MDC^3: A Novel Multimodal Dataset for Commercial Content Classification in Bengali

Anik Mahmud Shanto[1], **MST. Sanjida Jamal Priya**[1]
Fahim Shakil Tamim[1,2] and **Mohammed Moshiul Hoque**[1]
[1] Chittagong University of Engineering & Technology, Chittagong-4349, Bangladesh
[2] International University of Business Agriculture and Technology, Dhaka, Bangladesh
{u1904049, u1904057}@student.cuet.ac.bd,
fstamim9@gmail.com, moshiul_240@cuet.ac.bd

Abstract

Identifying commercial posts in resource-constrained languages among diverse and unstructured content remains a significant challenge for automatic text classification tasks. To address this, this work introduces a novel dataset named **MDC**^{3} (Multimodal Dataset for Commercial Content Classification), comprising 5,007 annotated Bengali social media posts classified as commercial and noncommercial. A comprehensive annotation guideline accompanies the dataset to aid future creation in resource-constrained languages. Furthermore, we performed extensive experiments on MDC^3 considering both unimodal and multimodal domains. Specifically, the late fusion of textual (mBERT) and visual (ViT) models (i.e., **ViT+mBERT**) achieves the highest F1 score of 90.91, significantly surpassing other baselines.

1 Introduction

Social media platforms are crucial for communication, enabling individuals and businesses to share diverse content. Commercial posts influence consumer behavior and brand perception, making their identification essential for transparency, consumer protection, and regulatory compliance (McQuarrie and Munson, 2014; Boerman and van Reijmersdal, 2016a). However, native advertising and influencer marketing blur the lines between ads and personal content, complicating detection (Boerman and van Reijmersdal, 2016b; Chia, 2012).

Detecting commercial posts is vital for targeted advertising, brand monitoring, and consumer behavior analysis. While most research focuses on English, Bengali social media lacks annotated datasets and faces challenges from multimodal content and cultural nuances. To address this, we introduce **MDC**^{3}, a dataset for classifying Bengali social media posts as commercial or non-commercial. Various unimodal and multimodal baselines are explored based on **MDC**^{3}.

Key contributions:

- Introduced **MDC**^{3}, a multimodal dataset with 5,007 labeled samples.
- Evaluated unimodal and multimodal baselines for Bengali commercial content classification.

2 Related Work

Social media's growing influence has made user-generated and influencer-created content pivotal in shaping consumer behavior (Gamage and Ashill, 2023). Influencers act as trusted figures but often blur the lines between personal and commercial content, complicating automatic detection (Vanninen et al., 2023; Weismueller et al., 2022; Ahammad et al., 2024). Subtle advertising strategies, such as conversational language and self-focused visuals, enhance audience engagement but hinder content classification (Hidarto and Andrieza, 2022; Kim et al., 2020). Multimodal approaches have shown promise in addressing these challenges. Vedula et al. (2017) leveraged text, audio, and video embeddings for ad effectiveness prediction, while Villegas et al. (2023) introduced datasets combining textual and visual modalities for better ad detection. These studies highlight the advantages of multimodal models over unimodal counterparts.

Beyond influencer marketing, multimodal research spans diverse applications, including

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 311–320

trend detection (Pandit et al., 2019), COVID-19 impact analysis (Unal et al., 2022), and gender-based communication studies (Hidarto and Andrieza, 2022). However, resource-constrained languages like Bengali remain underexplored, with most models trained on English-centric datasets, limiting their applicability to non-English contexts. This study addresses this gap by introducing a Bengali dataset for commercial content classification and proposing a multimodal approach that combines textual and visual features to enhance classification accuracy.

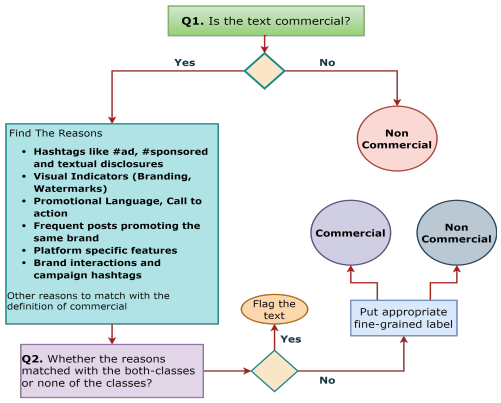


Figure 1: Data Annotation Process

3 Developement of MDC[3]

As per our investigation's outcome, no benchmark dataset is explicitly available for detecting influencer commercial content in Bengali. Therefore, this work presents a benchmark dataset, **MDC**[3] (Multimodal Commercial Content Classification Dataset), from Bengali social media posts comprising Facebook and Instagram posts categorized into two classes, *commercial* and *non-commercial*. The definitions of each class within the dataset are provided below, as described by (Villegas et al., 2023).

