SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene
University of Milano-Bicocca
Viale Sarca 336, 20126, Milan, Italy
elisabetta.fersini@unimib.it, francesca.gasparini@unimib.it
g.rizzi10@campus.unimib.it, a.saibene2@campus.unimib.it

Berta Chulvi, Paolo Rosso
Universitat Politècnica de València
Camino de Vera, Valencia, Spain
berta.chulvi@upv.es
prosso@dsic.upv.es

Alyssa Lees, Jeffrey Sorensen
Google Jigsaw,
111 8th Ave, New York, NY
alyssalees@google.com
sorenj@google.com

Abstract

The paper describes the SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI),which explores the detection of misogynous memes on the web by taking advantage of available texts and images. The task has been organised in two related sub-tasks: the first one is focused on recognising whether a meme is misogynous or not (Sub-task A), while the second one is devoted to recognising types of misogyny (Sub-task B). MAMI has been one of the most popular tasks at SemEval-2022 with more than 400 participants, 65 teams involved in Sub-task A and 41 in Sub-task B from 13 countries. The MAMI challenge received 4214 submitted runs (of which 166 uploaded on the leader-board), denoting an enthusiastic participation for the proposed problem. The collection and annotation is described for the task dataset. The paper provides an overview of the systems proposed for the challenge, reports the results achieved in both sub-tasks and outlines a description of the main errors for a comprehension of the systems capabilities and for detailing future research perspectives.

1 Introduction

Women have a strong presence online, particularly in image-based social media such as Twitter and Instagram: 78% of women use social media multiple times per day compared to 65% of men (Department, 2019). However, while new opportunities for women have been opened on the web, systematic inequality and discrimination offline is replicated in online spaces in the form of offensive contents against them (Frenda et al., 2019; Anzovino et al., 2018; Farrell et al., 2019; Plaza-Del-Arco et al., 2020; Gasparini et al., 2018). A popular communication tool in social media platforms are image macros popularly connoted as "memes" (Shifman, 2013). An internet meme is usually an image communicating pictorial content with an overlaid text that is added *a posteriori* by the meme author, with the main goal of being funny and/or ironic (Shifman, 2013). Although many memes are created with humorous intent, others have political or activist ambitions. Few familiar with the format would be surprised to learn that memes can be used to express hate against women, via sexist and aggressive messages in online environments (Paciello et al., 2021) that subsequently amplify the sexual stereotyping and gender inequality of the offline world (Franks, 2011). In order to counter this phenomenon, the Multimedia Automatic Misogyny Identification (MAMI) shared task has been organised at SemEval-2022 (Emerson et al., 2022). The proposed challenge consists of the identification of misogynous memes, taking advantage of both text and images available as sources of information. The task is organised around two main sub-tasks:

- **Sub-task A:** a basic task of misogynous meme identification, where a meme should be categorised either as misogynous or not misogynous;
- **Sub-task B:** an advanced task, where the type of misogyny should be recognised among potential overlapping categories such as stereotype, shaming, objectification and violence.

Some other tasks related to this topic, but that did not consider the same data and a multi-modal approach have been previously organised in the same area of interest, i.e. AMI@Evalita (Fersini et al., 2018a; Elisabetta Fersini, 2020), AMI@IberEval (Fersini et al., 2018b), HatEval (Basile et al., 2019) and OffenseEval (Zampieri

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 533 - 549
July 14-15, 2022 ©2022 Association for Computational Linguistics

(a) Shaming (b) Stereotype (c) Objectification (d) Violence

Figure 1: Examples of misogynous memes.

et al., 2020). However, the proposed MAMI challenge is a step forward the previous ones for two main reasons: (1) it is focused on multi-modality and (2) the type of misogynous contents are expressed in a completely different form, i.e. in the former challenge the presence of hateful contents was explicit within the text, while here it is often implicit.

2 Dataset and Annotation Process

Candidate memes have been collected by focusing on the following main types of misogyny:

- *Shaming*: The practice of criticising women who violate expectations of behaviour and appearance regarding issues related to gender typology (such as "slut shaming") or related to physical appearance (such as "body shaming") (Van Royen et al., 2018). This category focuses on content that seeks to insult and offend women because of some characteristics of their body or personality.

- *Stereotype*: a stereotype is a fixed, conventional idea or set of characteristics assigned to a woman (Eagly and Mladinic, 1989). A meme can use an image of a woman according to her role in the society (role stereotyping), or according to her personality traits and domestic behaviours (gender stereotyping).

- *Objectification*: A practice of seeing and/or treating a woman like an object (Szymanski et al., 2011).

- *Violence*: A meme that indicates physical and/or a call to violence against women (Andreasen, 2021).

Examples of the above mentioned types of misogynous memes are presented in Figure 1.

The procedure for collecting relevant memes for this shared task consisted of: (1) searching the most popular **social media platforms**, such as Twitter and Reddit; and (2) downloading samples from **websites** dedicated to meme creation and sharing, such as 9GaG, Knowyourmeme and Imgur, by site scraping and manual download. In both cases, in order to collect a proper number of misogynous memes, 4 main activities have been performed: (1) searching for threads dedicated to memes with women as the subject; (2) searching for threads or conversations dedicated to or written by persons who identify as anti-women or anti-feminist (such as the MGTOW website and the related threads on Reddit); (3) exploring discussions in recent events involving famous women (such as *Michelle Obama*); (4) searching by keywords and/or hashtags such as #girl, #girlfriend, #women, #feminist.
The final collection is composed of 15k memes that have been labelled by human annotators (duplicates have been previously removed). Among the labelled memes we obtained an adequate number of misogynous and non misogynous memes. The final benchmark dataset released for the MAMI challenge is composed of 10k memes for training and 1k for testing (balanced between classes). The dataset has been labelled using crowd-sourcing platforms according to the following primary questions[1]:
- Is this meme misogynous or not?
- If the meme is misogynous, what are the main categories to which the meme belongs (shaming,

[1]The prototype of the annotation interface and the annotation guidelines are reported in Appendix A

file_name	misogynous	shaming	stereotype	objectification	violence	Text Transcription
10846.jpg	1	0	1	1	1	SANDWICH!!!!!! don't make me tell you twice woman.

