# Poirot at SemEval-2022 Task 5: Leveraging Graph Network for Misogynistic Meme Detection

**Harshvardhan Srivastava**
Oracle India
Indian Institute of Technology, Kharagpur
harshvardhan.srivastava@oracle.com

## Abstract

In recent years, there has been an upsurge in a new form of entertainment medium called memes. These memes although seemingly innocuous have transcended the boundary of online harassment against women and created an unwanted bias against them. To help alleviate this problem, we propose an early fusion model for the prediction and identification of misogynistic memes and their type in this paper for which we participated in SemEval-2022 Task 5. The model receives as input meme image with its text transcription with a target vector. Given that a key challenge with this task is the combination of different modalities to predict misogyny, our model relies on pre-trained contextual representations from different state-of-the-art transformer-based language models and pre-trained image pre-trained models to get an effective image representation. Our model achieved competitive results on both SubTask-A and SubTask-B with the other competing teams and significantly outperforms the baselines.

## 1 Introduction

Meme culture in today's virtual climate gives us a variety of insight into the pop culture, general ideology and linguistic conversational manner of the generation. To understand the internet culture, it becomes essential to study memes (Shifman, 2013) and the impact it has on people on the internet. Some of the most popular communication tools on social media platforms are memes. Memes are essentially images characterized by the content of a picture overlaid with text that was introduced by people with the main purpose of being interesting and ironic. Women have a strong presence online, especially on image-based social media like Twitter, Snapchat, and Instagram. 78% of women use social media several times a day, compared to 65% of men (Fersini et al., 2022). While new opportunities are being opened up for women online, systematic inequality and discrimination are being replicated offline from these online spaces in the form of offensive content for women. Most of them were created to make funny jokes, but soon people began to use them as a form of hatred for women, leading to sexist and offensive messages in the online environment, and as a consequence, the sexual stereotyping and gender inequality of the offline world where sexuality stereotypes and gender inequality have been strengthened. This insensitive and obscene type of meme has a profound effect on a person's mental health and can exhibit harmful effects on cognitive and emotional processes leading to mental illnesses as shown in Paciello et al. (2021).

In this work, we present team Poirot's solution to SemEval - 2022 Task 5 competition as described in detail in Fersini et al. (2022). We focused our efforts on our primary approach of building a Multi-Modal module that uses features from both images and text. Furthermore, in this paper, we provide ablation studies on different modalities, relative importance of the different modalities and some training parameters, and show how by changing the module parameters, the predictions on the misogynistic identification of memes aggravates or allays.

## 2 Background

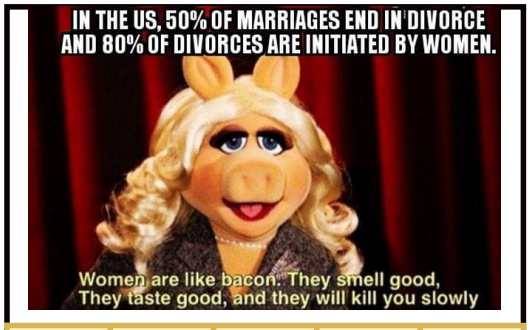
### 2.1 Task Description

The organisers have provided us with data tasked with the identification of misogynous memes, taking advantage of both text and images available as source of information. The task is comprised around two main sub-tasks:

- Sub-Task A: first task about misogynous meme identification, where a meme is categorized in a binary format; either as misogynous or not misogynous;
- Sub-Task B: second task, where the type of misogyny is recognized among potential over-



*Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 793 - 801




| misogynous | shaming | stereotype | objectification | violence |
|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 |

Figure 1: Misogyny in meme

lapping categories such as stereotype, shaming, objectification and violence as described in (Fersini et al., 2022).

The sub-tasks are arranged in an increasing range of difficulty. The competition is challenging, as identifying the misogynous nature of a meme is more complex in a multi-modal setting than performing the same task only on textual data. For memes, comprised of image and text information, a multi-modal approach for understanding both visual and textual cues is needed. Also, in sub-task B, the nature of problem difficulty is increased as the type of misogyny has to be identified, which can belong to multiple categories due to the nature of the dataset.

### 2.2 Dataset

The datasets for the competition provided by the task organisers are memes collected from the web and manually annotated via crowdsourcing platforms (Fersini et al., 2022). Each sample is supported by an image and the corresponding text transcription (if it exists) on the image. An example of the sample is given in Figure 1. The statistical information about the datasets can be found in Table 1.

