International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, Delaware, USA
Volume 13– No. 1, August 2025 – www.ijais.org

FusionGuard: A Multimodal Adversarially Aware Classifier for Robust Image-Text Classification

Emmanuel Ludivin Tchuindjang Tchokote
Department of Computer Engineering,
Faculty of Engineering and Technology (FET),
University of Buea, P.O. Box 63, Buea, Cameroon

Elie Fute Tagne
Department of Computer Engineering,
Faculty of Engineering and Technology (FET),
University of Buea, P.O. Box 63, Buea, Cameroon

ABSTRACT
From social media content moderation to medical image diagnosis and multimedia retrieval, multimodal classification models that integrate textual and visual information are increasingly becoming the center of interest. Despite their improved performance over unimodal systems, these models are often vulnerable to adversarial attacks that exploit modality-specific weaknesses that lead to misclassification and unreliable outcomes. Moreover, real-world datasets commonly suffer from class imbalance which further degrades model generalization. To address these challenges, the researchers developed FusionGuard, a novel multimodal classification framework that combines the complementary strengths of TinyBERT for text encoding and EfficientNet for image feature extraction within a hybrid fusion architecture. Furthermore, they incorporated adversarial training to enhance robustness against adversarial attacks. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) was used to mitigate class imbalance, alongside focal loss optimization to focus learning on difficult examples and reduce bias. The team obtained 80% accuracy score, 79.71% macro precision and 79.21% macro F1 score against FGSM attacks respectively. The results obtained establish their model as a reliable and fair solution for robust multimodal learning in security-critical applications.

General Terms
Multimodal Learning, Adversarial Training, Artificial Intelligence

Keywords
Multimodal Classification, Balanced-FGSM, SMOTE, Focal Loss, Hybrid Fusion, Hate Speech Detection.

1. INTRODUCTION
Social media has become a major communication channel in everyday life and it has drastically changed the way people interact and communicate with each other. Among the social media platforms [1], the most popular ones include Facebook, YouTube, WhatsApp, Instagram, TikTok, WeChat and Twitter. Such platforms are widely used as an important gateway for connecting people, spreading thoughts and linking business entities to customers. The proliferation of multimodal data has greatly motivated extensive research into models capable of jointly understanding and classifying such heterogenous inputs. Multimodal learning [2] has come to leverage the complementary strengths of each modality. For an image-text pair, the image provides a rich spatial and contextual cue, while the text offers semantic and descriptive insights that can be very useful in online hate detection.

Recent advances [3] in deep learning have popularized hybrid fusion architectures, where modality-specific encoders independently extract features before being combined for joint reasoning. In natural language, transformer-based models such as TinyBERT [4] have set new benchmarks by efficiently encoding textual semantics, while lightweight convolutional neural networks like EfficientNet [5] excel in extracting discriminative visual features under resource constraints settings. Despite the success registered by these multimodal classifiers, they still remain vulnerable to two key limitations that impede their deployment in a real-world, high stakes scenarios. The first limitation, adversarial attacks that exploit the intrinsic weaknesses of individual modalities by introducing subtle, often imperceptible perturbations designed to mislead models. The second limitation is the class imbalance, which is a pervasive problem that affects the fairness and generalization of models. Existing real world multimodal datasets like MAMI [6] and FBHM [7] have skewed class distributions, where minority classes are underrepresented. Techniques such as Synthetic Minority Oversampling Technique (SMOTE), class-balanced sampling, and focal loss have shown promising results in unimodal learning but require adaptation and integration for multimodal systems.

In this paper, a novel multimodal classification framework named FusionGuard was designed to comprehensively address the challenges mentioned above. The approach combines the TinyBERT text encoder and EfficientNet image encoder within a hybrid fusion architecture that integrates adversarially aware training to enhance resilience against modality-specific perturbations. It also integrated PCA and SMOTE-based oversampling technique to tackle class imbalance, while using focal loss optimization to improve the models learning on hard examples, which collectively helps to reduce bias. The rest of this study is organized as follows: in section 2, the review related work on multimodal fusion, adversarial robustness, and class imbalance is discussed. Section 3 explains the architecture of FusionGuard and its training methodology. Section 4 describes the implementation and the results obtained. Finally, Section 5 concludes and outlines future directions.

