International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(7s) 2026 pp. 894-905
https://doi.org/10.70917/ijcisim-2026-3161
Article

A BERT-Based Deep Learning Framework for Multimodal Hate Speech Detection

Anupama Jha[1], Alpna Sharma[2], Shailee Bhatia[3]

[1,2,3]VIPS-TC, GGSIP University, Delhi

Abstract: In the recent decade, with the development of social media applications, there has been an increasing dispersion of harmful and abusive content across various languages and media formats. And these contents often appear in multimodal forms where textual messages are in the combination with images, memes, or other visual elements. Prior studies reveal that traditional text-based detection methods often fail to capture the complex relationships among these modalities, making it difficult to identify implicit or context-dependent harmful expressions.

Through this study, dealing with the above challenges in mind, we have made a humble attempt to implement a transformer-based multimodal deep learning framework for detecting toxic/hate speech online content.

Two proposed approaches, Bidirectional Encoder Representations from Transformers (BERT) & multilingual BERT (mBERT), have been made for the extraction of textual features, while deep vision-based models have been made for visual features extraction. The outcome of both textual and visual representations are then incorporated by using multimodal fusion techniques for better understanding of complex patterns in online content.

The proposed model is evaluated on datasets specifically designed for the research domain of hate speech and harmful content detection, such as MMHS150K and the Facebook Hateful Memes datasets. The outcomes of this study show that the suggested multimodal approach is much more efficient compared to conventional approaches which employ only one channel. Using both textual and visual channels allows capturing contextual relationships and hints, like understanding sarcasm. Also, this approach supports safer, more reliable moderated systems that improve the identification of harmful and culturally nuanced content on social media.

Keywords: Multimodal Technique, Hate Speech Recognition, BERT, Multilingual- BERT(mBERT), Vision–Language Fusion, NLP, Social Media Analytics

1. Introduction

With the development of social media applications, it has been easy for people around the world to connect instantly; however, it has also led to a quicker spread of harmful, abusive, and hateful content online. Social media allows people to create and share a large amount of content every day. Because so much content is posted quickly, it is hard to monitor and control harmful material. As a result we are facing a big challenge for platforms, policymakers, and researchers in the field of Natural Language Processing (NLP), for detecting offensive, abusive or hateful content in different languages and formats[[endnoteRef:1]].

In social media environments, toxic, offensive and hate speech often appear in multimodal forms such as images, memes, emojis, and other visual elements that work together with text to express the meaning. Due to the fact that many of the traditional machine learning methods are primarily focused on textual features[[endnoteRef:2]], it makes it difficult to discover hidden or implicit forms of toxicity. In addition to that, these approaches might have difficulties in interpreting contents that convey damaging meaning through both text and visuals, as well as mixed languages and culturally distinct idioms[[endnoteRef:3]].

Detection models that can simultaneously analyse textual and visual data are becoming more and more necessary to get around these restrictions. By coalescing computer vision methods with language models, multimodal deep learning algorithms emerge as a noble solution. They would also be able to comprehend the complex patterns observed in multimodal web data using textual representation extracted by transformer-based models combined with visual feature extraction networks.

Two popular benchmark datasets, such as MMHS150K and the Facebook Hateful Memes dataset, have significantly contributed to the development of multimodal hate speech detection models capable of analysing text, images, and multilingual information together. Beyond these developments, we are still facing a number of challenges that need to be resolved, including identifying implicit toxicity, dealing with model bias, and managing cultural variations in the definition of harmful online behaviour [[endnoteRef:4]].

Therefore, this study aims to develop a robust multimodal and multilingual hate speech detection framework using deep learning techniques to improve detection accuracy and enhance cross-cultural generalisation.

2. Research Motivation

In this era of big data, a huge amount of data, also known as multimodal data, is generated in the form of images, videos, audio, and text, are packed with very high value insights. Analysing such kind of data is not effectively analysed with the traditional data processing techniques. Multimodal Deep learning has come out as a powerful technique for learning complex patterns from big data. However, analysing such big data remains a challenging research problem due to computational complexity. With this research work, we attempt to explore and develop effective multimedia deep learning models that can learn meaningful representations from multiple data modalities which improve the accuracy, robustness, and efficiency of multimedia data analysis by designing optimized deep learning architectures and can tackle the growing prevalence of multimodal data and the strengths of advanced deep learning approaches. The outcome of this research can contribute to academicians and researchers for analysing real-world applications.

