A Two-step Approach for Effective Detection of Misbehaving Users in Chats

Presented by: Esaú Villatoro-Tello

Co-authors: *Antonio Juárez-Gonzales*
Hugo Jair Escalante
Manuel Montes-y-Gómez
Luis Villaseñor-Pineda

Our team…

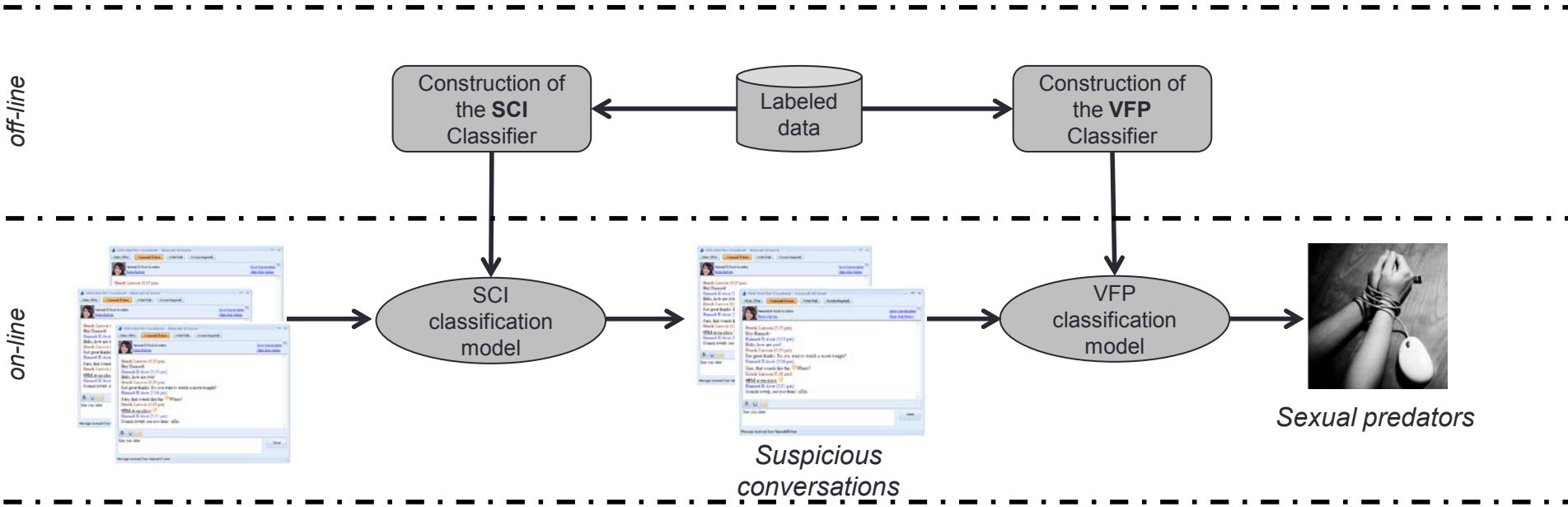
- Language and Reasoning Group
- Information Technologies Dept.
- Universidad Autónoma Metropolitana (UAM)-Cuajimalpa
- Mexico, DF.

- Language Technologies Lab.
- Computer Science Dept.
- Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
- Puebla, Mex.

Sexual Predators Identification

- Based on the following hypotheses:
  - Terms used in the process of child exploitation are categorically and psychologically different than terms used in general chatting
  - Predators usually apply the same course of conduct pattern when they are approaching a child

System Description



- Broadly speaking, our system faces the problem of sexual predators identification as a TC task by means of a supervised approach
- Two main stages: **S**uspicious **C**onversations **I**dentification and the **V**ictim **F**rom Predator disclosure

Our approach

- We face the problem as a TC task by means of a supervised approach

- Our system includes the following modules:
  - Filtering
  - Suspicious Conversations Identification (SCI classifier)
  - Victim From Predator disclosure (VFP classifier)

- Notice that no pre-processing stage is included, this means that we did not remove any *punctuation* marks, *stopwords* and neither apply a *stemming* process.