Additionally, we provide a quick look into the training dataset which has a significant data imbalance for *4* of labels in a number of samples belonging to each of the *4* given labels except the label "misogynous". This imbalance affects the performance of the models on the test set specially in the case of multilabel prediction as not equal training instances are available for each of the classes. The information about the number of samples belonging to each of the 5 classes for the training set is given in Table 1. This dataset imbalance is dealt with in section 3.3.

### 2.3 Evaluation Criteria

The teams' performance is evaluated by the macro F1 score for task A. For tasks B, the weighted F1 score is computed for each subtask (misogynous, shaming, stereotype, objectification, violence)(Fersini et al., 2022), and the average F1 score of these subtasks is used to rank the systems.

### 2.4 Related Work

Earlier, some other meme datasets have been created like the dataset created in Oriol et al. (2019) with the intention of automatically detecting hate speech, and the hateful memes dataset by Facebook (Kiela et al., 2020), which created a challenge set for multimodal classification of hatred in memes. Previous work encompassing categories like hate speech, sexism, and toxicity detection in memes has primarily been explored from a textual perspective using Natural Language Processing(NLP). However, recent methods are aiming to use multimodal approaches to solve the issue at hand. VL-BERT(Su et al., 2020) used the single-stream architecture, where a single Transformer is applied to both images and text. ViLBERT(Lu et al., 2019) and LXMERT(Tan and Bansal, 2019) introduced a two-stream architecture where two transformers are applied to images and text independently and later merged by a third transformer. ERNIE-ViL(Yu et al., 2021) incorporates structured knowledge obtained from scene graphs to learn joint representations of vision-language. Zia et al. (2021) presents the multimodal pipeline based on pre-trained visual and textual representations for the shared task involving the detection of hateful memes.

| Set | Number of Samples |
|---|---|
| `Trial` | 100 |
| `Train` | 10000 |
| `Test` | 1000 |

| Label | Positive Samples |
|---|---|
| *misogynous* | 5000 |
| *shaming* | 1274 |
| *stereotype* | 2810 |
| *objectification* | 2202 |
| *violence* | 953 |