3. Literature Survey

Due to the development in machine learning and deep learning techniques in recent years, hate speech detection has changed dramatically. In an effort to detect abusive language in social media posts, early methods mostly relied on conventional machine learning algorithms that employed manually created linguistic features, which often missed complex meanings and contextual relationships in online content. Text mining technology is more effective with the advent of transformer-based architectures and deep neural networks. The massive amount of multimedia data being produced on social media platforms has led to multimodal methods being prevalent today. This is because multimodal frameworks combine text and image analysis to increase the effectiveness of hate speech detection.

3.1 Text Based Hate Speech Detection and Unimodal Architectures

Text based analysis, in which models analysed just textual input, was the primary focus of early research on hate speech detection. Davidson et al. (2017) [[endnoteRef:5]] implemented using supervised machine learning methods to identify the sentiments of Twitter hate speech.

Because they process and interpret data from a single data modality, these early systems can be classified as unimodal architectures. The unimodal techniques in NLP applications, such as RNNs and LSTM(Long Short Term Memory) networks, are frequently employed to capture sequential dependencies in textual data in applications like text mining, sentiment analysis, and the detection of abusive language.

A unimodal architecture is a kind of machine learning framework that processes and analyses data from a single data modality—such as text, pictures, audio, or video. Models in deep learning and classic machine learning systems are usually built to handle a single kind of data at a time. For instance, picture recognition systems examine visual data, but text classification algorithms only consider textual inputs. To carry out certain prediction or classification tasks, these models discover patterns and correlations within a single data source [[endnoteRef:6]].

To extract semantic and contextual information from text data, unimodal systems frequently use models like recurrent neural networks (RNNs), long short-term memory networks (LSTMs), or transformer-based designs like BERT in natural language processing tasks [[endnoteRef:7]]. Convolutional neural networks (CNNs) and deep residual networks are also frequently employed in computer vision applications to process picture input and recognise visual patterns. With various tasks including sentiment analysis, object detection, and image classification, these models have proven to perform well [[endnoteRef:8]] [[endnoteRef:9]].

Unimodal techniques have several limitations despite their value, as content in social media conversion like Sarcasm, irony, slang terms, and code-mixed language, can make interpretation difficult.

3.2 Transformer Based Language Models

The performance of NLP (natural language processing) systems has significantly improved with the introduction of transformer-based designs.

The transformers are the most appropriate when looking to establish the connections within context and distant connections due to the use of the self-attention model. Some of the transformer models such as Bidirectional Encoder Representations from Transformers (BERT) and multilingual BERT (mBERT) are generally useful when mining multilingual texts or detecting hate speech.

Transformer-based models such as BERT and mBERT are ideal when it comes to multilingual text mining, sentiment analysis, and hate speech detection, because they are trained using multiple languages at once [[endnoteRef:10]].

3.3 BERT Based Textual Feature Extraction

One of the key models within the scope of NLP, particularly with respect to detection of hate speech, is that presented by Devlin et al. (2019) [11]. The model understands the actual meaning of texts and comprehends the complete meaning of the sentence through bidirectional processing, i.e., from left to right and vice versa. This capability allows the model to capture semantic and syntactic relationships more effectively, making it particularly useful for analysing informal language used in social media environments.

The methodology includes several steps, including Text pre processing (Cleaning, Normalization & tokenisation), contextual embedding (tokens into numbers) and the use of [CLS] token (Summary of the text).

In the proposed framework, BERT is used to extract meaningful textual features from user-generated content. The text processing pipeline generally begins with a preprocessing stage in which the raw text is cleaned and normalised. This step may include converting text to lowercase, removing unwanted characters such as punctuation, URLs, and special symbols, and performing tokenisation using the BERT tokenizer. The tokenizer splits the input text into subword units that are compatible with the BERT architecture.

After tokenisation, the input sequence is passed through the BERT model to obtain contextual embeddings. In many text classification tasks, the representation corresponding to the special [CLS] token is used as a global representation of the entire text sequence. This vector captures the contextual meaning of the sentence and can be used as input for downstream classification layers. Fine-tuning the BERT model on domain-specific datasets further improves its ability to recognise subtle patterns of abusive or hostile language commonly observed in online discussions.

3.4 Multimodal Hate Speech Detection

Because memes and multimedia posts are becoming more and more popular, damaging messages are often communicated by combining textual and visual components. Researchers/Academicians have been investigating multimodal learning frameworks which incorporate data from other modalities, including text and images, in an effort to overcome this constraint.

As previously mentioned, the transformer based models like BERT or RoBERTa are frequently used to extract textual features, whereas image processing networks like CNNs (Convolutional Neural Networks), ResNet, or Vision Transformers are used to acquire visual data.

Such models utilize high level feature representations to improve the performance. Three main stages are involved in fusion technique:

• In early fusion, low level characteristics from various modalities are merged before model training.

