MIMIC: Misogyny Identification in Multimodal Internet Content in Hindi-English Code-Mixed Language

AAKASH SINGH

Department of Computer Science, Banaras Hindu University, Varanasi-221005 (India)

DEEPAWALI SHARMA

Department of Computer Science, Banaras Hindu University, Varanasi-221005 (India)

VIVEK KUMAR SINGH

Department of Computer Science, University of Delhi, Delhi-110007 (India)

Over the years, social media has emerged as one of the most popular platforms where people express their views and share thoughts about various aspects. The social media content now includes a variety of components such as text, images, videos etc. One type of interest is memes, which often combine text and images. It is relevant to mention here that, social media being an unregulated platform, sometimes also has instances of discriminatory, offensive and hateful content being posted. Such content adversely affects the online well-being of the users. Therefore, it is very important to develop computational models to automatically detect such content so that appropriate corrective action can be taken. Accordingly, there have been research efforts on automatic detection of such content focused mainly on the texts. However, the fusion of multimodal data (as in memes) creates various challenges in developing computational models that can handle such data, more so in the case of low-resource languages. Among such challenges, the lack of suitable datasets for developing computational models for handling memes in low-resource languages is a major problem. This work attempts to bridge the research gap by providing a large-sized curated dataset comprising 5,054 memes in Hindi-English code-mixed language, which are manually annotated by three independent annotators. It comprises two subtasks: (i) Subtask-1 (Binary classification involving tagging a meme as misogynous or non-misogynous), and (ii) Subtask-2 (multi-label classification of memes into different categories). The data quality is evaluated by computing Krippendorff's alpha. Different computational models are then applied on the data in three settings: text-only, image-only, and multimodal models using fusion techniques. The results show that the proposed multimodal method using the fusion technique may be the preferred choice for the identification of misogyny in multimodal Internet content and that the dataset is suitable for advancing research and development in the area.

Keywords: Hindi-English code-mixed data, Low resource language, Misogynistic attitude, Misogyny detection, Multimodal content.

1 INTRODUCTION

The different kinds of social media platforms allow users to share a wide variety of content that may include text, images, videos, and more. Online communication now extends beyond the mere use of text, with images and videos being used quite often. Nowadays, memes have gained a lot of popularity on social media platforms. A meme typically consists of an image conveying pictorial content, with overlaid text added by the meme's creator. The memes are often targeted to be humorous or ironic [1]. However, sometimes they are also used by certain individuals to troll, engage in mockery, spread hate, incite violence, and target vulnerable communities [2,3]. Over the last few decades, numerous studies [4-7] have attempted to address inappropriate and harmful text on social media. There have been research efforts on automatic detection of such content focused mainly on the texts. However, the fusion of multimodal data (as in the case of memes) creates various challenges in developing computational models that can handle such data, more so in case of low-resource languages. As a result, the area of multimodal Internet content is relatively less explored.

It is evident that images hold a more profound impact than text, highlighting the need to address multimodal data effectively [8]. A concerning trend is the frequent sharing of memes that make fun of, objectify, and humiliate women. Expressions within these memes that display objectification, prejudice, humiliation, violence, or hatred towards women fall under the category of misogyny [9]. Women are active participants on social media, and the propagation of misogynistic multimodal content, particularly as memes can severely affect the mental health of the users [10]. Addressing the issue represents a crucial step in moving towards a society characterized by both peace and inclusivity, aligning with the United Nations Sustainable Development Goals (#5 and #11). Some studies [11] [15] highlight the necessity for the research community to approach this issue with

ACM 2375-4699/2024/04-ART
http://dx.doi.org/10.1145/3656169

innovation. Consequently, it becomes important to address and flag such misogynistic multimodal content to ensure a safer online environment.

Several previous studies [12-14] have focused on misogyny detection within textual data, primarily within the English language on various online social media platforms, with some recent efforts in handling the data from non-English languages as well [16-18]. The majority of the studies have however utilized only the textual data, resulting in lower accuracies of the models. There have been some studies [19-22] that focused on multimodal data to detect misogyny, but mainly for English language content. To the best of our knowledge, there are no previous studies on detecting misogyny in multimodal data involving Hindi-English code-mixed language. The absence of a suitable dataset may be one major reason for this. Therefore, there is a need to create a suitable dataset that can be used for research on misogyny detection in multimodal content that includes texts in low-resource languages.

This work attempts to bridge the research gap by providing a large-sized curated dataset of memes in Hindi-English code-mixed language. The dataset is suitably annotated by independent annotators and the quality of annotations is evaluated by computing standard measures of inter annotator agreement. Two subtasks are then proposed on the data, first involves a binary classification of identifying whether a meme is misogynous or not, and the second subtask involves categorizing the memes into various categories of misogyny. Different computational models are then applied on the data in three settings. First, only textual data is used to develop models for classification. Thereafter, the image data is used for developing image classification models. Finally, both textual and image data are used together to develop fused models for misogyny detection. To the best of our knowledge, this dataset is first such multimodal dataset in Hindi-English code-mixed language for misogyny detection. Further, the computational models developed provide a suitable method for detection and categorization of misogynous memes. The study is thus first such experimental work on misogyny detection in multimodal data (memes) involving Hindi-English code-mixed language.

The major contributions of this paper are as follows:

1. It presents a large-sized curated and annotated dataset of 5,054 memes in Hindi-English code-mixed language for misogyny detection. Two subtasks are provided in the dataset.
2. The quality of the dataset is evaluated by measuring the inter-annotator agreement, which provides confidence in the quality of the dataset developed.
3. Different text-only and image-only transformer-based models are implemented and explored on the developed dataset for the two subtasks.
4. A suitable model for misogyny detection in the code-mixed data is developed by combining both, the text and image models, with the help of late fusion technique.

The rest of the paper is organized as follows: Section 2 surveys some previous related work on misogyny detection, including the existing datasets developed for the purpose. Section 3 describes in detail the dataset construction, annotation procedure, and annotation quality. Section 4 presents the details of the computational models implemented. The experimental setup and the evaluation metric are described in Section 5, followed by the description of the results in Section 6. The results and findings are discussed in Section 7, which also includes a discussion of limitations and some open challenges of research in the area. The paper concludes in Section 8 with a summary of the work done and its importance.

2 RELATED WORK

Social media platforms have vast amounts of data in various forms, including text, images, videos, and their combinations. Unfortunately, a significant volume of undesirable content spreads on these platforms, and at various instances such content is directed at women. The term used to describe such hatred, prejudice, or stereotyping against women is referred to as "misogyny" [9,10]. Detecting and flagging such content on social media is important. Most of the research has been conducted in this direction on textual data, primarily in the English language. Numerous studies [11-14] have sought to identify misogyny in English-language social media texts. In addition to English, some research has explored the detection of misogynistic textual content in other languages [15-18]. Nevertheless, the research community has predominantly concentrated on textual data for misogyny detection. The memes have recently gained more popularity for communication and expressing opinions on social

media; however, they are not well-explored yet. Memes that have textual content in low-resource languages need particular attention.

There are a limited number of datasets available for automatic detection of multimodal misogynistic content. One such dataset [23] comprises 800 memes in the English language collected from popular social media platforms like Twitter, Facebook, Instagram, Reddit, and consulting websites. These memes are categorized into misogynistic and non-misogynistic, and their aggressiveness and irony levels are also determined. One of the major and often used datasets for detecting misogyny in multimodal English content is "Multimedia Automatic Misogyny Identification (MAMI)," which was featured as part of the SemEval-2022 Task [24]. This dataset includes 10,000 memes sourced from popular social media platforms such as Twitter and Reddit, as well as websites dedicated to meme creation and sharing. It involves two subtasks: (i) Subtask A, where memes are categorized as misogynous or non-misogynous, and (ii) Subtask B, which identifies the specific type of misogyny within misogynous memes, such as shaming, objectification, stereotypes, or violence. In addition to this, there is one more dataset released for LT-EDI@ACL-2024 on detecting misogyny in memes in Tamil and Malayalam[1], which contains approximately 2,000 memes with text in Tamil and 1,000 memes with text in Malayalam collected from various social media platforms. These memes are classified as either misogynistic or non-misogynistic. A summary of the major datasets on multimodal misogyny detection is provided in **Table 1**.

Table 1: Major Datasets on Multimodal Misogyny Detection

Name of the Dataset	**Description**				**Link**
	Source	**Size**	**Class/Labels**	**Language**	
Memes with text transcriptions for automatic detection of misogynistic content [23]	Facebook, Twitter, Instagram and Reddit	800 memes	Misogyny, Aggressiveness and irony	English	https://github.com/MIND-Lab/MEME
Multimodal Automatic Misogyny Identification (MAMI) [24]	Twitter, Reddit, 9GaG, Imgur	10,000 memes	**Subtask A:** Misogynous, Not Misogynous **Subtask B:** Shaming, Stereotype, Objectification, Violence	English	https://github.com/TIBHannover/multimodal-misogyny-detection-mami-2022
Identification of Misogynistic memes	Facebook, Instagram and other social media platforms	**Tamil:** about 2000 memes **Malayalam:** about 1000 memes	Misogyny, Non-Misogyny	Tamil, Malayalam	https://codalab.lisn.upsaclay.fr/competitions/16097

Many studies have been conducted on the MAMI dataset [24] for the detection of misogyny in multimodal English memes. Different methodologies have been employed to identify misogyny in multimodal data. **Table 2** displays a tabular representation of some of the major studies conducted on the MAMI dataset.

