Adolescent Online Safety: Expanding the Knowledge Base and Developing Machine Learning Algorithms

Carlson Sharpless
University of Central Florida
cwsharpless@knights.ucf.edu

Pallavi Dacre
University of Central Florida
pdacre19@knights.ucf.edu

Komila Khamidova
University of Central Florida
komilak@knights.ucf.edu

Avery Reyna
University of Central Florida
avery.reyna16@knights.ucf.edu

Abdulrahman Al Sumaih
University of Central Florida
aalsumaih@knights.ucf.edu

Christian Lozano
University of Central Florida
christianlozano@knights.ucf.edu

Afsaneh Razi
University of Central Florida
afsaneh.razi@knights.ucf.edu

Ashwaq Soubai
University of Central Florida
atalsoubai@knights.ucf.edu

Pamela Wisniewski
University of Central Florida
pamela.wisniewski@ucf.edu

Motivation and Big Idea

- Adolescent safety is an important aspect of modern-day social media design.
- There are many methods that can be used to improve social media safety, but one common example is the use of artificial intelligence models that detect potentially risky messages.
- While risk detection models are not new, most are based on public datasets, which can be unrepresentative of actual adolescent experiences. In particular, many are based around adults posing as teens for journalistic operations.

Research Questions

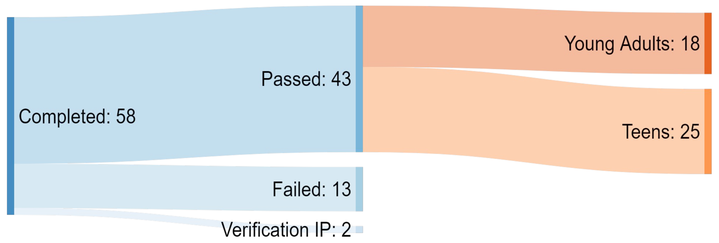
- How can we gather useful data from actual teens?
- How can we identify potentially risky messages in the data for use with machine-learning models?

Methodology

- We first had participants take an online survey where they describe their Instagram usage and their attitudes toward it.
- After that was completed, participants upload a copy of their Instagram data and flag conversations with potential risks.
- Finally, the messages are being qualitatively annotated by the research team to create machine learning training datasets.

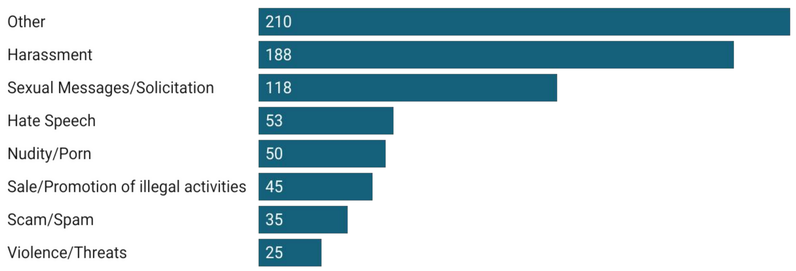
Findings

Status of Participants Who Have Completed the Study:



Number of Unsafe Messages by Risk Levels of 43 Participants

Number of Unsafe Messages by Risk Type of 43 Participants



- PAN12 Sexual Predator Identification (PAN12) dataset was used to better understand how Naive Bayes (NB) and Support Vector Machine (SVM) perform in the context of online risk detection.
- PAN12 is benchmark public dataset that will allow us to test various models.
- Our PAN12 models seek to identify messages sent by predators in online conversations.
- An accuracy score of about ~96% was obtained using SVM (Al Sumaih) and ~75% using NB (Dacre) when applied to the PAN12 dataset.

Discussion

Data Collection:

- Instagram is being used as a tool to promote the study to potential participants. Our goal is to have 180 verified participants.

Data Annotation:

- Messages which involve promotion of or engagement in dangerous and/or illegal activity will receive the highest risk level. Harassment (188) and sexual messages (118) were flagged the most by participants.

Machine Learning:

- Our NB implementation (naively) distinguishes predator messages from non-predator messages by counting the number of times a particular word appears in each type of message.
- Since predators may use similar language with multiple victims, this explains why NB performs well, despite its simplicity.
- We used PAN12 because messages in the dataset contain risk types that our algorithms will need to identify.

Acknowledgements

This research is partially supported by the U.S. National Science Foundation under grants IIP-1827700. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsor.