Hate Speech and Offensive Content Identification in Memes in Hindi and Gujarati using BERT-Based Approach

Peng Zhang, Gehao Lu*

School of Information Science and Engineering, Yunnan University, Kunming 650500, Yunnan, China.

Abstract

With the popularity of social media, memes have become an important form of information dissemination and online communication. However, memes are also often used to spread hate speech, offensive content, and false information, posing a serious threat to the online ecosystem. Especially in regional languages such as Hindi and Gujarati, which have relatively scarce resources, the automatic identification of such content is extremely challenging due to their multimodal combination of images and text characteristics, complex cultural backgrounds, and unique linguistic phenomena such as code-switching. Our team *Peng Zhang* proposes an advanced method based on BERT aimed at effectively identifying hate speech and offensive content in Hindi and Gujarati memes [1]. The core of this method lies in utilizing the powerful contextual understanding capabilities of pre-trained language models. We first extract text information from meme images and perform necessary preprocessing, followed by fine-tuning with a targeted BERT model to complete the classification task [2]. To verify the effectiveness of the method, we compared the performance of various BERT variant models, including multilingual BERT and monolingual BERT models specifically pre-trained for Indian languages such as Hindi or Gujarati. We mainly participated in the identification of memes for abuse, sentiment, sarcasm, and vulgarity detection. The team rank is determined based on the average of all four Macro F1 scores. Our team *Peng Zhang* ranked 11th on HASOC2025-meme (Gujarati) with a score of 0.55522 and 8th on HASOC2025-meme (Hindi) with a score of 0.57094. The above two tracks mainly employ transformer-based models. The code sources for this paper are available via GitHub.

Keywords

BERT, Multimodal Data, Natural Language Processing, Class Classification

1. Introduction

With the rapid growth of social media platforms [3] such as Twitter, Facebook, and Instagram, memes have emerged as a powerful medium for online communication. Their ability to convey complex ideas through a combination of text and visuals has made them a popular means of expression across diverse linguistic and cultural communities. However, this widespread use has also brought significant challenges. Memes are increasingly exploited to disseminate hate speech [4], offensive language, and harmful content, raising serious concerns about their impact on social harmony and digital safety [5].

In multilingual societies like India, the problem becomes even more complex due to the presence of multiple regional languages such as Hindi and Gujarati, where linguistic resources remain limited. Unlike English, which benefits from abundant annotated data and well-established NLP tools, these low-resource languages pose unique challenges including code-switching, dialectal variations, and cultural context dependencies. Consequently, the automatic detection of offensive or hateful content in such languages remains a non-trivial task [6].

The HASOC 2025 [7] shared task at the Forum for Information Retrieval Evaluation (FIRE) provides a standardized benchmark for addressing these challenges. It focuses on multimodal meme datasets in multiple languages, with subtasks covering sentiment detection, sarcasm identification, vulgarity recognition, abuse detection, and target community identification task. Our work participates in the Hindi and Gujarati tracks, proposing a BERT-based approach that leverages the power of pre-trained transformer models to handle the linguistic and contextual complexity of these datasets effectively.

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

*Corresponding author.

zpp1219@gmail.com (P. Zhang); glu@ynu.edu.cn (G. Lu)
https://github.com/zpp1219 (P. Zhang)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-02-24

All tasks constitute the HASOC 2025 [8] Meme repertoire in Hindi, Gujarati, Bodo, and Bangla languages. We did not participate in the target community identification task. In addition, our study did not investigate the Bodo and Bangla datasets.

2. Related Work

The rapid expansion of social media platforms has driven extensive research on detecting hate speech, offensive language, and harmful content across multiple modalities and languages. Prior studies can be broadly categorized into three research directions: (1) traditional text-based hate speech detection, (2) multimodal approaches integrating visual and textual cues, and (3) transformer-based methods leveraging large-scale pre-training.

2.1. Text-based Hate Speech Detection

Early research in hate speech detection relied heavily on classical machine learning techniques such as Support Vector Machines (SVM), Random Forests, and Logistic Regression [9, 10]. These methods utilized handcrafted features such as TF-IDF, word n-grams, and sentiment polarity. However, their effectiveness was limited due to the inability to capture contextual and semantic relationships in text, especially in morphologically rich or code-switched languages such as Hindi and Gujarati.

The advent of deep learning models marked a significant shift in this field. Convolutional Neural Networks (CNN) [11] and Recurrent Neural Networks (RNN) [12] enabled automatic feature extraction from text, improving robustness across domains. Despite these advances, these models struggled with understanding long-range dependencies and multi-lingual semantics, which paved the way for transformer-based approaches.

