THE WEB CONFERENCE ACM

CHARACTERIZING, DETECTING, AND PREDICTING
Online Ban Evasion

Manoj Niverthi*, Gaurav Verma*, Srijan Kumar
Georgia Institute of Technology

Georgia Institute of Technology
College of Computing

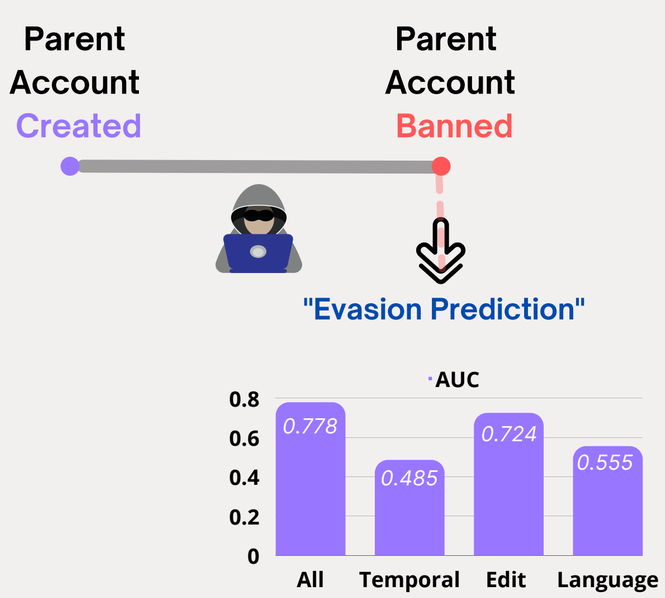
1 Introduction

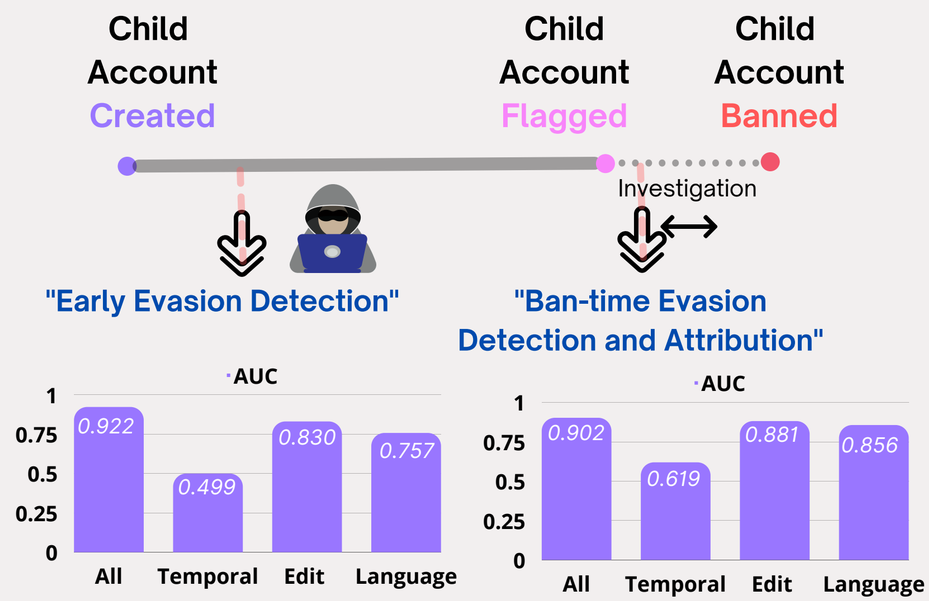
- Malicious actors can **evade bans** to continue **malicious activities**
- Only **10% of ban evaders** are reported by mods – Reddit Security
- Current moderation strategies are **manual** and **time-consuming**
- Envisioned moderation strategy
  - **Automated**, **evidence-based**, and **efficient**
  - **Final decision** lies with the moderators

2 Wikipedia Ban Evasion Dataset

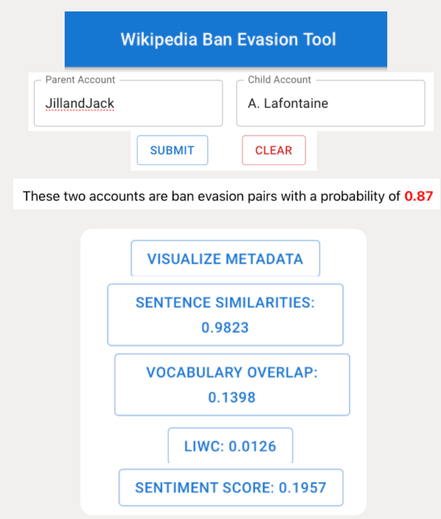
- **8,551** ban evasion pairs
- **Manual rigorous investigations** by Wikipedia users
  - Specifically covers "bad" ban evasion
- **All meta-data** for each user and **all the edit information**
- **Publicly available** for download
  - https://github.com/srijankr/ban_evasion

3 Ban Evasion Life Cycle





4 Live Demo



with **Zhen Jiang** and **Jio Oh**

5 Analyzing Ban Evasion

- Future ban evaders are **more active and less explicit** than non-evading malicious users
- **Similar motives** of ban evasion pairs
  - Same Wikipedia pages
  - Similar vocabulary
  - Similar psycholinguistic attributes
- Success of ban evasion pairs
  - **Fewer swear words**
  - **More objective language**
  - **Low username similarity**

Stop by our main conference presentation @ ACM WebConf
Apr 28, 16:45 CET
Apr 28, **10:45 ET**
Apr 28, **07:45 PT**
Apr 28**, 20:15 IST**