Table 1: Annotation format of the training and testing instances.

stereotype, objectification, violence)?
In the last case, i.e. related to the misogyny category, multiple overlapping labels have been considered. The memes were shown one at a time to avoid bias introduced by the annotators seeing multiple memes simultaneously.

Memes were annotated by 3 observers and the final label was given according to the majority of the labels (2/3). The text of the memes have been transcribed using Google Cloud Vision[2]. We report an example of a meme that has been provided to the participants as training example, which is composed of raw image (Figure 2) and the corresponding labels available through a csv file (Table 1).

Figure 2: Raw image (10486.jpg)

We estimated the inter-annotator agreement using the Fleiss-κ coefficient (Fleiss, 1971). In particular, we used the traditional Fleiss-κ measure for estimating the agreement related to the misogynous vs not misogynous annotation necessary for Sub-task A, while we adopted the Fleiss-κ with the MASI (Jaccard) index (Passonneau, 2006) to calculate the agreement between annotators on multiple (overlapping) annotations necessary for Sub-task B. Regarding the agreement on the misogynous vs not misogynous annotations, we estimated a coefficient equal to 0.5767, while for the type of misogyny labelling we derived a coefficient equal to 0.3373. We report in Table 2 the details about the dataset provided to the participants. The values of the Fleiss-κ measure suggest that the agreement for the misogynous labelling is moderate, denoting a quite simple task for humans, while the agreement for the type of misogyny annotation is fair, denoting a quite hard task.

[2]https://cloud.google.com/vision/docs/ocr

3 Evaluation Measures and Baseline

Sub-task A. Systems have been evaluated using macro-average F1-Measure. In particular, for each class label (misogynous and not misogynous) the corresponding F1-Measure has be computed, and the final score has been estimated as the arithmetic mean of the two F1-Measures. The baseline models used as benchmark with respect to the participants are:

- **Baseline Text**: a deep representation of text, a fine-tuned sentence embedding using the USE (Cer et al., 2018) pre-trained model;
- **Baseline Image**: deep representation of image content, based on a fine-tuned image classification model grounded on VGG-16 (Simonyan and Zisserman, 2014);
- **Baseline Image_Text**: a concatenation of the previous deep image and text representations through a single layer neural network.

We also used two multi-label models introduced for Sub-task B and detailed in the following paragraph.

Sub-task B. Systems have been evaluated using weighted-average F1-Measure. In particular, the F1-Measure has been computed for each label and then the average has been weighted by the number of true instances for each label. For Sub-task B, the baselines are grounded on:

- **Baseline Flat Multi-label**: 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 Multi-label**: a hierarchical multi-label model, based on text representations for predicting whether a meme is misogynous or not and, if misogynous, the corresponding type.

4 Participant Systems and Results

MAMI has been one of the most popular tasks in SemEval-2022, with 65 teams that joined Sub-task

	Misogyny Labelling (Sub-task A)			Type of Misogyny Labelling (Sub-task B)				
	Misogynous	**Not Misogynous**	**Fleiss-k Agreement**	**Shaming**	**Stereotype**	**Objectification**	**Violence**	**Fleiss-k Agreement**
Training Set	5000 (50%)	5000 (50%)	0.5767	1274 (25.48%)	2810 (56.20%)	2202 (44.04%)	953 (19.06%)	0.3373
Test Set	500 (50%)	500 (50%)		146 (29.20%)	350 (70.00%)	348 (69.60%)	153 (30.60%)	

Table 2: Dataset characteristics.

A and 41 teams that participated in Sub-task B. We received a total of 4,214 submissions, of which 166 submitted to the leader-board. Among the teams joining the MAMI challenge, 41 groups have provided the details about their participation (team name, number of team members, country, and description of their system). In Appendix B (Table 8), we report features about the teams that have provided team information for further analysis and discussion. On average, the teams are composed of 2 members, varying from 1-person teams (the most frequent case) to 7 members (the largest team). Regarding geographic distribution, the majority of the participants come from India (12 teams), followed by USA and Germany (5), UK and China (4), Italy and Spain (3) and the remaining countries with 1 team each.

As a general overview of the results, we report in Table 3 the mean, standard deviation (StDev), minimum, maximum, median and the first and third quartiles (Q1 and Q3) of the performance achieved by the participant teams.

In Sub-task A, we notice that the maximum value

	Min	Q1	Mean	Median	StDev	Q3	Max
Sub-task A	0.481	0.649	0.680	0.679	0.064	0.722	0.834
Sub-task B	0.467	0.634	0.663	0.680	0.059	0.706	0.731

Table 3: Basic statistics of the results for the participating systems in Sub-task A and Sub-task B, expressed in terms of macro-averaged and weighted-average F_1-score respectively.

(0.834) is much higher than the corresponding one in Sub-task B (0.731), while the difference is less evident when considering the mean (from 0.680 to 0.663) and the median value (from 0.679 to 0.680). When considering the max values, it emerges that Sub-task B seems to be more difficult than Sub-task A, while the median values indicates that for the 50% of the systems both tasks are equally challenging.

In regards to the models adopted by the participants, it has been observed that the majority of the teams exploited pre-trained models, distinguished in text-based, where the most used ones are based on BERT (Devlin et al., 2019) such as RoBERTa (Liu et al., 2019), and image-based models, where the most adopted ones are based on VisualBERT (Li et al., 2020a). Among these systems, considered by 90% of the teams, half of them adopted an ensemble strategy to make the final prediction. The remaining ones adopted either traditional neural networks (30%) or multi-task (20%) approaches to classify the memes. Few teams exploited models, such as CLIP (Radford et al., 2021) and ViLBERT (Lu et al., 2019), to jointly learn the characteristics of misogynous and not misogynous memes, and the related misogyny categories.

4.1 Sub-task A

Sub-task A was attempted by 65 teams, where 47 of them (72%) outperformed the best provided baseline, the Baseline Hierarchical Multi-label model, in terms of macro-averaged F_1-score. The highest score (0.834) has been obtained by the SRCB team (Zhang and Wang, 2022), which defined an ensemble model of deep multi-modal features with Multi Layer Perception (Kubat, 1999), Extreme Gradient Boosting (Chen and Guestrin, 2016) and Gradient-Boosted Decision Trees (Si et al., 2017).

