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: **Suspicious Conversations Identification** and **the Victim From 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

<table>
<thead>
<tr>
<th>Number of...</th>
<th>Original data</th>
<th>Filtered data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chat conversations</td>
<td>66,928</td>
<td>6,588</td>
</tr>
<tr>
<td>Users</td>
<td>97,690</td>
<td>11,038</td>
</tr>
<tr>
<td>Sexual Predators</td>
<td>148</td>
<td>136</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of...</th>
<th>Original data</th>
<th>Filtered data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chat conversations</td>
<td>155,129</td>
<td>15,330</td>
</tr>
<tr>
<td>Users</td>
<td>218,702</td>
<td>25,120</td>
</tr>
<tr>
<td>Sexual Predators</td>
<td>254</td>
<td>222</td>
</tr>
</tbody>
</table>

Training data       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\textsuperscript{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

## Training Results

- **SCI Results**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Weighting</th>
<th>Accuracy</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>binary</td>
<td>0.9848</td>
<td>0.9361</td>
</tr>
<tr>
<td>SVM</td>
<td>tf-idf</td>
<td>0.9883</td>
<td>0.9516</td>
</tr>
<tr>
<td>NN</td>
<td>binary</td>
<td>0.9874</td>
<td>0.9464</td>
</tr>
<tr>
<td>NN</td>
<td>tf-idf</td>
<td>0.9825</td>
<td>0.9254</td>
</tr>
</tbody>
</table>

- **VFP Results**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Weighting</th>
<th>Accuracy</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>binary</td>
<td>0.9148</td>
<td>0.9138</td>
</tr>
<tr>
<td>SVM</td>
<td>tf-idf</td>
<td>0.9259</td>
<td>0.9305</td>
</tr>
<tr>
<td>NN</td>
<td>binary</td>
<td>0.9407</td>
<td>0.9424</td>
</tr>
<tr>
<td>NN</td>
<td>tf-idf</td>
<td>0.9296</td>
<td>0.9337</td>
</tr>
</tbody>
</table>
# Test Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>F-measure(0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.9537</td>
<td>0.4055</td>
<td>0.5691</td>
<td>0.7507</td>
</tr>
<tr>
<td>SCI(NN-B)&amp;VFP(NN-TF-IDF)</td>
<td>0.9479</td>
<td>0.7874</td>
<td>0.8602</td>
<td>0.9107</td>
</tr>
<tr>
<td>SCI(NN-B)&amp;VFP(NN-B)</td>
<td>0.9804</td>
<td>0.7874</td>
<td>0.8734</td>
<td>0.9346</td>
</tr>
</tbody>
</table>
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 patterns 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