateful | 71% | 0.71 | 0.94 | 0.81 |
| Hateful |  | 0.75 | 0.30 | 0.43 |

**IndicBERT + CNN Fusion**  
Achieved best results: **71% accuracy**, using dropout and early stopping.

| **Class** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| 0 (Non-Hateful) | 64% | 0.64 | 1.00 | 0.78 |
| 1 (Hateful) |  | 0.00 | 0.00 | 0.00 |

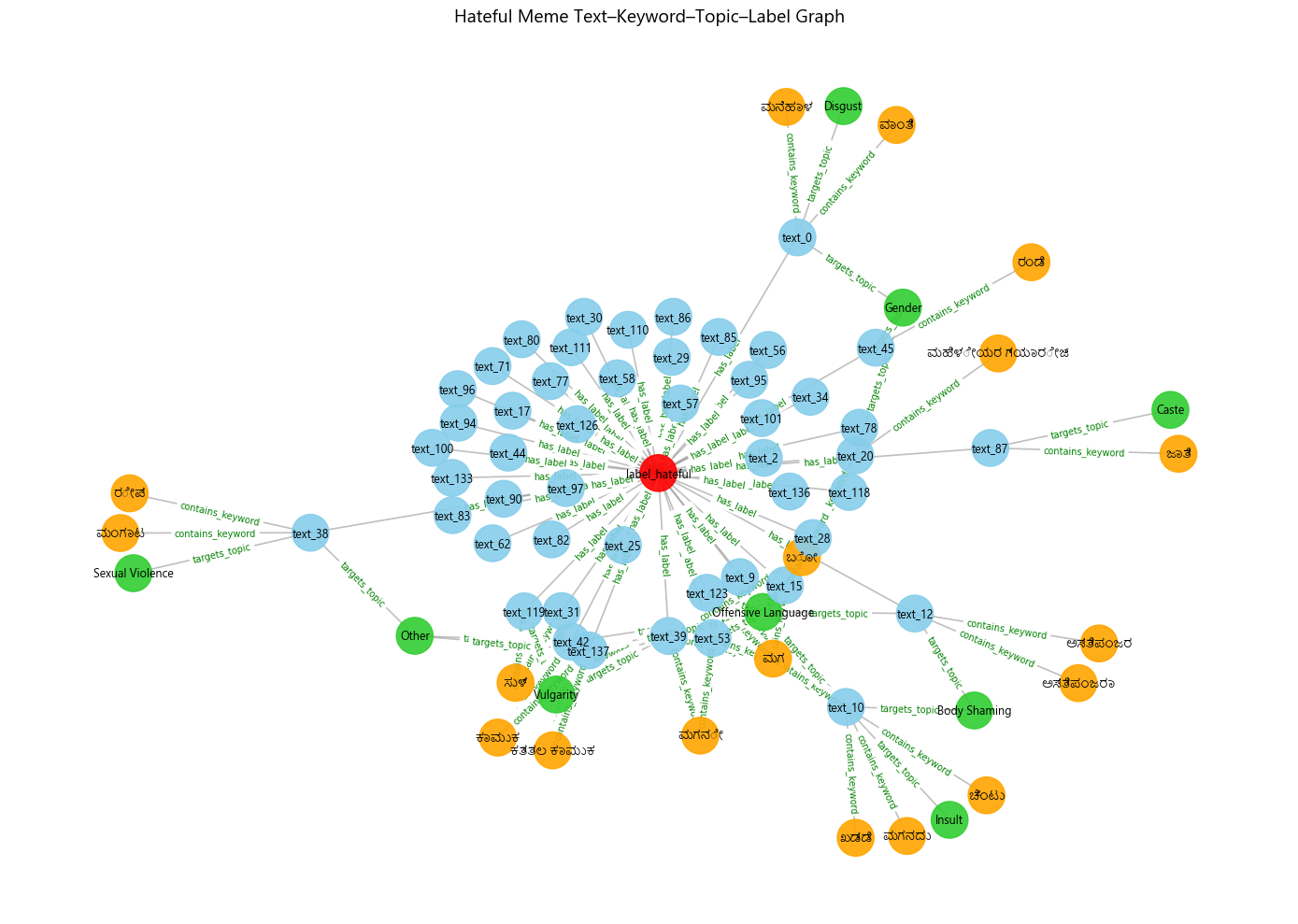
**mBERT + CNN Fusion**  
Fusion of 768-dim text embeddings (mBERT) and 512-dim CNN image embeddings. Achieved **64% accuracy**.