In one study [20], the Perceiver IO model was utilized as a late fusion approach for both subtasks, yielding a macro F1-score of 69.9% for Subtask A and a weighted F1-score of 69.3% for Subtask B. Another study [21] applied the Multimodal text-tags model (MTT) to classify whether a meme is misogynous or non-misogynous in Subtask A, achieving an AUC of 0.786. In Subtask A, a study [22] demonstrated that ENIE-Vil-large, coupled with word masking, image captions, and additional data from memes, and outperformed other approaches and achieved an F1-score of 79.3. In Subtask B, oversampling was combined with the ERNIE-Vil-large model, resulting in an F1-score of 72.8. Another study [25] employed the CLIP model with visual and language (VL) features and utilized Logistic Regression (LR) for Subtask A. Similarly, multiple other studies [26 -28] implemented the CLIP model for misogyny detection. A different approach was proposed in study [29], which introduced the Multimodal Multitask Variational AutoEncoder (MMVAE), yielding an F1 score of 0.723 for Subtask A and 0.634 for Subtask B. In another study [30], the LXMERT model with fusion techniques achieved F1-scores of 0.662 and 0.663 for Subtask A and Subtask B, respectively. The authors of study [31] suggested the use of BiLSTM with BERT embedding and incorporated nude detection for the first subtask. A study [32] extracted features using object attributes of fixed box sizes (OSCAR), with OSCARhm_pretrained_ens achieving the highest F1-score of 68.5 for Subtask A, and OSCARens obtaining a higher F1-score of

[1] https://codalab.lisn.upsaclay.fr/competitions/16097#participate

52.6 for Subtask B. For the first subtask, VisualBERT was experimented with R-CNN, resulting in an F1-score of 0.670 [33]. Likewise, in another study [34], the BERT model was employed for misogyny detection.

Table 2: Tabular representation of major studies on the MAMI dataset

Author	Models	Results	
		Subtask A	Subtask B
Sharma, M., Kandasamy, I., & Vasantha, W. B. (2022, July) [19]	**Subtask A:** Voting Ensemble of attention. **Subtask B:** BERT+ViT using attention	F1 score-=0.757	F1 score= 0.690
Attanasio, G., Nozza, D., & Bianchi, F. (2022) [20]	Perceiver IO (FCWA) both subtasks	Macro F1 score=69.9%	Weighted F1 score=69.3%
Rizzi, G., Gasparini, F., Saibene, A., Rosso, P., & Fersini, E. (2023) [21]	**Subtask A:** Multimodal Text-Tags Model (MTT)	AUC- 0.786	-
Zhou, Z., Zhao, H., Dong, J., Ding, N., Liu, X., & Zhang, K. (2022, July) [22]	**Subtask A:** ERNIE-Vil-large+WM+IC+AD+Emsembling+TS+PA **Subtask B:** ERNIE-Vil-large+Oversampling+PT+RC	F1-score=79.3	F1-score =72.8
Chen, L. (2022, July) [25]	**Subtask A:** CLIP VL feature+LR	Macro F1 score =0.778	-
Zhang, J., & Wang, Y. (2022, July).[26]	CLIP img+txt for both subtasks	Macro F1-score=0.834	Macro F1-score=0.731
Muti, A., Korre, K., & Barrón-Cedeño, A. (2022, July) [27]	Multi (BERT+CLIP) for both subtasks	Macro averaged F1 =0.727	Weighted F1-score=0.710
Arango, A., Perez-Martin, J., & Labrada, A. (2022, July). [28]	**Subtask A:** CLIP_sum	F1-score= 71%	-
Gu, Y., Castro, I., & Tyson, G. (2022, July) [29]	MMVAE for both subtasks	F1-score-=0.723	F1-score=0.634
Han, C., Wang, J., & Zhang, X. (2022, July). [30]	LXMERT with fusion	F1-score= 0.662	F1-score= 0.633
Cordon, P., Diaz, P. G., Mata, J., & Pachón, V. (2022, July). [31]	**Subtask A:** BiLSTM with BERT embedding+nude detection	F1-score=0.665	-
Agrawal, S., & Mamidi, R. (2022, July).[32]	**Subtask A:** OSCAR $_{hm\_pretrained\_ens}$ **Subtask B:** OSCAR $_{ens}$	Macro F1-score= 68.5	Macro F1-score= 52.6
Ravagli, J., & Vaiani, L. (2022, July) [33]	**Subtask A:** VisualBERT with R-CNN	F1-score= 0.670	-
Sharma, G., Gitte, G. S., Goyal, S., & Sharma, R. (2022, July)[34]	BERT for both subtasks	F1-score=66.24%	F1-score= 64.76%
Rao, A. R., & Rao, A. (2022, July) [35]	Weighted Average Ensemble (VisualBERT+MMBT) for both subtasks	F1-macro=0.761	F1-macro=0.705
Sivanaiah, R., Angel, S., Rajendram, S. M., & Mirnalinee, T. T. (2022, July).[36]	Ensemble (AlBERT+CNN) for both subtasks	F1-score=0.5223	F1-score=0.4673
Habash, M., Daqour, Y., Abdullah, M., & Al-Ayyoub, M. (2022, July). [37]	**Subtask A:** Ensemble(two MMBT + VisualBERT)	F1-score= 0.722	-
García-Díaz, J., Caparros-Laiz, C., & Valencia-García, R. (2022, July)[38]	Ensemble learning (LF, BF, BI) for both subtasks	Macro F1-score= 0.687	Macro F1-score= 0.663

Several studies [19], [35-38] have employed ensemble models to detect misogyny in memes. In the study [19], an ensemble model combining BERT and Vision Transformer (ViT) was utilized, employing a voting scheme of attention for Subtask A, resulting in an F1-score of 0.757. For Subtask B, the same study utilized the BERT+ViT ensemble with attention to classifying misogynous memes into further sub-categories, achieving an F1-score of 0.690. In another study [35], an ensemble model comprising VisualBERT and MMBT was experimented with, using a weighted average approach for both subtasks. This yielded F1-scores of 0.761 for Subtask A and 0.705 for Subtask B. Similarly, another study [36] implemented an ensemble model combining AlBERT and CNN for both tasks. For Subtask A, experimented with an ensemble incorporating MLP, and for Subtask B, majority voting was employed. In the study [37], an ensemble approach using two Multimodal Bi-Transformers (MMBT) along with VisualBERT was implemented for Subtask A, achieving an F1-score of 0.722. Additionally, a study [38] implemented ensemble learning that combined linguistic features (LF), BEiT (BI), and Bert Embeddings (BF) for the subtasks.

The existing research work predominantly concentrates on analyzing textual data from social media to identify misogyny. Although a limited number of studies have explored misogyny detection in multimodal content, such as memes, there remains a substantial research gap, especially in the case of non-English and code-mixed languages. The previous studies in the field of multimodal misogyny detection have mainly concentrated on the English language, with only a few recent studies extending their scope to languages like Tamil and Malayalam. A noticeable lack of resources exists for non-English languages, especially when considering Hindi and Hindi-English code-mixed multimodal content. The present work attempts to address this research gap by introducing a high-quality curated and annotated dataset of memes extracted from various popular social media platforms in Hindi-English code-mixed languages. The dataset's annotation quality is evaluated using established inter-annotator agreement measures. It is important to note that this dataset is the first such dataset in Hindi-English code-mixed language. Additionally, a methodology for identifying misogyny within multimodal content is proposed. Different transformer-based models for text-only and image-only content are implemented independently. Subsequently, a late fusion technique is employed to combine the text and image models, proposing a state-of-the-art multimodal approach for identifying misogyny in memes involving Hindi-English code-mixed content.

3 DATASET PREPARATION

3.1 Data Collection

The memes for the dataset are collected from the popular social media platforms such as Facebook, Instagram, Reddit and some other image-sharing platforms like Pinterest. The memes are collected through a mix of automated scraping and manual downloading from these platforms. **Figure 1** illustrates the various steps in the data collection process. To ensure a diverse range of misogynistic memes, several steps have been taken, namely: (i) searching for Facebook, and Instagram pages or groups specifically focused on memes featuring women; (ii) exploration of recent discussions related to well-known women in various fields, including actors, models, and artists; (iii) using targeted keywords and hashtags such as #girlfriend, #papa ki pari, #women, #feminist, #moti, #stree to search relevant meme content; (iv) searching of Reddit threads dedicated to or written by individuals who identify as anti-women or antifeminist for additional meme resources; and (v) locate Pinterest page IDs dedicated to sharing memes about women.