2. LITERATURE REVIEW
Given the growing interest in applying DMMs to real-world tasks, researchers have sought innovative strategies to address the fusion of heterogeneous data sources in socially impactful applications. One such effort is by Hossain et al [8], they proposed a comprehensive framework that utilizes the weighted ensemble technique to assign weights to the participating visual, textual and multimodal models for analyzing memes. They employed the state of the art visual (i.e., VGG19, VGG16, ResNet50) and textual (i.e., BERT, XLM-R, Distill-BERT) models to make the constituent

modules of the framework. Moreover, they used early and late fusion to combine the visual and textual features to develop the multimodal models. Extending beyond the fusion-based framework for meme analysis, Singh et al [9], went on to investigate the detection of hate speech in multimodal data that comprised of text-embedded images using advanced deep learning models. They conducted experiments using four state-of-the-art classifiers: XLM-Roberta-base, BiLSTM, XLNet base cased and AL-BERT, on the CrisisHateMM dataset. they found out that XLM-Roberta-base exhibits superior performance, outperforming the other classifiers across all evaluation metrics, including an impressive F1 score of 84.62%. However, these fusion approaches often overlook challenges related to robustness and fairness.

Adversarial attacks are a big danger to deep learning models because they create input changes that trick classifiers without changing the information that can be seen. While extensive research has explored adversarial robustness in unimodal domains (basically text and image), multimodal robustness remains relatively underexplored. Recent studies reveal that adversarial perturbations can be modality-specific or cross-modal, complicating defense strategies. While most adversarial training methods target unimodal data, Gan et al [10], did the first known effort on large-scale adversarial training for vision-and-language (V+L) representation learning. Instead of adding adversarial perturbations on image pixels and textual tokens, they proposed to perform adversarial training in the embedding space of each modality. In addition to developing robust multimodal models, understanding their vulnerabilities to adversarial attacks is crucial, especially in sensitive applications.

Aggarwal et al [11], conducted a use case study to analyze the vulnerabilities of existing hateful meme detection systems against external adversarial attacks. They proposed nine different attacks to investigate the vulnerabilities of the existing hate meme detection systems to human-induced adversarial attacks and found out that all of these models are highly vulnerable with a drop of as high 10% in macro-F1 performance in certain cases. Ailneni et al., [12] developed a novel approach to automatically uncover both Misogyny Problems (MPs) and their Frames of Misogyny (FoMs) from social media posts, particularly memes. Unlike prior works that rely on pre-annotated categories, their method uses Chain-of-Thought (CoT) prompting with large multimodal models (LMMs) and large language models (LLMs), including GPT-4o, to discover misogyny in a data-driven and minimally supervised fashion. The study demonstrated state-of-the-art performance on the MAMI dataset, surpassing previous benchmarks in both binary and multi-label misogyny classification.

Real world multimodal datasets are often imbalanced, with certain classes widely underrepresented due to either natural distribution biases or data collection constraints. Recent studies [13] employ techniques like Synthetic Minority Oversampling Technique (SMOTE) to mitigate class imbalance and also to improve models' generalization abilities. Also, focal loss has shown success in NLP and multimodal classification [14] by emphasizing harder on minority-class examples.

3. FusionGuard FRAMEWORK

FusionGuard as shown in Figure 1 is a multimodal classification framework designed to address three interrelated challenges commonly observed in image-text datasets namely: adversarial vulnerability and class imbalance. These issues often degrade the performance, generalizability, and fairness of multimodal systems. FusionGuard integrates three tightly coupled modules to mitigate these challenges:

•Hybrid Multimodal Fusion for robust and expressive feature integration,

•Adversarially Aware Training for resilience to perturbations, and

•Bias Mitigation via Resampling and Loss Adaptation to ensure class-level fairness.

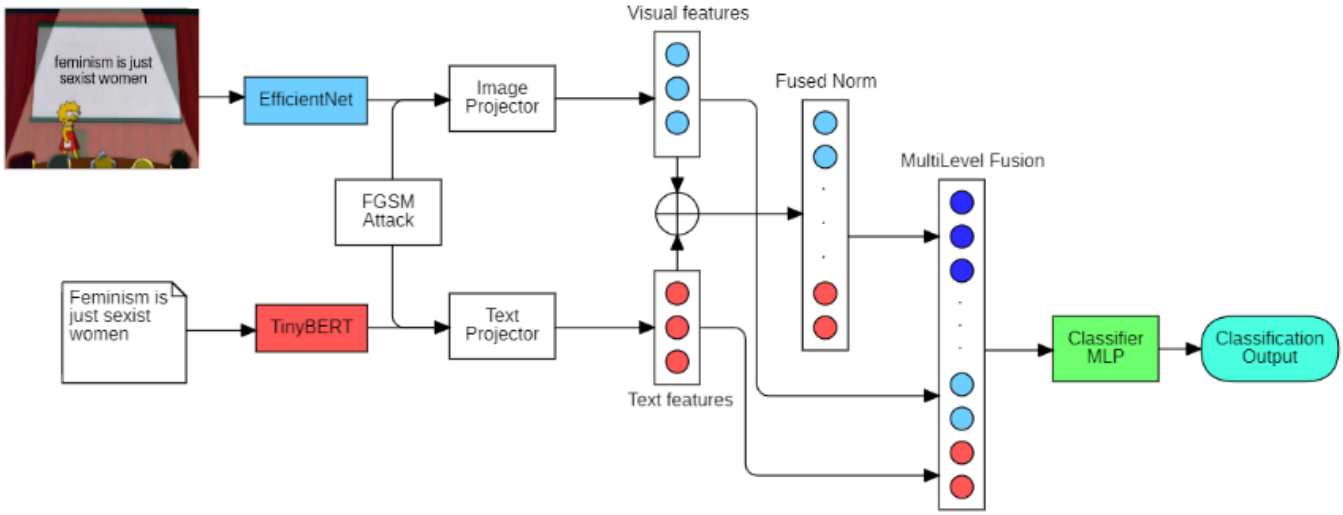

Fig 1: Overview of FusionGuard

The complete system pipeline is organized into the following stages:

3.1 Data Ingestion and Preprocessing

The input to FusionGuard is a dataset$D = \{(x_i^t, x_i^v, y_i)\}_{i=1}^{N}$, where x_i^t and x_i^v represent the textual and visual inputs respectively, and $y_i \in \{1, ..., C\}$ is the corresponding class label, the goal is to train a robust classifier $f: (x^t, x^v) \rightarrow y$ that resists adversarial attacks and performs well across all classes, especially the underrepresented ones.

For image data, it was resized to a fixed resolution of 224x224, normalized it using ImageNet mean and standard deviation, and lastly, it was further augmented using transformations such as random horizontal flipping, color jitter and random rotations to help improve the generalization of the model and also simulate real-world distortions. For the text data, the text was first cleaned using Ekphrasis, which standardizes irregular tokens, expands contractions, and normalizes emoticons, hashtags, and URLs. Moving further, the clean text was tokenized using the pretrained TinyBERT tokenizer that converts the clean input text into token IDs, segment embeddings, and attention masks.

3.2 Feature Extraction

The feature extraction module integrates the semantic representations from textual and visual modalities using pre-trained transformer and CNN backbones. Let $x^{(i)} = \left(v^{(i)}, t^{(i)}\right)$ denote the i-th multimodal sample, where $v^{(i)} \in \mathbb{R}^{H\times W\times 3}$ is the input image and $t^{(i)} \in \mathbb{T}$ is the associated tokenized text. The visual encoder f_v was implemented using EfficientNet-B0 to extract deep spatial features, and the visual input was projected to an embedding $\mathbf{v}_i \in \mathbb{R}^d$:

$$\mathbf{v}_i = W_v \cdot f_v\left(v^{(i)}\right), W_v \in \mathbb{R}^{1280\times d} \quad (1)$$

where d is the common embedding size (e.g., $d = 768$). Simultaneously, the textual features were extracted using a pre-trained TinyBERT encoder f_t, yielding a CLS-token embedding $\mathbf{t}_i' \in \mathbb{R}^{312}$, which was linearly projected into the same space:

$$\mathbf{t}_i = W_t \cdot \mathbf{t}_i', W_t \in \mathbb{R}^{312\times d} \quad (2)$$

The resulting unimodal embeddings $\mathbf{v}_i$ and $\mathbf{t}_i$ are concatenated and passed through a fusion network f_{fuse} to generate the joint representation

3.3 Hybrid Multimodal Fusion

A hybrid fusion approach was adopted to combine textual and visual representations at both feature and decision levels, allowing the model to capture both cross-modal dependencies and individual semantic features.

Let $\mathbf{v}_i, \mathbf{t}_i \in \mathbb{R}^d$ represent the visual and textual embeddings, respectively. These embeddings are initially concatenated and passed through a fusion network:

$$\mathbf{z}_i = f_{\text{fuse}}\left([\mathbf{v}_i \| \mathbf{t}_i]\right), \tilde{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} \quad (3)$$

This fused vector models early-stage interactions across modalities and is L2-normalized to promote training stability. In parallel, modality-specific branches refine the individual embeddings to preserve unique information from each source:

$$\mathbf{v}_i' = f_v^{\text{branch}}(\mathbf{v}_i), \mathbf{t}_i' = f_t^{\text{branch}}(\mathbf{t}_i) \quad (4)$$

The final multimodal representation $\mathbf{h}_i$ is formed by concatenating the refined visual and textual features with the normalized fused vector:

$$\mathbf{h}_i = [\mathbf{v}_i' \| \mathbf{t}_i' \| \tilde{\mathbf{z}}_i] \quad (5)$$