Table 1: Dataset and Labels Information

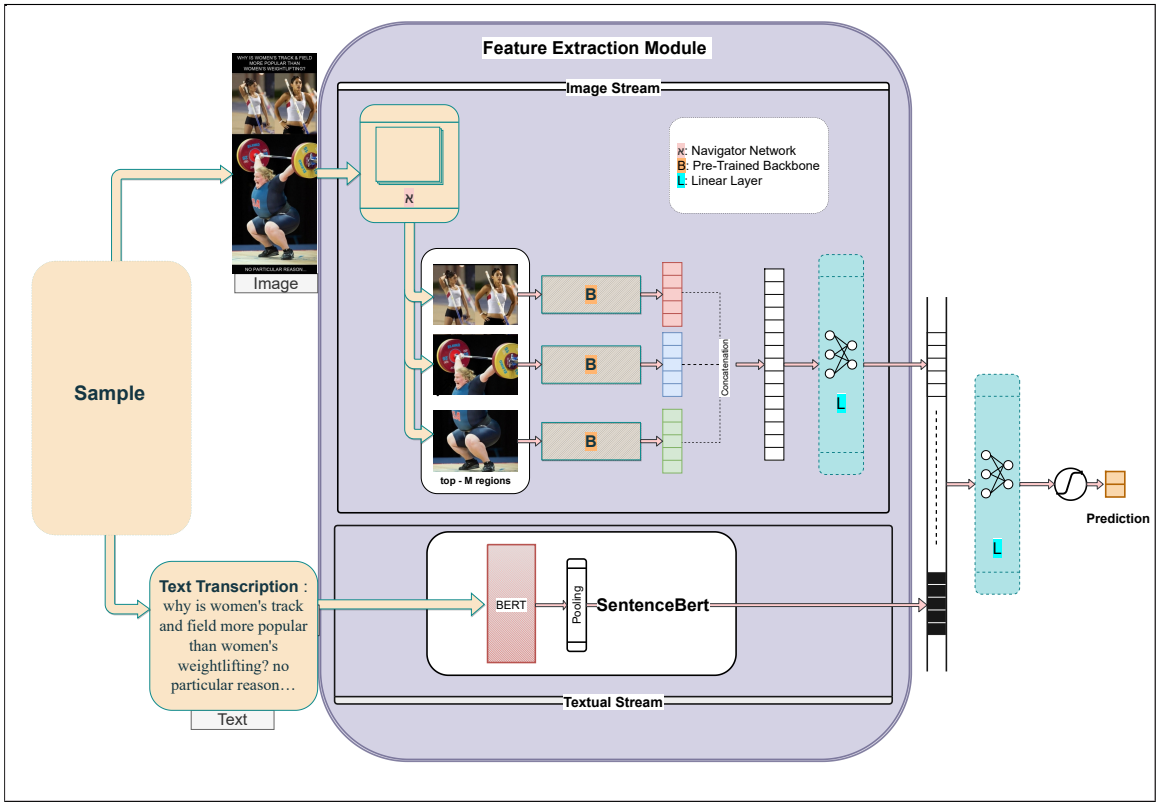



Figure 2: Overview of the binary model used for misogyny detection. The two modalities are passed simultaneously through the Feature Extraction Module in two separate paths and trained together and finally fused together and passed through a linear layer to get binary logits.

## 3 System Overview

The solution system comprises of 2 separate systems for both the sub-tasks. The approach can be broadly divided into binary approach and multi-label approach.

### 3.1 Binary Model

Broadly, the model consists of two modal information streams, text and image. The proposed approach leverages on multi-modal information to provide the classification of a sample. We exploit text transcriptions written in natural language jointly with visual information coming from the meme image. In the initial stage the pipeline is divided into two streams running in parallel which on later stage is joint together. The outline of the proposed architecture is shown in Figure 2. The sub-modules are described below:

1. **Image Features Extraction** : This stream is also separated into two sub-modules :

   - *pre-trained Representation Module* : We can use any backbone CNN base models to learn the features of an image. For our experimentation, we use ResNet-101 and ResNet-50 (He et al., 2016) as the backbone *B* model. The rationale behind choosing two separate backbone models is to compare generalised image representation as compared to domain-specific image representation in this case where outright misogyny in images can be detected quickly by looking at nsfw content in the image part of the meme. Thus if the input image $I \in \mathbb{R}^{d \times d}$, where d = 224, the output of the feature extractor would give us intermediate level features $\kappa \in \mathbb{R}^{D}$, where D = 2048,

   $$\kappa = B(I)$$

   - *Navigator Module* : We use Navigator Module from the NTS-net as described in (Yang et al., 2018) model to decouple the image into several parts. For an input image, the image is fed into the Navigator network to compute the informativeness of all regions. It is fed to the navigator network, which extracts meaningful

parts to separate *top-M* regions. The feature extractor extracts its deep feature map for each of those parts. These features are then concatenated $C$ together:

$$\kappa_i = \aleph(I), i \in [0, M-1]$$

$$f_v = C(\kappa_0, \kappa_1, \kappa_2, ...., \kappa_{M-1}) \in \mathbb{R}^{M \times D}$$

2. **Text Features Extraction:** : The second modality being the textual Stream, uses the SentenceBERT model (Reimers and Gurevych, 2019). SBert modifies the BERT network using a combination of siamese and triplet networks to derive semantically meaningful embedding of sentences. As a state of the art language model, BERT has greatly influenced results in the text classification task as shown in Minaee et al. (2021), we use SentenceBert $S$ model trained on Siamese BERT networks. Thus we convert the given text $T$ transcription into features vector $f_t$. Formally :

$$f_t = S(T) \in \mathbb{R}^E$$

where E = 768.

The image extraction part and text extraction part is clubbed together to form the **Feature Extraction Module**. If $C_{mc}$ is multimodal feature concatenation logic, then,

$$F = C_{mc}(f_v, f_t)$$

This module outputs a feature vector of size $N = E + D$ features. These features are then passed through a $f$(linear layer) to output logits which are then passed on to $\sigma$ function to generate predictions.

$$y_{pred} = \sigma(f(F))$$

### 3.2 Multi-Label Model

The multi-label model, keeps the feature extracting pipeline of the network in the binary model intact while changing the final output method by using Graph Neural Networks. The model consists of two essential parts : (i) Feature Extraction Module and (ii) Graph classification module. The overall architecture of our model is explained in the Fig. 3.