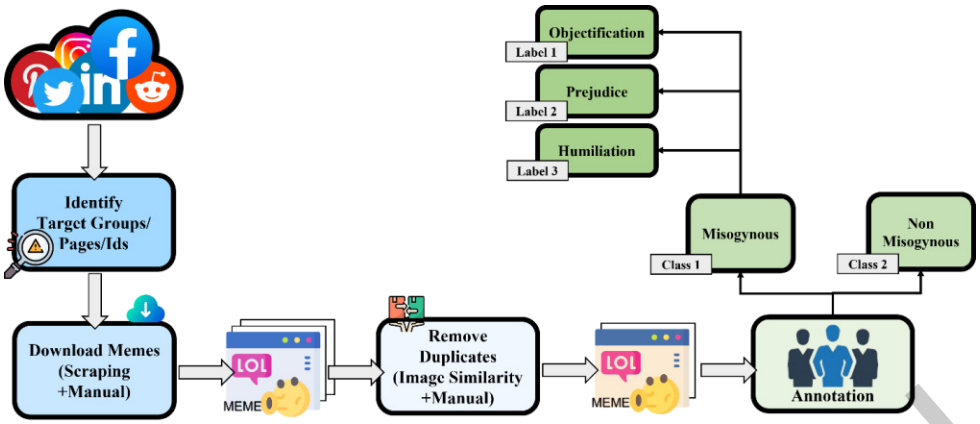

Figure 1. Flow-chart of data collection process

A total of 6,761 memes were initially downloaded. Following this, some duplicate versions and unclear (low-resolution) memes were removed, resulting in a final set of 5,054 unique memes. The textual content of these memes was transcribed using the EasyOCR library of Python, which is a library trained to support the extraction of more than 80 popular writing scripts including English (Latin) & Hindi (Devanagari). The dataset thus comprises of images and the associated Hindi-English code-mixed language text.

3.2 Data Annotation

The data collected was then annotated with the help of independent annotators. Two subtasks were conceptualized on the dataset. One is to identify whether a meme is misogynous or not, and the second is to identify the specific category (objectification, prejudice, humiliation) of misogynous memes. The annotation involved three qualified annotators proficient in both English and Hindi, with a thorough understanding of the annotation scheme confirmed before proceeding. Annotators were appropriately briefed about the annotation process and were shown sufficient number of examples for different categories. Sufficient time for annotation was provided to ensure effective and accurate annotations. Two key questions were designed for dataset annotation: (1) Is this meme misogynous or not? (2) If the meme is misogynous, what are the primary categories to which it should belong to (objectification, prejudice, and humiliation)? Additionally, annotators were instructed that for memes related to the misogyny category, multiple overlapping labels may be possible. **Figure 2** shows some example of misogynous and non-misogynous memes.
The following annotation schema was given for both subtasks to the annotators:

Subtask 1 (Binary Classification): In the first subtask the memes are classified into two categories: (i) misogynous or (ii) non-misogynous.

Misogynous: Memes are labeled as misogynous if any of the following conditions are met:

(i) Meme that conveys or reinforces anti-women attitudes, beliefs, or stereotypes about women.
(ii) Meme that promotes negative and prejudiced perceptions of women, often to undermine their worth, abilities, or dignity.
(iii) Meme uses humour, imagery or language to demean, objectify or marginalize women.

Non-Misogynous: Memes that do not exhibit anti-women sentiments are categorized as non-misogynistic

Subtask 2 (Multi-label classification): The memes are categorized into three categories: (i) Objectification, (ii) Prejudice, and (iii) Humiliation

Objectification: Objectification refers to the reduction of women to just objects, focusing on their physical appearance rather than recognizing their humanity [39]**.** Memes are categorized as objectification if any of the following conditions are met:

(i) Memes that reduce women to their physical attributes, such as body parts or appearance, without acknowledging their intellectual, emotional, or individual qualities.

(ii) Memes that sexualize women without any relevant context or purpose, present them solely as sexual objects.

(iii) Memes that dehumanize women by treating them as commodities for visual pleasure or entertainment, disregarding their autonomy and dignity.

Prejudice: Prejudice refers to preconceived opinions, attitudes, or judgments that are not based on reason or experience but are instead rooted in biased beliefs or stereotypes about women [40]. Memes fall into the category of prejudice if any of the following conditions are satisfied:

(i) Memes that propagate stereotypes by making broad and unfair generalizations about women, portraying them in limited and biased roles based on traditional or discriminatory views.

(ii) Memes that use offensive language or content that mocks women based on their gender reinforce negative biases and diminish the importance of women's contributions, aspirations, and reinforce discriminatory attitudes.

Humiliation: Humiliation refers to mocking women based on their behaviour, appearance, or choices [41]. Memes are classified as humiliation when any of the specified conditions are met:

(i) Memes that depict slut-shaming, motherhood shaming, and women being unfairly treated.

(ii) Memes that ridicule women based on their body weight, body shape, or perceived lack of conformity to societal beauty standards.

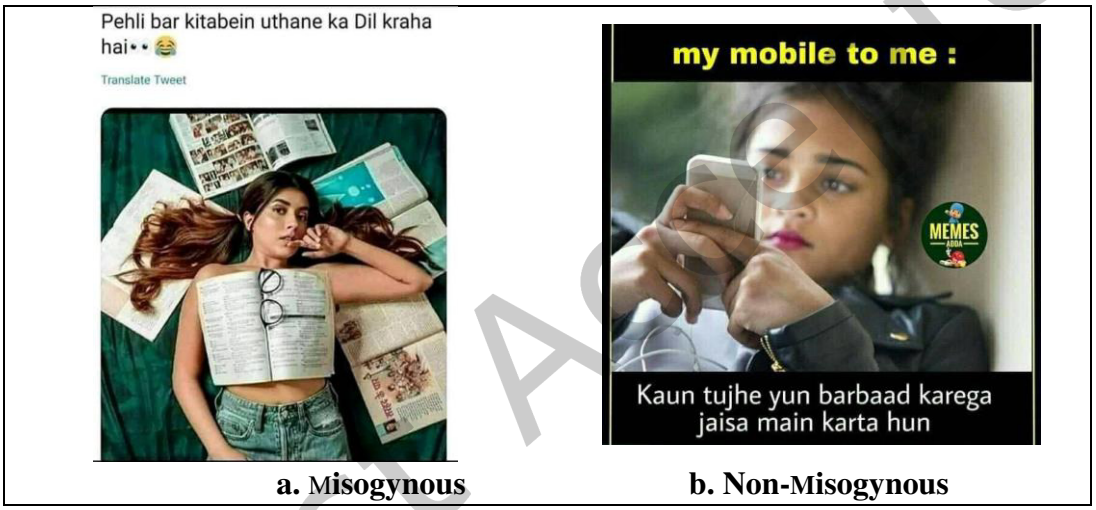

Figure 2. Examples from Misogynous and Non-Misogynous classes from the dataset

Figure 3 shows illustrative examples of the aforementioned categories (labels) of misogynous memes.

For evaluating the quality of annotations, the inter-annotator agreement was computed by using the standard measure of Krippendorff's alpha [42]. For subtask 1, the coefficient value was 0.83, while for subtask 2, the coefficient value was 0.67. These values of Krippendorff's alpha suggest that the agreement for subtask 1 is strong, while the agreement for subtask 2 is moderate. The values suggest that the annotation quality is acceptable for use of the dataset for experimental work.

3.3 Dataset Statistics

This section summarizes the key statistics of the dataset created. In sub-task 1, memes are classified into two disjoint sets: Misogynistic and Non Misogynistic. The data distribution here is quite balanced, with 2515 memes being in the first category and 2539 memes in the second category. In the sub-task 2, there are three classes of memes. The type with the largest count is observed to be the first label "Objectification" (n=1293), followed by the second label "Prejudice" (n=919), and the third "Humiliation" (n=393). **Figure 4** visually presents a summary of the distribution of the dataset. Upon closer examination of Figure 4, it becomes apparent that there are several overlaps among the existing labels. The largest one is between the label pairs (Objectification and prejudice) followed by (Prejudice and Humiliation) and (Objectification and Humiliation). There also exist memes (03 in count) that are labeled in all three categories.

a. Objectification **b. Prejudice** **c. Humiliation**

Figure 3. Examples of Objectification, Prejudice, and Humiliation labels from the dataset.

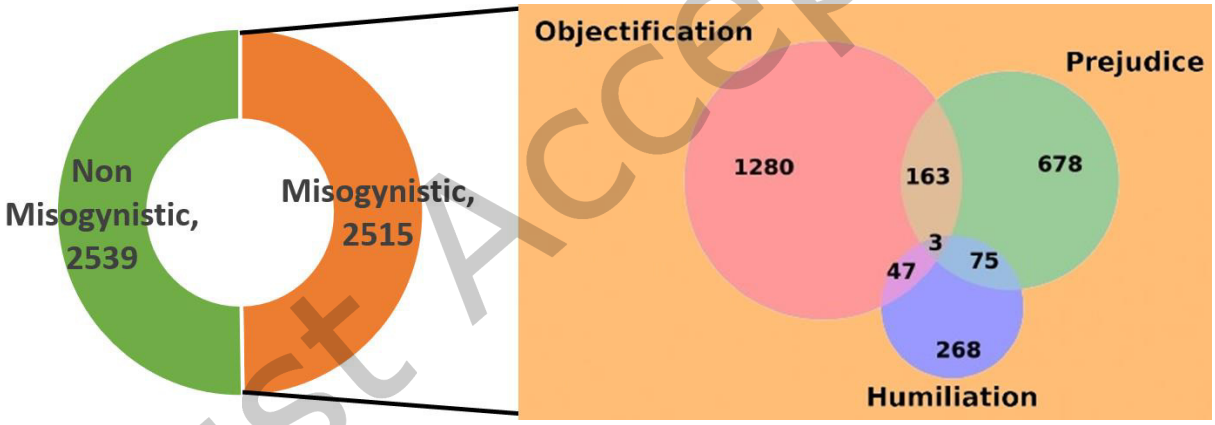

Figure 4. Dataset distribution summary

4 ALGORITHMIC MODELS

As the memes collected comprised text and image content, computational models can be designed to use them independently or together. For exploring the different possibilities of model design, first, the text and image content are handled separately as independent models, and thereafter both text and image content are used in multimodal model. The different computational models explored and implemented are described below.

