JRLV at SemEval-2022 Task 5: The Importance of Visual Elements for Misogyny Identification in Memes

Jason Ravagli
Università degli Studi di Firenze
Florence, Italy
jason.ravagli@stud.unifi.it

Lorenzo Vaiani
Politecnico di Torino
Turin, Italy
lorenzo.vaiani@polito.it

Abstract

Gender discrimination is a serious and widespread problem on social media and online in general. Besides offensive messages, memes are one of the main means of dissemination for such content. With these premises, the MAMI task was proposed at the SemEval-2022, which consists of identifying memes with misogynous characteristics. In this work, we propose a solution to this problem based on Mask R-CNN and VisualBERT that leverages the multimodal nature of the task. Our study focuses on observing how the two sources of data in memes (text and image) and their possible combinations impact performances. Our best result slightly exceeds the higher baseline, but the experiments allowed us to draw important considerations regarding the importance of correctly exploiting the visual information and the relevance of the elements present in the memes images.

1 Introduction

Even though many advances and initiatives have been carried on in the last decades, misogyny and gender discrimination still represent a serious social problem. Comments and posts with hate messages or bad jokes having the female gender as the target are published every day on social networks. Memes are the most popular means of communication on the internet. A meme basically is an image with overlayed text. The combination of image and text should transmit in a clear and direct way a message, which is often funny or ironic. However, many users use them to spread hate or discriminatory messages against certain categories of people. Considering the number of contents that are published online every moment, social network administrators require automatic tools to identify and remove this type of memes. Such tools can be an effective way to fight and stop sexist and misogynous manifestations online.

For the SemEval-2022 it has been proposed the Multimedia Automatic Misogyny Identification (MAMI) task (Fersini et al., 2022), which consists of identifying misogynous memes.

In this work, we present our proposed solution for subtask A of MAMI, which is a binary classification problem where a meme must be categorized as misogynous or non-misogynous. Our solution exploits the multimodal nature of the available data (a meme is represented by a pair image - overlayed text) by using a VisualBERT model together with a Mask R-CNN. Our solution classified 35^{th} out of 69 participants, getting an F1 score 2% higher than the best baseline. Beyond metric values, our work highlights the different importance of various visual elements and how their usage can affect the performances.

2 Task Overview

The purpose of MAMI is identifying misogyny contents and messages within memes through the use of visual and textual information. Organizers set up this task as a competition, providing an annotated dataset and proposing two different sub tasks.

2.1 Dataset

The training dataset contains 10000 memes. Each sample is composed of a pair (image, text) corresponding to the visual and textual information inside the meme.All texts transcripts are already extracted from pictures and available in English language. Data samples are labeled with five different binary tags, each aimed at identifying a different type of hate content: misogyny, objectification, violence, stereotype and shaming. More than one class can characterize a single data sample.

During the final phase of the MAMI challenge, task organizers released a test dataset, made by 1000 elements, which labels were released at the end of the competition only. Therefore, we decided to reserve 25% of the labeled training samples for the validation set to perform our experiments.

In addition to general information, we compute

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 610 - 617

some image-related statistics to analyze and compare all dataset splits. In particular, we use Mask R-CNN (He et al., 2018) classification and bounding-box regression branches to identify which, how many and how large are the most frequent elements within the images. We try two different Mask R-CNN pretraining: the former trained on COCO (Lin et al., 2015) (91 classes) while the latter trained on LVIS version 0.5 (Gupta et al., 2019) (1230 classes).

Computed statistics for both mask R-CNN pretrainings are reported in Table 1. Comparing the two, it is possible to observe how the COCO pretraining allows identifying elements in greater quantity and of greater dimensions than the LVIS one. This is a strange result considering the higher number of classes contained in LVIS. Looking in detail at the detected objects we noticed that the majority of elements identified by the COCO-pretrained network belong to the class "person" ($\sim 63\%$ for train and validation, $\sim 68\%$ for test). This category is not present among those of LVIS, which instead detects as most popular categories some types of clothing and jewelry, such as swim suits, dresses, necklaces and earrings (each appearing at most in $\sim 8\%$ of predicted classes). Moreover, COCO pretraining performs better also in other domains different from photos, such as drawings and cartoons, where LVIS pretraining is often not able to detect and extract any relevant patch. A last noteworthy fact is the dimensional discrepancy between the test and the other splits when using the COCO pretraining. Even though the detected object ratio is lower on the test set, the percentage of covered picture is higher, implying that elements in test images are larger but in less quantity than in the other splits.