1. **Feature Extraction Module** : Same as in binary model 3.1

2. **Graph Classification Module** : A graph has an effective message passing system, which can be modelled to find the inter-dependency of the labels amongst each other, and hence, efficiently capture the semantic importance of a label $u_1$ depending on co-occurring label $u_2$. We represent each node of the graph input to be a label, having the node features as GloVe embedding having $e$ features. Formally, we use Graph Network to learn the multi-label classification model to learn label representation:

$$L_{n+1} = \phi(L_n, A)$$

where $L_n \in \mathbb{R}^{u \times e}$ represents class label representation at *nth* graph layer, $\phi$ represents the message passing network and $A \in \mathbb{R}^{u \times u}$ represents the adjacency matrix. Through stacking multiple Graph Network Layers, we model the complex inter-relationships among classes.

**Creation of Adjacency Matrix** : We calculate the label adjacency matrix $\mathbb{A}$ by mining label co-occurrence patterns in the training and trial dataset. Let the label matrix $L_m \in \mathbb{R}^{n_s \times u}$, where $n_s$ are the number of training and trial samples. Then the co-occurrence matrix

$$A_{coo} = L_m^T \times L_m \in \mathbb{R}^{u \times u}$$

To create the adjacency matrix from the co-occurrence matrix and to remove the self node loop from the graph, we create a vector $N_u \in \mathbb{R}^u$, having

$$N_u[i] = A_{coo}[i][i]$$

Finally, the adjacency matrix $\mathbb{A}$ can be constructed as:

$$A_{ij} = \begin{cases} 0 & \text{if } i = j, \\ \frac{A_{coo}[i][j]}{N_u[i]} & \text{otherwise} \end{cases}$$

### 3.3 Multi-Label Classification Loss

We notice the imbalance present in data for different classes which we can see in Table 1, but the extent of imbalance is different for different labels.

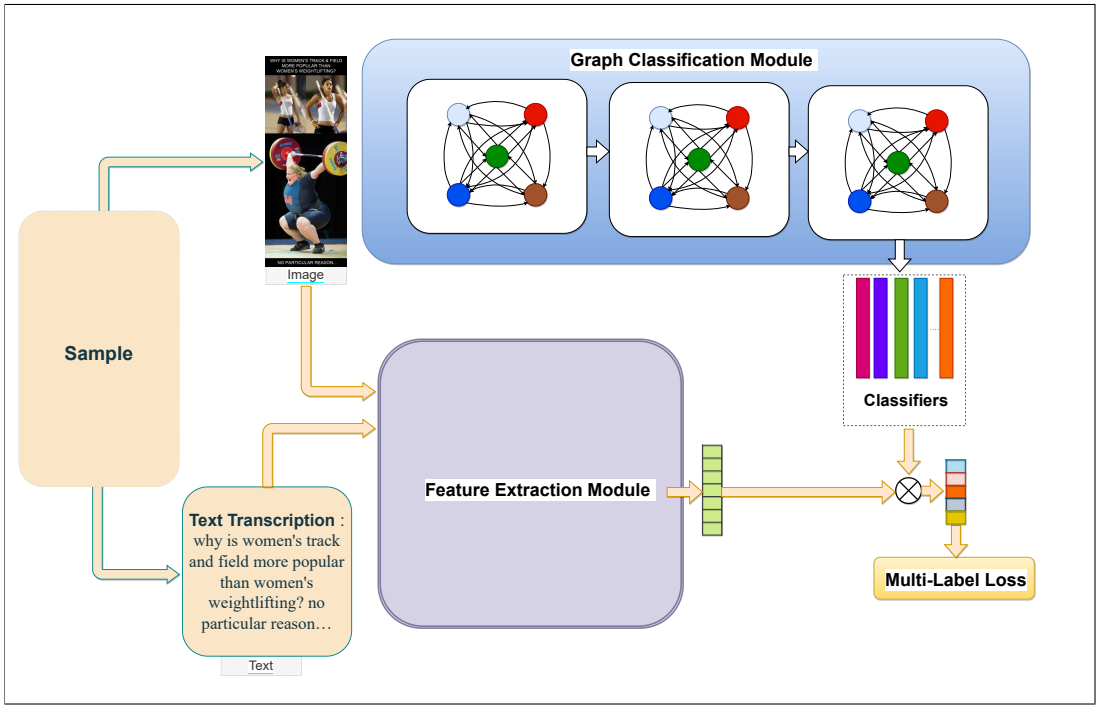



Figure 3: : The general architecture of our Multi-Modal-Multi-Label model. We use image features and text features from the *Feature Extraction* (FE) which has a pre-trained ResNet, pre-trained sentence transformer SBert. The features are passed through a 5-layer classifier stack generated from the Graph Classification Module which takes input the label's semantic information to generate the output multi-label prediction.