4.1 Text models

This section outlines the text models implemented for both subtasks. A total of four state-of-the-art models employed in the study are detailed below.

I. **BERT:** BERT, which stands for Bidirectional Encoder Representations from Transformers, was developed by Google and undergoes pre-training using Wikipedia and the Toronto Book corpus [43]. The BERT architecture consists of multiple Encoder layers, where each encoder is composed of two sub-layers: a feed-forward layer and a self-attention layer. Tokenization is performed using the BERT tokenizer, introducing two special tokens, [CLS] and [SEP], in each token

sequence. The [CLS] token is positioned at the sequence's beginning, while [SEP] is utilized to segment different sections. The tokenized data is then input into the BERT model for subsequent classification tasks. BERT's bidirectional approach allows it to capture contextual information from both preceding and succeeding words, enhancing its understanding of language variations and improving performance across various natural language processing tasks.

II. **ALBERT:** ALBERT stands for A Lite BERT, and it is a variant of the BERT model. ALBERT, introduced as a Lite BERT, focuses on improving the efficiency and scalability of the BERT model [44]. One of the main goals of ALBERT is to reduce the number of parameters in the model while maintaining or improving performance. This is achieved through various parameter-sharing techniques. ALBERT incorporates two types of parameter-sharing techniques. First, it shares parameters across layers, allowing for more efficient use of parameters throughout the model. Second, it shares parameters across attention heads, further reducing redundancy. It uses factorized embedding parameterization, separating the size of the hidden layer from the size of the embedding layer. Implementing ALBERT involves downloading pre-trained models, fine-tuning them on task-specific data, and using the trained model for inference.

III. **RoBERTa:** RoBERTa stands for Robustly Optimized BERT approach. It is based on Google's BERT model released in 2018 [45]. Like BERT, RoBERTa is pre-trained on large amounts of unlabeled text data. However, RoBERTa modifies the pre-training objectives to remove the next sentence prediction task and instead focuses on masked language modeling. It employs dynamic masking. This means that different training epochs use different masked words, which helps the model generalize better. RoBERTa uses larger batch sizes during pre-training, enabling more efficient use of parallel processing capabilities and contributing to improved training efficiency. It does not perform sentence pair classification during pre-training, and it treats sentences as individual sequences.

IV. **MuRIL:** MuRIL is a language model based on BERT, having undergone pre-training on a diverse array of languages. It is pre-trained from scratch using Wikipedia, Common Crawl, PMINDIA, and Dakshina corpora for 17 Indian languages: Assamese, Gujarati, Kashmiri, Bengali, Hindi, Kannada, Malayalam, Nepali, Marathi, Punjabi, Oriya, Sanskrit, Tamil, Sindhi, Telugu, Urdu, and English [46]. Throughout the training process, both translation and transliteration segments were integrated. The training data for the model includes monolingual segments and parallel segments, encompassing two types of parallel data: translated and transliterated content. The model was trained using a self-supervised masked language modeling task**.**

4.2 Image models

This section describes the image models used for the classification task. In total, four State-of-the-art (SOTA) models have been used in the study. Three of them belong to Convolutional Neural Network (CNN) based architecture and one has a transformer-based approach to handle classification. The choice of these models and their variations aligns with the study [47], in which the authors conducted a performance comparison of these contemporary classification models. These pre-trained models are accessible via TensorFlow Hub.

I. **Vision Transformer (ViT):** ViT works by leveraging the power of transformer architectures [48], initially designed for sequential data like text (Natural Language Processing) [49]. The image here is treated as a sequence of fixed-size non-overlapping patches, which are linearly embedded into high-dimensional vectors. These patch embeddings, along with positional encodings, are fed into a transformer encoder. The transformer's self-attention mechanism enables capturing global contextual information. This facilitates learning long-range dependencies crucial for understanding complex visual patterns. The positional encoding helps in maintaining spatial information of patches. Multiple transformer blocks process the sequence hierarchically, extracting increasingly abstract features. ViT provides an edge in terms of scalability, adaptability, and parallel processing.

II. **InceptionV3:** Inception as a model was developed by Google [50] and has performed exceptionally well in the ImageNet Visual Recognition Challenge (2014). The primary objective of the design was to go for a deeper convolutional network without getting into the traps of overfitting and exploding computational resources. The remedy proposed was a transition from fully connected network architectures to sparsely connected ones. InceptionV3 (2016) [51] improves upon the original model by introducing factorized convolutions, using 1xN and Nx1 convolutions to reduce complexity while

maintaining representation. It includes batch normalization, factorized 7x7 convolutions, an auxiliary classifier to tackle vanishing gradients, and global average pooling.

III. **EfficientNetV2:** EfficientNets are another approach by Google [52] to address the scaling problem of the convolutional neural network in a novel way. It introduces a compound scaling method that uniformly scales network dimensions, such as depth, width, and resolution, resulting in a balance between performance and computational cost. EfficentNetV2 [53] is a further improvement over its predecessor in terms of parameter efficiency and training speed. The developers employed a blend of Neural Architectural Search (NAS) and scaling to achieve the intended purpose. The MBConv blocks were replaced by combinations of MBConv and Fused-MBConv to improve the utilization of mobile /server accelerators.

IV. **Big Transfer (BiT):** It is yet another CNN-based architecture researched by the team Google Brains [54]. The development of this technique aimed to address the challenge of a limited supply of labeled data. The transfer learning-based conventional approaches have faced difficulties in applying knowledge learned from extensive datasets to smaller datasets. The BiT model architecture utilizes a conventional ResNet with augmented depth and width. The research explores two essential elements vital for transfer learning: Upstream pre-training and Downstream fine-tuning. The first concept emphasizes the correlation between dataset size and architecture scale, swapping ResNet's batch normalization with GroupNorm and Weight Standardization (GNWS). In the second approach, they introduced a cost-effective fine-tuning protocol named "BiT-HyperRule.

4.3 Proposed multimodal weighted late fusion method

Multimodal models are the models that integrate diverse types of data for comprehensive analysis and understanding. The study has considered memes which tend to have two types of data embedded, image and text. There exist various approaches to fuse learning from different modalities. The late fusion approach is the one that has been considered in the analysis. It was contemplated with the intention of harnessing and amplifying the capabilities of state-of-the-art models existing independently in both the image and text domains. **Figure 5** illustrates the foundational structure of the proposed method with a block diagram.

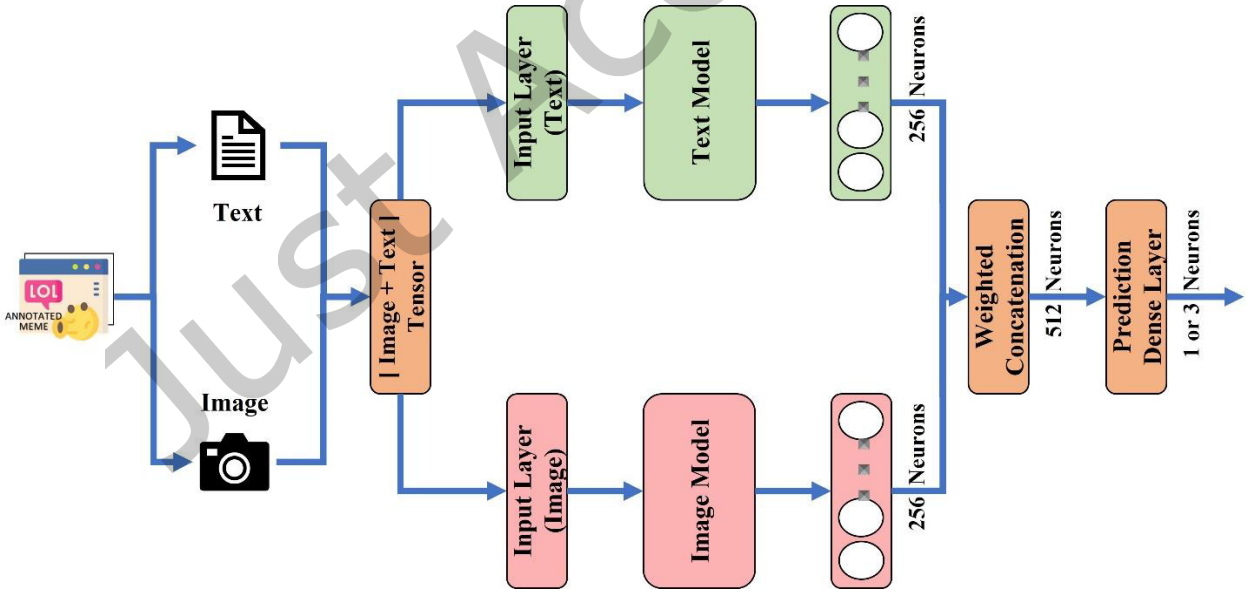

Figure 5: The framework of the multimodal approach used in the study (Block diagram).

