

# AUTAPP: Automated Detection of Online Harms for Social Media Applications

To learn more about **REPHRAIN**,  
our future plans and how to get  
involved:



[www.rephrain.ac.uk](http://www.rephrain.ac.uk)



[@REPHRAIN1](https://twitter.com/REPHRAIN1)



[rephrain-centre@bristol.ac.uk](mailto:rephrain-centre@bristol.ac.uk)

Rohit Nautiyal, Claudia Peersman and Minhao Zhang

# Overview

 Introduction

 Research Scope and Objectives

 Literature Review

 Next Steps



# Research Scope and Objectives



---

Potential role of automated methods to address online harms

---

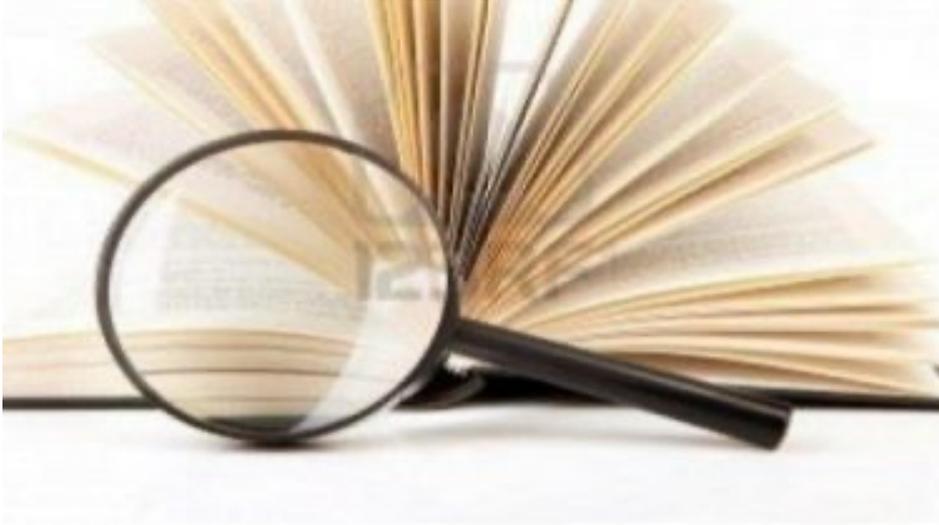
Abuse of commercially available PETs for criminal purposes in the context of online harms

---

Multidisciplinary perspective: literature review, develop intelligent technologies, semi-structured interviews



# Literature Review



- Background
- Approach
- Research Objectives
- Results
- Next steps

# Social media damages teenagers' mental health during Covid pandemic

The screenshot shows a BBC News article. The main headline is "Social media damages teenagers' mental health, report says". The sub-headline is "Record high number of recorded grooming crimes lead to calls for stronger online safety legislation". The article is dated 27 January 2021. There are two images: one of a young woman looking at her phone and another of a young girl looking at her phone. A green button labeled "Take action with us" is visible. The article text includes: "Teenagers' mental health is being damaged by heavy social media use, a report has found. Research from the Education Policy Institute and The Prince's Trust said wellbeing and self-esteem were similar in all children of primary school age. Boys and girls' wellbeing is affected at the age of 14, but girls' mental health drops more after that, it found. A lack of exercise is another contributing factor - exacerbated by the pandemic, the study said. According to the research: One in three girls was unhappy with their personal appearance by the age of 14, compared with one in seven at the end of primary school. The number of young people with probable mental illness has risen to one in six, up from one in nine in 2017."

- Increased social media use was linked to negative wellbeing and self-esteem
- Misinformation across social media major source of anxiety and stress
- 5441 cases of child abuse records during 2020 to 2021
- Increase of around 70% of reported crimes

Source1: <https://www.bbc.co.uk/news/technology-55826238>

Source2: <https://reutersinstitute.politics.ox.ac.uk/news/how-coronavirus-pandemic-changing-social-media>

# Social media damages teenagers' mental health during Covid pandemic (2)

## Half of children and teens exposed to harmful online content while in lockdown

New research by the British Board of Film Classification (BBFC) has shown that children and teens are being exposed to harmful or upsetting content while in lockdown, often on a daily basis.

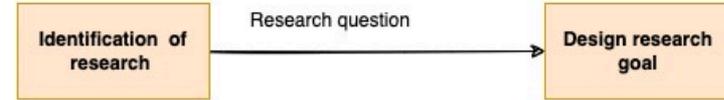
- 47% of teens say they have seen content online they wish they hadn't seen while in lockdown, and one in seven (13%) see harmful videos everyday.
- 14 year olds see the most harmful content, with a quarter saying they see inappropriate videos every day.
- The BBFC website and free app contains ratings info and age ratings so parents can help their children make informed viewing choices.

