INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

FLLM 2025
The 3rd International Conference on Foundation and Large Language Models

BJÖRN WINTERLEITNER, TIM SCHLIPPE & KRISTINA SCHAAFF

INVESTIGATING LARGE LANGUAGE MODELS FOR THE DETECTION OF CYBERBULLYING

Vienna, Austria
November 25, 2025

CONTENT

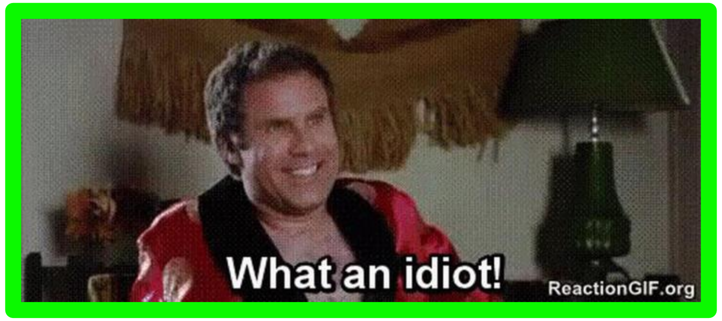
INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

1

INTRODUCTION

MOTIVATION

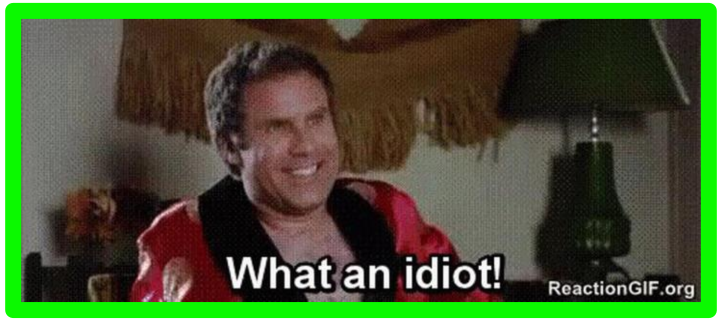
INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES



~50% OF TEENAGERS
HAVE EXPERIENCE WITH CYBERBULLYING *(Vogels, 2022)*

MOTIVATION

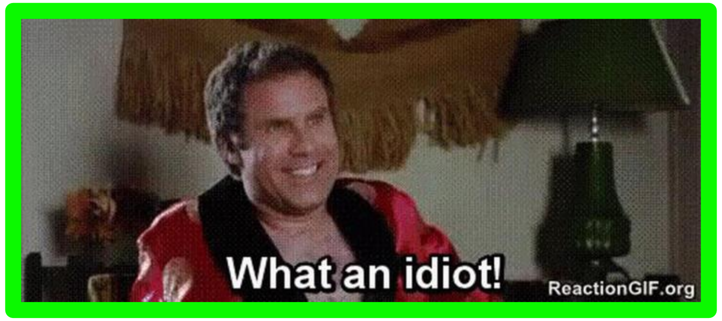
INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES



~50% OF TEENAGERS HAVE EXPERIENCE WITH CYBERBULLYING *(Vogels, 2022)*

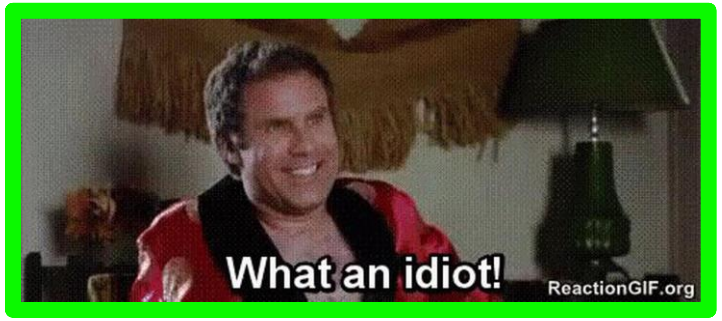
TROUBLE SLEEPING, DIFFICULTY CONCENTRATING, OR FEELINGS OF ANXIETY

MOTIVATION



CYBERBULLYING CAN BE MULTIMODAL! (*Vogels, 2022*)

MOTIVATION

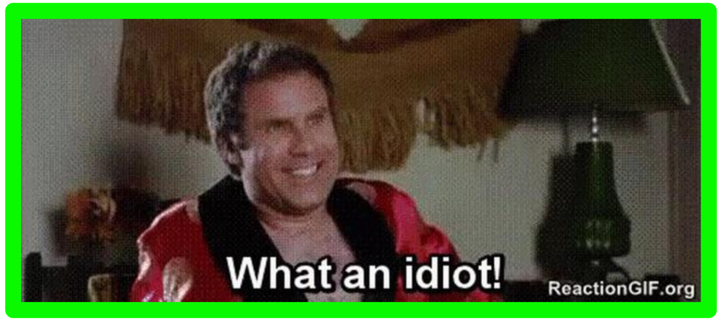


CYBERBULLYING CAN BE MULTIMODAL! *(Vogels, 2022)*

TEXT, IMAGES, VIDEOS

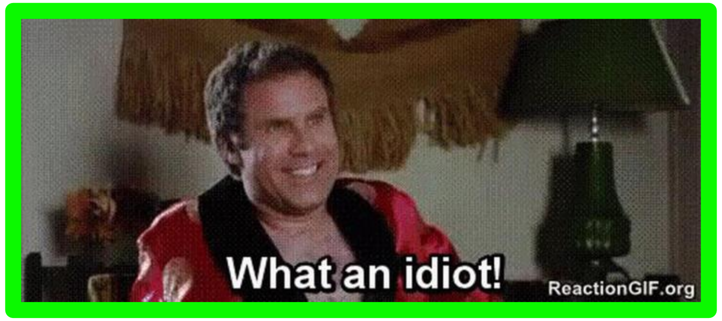
MOTIVATION

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES



CHALLENGE:
TOO MUCH CONTENT FOR MANUAL DETECTION (*Vogels, 2022*)

