Behavior Detection of Social Media Memes Using NLP

[1]Bhavya Jain, [2]Sudhanshu Yadav, [3]Dhruv Goyal, [4]Anand Gupta
[1]B.Tech Student, [2]B.Tech Student, [3]B.Tech Student, [4]Associate Professor
[1,2,3,4]Department of CSE,
[1,2,3,4]Netaji Subhash University of Technology (NSUT), Dwarka, India

Abstract - The rise in the number of social media users has led to an increase in the hateful content posted online. In countries like India, where multiple languages are spoken, these abhorrent posts are from an unusual blend of code-switched languages. This hate speech is depicted with the help of images to form “Memes" which create a long-lasting impact on the human mind. In this paper, we take up the task of hate and offense detection from multimodal data, i.e. images (Memes) that contain text in code-switched languages. We firstly present a novel triply annotated Indian political Memes (IPM) dataset, which comprises memes from various Indian political events that have taken place post-independence and are classified into three distinct categories. We also propose a binary-channeled CNN cum LSTM based model to process the images using the CNN model and text using the LSTM model to get state-of-the-art results for this task.

Index Terms – LSTM(Long Short-term Memory), CNN, Word embeddings, SRL(Semantic Role Labeling), Hate Speech Detection, Multimodal, Datasets, Code Switching, Indian Political Memes Classification, Social Media Analysis

I. Introduction

Internet memes have become a major form of communication and expression, they are an essential part of social media’s popular culture. But like with any other type of content in social media, they also can lead to the spread of hate speech and propaganda. In our approach, the input to the model is a meme image. The text is extracted from the memes through OCR extraction. The image and text features are formed using various feature extraction methods. Based on the design of the classifier, the image features are extracted through CNN model whereas the text features are extracted through LSTM model and SRL. Both these features are then combined to predict whether the meme is Hate-inducing, Satirical or Benign.

II. Literature Survey

The conventional method for meme categorization includes the CNN-LSTM combination channel, i.e., CNN for images features and LSTM for text features as shown in Figure 1. This does not include the semantic meaning of the memes and the F1-Score for the model is 0.781.
We aim at improving the metrics of the model by adding word embeddings (Global Vector embeddings) and SRL (Semantic Role Labeling) layers to the conventional model and improve the semantics of the classification results. This also includes classification using only single input by the user i.e., meme. The text is extracted by our enhanced model using OCR extraction.

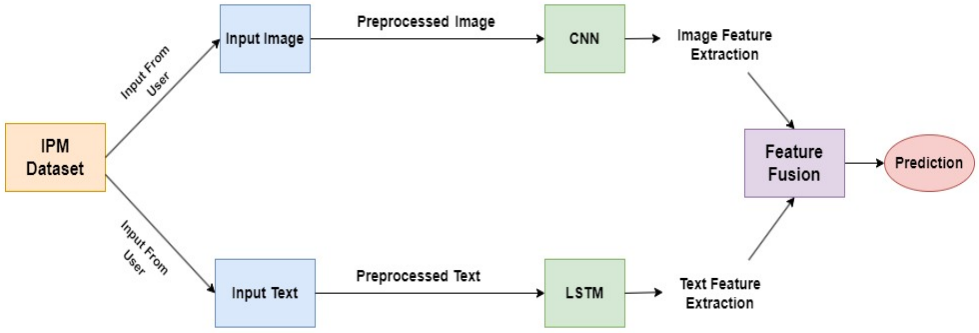

Fig.1 Basic LSTM-CNN model based on previous work

The significant contributions in our work are behavior of Meme prediction on IPM dataset. We aim at Classifying Memes into 3 categories namely:- Hate-inducing, Satirical or benign with an improved F1-score.

III. DATASET AND METHODOLOGY

The Indian Political Memes (IPM) Dataset was curated by scraping images using the "google_images_download" module with keywords related to famous Indian politicians, social activists, journalists, and political events post-independence. A total of 5000 images were scraped, and 1500 memes were randomly selected for annotation. Three annotators categorized these memes into the following groups:

(I) Hate Inducing
(II) Satirical
(III) Non-offensive

Label	IPM Dataset
Non-Offensive	339
Hate-Inducing	427
Satirical	452
Total	**1218**

Table 1: Class Distribution in Indian Political Memes (IPM) dataset

After removing blurred and without text images, the final dataset consisted of 1218 memes as shown in Table 1.

(1) PREPROCESSING IMAGES

To optimize the text extraction process from memes, we employ several image preprocessing techniques to enhance OCR accuracy:

(i) *Rescaling*: Certain images are resized to a larger dimension to improve the readability of text, especially in cases where the font size is small.

(ii) *Gaussian Blurring*: This technique is applied to reduce image noise, enhancing the clarity of text. By convolving the image with a Gaussian kernel, we achieve smoother gradients and improved OCR performance.

(iii) *Deskewing*: Images with skewed text are corrected by deskewing, which rotates the image to align the text horizontally. This step ensures better OCR recognition of the text orientation.

(iv) *Gaussian Adaptive Thresholding*: Thresholding is employed to convert the text into a binary format (black and white), aiding OCR algorithms in distinguishing text from the background more accurately.

(2) EXTRACTION OF TEXT IMAGES

Following the image preprocessing steps outlined earlier, the processed images are then subjected to an open-source OCR reader available at ocr.space. This OCR tool is utilized to extract text content from the memes accurately. The resulting text is then translated into English for further analysis and understanding. Below is a summary of the samples after text extraction from the image examples mentioned earlier, along with their respective English translations.

Hate Inducing *Benign* *Satirical*

Fig.2 Examples of a) Hate Inducing b) Benign and c) Satirical Memes in IPM

Figure 2 illustrates examples of non-offensive, satirical, and hate-inducing memes. These memes were part of the dataset that underwent a pipeline for extracting text from the images. Table 2 presents the extracted text from the images, their corresponding English translations, and the assigned labels. Various evaluation metrics have been computed for our dataset to assess its quality and performance.

Figure	Hinglish Text Extracted	English Translation	Label
Figure 1	Tu wohi secular ghoda hai na jo Eid Mubarak ke status daal raha hai	Are you that secular horse who put status of Eid Mubarak	Hate Inducing
Figure 2	Agar hum kare to kare kya bole to bole kya	What should we do What shall we speak	Non Offensive
Figure 3	WHAT DO YOU MEAN I CAN'T EAT AMERICA	What do you mean I can't eat America	Satirical

Table 2: Example of Hinglish text extraction from the memes with their respective English translations

(3) PREPROCESSING TEXT

The text extracted from memes underwent a processing pipeline aimed at converting them into semantic feature vectors.

(i) Initially, preprocessing steps were applied to remove non-essential elements such as hashtags (e.g., #politicsinIndia), URLs, user mentions (denoted by '@'), and numbers from the text, as they do not contribute relevant sentiment information. Additionally, stop words were removed using the NLTK library.

(ii) Emoticons (e.g., ":)", "XD") were replaced with their corresponding textual descriptions to capture the true emotions they convey.

(iii) Devanagari (Hindi) script comments were transliterated into Roman (English) script using the Indic-transliterate Python library.

(iv) Hinglish text was translated into English using an Xlit-Crowd Conversion Dictionary to ensure consistency in the language and facilitate further analysis.

IV. OUR MODEL

We propose a novel method which involves a binary channeled CNN-cum-LSTM model that takes the meme image as its input and finally concatenates the two channels(text and image) to produce the result. The model architecture is depicted in Figure 3.

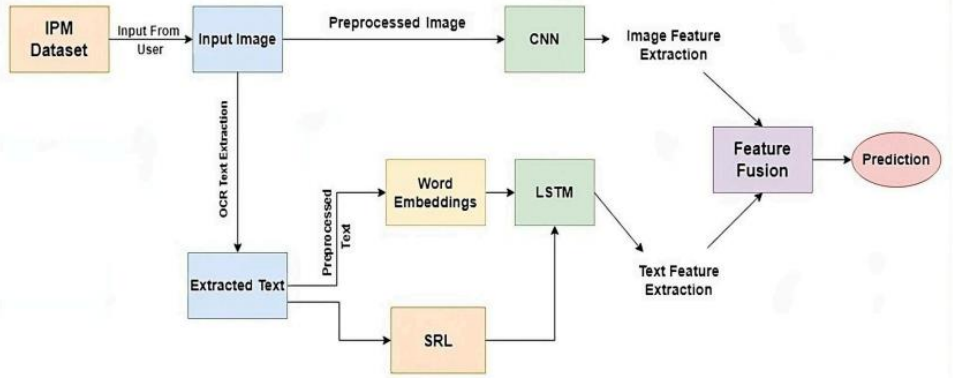

Fig.3 Proposed model architecture

The CNN Channel

The CNN is used to extract features from the input image, employing two convolutional layers, max-pooling layers, and dropout[8] layers in the code.

i.) Conv2D: The Conv2D algorithm uses a convolutional layer with 64 filters and a kernel size of (5, 5) followed by ReLu activation.

ii.) MaxPooling2D: This layer is used to down sample spatial dimensions.

iii.) Dropout : This layer is used for Regularisation. iv.) The process is repeated with another convolutional layer and max-pooling.

The final output (l_1) of the CNN part is then combined with the LSTM part.