- **Commercial (Com):** Commercial posts promote or endorse a brand or its products or services, a free product or service, or any other incentive.
- **Non-commercial (NCom):** Non-commercial posts refer to organic content such as personal ideas, comments, and life updates that do not aim to be monetized.

3.1 Data Collection and Annotation

From April to November 2024, we collected 5,007 multimodal influencer posts from Facebook (66.2%) and Instagram (33.8%), including 2,750 commercial and 2,257 non-commercial entries. These were sourced from Bangladeshi influencers and commercial pages. The dataset prioritizes authenticity by including only Bengali content with authentic or captured visuals. To ensure ethical standards, all data were sourced from publicly accessible domains, excluding entries without multimodal elements, unclear visuals, cartoons, or insufficient text.

Three experienced annotators annotated the dataset following clear guidelines on labeling (Figure 1), tool use, and quality standards. Annotators received training to ensure adherence to criteria for commercial and non-commercial content. The process was independent and there were regular meetings to resolve ambiguities and reach consensus. A senior professor with 10+ years of experience evaluated inter-annotator agreement using majority voting (Algorithm 1), ensuring the dataset's reliability. Appendix A describes the details of the majority voting algorithm.

We applied inter-annotator agreement standards (i.e., Cohen's kappa coefficient (Cohen, 1960), to measure the quality of the annotations. On the kappa scale, we achieved 0.86 implies an almost ideal agreement.

3.2 Statistics

We have stratified the dataset into three sets: train (60%), validation (20%), and test (20%). Table 1 demonstrates the statistics of the dataset.

Figure 2 illustrates samples of the dataset. The dataset is available on online: https://github.com/anik5099/Multimodal-Commercial-Content-Classification-Binary

4 Methodology

This section describes the baseline models for classifying commercial content, which in-

Split	Com	NCom	T_W	U_W
Train	1631	1372	57292	6930
Val	546	456	17843	4291
Test	573	429	20086	4454
Total	2740	2257	95221	15675

Table 1: Class distribution in train, validation, and test sets. The acronyms T_W, U_W denotes total words and unique words, respectively

Figure 2: Dataset samples.

clude unimodal (visual or textual) and multimodal (visual and textual) models. Figure 3 depicts the abstract process of commercial content classification employing textual and visual modalities. For this classification task, image and textual modalities have been trained separately. Then the unimodal models have been fused to produce multimodal classification. **Experimental Settings** is explained in Appendix 5.

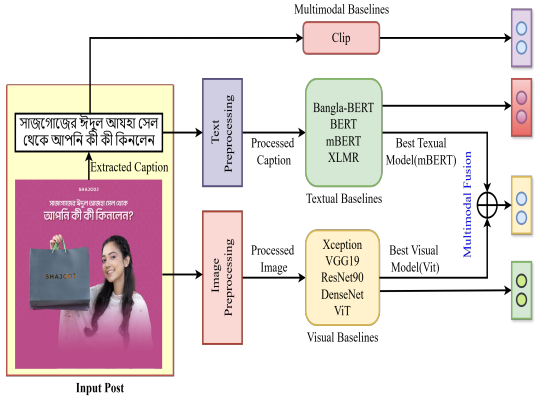


Figure 3: Abstract process of multimodal content classification

4.1 Data Preprocessing

For the multimodal commercial content classification task, separate pipelines have been applied to prepare text and image data. Captions have been cleaned, tokenized, and converted into input IDs and attention masks using a subword tokenizer for text data. Sequences have been padded or truncated to a fixed length. Images have been resized to 224×224, and normalized. Processed images have been converted to tensors. The dataset labels have been encoded (1 for commercial and '0' for non-commercial). The dataset has been split into 60-20-20 for training, validation and test set.