3.4 Class Imbalance Mitigation Module

To address class imbalance, SMOTE was applied within the fused multimodal feature space and Focal loss. Specifically, the complete set of normalized fusion embeddings $\{\tilde{\mathbf{z}}_i\}_{i=1}^N$ along with their associated labels $\{y_i\}_{i=1}^N$ were first extracted. To facilitate interpolation and reduce computational complexity, Principal Component Analysis (PCA) was employed to project the embeddings into a lower-dimensional subspace:

$$\mathbf{z}_i^{\text{PCA}} = P \cdot \tilde{\mathbf{z}}_i, P \in \mathbb{R}^{k\times d}, k < d \quad (6)$$

Within this reduced space, synthetic samples were generated for minority classes using the standard SMOTE procedure, which creates new data points by linearly interpolating between a given sample and one of its randomly selected nearest neighbors from the same class:

$$\mathbf{z}_{\text{new}} = \mathbf{z}_i^{\text{PCA}} + \lambda \cdot \left(\mathbf{z}_{\text{nn}}^{\text{PCA}} - \mathbf{z}_i^{\text{PCA}}\right), \lambda \sim \mathcal{U}(0,1) \quad (7)$$

Here, $\mathbf{z}_{\text{nn}}^{\text{PCA}}$ denotes a nearest neighbor of $\mathbf{z}_i^{\text{PCA}}$ sharing the same class label.

Following the synthetic data generation, the interpolated features were projected back into the original feature space via the inverse PCA transformation:

$$\tilde{\mathbf{z}}_{\text{new}} = P^{\top} \cdot \mathbf{z}_{\text{new}} \quad (8)$$

This augmentation strategy allowed for the expansion of underrepresented classes in a manner that is both semantically consistent and modality-aware, ensuring that the synthesized features retained the essential fused characteristics required for effective supervised learning.

Also, a class-balanced variant of the Focal Loss that adjusts gradient contribution was used based on class frequencies and sample difficulty. Given model predictions $\hat{y}_i \in \mathbb{R}^C$ and true label $y_i \in \{0, \ldots, C-1\}$, the Focal Loss is defined as:

$$\mathcal{L}_{\text{focal}} = -\alpha_{y_i}\left(1 - p_{y_i}\right)^{\gamma} \log\left(p_{y_i}\right) \quad (9)$$

where $p_{y_i} = \text{softmax}(\hat{y}_i)[y_i]$ is the predicted probability for the true class, $\gamma > 0$ is the focusing parameter that down-weights easy examples, and $\alpha_{y_i} \in [0,1]$ is the class-balancing weight.

3.5 Adversarially Aware Training

Adversarial training was incorporated in FusionGuard through multimodal perturbations in order to improve its robustness. The Balanced Fast Gradient Sign Method (Balanced-FGSM) applied in this paper is an extension of the classical FGSM adversarial attack, that is suitable for multimodal models, particularly those employing hybrid fusion strategies. In a standard FGSM attack, adversarial examples are generated by adding perturbations in the direction of the gradient loss with respect to the input, scaled by a perturbation magnitude ϵ:

$$x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y)) \quad (10)$$

However, in multimodal models with heterogeneous inputs such as images and text, applying a uniform perturbation magnitude may lead to disproportionate perturbations. This imbalance can either obscure the contribution of a modality with low gradient sensitivity or over distort a dominant one. To mitigate this, Balanced-FGSM was applied to distribute the perturbation budget proportionally to the influence of each modality on the model's output. In the Balanced-FGSM framework, given a loss function $\mathcal{L}$ and model parameters $\boldsymbol{\theta}$,

the gradients of the loss with respect to the image and text inputs are computed as:

$g_{\text{img}} = \nabla_{x_{\text{img}}} \mathcal{L}$ and $g_{\text{text}} = \nabla_{x_{\text{text}}} \mathcal{L}$, respectively. The total gradient magnitude is estimated using the ℓ_1-norm:

$$G = \|g_{\text{img}}\|_1 + \|g_{\text{text}}\|_1 \quad (11)$$

Each modality is then assigned a share of the total perturbation budget based on its normalized gradient magnitude:

$$\epsilon_{\text{img}} = \epsilon \cdot \frac{\|g_{\text{img}}\|_1}{G}, \epsilon_{\text{text}} = \epsilon \cdot \frac{\|g_{\text{text}}\|_1}{G} \quad (12)$$