The LSTM Channel

The LSTM is a machine learning technique used to identify and analyze sequential patterns and dependencies in the input text.

i.) Embedding: This layer maps words to dense vectors.
ii.) SRL: This layer assigns entity: role pair to form "who did what to whom"
iii.) Dropout: This layer is used for Regularisation.
iv.) LSTM: The LSTM layer with recurrent dropout and 64 units dropout is used.

The LSTM part's final output (l_2) is then combined with the CNN part.

Combination of Channel

The concatenate function is used to combine the outputs of the CNN and LSTM parts. The process involves adding additional dense layers (d) and the final output layer (output) for further processing and prediction. The model uses visual features from images and sequential patterns from text for prediction, generating concatenated output for classification tasks. The loss function used in the last layer was categorical cross-entropy. The output obtained after passing through the dense layers is one of the three classes, i.e., Hate Inducing, Satirical and Non-offensive.

V. EXPERIMENTATION AND RESULT

We analyze the results of our model on the IPM dataset as shown in Figure 4, the Meme is accurately classified as hate-inducing.

Fig.4 Meme Classification Result

Meme Classification results show better scores for Precision, Recall and F1-Score for our model with SRL and GloVe word embeddings implementation in the CNN-LSTM model than the conventional CNN-LSTM model. The metric comparisons of the conventional as well as our enhanced model are depicted through Figure 5 and Table 3.

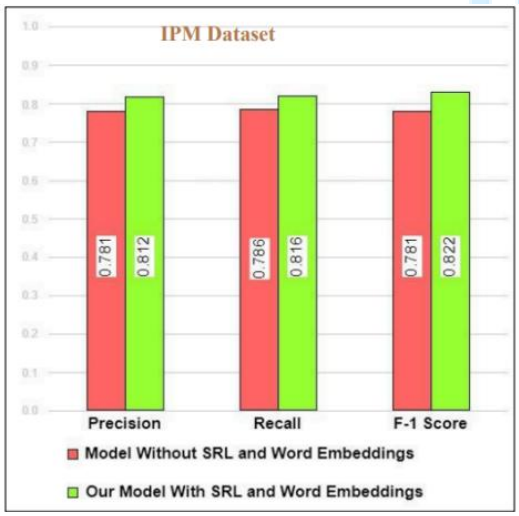

Fig.5 Metric Calculation

Features	Precision	Recall	F1-score
Our Model with Word Embeddings and SRL	0.812	0.816	0.822
Model without Word Embeddings and SRl	0.781	0.786	0.781

Table 3: Meme Prediction Results of our model

ERROR ANALYSIS

The potential factors contributing to incorrect predictions by our model include:

(i) *OCR limitations:* The OCR tool we utilized (ocr.space) may encounter challenges with accurately extracting text from memes with small fonts, slight blurriness, or text arranged vertically or diagonally, leading to errors in text recognition.

(ii) *Uncommon Hinglish words:* Our model may struggle with uncommon Hinglish words that arise from spelling variations, grammatical errors, or the blending of regional languages in memes. These variations can result in new words not present in standard dictionaries, impacting model accuracy.
(iii) *Disguised hate content:* Some memes may appear satirical to annotators but contain disguised hate speech directed towards individuals. Such subtle expressions of hate may not be correctly identified and classified by our model, leading to mispredictions.

VI. CONCLUSIONS

In this paper, we used a multi- channel CNN-LSTM model, which processes the images and text individually and combines the analysis from both channels to give the final classification result on the IPM dataset (Indian Political Memes) which consists of images (Memes) classified in three categories - Benign, Satirical and Hate Inducing. GloVe word embedding models are implemented along with Semantic Role Labelling which suggests that our model outperforms all the other models for Hinglish language Memes classification on the IPM dataset. We believe this method would be useful for hate speech detection for memes on various social media platforms.

VII. REFERENCES

[1] MacAvaneySean *et al., "*Hate speech detection: Challenges and solutions", PLoS One, 2019
[2] JakubowiczAndrew, "Trolling, hate speech and cyber racism on social media", Cosmopolitan Civil Societies: An Interdisciplinary Journal, 2017
[3] IslamNoman et al., "A survey on optical character recognition system", 2017
[4] Castaño DíazCarlos Mauricio, "Defining and characterizing the concept of internet meme", Ces Psicología, 2013
[5] PronobisAndrzej *et al., "*Large-scale semantic mapping and reasoning with heterogeneous modalities"
[6] KielaDouwe *et al. "*The hateful memes challenge: Detecting hate speech in multimodal memes", 2020
[7] FortunaPaula *et al., "*A survey on automatic detection of hate speech in text", ACM Computing Surveys, 2018
[8] BaltrušaitisTadas *et al., "*Multimodal machine learning: A survey and taxonomy", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018
[9] GravesAlex *et al.*, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures
Neural Networks",2005