The research, carried out by YouGov, has revealed:

- 47% of children and teens have seen content they had rather avoided
- 29% leaving them feeling uncomfortable
- 23% scared and 19% confused

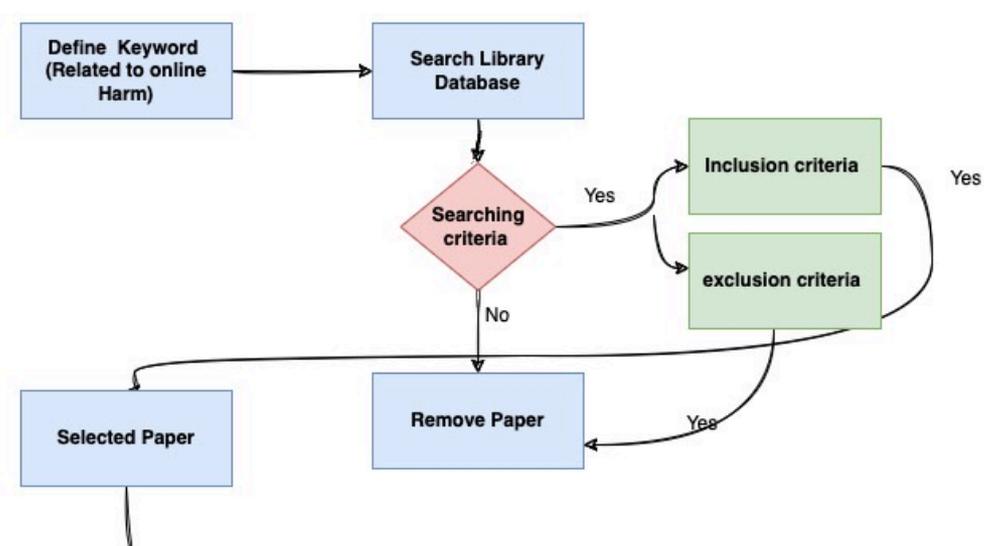
Online harm and detection Techniques

Research Definition



# Approach

Methodology



Outcome of SLR





# Research Questions

- RQ 1: What definitions of the different types of online harms are found in the literature?
- RQ 2: Which detection techniques are used to identify various type of online harm ?
- RQ 3: Which datasets are publicly available for training tools for detecting the different types of online harms in social media platforms ?

# Inclusion & Exclusion Criteria

## **Inclusion**

- All the published articles are written in English
- Articles which focus specifically on online harm detection in social media.
- All the published articles that relate to one or more research questions



# Inclusion & Exclusion (2)

## Exclusion

- Articles published in any language other than English
- Articles containing insufficient information regarding online harm detection on social media platforms
- Articles that occur more than once (deduplication)
- Articles which do not have any link with the given research questions

# Databases

- **Number of databases used : 5**

(IEEE Explore, ACM Digital, Libraries, Science Direct, Scopus, ACL Anthology)

- **Currently used database: 3**

(IEEE Explore, ACL Anthology, ACM Digital Library)

- **Total number of relevant studies : 110**

# Type of Online Harm and Used Methods

Type of Harm	Title	Detection Methods
Child Grooming	Detection of Cyber Grooming in Online Conversation	<ol style="list-style-type: none"><li>1) Message-Based Detection</li><li>2) Author-Based Detection</li><li>3) Conversation-Based Detection</li></ol>
Child abusing	Child Sexual Abuse Detection in Image and Video	<ol style="list-style-type: none"><li>1) Skin Detection-Based methods</li><li>2) Image Descriptor Based methods</li><li>3) Video-Based Methods</li><li>4) Deep learning-Based Methods</li></ol>
Sexual harassment	Distributional Semantics Approach to Detect Intent in Twitter Conversations on Sexual Assaults	Model for a Twitter post using semantic features of the intent senses learned with CNN
Hate speech	Hate Speech Detection in Twitter using Natural Language Processing	Deep learning and CNN used for hate speech

# Type of Online Harm and Used Methods (2)

Type of Harm	Title	Detection Methods
Cyber bullying	Unsupervised Cyber Bullying Detection in Social Networks	Syntactic and Semantic analysis Natural language processing(NLP), GHSOM network algorithm, SOMToolbox2 framework
Extremist	Social Big Data Mining Framework for Extremist Content Detection in Social Networks	Framework for opinion mining, Graph API,NLP for content analysis
Child Grooming	Detecting Nastiness in Social Media	NLP approach, classic features, newer features, supervised classification algorithm