The adversarial examples for both modalities are generated as follows:

$$x_{\text{img}}^{\text{adv}} = x_{\text{img}} + \epsilon_{\text{img}} \cdot \text{sign}(g_{\text{img}}),\ x_{\text{text}}^{\text{adv}} = x_{\text{text}} + \epsilon_{\text{text}} \cdot \text{sign}(g_{\text{text}}) \quad (13)$$

3.6 Training Procedure

As shown in Figure 2, the method starts with loading the raw MAMI dataset which was obtained by merging the train and val csv files, then a stratified splitting procedure was applied to ensure class balance in the validation set by typically selecting 20 samples per class for the validation set while allocating the remainder for training. Each sample was then passed through a hybrid backbone composed of TinyBERT for textual embedding and EfficientNet for visual features, whose outputs are merged via a learned fusion module. The fused representations are then extracted and passed through PCA to reduce its dimensionality, thereby it helps to optimize the subsequent application of SMOTE in the shared embedding space.

A Balanced FGSM step was also included to improve the model's robustness against adversarial perturbations. This simply involves computing the gradients of the loss function with respect to both text and image inputs and distributing the perturbation magnitude proportionally to each modality based on their respective gradient norms. If adversarial training mode is enabled, the perturbed inputs are reintroduced into the SMOTE-balanced dataloader to ensure that the training process accounts for both synthetic diversity and adversarial resilience. The training was conducted using the AdamW optimizer, with cosine learning rate scheduling and mixed-precision support to improve computational efficiency. Upon completion, the best performing model is saved based on the macro F1 score.

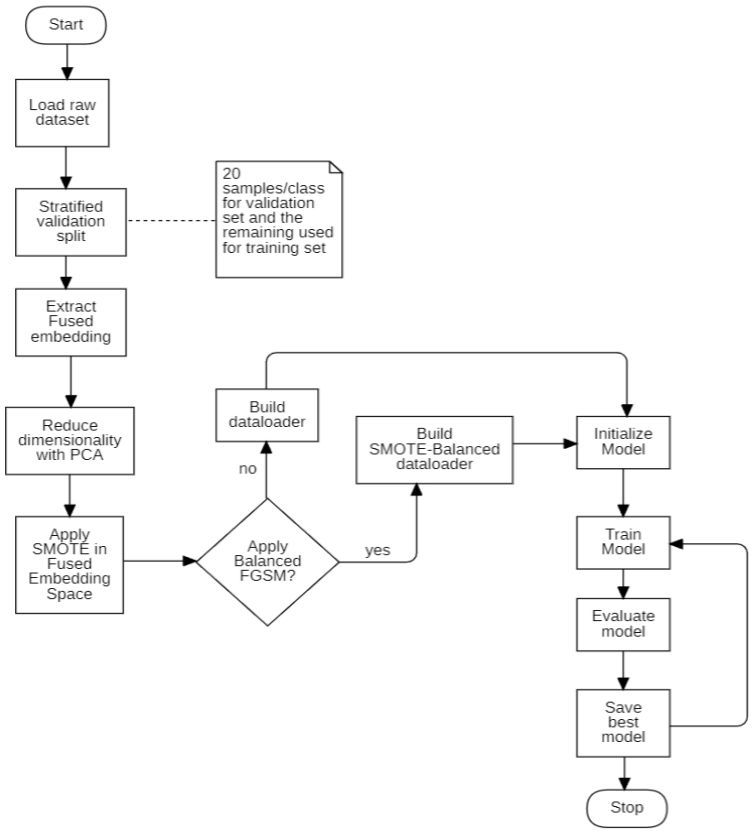

Fig 2: Training Pipeline

4. IMPLEMENTATION AND RESULTS

4.1 Dataset

In this study, the Multimedia Automatic Misogyny Identification (MAMI) dataset used was provided by Fersini et al., [6] as part of the SemEval 2022. The dataset was obtained upon completion of an online form available at https://forms.gle/AGWMiGicBHiQx4q98. The dataset contains 10,000 memes for training and 1000 memes for testing that was collected by targeting the following key categories of misogynistic content as investigated by Rizzi et al., [15] namely, other misogynous, shaming, stereotype, objectification, violence. One noticeable thing was that the dataset exhibits a significant imbalance with a heavy skew towards non-hateful classes as shown in Figure 3. The training and validation set were joined together to create a new

validation set with 20 examples set per class and the rest used for training.