This knowledge can be passed to the neural network in terms of class weights in order to penalize adequately. Let $n_s$ be the number of samples in the dataset. We calculate the weighted importance of a class using the below equations:

$$W_p[i] = \frac{N_p[i]}{n_s}, W_n[i] = \frac{N_n[i]}{n_s}$$

where $N_p[i]$, is the number of positive samples for class *i* and $N_n[i]$, is the number of negative samples for class *i*.

$$-dl = W_p * y * \log(p) + W_n * (1-y) * \log(1-p)$$

where y is the ground truth and p is the predicted output. The calculated weights are shown in table 2.

| | Weights | |
|---|---|---|
| **Label** | **Positive** ($W_p$) | **Negative** ($W_n$) |
| misogynous | 1.000 | 1.000 |
| shaming | 3.924 | 0.573 |
| stereotype | 1.779 | 0.695 |
| objectification | 2.270 | 0.641 |
| violence | 5.246 | 0.552 |

Table 2: Calculated weights for regularizing cross-entropy loss in the custom loss function

## 4 Experimental Setup

### 4.1 Baselines

We use the baseline provided by the task organisers [1] which depend on the Sub-Task and use a different set of features for different tasks:

1. Sub-Task A: Misogynous Meme Identification
   - ***Baseline-Text***: deep representation of text, i.e. a fine-tuned sentence embedding using the USE(Cer et al., 2018) pre-trained model
   - ***Baseline-Image***: deep representation of image content, i.e. based on a fine-tuned image classification model grounded on VGG-16(Liu and Deng, 2015).
   - ***Baseline-IT***: concatenation of deep image and text representations, i.e. based a single layer neural network.
2. Sub-Task B : Type of Misogynous Meme Identification

[1] https://github.com/MIND-Lab/MAMI

- ***Baseline-Flat***: a multi-label model, based on the concatenation of deep image and text representations, for predicting simultaneously if a meme is misogynous and the corresponding type
- ***Baseline-Hierarchical***: a hierarchical multi-label model, based on text representations, for predicting if a meme is misogynous or not and, if misogynous, the corresponding type.

### 4.2 Hyperparameters and Implementation Details

Before passing the text transcription to the text stream, we apply some basic text preprocessing to all our sentences. First, we normalize all the sentences by converting all white-space characters to spaces. Also, in the image stream, before passing the image to the navigator network, we resize the image to size [224,224] for uniformity and perform random crops and flips before feeding it to the network. When concatenating the image features with textual features, we use a parameter $\lambda$ to combine the two together. The final feature vector is formulated as following,

$$F = [(1 - \lambda) \times f_v; \lambda \times f_t]$$

We adopt a 2-layer graph network for our best performing system. For node features, we use the 300-Dimensional GloVe embeddings (Pennington et al., 2014) trained on the Wikipedia Dataset. Table 3 contains the list of general hyperparameters we used. We implement the network based on PyTorch.

| Parameter Name | Value |
|---|---|
| Optimizer | AdamW |
| Pre-Trained BERT LR | 2e-4 |
| Navigator Network LR | 1e-3 |
| Graph Learning Rate | 1e-2 |
| Graph Layer 1 Dim | 512 |
| Graph Layer 2 Dim | 2048 |
| $\lambda$ (Concatenation Parameter) | 0.7 |

Table 3: Major hyperparameters used

## 5 Results

Table 5 and 6 compares the macro and weighted $f_1$ scores of our best performing models on the binary classification task and the multi-label task respectively alongside the score achieved by the baseline models. We also present several ablations for the best performing models on $\lambda$ parameter and its effect on the final score achieved by the model. The $\lambda$ ablations can be found in Table 4.

