Towards the Early Detection of Child Predators in Chat Rooms: A BERT-based Approach

Sinchana Kumbale
Manipal Institute of Technology
sinchana.kumbale@gmail.com

Smriti Singh
Manipal Institute of Technology
smritisingh26@yahoo.com

Abstract

The identification of child predators in chat rooms has been a concerning issue for a long time. The COVID-19 pandemic has caused an increase in internet usage and thus, an increase in the incidents of child exploitation. Hence, there is an urgent requirement to develop a system that can protect children from sexual exploitation. In this work, we propose a BERT-based approach to the early detection of child predators using the PAN’12 dataset. Our methodology makes use of a minimum number of dialogues and is able to identify child predators in chats within 45 minutes. An extensive comparison against other deep learning methods suited to text classification tasks and a detailed qualitative analysis makes the case for our proposed methodology.

1 Introduction and Related Work

The global COVID-19 pandemic, which has restricted scores of people to their homes, has escalated internet usage. The massive engagement of children online combined with the isolation has raised valid concerns on the increased exposure of children to sexual predators (Parks et al., 2020). A recent report stated that the incidents of child exploitation nearly doubled during the pandemic (Today, 2020). The psychological impact faced by children and adolescents who are subject to sexual abuse (Guerra et al., 2018) is often deep-rooted. This impact could be mitigated via the early detection of child grooming and predatory intent in online chat rooms. Early detection of predatory intent could save children from becoming the victims of sexual abuse. The course of the life of many young adults could change through such early mediation. The automation of this task is necessary, since there are too many conversations that occur online for each of them to be monitored manually. Our model aims to detect predatory intent in online conversations early on, thereby packing the potential to be deployed in real time and save children from sexual predators.

Previous works along the same lines employ several approaches, including Chained Classifiers (Escalante et al., 2013), support vector machines (SVM) (Parapar et al., 2014). Recent efforts leveraged random forests (CARDEI and Rebedea, 2017) and Naive Bayes classifiers (Dhouioui and Akaichi, 2016). The state of art paper (Fauzi and Bours, 2020) applied ensemble learning with soft voting. All previous approaches mentioned display a high level of accuracy but are applicable on completed conversations and thus have little real-life usage. As explored by (Fauzi and Bours, 2020) and their suggestions for future work, our model focuses on the early detection of predators by minimising the number of dialogues and reducing the time duration of conversations. Our work most closely resembles that of (Vogt et al., 2021), wherein they investigate a BERT-based approach to sexual predator detection using dynamic conversation length. However, our method focuses on conversation duration as opposed to conversation length and investigates whether simpler pipelines can be used to leverage a competitive accuracy and a computationally efficient methodology. Our motivation behind investigating a potentially computationally efficient methodology is to develop a model that is not only accurate, but deployable in real time in the near future.
2 Methodology

In this section, we describe the dataset used, our preprocessing pipeline and the steps taken to develop our model. We also elaborate on the experiments we conduct as a part of our study.

2.1 Corpus:
The corpus used is from the first international competition on sexual predator identification (PAN’12), held within the Conference and Labs Evaluation Forum (CLEF) (Zenodo, 2012). It consists of several conversations including those between known predators and their victims. We tailor the dataset to this specific problem statement by reducing each conversation to under 45 minutes. This time duration was empirically obtained, by manually scouring through 100 randomly sampled conversations in the dataset and averaging the time taken for the conversation to take a predatory turn. Further, conversations with lesser than 8 dialogues were removed to ensure that the dataset only included contextually rich communication. This number has also been obtained empirically, as described above.

Next, we investigate the various possibilities of annotating the data. Annotating at a conversation level is logically flawed as only parts of a conversation can be termed predatory. Thus, we annotate conversations at a message level, to better train our model to detect predatory language. The annotation process has two stages. First, we identify the messages sent by the 142 unique predator IDs as given in the training corpus of the dataset. Then, we manually check these messages and identify the messages that are non predatory. We define a message as predatory if it contains sexual content, inappropriate advances, or reference to nudity/sexual activity. Thus, our modified dataset consists of 39548 conversations and 102616 messages, 48280 of which are predatory.