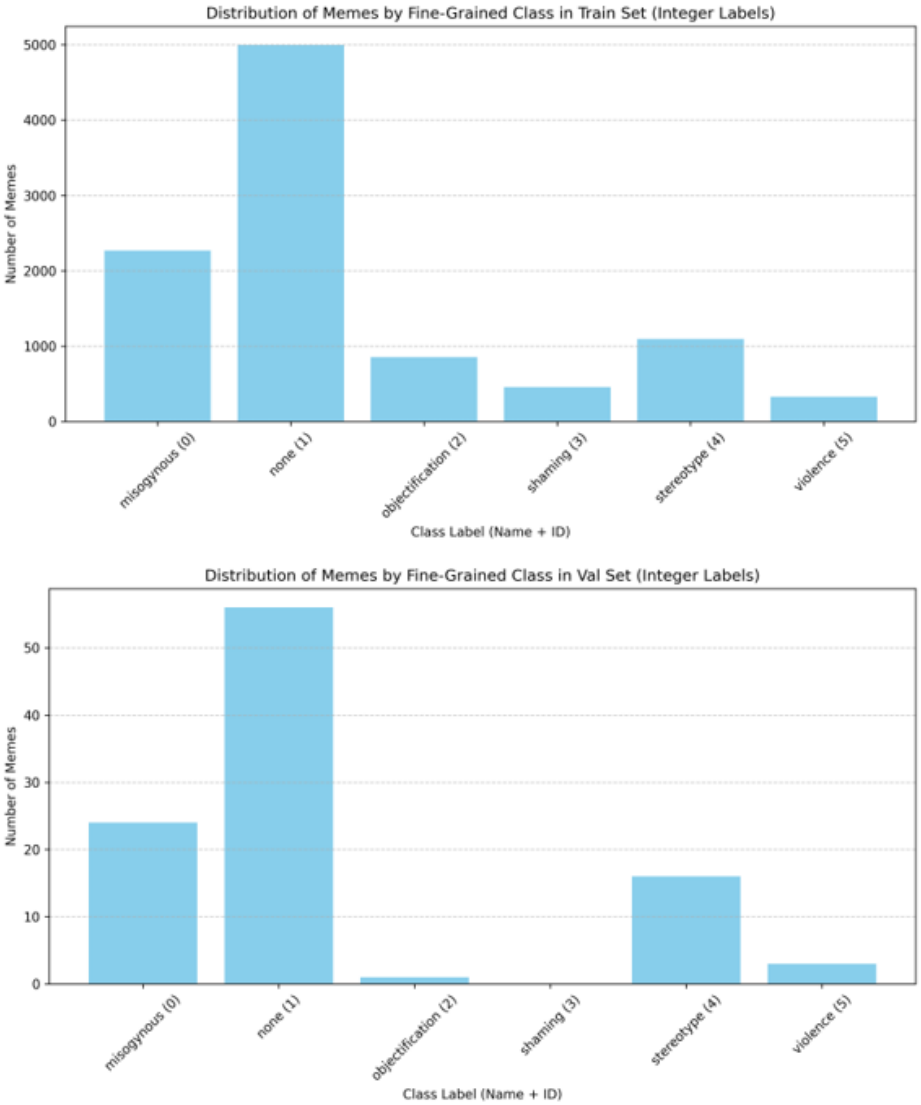

Fig 3: Distribution of Memes by Fined-Grained Class for both Training and Validation set.

4.2 EXPERIMENTAL SETUP

Extensive experiments were conducted on a custom-built MAMI dataset and the evaluation was performed for both clean and adversarial settings. The method adopts a hybrid fusion approach that merges the strength of early and late fusion to effectively leverage the complementary information present in both visual and textual modalities. SMOTE was applied to the textual and visual embeddings independently to balance the class distribution, and Balanced-FGSM was applied to the multimodal data to improve robustness. Training of the model was performed using the AdamW optimizer with a learning rate of 1e-5, weight decay of 0.001, and a linear warm-up over the first 7% of training steps. FusionGuard was trained for up to 25 epochs using automatic mixed precision (AMP) and apply early stopping based on validation macro F1 score. To address label imbalance, the Focal Loss function was used and the text inputs were tokenized using the TinyBERT tokenizer with a maximum sequence length of 128 tokens. The batch size is 8, and all experiments are implemented in PyTorch 2.0 with Python 3.11, running on a system equipped with an NVIDIA RTX A2000 GPU and 32GB RAM under Windows 10. The best-performing model is selected based on macro-averaged F1 score and accuracy on the validation set.

4.3 CLASSIFICATION RESULTS

In this section, to assess the effectiveness of FusionGuard, a series of experiments was conducted on the MAMI dataset. The model was trained using a stratified validation split (20 samples per class) and evaluated based on macro-averaged precision, F1-Score and overall accuracy. The best performance by FusionGuard selected via early stopping based on validation macro-f1 score is a macro F1 score of 79.12%, a macro precision of 79.91% and an overall accuracy of 80.00% on the validation set as shown in Figure 4. These results highlight the model's capability to generalize across all classes, including minority ones which is crucial.