| | | **Binary** | **Multi-Label** |
|---|---|---|---|
| **Backbone Model** | $\lambda$ | $f_1^{macro}$ | $f_1^{weighted}$ |
| ResNet-101$_{imagenet}$ | 0.1 | 0.601 | 0.590 |
| | 0.2 | 0.619 | 0.597 |
| | 0.3 | 0.645 | 0.622 |
| | 0.4 | 0.689 | 0.628 |
| | 0.5 | 0.702 | 0.641 |
| | 0.6 | 0.736 | **0.645** |
| | 0.7 | **0.741** | 0.643 |
| | 0.8 | 0.728 | 0.631 |
| | 0.9 | 0.702 | 0.612 |
| ResNet-50$_{nsfw}$ | 0.1 | 0.611 | 0.591 |
| | 0.2 | 0.620 | 0.595 |
| | 0.3 | 0.691 | 0.612 |
| | 0.4 | 0.703 | 0.632 |
| | 0.5 | 0.736 | 0.634 |
| | 0.6 | 0.749 | **0.635** |
| | 0.7 | **0.759** | 0.632 |
| | 0.8 | 0.734 | 0.623 |
| | 0.9 | 0.698 | 0.601 |

Table 4: $\lambda$ effect on model performance

ResNet-101$_{imagenet}$ uses the backbone *B* pretrained on the ImageNet dataset, while ResNet-50$_{nsfw}$[2] model uses the backbone fine-tuned on around 40GB of *nsfw* data. We divide our model for multi-label classification according to different types of loss used during the training stage.

| **SubTask-A** | |
|---|---|
| Model | $f_1^{macro}$ |
| Baseline-Text | 0.640 |
| Baseline-Image | 0.639 |
| Baseline-IT | 0.543 |
| Ours(ResNet-101$_{imagenet}$) | 0.751 |
| Ours(ResNet-50$_{nsfw}$) | **0.759** |

Table 5: Comparing the $f_1^{macro}$ of our methods and the baselines for binary classification task.

### 5.1 Task Results

**Subtask-A**: The results of the experiments for the binary classification task can be seen in Table 5.

[2]https://github.com/emiliantolo/pytorch_nsfw_model

| SubTask-B | |
|---|---|
| Model | $f_1^{weighted}$ |
| Baseline-Flat | 0.421 |
| Baseline-Heirarchical | 0.621 |
| Ours(ResNet-101$_{imagenet}$) | |
| + `SM Loss` | 0.641 |
| + `Custom Loss` | **0.645** |
| Ours(ResNet-50$_{nsfw}$) | |
| + `SM Loss` | 0.632 |
| + `Custom Loss` | 0.638 |

Table 6: Comparing the $f_1^{weighted}$ of our methods and the baselines for multi-label classification task.`SM Loss` refers to multi-label SoftMargin loss

| | | Multi-Label |
|---|---|---|
| **Backbone Model** | **Graph Depth** | $f_1^{weighted}$ |
| ResNet-101$_{imagenet}$ | *2-layer* | **0.644** |
| | *3-layer* | 0.644 |
| | *4-layer* | 0.632 |
| | *5-layer* | 0.628 |
| ResNet-50$_{nsfw}$ | *2-layer* | 0.641 |
| | *3-layer* | **0.643** |
| | *4-layer* | 0.629 |
| | *5-layer* | 0.621 |

Table 7: Graph Network Depth effect on model performance

For Subtask-A, the models that used multi-model training and an additional navigator network on the image end outperformed the single modality models and the simple multimodal concatenation model. One of the major reasons for our model outperforming the image-only baseline model could be that the navigator network learns to recognise relevant parts of the image as compared to passing the complete image as one. Amongst the model using *ResNet* backbone, the model fine-tuned on *nsfw* had an edge over the model which had been pre-trained on *imagenet* dataset. This can indicate that there is an indicator of women's image representation with the meme being a misogynistic one.

**Subtask-B**: The results of the experiments for the multi-label classification task can be seen in Table 6. For Subtask-B, the models that used graph network to create independent classifiers and an additional navigator network on the image end outperformed the models using the simple multi-modal concatenation model with classification head and the hierarchical multi-label model using text representations. Amongst the model using *ResNet* backbone, the model fine-tuned on *nsfw* dataset performed poorer to the model which had been pre-trained on *imagenet* dataset. This can be an indicator that general feature representations are perhaps more important for the identification of the specificity of misogyny as compared to that of the fine-tuned feature representations.

