



Strategized Counter Trolling

*By: Akshaya Sundaram, Ripunjay Sharma,
Subha Ramesh,
Sharvari Thippanna*

Background

- **Psychological** - Trolls are motivated by desire to hurt / anger victims
 - **Humor** - Strategy to counter trolls
 - 'Zero Trollerance' campaign counter trolled trolls with videos
 - Recent work has proved that counter-narratives are effective in hate countering
1. <https://www.bbc.com/future/article/20160318-what-is-the-best-way-to-stop-internet-trolls>
 2. Buckels, Erin E., et al. "Internet trolling and everyday sadism: Parallel effects on pain perception and moral judgment." *Journal of personality* 87.2 (2019): 328-340.



Motivation

Freedom of
Speech

De-escalate
trolling



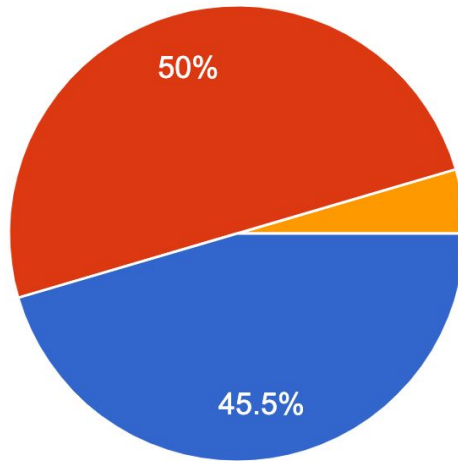
Survey

- Conducted survey to **assess trolling behavior** and to design an effective counter trolling system
- Participants were asked to provide response to the following questions :
 - Have you come across someone being trolled on the Internet ?
 - Do you think Internet trolls have turned more offensive (racially / sexually) over the past couple of years ?
 - Do you think a support system is needed to tackle trolling ?
 - Would you prefer an automated system that trolls the troller or a platform that automatically deletes trolls ?
- **22** respondents recorded

Survey Results

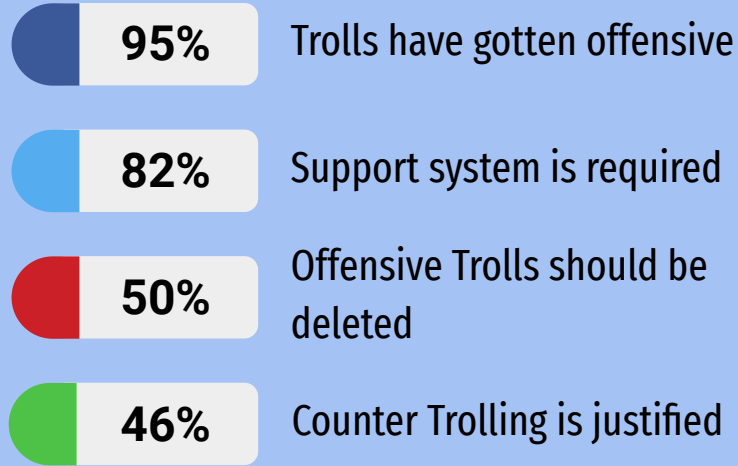
Would you prefer an automated system that trolls the troller or a platform that automatically deletes trolls ?

22 responses



- System that automatically trolls the troller
- Platform that automatically deletes trolls
- Again, I don't think that all trolls should be banned or deleted. Freedom of expression is important too. Nowadays we see lot of trolls and memes that do mock the wrong moves of celebrities or governments which is needed too. I think having humans in the loop who are politically neutral and diverse in gende...

Survey Results



Counter Troll Strategies



Memes

Jokes

Quotes

Toxic Flag

- Memes, Jokes and Quotes were chosen generically
- Designed to counter troll without inciting toxic speech
- Toxic Content Flag can be used for offensive or inappropriate comment


