Machine Learning Algorithms for Predators Detection in Online Chat Conversations

Ken Jen Lee, Romario Timothy Vaz, Yipeng Ji
{kj24lee, r4vaz, y43ji}@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Abstract

This paper studies three machine learning algorithms for identifying suspicious online conversations, defined by the presence of sexual predators from a corpus of online conversations. It has been found that 13% of youth Internet users received unwanted sexual solicitations, with an estimate of up to 3.72 million cases in 5 years by (Wolak, Mitchell, and Finkelhor 2006). This is despite many youths not disclosing their encounters, making child luring a serious social issue. Three approaches were selected to identify suspicious online conversations: 1) A linear Support Vector Machine (SVM); 2) A Generative Adversarial Network (GAN) -assisted Multi-Layer Perceptron (MLP); 3) A one-dimensional Convolutional Neural Net (CNN). Additionally, the word features that were identified as being the most prominent when detecting suspicious online conversations were derived by analyzing the CNN. Among the models, it is shown that the CNN performs the best, slightly outperforming the SVM, followed by the GAN-assisted MLP. Moreover, interesting insights were made by the features identified as being the most significant. Potential important information on sentiments of suspicious online conversations were revealed, alongside locations and other personal information of their authors.

Introduction

In this "Machine Learning for Social Good" paper, conventional and novel machine learning algorithms are used to detect sexual predators in online conversations. In a recent study conducted by (Anderson and Jiang 2018), 95% of American teens aged 13-17 have access to a smartphone and 45% of them are on the internet almost constantly. According to the National Center for Missing and Exploited Children operated CyberTipline, a public mechanism in the United States for reporting instances of suspected child sexual exploitation, online enticement is among the reasons accounting for most of the 10 million reports received in 2017 (for Missing & Exploited Children 2017). Due to the natural vulnerability of teenagers, sexual predators often seek their internet-initiated sex crimes victim in an online chat room and build relationships by offered gifts or money and naked pictures exchange (of Justice 2018). As such, detecting sexual predators in the early-stage of online conversations before the possible physical crimes happen are an important issue that should be solved. By finding potential solutions to this problem and applying them in real life scenarios, we will be able to protect young children and teenagers from online sexual grooming and any further possible physical harms, thus making the internet a safer place.

This paper explored multiple algorithms to identify online conversations with sexual predators, including a Support Vector Machine (SVM), a Generative Adversarial Network (GAN)-assisted Multilayer Perceptron (MLP) and a Convolutional Neural Net (CNN). The (PAN 2012) Sexual Predator Identification training and testing datasets were used for this paper. Accuracy, precision and recall were used as performance evaluation metrics of the algorithms. It is hoped that by completing this project, attention could be brought to the urgent problem that online predators pose and for the usefulness of machine learning algorithm in tackling this problem to be proven.

The contributions of this paper can be summarized in the following points:

- Validating the effectiveness of SVM in classifying conversations containing predators.
- Investigating the effectiveness of non-conventional text classification methods including a GAN-assisted MLP and a one-dimensional CNN to identify online predators in conversations.
- Identifying word features that are the most prominent in detecting suspicious conversations and analyze them for meaningful insights.

Related Work

In (Javier Parapar and Barreiro 2012; Esa Villatoro-Tello and Villaseor-Pineda 2012; Morris and Hirst 2012; Meyer 2015), SVM was used for the purpose of detecting sexual predators. The same dataset was used by this paper as well, allowing for better insights into the performance of SVMs in tackling the same issue. Particularly, (Esa Villatoro-Tello and Villaseor-Pineda 2012) proposed a two stages structure involving the Suspicious Conversation Identifier (SCI) and Victim From Predator disclosure (VFP). While the former separates conversations that involve sexual predators, the latter takes in conversations identified by the SCI to have sexual predators and identifies the specific author(s) who may be sexual predators. The authors also compared the performance of SVM and a single hidden layer neural net for both stages. The reason why this paper was chosen specifically to model after is the fact that they scored the highest in the PAN 2012 competition to identify predators (PAN 2012).

On the other hand, (Hojjat Aghakhani and Vigna 2018) details GAN's use for the purpose of textual anomaly detection, providing valuable information regarding the design of the GAN-assisted MLP. Moreover, (Kim 2014; Yang Liu and Zhang 2018; Hughes M and T 2017) show both CNNs and altered CNNs can perform well on complicated text classification tasks, such as sentence classification and medical text classification. In particular, (Kim 2014) demonstrates that basic CNN models, such as feed-forward CNNs, can be relatively simple yet very powerful and effective models for text classification which inspires us to apply a one-dimensional CNN for our task.

(authors 2018; Dan Li and Ng 2019; Houssam Zenati and Chandrasekhar 2018) propose different approaches for non-textual anomaly detection using various GAN structures, which is useful in providing ideas on how GAN can be optimized for our task in an anomaly detection context. Particularly, (Dan Li and Ng 2019) explores the usage of LSTM-RNN with GAN for multivariate time series tasks, but looks to be useful for textual tasks as well. Lastly, (Hugo Jair Escalante and y Gomez 2013; Darnes Vilario and Len 2012) are papers that aims to identify sexual predators using the PAN 2012 dataset as well, using other approaches like Bayes classifier (Darnes Vilario and Len 2012) and specific feature engineering (Hugo Jair Escalante and y Gomez 2013). These will be useful for an idea of alternative approaches to the same issue.

