TIB-VA at SemEval-2022 Task 5: A Multimodal Architecture for the Detection and Classification of Misogynous Memes

Dănilă Mihaela-Alexandra
Manole Patricia-Theodora
Preda Alexandru-Florin

Motivation

The spread of misogynistic content on social media platforms such as Facebook, Twitter, and Instagram has grown significantly, impacting millions of users daily. It is crucial to identify and limit such harmful content to foster a respectful and safe online community.

We wanted to use this multimodal in an app, in order to have a safe platform where misogynistic content is blurred.

Introduction

Misogynistic memes are particularly challenging to detect because they combine text and images.Traditional text-based detection methods are inadequate for this complex, multimodal format.

The research by Hakimov et al. introduces an advanced multimodal architecture that seamlessly integrates text and image analysis.This state-of-the-art approach effectively identifies and classifies misogynistic memes, marking a significant step.forward in combating online hate speech.

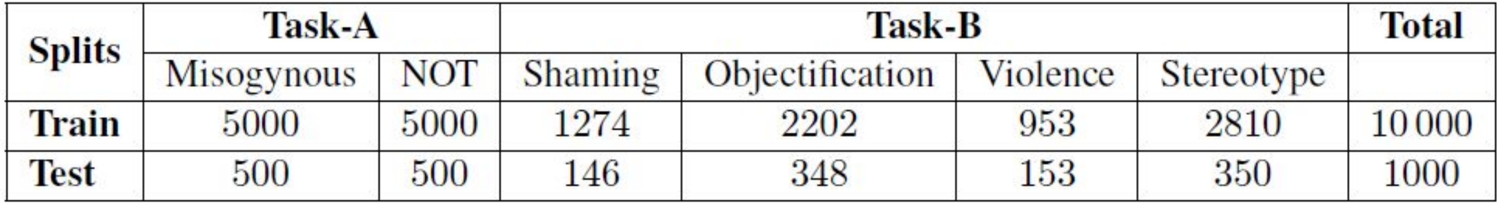
Task and Dataset

The challenge contains 2 tasks:

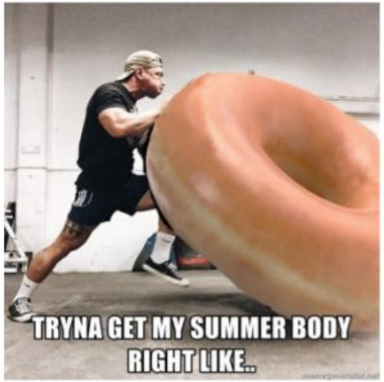
- Task-A: Binary classification of misogyny.
- Task-B: Multi-label classification of misogyny subtypes (stereotype, shaming, objectification, violence)

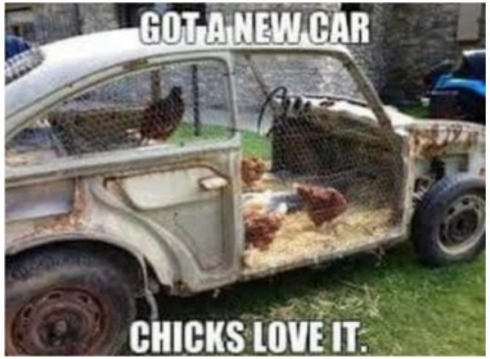
Dataset details- MAMI(Multimedia Automatic Misogyny Identification):

- Training set: 10,000 samples.
- Test set: 1,000 samples.



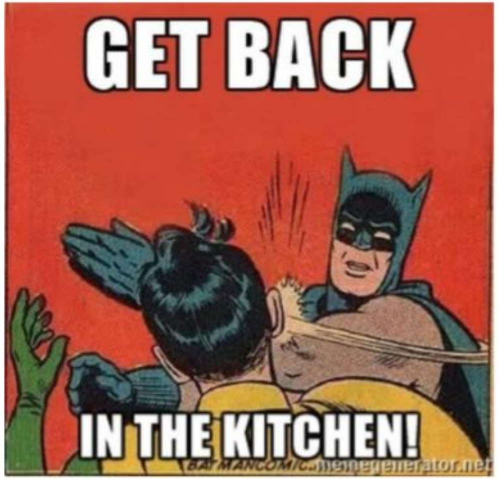
Splits	Task-A		Task-B				Total
	Misogynous	NOT	Shaming	Objectification	Violence	Stereotype	
Train	5000	5000	1274	2202	953	2810	10 000
Test	500	500	146	348	153	350	1000