The framework intakes batches of memes (images and their respective text) as input tensors. This tensor is in the form of a zipped list, where the first element represents image data and the second text. Based on their positional arguments, they are in turn fed to the respective input layer. The input layer is also responsible for some parts of the preprocessing of the data including setting shape and datatype of the tensor. Then comes the phase where the data tensor has to be passed through a pre-trained model of the respective modality. To do this, the selection of two models (one for image and one for text) is necessary. The study

examines all possible combinations of the top-performing image and text models mentioned earlier. The output of these two selected models is then passed to a dense layer of an equal number of neurons (256). The intention behind this was to ensure that the learnings of both models get equal weight (initially), as different models have different output parameters. The study proposes utilizing a weighted concatenation technique to integrate the knowledge from both modalities. It empowers the model with the ability to determine the priority of each modality in various scenarios. In the end, a final prediction dense layer was placed to decide the class/labels based on all prior learnings.

5 EXPERIMENTAL SETUP

The description of the experimental setup and configurations employed for training the models discussed in Section 4 is provided here. The study adhered to the training configurations suggested by [55]. To ensure consistency and to facilitate an unbiased comparison, the same configurations were applied across all models (Uni+Multimodal). Rigorous 5-fold cross-validation techniques were employed for model evaluation. This ensured that the models were tested on every available data point, thereby eliminating any possibilities of data dependency or overfitting. A gist of the training and evaluation process is provided in **Figure 6.** The experiment was performed on the Nvidia A5000 RTX graphics unit with 24GB of memory as the hardware. The primary Python frameworks (Software) employed were Scikit-learn and TensorFlow 2.0.

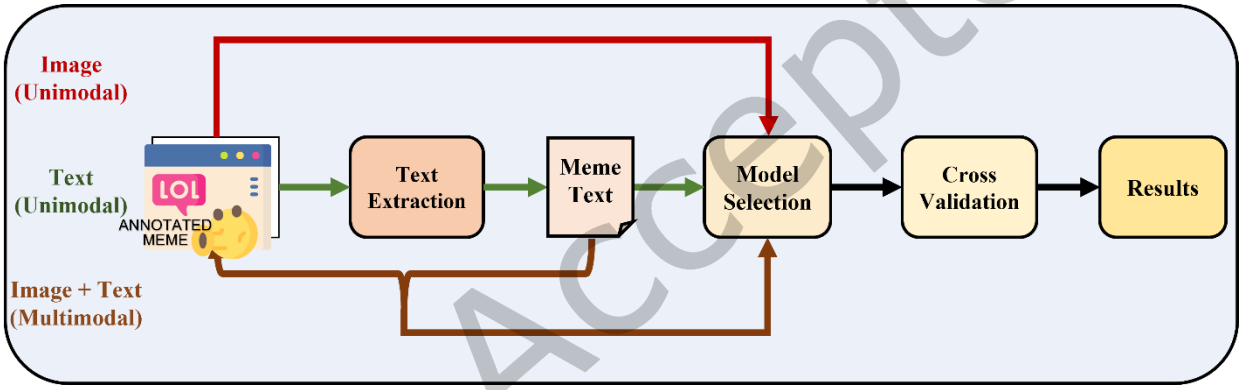

Figure 6. Block diagram of experimental setup and evaluation.

5.1 Model Compilation

The models used in the study were supplied with input in batches of size 32 during the training and testing phase. The input to image models consists of images with color three channels (RGB) that have been resized to 224x224. This measure was taken because all the image models used were primarily pre-trained on images of comparable size. Similarly, text underwent preprocessing before being provided to text models. All text models employed in the study had their preprocessing layer that received uncased meme text. A train-validation split of 0.2 was used in the training phase. To calculate the loss incurred at each epoch, the Binary cross-entropy function was utilized. It measures the difference between two probability distributions for a given random variable or set of values [56]. This loss function has served in both tasks (Binary and Multi-label classification). To minimize the loss function, the study used Rectified-Adam as an optimizer. It effectively addresses the initial training stage's significant variance issue observed in Adam, especially when working with a constrained dataset [57]. Callbacks were used to ensure that the models reached their optimum training saturation. The study employs two Keras pre-defined callbacks. ReduceLROnPlateau callback to control the learning rate dynamically and EarlyStopping callback to prevent the training process from overfitting and to stop at the best epoch.

5.2 Evaluation Metric

To quantify the model's learning performance and predictive power various evaluation metrics are used. In this study, we have used two popular metrics, Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and F1- score. The AUC-ROC is a crucial metric for evaluating the performance of a deep learning model, especially in binary classification tasks. The ROC

curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) at various threshold values. The AUC-ROC represents the area under this curve, indicating the model's ability to discriminate between positive and negative instances.

$$True\ Positive\ Rate\ (TPR) = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive}{False\ Positive + True\ Negative}$$

The ROC curve is generated by calculating TPR and FPR at different classification thresholds. AUC-ROC, ranging from 0 to 1, quantifies the area under this curve. F1-Score is another popular metric in this series that combines precision and recall into a single value.

$$Precision\ (P) = \frac{True\ Positive}{True\ Positive + False\ Positives}$$

$$Recall\ (R) = \frac{Ture\ Positive}{True\ Positive + False\ Negative}$$

The F1-score is the harmonic mean of precision and recall and is calculated using the formula:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

The F1-score ranges from 0 to 1, where a higher value indicates better model performance in terms of both precision and recall. It is particularly useful when there is an imbalance between the number of positive and negative instances in the dataset. In the context of multiclass classification, two variations of the F1-score are often employed: macro-averaged F1-score and weighted F1-score. Here, the first one is the unweighted average of class-wise F1-score while the latter incorporates class weights when calculating the average.

6 RESULTS

A detailed analysis of the results obtained in the study is described here. For each classification task (Binary and multi-label) we have applied all the discussed image and text models in section 4. In the case of multimodal, 4 combinations of the top 2 performing unimodal models were analyzed. In this case, it was found to be EfficientNetV2 and BiT for image and MuRIL and RoBERTa for text.

6.1 Sub-Task 1 (Binary Misogyny Classification):

The primary objective here is to classify a meme into one of two classes (binary) - Misogynistic and Non-Misogynistic. As noted in Section 3, the classes are evenly distributed, hence no imbalance class handling technique is required. Training with unimodal data was first considered to identify the top performers in each modality. **Table 3** presents the performance of each model on various metrics discussed in Section 5. The first value in the table cell represents the average (mean) of all results obtained in 5 runs of the model in 5-fold cross-validation. While the second value shows the standard deviation (SD) from the mean.

Table 3. Model performance across different metrics in binary classification (Sub-Task 1).

Modality	Modal	Macro F1-Score [SD]	Weighted F1-Score [SD]	AUC [SD]
Image (Unimodal)	**InceptionV3**	0.6549 [0.0239]	0.6545 [0.0224]	0.7541 [0.0210]
	ViT	0.4924 [0.0288]	0.4929 [0.0291]	0.5049 [0.0302]
	EfficientNetV2	0.6957 [0.0217]	0.6961 [0.0228]	0.7659 [0.0227]
	BiT	0.7132 [0.018]	0.7143 [0.011]	0.79 [0.017]
Text (Unimodal)	**ALBERT**	0.6065 [0.0272]	0.6066 [0.0272]	0.6638 [0.0279]
	BERT	0.6125 [0.0226]	0.6126 [0.0214]	0.6668 [0.0196]
	RoBERTa	0.6209 [0.0166]	0.6207 [0.0149]	0.688 [0.0154]
	MuRIL	0.667 [0.0138]	0.6669 [0.0145]	0.7233 [0.0147]
Image + Text (Multimodal)	**EfficientNetV2 + RoBERTa**	0.7183 [0.0126]	0.7182 [0.0133]	0.801 [0.0118]
	EfficientNetV2 + MuRIL	0.72 [0.02]	0.7197 [0.024]	0.8064 [0.0182]
	BiT + RoBERTa	0.7244 [0.0281]	0.725 [0.0296]	**0.8181** [0.0244]
	BiT + MuRIL	**0.7319** [0.0218]	**0.7319** [0.0203]	0.8168 [0.0235]

Figure 7 represents a pictorial summary of the observations. Observing from both the table and the figure, one can infer that multimodal models have excelled in both performance metrics when compared to unimodal models. A closer observation also reveals that in unimodal the CNN-based image models have performed well when compared to text models. To add, the transformer-based image model (ViT) has failed miserably in this binary classification task. The overall outperformers were found to be the multimodal-based BiT+RoBERTa in AUC (AUC-ROC) and BiT+MuRIL in F1-score.