2.2 Sub-tasks

Regarding the goals of the challenge, the MAMI task is split up into two sub-tasks:

- Sub-task A consists of a binary classification problem, where each sample/meme must be categorized as misogynous or non-misogynous. The two classes are balanced.

- Sub-task B is a multi-class and multi-label classification problem, where each meme/sample must be assigned labels belonging to 5 different classes: misogynous, stereotype, shaming, objectification, and violence. It is an advanced task since the labels identify more in-depth the type of offensive content and the classes are unbalanced.

Our team worked on a solution for sub-task A only.

3 Related Works

Hate speech detection in text data has been deeply studied in the last few years. State-of-the-art approaches usually apply transformers-based methods achieving impressive results. (Mozafari et al., 2020) for example proposed different hybrid architectures to fine-tune a BERT model (Devlin et al., 2019) for detecting offensive tweets. However, identifying hate content in multimodal data requires correlating visual and textual information and introduces an additional challenge. Multimodal transformers, such as VisualBERT (Li et al., 2019) and ViLBERT (Lu et al., 2019), are currently the state-of-the-art models for these types of problems.

In 2020 Facebook AI organized the Hateful Memes Challenge (Kiela et al., 2021). The competition was similar to MAMI and consisted of categorizing a meme as hateful or non-hateful. The dataset contained memes with various types of hate messages (e.g. targeting an ethnicity, a religion, or the sexual orientation of people). Winning solutions all used ensembles of multimodal transformers-based networks. (Zhu, 2020) won the competition using a complex and task-specific pipeline to extract additional information from the data with which to fine-tune multimodal transformers networks. (Muennighoff, 2020) and (Velioglu and Rose, 2020), who ranked respectively second and third, proposed simpler ensemble methods based on VisualBERT and VilBERT-derived architectures.

Many other works regarding meme analysis for hate speech detection can be found in the literature. Each has its own peculiarities, often related to information extraction techniques, especially from the visual content. Among them, the most used are text removal, object detection, image captioning, and entity recognition, such as in (Deshpande and Mani, 2021), (Pramanick et al., 2021) and (Lee et al., 2021).

4 Method

In this section, we present architectures and preprocessing steps used to address sub-task A of MAMI.

Split	Size	COCO			LVIS		
		Detected Objects	Objects Ratio	Selected Area	Detected Objects	Objects Ratio	Selected Area
Train	7500	21480	2.86	51.55%	13665	1.82	13.20%
Validation	2500	7174	2.87	52.31%	4460	1.78	12.79%
Test	1000	2490	2.49	60.89%	1439	1.44	12.29%

Table 1: dataset statistics related to image content. For each split and for each pretraining version of Mask R-CNN are reported the total number of detected objects, the average number of elements per image and the average percentage of pixels per image contained in all the selected boxes .

First of all, we exploit single modalities separately, through state-of-the-art deep models, to investigate their relevance and the quantity of information they carry on. Subsequently, we combine image and text embeddings to estimate the performance of a multimodal approach.

The strategies mentioned above can be considered as three baseline models in which the information from each modality is retrieved without taking into account the other one and then used both alone or together. Afterward, we adopt early fusion methods to create a more informative embedding of a whole meme, extracting information jointly from both modalities with the same model. This allows to directly obtain a final representation depending on both text and image, thus removing the need for a fusion point between unimodal embeddings.

4.1 Text Encoding

Transformer architectures (Vaswani et al., 2017) are currently the state-of-the-art models in NLP as regards the generation of textual embedding. Among them, BERT (Devlin et al., 2019) is the most renowned sentence encoder, with top-level performances in encoding sentence semantics.