4.2 Unimodal Baselines

This work explores several unimodal (visual or textual) and multimodal (visual and textual) baselines to classify the commercial content in Bengali. The **Hyperparameter Settings** for unimodal and multimodal models have been described in section B.

4.2.1 Visual Modality

We fine-tuned prominent convolutional neural network (CNN) architectures to classify visual data. The input images were resized to $224 \times 224 \times 3$ and preprocessed using standard normalization techniques. Specifically, we employed the following CNN models:

- **Xception** (Chollet, 2017): A depthwise separable convolutional network optimized for computational efficiency.
- **VGG19** (Simonyan and Zisserman, 2015): Known for its more profound architecture, this model emphasizes hierarchical feature extraction.
- **ResNet50** (He et al., 2015): A residual network addressing vanishing gradient issues through skip connections.
- **DenseNet** (Huang et al., 2018): A densely connected architecture designed to enhance feature reuse across layers.
- **ViT** (Dosovitskiy et al., 2021): Vision Transformer (ViT) uses self-attention to

process image patches as sequences, achieving competitive performance in image recognition tasks compared to traditional CNNs.

The top layers of each model were replaced with a dense layer of 32 neurons and a sigmoid layer for binary classification. These architectures were fine-tuned on our dataset using transfer learning techniques.

4.2.2 Textual Modality

Transformer-based models have been proven superior in many text classification tasks (Shanto et al., 2024; Chowdhury et al., 2024; Tamim et al., 2023a,b). Therefore, for text classification, we utilized transformer-based architectures, fine-tuned for our specific task:

- **BERT** (Devlin et al., 2019): A bidirectional transformer pre-trained on large-scale English text corpora.
- **mBERT** (Devlin et al., 2019): A multilingual version of BERT designed for cross-lingual tasks.
- **Bangla-BERT** (Bhattacharjee et al., 2022): A language-specific transformer model pre-trained on Bangla text.
- **XLM-Roberta** (Conneau et al., 2020): A cross-lingual transformer optimized for multilingual tasks.

Each model processed tokenized text sequences and generated contextual embeddings of size 768. These embeddings were passed through a dense layer with 32 neurons, followed by a sigmoid layer for classification.

4.3 Multimodal Baselines

This work exploited several multimodal techniques to analyze the multimodal data (visual and textual). A dense layer with 768 neurons was applied separately to the visual and textual modalities for model construction. The outputs from these layers were then concatenated to create a combined representation of visual and textual features. This combined feature was further processed through another dense layer of 768 neurons, followed by a softmax layer to classify posts into commercial or non-commercial categories. We have used the late fusion technique for measuring baselines because late fusion is more interpretable and allows each modality to leverage its unique characteristics. Besides, Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) was employed to align visual and textual data, leveraging contrastive learning techniques (Chen et al., 2020).

The late fusion process can be mathematically expressed as follows:

$$\mathbf{z}_{\text{combined}} = \text{Concat}\left(\mathbf{z}_{\text{visual}}, \mathbf{z}_{\text{text}}\right) \quad (1)$$

Here, the visual and textual features are extracted as:

$$\mathbf{z}_{\text{visual}} = \sigma(\mathbf{W}_{\text{visual}}\mathbf{x}_{\text{visual}} + \mathbf{b}_{\text{visual}}) \quad (2)$$

$$\mathbf{z}_{\text{text}} = \sigma(\mathbf{W}_{\text{text}}\mathbf{x}_{\text{text}} + \mathbf{b}_{\text{text}}) \quad (3)$$

where σ is the activation function (Softmax), $\mathbf{W}_{\text{visual}}$ and $\mathbf{W}_{\text{text}}$ are the weights, and $\mathbf{b}_{\text{visual}}$ and $\mathbf{b}_{\text{text}}$ are the biases for the respective modalities. Concat denotes the concatenation operation.

The concatenated feature vector $\mathbf{z}_{\text{combined}}$ is passed through a dense layer followed by a softmax layer for classification:

$$\mathbf{y}_{\text{pred}} = \text{Softmax}(\mathbf{W}_{\text{class}}\mathbf{z}_{\text{combined}} + \mathbf{b}_{\text{class}}) \quad (4)$$

Here, $\mathbf{W}_{\text{class}}$ and $\mathbf{b}_{\text{class}}$ represent the weights and biases of the classification layer, respectively. The Softmax function maps the output to probabilities for the two categories (commercial and non-commercial).