**IV. KEYWORD MAPPING AND KNOWLEDGE GRAPH**

We developed a curated Kannada hate keyword list (~35 words), e.g., “ರಂಡಿ”, “ಮಗನೇ”, “ಜಾತಿ”. Each was linked to a semantic topic like *Gender*, *Politics*, *Vulgarity*, etc. Using networkx, we constructed a knowledge graph with: - **Text nodes** = individual memes - **Keyword nodes** = hate keywords - **Topic nodes** = semantic categories - **Label nodes** = Hateful/Non-Hateful

Edges represented: - contains\_keyword - targets\_topic - has\_label

The graph was visualized using matplotlib with improved layout, color, and Kannada font rendering.



A diagram of a network

AI-generated content may be incorrect.

A diagram of a network

AI-generated content may be incorrect.

A diagram of a network

AI-generated content may be incorrect.

A diagram of a network

AI-generated content may be incorrect.

A diagram of a network

AI-generated content may be incorrect.

**VIII. CHALLENGES** - OCR struggled with decorative Kannada fonts. - Manual annotation was labor-intensive. - Small dataset limited model generalization. - Knowledge graph required manual keyword engineering.

**IX. CONCLUSION**

This study demonstrates a comprehensive framework for Kannada hateful meme detection using OCR, manual annotation, knowledge graphs, and multimodal learning. The fusion of IndicBERT + CNN showed the highest performance. The knowledge graph added interpretability by mapping hateful keywords to semantic concepts.

**X. FUTURE WORK** - Increase dataset size (>500 memes) - Fine-tune OCR for Kannada fonts - Explore CLIP, Flamingo for cross-modal fusion - Enrich knowledge graphs with Wikidata triples

**REFERENCES**

[1] D. Kiela et al., “The Hateful Memes Challenge,” NeurIPS 2020.  
[2] S. Suryawanshi et al., “MultiOFF Dataset,” ELRA, 2020.  
[3] B. R. Chakravarthi et al., “DravidianCodeMix,” LREC, 2022.  
[4] A. Hegde and S. Lakshmaiah, “Trolling Memes in Kannada,” 2024.  
[5] D. Kakwani et al., “IndicNLPSuite,” arXiv:2005.00085, 2020.  
[6] P. Ajawan et al., “Krishiq-BERT,” JIEI Series B, 2024.  
[7] C. Koutlis et al., “MemeFier,” arXiv:2304.02906, 2023.  
[8] S. Sharma et al., “MEMEX,” ACL, 2023.  
[9] L. Shang et al., “KnowMeme,” IEEE eScience, 2021.  
[10] R. Kumar et al., “MemeGraphs,” arXiv:2305.18391, 2023.  
[11] L. Tommasini et al., “IMKG,” Springer, 2023.  
[12] A. Singh and R. Thakur, “Federated Hate Detection,” NAACL, 2024.  
[13] S. Dandapat et al., “Hate Speech in Indian Languages,” Springer, 2024.  
[14] A. Sharma et al., “HASTIKA,” LRE, 2025.

**Appendix**  
- Dataset: meme\_extracted.xlsx  
- Notebooks: image\_CNN, mBERT, IndicBERT, fusion models  
- Graph: hateful\_meme\_graph.gml