Title: Inclusion criteria and PRISMA diagram appendix for “The unappreciated role of intent in algorithmic moderation of abusive content on social media”
Authors: Xinyu Wang (1), Sai Koneru (1), Pranav Narayanan Venkit (1), Brett Frischmann (2), Sarah Rajtmajer (1)
Date: July 29th, 2025
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix C: Inclusion criteria and PRISMA diagram

Query for dataset papers

KEY (“social media” AND “dataset” AND (“NLP” OR “Natural Language Processing”) AND (“hate speech” OR “abus*” OR “offens*” OR “cyberbully*”)) OR TITLE (“social media” AND “dataset” AND (“hate speech” OR “abus*” OR “offens*” OR “cyberbully*”)) AND (LIMIT-TO (LANGUAGE, “English”))

Inclusion criteria for dataset papers

We applied the following inclusion criteria:

1) The paper presents a novel dataset for which annotation procedures are described.
2) The dataset is intended for training and testing algorithm(s) aimed at abuse detection.
3) The dataset is curated from one or more widely used social media platforms.
4) The dataset is in English.[1]
5) The dataset includes textual content.

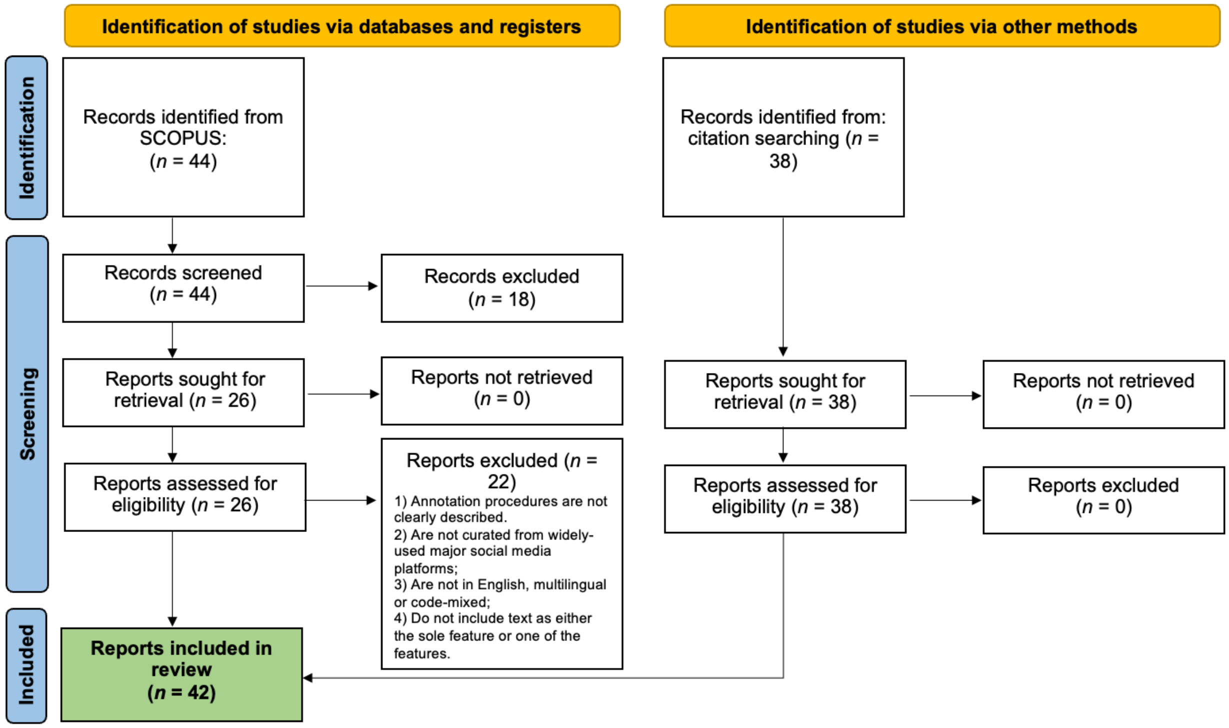
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from SCOPUS: (n = 44)
Records identified from: citation searching (n = 38)
Records screened (n = 44)
Records excluded (n = 18)
Reports sought for retrieval (n = 26)
Reports not retrieved (n = 0)
Reports sought for retrieval (n = 38)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 26)
Reports excluded (n = 22)
1) Annotation procedures are not clearly described.
2) Are not curated from widely-used major social media platforms;
3) Are not in English, multilingual or code-mixed;
4) Do not include text as either the sole feature or one of the features.
Reports assessed for eligibility (n = 38)
Reports excluded (n = 0)
Reports included in review (n = 42)

Figure C1. PRISMA diagram for the selection of papers presenting labeled datasets for online abuse.

[1] Focusing on monolingual settings allows us to address these issues directly before extending analyses to multiple languages, where cultural variations and linguistic nuances further complicate intent inference.

Query for algorithm papers

KEY (“social media” AND (“NLP” OR “Natural Language Processing”) AND “de-taction” AND (“hate speech” OR “abus*” OR “offens*” OR “cyberbully*”)) OR TITLE (“social media” AND “detection” AND (“hate speech” OR “abus*” OR “offens*” OR “cyberbully*”)) AND (LIMIT-TO (LANGUAGE, “English”))

Inclusion criteria for algorithm papers

Similar to our survey of datasets, we removed papers that were not accessible, not written in English, or which did not describe an algorithm for detection of online abuse. We removed models designed for multilingual or non-English tasks.

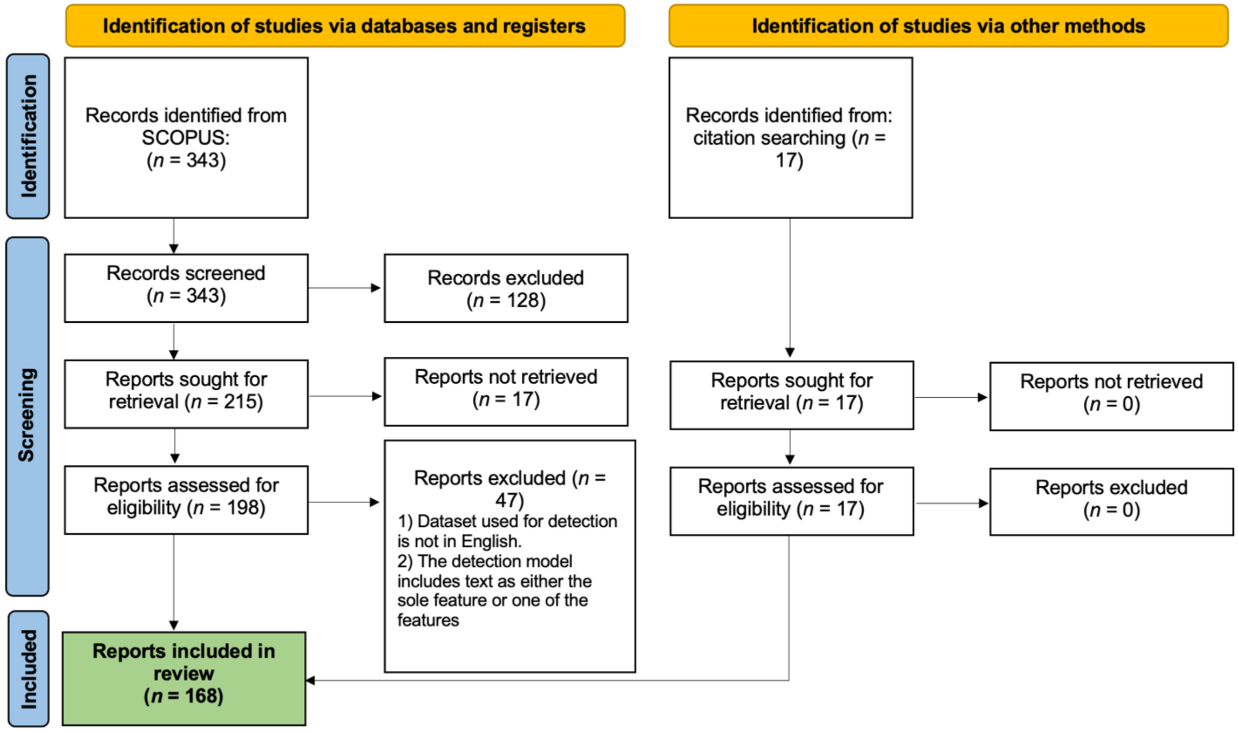
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from SCOPUS: (n = 343)
Records identified from: citation searching (n = 17)
Records screened (n = 343)
Records excluded (n = 128)
Reports sought for retrieval (n = 215)
Reports not retrieved (n = 17)
Reports sought for retrieval (n = 17)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 198)
Reports excluded (n = 47)
1) Dataset used for detection is not in English.
2) The detection model includes text as either the sole feature or one of the features
Reports assessed for eligibility (n = 17)
Reports excluded (n = 0)
Reports included in review (n = 168)

Figure C2. PRISMA diagram for the selection of papers presenting detection algorithms for online abuse.