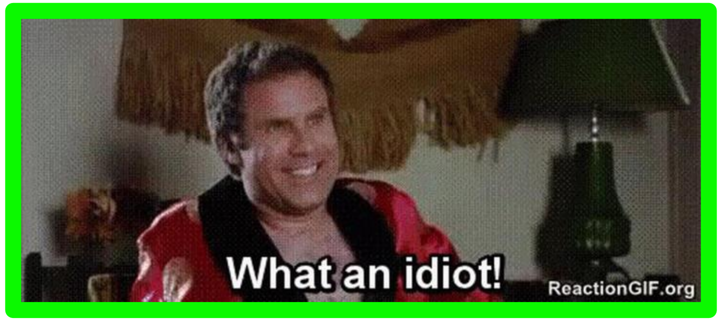
MOTIVATION



CHALLENGE:
TOO MUCH CONTENT FOR MANUAL DETECTION *(Vogels, 2022)*

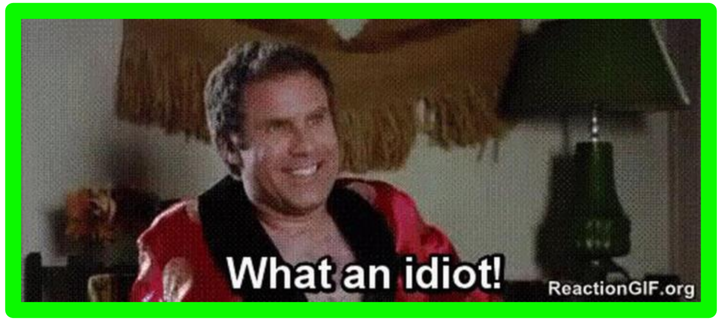
TOO MUCH ONLINE CONTENT TO MANUALLY MONITOR + ADDRESS EVERY CYBERBULLYING POST

MOTIVATION



SOLUTION:
AI TO DETECT CYBERBULLYING AUTOMATICALLY

MOTIVATION



SOLUTION: AI TO DETECT CYBERBULLYING AUTOMATICALLY

AUTOMATICALLY DETECT CYBERBULLYING AND ADDRESS IT MORE EFFECTIVELY AND AT SCALE

2

RELATED WORK

RELATED WORK

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

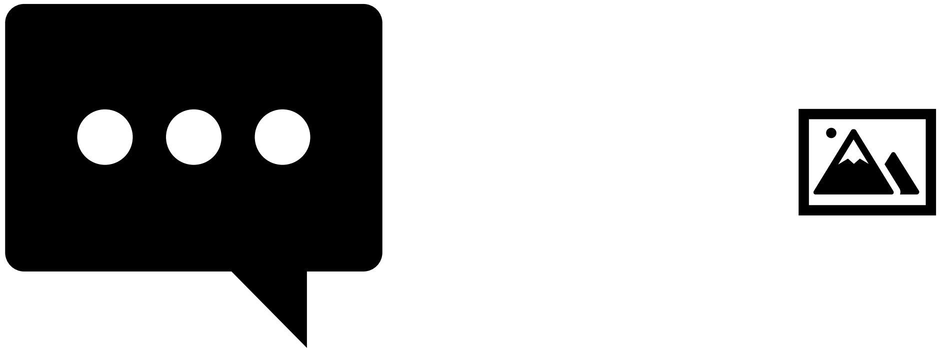
TRADITIONAL +
TRANSFORMERS vs. LLMs

e.g., Nina-Gutiérrez et al. (2024), Shekhar (2024), Ramos at al. (2024)

RESEARCH ON CYBERBULLYING WITH LLMS LIMITED

RELATED WORK

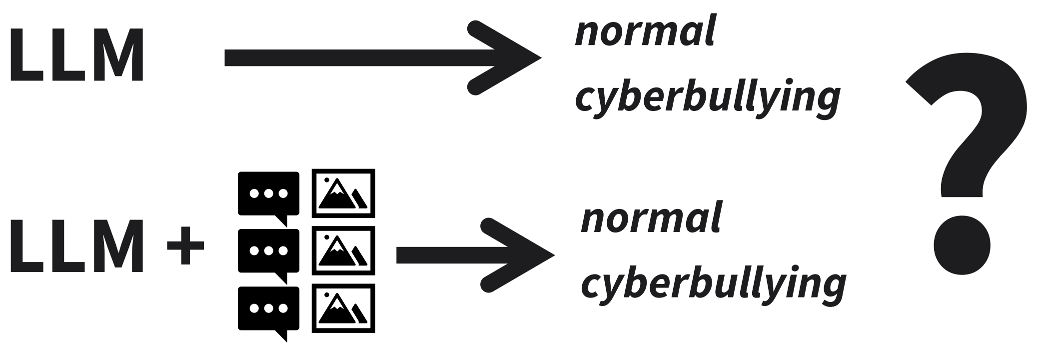
INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES



e.g., Batani et al. (2022), Islam et al. (2020), Raj et al. (2021), Alkomah et al. (2022), Atoum (2020), Van Bruwaene et al. (2020)