2.2. Multimodal Meme Analysis

In recent years, memes have become a prevalent form of communication on social media, combining both image and text to convey emotion and opinion. This has prompted a new line of research on multimodal hate speech detection [13, 14]. These studies incorporate both visual features (from CNNs or vision transformers) and textual embeddings (from BERT-like models) to understand the implicit and explicit cues in memes.

For example, Kiela et al. [15] (2020) introduced the Hateful Memes Challenge, which emphasized the difficulty of multimodal reasoning where neither image nor text alone is sufficient to infer hatefulness. Similarly, Zhang et al. [16] (2021) proposed a cross-modal fusion model that integrates attention-based representations from both modalities. However, most existing datasets and models are focused on English, leaving regional languages underrepresented in meme-based hate detection.

2.3. Transformer-based Models for Offensive Language

Transformer-based models such as BERT [17], RoBERTa [18], and ALBERT [19] have revolutionized natural language processing by leveraging contextualized embeddings through self-attention mechanisms. These models achieve state-of-the-art performance across a variety of text classification tasks, including hate speech and offensive content detection.

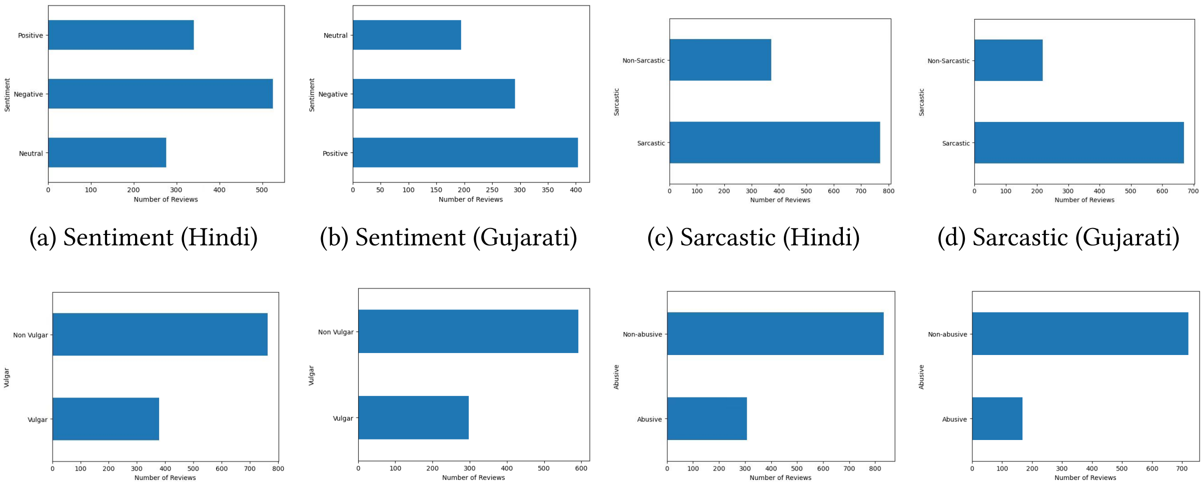
For Indian languages, multilingual pre-trained models such as mBERT and MuRIL [20] have enabled transfer learning across low-resource languages. Moreover, monolingual adaptations such as *HindBERT*, *GujaratiBERT*, and *HindRoBERTa* developed by L3Cube–Pune [21] have shown improved results by capturing language-specific semantics. These models mitigate the challenges posed by code-switching and morphological variations, which are common in Hindi and Gujarati social media discourse.

2.4. HASOC and Indic NLP Research

The HASOC (Hate Speech and Offensive Content) shared tasks under the Forum for Information Retrieval Evaluation (FIRE) series have played a crucial role in promoting research on multilingual and multimodal hate speech detection. The earlier editions (HASOC 2019–2024) primarily focused on text-based datasets in languages such as English, Hindi, and German [22]. Later versions, including HASOC 2025, introduced meme-based multimodal datasets, extending the task to include sentiment, sarcasm, vulgarity, and abuse detection.

Our work builds upon these foundations by evaluating multiple transformer architectures—including BERT, RoBERTa, ALBERT, and DistilBERT—on the HASOC 2025 Hindi and Gujarati meme datasets. Unlike previous studies, which mainly addressed monolingual English data or textual hate speech, we explore the under-studied domain of multimodal meme classification in low-resource Indic languages [23].