This formulation effectively integrates features from both modalities, showcasing the power of late fusion for joint multimodal learning.

5 Experimental Setup

This section describes the summary of the experimental setup while training and evaluating our model on the dataset. The simulation was run on a personal computer with an NVIDIA GeForce GTX 2060 GPU and an Intel Core i7-9700 CPU running at 3.00 GHz. Additionally, a Kaggle Notebook with a P100 GPU was uti-

App	A(%)	P(%)	R(%)	F1(%)
Visual Only				
Xception	64.17 ± 0.87	65.78	62.67	63.59 ± 0.77
VGG19	73.65 ± 0.49	69.70	69.66	73.53 ± 0.36
VGG16	74.18 ± 0.60	73.18	70.66	74.23 ± 0.58
ResNet	79.14 ± 0.21	73.86	68.66	79.11 ± 0.16
DenseNet	67.84 ± 0.54	63.03	63.07	67.90 ± 0.43
ViT	**81.70** ± 0.013	**82.85**	**81.94**	**81.71** ± 0.011
Textual Only				
B-BERT	84.56 ± 0.01	83.60	82.91	83.21 ± 0.03
BERT	81.18 ± 0.04	83.12	79.44	81.14 ± 0.01
mBERT	**86.83** ± 0.003	**91.10**	**84.27**	**87.43** $\pm .006$
XLM-R	74.25 ± 0.27	74.08	94.44	81.95 ± 0.09
Multimodal				
CLIP	77.94 ± 0.01	77.88	77.94	77.75 ± 0.00
ViT+mBERT	**90.92** ± 0.001	**90.91**	**90.92**	**90.91** $\pm .001$

Table 2: Performance comparison of unimodal and multimodal models on the test set. The symbols A, P, R, and F1 denote accuracy, precision, recall, and F1-score, respectively. The standard deviation (±) with three random seeds is reported.

lized to ensure sufficient processing capability.

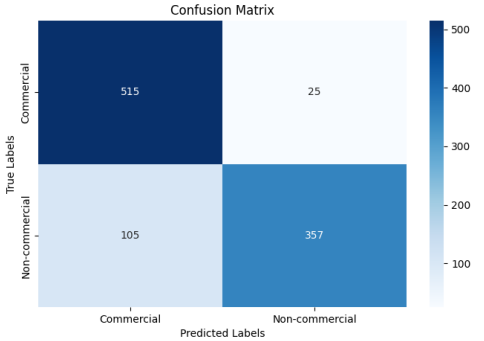
5.1 Results

Table 2 provides an overview of the performance of various unimodal and multimodal models. Among the visual-only models, ViT emerges as the top performer with an F1 score of 82.03, surpassing other models like ResNet and Xception. On the textual side, m-BERT outshines all other unimodal models, achieving a remarkable F1 score of 87.43. However, the proposed model **ViT+mBERT** demonstrates the most significant advancement, achieving an F1 score of **90.92**, marking an improvement of several percentage points over the best baseline model. This result underscores the proposed approach's superior effectiveness in leveraging visual and textual modalities. The CLIP model showed inferior performance (F1 score of 77.75%) due to the domain gap in pretraining, limited fine-tuning, lack of modality-specific processing, and insufficient task-specific adaptation.

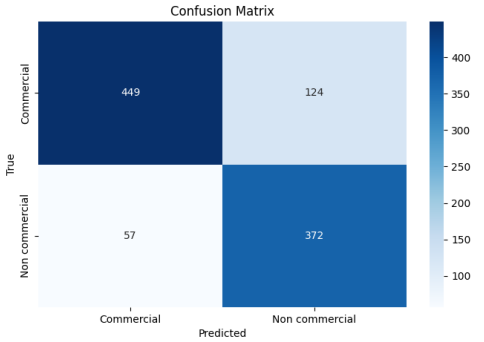
6 Error Analysis

In our study on the classification of multimodal commercial content from social media, we performed an extensive error analysis to identify the strengths and weaknesses of our proposed model. The analysis was conducted quantitatively and qualitatively to understand the model's performance comprehensively.