We preprocess the modified dataset using the following steps:

- We cleaned the text by tokenising it and removing unwanted tags
- We replaced contradictions using the python package contradictions, which utilises a dictionary to replace it with the corresponding word sequence.
- We converted all characters to lowercase and removed stop words, reducing instances of low information words.
- Then, we lemmatise and normalise the text using the NLTK library as chat-based conversations are usually abound with abbreviations.

After these preprocessing steps, we fed the messages into the model aiming to detect predatory intent. The detection of predators in a limited window is a complicated task even for humans, so we leverage various deep learning methods during modelling.

2.2 Modelling

We experiment with three different deep learning models, namely, BERT (Devlin et al., 2018), a Bi-LSTM and an RNN, in an effort to identify the best performing model. We choose these models due to their high accuracy on text classification tasks. RNN architecture utilises previous cases and inputs to formulate newer states. The ability to "remember" previous inputs deems recurrent neural networks a suitable choice for this task. A bi-LSTM is a special kind of RNN, that can also decide which information to remember and which to forget. It can read text in a bidirectional manner, giving it a deeper understanding of text compared to other models. BERT’s model architecture is a multi-layer bidirectional transformer (Devlin et al., 2018) with a large number of hidden layers. Due to its ability to read text bidirectionally, it also has a deeper understanding of semantic relations between text, making it suitable for many text classification problems.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.
Our modelling consists of two separate experiments. On one hand, we train the aforementioned models on the modified dataset. Then, to establish a standard for comparison, we also train these models on the original, unshortened dataset. To offer fair comparison, we ensure that we remove conversations with lesser than 8 dialogues even in the unshortened dataset to retain only contextually relevant conversations.

For BERT, we use the base model, choose the recommended learning rate (Devlin et al., 2018) of $2e^{-5}$ and train over 1 epoch. We evaluate the models on the test-set taken from the PAN’12 dataset across four accuracy metrics: Accuracy, precision, recall and F1 score. Our results are showcased in Table 1.

### 3 Results

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Accuracy</th>
<th>F1 Scores</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>0.74 ± 0.00</td>
<td>0.74 ± 0.02</td>
<td>0.735 ± 0.03</td>
<td>0.735 ± 0.02</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>0.64 ± 0.03</td>
<td>0.62 ± 0.10</td>
<td>0.67 ± 0.05</td>
<td>0.63 ± 0.20</td>
</tr>
<tr>
<td>RNN</td>
<td>0.66 ± 0.03</td>
<td>0.645 ± 0.07</td>
<td>0.69 ± 0.06</td>
<td>0.655 ± 0.20</td>
</tr>
</tbody>
</table>

Table 1: Modified Dataset

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Accuracy</th>
<th>F1 Scores</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>0.73 ± 0.00</td>
<td>0.735 ± 0.005</td>
<td>0.73 ± 0.01</td>
<td>0.735 ± 0.005</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>0.65 ± 0.03</td>
<td>0.635 ± 0.08</td>
<td>0.68 ± 0.07</td>
<td>0.65 ± 0.21</td>
</tr>
<tr>
<td>RNN</td>
<td>0.69 ± 0.01</td>
<td>0.69 ± 0.04</td>
<td>0.70 ± 0.08</td>
<td>0.695 ± 0.14</td>
</tr>
</tbody>
</table>

Table 2: Original Dataset

On comparing the various models we train on our modified dataset, we find that BERT outperforms the other models with an average accuracy gain of 9% (Table 1). To further examine the accuracy of the model, we perform qualitative error analysis by randomly sampling wrongly classified messages and find that the primary confounding variables are context and consent. Messages like, “Would you like to meet soon?” and "I love sucking your p**ssy" are marked as predatory. While it is very likely that such messages can be predatory, they could also be consensually exchanged. We attribute these errors to the context of similar language in the training set.

Further, we find that training on our dataset actually yields slightly higher accuracy when compared to training on the original, unshortened dataset, as illustrated in Table 2. This proves that shortening the conversation using our empirically established time limit does not have a negative effect. Thus, our proposed methodology is able to improve accuracy while improving computational efficiency at the training stage.

### 4 Conclusion and Future Work

Identification of predators with limited context and a defined time limit is a complicated task even for humans. In this paper, we present a BERT-based approach to identify predatory intent early on (45 minutes from the start of the conversation) within online chats. We believe that future endeavours can focus on improving efficiency while training the model to retain context-specific information. The deployability of this model in real time should also be investigated.

### References