The class-wise precision, recall, and F1-scores are shown in **Table 4**. Clearly, the previously discussed multimodal-based models (BiT+RoBERTa & BiT+MuRIL) seem to compete with each other across various parameters in this context. The deep and optimized architecture of RoBERTa and the language-specific training of MuRIL when combined with the efficient downstream fine-tuning technique of BiT results in better classification.

6.2 Sub-Task 2 (Multi-label Misogyny Classification):

The second task involves the identification of correct labels for a given misogynous meme. Hence the available data point for this task becomes half, i.e. memes in the misogyny class (n = 2515). The problem is imbalanced in terms of class distribution as can be observed in Section 3. However, in this study, we have not used any imbalanced dataset handling techniques. All models were trained with the available data points without any augmentation. The list of model performance summaries on test data can be found in **Table 5**.

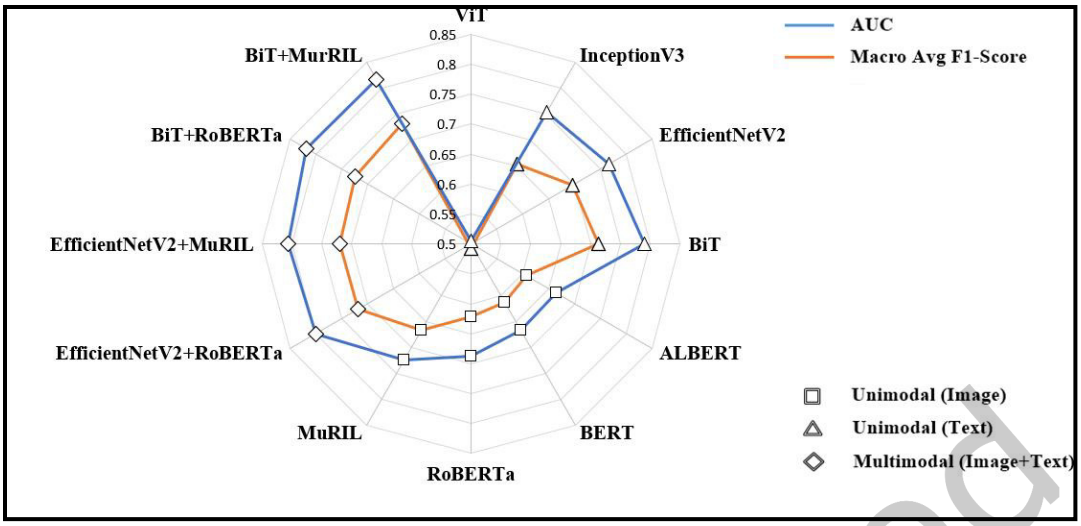

Figure 7. Depicting the performance of each model across diverse parameters (AUC & Macro Avg. F1-score) through a comparative spider chart specifically for sub-task 1.

Table 4. Class-wise values of Precision, Recall, and F1 scores for Sub-Task 1

Modality	Modal	Precision [SD]		Recall [SD]		F1-Score [SD]	
		0	1	0	1	0	1
Image (Unimodal)	InceptionV3	0.7225 [0.0408]	0.6407 [0.0554]	0.5514 [0.1257]	0.7772 [0.0959]	0.6151 [0.0572]	0.6946 [0.0223]
	ViT	0.5073 [0.0299]	0.4998 [0.0363]	0.5336 [0.1396]	0.4712 [0.1477]	0.5095 [0.0766]	0.4753 [0.0791]
	EfficientNetV2	0.6933 [0.025]	0.6996 [0.0275]	0.7081 [0.037]	0.6836 [0.0286]	0.7003 [0.0271]	0.6911 [0.0228]
	BiT	0.7488 [0.0445]	0.6964 [0.0139]	0.6681 [0.0645]	0.7603 [0.093]	0.7021 [0.0214]	0.7242 [0.0456]
Text (Unimodal)	ALBERT	0.6315 [0.0548]	0.6095 [0.0197]	0.5963 [0.108]	0.6295 [0.1445]	0.6036 [0.0368]	0.6095 [0.0711]
	BERT	0.6265 [0.0346]	0.6065 [0.0225]	0.5874 [0.0543]	0.6419 [0.0756]	0.6038 [0.0255]	0.6213 [0.041]
	RoBERTa	0.6229 [0.0345]	0.625 [0.0401]	0.6375 [0.0481]	0.609 [0.052]	0.6276 [0.0129]	0.6142 [0.0239]
	MuRIL	0.6831 [0.0321]	0.657 [0.0113]	0.6386 [0.0367]	0.6978 [0.0542]	0.6585 [0.0141]	0.6755 [0.0236]
Image + Text (Multimodal)	EfficientNetV2 + RoBERT	0.7338 [0.035]	0.7088 [0.0151]	0.6969 [0.039]	0.7416 [0.0534]	0.7132 [0.0137]	0.7234 [0.0201]
	EfficientNetV2 + MuRIL	0.7181 [0.0453]	0.7444 [0.0691]	0.7666 [0.0908]	0.6824 [0.0801]	0.7356 [0.0267]	0.7044 [0.0238]
	BiT + RoBERTa	**0.7755** [0.0463]	0.7115 [0.0502]	0.6651 [0.1239]	**0.791** [0.0899]	0.7056 [0.0598]	**0.7432** [0.0226]
	BiT + MuRIL	0.706 [0.0323]	**0.7714** [0.025]	**0.7987** [0.0352]	0.6688 [0.0444]	**0.7487** [0.0231]	0.7152 [0.0239]

Table 5. Model performance across different metrics in multi-label classification (Sub-Task 2).

Modality	Modal	Macro Avg. F1-Score [SD]	Weighted Avg. F1-Score [SD]	AUC [SD]
Image (Unimodal)	**InceptionV3**	0.4375 [0.0122]	0.5544 [0.0123]	0.7678 [0.0246]
	ViT	0.4907 [0.0147]	0.5971 [0.0166]	0.7914 [0.016]
	EfficientNetV2	0.5025 [0.0224]	0.5795 [0.0098]	0.7702 [0.0192]
	BiT	0.5154 [0.0184]	0.6056 [0.0178]	0.7852 [0.0176]
Text (Unimodal)	**ALBERT**	0.3687 [0.0235]	0.5097 [0.0202]	0.7567 [0.0266]
	BERT	0.3584 [0.0148]	0.4928 [0.0114]	0.753 [0.0299]
	RoBERTa	0.3432 [0.0483]	0.4792 [0.0343]	0.7595 [0.0235]
	MuRIL	0.3692 [0.0321]	0.5125 [0.0284]	0.7465 [0.0197]
Image + Text (Multimodal)	**EfficientNetV2 + RoBERTa**	0.4426 [0.0369]	0.5576 [0.0367]	0.7743 [0.0194]
	EfficientNetV2 + MuRIL	0.4872 [0.0325]	0.5949 [0.0247]	0.7837 [0.0169]
	BiT + RoBERTa	**0.527** [0.0415]	**0.6303** [0.0244]	**0.8183** [0.0117]
	Bit + MuRIL	0.5217 [0.0257]	0.6159 [0.0254]	0.815 [0.0176]

The performance of each model can also be tracked in **Figure 8**. In this task also, the multimodal-based models are seen to perform better than unimodal models. To add, we have a clear winner, i.e. (BiT+RoBERTa) surpassing other models in every metric. The performance of BiT+ MuRIL was also observed to be very close to the best performer except in the weighted average F1 score. The trend of image-based models outperforming text-based models persists in this task as well. Text-based models can also be seen to perform miserably in terms of weighted F1-score. **Table 6** presents the class-wise values of Precision, Recall, and F1 scores for Sub-Task 2 for a more detailed understanding of the model performance.

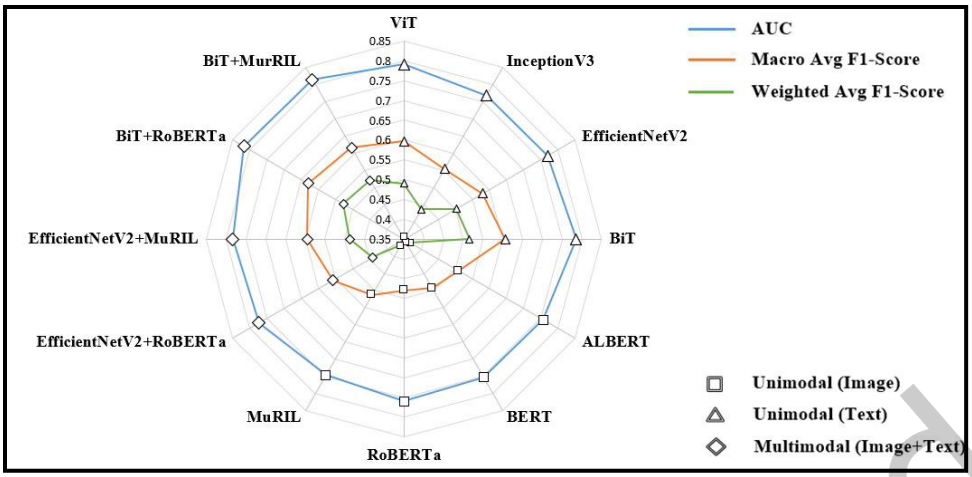

Figure 8. Depicting the performance of each model across diverse parameters (AUC, Macro Avg. F1-score & Weighted Avg. F1-score) through a comparative spider chart specifically for sub-task 2.