TEXT FEATURES ARE MORE INVESTIGATED

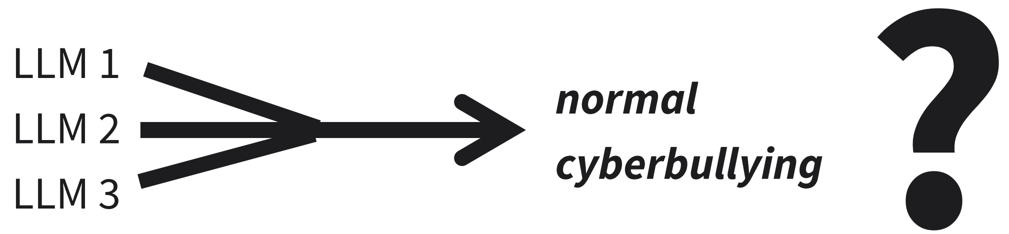
RELATED WORK



e.g., Bauer et al. (2024), Bucciarelli at al. (2024)

ZERO-SHOT VS. FEW-SHOT
ONLY FOR SENTIMENT ANALYSIS

RELATED WORK



e.g., Gomez et al. (2019), Mahlangu & Tu (2019), Yadav et al. (2020), Mahat (2021), Roy et al. (2020), Bharti et al. (2021)

LLM COMBINATION REMAIN UNEXPLORED

3

EXPERIMENTAL SETUP

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

EXPERIMENTAL SETUP: LLMS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Model	Provider	Context Window	Max Images	API Access
GPT-4o	OpenAI	128k	16	Paid
GPT-4o mini	OpenAI	128k	16	Paid
Gemini Flash-1.5	Google	1,000k	3	Free
Claude 3.5 Sonnet	Anthropic	200k	8	Paid
Pixtral-12b	Mistral	n.a.	8	Free

EXPERIMENTAL SETUP: LLMS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Model	Provider	Context Window	Max Images	API Access
GPT-4o	OpenAI	128k	16	Paid
GPT-4o mini	OpenAI	128k	16	Paid
Gemini Flash-1.5	Google	1,000k	3	Free
Claude 3.5 Sonnet	Anthropic	200k	8	Paid
Pixtral-12b	Mistral	n.a.	8	Free

TEXT + IMAGE, PRIORITIZED FREE ACCESS

EXPERIMENTAL SETUP: PROMPT FOR ZERO-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

You are an AI assistant tasked with identifying cyberbullying in social media posts.

Your task is to analyze the following post and determine if it is *normal* or *offensive*. *offensive* posts are those that demean, mock, discriminate against, or insult a person or group of people. *normal* posts do not contain *offensive* content.

Based solely on the content provided, classify the post as either *normal* or *offensive*. Respond with just one word: either *normal* or *offensive*.

Post: [TEXT and/or IMAGE]

ZERO-SHOT

EXPERIMENTAL SETUP: PROMPT FOR FEW-SHOT

You are an AI assistant tasked with identifying cyberbullying in social media posts.

Your task is to analyze the following post and determine if it is *normal* or *offensive*. *offensive* posts are those that demean, mock, discriminate against, or insult a person or group of people. *normal* posts do not contain *offensive* content.

Below are examples of posts and their classifications:
Example 1 (*normal*): [TEXT and/or IMAGE]
Example 2 (*offensive*): [TEXT and/or IMAGE]
[Examples 3-6 follow the same pattern]

Now, analyze the following post and determine if it is *normal* or *offensive*. Respond with just one word: either *normal* or *offensive*.

Post: [TEXT and/or IMAGE]

FEW-SHOT

EXPERIMENTAL SETUP: DATA

506 posts

(*Toraman et al., 2022*)

6 few-shot
- 3 *normal*
- 3 *offensive*

500 test
- 250 *normal*
- 250 *offensive*

EXPERIMENTAL SETUP: METRIC + MODALITIES

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

F1 SCORE

Text-only

Image-only

Text + Image

4

RESULTS

RESULTS: ZERO-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39

RESULTS: ZERO-SHOT

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	$\boxed{77.06}$
	Pixtral-12b	84.39

RESULTS: ZERO-SHOT

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39

RESULTS: ZERO-SHOT

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39

STRONG TEXT COMPREHENSION CAPABILITIES

RESULTS: ZERO-SHOT

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**

RESULTS: ZERO-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**

< 50%

RANDOM F1 = 50%

RESULTS: ZERO-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**

< 50%

RANDOM F1 = 50%

RESULTS: ZERO-SHOT

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**

< 50%

RANDOM F1 = 50%

FUNDAMENTAL LIMITATION IN VISUAL PROCESSING

RESULTS: ZERO-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	$\boxed{\textbf{50.51}}$

< 50%

RANDOM F1 = 50%

HUMAN F1 = 65%

DIFFICULT FOR LLMS AND FOR HUMANS

RESULTS: ZERO-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**
Text + Image	Claude Sonnet 3.5	79.74
	Gemini Flash 1.5	77.15
	GPT-4o mini	79.48
	GPT-4o	**81.89**
	Pixtral-12b	72.88

RESULTS: ZERO-SHOT

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**
Text + Image	Claude Sonnet 3.5	79.74
	Gemini Flash 1.5	77.15
	GPT-4o mini	79.48
	GPT-4o	**81.89**
	Pixtral-12b	72.88

Ø 82.96%

Ø 78.22%

RESULTS: ZERO-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	$\boxed{77.06}$
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**
Text + Image	Claude Sonnet 3.5	79.74
	Gemini Flash 1.5	77.15
	GPT-4o mini	79.48
	GPT-4o	$\boxed{\mathbf{81.89}}$
	Pixtral-12b	72.88

Ø 82.96%

Ø 78.22%

RESULTS: ZERO-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	$\boxed{77.06}$
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**
Text + Image	Claude Sonnet 3.5	79.74
	Gemini Flash 1.5	77.15
	GPT-4o mini	79.48
	GPT-4o	$\boxed{\mathbf{81.89}}$
	Pixtral-12b	72.88