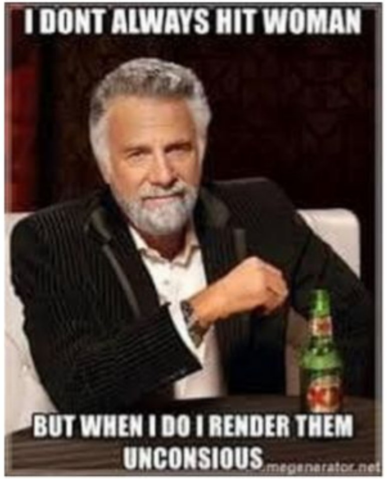




Label: not misogynous

Label: not misogynous



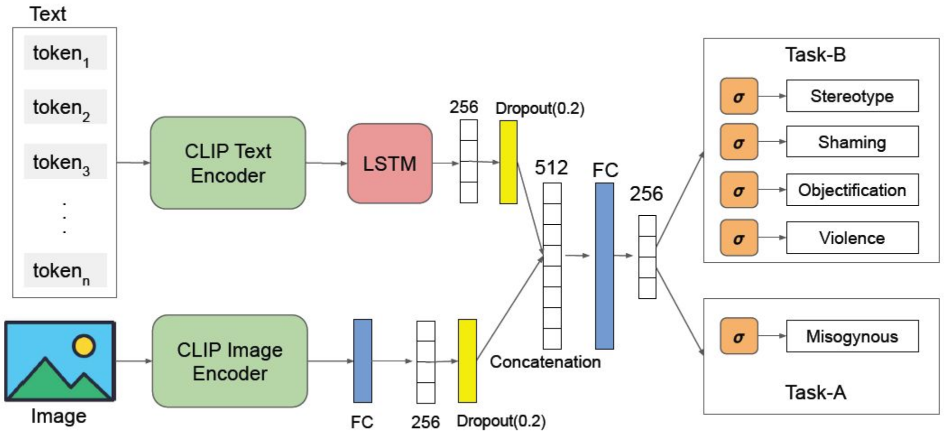


Label: misogynous (stereotype, violence)

Label: misogynous (violence)

Proposed Model Architecture

- **Neural model that utilizes CLIP** for pre-trained multimodal features:
  - Separate encoders for text and image components.
  - LSTM layer for text context representation.
- **Combination** of text and image features via concatenation and fully connected layers.



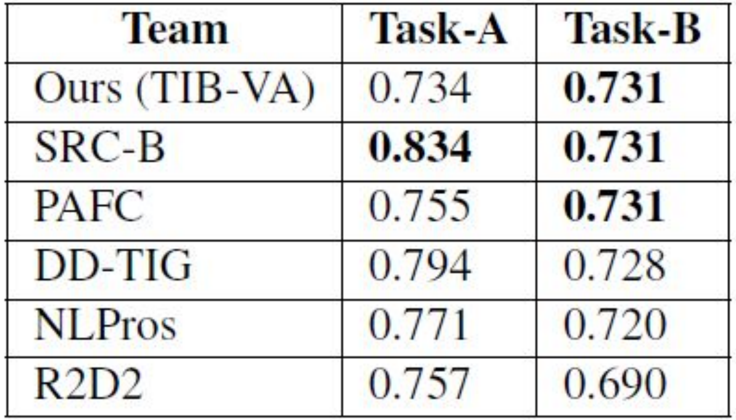
Experimental Setup and results

Training Details:

- **Optimizer**: Adam.
- **Learning Rate**: 1e-4.(decreased by half every 5 epochs)
- **Batch Size**: 64.(max 20 epochs)
- **Validation Split**: 10% of the training data.

Evaluation Metrics: Macro F1 for Task-A and Weighted F1 for Task-B.