Table 6. Class-wise values of Precision, Recall, and F1 scores for Sub-Task 2

Modality	Modal	Precision [SD]			Recall [SD]			F1-Score [SD]		
		0	1	2	0	1	2	0	1	2
Image (Unimodal)	InceptionV3	0.6785 [0.0133]	0.5716 [0.0458]	0.5026 [0.122]	0.7912 [0.0594]	0.3802 [0.0475]	0.0764 [0.0185]	0.7291 [0.0235]	0.4532 [0.0256]	0.1302 [0.0257]
	ViT	0.7382 [0.0163]	0.5567 [0.0333]	0.4918 [0.1083]	0.7461 [0.0323]	0.5081 [0.0753	0.1279 [0.0285]	0.7419 [0.0236]	0.5295 [0.0346]	0.2005 [0.0419]
	EfficientNetV2	0.7314 [0.037]	0.5395 [0.0206]	0.4004 [0.0623]	0.729 [0.0234]	0.5004 [0.0287]	0.1974 [0.0612]	0.7292 [0.0143]	0.5185 [0.0166]	0.2599 [0.0583]
	BiT	0.7093 [0.031]	0.5877 [0.033]	0.448 [0.0969]	0.794 [0.0295]	0.469 [0.1155]	**0.2211** [0.0461]	0.7484 [0.0185]	0.5109 [0.073]	**0.2868** [0.0328]
Text (Unimodal)	ALBERT	0.6347 [0.0157]	0.5243 [0.0496]	0.0 [0.0]	**0.832** [0.0854]	0.3236 [0.0995]	0.0 [0.0]	0.7174 [0.0213]	0.3886 [0.0807]	0.0 [0.0]
	BERT	0.6532 [0.0225]	0.5513 [0.0749]	0.0 [0.0]	0.7306 [0.0919]	0.3209 [0.0912]	0.0 [0.0]	0.6858 [0.0318]	0.3894 [0.0714]	0.0 [0.0]
	RoBERTa	0.6546 [0.0438]	**0.6389** [0.1838]	0.2 [0.4]	0.7576 [0.139]	0.2791 [0.1667]	0.0025 [0.005]	0.6926 [0.0359]	0.3321 [0.1776]	0.0049 [0.0099]
	MuRIL	0.6549 [0.0145]	0.5147 [0.0333]	0.0 [0.0]	0.8239 [0.0959]	0.3261 [0.1478]	0.0 [0.0]	0.7266 [0.0333]	0.3811 [0.1165]	0.0 [0.0]
Image + Text (Multimodal)	EfficientNetV2 + RoBERTa	0.7129 [0.0304]	0.5508 [0.0451]	0.4099 [0.1091]	0.7319 [0.0271]	0.421 [0.0965]	0.0841 [0.0373]	0.722 [0.0248]	0.4705 [0.0666]	0.1354 [0.0508]
	EfficientNet + MuRIL	0.7151 [0.0231]	0.6018 [0.0658]	0.4118 [0.0987]	0.7734 [0.0354]	0.4927 [0.1004]	0.1337 [0.0702]	0.742 [0.0046]	0.5291 [0.0487]	0.1905 [0.0906]
	Bit + RoBERTa	0.7454 [0.0279]	0.6132 [0.032]	**0.7391** [0.1499]	0.7649 [0.0526]	**0.6025** [0.0436]	0.1436 [0.085]	**0.7532** [0.018]	**0.6056** [0.0125]	0.2221 [0.1101]
	Bit + MuRIL	**0.7531** [0.0335]	0.5887 [0.0257]	0.675 [0.1813]	0.7395 [0.0436]	0.5285 [0.0917]	0.1708 [0.0322]	0.7447 [0.0186]	0.552 [0.0473]	0.2684 [0.042]

7 DISCUSSION

Social media is now full of instances of hateful, abusive, and discriminatory content. This content is usually in the form of texts, images, videos or a mix of them. A part of this content targets women and tries to objectify or discriminate them in different ways. One such instance is misogyny that needs to be identified for taking corrective actions. However, research efforts on design

of automatic detection methods for multimodal content involving low-resource language content is not that well-developed. This study, therefore, attempted to focus on detecting misogyny from memes in Hindi-English code-mixed language. A large-sized dataset, consisting of 5,054 memes, is created specifically for detecting misogyny in multimodal content (memes) in Hindi-English code-mixed language. This dataset is then used to design suitable computational models for the automatic detection of such content. First, unimodal models for text and image data are implemented separately. Subsequently, a multimodal method is proposed that utilizes a late fusion technique to combine text and image features. The results obtained confirm the suitability of the developed dataset for further use and research in the area.

In this context, it's important to observe that, when focusing on unimodality, image-only models typically outperform their text-only counterparts. This strengthens and aligns with the belief that the visual characteristics of a meme dominate in categorizing it as misogynous. In addition, the effectiveness of text-only models is further constrained by the efficiency of the employed text extraction model. However, one should also not neglect the fact that the interpretation of a meme changes radically when some maliciously intended text is added. Thus, to grasp the correct intent of the meme, both its visual components and the semantics of its text play a crucial role. This gives the models trained on multiple modalities an edge in the classification task. The multimodal approach proposed in the study has exhibited superior performance in both of the subtasks. However, there may still be several possibilities for further improving the models. One may try to implement a better multilingual text extraction model, coupled with improvements in each modality-based model. Furthermore, augmenting the data points in the target class, either through manual means or synthetic methods, may further contribute to enhancing the classification accuracy.

The present work has thus contributed to the research effort toward the automatic detection of misogynous content by providing a suitable large-sized dataset and appropriate computational models. The results obtained show good performance indicating the usefulness and contribution of the work done. To the best of our knowledge, this dataset developed is the first such multimodal dataset involving images and the associated text in Hindi-English code-mixed language. Thus, it can be used to further advance the research in the area. The present work, however, has certain limitations too that can be addressed in future works. One of the challenges in detecting misogyny in Hindi-English code-mixed memes is the limited size of available datasets. To improve model performance, more data should be collected and added to the training set. Additionally, accurately extracting the code-mixed text from memes, especially in low-resource languages, poses a significant challenge. Misogynistic memes often use sarcasm, which can be difficult for models to identify. Therefore, models need to be trained to effectively recognize and interpret sarcasm in memes to improve the accuracy of misogyny detection.

The research in the area of automatic detection of misogynous multimodal content has several additional challenges to be addressed. Some of these challenges that can be addressed in future research may be listed as follows:

- **Slang and Informal Language**: Memes often use slang and informal language, which can be difficult for models to interpret, especially when it is mixed with multiple languages.
- **Misogynistic Humor:** Detecting misogyny requires understanding subtle forms of humor and sarcasm, which can be challenging for models, especially in a cross-language context.
- **Discrimination Intersectionality:** Misogyny intersects with other forms of discrimination, such as racism, homophobia, and transphobia. Identifying and disentangling these intersecting dimensions from content requires sophisticated algorithms capable of recognizing multiple layers of bias and discrimination.
- **Multimodality:** Memes combine text and images, however, there exist other forms of multimodal data such as animated GIFs, short videos, and reels. Understanding the relationship among these modalities is crucial for detecting misogyny accurately. This requires suitable computational methods that can process different modalities effectively.
- **Explainability:** The explainability of models for misogyny detection from memes in Hindi-English code-mixed language is crucial for understanding how these models make decisions. Given the complexity of code-mixed memes and the variations of language and culture they involve, explainability can help users trust the models' outputs and identify potential biases or errors in their judgments.

8 CONCLUSION

The paper presents research efforts toward the automatic detection of misogyny in multimodal Internet content. In this regard, a large-sized curated and annotated corpus of memes involving Hindi-English code-mixed language text is developed. The inter-annotator agreement computed confirms the suitability of the developed dataset. Two subtasks are provided in the dataset, one each involving binary classification and multiclass classification. Different computational models, including text-only, image-only and multimodal approaches, are applied on the dataset for detecting and categorizing the misogynistic memes. The image-

only model shows better performance than text-only models in both subtasks. The multimodal models demonstrate superior performance compared to the implemented unimodal models in both subtasks. In Subtask-1, the multimodal-based BiT+RoBERTa and BiT+MuRIL show good performance. In Subtask-2, BiT+RoBERTa consistently outperformed other models in various metrics such as Macro-average F1-score, Weighted Average F1-score, and AUC. The results firmly establish that it is important to consider both the textual and image content of the memes for better classification and understanding. Overall, the newly developed dataset, coupled with implemented multimodal models, serves as a valuable resource for advancing the research of automatic detection and categorization of misogynistic memes involving Hindi-English code-mixed language. In addition, this framework can have practical applications in real-world settings, where memes posted on social media content can be scanned by the model trained on the developed dataset and memes that are potentially misogynistic can be identified and appropriately dealt with. Such development will help in online mental well-being of women using different social media platforms.