Our text-based baseline model exploits BERT encoder to convert the input sentence into a 768-dimensional embedding vector, obtained with a mean pooling operation over all the output tokens. The classification is accomplished by a single fully connected layer. Both sentence encoder and classification head are trained end-to-end.

4.2 Image Encoding

Convolutional Neural Networks (LeCun et al., 1999) are the type of model most commonly employed in image analysis. CNNs consist of a stack of convolutional layers that extract visual elements from the image, assigning each an appropriate relevance. Among all CNNs, VGG16 (Simonyan and Zisserman, 2015) is one of the most known and used.

Our image-based baseline end-to-end model uses the VGG16 feature extractor along with a multi layer perceptron to obtain a 2048-dimensional embedding vector from the entire original input image. Once again, the classification is accomplished by an additional fully connected layer, as happens for the text, and the two pieces of architecture are trained end-to-end.

4.3 Multimodal Fusion

As a naive multimodal solution, we combine each modality embedding extracted with the previously described techniques. Adopting this implementation we build an end-to-end architecture with two input branches, one for each modality, and a single classification head that takes as input the concatenation of text and image representation vectors, thus setting a late fusion point in the system.

4.4 Multimodal Extraction

Baseline architectures presented in previous subsections suffer from several disadvantages. First and foremost, the choice of fusion point is a critical decision during the construction of multimodal architectures. A late fusion point does not allow information contained in one modality to affect the embedding creation of the others, limiting the influence between modalities only through the back-propagation steps along with the end-to-end architecture. On the other hand, an early fusion has proven to be the best choice for a wide range of other tasks, such as in (Barnum et al., 2020) and (Peinelt et al., 2020).

Another weakness is the usage of the entire input image. The overlayed text in memes is often characterized by a showy style. Since transcriptions are available and typography is unlikely to bring any useful information about hate content,

the text should be removed from the pictures, or simply not considered, to avoid affecting the image representations with noise data.

VisualBERT (Li et al., 2019) is a novel transformer designed to address vision-and-language tasks. It reuses self-attention mechanism to align elements of the input text and regions in the input image. VisualBERT is able to address all the weaknesses highlighted above. Input tokens originate from both text and image concurrently, implicitly setting an early fusion point at the model input stage. Furthermore, image input tokens are typically extracted from the original picture as small patch representations, instead of using the whole figure, allowing to focus on important details and ignoring text, background and noise elements. This architecture has been successfully applied to other tasks involving memes (Velioglu and Rose, 2020).

The criterion adopted to select the input patches is based on Mask R-CNN. The same pretrainings mentioned in the dataset sections are used to retrieve multiple Region Of Interest (ROI) in the picture. Then, if a ROI overcomes a fixed confidence threshold, it is select as input token for VisulBERT.

Similarly to BERT, VisualBERT provides a 768-dimensional vector for each input tokens, based on both image and text information, that is finally fed as input for the fully connected classification layer.

5 Experimental Results

5.1 Experimental Design

We try several configuration of the proposed architecture. During the ROIs identification step we exploit both COCO and LVIS pretraining, feeding the model with the patches coming from both Mask R-CNN versions, both separately and simultaneously. Another setting of our architecture concerns the use of transformer output embeddings as input for the classification head: we feed the final fully connected layer with both the CLS token embedding only and the average of all input token embeddings. During our experiments, one of the major issues we faced was the small dimension of the dataset and, consequently, the risk of overfitting. Thus, we used our validation set, obtained with an hold-out splitting as described in 2.1, to monitor the model performance during training. Then, the test set is evaluated selecting the best model according to the results obtained on validation set. Task organizers designated the F1 score as the official metric for the competition, so we used it for model selection.

Algorithms' configuration. We train our models for 25 epochs to minimize BCE loss with AdamW (Loshchilov and Hutter, 2019), using a batch size of 32 and 10^{-5} as learning rate. Many of the best models were obtained within of the first 10 epochs, so there were no need for longer training sessions. For the sake of reproducibility, the code is publicly available in the project repository [1].