Ø 82.96%

Ø 78.22%

IMAGES OFTEN DO NOT PROVIDE ADDITIONAL VALUABLE INFORMATION

RESULTS: FEW-SHOT

Content	Model	F1 (%)	F1 (%)
Text-only	Claude Sonnet 3.5	82.52	73.17
	Gemini Flash 1.5	**85.64**	54.74
	GPT-4o mini	85.17	66.00
	GPT-4o	77.06	**78.12**
	Pixtral-12b	84.39	75.51
Image-only	Claude Sonnet 3.5	43.32	49.06
	Gemini Flash 1.5	42.81	41.66
	GPT-4o mini	47.26	49.34
	GPT-4o	45.78	**51.35**
	Pixtral-12b	**50.51**	36.08
Text + Image	Claude Sonnet 3.5	79.74	70.64
	Gemini Flash 1.5	77.15	62.15
	GPT-4o mini	79.48	70.18
	GPT-4o	**81.89**	**78.33**
	Pixtral-12b	72.88	70.58

RESULTS: FEW-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)	F1 (%)
Text-only	Claude Sonnet 3.5	82.52	73.17
	Gemini Flash 1.5	**85.64**	54.74
	GPT-4o mini	85.17	66.00
	GPT-4o	77.06	**78.12**
	Pixtral-12b	84.39	75.51
Image-only	Claude Sonnet 3.5	43.32	49.06
	Gemini Flash 1.5	42.81	41.66
	GPT-4o mini	47.26	49.34
	GPT-4o	45.78	**51.35**
	Pixtral-12b	**50.51**	36.08
Text + Image	Claude Sonnet 3.5	79.74	70.64
	Gemini Flash 1.5	77.15	62.15
	GPT-4o mini	79.48	70.18
	GPT-4o	**81.89**	**78.33**
	Pixtral-12b	72.88	70.58

FEW-SHOT DID NOT CONSISTENTLY IMPROVE F1 OVER ZERO-SHOT

RESULTS: FEW-SHOT

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**
Text + Image	Claude Sonnet 3.5	79.74
	Gemini Flash 1.5	77.15
	GPT-4o mini	79.48
	GPT-4o	**81.89**
	Pixtral-12b	72.88

Ø 82.96%

F1 (%)
73.17
54.74
66.00
78.12
75.51

Ø 69.50%

RESULTS: FEW-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**
Text + Image	Claude Sonnet 3.5	79.74
	Gemini Flash 1.5	77.15
	GPT-4o mini	79.48
	GPT-4o	**81.89**
	Pixtral-12b	72.88

Ø 82.96%

F1 (%)
73.17
54.74
66.00
78.12
75.51

Ø 69.50%

RESULTS: FEW-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)	F1 (%)
Text-only	Claude Sonnet 3.5	82.52	73.17
	Gemini Flash 1.5	**85.64**	54.74
	GPT-4o mini	85.17	66.00
	GPT-4o	77.06	**78.12**
	Pixtral-12b	84.39	75.51
Image-only	Claude Sonnet 3.5	43.32	49.06
	Gemini Flash 1.5	42.81	41.66
	GPT-4o mini	47.26	49.34
	GPT-4o	45.78	**51.35**
	Pixtral-12b	**50.51**	36.08
Text + Image	Claude Sonnet 3.5	79.74	
	Gemini Flash 1.5	77.15	
	GPT-4o mini	79.48	
	GPT-4o	**81.89**	
	Pixtral-12b	72.88	

Ø 45.93%

Ø 45.49%

RESULTS: FEW-SHOT

Content	Model	F1 (%)
Text-only	Claude Sonnet 3.5	82.52
	Gemini Flash 1.5	**85.64**
	GPT-4o mini	85.17
	GPT-4o	77.06
	Pixtral-12b	84.39
Image-only	Claude Sonnet 3.5	43.32
	Gemini Flash 1.5	42.81
	GPT-4o mini	47.26
	GPT-4o	45.78
	Pixtral-12b	**50.51**
Text + Image	Claude Sonnet 3.5	79.74
	Gemini Flash 1.5	77.15
	GPT-4o mini	79.48
	GPT-4o	**81.89**
	Pixtral-12b	72.88

Ø 45.93%

F1 (%)
73.17
54.74
66.00
78.12
75.51
49.06
41.66
49.34
51.35
36.08

Ø 45.49%

RANDOM F1 = 50%

HUMAN F1 = 65%

RESULTS: FEW-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)	F1 (%)
Text-only	Claude Sonnet 3.5	82.52	73.17
	Gemini Flash 1.5	**85.64**	54.74
	GPT-4o mini	85.17	66.00
	GPT-4o	77.06	**78.12**
	Pixtral-12b	84.39	75.51
Image-only	Claude Sonnet 3.5	43.32	49.06
	Gemini Flash 1.5	42.81	41.66
	GPT-4o mini	47.26	49.34
	GPT-4o	45.78	**51.35**
	Pixtral-12b	**50.51**	36.08
Text + Image	Claude Sonnet 3.5	79.74	70.64
	Gemini Flash 1.5	77.15	62.15
	GPT-4o mini	79.48	70.18
	GPT-4o	**81.89**	**78.33**
	Pixtral-12b	72.88	70.58

Ø 69.50%

Ø 78.22%

Ø 70.37%

RESULTS: FEW-SHOT

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

Content	Model	F1 (%)	F1 (%)
Text-only	Claude Sonnet 3.5	82.52	73.17
	Gemini Flash 1.5	**85.64**	54.74
	GPT-4o mini	85.17	66.00
	GPT-4o	77.06	**78.12**
	Pixtral-12b	84.39	75.51
Image-only	Claude Sonnet 3.5	43.32	49.06
	Gemini Flash 1.5	42.81	41.66
	GPT-4o mini	47.26	49.34
	GPT-4o	45.78	**51.35**
	Pixtral-12b	**50.51**	36.08
Text + Image	Claude Sonnet 3.5	79.74	70.64
	Gemini Flash 1.5	77.15	62.15
	GPT-4o mini	79.48	70.18
	GPT-4o	**81.89**	**78.33**
	Pixtral-12b	72.88	70.58