Data availability

The dataset is accessible at https://www.kaggle.com/datasets/aakash941/mimic-dataset

References

[1] Shifman, L. (2013). Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of computer-mediated communication*, *18*(3), 362-377.

[2] Sharma, D., Gupta, V., & Singh, V. K. (2022, December). Detection of homophobia & transphobia in Malayalam and Tamil: Exploring deep learning methods. In *International Conference on Advanced Network Technologies and Intelligent Computing* (pp. 217-226). Cham: Springer Nature Switzerland.

[3] Sharma, D., Singh, A., & Singh, V. K. (2024). THAR-Targeted Hate Speech Against Religion: A high-quality Hindi-English code-mixed Dataset with the Application of Deep Learning Models for Automatic Detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[4] Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive language detection using multi-level classification. In Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, Canada, May 31–June 2, 2010. Proceedings 23 (pp. 16-27). Springer Berlin Heidelberg.

[5] Chakraborty, A., Joardar, S., & Sekh, A. A. (2023). Ensemble Classifier for Hindi Hostile Content Detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[6] Sharma, D., Singh, V. K., & Gupta, V. (2023). TABHATE: A Target-based Hate Speech Detection Dataset in Hindi. *Research Square*, 1-12.

[7] Sharma, D., Gupta, V., & Singh, V. K. (2024). Abusive comment detection in Tamil using deep learning. In *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications* (pp. 207-226). Morgan Kaufmann.

[8] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, *37*, 98-125.

[9] Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny detection in twitter: a multilingual and cross-domain study. *Information processing & management*, *57*(6), 102360.

[10] Paciello, M., D'Errico, F., Saleri, G., & Lamponi, E. (2021). Online sexist meme and its effects on moral and emotional processes in social media. *Computers in human behavior*, *116*, 106655.

[11] Singh, A., Kanaujia, A., & Singh, V. K. (2022). Research on Sustainable Development Goals: How has Indian Scientific Community Responded?. *Journal of Scientific & Industrial Research*, *81*(11), 1147-1161.

[12] Ahluwalia, R., Soni, H., Callow, E., Nascimento, A., & De Cock, M. (2018). Detecting hate speech against women in english tweets. *EVALITA Evaluation of NLP and Speech Tools for Italian*, *12*, 194.

[13] Pascucci, A., Manna, R., Masucci, V., & Monti, J. (2020, May). The role of computational stylometry in identifying (misogynistic) aggression in english social media texts. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 69-75).

[14] Goenaga, I., Atutxa, A., Gojenola, K., Casillas, A., de Ilarraza, A. D., Ezeiza, N., ... & Perez-de-Viñaspre, O. (2018, September). Automatic Misogyny Identification Using Neural Networks. In *IberEval@ SEPLN* (pp. 249-254).

[15] Singh, A., Kanaujia, A., Singh, V. K., & Vinuesa, R. (2023). Artificial intelligence for Sustainable Development Goals: Bibliometric patterns and concept evolution trajectories. *Sustainable Development*, in press. *DOI: 10.1002/sd.2706*

[16] García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R., & Valencia-García, R. (2021). Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, *114*, 506-518.

[17] Rahali, A., Akhloufi, M. A., Therien-Daniel, A. M., & Brassard-Gourdeau, E. (2021, October). Automatic Misogyny Detection in Social Media Platforms using Attention-based Bidirectional-LSTM. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2706-2711). IEEE.

[18] Aldana-Bobadilla, E., Molina-Villegas, A., Montelongo-Padilla, Y., Lopez-Arevalo, I., & S Sordia, O. (2021). A language model for misogyny detection in latin american spanish driven by multisource feature extraction and transformers. *Applied Sciences*, *11*(21), 10467.

[19] Sharma, M., Kandasamy, I., & Vasantha, W. B. (2022, July). R2d2 at semeval-2022 task 5: Attention is only as good as its values! a multimodal system for identifying misogynist memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 761-770).

[20] Attanasio, G., Nozza, D., & Bianchi, F. (2022). MilaNLP at semeval-2022 task 5: Using perceiver IO for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

[21] Rizzi, G., Gasparini, F., Saibene, A., Rosso, P., & Fersini, E. (2023). Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, *60*(5), 103474.

[22] Zhou, Z., Zhao, H., Dong, J., Ding, N., Liu, X., & Zhang, K. (2022, July). DD-TIG at semeval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 563-570).

[23] Gasparini, F., Rizzi, G., Saibene, A., & Fersini, E. (2022). Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, *44*, 108526.

[24] Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P. & Sorensen, J. (2022, July). SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 533-549).

[25] Chen, L. (2022, July). RIT boston at semeval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 636-641).

[26] Zhang, J., & Wang, Y. (2022, July). SRCB at semeval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 585-596).

[27] Muti, A., Korre, K., & Barrón-Cedeño, A. (2022, July). UniBO at semeval-2022 task 5: A multimodal bi-transformer approach to the binary and fine-grained identification of misogyny in memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 663-672).

[28] Arango, A., Perez-Martin, J., & Labrada, A. (2022, July). Hateu at semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 581-584).

[29] Gu, Y., Castro, I., & Tyson, G. (2022, July). MMVAE at semeval-2022 task 5: A multi-modal multi-task VAE on misogynous meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 700-710).

[30] Han, C., Wang, J., & Zhang, X. (2022, July). YNU-HPCC at semeval-2022 task 5: Multi-modal and multi-label emotion classification based on LXMERT. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 748-755).

[31] Cordon, P., Diaz, P. G., Mata, J., & Pachón, V. (2022, July). I2c at semeval-2022 task 5: Identification of misogyny in internet memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 689-694).

[32] Agrawal, S., & Mamidi, R. (2022, July). Lastresort at semeval-2022 task 5: Towards misogyny identification using visual linguistic model ensembles and task-specific pretraining. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 575-580).

[33] Ravagli, J., & Vaiani, L. (2022, July). JRLV at semeval-2022 task 5: The importance of visual elements for misogyny identification in memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 610-617).

[34] Sharma, G., Gitte, G. S., Goyal, S., & Sharma, R. (2022, July). IITR codebusters at semeval-2022 task 5: Misogyny identification using transformers. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 728-732).

[35] Rao, A. R., & Rao, A. (2022, July). ASRtrans at semeval-2022 task 5: Transformer-based models for meme classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 597-604).

[36] Sivanaiah, R., Angel, S., Rajendram, S. M., & Mirnalinee, T. T. (2022, July). TechSSN at semeval-2022 task 5: Multimedia automatic misogyny identification using deep learning models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 571-574).

[37] Habash, M., Daqour, Y., Abdullah, M., & Al-Ayyoub, M. (2022, July). YMAI at SemEval-2022 Task 5: Detecting Misogyny in Memes using VisualBERT and MMBT MultiModal Pre-trained Models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 780-784).

[38] García-Díaz, J., Caparros-Laiz, C., & Valencia-García, R. (2022, July). UMUTeam at SemEval-2022 Task 5: Combining image and textual embeddings for multi-modal automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 742-747).

[39] Szymanski, D. M., Moffitt, L. B., & Carr, E. R. (2011). Sexual objectification of women: Advances to theory and research 1ψ7. *The Counseling Psychologist*, *39*(1), 6-38.

[40] Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and social psychology bulletin*, *15*(4), 543-558.

[41] Van Royen, K., Poels, K., Vandebosch, H., & Walrave, M. (2018). Slut-Shaming 2.0. *Sexting: Motives and risk in online sexual self-presentation*, 81-98.

[42] Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability.

[43] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[44] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

[45] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[46] Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., ... & Talukdar, P. (2021). Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

[47] Singh, A., & Singh, V. K. (2022, December). Exploring Deep Learning Methods for Classification of Synthetic Aperture Radar Images: Towards NextGen Convolutions via Transformers. In *International Conference on Advanced Network Technologies and Intelligent Computing* (pp. 249-260). Cham: Springer Nature Switzerland.

[48] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Preprint arXiv:2010.11929*. Retrieved from https://arxiv.org/abs/2010.11929.

[49] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. Retrieved from https://proceedings.neurips.cc/paper/7181-attention-is-all**.**

[50] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html

[51] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html

[52] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. *In International conference on machine learning* (pp. 6105-6114). PMLR.

[53] Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. *International Conference on Machine Learning*, 10096–10106. PMLR. Retrieved from http://proceedings.mlr.press/v139/tan21a.html

[54] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2020). Big Transfer (BiT): General Visual Representation Learning. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 491–507). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58558-7_29

[55] Singh, A., & Singh, V. K. (2023). A hybrid transformer–sequencer approach for age and gender classification from in-wild facial images. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-023-09087-7

[56] Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press. Retrieved from https://books.google.com/books?hl=en&lr=&id=RC43AgAAQBAJ&oi=fnd&pg=PR7&dq=%5B42%5D.%09Murphy,+K.+P.+(2012).+Machine+learning:+a+probabilistic+perspective.+MIT+press.&ots=unfw8yMo19&sig=5b7qpO7TPkR6tiVbiHNtSfu-8YU

[57] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2021, October 25). *On the Variance of the Adaptive Learning Rate and Beyond*. arXiv. Retrieved from http://arxiv.org/abs/1908.03265