Hardware Settings. Experiments were perfomed on a machine equipped with AMD® Ryzen 9® 3950X CPU, Nvidia® RTX 3090 GPU, and 128 GB of RAM running Ubuntu 21.10.

5.2 Results

Table 2 shows F1 scores on validation and test set of our baselines and proposed methods. We also report the best organizers baseline result, which was released at the end of competition.

As we can see, VisualBERT methods perform slightly better than our baseline approaches. Among them, the unimodal BERT and the multimodal fusion method turn out to be our best baselines on the validation and test sets respectively, while exploiting visual information only gives unsatisfactory results. This leads us to formulate some considerations about the nature of the task. First of all, not in all memes the quantity of relevant information is equally distributed between the two data sources (image and text). Indeed, some memes can be formed by a neutral image but a very offensive text or by a neutral or ironic test and a tacky image. The former can be successfully recognized by a text-only approach, which can manage to classify also the latter if correctly trained on identifying malicious messages in ironic texts. On the contrary, image-only methods can only spot misogyny in memes with explicit offensive images. Furthermore, simple approaches that focus on whole images like our VGG16 will fail to link different objects in the image scene and to understand the context depicted in it. As a result, the network will work properly only if the meme image clearly contains the target of the hate message (e.g. a woman on the foreground).

However, in some misogynous memes the hate message is formed by the combination of an apparently innocent meme and a neutral text. VisualBERT should be capable of merging the two

[1]The GitHub project repository is available at https://github.com/VaianiLorenzo/SemEval-2022-MAMI

sources of information and putting them into context, identifying these subtle cases. Indeed, it significantly outperforms baseline methods in both validation and test set.

Focusing on the VisualBERT-based methods, we found that using the CLS token over the average of the output embeddings for the classification gave better results in general. We can see from Table 2 that all the three pretraining modalities for Mask R-CNN led to similar results on the validation set, with the LVIS one performing slightly better and achieving an F1 score of 0.823. However, when evaluating on the test set, VisualBERT with Mask R-CNN pretrained on the COCO dataset performed the best, with an F1 score of 0.67 and an improvement of 4% over the other two methods and 2% compared with task organizers baseline. According to the results, giving VisualBERT features from both pretrained modalities gave no improvement. The motivation behind this can be the high overlap probability between patches: combining them brings no additional information and introduce redundancy.

The performance gap between the COCO-pretrained model and the LVIS one on the test set compared to the validation can be explained by the statistics described in Section 2.1. The pretraining on COCO allows the Mask R-CNN to detect ROIs containing whole people, that is, it allows to provide to VisualBERT features regarding the potential targets of the misogynous messages. Since more people are found in the test set images, VisualBERT can exploit the visual information in a more effective way in those memes. Some examples of critical types of memes discussed in this section and their analysis can be found in the Appendix A.

Due to the limited number of examples, our models showed signs of overfitting after a few epochs. Using the whole available training set brought no improvements. Hence, we tried some solutions to help the model to generalize better on unseen data. We did a pre-train phase where we applied our VisualBERT to the Hateful Memes task before fine-tuning it on the MAMI one with a lower learning rate. Unfortunately, this did not bring any improvement. The model only converged faster on the MAMI training set, maybe due to a similarity between the data of the two competitions. Afterward, we tried to keep the VisualBERT layers frozen and gradually unfreezing them during training, but also in this case we got no improvements.

Model	**Val F1 score**	**Test F1 score**
Task Baseline	-	0.650
BERT	0.786	0.607
VGG16	0.696	0.571
BERT + VGG16	0.756	0.628
VB (COCO)	0.811	**0.670**
VB (LVIS)	<u>0.823</u>	0.631
VB (LVIS + COCO)	0.818	0.632

Table 2: F1 scores of our VisualBERT-based architectures compared with both our and organizers baselines. Best results obtained for validation and tests splits are underlined and highlighted in bold respectively.