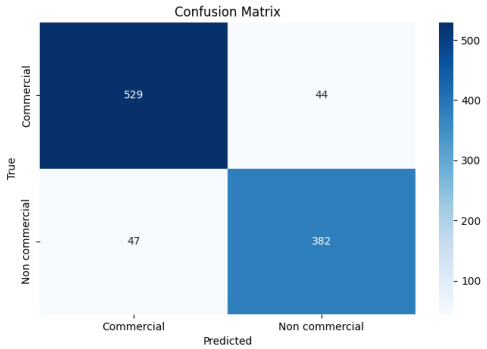
Quantitative Analysis: During the evaluation, we conducted a detailed quantitative analysis using confusion matrices to assess the model's performance.



(a) Best textual model



(b) Best visual model



(c) Proposed model

Figure 4: Confusion matrices of employed models

The confusion matrices are shown in Figure 4 states that the proposed multimodal model **ViT+mBERT**, which integrates both text and visual data, achieved an accuracy of 90.92%. This represents a significant improvement over unimodal models that rely

solely on visual or textual data, demonstrating the model's effectiveness in accurately classifying commercial content on social media platforms. The analysis revealed specific challenges, particularly with subtle commercial content. Posts that casually mentioned products without explicit promotional intent were occasionally misclassified as non-commercial despite the presence of brand-related keywords. For instance, these errors occurred in 7.68% of the cases. Conversely, non-commercial posts that included neutral or critical discussions of products were sometimes wrongly categorized as commercial, likely due to an overreliance on product-related keywords. This misclassification was observed in 10.95% of the cases. Moreover, the complexity of multimodal data led to further misclassification issues. Specifically, posts with ambiguous text and neutral visuals were prone to errors, which occurred 15.87% of the time. This underscores the need for more sophisticated textual and visual data integration to improve classification accuracy and reduce errors. Table 3 shows the error rate of both unimodal and multimodal models. The detailed **Qualitative Analysis** is explained in Appendix C.

Approach	% of Error
Visual Only	
Xception	35.83
VGG19	26.35
VGG16	25.82
ResNet	20.86
DenseNet	32.16
ViT	**18.30**
Textual Only	
BERT	18.82
XLM-R	25.75
B-BERT	15.44
m-BERT	13.17
Multimodal	
CLIP	22.25
ViT+mBERT	**9.09**

Table 3: Error rate of employed models

7 Conclusion

This paper proposed a multimodal framework for detecting commercial content in Bengali social media posts, evaluated on the newly developed **MDC**[3] dataset with 5,007 posts labeled as commercial and non-commercial. The study utilized models such as mBERT for textual features and ViT for visual features. Results show that multimodal approaches significantly outperform unimodal methods, with ViT+mBERT achieving the best performance. Error analysis identified challenges in detecting subtle advertising styles. Future work will expand the dataset, incorporate diverse domains, and explore advanced fusion techniques to improve model robustness and performance. Moreover, explainability analysis will also be included to improve the model's clarity.

Limitations

The proposed methodology utilized a late fusion, which possesses several limitations. The class imbalance within the dataset may lead to biased predictions toward the more prevalent class, thereby compromising the model's performance. The explainability of how the model mitigates bias is not explained in the paper. The dynamic nature of social media content may hinder the model's ability to generalize to novel content types not sufficiently represented in the training data. While the late fusion technique effectively merges visual and textual features, it may not fully capture the intricate interdependencies between these modalities, thus limiting the model's capacity to generate optimal predictions.

References

Tanzin Ahammad, Zaima Sartaj Taheri, and Fahim Shakil Tamim. 2024. Sentiment classification of multi-modal memes using deep learning and transformer techniques. In *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, pages 1–6.

Nadir On The Go Bangla. 2024. [Online; accessed 24-August-2024]. [link].

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *Preprint*, arXiv:2101.00204.

S. C. Boerman and E. A. van Reijmersdal. 2016a. The effects of disclosure of sponsored content on social media influencers' credibility and engagement. *Journal of Advertising*, 45(4):466–475.

Sophie C. Boerman and Eva A. van Reijmersdal. 2016b. The effects of the standardized disclosure for online sponsored content: A systematic review. *Journal of Advertising*, 45(4):458–472.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Aleena Chia. 2012. Welcome to me-mart: The politics of user-generated content in personal blogs. *American Behavioral Scientist*, 56(4):421–438.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. *Preprint*, arXiv:1610.02357.

Rafsan The Chotovai. 2024. [Online; accessed 2-January-2025]. [link].