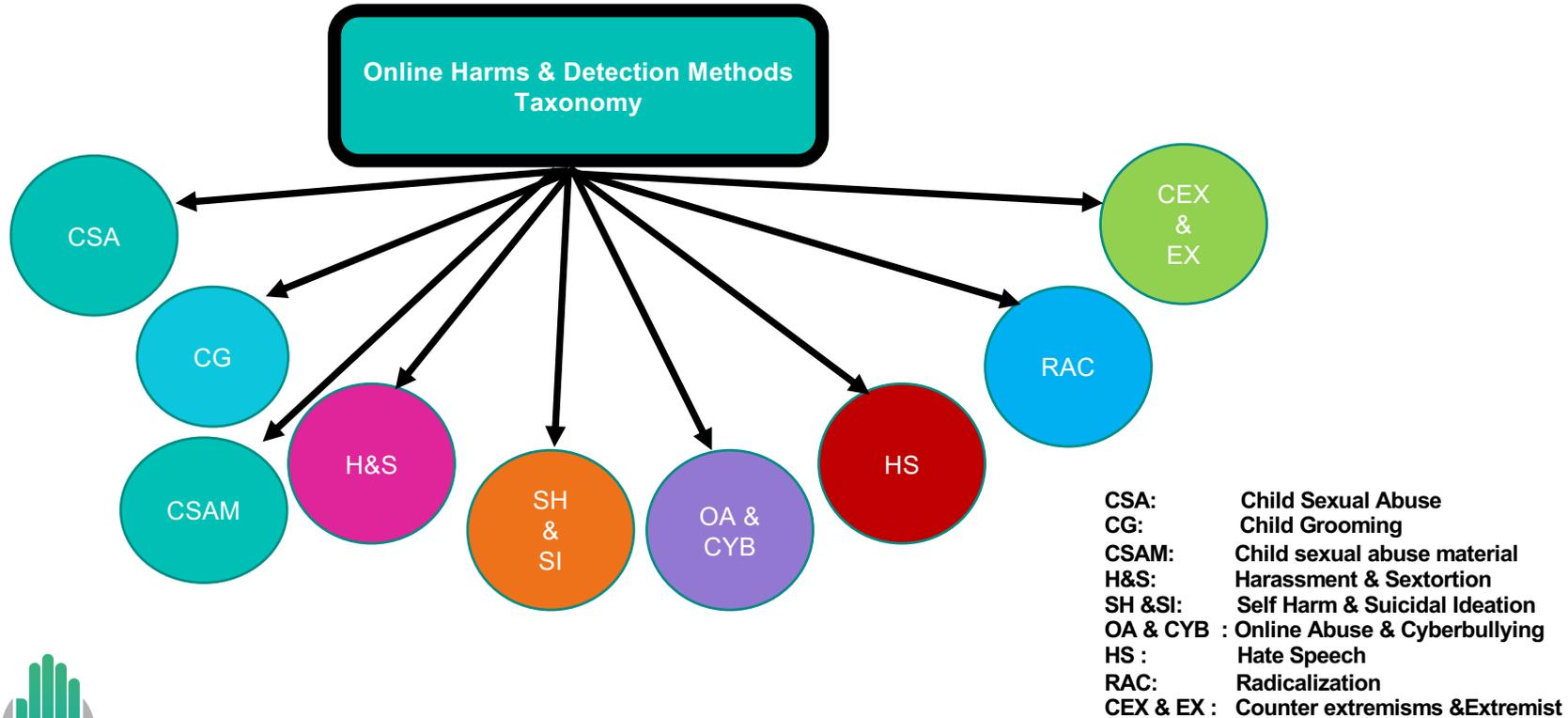
# Analysis of Methods and Harms

Type of Harm	ML	Detection Tool	NLP/Text Mining	Deep Learning	Image/Video Analysis	Framework
Child Grooming	YES	YES	YES	YES		
CSAM	YES	YES	YES	YES	YES	
Self harm /Suicidal Ideation	YES			YES		YES
Harassment, Sextortion	YES	YES	YES		YES	
Online Abuse/Cyberbullying	YES		YES		YES	
Hate speech	YES		YES	YES	YES	
Radicalisation	YES		YES			YES
Counter Extremisms/Extremist	YES	YES		YES		YES