6 Conclusions and Discussion

In this work we demonstrate how leveraging the multimodal nature of data allows to achieve a significant boost in performance when facing tasks involving memes. Moreover, we confirm that an early fusion point between modalities implemented using VisualBERT, the current state-of-the-art architecture for vision-and-language tasks, can lead to satisfactory results to the MAMI task. The reliability of this model allows us to focus on other crucial aspects of the problem, such as the data to use as input to the network. While unimodal transformers successfully identify offensive content in textual data, state-of-the-art computer vision models obtain lower results on this task when analyzing visual information only. Catching the message of a meme requires understanding the context created by the text and the various elements depicted in the image. A deep learning model will struggle to form this context if the visual data are analyzed as a whole. Therefore, we have to first extract the relevant parts from the image and use their features separately as input to our multimodal network. By doing this, VisualBERT is capable of correlating text and visual scene to create the aforementioned context. Our experiments showed that the most relevant elements in the images are people, likely the target or the offender mentioned in the text.

Despite these discoveries, the extraction of information from the pictures remains the main obstacle of this task. As a future work, we plan to thoroughly investigate the content analysis of images. In particular we would like to enrich the preprocessing stage, performing captioning and entity detection of the image, in order to transpose some visual information in textual format.

References

George Barnum, Sabera Talukder, and Yisong Yue. 2020. On the benefits of early fusion in multimodal representation learning.

Tanvi Deshpande and Nitya Mani. 2021. An interpretable approach to hateful meme detection.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Aurora Saibene Giulia Rizzi, 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.

Agrim Gupta, Piotr Dollár, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2018. Mask r-cnn.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes.

Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio. 1999. Object recognition with gradient-based learning. In *Feature Grouping*. Springer.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. *Proceedings of the 29th ACM International Conference on Multimedia*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII*, pages 928–940, Cham. Springer International Publishing.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visiolinguistic models applied to hateful memes.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. Better early than late: Fusing topics with word embeddings for neural question paraphrase identification.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets.

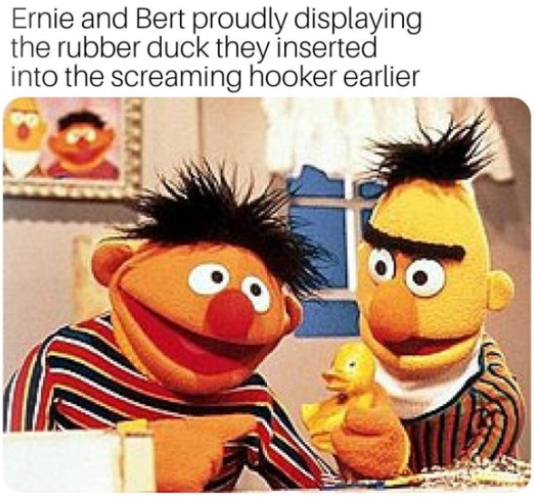
Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution.

A Memes Examples



(a) Meme with a neutral image and offensive text.

(b) Meme with neutral/ironic text and tacky image.

Figure 1: two example of misogynous memes where one of the two modalities, image in (a) and text in (b), contains neutral and /or not useful information according to the task goal. These memes highlight the importance of a multimodal approach.

(a) misogynous meme.

(b) not misogynous meme.

Figure 2: Two memes from test set depicting a woman in the foreground. The image-only baseline model classifies both memes as misogynous, probably due to the presence of a woman as main element in the picture. On the other hand, our best performing model predicts the appropriate label, demonstrating that the text is definitely needed to correctly classify these type of memes.