Md. Sajid Alam Chowdhury, Mostak Chowdhury, Anik Shanto, Hasan Murad, and Udoy Das. 2024. Fired_from_NLP at AraFinNLP 2024: Dual-phase-BERT - a fine-tuned transformer-based model for multi-dialect intent detection in the financial domain for the Arabic language. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 410–414, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Neha Fun Fitness. 2024. [Online; accessed 24-September-2024]. [link].

Thilini Chathurika Gamage and Nicholas Jeremy Ashill. 2023. # sponsored-influencer marketing: effects of the commercial orientation of influencer-created content on followers' willingness to search for information. *Journal of Product & Brand Management*, 32(2):316–329.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Anderson Hidarto and Aryani Andrieza. 2022. Gender differences in influencer advertisements on instagram: A multimodal perspective. *Journal of Language and Literature*, 22(1):220–237.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely connected convolutional networks. *Preprint*, arXiv:1608.06993.

Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han, and Wei Wang. 2020. Multimodal post attentive profiling for influencer marketing. In *Proceedings of The Web Conference 2020*, pages 2878–2884.

E. F. McQuarrie and J. M. Munson. 2014. Fusing advertising and entertainment: A review of the literature. *Journal of Advertising Research*, 54(3):234–249.

Sushain Pandit, Fang Wang, Vijay Ekambaram, and Sarbajit K Rakshit. 2019. Trend identification and modification recommendations based on influencer media content analysis. US Patent App. 16/449,419.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Shajgoj. 2024. [Online; accessed 24-October-2024]. [link].

Anik Shanto, Md. Sajid Alam Chowdhury, Mostak Chowdhury, Udoy Das, and Hasan Murad. 2024. Fired_from_NLP at SemEval-2024 task 1: Towards developing semantic textual relatedness predictor - a transformer-based approach. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 859–864, Mexico City, Mexico. Association for Computational Linguistics.

Shorodindu. 2024. [Online; accessed 1-January-2025]. [link].

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Preprint*, arXiv:1409.1556.

Fahim Shakil Tamim, Sourav Saha, Avishek Das, and Mohammed Moshiul Hoque. 2023a. Detecting violence inciting texts based on pre-trained transformers. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Fahim Shakil Tamim, Zaima Sartaj Taheri, and Mohammed Moshiul Hoque. 2023b. Detecting signs of depression from social media texts using generalized autoregressive pretraining transformer model. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Mesut Erhan Unal, Adriana Kovashka, Wen-Ting Chung, and Yu-Ru Lin. 2022. Visual persuasion in covid-19 social media content: A multi-modal characterization. In *Companion Proceedings of the Web Conference 2022*, pages 694–704.

Heini Vanninen, Joel Mero, and Eveliina Kantamaa. 2023. Social media influencers as mediators of commercial messages. *Journal of Internet Commerce*, 22(sup1):S4–S27.

N Vedula, W Sun, H Lee, H Gupta, M Ogihara, J Johnson, G Ren, and S Parthasarathy. 2017. Multimodal content analysis for effective advertisements on youtube. arxiv.

Danae Sánchez Villegas, Catalina Goanta, and Nikolaos Aletras. 2023. A multimodal analysis of influencer content on twitter. *Preprint*, arXiv:2309.03064.

Jason Weismueller, Richard L Gruner, and Paul Harrigan. 2022. Consumer engagement in influencer marketing video campaigns: An abstract. In *Academy of Marketing Science Annual Conference*, pages 71–72. Springer.

A Annotation with Majority Voting

Algorithm 1 determines whether influencer posts on social media are commercial or non-commercial using majority voting from three annotators.

Algorithm 1 Majority Voting with 3 Annotators

Require: A set of posts $P = \{p_1, p_2, \ldots, p_n\}$. Each post p_i has three labels L_1, L_2, L_3 given by three annotators.

Ensure: Final labels for each post indicate "Commercial" (C) or "Non-Commercial" (NC).

function MajorityVoting($annotations$)
  $final\_labels \leftarrow []$
  for all $annotation \in annotations$ **do**
    $C, NC \leftarrow 0, 0$
    for all $label \in annotation$ **do**
      if $label ==' C'$ **then** $C \leftarrow C + 1$
      else if $label ==' NC'$ **then** $NC \leftarrow NC + 1$
    end for
    if $C \geq 2$ **then** $final\_labels$.append($'Commercial'$)
    else $final\_labels$.append($'Non-Commercial'$)
  end for
  return $final\_labels$
end function

B Hyperparameter Configuration

Different hyperparameters are tuned for visual and textual models. The best multimodal model's hyperparameters are also tuned based on the training dataset.