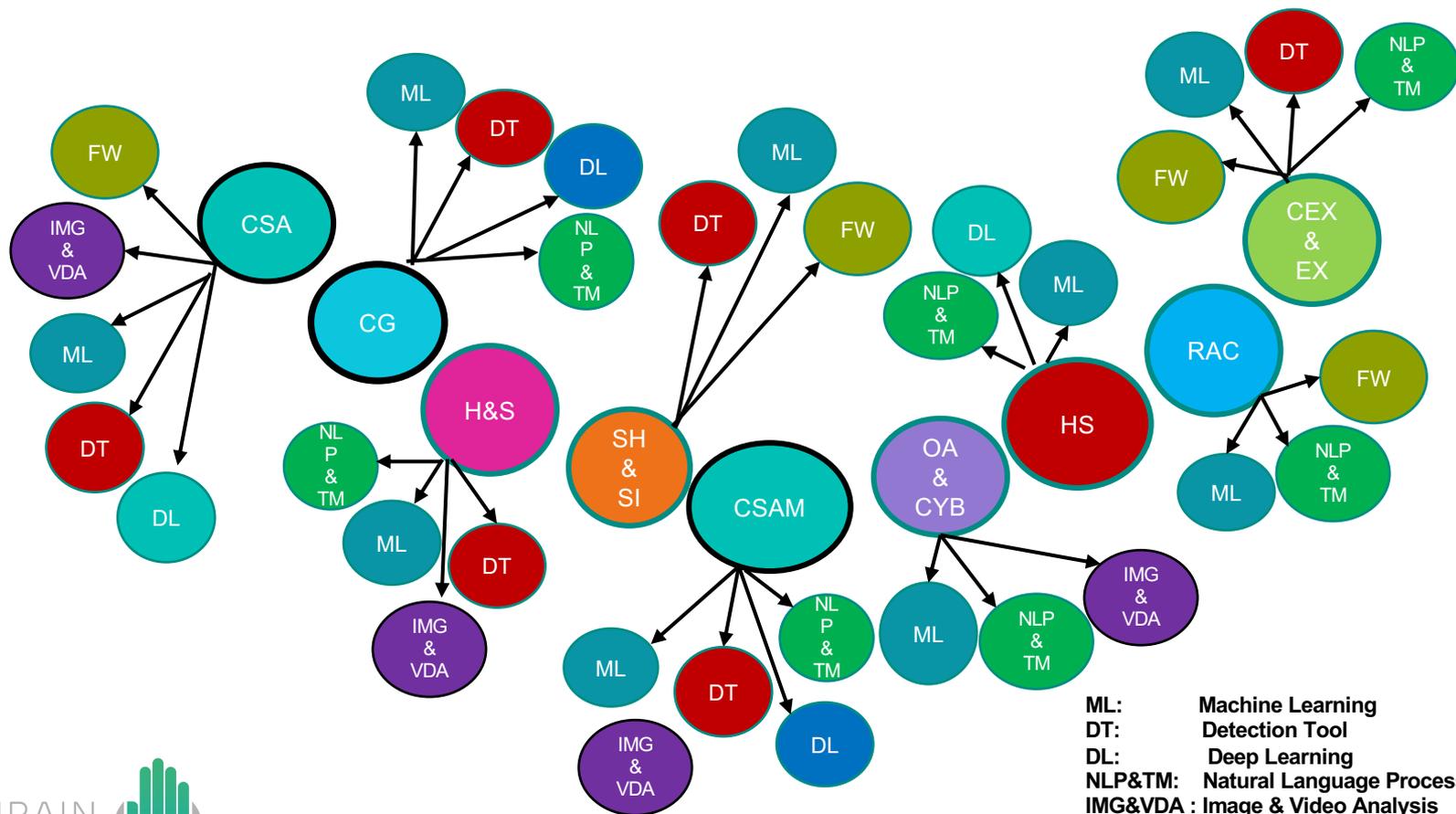
Framework : Used for content and context detection

Detection Tool: Pre-existing tools for feature extraction and dataset

# Detection Methods and Harms Taxonomy



# Detection Methods and Harms Taxonomy



**ML:** Machine Learning  
**DT:** Detection Tool  
**DL:** Deep Learning  
**NLP&TM:** Natural Language Processing & Text Mining  
**IMG&VDA :** Image & Video Analysis  
**FW:** Framework