We report in Table 4 the Top-10 teams in Sub-task A, ranked according to macro-average F_1-score (the overall leader-board is reported in Appendix C.) Regarding the top-3 systems, DD-TIG

	Team Name
1	SRCB (Zhang and Wang, 2022)
2	DD-TIG (Zhou et al., 2022)
3	RIT Boston (Chen and Chou, 2022)
4	NLPros
5	ASRtrans (Rao and Rao, 2022)
6	Poirot (Srivastava, 2022)
7	R2D2 (Sharma et al., 2022b)
8	PAIC (ZHI et al., 2022)
8	ymf924
8	RubCSG (Yu et al., 2022)
9	hate-alert
10	AMS_ADRN (Li et al., 2022)

Table 4: Top-10 teams in Sub-task A, ranked according to macro-average F_1-score.

(Zhou et al., 2022), ranked second place by defining an ensemble of different pre-trained models: (1) ERNIE-Vil (Yu et al., 2021), which incorporates structured knowledge obtained from scene graphs to learn joint representations of vision-language; (2) Uniter (Chen et al., 2020), which learns a joint multi-modal embedding through a Transformer-based architecture over four image-text datasets; (3) VisualBERT (Li et al., 2020a), which is composed of a stack of Transformer layers that implicitly align elements of an input text and regions in an associated input image with self-attention; (4) Oscar (Li et al., 2020b), which exploits object tags detected in an image as anchor point to learn the alignment with the caption fragments.

RIT Boston (Chen and Chou, 2022) ranked third and used OpenAI's CLIP model (Radford et al., 2021) to obtain high-quality multi-modal features and then used a logistic regression (LR) model to make a binary classification. In their model, a data-centric AI principle was used to further improve performance by manually rating a subset of test data and adding this extra data into the train set.

4.2 Sub-task B

Sub-task B was attempted by 41 teams, where 35 of them (85%) outperformed the best MAMI baseline, which also in this case is the Baseline Hierarchical Multi-label model. We report in Table 5 the Top-10 teams in Sub-task B, ranked according to weighted-average F_1-score (the overall leaderboard is reported in Appendix C). The highest result (0.731) has been obtained by three teams, i.e., SRCB (Zhang and Wang, 2022), TIB-VA (Hakimov et al., 2022) and PAIC (ZHI et al., 2022). The SRCB team (Zhang and Wang, 2022) adopted the same ensemble model used for Sub-task A. The system developed by TIB-VA is instead based on a Deep Learning model grounded on CLIP image and text features combined with a LSTM (Hochreiter and Schmidhuber, 1997), while PAIC (ZHI et al., 2022) did not provide any information about their approach. In second place, the ymf924 team did not provide any information about their approach, while in third place is the DD-TIG (Zhou et al., 2022) team with the same approach used for Sub-task A.

	Team Name
1	SRCB (Zhang and Wang, 2022)
	TIB-VA (Hakimov et al., 2022)
	PAIC (ZHI et al., 2022)
2	ymf924
3	DD-TIG (Zhou et al., 2022)
4	NLPros
5	QMUL
6	Unibo (Muti et al., 2022)
7	RubCSG (Yu et al., 2022)
8	AMS_ADRN (Li et al., 2022)
9	taochen (Tao and jae Kim, 2022)
10	ASRtrans (Rao and Rao, 2022)

Table 5: Top-10 teams in Sub-task B, ranked according to weighted-average F_1-score.

In general, the most predominant models for addressing Sub-task B are multi-class approaches, multi-task learning, and/or ensemble methods, where the feature space for learning has been derived either by image and text pre-trained models or by a joint embedding space.

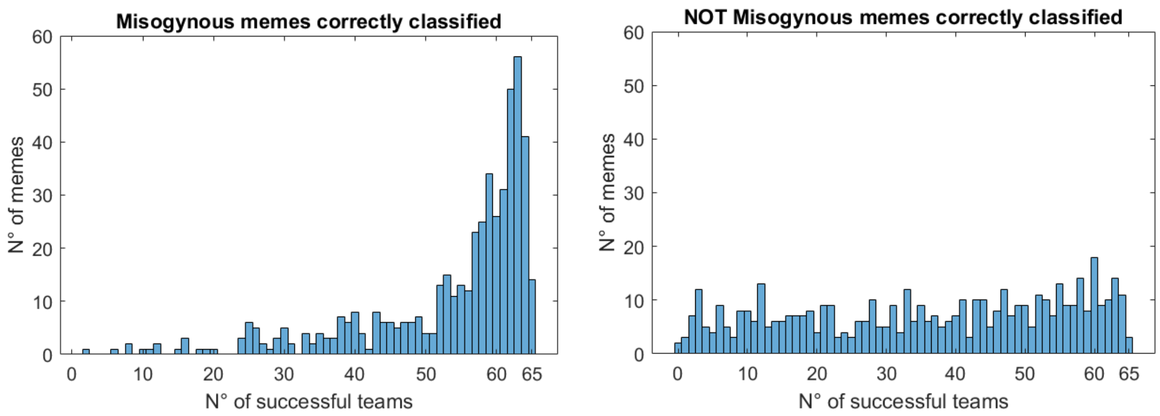
5 Error Analysis

In order to gain deeper insight into the prediction capabilities of the systems and delineate the open issues about the recognition and classification of misogynous memes, we conducted a detailed error analysis on both sub-tasks, considering all participating teams. The error distributions and the types of the most common errors in regards to the labels to be predicted are detailed in the following subsections. We considered memes misclassified by at least 25%, 50% and 75% of the teams, distinguishing False Positive (FP) and False Negative (FN), according to the labels available in each sub-task. For the memes misclassified by at least 75% of the teams, we reported the most frequent types of errors by analysing the visual and textual content of the memes.

5.1 Sub-task A

In Figure 3, the distribution of correct classifications with respect to the number of successful teams is reported for misogynous and not misogynous memes. The distribution of correctly classified misogynous memes (Figure 3(a)) is uni-modal and peaked towards higher values, implying that most memes have been correctly classified by most teams. On the other hand, considering the not misogynous ones, the distribution is more uniform (Figure 3(b)), denoting that in general the models are more recall than precision oriented. There are 14 memes out of 500 (2.8%) correctly classified as misogynous by all the teams (Figure 3(a), last bin), while no one is misclassified by all the teams. In the worst case, only one misogynous meme was misclassified by 63 out of 65 teams.