Ø 69.50%

Ø 78.22%

Ø 70.37%

INTERPRETATION: ZERO-SHOT VS. FEW-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Content	Model	F1 (%)	F1 (%)
Text-only	Claude Sonnet 3.5	82.52	73.17
	Gemini Flash 1.5	**85.64**	54.74
	GPT-4o mini	85.17	66.00
	GPT-4o	77.06	**78.12**
	Pixtral-12b	84.39	75.51
Image-only	Claude Sonnet 3.5	43.32	49.06
	Gemini Flash 1.5	42.81	41.66
	GPT-4o mini	47.26	49.34
	GPT-4o	45.78	**51.35**
	Pixtral-12b	**50.51**	36.08
Text + Image	Claude Sonnet 3.5	79.74	70.64
	Gemini Flash 1.5	77.15	62.15
	GPT-4o mini	79.48	70.18
	GPT-4o	**81.89**	**78.33**
	Pixtral-12b	72.88	70.58

INTERPRETATION: ZERO-SHOT VS. FEW-SHOT

‘KNOWLEDGE PRIORS’ EFFECT: PRE-TRAINED OVER FEW-SHOT

(Chochlakis et al., 2024)

	GPT-4o	77.06	**78.12**
	Pixtral-12b	84.39	75.51
Image-only	Claude Sonnet 3.5	43.32	49.06
	Gemini Flash 1.5	42.81	41.66
	GPT-4o mini	47.26	49.34
	GPT-4o	45.78	**51.35**
	Pixtral-12b	**50.51**	36.08
Text + Image	Claude Sonnet 3.5	79.74	70.64
	Gemini Flash 1.5	77.15	62.15
	GPT-4o mini	79.48	70.18
	GPT-4o	**81.89**	**78.33**
	Pixtral-12b	72.88	70.58

INTERPRETATION: ZERO-SHOT VS. FEW-SHOT

‘KNOWLEDGE PRIORS’ EFFECT: PRE-TRAINED OVER FEW-SHOT

(Chochlakis et al., 2024)

FEW-SHOT EXAMPLES MAY NOT BE REPRESENTATIVE ENOUGH

	GPT-4o	77.06	**78.12**
	GPT-4o	45.78	**51.35**
	Pixtral-12b	**50.51**	36.08
Text + Image	Claude Sonnet 3.5	79.74	70.64
	Gemini Flash 1.5	77.15	62.15
	GPT-4o mini	79.48	70.18
	GPT-4o	**81.89**	**78.33**
	Pixtral-12b	72.88	70.58

INTERPRETATION: ZERO-SHOT VS. FEW-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

‘KNOWLEDGE PRIORS’ EFFECT: PRE-TRAINED OVER FEW-SHOT

(Chochlakis et al., 2024)

GPT-4o	77.06	**78.12**

FEW-SHOT EXAMPLES MAY NOT BE REPRESENTATIVE ENOUGH

GPT-4o	45.78	**51.35**

6 EXAMPLES MAY BE INSUFFICIENT TO LEARN

GPT-4o	**81.89**	**78.33**
Pixtral-12b	72.88	70.58

INTERPRETATION: ZERO-SHOT VS. FEW-SHOT

‘KNOWLEDGE PRIORS’ EFFECT: PRE-TRAINED OVER FEW-SHOT

(Chochlakis et al., 2024)

GPT-4o	77.06	**78.12**

FEW-SHOT EXAMPLES MAY NOT BE REPRESENTATIVE ENOUGH

GPT-4o	45.78	**51.35**

6 EXAMPLES MAY BE INSUFFICIENT TO LEARN

GPT-4o	**81.89**	**78.33**
Pixtral-12b	72.88	70.58

OVERFITTING: LLMS MAY BE TOO SENSITIVE TO EXAMPLES

RESULTS: LLM FUSION WITH MAJORITY VOTING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Method	Content	All 5	Top 3	Best
zero-shot	Text	84.77	**87.59**	85.64
	Image	43.90	46.38	50.51
	Text + Image	81.95	81.43	81.89

RESULTS: LLM FUSION WITH MAJORITY VOTING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Method	Content	All 5	Top 3	Best
zero-shot	Text	84.77	**87.59**	85.64
	Image	43.90	46.38	50.51
	Text + Image	81.95	81.43	81.89

RESULTS: LLM FUSION WITH MAJORITY VOTING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Method	Content	All 5	Top 3	Best	
zero-shot	Text	84.77	**87.59**	85.64	
	Image	43.90	46.38	50.51	PIXTRAL-12B
	Text + Image	81.95	81.43	81.89	

RESULTS: LLM FUSION WITH MAJORITY VOTING

Method	Content	All 5	Top 3	Best
zero-shot	Text	84.77	**87.59**	85.64
	Image	43.90	46.38	50.51
	Text + Image	$\boxed{81.95}$	81.43	$\boxed{81.89}$

RESULTS: LLM FUSION WITH MAJORITY VOTING

Method	Content	All 5	Top 3	Best
zero-shot	Text	84.77	**87.59**	85.64
	Image	43.90	46.38	50.51
	Text + Image	81.95	81.43	81.89

RESULTS: LLM FUSION WITH MAJORITY VOTING

Method	Content	All 5	Top 3	Best
zero-shot	Text	84.77	**87.59**	85.64
	Image	43.90	46.38	50.51
	Text + Image	81.95	81.43	81.89
few-shot	Text	71.73	77.13	78.12
	Image	47.19	51.59	51.35
	Text + Image	72.01	74.22	**78.33**