Results



Team	**Task-A**	**Task-B**
Ours (TIB-VA)	0.734	**0.731**
SRC-B	**0.834**	**0.731**
PAFC	0.755	**0.731**
DD-TIG	0.794	0.728
NLPros	0.771	0.720
R2D2	0.757	0.690

Strengths

Strengths of the Research

- **Combination**: Effectively combines textual and visual modalities.
- **Pre-trained Models**: Leverages CLIP for robust feature extraction.
- **Reproducibility**: Publicly available codebase.

Weaknesses

Weaknesses and Areas for Improvement

- **Performance on Task-A**: Moderate performance indicates room for optimization.
- **Dataset Limitations**: Reliance on predefined misogyny subtypes may overlook nuanced content.
- **Future Integration**: Potential for integrating additional modalities like audio or metadata.

Future Directions

Impact and Future Directions

- **Online Safety**: Enhancing online safety through scalable hate speech detection systems.
- **Future Work**: Exploring diverse multimodal features that measure different aspects of visual content such as violence, nudity or specific objects and scene-specific content.

Our contribution

Our objective is to design a user interface inspired by a social media page feed, integrating mechanisms to detect misogynistic content.

We aim to allow users to seamlessly use the app while ensuring that misogynistic images are automatically blurred to create a safer online environment.

Our focus is on automatically detecting and blurring misogynistic images.

We strive to maintain a familiar and engaging user experience

Code modifications

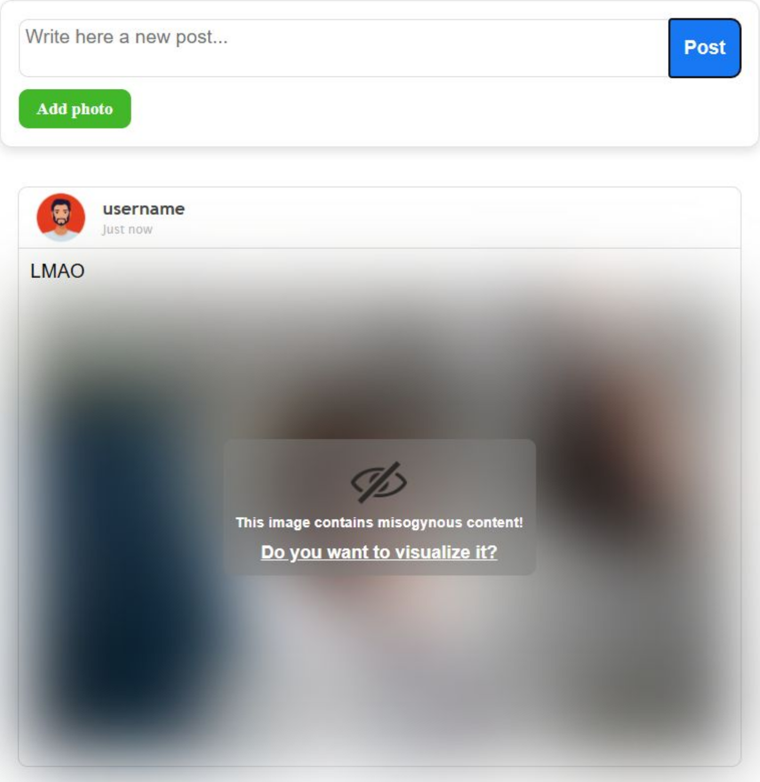
We obtained better results on Task-A using vit14, but our scores on Task-B were not as good.

Ours Results	Task-A	Task-B
ResNet50	0.83472	0.60974
ResNet504	0.84179	0.59460
vit14	**0.86397**	0.62102
vit32	0.84580	0.61527

Integrating the software artefact

- For interaction with the software artefact, we have created a React app that simulates a social media platform where the offensive content is censored and it can be viewed only after acknowledge of the character of that content.
- By creating this interface, we show the utility and great level of usability of the model discussed in the paper.

Feed Page Design



Write here a new post...

Post

Add photo

username

Just now

LMAO

This image contains misogynous content!

Do you want to visualize it?