In Table 6 the error distribution of Sub-task A is reported, considering the misclassification of misogynous memes and not misogynous ones sepa-



(a) (b)

Figure 3: Distributions of correct classifications with respect to the number of successful teams for misogynous (a) and not misogynous (b) memes.

Teams	Misogynous memes predicted as NOT Misogynous (FN)		NOT Misogynous memes predicted as Misogynous (FP)		Overall misclassified memes (FP+FN)	
25% (16 teams)	128	25.60%	340	68.00%	468	46.80%
50% (33 teams)	46	9.20%	220	44.00%	266	26.60%
75% (49 teams)	12	2.40%	109	21.80%	121	12.10%

Table 6: Error distribution on Sub-task A.

rately, and finally the overall errors. In general, the percentage of classification errors of non misogynous memes are higher than misogynous ones, confirming that the methods are more precision than recall oriented. This suggests that most of the systems tends to be biased towards the misogyny category due to the presence of text or images that mislead the systems. Focusing on the memes misclassified by at least 75% of the teams, the most frequent types of errors can be summarised in the following paragraphs.

Misogynous memes predicted as NOT misogynous (FN). Twelve memes belong to this set. Five of them involve sexual objectification, that requires correlation of textual and visual content to classify. In particular, the meme depicted in Figure 4 is characterised by a neutral text and depicts a neutral object. In this case, the shape of the object together with the text needs to be correlated to grasp the sexual meaning. This meme was correctly classified by only 6 teams out of 65. Another group of misclassified memes, corresponding to one third of this set, is related to violence, both physical (visually represented), and sexual, which is less explicitly evoked.

Figure 4: A misogynous meme classified as a non misogynous one (Raw image: 17013.jpg).

NOT Misogynous memes predicted as misogynous (FP). 109 NOT misogynous memes were incorrectly predicted by at least 75% of the teams. The majority of the misclassified memes contain textual or visual content that are often contained in misogynous memes. For example, 38% of the memes contain words and phrases such as “woman, man, fat, boobs, kitchen, dishwasher, chicks, make me a sandwich, . . . ”, and 31% depict close up images of women, which often emphasise the neck-

Figure 5: Example of meme with an antithetical content (Raw image: 15138.jpg).

Figure 6: Most common example of Shaming meme misclassified as NOT Shaming (Raw image: 15559.jpg)

line, or depict faces with evident makeup. An interesting group of misclassified memes (7 out of 109) shows antithetical content. In general, most of the visual and textual information recall typical misogynous memes (with viral phrases such as "back to the kitchen" or depicting misogynous scenes such as physical violence), however additional information both visual and textual, with an opposite meaning, changes the overall message conveyed, as depicted in the example in Figure 5.

Memes featuring famous characters or actors who are often depicted associated to messages of all kinds, such as Ryan Gosling with the "hey girl" memes, Dwight Schrute or Willy Wonka, are also frequently misclassified (about 10%). Finally it is worth noting that other misclassified memes are those that convey feminist ideals and content.

5.2 Sub-task B

We report in Table 7, the error distribution of Sub-task B, accordingly to the labels predicted (i.e., Stereotype, Violence, Shaming and Objectification). The first interesting insights involve the misogyny categories that are misclassified by at least 75% of the teams, in a ranked order: Objectification (14.60% of memes are wrongly classified by at least 31 teams in the over 41 participating teams), Stereotype (13.10%), Violence (3.30%) and Shaming (3.2%). A further interesting insight relates to the ability of the models with respect to the False Negative (FN) and the False Positive (FP) of each class. While for Shaming and Violence the percentage of FP (0.82% and 0.12% respectively) is much lower than the percentage of FN (17.2% and 20.92%), for Stereotype and Objectification the opposite is true, where FP (27.71% and 35.92% respectively) rates are much higher then FN (5.23% and 3.22%). We analysed the most predominant errors, with respect to each misogyny category.

Shaming. Regarding the first misogyny category, the most frequent error by at least 75% of the teams relates to the classification of Shaming memes as NOT Shaming (17.12%). The majority of the memes wrongly classified relates to the concept of *fat shaming* where overweight women are compared, implicitly or explicitly, to a narrow standard. An example of such errors is reported in Figure 6.

Violence. With the Violence category, the most frequent error by at least 75% of the teams relates to the classification of Violence memes as NOT Violence ones (20.92%). In this case, the majority of the memes wrongly classified as NOT Violence relates to the concept of *physical assault* typically depicted with a violent image (e.g., woman with bruises) but with neutral text (e.g., "don't tell her twice") or by a neutral image (e.g., standing men) coupled with a violent text (e.g., "women need a good beating once in a while"). An example of a misclassified violent meme is shown in Figure 7.

Stereotype. In the Stereotype category, the most frequent error by at least 75% of the teams relates to the classification of NOT Stereotype memes as Stereotype ones (27.71%). In this case, the most frequent misclassification concerns memes that are related to the concept of *men in the kitchen*, where the image typically represents men and the text is related to the stereotype of woman in kitchen ("cooking"). An example of such errors is re-

Teams	Shaming predicted as NOT Shaming (FN)		NOT Shaming predicted as Shaming (FP)		Overall misclassified Shaming memes (FP+FN)	
25% (11 teams)	92	63.01%	143	16.74%	235	23.50%
50% (21 teams)	59	40.41%	44	5.15%	103	10.30%
75% (31 teams)	25	17.12%	7	0.82%	32	3.20%
	Violence predicted as NOT Violence (FN)		NOT Violence predicted as Violence (FP)		Overall misclassified Violence memes (FP+FN)	
25% (11 teams)	90	58.82%	32	3.78%	122	12.20%
50% (21 teams)	65	42.48%	6	0.71%	71	7.10%
75% (31 teams)	32	20.92%	1	0.12%	33	3.30%
	Stereotype predicted as NOT Stereotype (FN)		NOT Stereotype predicted as Stereotype (FP)		Overall misclassified Steretype memes (FP+FN)	
25% (11 teams)	236	36.31%	278	79.43%	514	51.40%
50% (21 teams)	94	14.46%	190	54.29%	284	28.40%
75% (31 teams)	34	5.23%	97	27.71%	131	13.10%
	Objectification predicted as NOT Objectification (FN)		NOT Objectification predicted as Objectification (FP)		Overall misclassified Objectification memes (FP+FN)	
25% (11 teams)	151	23.16%	260	74.71%	411	41.10%
50% (21 teams)	65	9.97%	205	58.91%	270	27.00%
75% (31 teams)	21	3.22%	125	35.92%	146	14.60%

Table 7: Error distribution on Sub-task B.