# Datasets



Type of Harm	Dataset name	Size Of Dataset	Accessibility	Limitation	Benefits
Hate Speech	Contextual Abuse Dataset	25000 Redditt entries	Publicly Accessible	Distinct for specific category	Contain rational, labeled ,annotated
Hate Speech	Implicit Hate Speech	22,0056 tweets, 6,346 implicit tweets	Publicly Accessible		Multilingual Detection
Harassment	Instagram social media data	80,000 posts, 1,262 tweets	On request Accessible	Identification of nodes	Useful for source and target connection
Online Abuse	ConvAbuse	4,185 chats and conversations	Publicly Accessible	Gender based	Detecting harmful content and human-machine conversational AI

# Datasets (2)

Type of Harm	Dataset name	Size Of Dataset	Accessibility	Limitation	Benefits
Online Abuse	SWAD	13,240 Tweets	Publicly Accessible	Some text are short and context missing	Presence of hashtags for context nature understanding
Toxic Language	SemEval-2019 Task 5	19,600 tweets,	Publicly Accessible	Detection against micro-blogging texts	SemEval 2019 multilingual dataset
Toxic Language	ALONE	688 Tweets	Publicly Accessible	Binary toxic or non toxic	Multimodal (text, images, emojis, metadata)
Offensive Language	Offensive posts	14,100 Posts	Publicly Accessible	Branching structure of tasks	Target identification

# Total Number of Data sets (27)

Type of Harm	Count
Hate speech/Toxic Language	8
Child Grooming	2
Child Sexual Abuse Media	1
Counter Extremism /Extremism	3
Cyber Bullying	2
Online Harassment /Abuse	8
Suicidal Ideation/Self Harm	3

# Examples of Online Harm Detection Tools

Tool Name	Description	Citation
Anti-Grooming Tool	Identifying online grooming in real time and scan text-based chat for possible grooming.	Thorn Grooming Classifier Robertson Wang   07-2021
Perspective API:	Perspective free API that uses machine learning to identify "Toxic" comments.	Hosseini, Hossein, et al. "Deceiving google's perspective api built for detecting toxic comments.
MIND TIME	The tool aims to detect Borderline personality disorder symptoms and signs of Self-harm ideation and allows users to make notes of daily events, experiences, thoughts, and feelings.	Andrews, Dittin, et al. "Child online safety in indian context."
Online hate Index (OHI)	Designed to transform the human understanding of hate speech via machine learning into a scalable tool that can be deployed on internet content to discover the scope and spread of online hate speech.	von Vacano, Claudia, Abigail T. De Kosnik, and Stephen Best. "Digital Humanities at Berkeley and the Digital Life Project.

# Examples of Online Harm Detection Tools(2)

Tool Name	Description	Citation
HATEMETER	Detect Anti Muslim hate speech using machine learning and NLP techniques platform available in English, French and Italian	Di Nicola, Andrea, et al. "HATEMETER: Hate speech tool for monitoring
COSMOS	Collect and analyses data from Twitter or social media platform in real-time by keyword specification using sentiment analysis and natural language processing.	Ampofo, Lawrence, et al. "Text mining and social media: When quantitative meets qualitative and software meets people."
MANDOLA	Detect hateful content through a combination of sentiment analysis, natural language processing. Machine learning and deep learning	Astorena, C.M.; Abundez, I.M.; Alejo, R.; Granda-Gutiérrez, E.E.; Rendón, E.; Villegas, O. Deep Neural Network for Gender-Based Violence Detection on Twitter Messages.

# Findings

- It is difficult to find standard definitions of each type of online harm in the context of social media.
- Most publicly available datasets & tools are on hate speech, toxic language & abusive language.
- Lack of data for some types of online harms.
- Similar text mining problems within different types of online harms:
  - use of persuasion (e.g. grooming, radicalisation)
  - direct, explicit language (cyber-bullying, hate speech, toxic language)



# Next Steps: Experimental Setup

- Setting up machine learning pipelines for performing online harms detection experiments → first step: hate speech/toxic language detection.
- Cross platform experiments on multiple datasets of hate speech, toxic language.
- Develop a framework for training online harm detection tools from the point of view of different text mining problems, rather than for each type of online harms.

# Future Research

## It would be interesting to explore:

- More efficient/effective techniques.
- Create new datasets for future research.
- Develop experimental matrices/frameworks for online harm detection.
- Model performance check on implicit content-based datasets.

# Questions?

Contact:

[rohit.nautiyal@bristol.ac.uk](mailto:rohit.nautiyal@bristol.ac.uk),  
[claudia.peersman@bristol.ac.uk](mailto:claudia.peersman@bristol.ac.uk),  
[minhao.zhang@bristol.ac.uk](mailto:minhao.zhang@bristol.ac.uk)