Application Structure

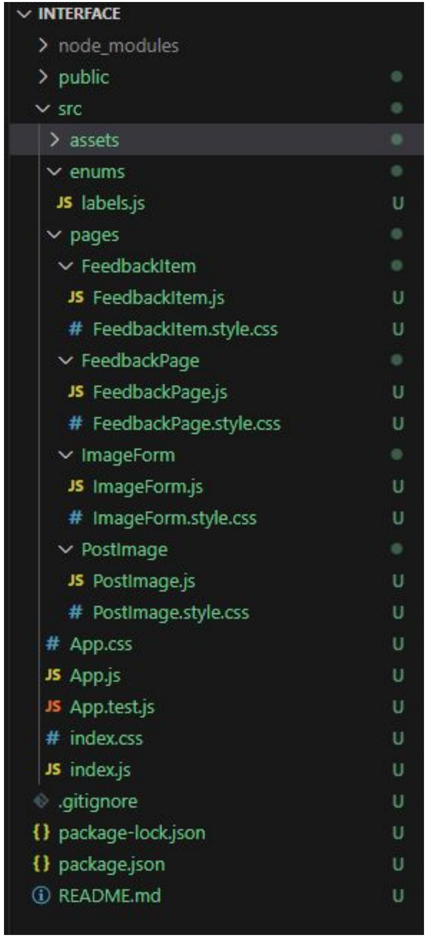
- The application follows a simple structure that contains the following components:

    - the ***FeedbackPage*** components that contains a typical pageof a social media platform where all the new posts are listed

    - The ***PostImage*** form component that contains a basic form with the upload image functionality and a textarea for the descriptionof that image

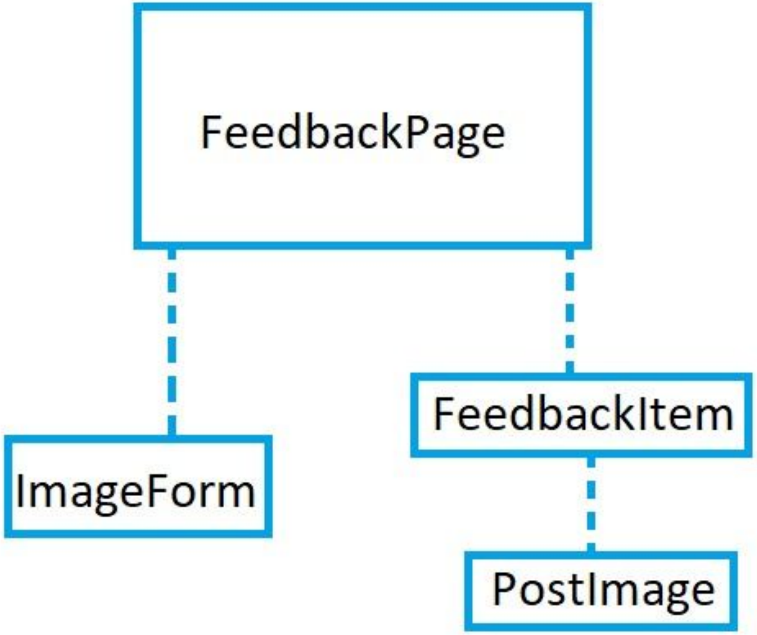
    - The ***FeedbackItem*** component contains a single prototype post that is styled in the typical way; it is dynamically rendered for each uploaded picture

    - The ***PostImage*** which is a presentational component that renders the image blured/unblured depending on the given flags



```
INTERFACE
  > node_modules
  > public                      ●
  v src                         ●
    > assets                    ●
    v enums                     ●
      JS labels.js              U
    v pages                     ●
      v FeedbackItem            ●
        JS FeedbackItem.js      U
        # FeedbackItem.style.css U
      v FeedbackPage            ●
        JS FeedbackPage.js      U
        # FeedbackPage.style.css U
      v ImageForm               ●
        JS ImageForm.js         U
        # ImageForm.style.css   U
      v PostImage               ●
        JS PostImage.js         U
        # PostImage.style.css   U
    # App.css                   U
    JS App.js                   U
    JS App.test.js              U
    # index.css                 U
    JS index.js                 U
  .gitignore                    U
  {} package-lock.json          U
  {} package.json               U
  ⓘ README.md                   U
```