Figure 7: Most common example of Violence meme misclassified as NOT Violent (Raw image: 16067.jpg)

ported in Figure 8. The analysis of the errors in the stereotyped category is controversial and interesting. Some of the memes that our annotators have labelled as non-stereotypical could be considered expressions of benevolent sexism (Glick and Fiske, 1996). Benevolent sexism is a subtle form of prejudice, which apparently values women more than men but does it connecting this positive evaluation to their traditional roles. This is a manifestation of sexism that is difficult to detect and it is still not consensual in society. In fact, these memes were considered by our annotators not to be an expression of stereotype. The task team decided to keep the annotators' view that reflects the majority thinking in society today, however, the models seem to have detected benevolent sexism and the errors go in that direction. If models are only detecting the kitchen scenario or a more subtle form of prejudice is an intriguing question for future research.

Figure 8: Most common example of NOT misogynous and NOT Stereotype meme misclassified as Stereotype (Raw image: 15137.jpg)

Objectification. In the Objectification category, the most frequent error by at least 75% of the teams relates to the classification of NOT Objectification memes as Objectification (35.92%). In this case, there is not a predominant archetype over the others that confounds the majority of the models.

6 Conclusions

The high number of participating teams at the MAMI challenge at SemEval-2022 confirms the growing interest of the research community not only in detecting abusive language but also pictorial content as sources of information. Overall, results and error analysis confirm that the detection of misogynous memes is challenging, with many open issues that need to be addressed. First of all, the fact that the most predominant error in misogyny recognition relates to the misclassification of NOT misogynous memes as misogynous ones suggests that some potential issues could be related to biased models. The research community is therefore encouraged to pay attention not only to accuracy metrics, but also to ensure models are unbiased before applying them in a real context. Another open issue relates to the capability of the systems to model the dynamics of the memes. Every day different memes, with different images and different text are generated on the web and shared online.

References

Samyak Agrawal and Radhika Mamidi. 2022. Lastresort at semeval-2022 task 5: Towards misogyny identification using visual linguistic model ensembles and task-specific pretraining. In *The 16th International Workshop on Semantic Evaluation*.

Maja Brandt Andreasen. 2021. 'rapeable'and 'unrapeable'women: the portrayal of sexual violence in internet memes about# metoo. *Journal of Gender Studies*, 30(1):102–113.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Ayme Arango, Jesus Perez-Martin, and Arniel Labrada. 2022. Hateu at semeval-2022 task 5: Multimedia automatic misogyny identification. In *The 16th International Workshop on Semantic Evaluation*.

Giuseppe Attanasio, Debora Nozza, and Federico Bianchi. 2022. MilaNLP at semeval-2022 task 5: Using perceiver IO for detecting misogynous memes with text and image modalities. In *The 16th International Workshop on Semantic Evaluation*.

Shubham Kumar Barnwal, Ritesh Kumar, and Rajendra Pamula. 2022. IIT DHANBAD CODECHAMPS at semeval-2022 task 5: MAMI - multimedia automatic misogyny identification. In *The 16th International Workshop on Semantic Evaluation*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mitra Behzadi, Ali Derakhshan, and Ian Harris. 2022. Mitra behzadi at semeval-2022 task 5 : Multimedia automatic misogyny identification method based on CLIP. In *The 16th International Workshop on Semantic Evaluation*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Lei Chen and Hou Wei Chou. 2022. RIT boston at semeval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle. In *The 16th International Workshop on Semantic Evaluation*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Pablo Cordon, Pablo Gonzalez Diaz, Jacinto Mata, and Victoria Pachón. 2022. I2c at semeval-2022 task 5: Identification of misogyny in internet memes. In *The 16th International Workshop on Semantic Evaluation*.

Statista Research Department. 2019. Share of u.s. adults who use social media 2019, by gender.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Alice H Eagly and Antonio Mladinic. 1989. Gender stereotypes and attitudes toward women and men. *Personality and social psychology bulletin*, 15(4):543–558.

Paolo Rosso Elisabetta Fersini, Debora Nozza. 2020. Ami @ evalita2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. CEUR.org.

Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, and Nathan Schneider. 2022. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, pages 87–96.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@ SEPLN*, pages 214–228.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Mary Anne Franks. 2011. Unwilling avatars: Idealism and discrimination in cyberspace. *Colum. J. Gender & L.*, 20:224.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

José Antonio García-Díaz, Camilo Caparros-Laiz, and Rafael Valencia-García. 2022. UMUTeam at semeval-2022 task 5: Combining image and textual embeddings for multi-modal automatic misogyny identification. In *The 16th International Workshop on Semantic Evaluation*.

Francesca Gasparini, Ilaria Erba, Elisabetta Fersini, and Silvia Corchs. 2018. Multimodal classification of sexist advertisements. In *ICETE (1)*, pages 565–572.

Peter Glick and Susan T Fiske. 1996. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3):491–512.

Qin Gu, Nino Meisinger, and Anna-Katharina Dick. 2022a. Qinian at semeval-2022 task 5: Multi-modal misogyny detection and classification. In *The 16th International Workshop on Semantic Evaluation*.

Yimeng Gu, Ignacio Castro, and Gareth Tyson. 2022b. MMVAE at semeval-2022 task 5: A multi-modal multi-task VAE on misogynous meme detection. In *The 16th International Workshop on Semantic Evaluation*.

Mohammad Habash, yahya Daqour, Malak AG Abdullah, and Mahmoud Al-Ayyoub. 2022. YMAI at semeval-2022 task 5: Detecting misogyny in memes using visualBERT and MMBT multimodal pre-trained models. In *The 16th International Workshop on Semantic Evaluation*.

Sherzod Hakimov, Gullal Singh Cheema, and Ralph Ewerth. 2022. TIB-VA at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. In *The 16th International Workshop on Semantic Evaluation*.