### 5.2 Ablation Studies

In this section, we perform ablation studies from two different aspects, particularly including the sensitivity of the classification models to effects of $\lambda$ when concatenating the two different types of modalities, visual and textual together, to determine the relative importance of the two with respect to each other, and the other being the depths of Graph Classification Module which we use for the multi-label classification model.

**Effects of different threshold values $\lambda$ :** We vary the values of the threshold concatenation parameter $\lambda$ from 0.1 to 0.9 in steps of 0.1. $\lambda = 0$ corresponds to building the entire feature vector from the visual stream while $\lambda = 1$ corresponds to the entire information coming from the textual stream. The results are shown in table 4, where the performance of the two models based on ResNet-101 backbone are compared pre-trained on two different datasets. It can be observed that the textual stream information is of higher importance in both the classification problem as the performance boost is skewed for roughly $\lambda = [0.6, 0.7]$. It may be due to the fact that in the images as well, a good amount of information that is used to recognise the misogyny of the meme is cognitively of the textual nature, while the image content of the meme is lesser in comparison to its textual counterpart. It may also be that the image content is not of high quality.

**Effects of different depth of Graph Classification Network:** We vary the values of the number of layers of the graph network from 2 to 5 and observe its effect on the model performance. For the two-layer model, the output dimensions of the layers are 512, 2048, for the three-layer model, the output dimensionalities are 512, 1024 and 2048 for the sequential layers, for the four-layer model, the dimensionalities are 512, 1024, 1024 and 2048, and for the five-layer model, the output dimensions

are 512, 1024, 1024, 1024, 2048. As shown in table 7, when the number of graph convolution layers increases, multi-label recognition performance drops on both datasets. The possible reason for the performance drop may be that when using more GCN layers, the propagation between nodes will be accumulated, which can result in over-smoothing.

**Effects of using Custom Loss Function:** We compare the results for multi-label classification with two types of losses : (i) MultiLabel Soft Margin Loss (`SM Loss`); (ii) Custom Loss as described in 3.3. From table 6, we can see that the `Custom Loss` out performs the `SM Loss` in the experimental runs.

The result can be explained by the fact that weighted classes affect the loss value for positive as well as negative labels.

**(i)** If the model predicts a positivity for the label which has a higher positive weightage the loss value would increase, thereby forcing the model to not favour one particular label. Similarly, when the model predicts a negative value for a particular label that has a higher negative weightage, the loss value would increase, forcing the model to not favour negativity of a particular label.

**(ii)** If the model predicts a positivity for the label which has a lower positive weightage, the loss value would decrease, thereby forcing the model to predict favourably for that particular label. Similarly, when the model predicts a negative value for a particular label that has a lower negative weightage, the loss value would decrease, forcing the model to favour the negativity of that particular label.

## 6 Conclusion

We have described the systems developed by as to solve the Multimedia Automatic Misogyny Identification challenge at Semeval 2022. In our best performing submission for SubTask-A, we framed the problem as a binary classification task and used two separate streams of information simultaneously to identify misogyny, while for our model for SubTask-B, we tried to find the semantic relation between the type of misogyny and their relative importance to solve the problem for Multi-Label classification. By making use of powerful, state-of-the-art, pre-trained models for text and images, our models were able to achieve a high F1 score for both the tasks. Our best performing model ranked 3rd out of the 10 teams submissions on SubTask-A and 22nd out of 30 team submissions on SubTask-B.

As part of future work, we aim to explore alternate approaches to model the multi-label dependencies using Knowledge-Graph and GAT Networks. Also, there seems to be a problem of oversmoothing when increasing the depth of the Graph Classification Module, which we aim to resolve using effective Normalization layers between the graph layers.

## References


Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Shuying Liu and Weihong Deng. 2015. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

Benet Oriol, Cristian Canton-Ferrer, and Xavier Giró i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. In

*NeurIPS 2019 Workshop on AI for Social Good*, Vancouver, Canada.

Marinella Paciello, Francesca D'Errico, Giorgia Saleri, and Ernestina Lamponi. 2021. Online sexist meme and its effects on moral and emotional processes in social media. *Comput. Hum. Behav.*, 116:106655.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Limor Shifman. 2013. Memes in a Digital World: Reconciling with a Conceptual Troublemaker. *Journal of Computer-Mediated Communication*, 18(3):362–377.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. 2018. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. In *AAAI*.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219.