Overview of our system

Counter Trolling - Trobo

Please feed your comment here

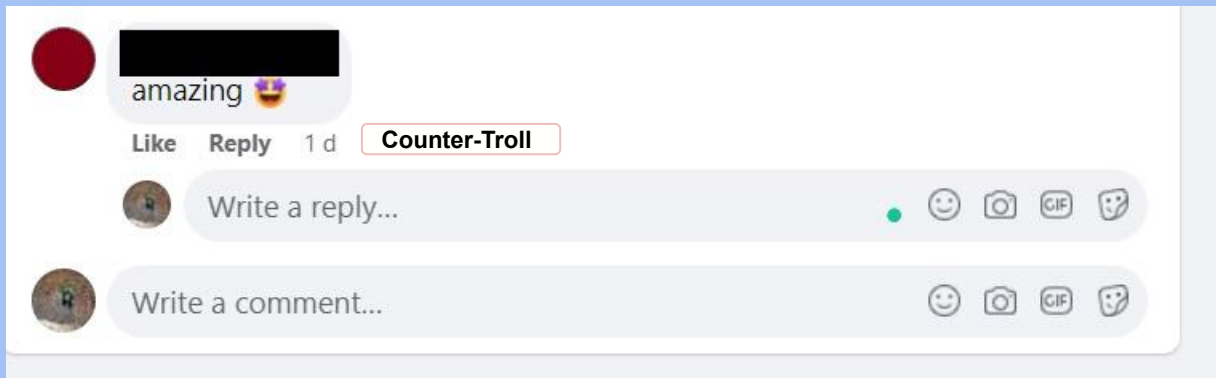


Post

-  **User** 3/10/2022
Perspective API to detect Toxicity in comments.
Reply Edit Delete **Counter-Troll** ▼
-  **User** 3/10/2022
Designed using React.js for front-end and Python Flask for backend.
Reply Edit Delete **Counter-Troll** ▼
-  **User** 3/10/2022
Hello
Reply **Counter-Troll** ▼

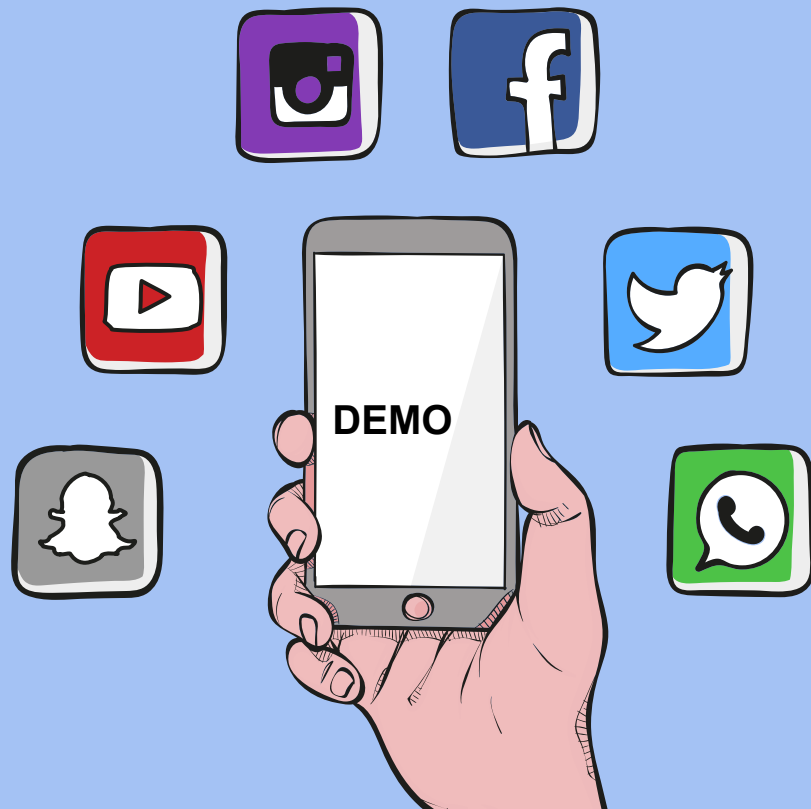
Counter- Troll dropdown provides options to select 4 different strategies:

- Jokes
- Memes
- Quotes
- Toxic Flag



Incorporate into existing
platforms





Possible Improvements

Language Support

Extend support to other languages like hi, nl, fr, de which the Perspective API supports

Media support

Extend support for detecting trolling in images and other media



Incorporate attributes

Enable fine-grained attributes like PROFANITY, IDENTITY_ATTACK and THREAT

Feedback to API

Incorporate feedback system to API in case of false positives



Thank you!