Chao Han, Jin Wang, and Xuejie Zhang. 2022. YNU-HPCC at semeval-2022 task 5: Multi-modal and multi-label emotion classification based on LXMERT. In *The 16th International Workshop on Semantic Evaluation*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Álvaro Huertas-García, Helena Liz, Guillermo Villar-Rodríguez, Alejandro Martín, Javier Huertas-Tato, and David Camacho. 2022. AIDA-UPM at semeval-2022 task 5: Exploring multimodal late information fusion for multimedia automatic misogyny identification. In *The 16th International Workshop on Semantic Evaluation*.

Milan Kalkenings and Thomas Mandl. 2022. University of hildesheim at semeval-2022 task 5: Combining deep text and image models for multimedia misogyny detection. In *The 16th International Workshop on Semantic Evaluation*.

Miroslav Kubat. 1999. Neural networks: a comprehensive foundation by simon haykin, macmillan, 1994, isbn 0-02-352781-7. *The Knowledge Engineering Review*, 13(4):409–412.

Da Li, Ming Yi, and Yukai He. 2022. AMS_ADRN at semeval-2022 task 5: A suitable image-text multimodal joint modeling method for multi-task misogyny identification. In *The 16th International Workshop on Semantic Evaluation*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020a. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Gustavo Lorentz and Viviane Moreira. 2022. INF-UFRGS at semeval-2022 task 5: analyzing the performance of multimodal models. In *The 16th International Workshop on Semantic Evaluation*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Shankar Mahadevan, Sean Benhur, Roshan Nayak, Malliga Subramanian, Kogilavani Shanmugavadivel, Kanchana Sivanraju, and Bharathi Raja Chakravarthi. 2022. Transformers at semeval-2022 task 5: A feature extraction based approach for misogynous meme detection. In *The 16th International Workshop on Semantic Evaluation*.

Paridhi Maheshwari and Sharmila Reddy Nangi. 2022. Teamotter at semeval-2022 task 5: Detecting misogynistic content in multimodal memes. In *The 16th International Workshop on Semantic Evaluation*.

Ahmed Mahmoud Mahran, Carlo Alessandro Borella, and Konstantinos Perifanos. 2022. Codec at semeval-2022 task 5: Multi-modal multi-transformer misogynic meme classification framework. In *The 16th International Workshop on Semantic Evaluation*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Arianna Muti, Katerina Korre, and Alberto Barrón-Cedeño. 2022. UniBO at semeval-2022 task 5: A multimodal bi-transformer approach to the binary and fine-grained identification of misogyny in memes. In *The 16th International Workshop on Semantic Evaluation*.

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. *Computers in Human Behavior*, 116:106655.

Andrei Paraschiv, Mihai Dascalu, and Dumitru Clementin Cercel. 2022. UPB at semeval-2022 task 5: Enhancing UNITER with image sentiment and graph convolutional networks for multimedia automatic misogyny identification. In *The 16th International Workshop on Semantic Evaluation*.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Flor-Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2020. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–19.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Tathagata Raha, Sagar Joshi, and Vasudeva Varma. 2022. IIITH at semeval-2022 task 5: A comparative study of deep learning models for identifying misogynous memes. In *The 16th International Workshop on Semantic Evaluation*.

Ailneni Rakshitha Rao and Arjun Rao. 2022. ASRtrans at semeval-2022 task 5: Transformer-based models for meme classification. In *The 16th International Workshop on Semantic Evaluation*.

Jason Ravagli and Lorenzo Vaiani. 2022. JRLV at semeval-2022 task 5: The importance of visual elements for misogyny identification in memes. In *The 16th International Workshop on Semantic Evaluation*.

Edgar Roman-Rangel, Jorge Fuentes-Pacheco, and Jorge Hermosillo Valadez. 2022. Vision-language approach to recognize misogynous content in memes. In *The 16th International Workshop on Semantic Evaluation*.

Gagan Sharma, Gajanan Sunil Gitte, Shlok Goyal, and Raksha Sharma. 2022a. IITR codebusters at semeval-2022 task 5: Misogyny identification using transformers. In *The 16th International Workshop on Semantic Evaluation*.

Mayukh Sharma, Ilanthenral Kandasamy, and Vasantha W B. 2022b. R2d2 at semeval-2022 task 5: Attention is only as good as its values! a multimodal system for identifying misogynist memes. In *The 16th International Workshop on Semantic Evaluation*.

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

Si Si, Huan Zhang, S Sathiya Keerthi, Dhruv Mahajan, Inderjit S Dhillon, and Cho-Jui Hsieh. 2017. Gradient boosted decision trees for high dimensional sparse output. In *International conference on machine learning*, pages 3182–3190. PMLR.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Rajalakshmi Sivanaiah, Angel Deborah S, Sakaya Milton Rajendram, and Mirnalinee T T. 2022. TechSSN at semeval-2022 task 5: Multimedia automatic misogyny identification using deep learning models. In *The 16th International Workshop on Semantic Evaluation*.

Harshvardhan Srivastava. 2022. Poirot at semeval-2022 task 5: Leveraging graph network for misogynistic meme detection. In *The 16th International Workshop on Semantic Evaluation*.

Dawn M Szymanski, Lauren B Moffitt, and Erika R Carr. 2011. Sexual objectification of women: Advances to theory and research $1\psi7$. *The Counseling Psychologist*, 39(1):6–38.

Chen Tao and Jung jae Kim. 2022. taochen at semeval-2022 task 5: Multimodal multitask learning and ensemble learning. In *The 16th International Workshop on Semantic Evaluation*.

Kathleen Van Royen, Karolien Poels, Heidi Vandebosch, and Michel Walrave. 2018. Slut-shaming 2.0. In *Sexting*, pages 81–98. Springer.

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 graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

Wentao Yu, Benedikt Boenninghoff, Jonas Röhrig, and Dorothea Kolossa. 2022. RubCSG at semeval-2022 task 5: Ensemble learning for identifying misogynous MEMEs. In *The 16th International Workshop on Semantic Evaluation*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

Jing Zhang and Yujin Wang. 2022. SRCB at semeval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *The 16th International Workshop on Semantic Evaluation*.