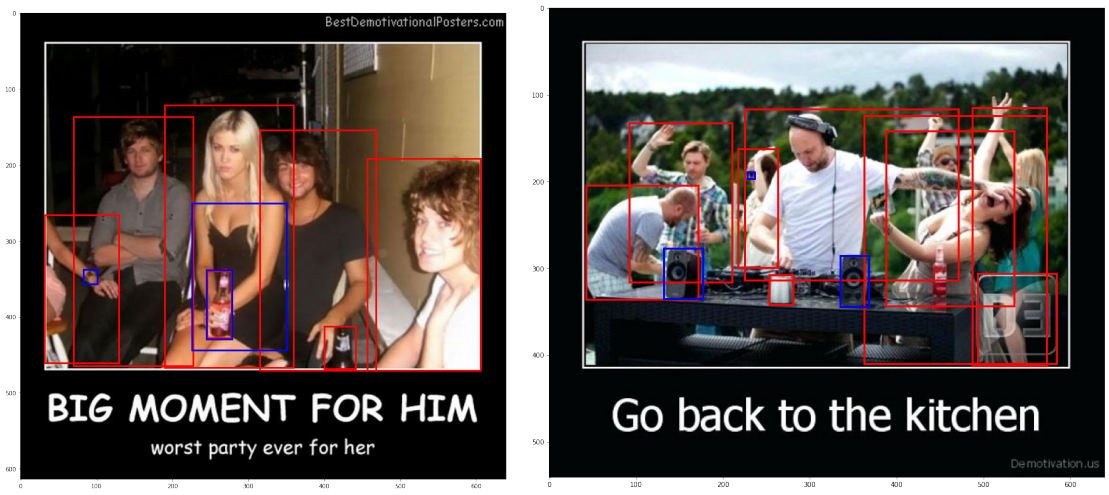


Figure 3: Two misogynous memes depicting several people. The highlighted boxes are the Mask R-CNN ROIs: red ones from the COCO pretraining and blue ones from the LVIS pretraining. As we can see, people are detected only by the former, while the latter retrieves mainly small objects. Moreover, it is possible to notice blue boxes are often contained in red ones, justifying why using COCO and LVIS pretraining jointly there is no improvement in performances.

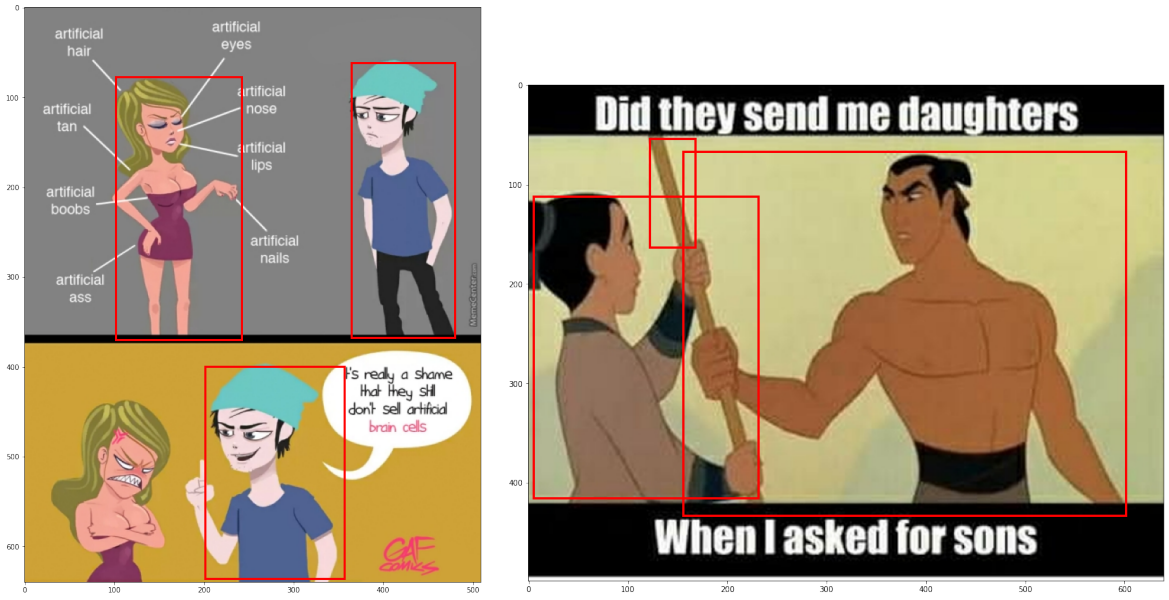


Figure 4: Two misogynous memes with cartoon style. The bounding boxes are all retrieved with Mask-R CNN pretrained on COCO, which is able to detect people also in images other than photos. On the contrary, the LVIS-pretrained model is often not able to identify any details in these kind of pictures.