Textual Models: The textual models leverage transformer-based architectures like Bangla-BERT and m-BERT, requiring specific configurations to handle tokenized text effectively. The hyperparameters were fine-tuned to ensure optimal training for the text modality. Table 4 gives a brief overview of the parameter setups we have used in the model.

Parameter	Value
Learning Rate	5×10^{-5}
Optimizer	AdamW
Batch Size	16
Number of Epochs	8
Loss Function	CrossEntropyLoss
Maximum Sequence Length	128
Warmup Steps	500
Weight Decay	1×10^{-2}

Table 4: Hyperparameter configurations for textual models

Visual Models: For visual models, the configurations were adapted to focus on efficient processing of high-dimensional image data. Specific adjustments were made to cater to the requirements of convolutional networks. Table 5 gives a brief overview of the parameter setups we have used in the model.

Parameter	Value
Learning Rate	1×10^{-4}
Optimizer	SGD
Batch Size	64
Number of Epochs	20
Loss Function	BinaryCrossEntropyLoss
Image Size	224×224
Weight Decay	5×10^{-4}
Momentum	0.9

Table 5: Hyperparameter configurations for visual models

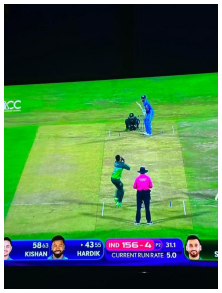
Multimodal Models: The multimodal models were designed to effectively integrate visual and textual features, leveraging their complementary nature for improved classification performance. The hyperparameters were carefully chosen to balance the unique requirements of each modality while optimizing the late fusion process. Table 6 gives a brief overview of the parameter setups we have used in the model.

C Qualitative Analysis

The qualitative analysis revealed that the model effectively identifies straightforward commercial content by integrating textual and visual cues, accurately classifying posts with explicit promotional language and product images. However, it struggles with subtle commercial content and ambiguous multimodal posts.

Parameter	Value
Learning Rate	2×10^{-5}
Optimizer	AdamW
Batch Size	32
Number of Epochs	10
Loss Function	BinaryCrossEntropyLoss
Fusion Method	Late Fusion (Concatenation)
Visual Feature Dimension	512
Textual Feature Dimension	768
Combined Feature Dimension	768
Dropout Rate	0.3
Weight Decay	1×10^{-3}

Table 6: Hyperparameter configurations for multimodal models



দাদাদের ৩৫০+ রান হলেই একটি পিৎজার সাথে আরেকটি ফ্রি
(If Dadas score 350+ runs, Buy one pizza, get one free)

(a)
Textual: Com (✓)
Visual: Non-Com (✗)
ViT+mBERT: Com (✓)

দেশাল ঈদ কালেকশন
(Deshal Eid Collection)

(b)
Actual: Com
Predicted: Non-Com

Figure 5: Example (a) illustrates a picture where the proposed method produces better predictions, and example (b) illustrates a wrongly classified sample. The symbols (✓) and (✗) indicate the correct and incorrect prediction.

For example, posts with brand mentions or subtle product placements and those with abstract images and vague text were often misclassified. These findings highlight the need for improved semantic understanding and differentiation between neutral and promotional content. Figure 5 depicts the label of data samples predicted by the proposed model.

D Social Media Profiles and Activity

The developed dataset is dedicated to multi-modal commercial content classification tasks. For developing the dataset, we have collected data samples from many social media pages. Table 7 shows some data sources from where data have been collected.

Name	Type	Affiliation	Popularity
Shajgoj (Shajgoj, 2024)	FP/IG	Beauty	2.1M
Nadir On The Go Bangla (Bangla, 2024)	FP	Travel	2.8M
Neha Fun & Fitness (Fitness, 2024)	FP/IG	Fitness	1.9M
Rafsan The Choto-vai (Chotovai, 2024)	FP/IG	Food	4.3M
Shorodindu (Shorodindu, 2024)	FP	Lifestyle	620k

Table 7: Social media profiles and activity