JIN MEI ZHI, Zhou Mengyuan, Mengfei Yuan, Dou Hu, Xiyang Du, Lianxin Jiang, Yang Mo, and XiaoFeng Shi. 2022. PAIC at semeval-2022 task 5: Multimodal misogynous detection in MEMES with multitask learning and multi-model fusion. In *The 16th International Workshop on Semantic Evaluation*.

Ziming Zhou, Han Zhao, Jingjing Dong, Ning Ding, Xiaolong Liu, and Kangli Zhang. 2022. DD-TIG at semeval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification. In *The 16th International Workshop on Semantic Evaluation*.

A Annotation Guidelines

We report here the annotation guidelines provided to the annotators participating in the crowdsourcing annotation process of the collected memes. Since some memes contain sensitive content, we provided an explicit advisory message to the annotators.

A.1 Overview

The job aims at labelling English memes shared by users on the web as misogynous or not misogynous. The first step is about collecting socio-demographic information about the annotators:

- Gender: indicate your gender as female, male, unspecified
- Age: please choose your age range between 18-15, 25-35, 35-45, 45-60, over-60
- Location: please indicate your country of birth

The second step is about misogyny labelling. Annotators have to decide whether a meme is misogynous or not. If a meme is labelled as misogynous, then two other questions will be answered:

- Type of misogyny: the annotator should indicate (multiple choice) if the meme represents shaming, stereotype, objectification and/or violence.
- Misogyny rating: the annotator should provide a rating about how much the meme is misogynous using stars, i.e. *, ** or ***.

A.2 Guidelines and examples

Misogyny Labelling. Looking at a meme at a time, annotators should label it as misogynous or not according to the following definitions:

- ***Misogynous***: a meme is misogynous if it conceptually describes an offensive, sexist or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group of women. Misogyny can be expressed in the form of shaming, stereotype, objectification and/or violence.
- ***Not Misogynous***: a meme that does not express any form of hate against women.

<u>Remark</u>: a meme is NOT misogynous if it is conceptually not related to women or even if it is related to women, but it does not represent an offensive, sexist or hateful concept against women.

Type of misogyny. If a meme is considered misogynous, then the annotator has to choose one or more types of misogynous categories, according to the following definitions:

- ***Shaming:*** memes aimed at insulting and offending women because of some characteristics of the body. These types of misogynous memes are related to denigrating the physical appearance of women (body shaming).
- ***Stereotype*:** memes are aimed at representing a fixed idea or set of characteristics assigned to women. These types of memes convey the image of women according to their role in the society (i.e., Role Stereotyping), to her personality traits and domestic behaviours (i.e., Gender Stereotyping) or to fixed ideological characteristics related to women's rights (i.e., Feminism Stereotype).
- ***Objectification:*** it is a practice of seeing and/or treating a woman like an object. These types of memes usually report an over-appreciation of women's physical appeal, depicting woman as an object (sexual objectification or human being without any value as a person).
- ***Violence:*** indicates a physical or verbal violence represented by textual or visual content. These types of misogynous memes are aimed at showing violence against women or at alluding to the intent of physically assert power over women.

Misogyny Rating. If a meme is considered misogynous then the annotator has to indicate, according to his/her opinion, how misogynistic it is using a 1 to 3 ratings: * indicates weak misogyny, ** means medium misogyny, *** means strong misogyny.

B Team Information

We report here the details provided by those teams that have responded to a request for team information.

Team Name	Country	Members
InfUfrgs	Brazil	1
HateU	Chile	3
AMS_ADRN	China	3
DD-TIG		1
SRC-B		6
YNU-HPCC		3
TIB-VA	Germany	2
qinian		3
Hildesheim		1
RubCSG		4
TechSSN	India	4
IITR CodeBusters		3
IIT DHANBAD CODECHAMPS		1
LastResort		1
SSN_NLP_MLRG		2
Gini_us		3
ASRtrans		1
IIITG-ADBU		1
Transformers		7
R2D2		3
Poirot		1
Tathagata Raha		1
JRLV	Italy	2
Unibo		3
Triplo7		1
YMAI	Jordan	2
UAEM-ITAM	Mexico	3
UPB	Romania	1
taochen	Singapore	1
UMUTeam	Spain	1
AIDA-UPM		6
I2C		1
NLPros	UK	5
MMVAE		1
codec		1
QMUL		1
Mitra Behzadi	USA	1
RIT Boston		2
Charicfc		1
Stanford MLab		5
TeamOtter		2

Table 8: Team characteristics.

C Leader-boards

C.1 Leader-board of Sub-task A

We report in Table 9 the leader-board for Sub-task A. Team Names marked with * have submitted team name and additional information for further analysis and discussion. For those teams that have not provided the Team Name, we maintained the user name used on Codalab for submitting their predictions.

To produce the reported leader-board, we filtered the ranking defined by the evaluated metrics to maintain only the highest achieved score per group. Afterwards, we scrolled through this ranking from top to bottom in order to create clusters based on the obtained scores and the statistical difference resulting from the application of the McNemar test (McNemar, 1947).
In particular, starting from the first entry in the ranking, we have included in the same cluster the groups that presented (1) the same score or (2) had a statistical equality in performance.
As stated before, statistical equality was computed with a pairwise analysis performed with the McNemar test: we evaluated the equality in performance of the analysed algorithm with the algorithm that obtained the highest score within the cluster, considering a value of alpha equal to 0.05. According to this criterion, in the event that the algorithm under analysis could not be included in the cluster, a new one was created; the subsequent ones would have been compared with the latter.

Notice that in the leader-board were maintained all the baseline results for comparison.

C.2 Leader-board of Sub-task B

We report in Table 10 the leader-board for Sub-task B. Team Names marked with * have submitted team name and additional information for further analysis and discussion. For those teams that have not provided the Team Name, we maintained the user name used on Codalab for submitting their predictions.

To obtain the reported leader-board, a similar approach to the one used for Sub-task A has been adopted. A McNemar test (McNemar, 1947) was adopted to evaluate the similarity in performance for the identification of every single type of misogyny. Two algorithms have been considered statistically equal in performance if there was statistical significance in all 4 tests (i.e., if there was a statistical significance for the performance related to all 4 types of misogyny). Thus, a difference in performance for the prediction of only one of the four types has been valued sufficient to consider the analysed algorithm as statistically unequal. As for Sub-task A, the grouping depends on statistical equality and on the scores obtained.