3. Exploratory Data Analysis



(a) Sentiment (Hindi) (b) Sentiment (Gujarati) (c) Sarcastic (Hindi) (d) Sarcastic (Gujarati)

(e) Vulgar (Hindi) (f) Vulgar (Gujarati) (g) Abusive (Hindi) (h) Abusive (Gujarati)

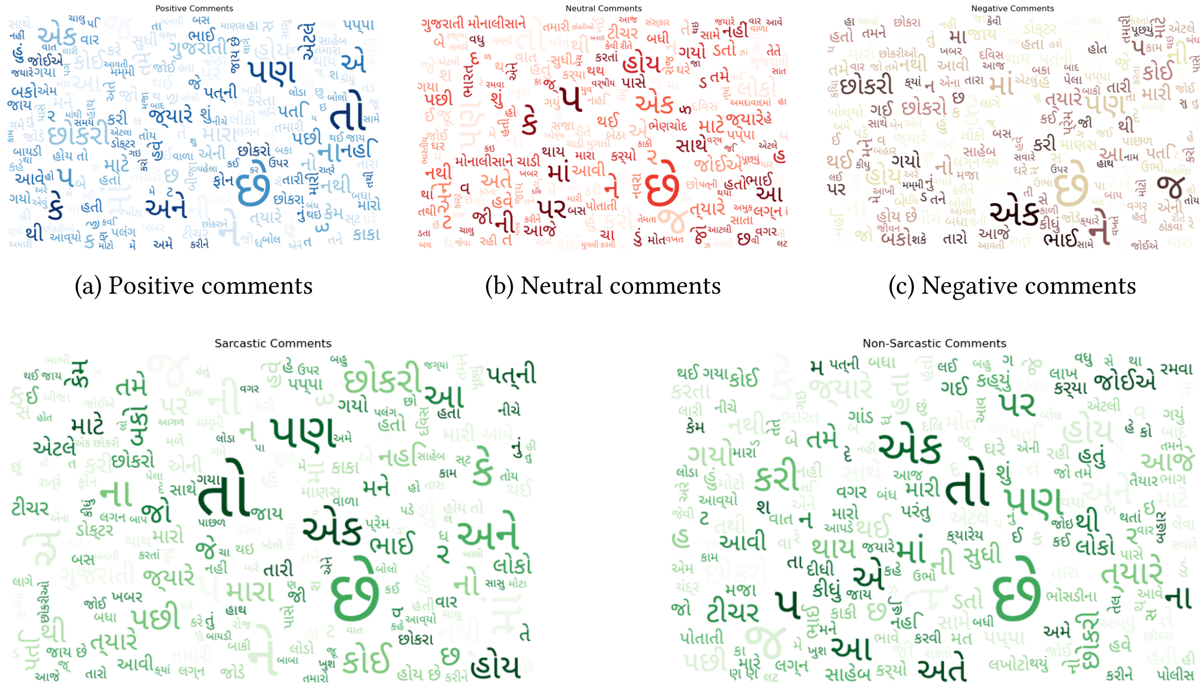
Figure 1: Details of HASOC2025-meme (Hindi) and HASOC2025-meme (Gujarati) datasets across different categories provided by the organizer.

For these tasks, participants are not allowed to use any external resources and datasets. So we only used the training set provided by the official. The test dataset is also provided by the official for the evaluation of models and the submitted result documents. The training and test sets for Hindi and Gujarati languages provided by HASOC 2025-meme organizer, as well as the labels situation in the training sets, are respectively presented in Figure 1.

4. Graphical Representation

The following images are the word-cloud for various types of memes given in the text dataset. It is a graphical representation of word frequency of different words used in each category of the content in memes on official resources and datasets.

The following is the situation of the word cloud generated by sentiment detection and sarcasm detection in the training set for the Gujarati language provided by HASOC 2025-meme organizer. See Figure 2.



(a) Positive comments (b) Neutral comments (c) Negative comments

(d) Sarcastic comments (e) Non-sarcastic comments

Figure 2: Word clouds illustrating representative terms for different sentiment and sarcasm categories in the HASOC2025-meme dataset.

5. Dataset

The number of training datasets and test datasets provided by the official organizers for Hindi and Gujarati languages in the HASOC 2025 competition task, as well as the number of labels for each sub-task include Sentiment detection, Sarcasm detection, Vulgarity detection, and Abuse detection in the training set, are shown in Tables 1.

Table 1
Hindi and Gujarati Language Data Analysis.

Details	Hindi		Gujarati	
	Posts in Train Data	Posts in Test Data	Posts in Train Data	Posts in Test Data
HASOC2025-meme	Total = 1141	769	Total = 889	604
Sentiment detection (Positive)	340		404	
Sentiment detection (Neutral)	276		194	
Sentiment detection (Negative)	525		291	
Sarcasm detection (Sarcastic)	770		670	
Sarcasm detection (Non-Sarcastic)	371		219	
Vulgarity detection (Vulgar)	378		297	
Vulgarity detection (Not Vulgar)	763		592	
Abuse detection (Abusive)	308		168	
Abuse detection (Non-abusive)	833		721	

6. Methodology

6.1. Overview

Our methodology integrates supervised learning techniques with advanced transformer-based architectures for hate speech and offensive content detection in the HASOC 2025 Hindi and Gujarati dataset. Using the Ktrain Training Aid Library, we simplify fine-tuning, hyperparameter optimization, and deployment. Specifically, we introduce the transformer architecture as the foundation, followed used by three representative pre-trained models: BERT, RoBERTa, and ALBERT, which are fine-tuned for classification tasks under the same supervised learning setup. The overall framework flowchart is shown in Figure 3.