Application Structure



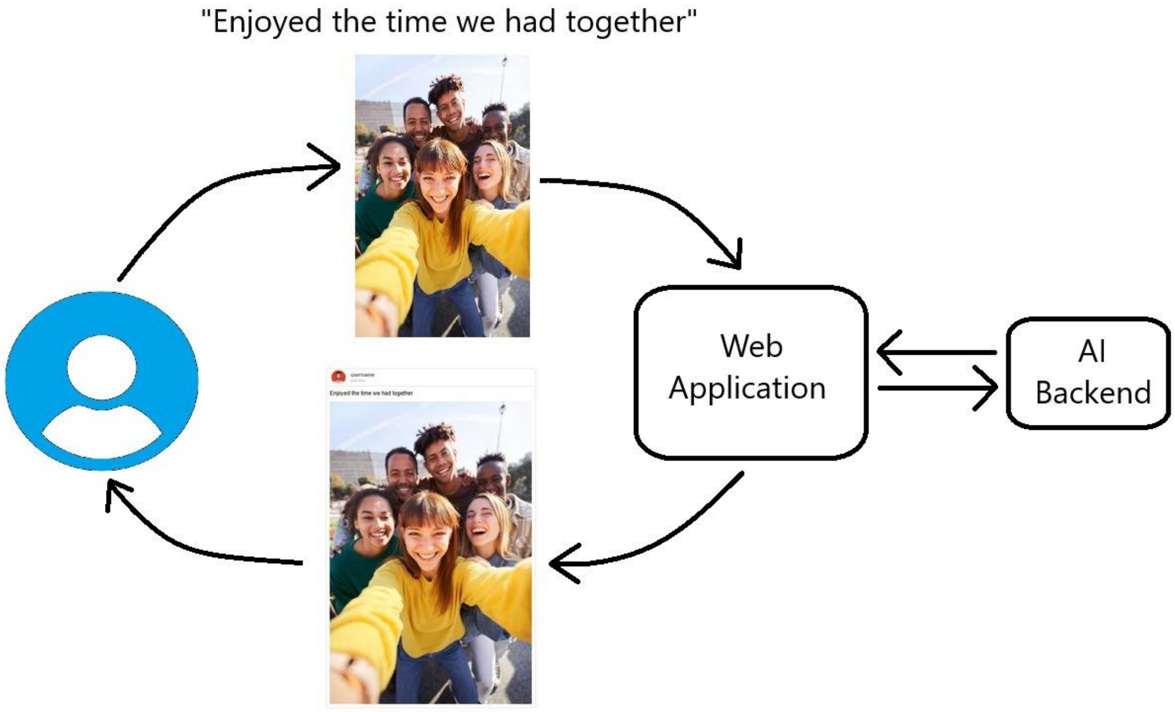
FeedbackPage

ImageForm

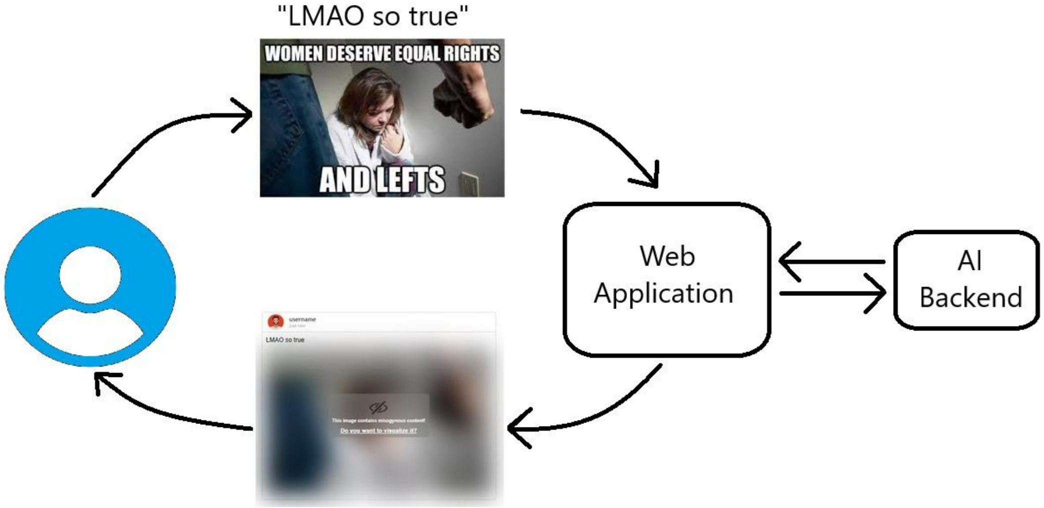
FeedbackItem

PostImage

Misogynistic Image Handling


"Enjoyed the time we had together"
username
Enjoyed the time we had together
Web Application
AI Backend

Misogynistic Image Handling



"LMAO so true"

Web Application

AI Backend

username
Just now

LMAO so true

This image contains misogynous content!