RESULTS: LLM FUSION WITH MAJORITY VOTING

Method	Content	All 5	Top 3	Best
zero-shot	Text	84.77	**87.59**	85.64
	Image	43.90	46.38	50.51
	Text + Image	81.95	81.43	81.89
few-shot	Text	71.73	77.13	78.12
	Image	47.19	51.59	51.35
	Text + Image	72.01	74.22	**78.33**
global top 3	Cross-method & modality		$\boxed{83.47}$	**85.64**

GEMINI FLASH 1.5 WITH ZERO-SHOT + TEXT: 85.64%

GPT-4O MINI WITH ZERO-SHOT + TEXT: 85.17%

GPT-4O MINI WITH ZERO-SHOT + TEXT+IMAGE: 85.89%

RESULTS: LLM FUSION WITH MAJORITY VOTING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Method	Content	All 5	Top 3	Best
zero-shot	Text	84.77	**87.59**	85.64
	Image	43.90	46.38	50.51
	Text + Image	81.95	81.43	81.89
few-shot	Text	71.73	77.13	78.12
	Image	47.19	51.59	51.35
	Text + Image	72.01	74.22	**78.33**
global top 3	Cross-method & modality		$\boxed{83.47}$	**85.64**

GEMINI FLASH 1.5 WITH ZERO-SHOT + TEXT: 85.64%

GPT-4O MINI WITH ZERO-SHOT + TEXT: 85.17%

GPT-4O MINI WITH ZERO-SHOT + TEXT+IMAGE: 85.89%

MAJORITY VOTING NOT A UNIVERSAL SOLUTION

6

CONCLUSION AND FUTURE WORK

CONCLUSION AND FUTURE WORK

Conclusion

- **LLMs perform very well on text-based cyberbullying detection**
  - ➔ achieving up to 85.64% F1 in zero-shot settings,
  - ➔ but struggle with image-only inputs

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- **LLMs perform very well on text-based cyberbullying detection**
  ➔ achieving up to 85.64% F1 in zero-shot settings,
  ➔ but struggle with image-only inputs
- **Multimodal (text+image) inputs did not outperform text-only**
  ➔ suggesting visual cues are inherently difficult to interpret for both humans and LLM

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- **LLMs perform very well on text-based cyberbullying detection**
  - ➔ achieving up to 85.64% F1 in zero-shot settings,
  - ➔ but struggle with image-only inputs
- **Multimodal (text+image) inputs did not outperform text-only**
  - ➔ suggesting visual cues are inherently difficult to interpret for both humans and AI
- **Few-shot learning did not improve performance**
  - ➔ with most LLMs performing worse than in *zero-shot*;
  - ➔ only GPT-4o showed slight gains

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- **LLMs perform very well on text-based cyberbullying detection**
  - ➔ achieving up to 85.64% F1 in zero-shot settings,
  - ➔ but struggle with image-only inputs
- **Multimodal (text+image) inputs did not outperform text-only**
  - ➔ suggesting visual cues are inherently difficult to interpret for both humans and AI
- **Few-shot learning did not improve performance**
  - ➔ with most LLMs performing worse than in *zero-shot*;
  - ➔ only GPT-4o showed slight gains
- **Ensembling via majority voting showed limited value**
  - ➔ with only top 3 majority voting resulting in small improvements

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- **LLMs perform very well on text-based cyberbullying detection**
  ➔ achieving up to 85.64% F1 in zero-shot settings,
  ➔ but struggle with image-only inputs
- **Multimodal (text+image) inputs did not outperform text-only**
  ➔ suggesting visual cues are inherently difficult to interpret for both humans and AI
- **Few-shot learning did not improve performance**
  ➔ with most LLMs performing worse than in *zero-shot*;
  ➔ only GPT-4o showed slight gains
- **Ensembling via majority voting showed limited value**
  ➔ with only selective voting among the top 3 models yielding small improvements

Future Work

- **Explore advanced ensemble or multi-agent methods**
  ➔ Move beyond majority voting to collaborative reasoning between LLMs

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- **LLMs perform very well on text-based cyberbullying detection**
  ➔ achieving up to 85.64% F1 in zero-shot settings,
  ➔ but struggle with image-only inputs
- **Multimodal (text+image) inputs did not outperform text-only**
  ➔ suggesting visual cues are inherently difficult to interpret for both humans and AI
- **Few-shot learning did not improve performance**
  ➔ with most LLMs performing worse than in *zero-shot*;
  ➔ only GPT-4o showed slight gains
- **Ensembling via majority voting showed limited value**
  ➔ with only selective voting among the top 3 models yielding small improvements

Future Work

- **Explore advanced ensemble or multi-agent methods**
  ➔ Move beyond majority voting to collaborative reasoning between LLMs
- **Combine LLMs with specialized vision models**
  ➔ Leverage strengths of both to address limitations in image-only classification.

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- **LLMs perform very well on text-based cyberbullying detection**
  ➔ achieving up to 85.64% F1 in zero-shot settings,
  ➔ but struggle with image-only inputs
- **Multimodal (text+image) inputs did not outperform text-only**
  ➔ suggesting visual cues are inherently difficult to interpret for both humans and AI
- **Few-shot learning did not improve performance**
  ➔ with most LLMs performing worse than in *zero-shot*;
  ➔ only GPT-4o showed slight gains
- **Ensembling via majority voting showed limited value**
  ➔ with only selective voting among the top 3 models yielding small improvements

Future Work

- **Explore advanced ensemble or multi-agent methods**
  ➔ Move beyond majority voting to collaborative reasoning between LLMs
- **Combine LLMs with specialized vision models**
  ➔ Leverage strengths of both to address limitations in image-only classification.
- **Fine-tune LLMs on cyberbullying data**
  ➔ Improve performance for both text and multimodal inputs through task-specific adaptation

THANK YOU

Tim Schlippe
tim.schlippe@iu.org

REFERENCES

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

[1] C. Toraman, F. Şahinuç, and E. H. Yilmaz, “Large-Scale Hate Speech Detection with Cross-Domain Transfer,” *PloS one*, vol. 17, no. 9, p. e0273224, 2022.