6.2. Supervised Learning

The problem is formulated as a supervised learning task, where each text input X is mapped to a corresponding label Y (e.g., hate speech or non-hate speech). The learning objective is to minimize the cross-entropy loss between predicted and true labels:

$$L = -\sum_i y_i \log(\hat{y}_i)$$

This allows the model to learn discriminative features for accurate classification. Supervised Learning is one of the most common types of methods in machine learning. It trains the model with labeled data to enable the model to predict the output from the input.

The following is a systematic introduction to its advantages and disadvantages: The advantages include high prediction accuracy, clear task objectives, mature algorithms, relatively good interpretability, and strong generalization ability (when there are sufficient samples). However, there are also shortcomings. For instance, relying on a large amount of labeled data makes it difficult to handle situations with insufficient labeling, unable to deal with unknown categories, possibly affected by data bias, and the generalization ability depends on the representativeness of the training set. Some methods have poor interpretability.

6.3. Transformer Architecture

The transformer architecture [24] employs self-attention mechanisms and feed-forward layers to model long-range dependencies in text sequences. Unlike recurrent networks, transformers process input tokens in parallel, improving both efficiency and representation quality. Key components include:

- **Multi-head Self-Attention:** Captures contextual relationships between all tokens in a sequence.
- **Positional Encoding:** Preserves word order information otherwise lost in parallel processing.
- **Feed-forward Layers & Layer Normalization:** Enhance representation learning and stability.

This architecture forms the backbone of modern NLP (natural language processing) models, enabling effective transfer learning when combined with large-scale pre-training.

6.4. Pre-train Models

Building on the transformer backbone, we adopt three widely used pre-trained models:

- **BERT (Bidirectional Encoder Representations from Transformers):** Utilizes bidirectional attention and is pre-trained on masked language modeling and next sentence prediction tasks.
- **RoBERTa (Robustly Optimized BERT Pretraining Approach):** Improves upon BERT by using dynamic masking, removing next sentence prediction, and training with larger corpora for longer durations.
- **ALBERT (A Lite BERT):** Reduces parameter size via factorized embedding parameterization and cross-layer parameter sharing, offering computational efficiency while maintaining accuracy.

Each model is fine-tuned on the HASOC 2025 Hindi and Gujarati dataset for comparative evaluation.

6.5. Ktrain Training Aid Library

The Ktrain Training Aid Library provides a high-level interface for training and fine-tuning models in TensorFlow and PyTorch. It automates key steps including data preprocessing, learning rate scheduling, and performance monitoring, accelerating experimentation while ensuring reproducibility [25].