Filtering

- This stage aims to:
  - Help us focusing only in the most important cases
  - Reduce the computational cost for automatically processing all the information
- It removes the conversation that accomplish:
  - Conversations that had only one participant
  - Conversations that had less than 6 interventions per user
  - Conversations that had long sequences of unrecognized characters
- Results

Number of...	*Original data*	*Filtered data*
Chat conversations	66,928	6,588
Users	97,690	11,038
Sexual Predators	148	136

Training data

Number of...	Original data	Filtered data
Chat conversations	155,129	15,330
Users	218,702	25,120
Sexual Predators	254	222

Test data

SCI Classifier

- The goal of the SCI classifier is to learn a model that allows to distinguishing between **general chatting** from possible cases of online **child exploitation**
  - We labeled as *suspicious conversations* those were at least one predator appears (5,790 non-suspicious, 798 suspicious)
- In other words, the SCI classifier works as a filter, allowing the VFP classifier to focus only on conversations that potentially include **sexual predators**
- Configuration:
  - BOW representation
  - Boolean and TF-IDF weighting

VFP Classifier

- The goal of the VFP classifier is to point at the potential **predator** from a conversation that was previously labeled as a **suspicious chat**
  - We labeled as *victims* those users that had a conversation with a *predator* (194 victims, 136 predators)
- The associated problem is less complex than trying to discriminate between *predators* and *normal users* directly
- Configuration:
  - BOW representation
  - Boolean and TF-IDF weighting

Classification Methods

- Two classifiers from the CLOP toolbox[1] were used in the text classification task:
  - Neural Networks (NN) - The NN classifier was set as a two layer neural network with a single hidden layer of 10 units.
  - Support Vector Machines (SVM) - For the SVM we tried linear and polynomial kernels
- During the development phase we adopted two-fold cross validation to estimate the performance of our methods using training data only

[1] *A. Saffari and I Guyon. Quick Start Guide for CLOP. Technical report, Graz-UT and CLOP-INET, May, 2006.*

Training Results

- SCI Results

Algorithm	Weighting	Accuracy	F-measure
SVM	*binary*	0.9848	0.9361
SVM	*tf-idf*	**0.9883**	**0.9516**
NN	*binary*	**0.9874**	**0.9464**
NN	*tf-idf*	0.9825	0.9254

- VFP Results

Algorithm	Weighting	Accuracy	F-measure
SVM	*binary*	0.9148	0.9138
SVM	*tf-idf*	**0.9259**	**0.9305**
NN	*binary*	**0.9407**	**0.9424**
NN	*tf-idf*	0.9296	0.9337

Test Results

Method	Precision	Recall	F-measure	F-measure(0.5)
Baseline	0.9537	0.4055	0.5691	0.7507
SCI(NN-B)&VFP(NN-TF-IDF)	0.9479	0.7874	0.8602	0.9107
SCI(NN-B)&VFP(NN-B)	**0.9804**	**0.7874**	**0.8734**	**0.9346**

Identifying predators' bad behavior

- It has been shown that every predator follows three main stages when approaching a child:
  - gain access to the victim
  - involve the victim in a deceptive relationship
  - launch and prolong a sexually abuse relationship
- Based on these facts, we believe that if we can generate *language models* from each one of the stages mentioned above, we will be able to find those lines that represent a bad behavior

Our approach…

- We approached the line-detection task as:
  - We automatically divide all the conversations where a predator appears in three sections without considering any type of contextual frontiers
  - Next we generated the language model (*lm*) of the 2nd and the 3rd parts
  - Finally, we compute the perplexity against the *lm* of each one of its interventions, and we delivered as the most distinctive lines of bad behaving those with the minor perplexity value

Conclusions

- Our proposal differs from traditional approaches in that it divides the problem in two stages:
  - The *Suspicious Conversations Identification* (SCI) stage
  - The *Victim From Predator disclosure* (VFP) stage

- Performed experiments showed that it is possible to train a classifier to:
  - Learn those particular terms that turn a chat conversation into a possible case of *online child exploitation*
  - Learn the behavioral patters of predators during a chat conversation allowing us to accurately distinguish victims from *predators*

Thank you!

- Contact information:
  - Esaú Villatoro Tello : evillatoro@correo.cua.uam.mx
  - Manuel Montes y Gómez: mmontesg@ccc.inaoep.mx