• During the learning process, learnt representations from several modalities are combined by intermediate fusion.

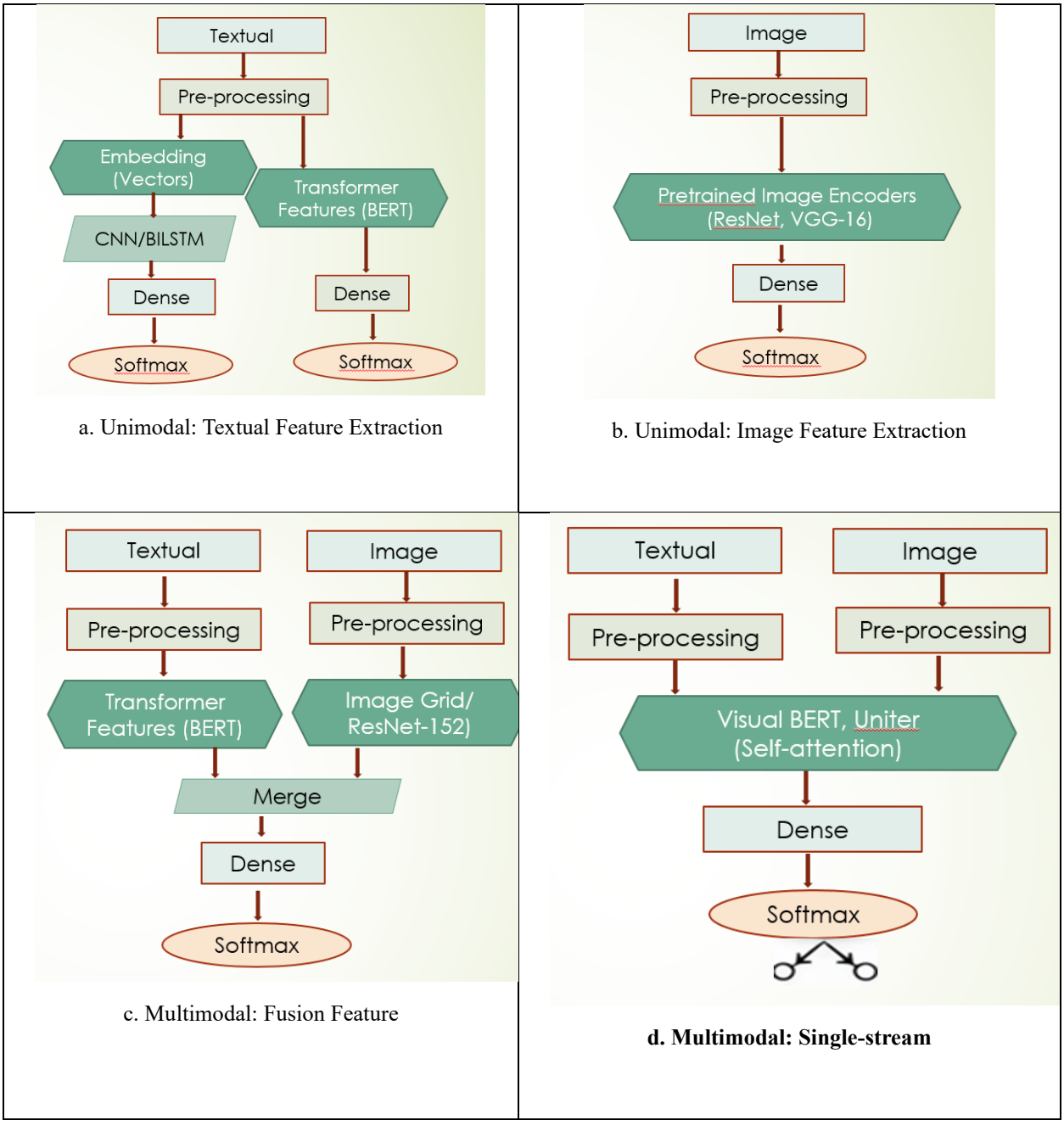
• Late fusion is the process of combining predictions from individual modality specific models.

This fusion method builds up data understanding by identifying complementary links between textual and visual information.

A number of standard datasets have been established to support multimodal research. Among these is the Facebook Hateful Memes dataset, which was put forth by Kiela et al.(2020) and includes memes with hateful content annotations [[endnoteRef:11]].

Vision-language models (VLMs) like VisualBERT, ViLT, and CLIP, which are intended to learn joint representations from both text and images, have also been the subject of recent research. The diagram in figure 1 below showing the architecture of Unimodal and Multimodel Deep learning framework.