Notice that in the leader-board were maintained all the baseline results for comparison.

Leaderboard Sub-task A		
	Team Name	Macro-average F_1-score
1	SRCB* (Zhang and Wang, 2022)	0.834
2	DD-TIG* (Zhou et al., 2022)	0.794
	RIT Boston* (Chen and Chou, 2022)	0.778
	NLPros*	0.771
3	ASRtrans* (Rao and Rao, 2022)	0.761
	Poirot* (Srivastava, 2022)	0.759
	R2D2* (Sharma et al., 2022b)	0.757
	PAIC (ZHI et al., 2022)	0.755
	ymf924	0.755
	RubCSG* (Yu et al., 2022)	0.755
	hate-alert	0.753
	AMS_ADRN* (Li et al., 2022)	0.746
	TIB-VA* (Hakimov et al., 2022)	0.734
4	union	0.727
	Unibo* (Muti et al., 2022)	0.727
	MMVAE* (Gu et al., 2022b)	0.723
	YMAI* (Habash et al., 2022)	0.722
	Transformers* (Mahadevan et al., 2022)	0.718
	taochen* (Tao and jae Kim, 2022)	0.716
	codec* (Mahran et al., 2022)	0.715
	QMUL*	0.714
	UPB* (Paraschiv et al., 2022)	0.714
	HateU* (Arango et al., 2022)	0.712
	yuanyuanya	0.708
	Triplo7* (Attanasio et al., 2022)	0.699
	InfUfrgs* (Lorentz and Moreira, 2022)	0.698
	Mitra Behzadi* (Behzadi et al., 2022)	0.694
	Gini_us*	0.692
5	riziko	0.687
	UMUTeam* (García-Díaz et al., 2022)	0.687
	Tathagata Raha* (Raha et al., 2022)	0.687
	LastResort* (Agrawal and Mamidi, 2022)	0.686
	TeamOtter* (Maheshwari and Nangi, 2022)	0.679
	ShailyDesai	0.677
	JRLV* (Ravagli and Vaiani, 2022)	0.670
	I2C* (Cordon et al., 2022)	0.665
	qinian* (Gu et al., 2022a)	0.665
	A.111	0.662
	IITR CodeBusters* (Sharma et al., 2022a)	0.662
	YNU-HPCC* (Han et al., 2022)	0.662
	WeiLW	0.661
	SSN_NLP_MLRG*	0.658
	UNIBUC-FMI	0.657
6	IIT DHANBAD CODECHAMPS* (Barnwal et al., 2022)	0.656
	Sattiy	0.655
	lianlio	0.654
	thisisatharva	0.650
	Baseline_Hierarchical_M.	0.650

Table 9 Continued from previous page

Leaderboard Sub-task A		
	Team Name	**Macro-average F_1-score**
6	IIITG-ADBU*	0.649
	UAEM-ITAM* (Roman-Rangel et al., 2022)	0.641
	Baseline_Image	0.640
	Baseline_Text	0.639
	Yet	0.639
	RaNdom	0.638
	AIDA-UPM* (Huertas-García et al., 2022)	0.636
	vishesh_gupta	0.634
	Levante	0.634
	Aily	0.632
	Charicfc*	0.620
	Stanford MLab*	0.619
7	rhitabrat	0.609
	Will To Live	0.606
	Hildesheim* (Kalkenings and Mandl, 2022)	0.603
	SakshiSingh	0.579
8	*Baseline_Image_Text*	0.543
	areen	0.524
	TechSSN* (Sivanaiah et al., 2022)	0.522
9	UET	0.481
10	*Baseline_Flat_Multilabel*	0.437

Table 9: Leader-board of Sub-task A.

Leaderboard of Sub-task B		
	Team Name	**Weighted-average F_1-score**
1	SRCB* (Zhang and Wang, 2022)	0.731
	TIB-VA* (Hakimov et al., 2022)	0.731
	PAIC (ZHI et al., 2022)	0.731
	ymf924	0.730
2	DD-TIG* (Zhou et al., 2022)	0.728
	NLPros*	0.720
3	QMUL*	0.713
4	Unibo* (Muti et al., 2022)	0.710
5	RubCSG* (Yu et al., 2022)	0.709
5	AMS_ADRN* (Li et al., 2022)	0.708
6	taochen* (Tao and jae Kim, 2022)	0.706
7	ASRtrans* (Rao and Rao, 2022)	0.705
8	codec* (Mahran et al., 2022)	0.698
9	Transformers* (Mahadevan et al., 2022)	0.695
10	Triplo7* (Attanasio et al., 2022)	0.693
	LastResort* (Agrawal and Mamidi, 2022)	0.692
	R2D2* (Sharma et al., 2022b)	0.690
	hate-alert	0.690
11	RIT Boston* (Chen and Chou, 2022)	0.689
12	Mitra Behzadi* (Behzadi et al., 2022)	0.681

Table 10 Continued from previous page

Leaderboard Sub-task B		
	Team Name	**Weighted-average F_1-score**
13	TeamOtter*(Maheshwari and Nangi, 2022)	0.680
14	Tathagata Raha* (Raha et al., 2022)	0.679
15	UPB* (Paraschiv et al., 2022)	0.673
16	riziko	0.668
17	UMUTeam* (García-Díaz et al., 2022)	0.663
18	UAEM-ITAM* (Roman-Rangel et al., 2022)	0.646
19	RaNdom	0.643
	qinian* (Gu et al., 2022a)	0.637
	UNIBUC-FMI	0.637
20	IITR CodeBusters* (Sharma et al., 2022a)	0.635
21	MMVAE* (Gu et al., 2022b)	0.634
	Yet	0.634
22	YNU-HPCC* (Han et al., 2022)	0.633
	Poirot* (Srivastava, 2022)	0.632
23	AIDA-UPM* (Huertas-García et al., 2022)	0.629
24	*Baseline_Hierarchical_M.*	0.621
25	YMAI* (Habash et al., 2022)	0.592
26	yuanyuanya	0.584
27	Stanford MLab*	0.563
28	UET	0.499
29	TechSSN* (Sivanaiah et al., 2022)	0.467
30	*Baseline_Flat_Multilabel*	0.421

Table 10: Leader-board of Sub-task B.