In Table 1, the researchers compared the model FusionGuard against other works on memes of the MAMI validation data on multilabel classification task.

Table 1. Multilabel classification performance comparison

with existing works using a 10-fold cross-validation on the memes of the MAMI validation data

Author	Model	Macro- F1 Scores(%)	Accuracy (%)	Precision (%)
Zhang et al, [16]	SRCB	73.1	N/A	N/A
Ailneni et al, [12]	Dis-MP&F	74.5	N/A	N/A
FusionGuard	Distill BERT-Efficie ntNet	**79.12**	**80.00**	**79.91**

In Figure 4, the label Hybrid reports the performance of the model on the stratified MAMI dataset. On the other hand, the label Hybrid+SMOTE reports the performance of the model after balancing the train set. This result leads to better results, confirming the effectiveness of embedding level oversampling. Finally, the label FusionGuard, integrates the Focal Loss to focus the learning process on hard and minority samples results in the best overall performance, particularly improving the macro F1-score and reducing misclassification in underrepresented classes and also adversarial training against FGSM attacks.

This ablation study reveals the incremental contribution of each component: the based hybrid architecture achieved 67.25% macro F1, adding SMOTE oversampling improved performance to 80.12%, and integrating Focal Loss with adversarial training yielded the final 79.91%.

For qualitative examples obtained from these experiments. The results of the model on the test set are presented in figure 5, these evaluation on the test set further demonstrated the robustness of FusionGuard, with a confidence mean of 71.75 % ± 19.08%, an average confidence drop of -2.01%, and a robustness rate of 93.20%, indicating a reliable prediction behavior. For attacks on the test set, only images were perturbed, not text tokens since they are discrete and don't support gradient based perturbation.

5. CONCLUSION AND FUTURE WORKS

In this paper, FusionGuard was developed, which is a robust multimodal classification framework that leverages the complementary strengths of vision and language representations through a hybrid fusion architecture that combines TinyBERT for textual embeddings and EfficientNet for visual features. Class imbalance was addressed and equally generalization of the model was enhanced by applying a SMOTE-based oversampling strategy in the fused embedding space. Also, a Balanced Fast Gradient Sign Method was incorporated to build the robustness of the model, Balanced FGSM distributed the perturbation magnitude across modalities based on gradient sensitivity. FusionGuard was trained using a stratified split strategy and evaluated with early stopping and macro F1 based model selection. The best model achieved 80% accuracy score, 79.71% macro precision and 79.21% macro F1 score against FGSM attacks respectively in adversarial attack settings.

While FusionGuard showed a strong generalization and a good resilience to class imbalance, several avenues remain open for future research. First, a look up to the integration of an adaptive adversarial training setup should be made, where the perturbation budget is learned dynamically. This may further improve the robustness of the model without manually tuning the perturbation factor. Secondly, it will be an interesting thing to expand the fusion mechanism to include cross-modal attention or co-attention layers that could further enhance inter-modality interaction and improve fine-grained semantic alignment. Also, performing cross dataset validation on other Misogynistic dataset to show the models consistent performance is needed. This will also help to confirm the generalizability beyond MAMI. Finally, it will be of good intention to extend the scope of these attacks and train the model on stronger attacks like PGD, C&W.