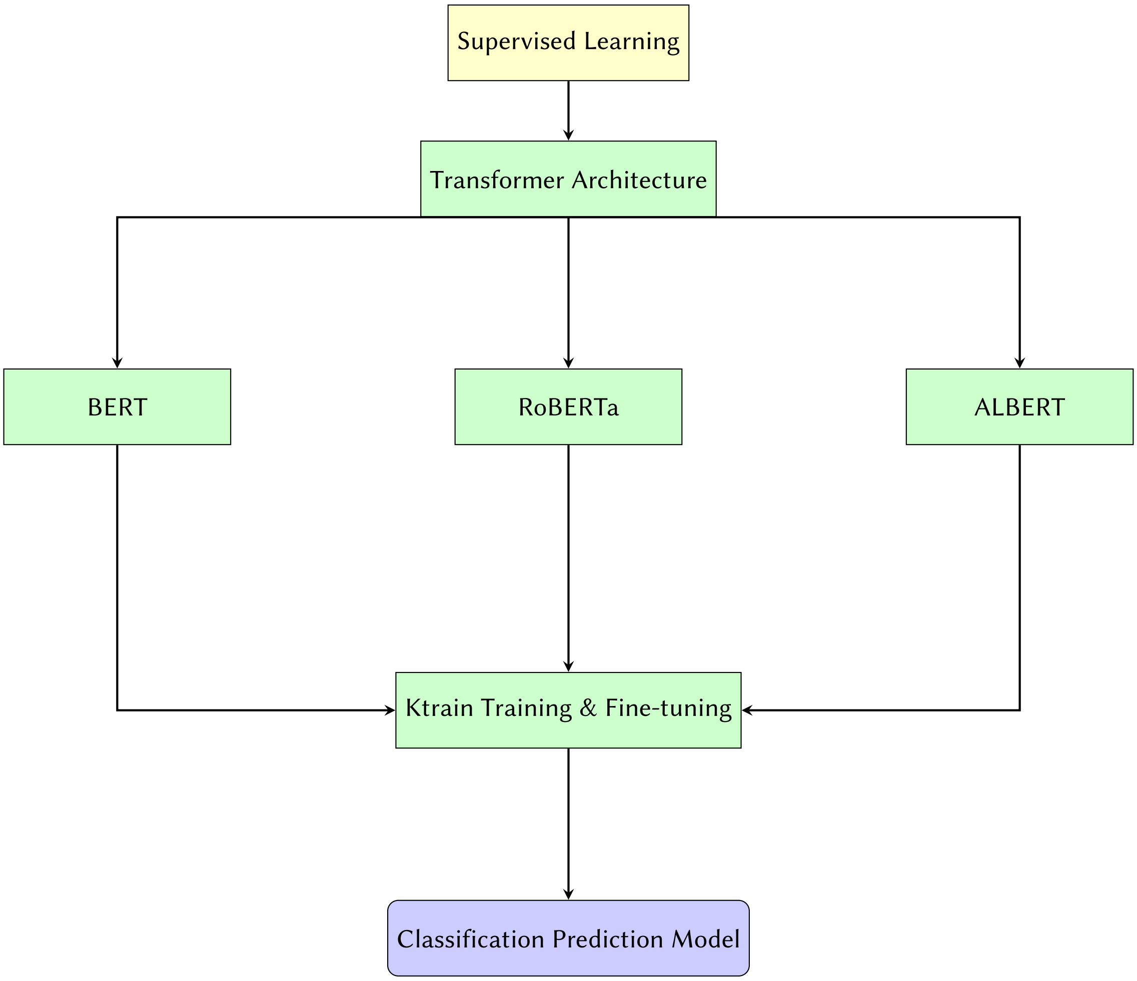

Supervised Learning
Transformer Architecture
BERT
RoBERTa
ALBERT
Ktrain Training & Fine-tuning
Classification Prediction Model

Figure 3: Overall flow and methodology of the experiment.

7. Bert Models

7.1. models–distilbert–distilbert-base-uncased

DistilBERT [26] is a lightweight transformer model that is smaller and faster than BERT while retaining most of its performance. It is pre-trained on the same corpus in a self-supervised manner, using the BERT-base model as a teacher. This means that no human-labeled data is required, the model leverages raw text and an automatic process to generate inputs and labels, allowing it to utilize large publicly available datasets efficiently. Specifically, DistilBERT is pre-trained with three objectives: **Distillation Loss:** The model is trained to produce output probabilities similar to the BERT-base teacher model, effectively transferring knowledge from the larger model. **Masked Language Modeling (MLM):** Following BERT's original pre-training, 15% of input tokens are randomly masked, and the model predicts these masked tokens using the full bidirectional context. Unlike traditional RNNs, which process tokens sequentially, or autoregressive models like GPT, which predict future tokens, MLM enables the model to learn deep bidirectional representations. **Cosine Embedding Loss:** The model is encouraged to generate hidden states that closely match those of the BERT-base model, further aligning its internal representations with the teacher.

This combination of objectives allows DistilBERT to achieve high efficiency with minimal performance loss, making it particularly suitable for resource-constrained environments. In some public datasets, its execution effect is shown in Table 2.

Table 2
DistilBERT model performance table. The task metric is classification accuracy.

Task | Dataset	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE
Classification	82.2	88.5	89.2	91.3	51.3	85.8	87.5	59.9

7.2. models–l3cube-pune–hindi-bert-v2

HindBERT is a Hindi-specific BERT model designed for natural language processing tasks in the Hindi language. It is based on the multilingual BERT architecture (Google's mBERT / MuRIL-base-cased) and has been fine-tuned on publicly available Hindi monolingual corpora to better capture the linguistic characteristics and nuances of Hindi text.

Unlike standard multilingual BERT, which is pre-trained on text from multiple languages and may not capture language-specific patterns effectively, HindBERT leverages monolingual fine-tuning to enhance performance for Hindi-specific tasks. This enables the model to learn richer semantic and syntactic representations that are more aligned with the Hindi language.

HindBERT has been successfully applied to various NLP tasks such as text classification, sentiment analysis, and offensive language detection in Hindi. Fine-tuning on task-specific datasets allows the model to adapt to domain-specific language usage and achieve improved classification accuracy.

The model is publicly available and can be accessed via HindBERT project link.