[2] S. Hinduja and J. W. Patchin, *Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying*, 2nd ed. Thousand Oaks, California: Corwin, 2015.

[3] UNICEF, “Cyberbullying: What Is It and How to Stop It,” 2024, accessed: 2024-05-15. [Online]. Available: https://www.unicef.org/end-violence/how-to-stop-cyberbullying

[4] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, “Cyberbullying: Its Nature and Impact in Secondary School Pupils,” *Journal of Child Psychology and Psychiatry*, vol. 49, no. 4, pp. 376–385, 2008.

[5] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, “Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research Among Youth,” *Psychological Bulletin*, vol. 140, no. 4, pp. 1073–1137, 2014.

[6] E. Aboujaoude, M. W. Savage, V. Starcevic, and W. O. Salame, “Cyberbullying: Review of an Old Problem Gone Viral,” *Journal of Adolescent Health*, vol. 57, no. 1, pp. 10–18, 2015.

[7] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson, “Risks and Safety on the Internet: The Perspective of European Children,” *EU Kids Online*, 2011. [Online]. Available: https://www.lse.ac.uk/media-and-communications/assets/documents/research/eu-kids-online/reports/EU-Kids-Online-Final-Report.pdf

[8] S. K. Schneider, L. O'Donnell, A. Stueve, and R. W. S. Coulter, “Cyberbullying, School Bullying, and Psychological Distress: A Regional Census of High School Students,” *American Journal of Public Health*, vol. 102, no. 1, pp. 171–177, 2012.

[9] J. W. Patchin and S. Hinduja, “Traditional and Nontraditional Bullying Among Youth: A Test of General Strain Theory,” *Youth & Society*, vol. 43, no. 2, pp. 727–751, 2011.

[10] M. F. Wright, “The Relationship Between Young Adults' Beliefs About Anonymity and Subsequent Cyber Aggression,” *Cyberpsychology, Behavior, and Social Networking*, vol. 16, no. 12, pp. 858–862, 2013.

[11] C. P. Bartlett, C. C. DeWitt, B. Maronna, and K. Johnson, “Social Media Use as a Tool to Facilitate or Reduce Cyberbullying Perpetration: A Review Focusing on Anonymous and Nonanonymous Social Media Platforms,” *Violence and Gender*, vol. 5, no. 3, pp. 147–152, 2018.

[12] S. Aslam, “Twitter Statistics 2024: How Many People Use Twitter?” 2024, accessed: 2024-04-10. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics

[13] R. Spence, A. Bifulco, P. Bradbury, E. Martellozzo, and J. DeMarco, “The Psychological Impacts of Content Moderation on Content Moderators: A Qualitative Study,” *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, vol. 17, 09 2023.

[14] E. Parliament and C. of the European Union, “Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services (Digital Services Act),” 2022, accessed: 15 March 2025. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX\%3A32022R2065

[15] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support Vector Machines,” *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[16] K. Reynolds, A. Edwards, and L. Edwards, “Using Machine Learning to Detect Cyberbullying,” *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, vol. 2, 12 2011.

REFERENCES

[17] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, “Learning from Bullying Traces in Social Media,” in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, 2012, pp. 656–666. [Online]. Available: https://aclanthology.org/N12-1084

[18] M. Dadvar, F. d. Jong, R. Ordelman, and D. Trieschnigg, “Improving Cyberbullying Detection with User Context,” in *Advances in Information Retrieval.* Springer Berlin Heidelberg, 2013, pp. 693–696.

[19] M. A. Rahman, M. S. Hossain, N. A. Alrajeh, and F. Alghamdi, “Cyberbullying Detection from Textual Comments Using Traditional Machine Learning and Deep Learning Approaches,” in *2021 International Conference on Information and Communication Technology Convergence (ICTC).* IEEE, 2021, pp. 191–196.

[20] A. Muneer and M. S. Fati, “A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter,” *Future Internet*, vol. 12, no. 11, p. 187, 2020.

[21] G. Salton and C. Buckley, “Term Weighting Approaches in Automatic Text Retrieval,” USA, Tech. Rep., 1987.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[23] J. O. Atoum, “A Cyberbullying Detection Model Using Machine Learning Methods in Social Networks,” in *2020 International Conference on Computer Science and Software Engineering (CSASE).* IEEE, 2020, pp. 254–259.

[24] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, “Cyberbullying Detection on Social Networks Using Machine Learning Approaches,” in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1–6.

[25] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, “Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques,” *Electronics*, vol. 10, no. 22, p. 2810, 2021.

[26] F. Alkomah, S. Salati, and X. Ma, “A New Hate Speech Detection System based on Textual and Psychological Features,” *International Journal of Advanced Computer Science and Applications*, vol. 13, 01 2022.

[27] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, “Prediction of Cyberbullying Incidents in a Media-Based Social Network,” in *The 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM ’16. IEEE Press, 2016, p. 186–192.

[28] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, “Exploring Hate Speech Detection in Multimodal Publications,” 2019.

[29] A. Botelho, S. Hale, and B. Vidgen, “Deciphering Implicit Hate: Evaluating Automated Detection Algorithms for Multimodal Hate,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* Association for Computational Linguistics, 08 2021, pp. 1896–1907. [Online]. Available: https://aclanthology.org/2021.findings-acl.166

[30] B. Vidgen, H. Margetts, and A. Harris, “How Much Online Abuse is There? A Systematic Review of Evidence for the UK,” Nov. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3582599

[31] G. Sahu, R. Cohen, and O. Vechtomova, “Towards A Multi-Agent System for Online Hate Speech Detection,” 2021.