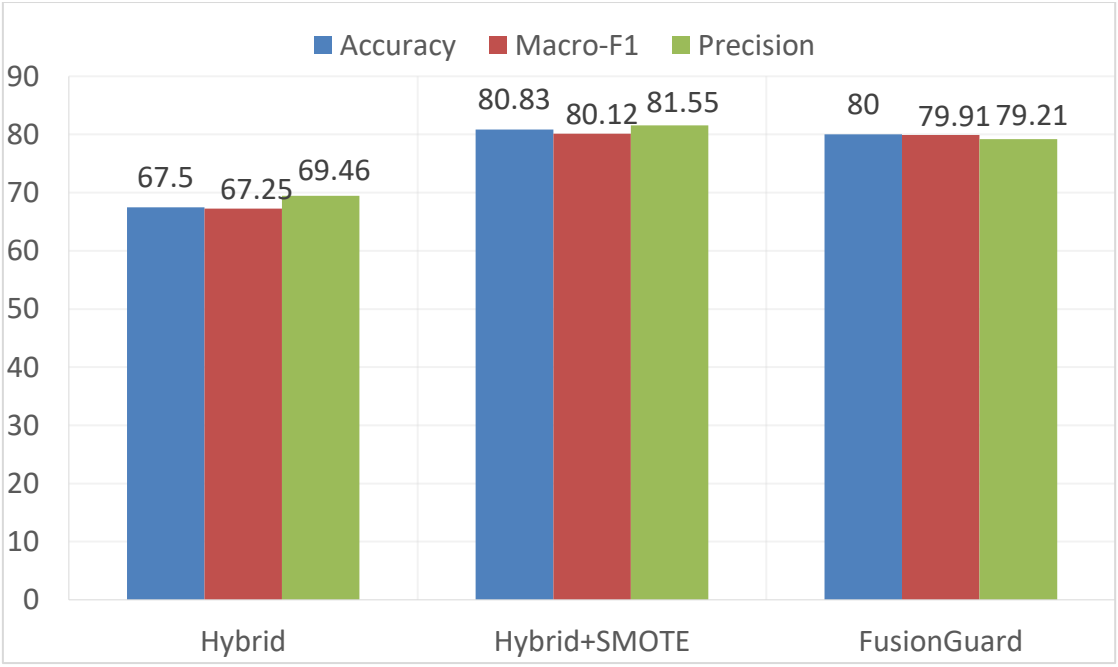

Fig 4: Results of Ablation study

Fig 5: Some examples of predicted classes after adversarial attacks

6. REFERENCES

[1] I. Alsmadi, K. Ahmad, M. Nazzal, F. Alam, A.-F. Ala, A. Khreishah and A. Algosaibi, "Adversarial NLP for Social Network Applications: Attacks, Defenses and Research Directions," IEEE Transactions on Computational Social Systems, 2022.

[2] M. Suzuki and Y. Matsuo, "A Survey of Multimodal Deep Generative Models," Journal of Advanced Robotics, Taylor&FrancisOnline, 2022.

[3] U. R. Mohammad Zia, Z. Sufyaan, M. Areeb, M. Musharaf and K. Nagendra, "A context-aware attention and graph neural network-based multimodal framework for misogyny detection," Information Processing and Management, Elsevier, 2025.

[4] J. Xiaoqi, Y. Yichun, S. Lifeng, J. Xin, C. Xiao, L. Linlin, W. Fang and L. Qun, "TinyBERT: Distilling BERT for Natural Language Understanding," EMNLP, 2020.

[5] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proceedings of the 36th International Conference on Machine Learning, Long Beach, 2019.

[6] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi and P. Rosso, "SemEval-2022 Task 5: Multimedia automatic misogyny identification," in 16th International workshop on semantic evaluation. Association for Computational Linguistics, 2022.

[7] D. Kiela, A. Mohan, H. Firooz, V. Goswani, A. Singh, P. Ringshia and D. Testuggine, "Hateful Memes Challenge and dataset for research on harmful multimodal content," in 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada., 2020.

[8] E. Hossain, O. Sharif, M. M. Hoque, M. A. A. Dewan, N. Siddique and M. A. Hossain, "Identification of Multilingual Offense and Troll from Social Media Memes Using Weighted Ensemble of Multimodal Features," Journal of King Saud University-Computer and Information Sciences, Elsevier, pp. 6605-6623, 2022.

[9] K. Singh, V. Vajrobol and N. Aggarwal, "Multimodal Hate Speech Event Detection 2023: Detection of Hate Speech and Targets using Xlm-Roberta-base," in Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, Varnia, Bulgaria, 2023.

[10] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng and J. Liu, "Large-Scale Adversarial Training for Vision-and-Language Representation Learning," in 34th Conference on Neural Information Processing Systems , Vancouver, 2020.

[11] P. Aggarwal, M. D. P. C., S. Punyajoy, M. Binny, Z. Torsten and M. Animesh, "HateProof: Are Hateful Meme Detection Systems really Robust?," in Proceedings of the ACM Web Conference 2023, 2023.

[12] A. Rakshitha Rao and H. Sanda M, "Automatically Discovering How Misogyny is Framed on Social Media," in Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, 2025.

[13] T. T. Emmanuel Ludivin and F. T. Elie, "Effective Multimodal Hate Speech Detection on Facebook Hate Memes Dataset using Incremental PCA, SMOTE, and

Adversarial Learning," Machine Learning with Applications , 2025.

[14] Y. Chen, W. Zhang, H. Zhang, D. Qu and X.-K. Yang, "Task-based Meta Focal Loss for Multilingual Low-resource Speech Recognition," ACM Trans. Asian Low-Resour. Lang. Inf. Process, 2023.

[15] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso and E. Fersini, "Recognizing misogynous memes: Biased models and tricky archetypes," Journal of Information Processing and Management, 2023.

[16] J. Zhang and Y. Wang, "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), Seattle, United States, 2022.