Identifying Online Misbehavior

Sanjana Cheerla, Vignitha Ampally, Vaibhav Garg, Saikath Bhattacharya, Dr. Munindar P. Singh

Introduction

Online misbehavior: when an app user intimidates or causes discomfort to another app user, e.g., by asking for inappropriate images, giving threat messages, and stalking.

Research question: How can we help app distribution platforms and users identify apps that facilitate misbehavior?

Opportunity: The requisite knowledge exists in app reviews.

Contribution: A natural language processing framework.

Data Set

- Created a comprehensive definition of online misbheavior.
- Made a list of apps typically used for online misbehavior and fetched similar apps.
- Retrieved application reviews from the App Store for each app.
- Filtered out reviews based on determined misbehavior keywords and reviews with less than 3-star ratings.
- Annotated each of the reviews based on the definition from Step 1: 0 if no indication of misbehavior, 1 otherwise.
- Refined the steps above based on how the agreement score is and as well as how relevant the gathered reviews are.

Example 1: Online Misbehavior Incidents

Sending Intimate Images
"I've been logged out of my accounts countless of times because I'm simply not interested in the people that text me so they report me when I leave them on seen. So many men are sending me nudes or bullying me on this app. This app is terrible don't use it..."

Asking Intimate Images
"This app is filled with disgusting creeps that ask for your Snapchat twitter Facebook and insta they will ask for a picture of you deny it they will ask for your name deny it and they will ask for your nudes please deny it..."

Sending Threats
"This is the absolute worst app I have ever been on. The rape culture is unreal. A guy just commented on a post of mine saying he wanted to beat and rape me..."

Stalking Activity
"...I had a person cyberstalk me for weeks and ... I feared for my life and you get off on a technicality. I'm going to the Police tomorrow to report my cyber stalker. I wish SKOUT actually cared about it's user as it claims it does."

Stalking Hazard
"I don't like how the update keeps asking you to add your location... And you can not click anything besides YES!!! I don't want people knowing where I am!! This is a safety concern especially for girls!!!!!! I liked the old one better and may just delete this... I'm not promoting stalking!!"

Table 1: Class distribution in labeled data.

	no misbehavior reported	misbehavior reported	total
train data	80	205	285
test data	348	1366	1714
total	428	1571	

Table 2: Model performance for no misbehavior class.

	Precision	Recall	F1 score
Glove + SVC	19.835	41.379	26.816
USE + SVC	**56.681**	**70.960**	**59.709**
Word2Vec + KNN	19.882	44.828	27.489

Table 3: Model performance for misbehavior class.

	Precision	Recall	F1 score
Glove + SVC	79.352	57.934	66.610
USE + SVC	**91.791**	**83.163**	**87.250**
Word2Vec + KNN	79.288	53.807	64.108

Table 4: Overall performance for all models.

	Precision	Recall	F1 score
Glove + SVC	54.142	46.713	58.530
USE + SVC	**80.630**	**73.480**	**81.658**
Word2Vec + KNN	51.984	45.799	56.573

Models / Training

We have created/currently working on the 4 models listed below:

1. GloVe with SVC classifier
2. Universal Sentence Encoder with SVC classifier
3. Word2Vec with K-Nearest Neighbors classifier.
4. In Progress: SentiBERT

To normalize the data we preprocessed it by removing extraneous words, removing punctuation, and lowercasing the text. Then the data was put through the aforementioned encoders and classifiers.

Results / Conclusion

The image above depicts the results from the classifiers. As per the data, the Universal Sentence Encoder performed the best in all categories (Tables 2-4).

Identifying online misbehavior will aid app store platforms and app store developers in determining how the app needs to be updated. Furthermore, this research can aid in identifying misbehavior in other domains.

In the future, we will be incorporating more annotated data to increase the size of the training set. This will help us to train a more generalized model to identify misbehavior. We are also going to identify which category of misbheavior each review is classified as, categories are defined as those shown in Example 1.

HOTSOS 2022
Hot Topics in the Science of Security (HotSoS) Symposium
April 5-7, 2022 | Hosted by the National Security Agency

NC STATE UNIVERSITY
Department of Computer Science