7.3. models–l3cube-pune–hindi-roberta

HindRoBERTa is a Hindi-specific RoBERTa model derived from the multilingual xlm-roberta-base and fine-tuned on publicly available monolingual Hindi corpora. This adaptation enables the model to better capture linguistic nuances of Hindi, outperforming general multilingual counterparts on Hindi-specific tasks .

Developed by the L3Cube–Pune team, HindRoBERTa is part of their suite of Indic language models, which includes HindBERT, HindAlBERT, and DevBERT, designed to address diverse Hindi and Devanagari-script tasks.

7.4. models–l3cube-pune–hindi-albert

HindAlBERT is a Hindi-specific ALBERT model adapted from the multilingual MuRIL-base-cased transformer and further fine-tuned on publicly available monolingual Hindi corpora. This targeted adaptation enables the model to better capture the unique linguistic characteristics of Hindi while preserving ALBERT's architectural efficiency.

As part of L3Cube–Pune's suite of Indic models, HindAlBERT offers enhanced performance on Hindi downstream tasks compared to general-purpose multilingual models. It complements related models such as Hindi BERT, Hindi RoBERTa, and DevBERT, which is a Devanagari-script model trained jointly on Hindi and Marathi data.

7.5. models–microsoft–xtremedistil-l6-h256-uncased

XtremeDistilTransformers is a distilled, task-agnostic transformer model that leverages task transfer to learn a small universal model, which can be applied to various tasks and languages.

The model combines task transfer with multi-task distillation techniques from the papers XtremeDistil: Multi-stage Distillation for Massive Multilingual Models [27] and MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers [28], implemented using the following GitHub code[1].

[1] GitHub: https://github.com/microsoft/xtreme-distil-transformers

The l6-h256 checkpoint, with 6 layers, a hidden size of 256, and 12 attention heads, has millions parameters and provides a 5.3× speedup over BERT-base. Other available checkpoints include xtremedistil-l6-h384-uncased and xtremedistil-l12-h384-uncased.

7.6. models–microsoft–xtremedistil-l6-h384-uncased

XtremeDistilTransformers is a distilled, task-agnostic Transformer model designed to learn a small universal model that can be applied to a wide range of tasks and languages. It leverages task transfer and multi-task distillation techniques to compress large pre-trained transformers into efficient models while maintaining high performance across tasks.

The model builds on insights from several distillation approaches, including XtremeDistil: Multi-stage Distillation for Massive Multilingual Models and MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, combining task transfer with deep self-attention distillation to produce a lightweight yet versatile model.

We employ the l6-h384 checkpoint, which contains 6 layers, a hidden size of 384, and 12 attention heads, corresponding to 22 million parameters. This model achieves a 5.3× speedup over BERT-base, making it suitable for resource-constrained environments. Other available checkpoints include xtremedistil-l6-h256-uncased and xtremedistil-l12-h384-uncased.

The model demonstrates competitive performance on standard NLP benchmarks such as GLUE and SQuAD-v2, confirming its effectiveness for both classification and question-answering tasks.

7.7. models–microsoft–xtremedistil-l12-h384-uncased

XtremeDistilTransformers is a distilled, task-agnostic Transformer model that leverages task transfer to learn a small universal model applicable to a wide range of tasks and languages, as described in the paper XtremeDistilTransformers: Task Transfer for Task-Agnostic Distillation [29].

The model combines task transfer with multi-task distillation techniques from the papers XtremeDistil: Multi-stage Distillation for Massive Multilingual Models and MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, using the implementation provided in the associated HuggingFace repository[2].

The l12-h384 checkpoint, which has 12 layers, a hidden size of 384, and 12 attention heads, contains millions parameters and provides a 5.3× speedup over BERT-base. Other available checkpoints include xtremedistil-l6-h256-uncased and xtremedistil-l6-h384-uncased.

The comparison results of the relevant models in terms of speed and params. When the classification task metric of the public dataset is classification accuracy, their execution performance are also shown in Table 3.

Table 3
Model Performance Comparison.

Models	#Params	Speedup	MNLI	QNLI	QQP	RTE	SST	MRPC	SQUAD2	Avg
BERT	109	1x	84.5	91.7	91.3	68.6	93.2	87.3	76.8	84.8
DistilBERT	66	2x	82.2	89.2	88.5	59.9	91.3	87.5	70.7	81.3
TinyBERT	66	2x	83.5	90.5	90.6	72.2	91.6	88.4	73.1	84.3
MiniLM	66	2x	84.0	91.0	91.0	71.5	92.0	88.4	76.4	84.9
MiniLM	22	5.3x	82.8	90.3	90.6	68.9	91.3	86.6	72.9	83.3
XtremeDistil-l6-h256	13	8.7x	83.9	89.5	90.6	80.1	91.2	90.0	74.1	85.6
XtremeDistil-l6-h384	22	5.3x	85.4	90.3	91.0	80.9	92.3	90.0	76.6	86.6
XtremeDistil-l12-h384	33	2.7x	87.2	91.9	91.3	85.6	93.1	90.4	80.2	88.5

[2] HuggingFace: https://huggingface.co/microsoft/xtremedistil-l12-h384-uncased

7.8. models–l3cube-pune–gujarati-bert

GujaratiBERT is a Gujarati BERT model trained on publicly available Gujarati monolingual datasets. Preliminary details on the dataset, models, and baseline results can be found in their paper [30].

The monolingual Hindi BERT models currently available on the model hub do not perform better than the multi-lingual models on downstream tasks. They present L3Cube-HindBERT, a Hindi BERT model pre-trained on Hindi monolingual corpus. Further, since Indic languages [31], Hindi and Marathi share the Devanagari script, they train a single model for both languages. They release DevBERT, a Devanagari BERT model trained on both Marathi and Hindi monolingual datasets. They evaluate these models on downstream Hindi and Marathi text classification and named entity recognition tasks. The HindBERT and DevBERT-based models show significant improvements over multi-lingual MuRIL, IndicBERT, and XLM-R. Based on these observations they also release monolingual BERT models for other Indic languages Kannada, Telugu, Malayalam, Tamil, Gujarati, Assamese, Odia, Bengali, and Punjabi.

8. Result

In our Hindi language task, since the single-language pre-trained model was fully trained in the relevant hindi language content, our group 8th on HASOC2025-meme (Hindi) with a score of 0.57094. The performance of these models proposed by our group on the test set is shown in Table 4.

Table 4
Overall HASOC2025-meme (Hindi) test results based on the average of all four Macro F1 score for our proposed model.

Model	Macro-F1
models–distilbert–distilbert-base-uncased	0.54888
models–l3cube-pune–hindi-bert-v2	0.57094
models–l3cube-pune–hindi-roberta	0.52088
models–l3cube-pune–hindi-albert	0.49045

In our Gujarati task, since the monolingual pre-trained model is not sufficiently trained on the relevant language content of Gujarati. This leads to better results using multilingual models such as models–microsoft–xtremedistil-l6-h256-uncased[3]. Ultimately, our group was ranked 11th on HASOC2025-meme (Gujarati) with a score of 0.55522. The performance of the model proposed by our group on the test set is shown in Table 5.

Table 5
Overall HASOC2025-meme (Gujarati) test results based on the average of all four Macro F1 score for our proposed model.

Model	Macro-F1
models–distilbert–distilbert-base-uncased	0.51366
models–microsoft–xtremedistil-l6-h256-uncased	0.55522
models–microsoft–xtremedistil-l6-h384-uncased	0.54059
models–microsoft–xtremedistil-l12-h384-uncased	0.45715
models–l3cube-pune–gujarati-bert	0.48358

It is worth mentioning that after our previous presentation of category labels, we found that there is a problem of category imbalance in the classification of these texts identified from memes images. To solve these problems, we introduced the technologies of class_weights and RandomOverSampler, which enabled these models to perform better.

[3] HuggingFace: https://huggingface.co/microsoft/xtremedistil-l6-h256-uncased

9. Conclusion

Our approach mainly introduces category weights when addressing the issue of category imbalance, giving larger weights to a few labeled samples and smaller weights to the majority labeled samples. This enables the model to better identify the samples in the test set after learning from these samples in the training set. We also employed RandomOverSampler to balance the original train dataset in the model's training process. These methods have been experimentally proven to be effective on the average of all four Macro F1 score metric for these tasks.

During the process of training the model, the first 8 layers are frozen first to train only the classification heads, and then all layers are thawed for fine-tuning. This enables the model to predict better on the test set samples. Introducing the strategies of EarlyStopping and ReduceLROnPlateau during the training process of the models enables the models we proposed to achieve the best results.

Acknowledgments

Thanks to the support of the official open source packages ktrain and keras, as well as the release of pre-training distillation version Bert. We can achieve good results with simple fine-tuning after the model has been pre-trained.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[2] M. Das, A. Mukherjee, Banglaabusememe: A dataset for bengali abusive meme classification, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 15498–15512.

[3] A. Parikh, H. Desai, A. S. Bisht, Da master at hasoc 2019: Identification of hate speech using machine learning and deep learning approaches for social media post (2019). URL: http://ceur-ws.org/Vol-2517/T3-18.pdf.

[4] K. Ghosh, S. Saha, T. Mandl, S. Modha, Findings from shared tasks on hate speech detection: Performance patterns for low-resource languages, Pattern Recognition Letters (2025). URL: https://www.sciencedirect.com/science/article/pii/S0167865525003150. doi:https://doi.org/10.1016/j.patrec.2025.09.004.