[32] S. Paul, S. Saha, and M. Hasanuzzaman, “Identification of Cyberbullying: A Deep Learning Based Multimodal Approach,” *Multimedia Tools and Applications*, vol. 81, pp. 1–20, 2022.

REFERENCES

[33] J. Heine, K. Schaaff, and T. Schlippe, "Investigating Text and Image Features for the Detection of Cyberbullying," in *The 8th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2024)*, Okayama, Japan, 2024.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[36] T. Ahmed, M. A. Kabir, T. Islam, M. S. Mahmud *et al.*, "Detection of Cyberbullying from Social Media Text Using Deep Transformer Model Ensembles," in *2022 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*. IEEE, 2022, pp. 416–421.

[37] W. Tapaopong, K. Panarat, P. Malasri, and K. Sirikanchana, "Evaluating Pre-trained Transformer Models for Cyberbullying Detection on Social Media," *Information Processing & Management*, vol. 61, no. 2, p. 103566, 2024.

[38] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A benchmark dataset for explainable hate speech detection," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 1, pp. 1–38, 2022.

[39] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media," in *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2019, pp. 1415–1420.

[40] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection," in *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1, 2021, pp. 1667–1682.

[41] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *The International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 512–515, 2017.

[42] S. Suryawanshi, B. R. Chakravarthi, P. Verma, M. Arcan, J. P. McCrae, and P. Buitelaar, "Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text," in *The Second Workshop on Trolling, Aggression and Cyberbullying*, 2020, pp. 32–41.

[43] D. Maity, A. A. Deshmukh, and S. K. Bharti, "MultiBully: A Multi-Task Learning Framework for Cyberbullying Detection," in *The 2022 International Conference on Multimedia Retrieval*, 2022, pp. 588–597.

[44] E. Nina-Gutiérrez, E. Laisa, and A. Molina-Villegas, "Leveraging Large Language Models for Multilingual Cyberbullying Detection," in *The 2024 International Conference on Natural Language Processing and Artificial Intelligence*, 2024, pp. 127–136.

[45] L. Hanu and U. team, "Detoxify: Toxic Comment Classification," GitHub repository, 2020. [Online]. Available: https://github.com/unitaryai/detoxify

REFERENCES

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

[46] G. G. Team *et al.*, “Gemini: A Family of Highly Capable Multimodal Models,” Google, Tech. Rep., 2023. [Online]. Available: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

[47] OpenAI, “GPT-3.5,” *OpenAI Documentation*, 2022. [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5

[48] A. Lees, D. Borkan, I. Kivlichan, J. Nario, and T. Goyal, “A New Perspective API: Efficient Multi-lingual Toxicity Measurement,” *arXiv preprint arXiv:2202.11176*, 2022. [Online]. Available: https://arxiv.org/abs/2202.11176

[49] OpenAI, “Moderation API,” OpenAI Documentation, 2022. [Online]. Available: https://platform.openai.com/docs/guides/moderation

[50] Anthropic, “Content Safety API,” Anthropic Documentation, 2023. [Online]. Available: https://docs.anthropic.com/claude/docs/content-safety

[51] OpenAI, “GPT-3.5 Turbo,” OpenAI Documentation, 2023. [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5-turbo

[52] S. Shekhar, “Evaluating Large Language Models for Cyberbullying Detection in Social Media,” *Journal of Cybersecurity and Privacy*, vol. 4, no. 1, pp. 18–35, 2024.

[53] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. C. Singh, A. Rouditchenko, V. Sanh, L. Chesnut, T. Darcet, S. Shakeri, K. Lyu, M. Ott, H. Touvron, T. Scialom, S. J. Hosseini, S. Ramaswamy, Y. Tay, X. Wang, M. Dehghani, M. Shah, Y. Liu, M. De Menten, A. Jin, A. R. Goyal, M. Lewis, K. Hoffman, P. Kohli, and D. Kiela, “Mistral 7B,” *arXiv preprint arXiv:2310.06825*, 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[54] LMSYS Org, “Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality,” LMSYS Org, 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[55] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “LLaMA: Open and Efficient Foundation Language Models,” *arXiv preprint arXiv:2302.13971*, 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[56] G. Ramos, T. Ferreira, R. Santos, and R. Prati, “Comparative Analysis of Large Language Models and Traditional Machine Learning for Hate Speech Detection in Portuguese,” in *The 2024 Brazilian Conference on Intelligent Systems*, 2024, pp. 215–226.

[57] J. Bauer, T. Schuster, E. Filtz, M. Koutraki, and A. Fensel, “Large Language Models for Automated Sentiment Analysis: Evaluating Zero-Shot and Few-Shot Learning Approaches,” in *The 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 43–51.

[58] S. Bucciarelli, M. Bertini, and A. Del Bimbo, “Investigating Large Vision-Language Models for Image Captioning: Capabilities and Limitations,” in *The 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024, pp. 1156–1165.

[59] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning,” *arXiv preprint arXiv:2304.08485*, 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[60] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved Baselines with Visual Instruction Tuning,” *arXiv preprint arXiv:2310.03744*, 2023. [Online]. Available: https://arxiv.org/abs/2310.03744

[61] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common Objects in Context,” in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755. [Online]. Available: https://arxiv.org/abs/1405.0312

REFERENCES

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

[62] G. Chochlakis, A. Potamianos, K. Lerman, and S. Narayanan, "The Strong Pull of Prior Knowledge in Large Language Models and Its Impact on Emotion Recognition," in *The 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Los Alamitos, CA, USA: IEEE Computer Society, September 2024, pp. 318–326. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ACII63134.2024.00041