Do you want to visualize it?

Why is AI necessary in our application?

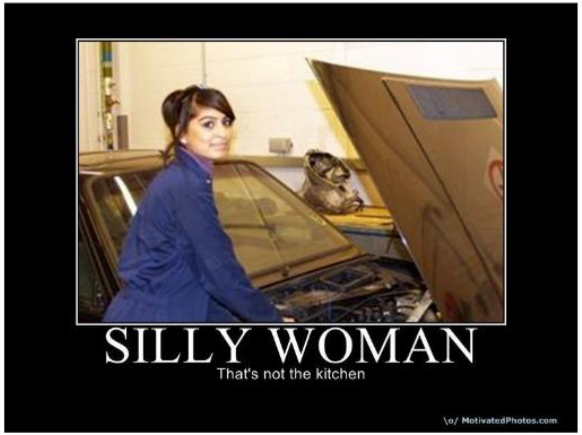
- Other ways of detecting offensive content are inferior.
- Non-AI approaches can iterate through the text and look for keywords that indicate that the content is offensive. However, they can not relate the description to the image and interpret the relationships between the parts of speech, which makes them inferior.

Why is AI necessary in our application?

Search on the image caption for key words in offensive language

Counter-example →

"So true, it's a better place for you"

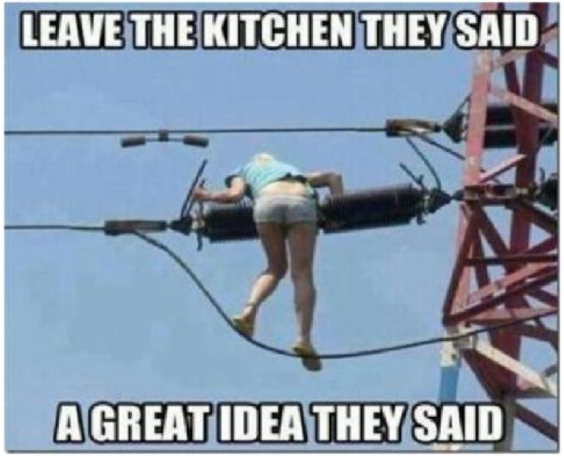


Why is AI necessary in our application?

Search in the image for mysogynistic key words such as the co-presence of the word "women" with words like "dumb", "silly" and so on

Counter-example

Women don't belong outside



Why is AI necessary in our application?

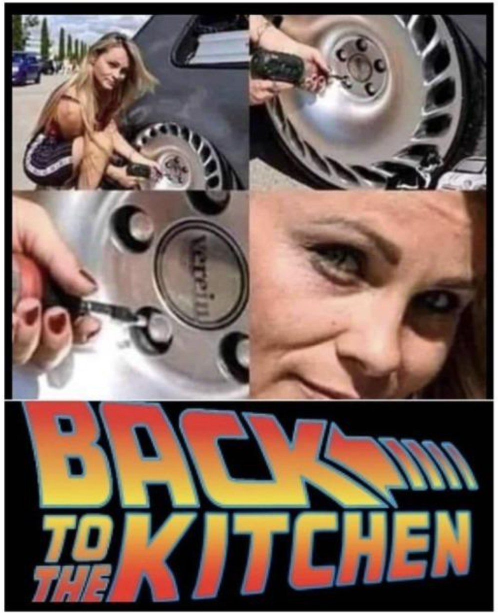
Search in the image for mysogynistic key words such as the co-presence of the word "women" with words like "dumb", "silly" and so on

&

Search on the image caption for key words in offensive language

Counter-example →

"Can't wait for Back to the Kitchen"



Future work

Language and Cultural Adaptability:

Expanding the app’s capabilities to detect misogynistic content across multiple languages and cultural contexts will make it more globally applicable and inclusive.

Expand Accessibility:

Develop mobile and desktop versions of the app for wider user reach.