[5] M. Yi, Myung, J. Lim, H. Ko, J. Shin, Method of profanity detection using word embedding and lstm (2021). URL: https://doi.org/10.1155/2021/6654029.

[6] S. Modha, T. Mandl, P. Majumder, D. Patel, Tracking hate in social media: Evaluation, challenges and approaches (2020). URL: https://link.springer.com/article/10.1007/s42979-020-0082-0.

[7] K. Ghosh, M. Das, M. Narzary, S. Saha, S. Barman, A. Mukherjee, S. Modha, D. Ganguly, U. Garain, S. Jaki, T. Mandl, Overview of the hasoc track at fire 2025: Abusive meme identification — shadows behind the laughter, in: K. Ghosh, T. Mandl, S. Pal, S. Majumdar, A. Chakraborty (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2025) December 17-20, Varanasi , India, CEUR-WS.org, 2025.

[8] K. Ghosh, M. Das, S. Patel, N. Bhandary, A. Das, A. Mukherjee, S. Modha, D. Ganguly, U. Garain, S. Jaki, T. Mandl, Overview of the hasoc track at fire 2025: Abusive meme identification — shadows

behind the laughter, in: FIRE '25: Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation. December 17-20, Varanasi , India, Association for Computing Machinery (ACM), New York, NY, USA, 2025.

[9] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM), 2017, pp. 512–515.

[10] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (2017) 1–10.

[11] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on twitter using a convolution–gru based deep neural network, in: European Semantic Web Conference, Springer, 2018, pp. 745–760.

[12] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 759–760.

[13] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, D. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems (NeurIPS), 2020.

[14] S. Pramanick, A. Sharma, M. Das, U. Garain, T. Mandl, S. Modha, T. Chakraborty, Detecting harmful memes and their targets, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4843–4851.

[15] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS) Workshop on Multimodal Learning, 2020. URL: https://hatefulmemeschallenge.com/.

[16] L. Zhang, L. Chen, C. Zhou, F. Yang, X. Li, Exploring graph-structured semantics for cross-modal retrieval, in: Proceedings of the 29th ACM International Conference on Multimedia, MM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 4277–4286. URL: https://doi.org/10.1145/3474085.3475567. doi:10.1145/3474085.3475567.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, NAACL-HLT (2019) 4171–4186.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[19] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, International Conference on Learning Representations (ICLR) (2020).

[20] S. Khanuja, S. Dandapat, A. Srinivasan, K. Bali, M. Choudhury, Muril: Multilingual representations for indian languages, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2108–2129.

[21] O. Gokhale, A. Kane, S. Patankar, T. Chavan, R. Joshi, Spread love not hate: Undermining the importance of hateful pre-training for hate speech detection, arXiv preprint arXiv:2210.04267 (2022).

[22] S. Modha, T. Mandla, D. Nandini, D. Patel, P. Majumder, Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages, in: Forum for Information Retrieval Evaluation (FIRE 2020), 2020, pp. 29–42.

[23] M. Das, S. Banerjee, A. Mukherjee, Data bootstrapping approaches to improve low resource abusive language detection for indic languages, in: Proceedings of the 33rd ACM conference on hypertext and social media, 2022, pp. 32–42.

[24] R. Child, S. Gray, A. Radford, I. Sutskever, Generating long sequences with sparse transformers, Machine Learning (2019). URL: https://arxiv.org/abs/1904.10509. doi:https://doi.org/10.48550/arXiv.1904.10509.

[25] J.-C. Mensonides, P.-A. Jean, A. Tchechmedjiev, S. Harispe, Imt mines ales at hasoc 2019: Automatic hate speech detection (2019). URL: http://ceur-ws.org/Vol-2517/T3-13.pdf.

[26] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2020). URL: https://arxiv.org/pdf/1910.01108.pdf.

[27] S. Mukherjee, A. Hassan Awadallah, XtremeDistil: Multi-stage distillation for massive multilingual models, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2221–2234. URL: https://aclanthology.org/2020.acl-main.202/. doi:10.18653/v1/2020.acl-main.202.

[28] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL: https://arxiv.org/abs/2002.10957. arXiv:2002.10957.

[29] S. Mukherjee, A. H. Awadallah, J. Gao, Xtremedistiltransformers: Task transfer for task-agnostic distillation, 2021. arXiv:2106.04563.

[30] R. Joshi, L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages, arXiv preprint arXiv:2211.11418 (2022).

[31] K. Ghosh, N. K. Singh, J. Mahapatra, et al., Safespeech: a three-module pipeline for hate intensity mitigation of social media texts in indic languages, Social Network Analysis and Mining 14 (2024). URL: https://doi.org/10.1007/s13278-024-01393-9. doi:10.1007/s13278-024-01393-9.