Figure 1: Architecture of Unimodal and Multimodal Deep Learning Framework showing text and image feature extraction and fusion.



a. Unimodal: Textual Feature Extraction

b. Unimodal: Image Feature Extraction

c. Multimodal: Fusion Feature

d. Multimodal: Single-stream

Table 1 below summarizes the review of some of the research papers studied based on different deep learning approaches used for Hate Speech Detection Analysis in the recent years.

TABLE 1: LIST OF EXISTING APPROACHES AND THEIR KEY ATTRIBUTES USED IN HATE SPEECH DETECTION

Author(s) & Year	Method / Model	Dataset	Key Contribution	Limitation
Davidson et al. (2017)	TF-IDF + SVM / Logistic Regression	Twitter Hate Speech Dataset	Early machine learning approach for detecting hate speech and offensive language	Limited ability to capture context and sarcasm
Devlin et al. (2019)	BERT Transformer Model	Large Pre-training Corpora	Introduced contextual embeddings that significantly improved NLP tasks	Primarily text-focused; not designed for multimodal data
Conneau et al. (2020)	Multilingual Transformer (mBERT / XLM-R)	Multilingual Corpora	Enabled cross-lingual representation learning for multiple languages	Performance may vary for low-resource languages
Kiela et al. (2020)	Multimodal Baselines (CNN + BERT)	Facebook Hateful Memes Dataset	Introduced benchmark dataset for multimodal hate speech detection	Complex contextual memes remain difficult
Gomez et al. (2020)	Multimodal Deep Learning	MMHS150K Dataset	Large dataset combining text and image information from social media	Dataset imbalance and annotation challenges
Li et al. (2021)	VisualBERT	Hateful Memes Dataset	Joint modelling of textual and visual features using transformer architecture	High computational requirements
Kim et al. (2021)	ViLT (Vision-Language Transformer)	Multimodal datasets	Faster vision-language transformer without heavy image feature extraction	Requires large training data

Radford et al. (2021)	CLIP	Large-scale image-text dataset	Powerful joint image–text representation learning	Not specifically designed for hate speech tasks
Suryawanshi et al. (2020)	Multimodal Abuse Detection	MAMI Dataset	Dataset and benchmark for detecting misogynous and abusive memes	Focus limited to specific abuse categories
Recent Multimodal Studies (2022-2024)	BERT + CNN / Vision Transformers	MMHS150K, Hateful Memes	Integration of textual and visual features improves hate speech detection accuracy	Challenges remain in detecting implicit or culturally nuanced hate

4. Dataset Description

This research study utilises two datasets—MMHS150K and Facebook Hateful Memes (FB Hate)—which provide complementary multimodal and multilingual features that are crucial for constructing and testing strong hate speech recognition models [[endnoteRef:12]].

4.1 MMHS150K

The MMHS150K dataset features over 150,000 meme instances acquired from twitter API, with each sample comprising a picture alongside its associated textual component. The samples that we have used in this study, are classified into two categories—hateful and non-hateful—making the dataset ideally suited for large-scale investigations on multimodal hate speech identification and meme understanding.

4.2 Facebook Hateful Memes

The Facebook Hateful Memes dataset includes ten thousand meme samples in order to guarantee class balance for fair evaluation. Each sample consists of an image–caption pair, facilitating multimodal analysis. This dataset is used as a standard benchmark for evaluating models that humbly attempt to capture subtle or implicit malicious intent initiating from the interaction between visual and textual information.

5. Proposed Methodology

In this research study, we humbly attempt to propose a methodology MMTox - A Transformer-Based Multimodal Framework for detection of online Toxic and Hate Speech Content. The following steps have to be proposed for:

5.1 Tokenization Using BERT/mBERT- Algorithmic Steps:

i) Text Normalization First, the input text sequence X={x1,x2,…,xn}, is translated into a standardised representation. This method includes

a) Converting all characters to lowercase (for uncased versions),

b) Removing control or non-printable characters, and

c) Performing whitespace normalisation to maintain uniform token bounds.

ii) Basic Tokenization

Based on punctuation and whitespace, divide the normalized text into preliminary tokens T = {t1, t2,..., tm}. The normalized text is divided into a preliminary token set T={t1,t2,…,tm} by splitting on whitespace and punctuation marks.

iii) Subword Tokenization (WordPiece)

Each token ti is then broken into smaller subword components using a fixed, preset vocabulary (V). The subwords (sj) are selected to optimize the likelihood:

$$t_i = \arg \max_{s_1,\ldots,s_k \in V} \prod_{j=1}^{k} P(s_j)$$

where P(sj) denotes the probability of the subword learned during pretraining. Unknown tokens are divided into smaller subwords to avoid out-of-vocabulary concerns.

iii) Special Token Insertion

Add special tokens for the sequence structure:

a. *[CLS] is at the beginning*

b. *[SEP] is at the end or between sentence pairs Finally, the token sequence:*

$$T' = \{[CLS], s_1, s_2, \ldots, s_k, [SEP]\}$$

iv) Token-to-ID Mapping

Map each token (si) to a unique integer ID:

$$\text{input\_ids} = \{id(s_1), id(s_2), \ldots, id(s_k)\}$$

v) Attention Mask Generation

Create a mask (A) distinguishing real tokens from padding:

$$A_i = \begin{cases} 1, & \text{if token is valid} \\ 0, & \text{if token is padding} \end{cases}$$

vi) Segment (Token Type) Embeddings

For sentence-pair challenges, assign embeddings Si ∈ {0,1} to indicate sentence membership.

Mathematical Foundation

The input token embeddings $\boldsymbol{Ei}$ fed into BERT/mBERT are created by the total of three embeddings, by which token, location, and segment information are encoded.

The input representation in BERT

Ei = Token Embedding + Segment Embedding + Position Embedding

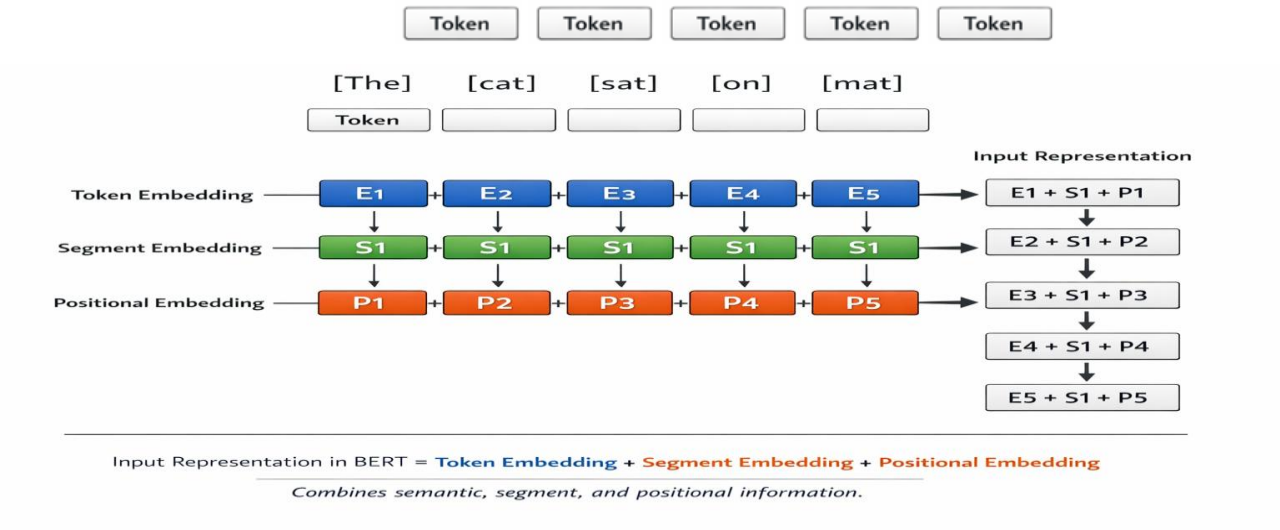
Where,

- **Token Embeddings (E)**: Capture the semantic meaning of each word.
- **Segment Embeddings (S)**: Specify the sentence or segment a token belongs to, assisting jobs with numerous sentences.
- **Positional Embeddings (P)**: Encode the position of each token within the sequence, preserving word order.

The sentence is converted into tokens like that:

Sentence: Hate speech undermines social cohesion

Tokens: [Hate] [speech] [undermines] [social] [cohesion]



The sequence E1, E2, ….. , EL is then passed into the Transformer encoder for contextual representation.

5.2 Data Preprocessing

Text Processing

The text is tokenized using the BERT/mBERT tokenizer to construct context-aware subword tokens. Special characters are abolished, whereas emojis are maintained since they transmit substantial emotional or implicit hate information. Additionally, code-mixed text is normalised to avoid errors in spelling, transliteration, and script usage.

Image Processing

All meme photos are downsized to 224 × 224 pixels to fit with the input specifications of ResNet-based systems. They are then adjusted using standard mean–variance scaling to stabilise the training process and increase model robustness. Additionally, to reduce overfitting and increase generalisation, data augmentation techniques—including random horizontal flips, rotations, scaling, and colour jittering—are applied during training.

5.3 Model Architecture

5.3.1 Text-Only Baseline Model

The text-processing component leverages mBERT to obtain multilingual contextual representations. These representations are subsequently fed into fully connected layers and a Softmax classifier. This text-only model is utilized as a baseline for comparing the effectiveness of visual and multimodal feature integration.

5.3.2 Image-Only Model

Visual features are extracted using **ResNet50**, pretrained on ImageNet. The final convolutional features are flattened and fed into dense layers. A Softmax output layer predicts the hate speech category based solely on image content11.

5.3.3 Multimodal Fusion Model

Two fusion strategies are implemented to integrate text and image modalities:

5.3.3.1 **Early Fusion:** Text embeddings from mBERT and image embeddings from ResNet50 are concatenated at the feature level. The combined feature vector is passed through dense layers and classified via Softmax.

5.3.3.2 **Late Fusion:** Independent classifiers are trained for text and image modalities.

The final predictions or logits from both classifiers are merged (e.g., weighted averaging or concatenation). A meta-classifier produces the final label [[endnoteRef:13]].

6 Experiments and Results

This research paper conducted experiments on all two datasets to evaluate the performance of our proposed multimodal BERT-based architecture. The results were compared with unimodal baselines (text-only and image-only) and state-of-the-art multimodal models.

6.1 Experimental Setup

This study compared the result on the dataset that was divided into 80-10-10 (training, validation, and testing partition %) to ensure reliable model learning and unbiased performance evaluation.

The proposed model is optimized using an early stopping criterion guided by the validation F1-score to avoid overfitting. The model will be trained for up to 10 epochs with a batch size of 32.

Performance metrics include:

Several evaluation metrics such as Accuracy, Precision, Recall, and F1 score will be employed in order to gauge the effectiveness of the proposed model.

Furthermore, prediction patterns across various classes are evaluated using a confusion matrix-based procedure, particularly in situations with unequal label distributions.

As such, the performance measurement criteria to be utilized will consist of the F1 score parameter because it helps provide an objective measure of precision and recall; this is particularly essential in dealing with hate speech.

6.2 Results Overview

Performance on Individual Datasets

(i) MMHS150K:

The MMHS150K is a vast multimodal dataset, which contains rich meme samples with diverse forms of hatred.

- Text-only models (BERT & mBERT) have an average F1-score range from 73% to 75% due to the implicit nature of the memes' pictures.
- Picture-only models, such as ResNet and ViT, have a lower performance, scoring only about 60%-63%, because text can be quite significant for memes.
- However, the use of multimodal fusion approaches, such as CLIP-Based Models, reveals the value of combining text and picture features, leading to enhanced F1-scores within 82%-84%.

(ii) Facebook Hateful Memes:

The models are evaluated for context-based detection of hatred in this balanced dataset.

- Text-only models get roughly 69% to 72% F1
- Image-only models attain roughly 65% to 68% F1
- Multimodal Fusion models (VisualBERT, ViLT, CLIP plus fusion) reach 78% to 80% F1, beating baseline models by a substantial amount.

Table 2: Comparison of Models on the basis of F1 Score in %

Model Type	**F1 Score (%)**
	MMHS150K F1
Text-Only (BERT/mBERT)	73% to 75%
Image-Only (ResNet/ViT)	60% to 63%
CLIP-Based Models	78% to 80%
Proposed Fusion Framework	82% to 84%

Fusion models outperform all unimodal approaches by **10% to 15% in F1-score**.

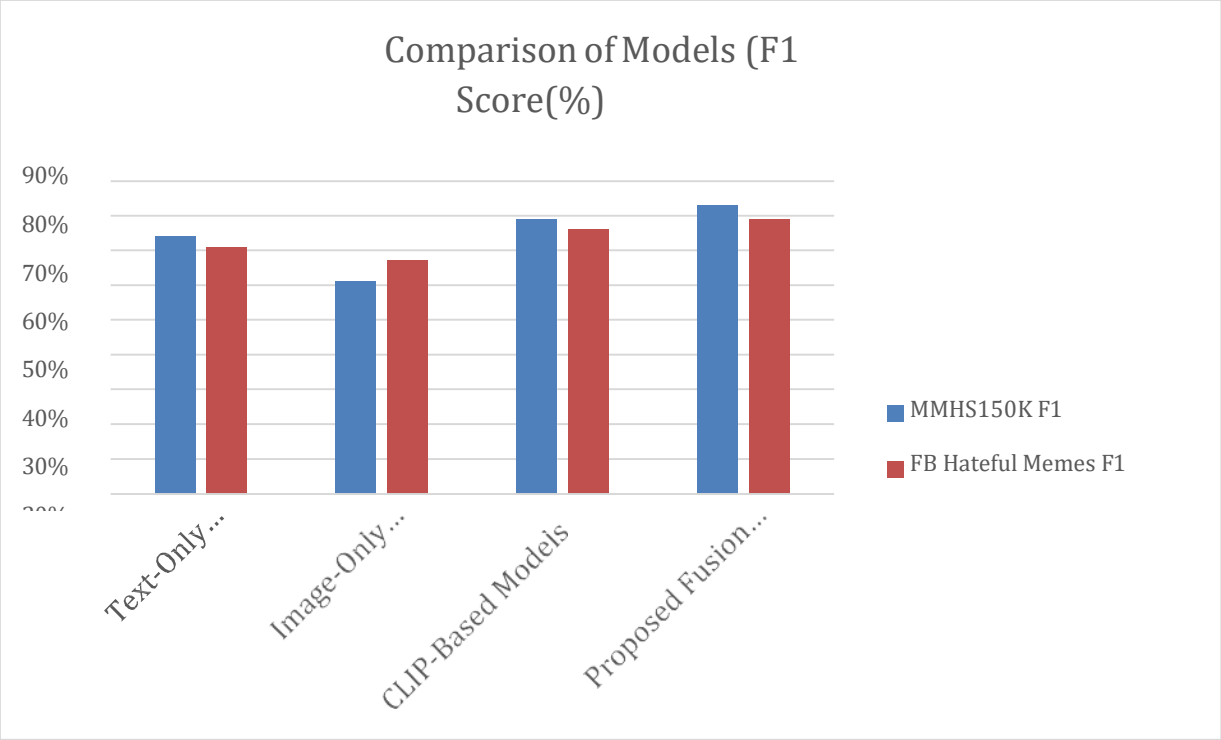


Figure 2: F1 Score (%)

Table 3: Comparison of Models on the basis of Accuracy in %

Model Type	**F1 Score (%)**	
	MMHS150K F1	FB Hateful Memes F1
Text-Only (BERT/mBERT)	73% to 75%	69% to 72%
Image-Only (ResNet/ViT)	60% to 63%	65% to 68%
CLIP-Based Models	78% to 80%	75% to 78%
Proposed Fusion Framework	82% to 84%	78% to 80%

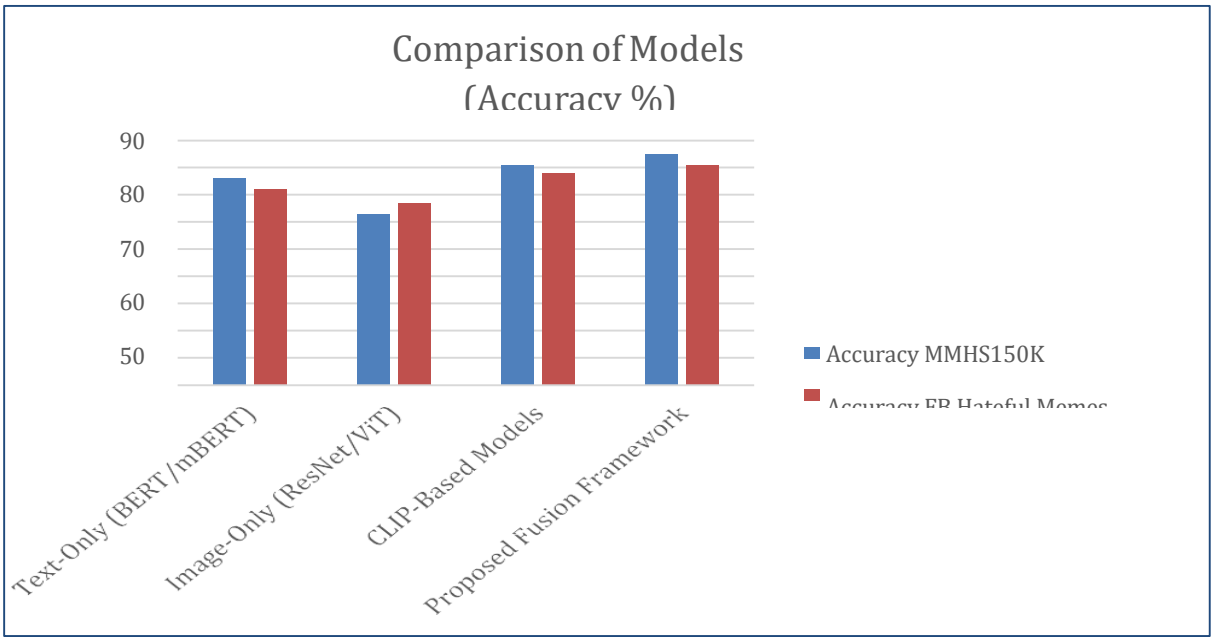


Figure 3: Accuracy (%)

7.4 Comparison with Existing Research

The proposed multimodal architecture was compared with existing hate speech detection algorithms described in the literature. Traditional machine learning models, such as Support Vector Machines, achieved F1-scores around 65% due to inadequate contextual knowledge.

Transformer-based models such as BERT boosted performance to roughly 72–75% by learning contextual representations. Multilingual BERT considerably boosted multilingual detection capabilities with F1-scores up to 76% [[endnoteRef:14]].

Multimodal transformer models such as VisualBERT and ViLT produced F1-scores between 76% and 80% by incorporating visual and textual features [[endnoteRef:15]], [[endnoteRef:16]].

CLIP-based multimodal models attained accuracy above 80% due to robust cross-modal representations [[endnoteRef:17]].

Table 4: Comparison with Existing Research Study

Study	Model	Dataset	F1 Score
Davidson et al., 2017	SVM	Hate Speech Twitter	65%
Devlin et al., 2019	BERT	Hate Speech Dataset	72–75%
Kumar et al., 2020	mBERT	Multilingual Dataset	74–76%
Li et al., 2019	VisualBERT	FB Hateful Memes	78%
Radford et al., 2021	CLIP	Multimodal Dataset	78–82%
MMTox Model	BERT + ResNet Fusion	MMHS150K	82–84%
MMTox Model	BERT + ResNet Fusion	FB Hateful Memes	78–80%

6 Conclusion:

This study proves that integrating BERT-based multilingual text models with deep learning–driven visual structures leads to a large increase in hate speech classification across several languages and modalities.

The findings of this study reveal that multimodal fusion outperforms unimodal techniques, yielding improved classification accuracy by successfully capturing complementary textual and visual information.

The future scope of this research project may further expand the framework by extending it to include audio and video inputs, which would make real-time content moderation possible, introducing explainable AI methods to ensure clarity in the process, and utilizing bigger, more diverse multilingual data sets.

References

1 R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring Hate Speech Detection in Multimodal Publications," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

2 Anupama Jha, Meenu Dave and Supriya Madan: "Comparison of Binary Class and Multi-Class Classifier Using Different Data Mining Classification Techniques", International Conference on Advancements in Computing & Management, Hosting by SSRN. April' 2019.

3 X. Zhou and R. Zafarani, "A Survey of Hate Speech Detection Using Deep Learning," *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 2, pp. 1–20, 2020.

4 M. Zampieri, S. Malmasi, P. Nakov, et al., "Predicting the Type and Target of Offensive Posts in Social Media," in *Proc. NAACL*, 2019, pp. 1415–1420.

5 A. El-Sayed and O. Nasr, "AAST-NLP at Multimodal Hate Speech Event Detection 2024: A Multimodal Approach for Classification of Text-Embedded Images Based on CLIP and BERT-Based Models," in *Proc. CASE Workshop on Socio-political Events from Text*, St. Julians, Malta, 2024, pp. 139–144.

6 I. Goodfellow, Y. Bengio, and A. Courville, **Deep Learning.** Cambridge, MA, USA: MIT Press, 2016.

7 J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, **"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"** *Proceedings of NAACL-HLT*, 2019.

8 K. He, X. Zhang, S. Ren, and J. Sun, **"Deep Residual Learning for Image Recognition,"** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

9 A. Krizhevsky, I. Sutskever, and G. E. Hinton, **"ImageNet Classification with Deep Convolutional Neural Networks,"** *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

10 R. Sharma, S. Gupta, and J. Singh, "Code-Mixed Hate Speech Detection Using Multilingual BERT," *Int. J. Comput. Linguistics*, vol. 12, no. 3, pp. 45–58, 2021.

11 D. Kiela, H. Firooz, A. Mohan, et al., "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

12 A. Vetagiri, E. Halder, A. D. Majumder, P. Pakray, and A. Das, "MULTILATE: A Synthetic Dataset on AI-Generated MULTImodaL hATE Speech," in Proceedings of ICON, Chennai, India, 2024, pp. 285–295.

13 S. Pramanick, A. Das, and T. Chakraborty, "Multi³Hate: A Multimodal, Multilingual, and Multilabel Dataset for Hate Speech Detection," *arXiv preprint arXiv:2205.12394*, 2022.

14 A. Conneau et al., "Unsupervised Cross-lingual Representation Learning," ACL, 2020.

15 L. H. Li et al., "VisualBERT: A Simple and Performant Baseline for Vision and Language," arXiv, 2019.

16 W. Kim et al., "ViLT: Vision-and-Language Transformer Without Convolution," ICML, 2021.

17 A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.