Data Description

In this section, more details are provided on where the data for this paper comes from and the process of filtering this data before they are used for training purposes.

Data Source

The dataset provided for (PAN 2012) under the category Sexual Predator Identification was used. Both training and testing data are available and both datasets contain two relevant files. The first is an XML file containing all conversations. Each conversation has a variable number of messages and each message contains the text sent and author's id. The second file contains author ids of authors who have been labelled as sexual predators. The training set contains 66,927 conversations in total, with 142 predators and 97,547 authors who are non-predators. The testing data, on the other hand, contains 155,128 conversations with 254 predators identified. The provision of author ids of users who are sexual predators is significant as that provides this project with properly labelled data. This allows models to understand the structure of conversations where sexual predators are involved. From those conversations, further classifications can be made as messages by all sexual predators in these conversations can be identified as well. There are also some limitations for this dataset, including its imbalance as there are way more normal conversations than conversations that involve sexual predators.

Data Preparation

In the data pre-processing stage, unlike many other text analysis studies, we preserved chats with emoticons, abbreviations, grammatical errors and intentional misspelled words and phrases because those provide valuable information for distinguishing perpetrators. For instance, :-* (kiss) and xoxo are both indications of soft sexual introductions as shown by (Esa Villatoro-Tello and Villaseor-Pineda 2012). In the data pre-filtering stage, conversations in the training dataset meeting any of the following criteria were removed: 1) conversations with only one author; 2) conversations with each author having six or less messages; 3) conversations with long sequences of unrecognized characters as used in (Esa Villatoro-Tello and Villaseor-Pineda 2012). The third criterion is defined as having more than 60% of any particular message with lengths exceeding 20 characters be made up of non-alphanumeric characters. After filtering, 78.7% conversations involving only non-predators (normal conversations) and 55.3% conversations involving sexual predators (suspicious conversations) were removed, this means 52,224 out of 66,927 conversations were removed with 14,703 remaining. Furthermore, while 74.4% of authors who are not predators were removed, only 3.5% of predators were removed during the dataset. This means the filtered data has a total of 25,099 authors, among which 137 have been labelled as predators. Then, for each conversation, all messages were concatenated into a single string and given a label of 1 if one or more authors in the conversation is a predator, or 0 otherwise. Lastly, the vector of strings (one for each conversation) is vectorized using the TF-IDF weighing scheme. This resulted in a dataset with 121,394 features.

Predictive Models

Linear SVM

The first model implemented is a linear SVM. SVMs have been proven by (Joachims 1998) to be an effective algorithm for text classification purposes. Primary reasons for this include the high number of dimensions of the dataset (121,394 dimensions) and the sparse nature of the feature values. As such, a linear SVM with a square hinge loss, solving the dual optimization problem with a C coefficient of 2.90 was found to be the best using 10-fold cross validation.

GAN-Assisted MLP

The intuition behing using this algorithm is the unbalanced amount of normal and suspicious conversations. Particularly, there are many more normal conversations than suspicious conversations in the training dataset. Hence this algorithm aims to create fake suspicious conversations to counter that imbalance during training.

This algorithm uses a pre-trained GAN to train a MLP with a hidden layer containing 10 units. The algorithm contains the generator network, G, that produces fake suspicious conversations, the discriminator, D', that differentiates fake

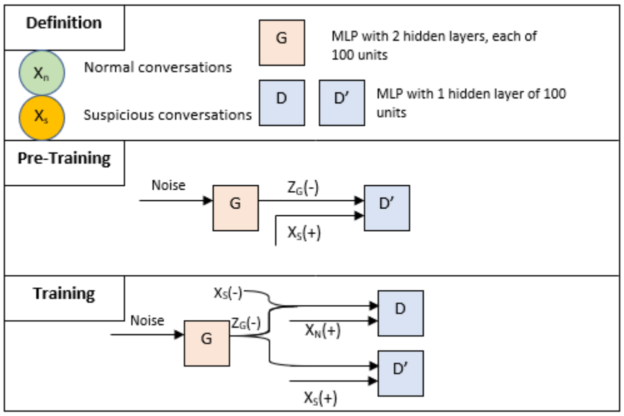

Figure 1: GAN-assisted MLP Structure and Training Procedure

suspicious conversations produced by G from actual suspicious conversations, and the MLP, D, that differentiates between suspicious conversations and normal ones. While D' also has a single hidden layer with 10 units, G has 2 hidden layers, each with 100 units. Initially, G and D' are pretrained with their respective objective functions below:

$$\begin{gathered} min(-_{S\sim G_\alpha}[logD'(S)]) \\ min(-_{S\sim X_S}[logD'(S)] -_{S\sim G_\alpha} [1 - logD'(S)]) \end{gathered} \tag{1}$$

Then, G, D and D' are all trained together. While D' has the same objective function, the objective functions of G and D are as follows:

$$\begin{gathered} min(-_{S\sim G_\alpha}[logD'(S)] -_{S\sim G_\alpha} [logD(S)]) \\ min(-_{S\sim X_N}[1 - logD(S)] -_{S\sim X_S} [logD(S)] \\ -_{S\sim G_\alpha}[logD(S)]) \end{gathered} \tag{2}$$

It should be noted that X_N denotes the set of normal conversations in the training dataset, X_S denotes the set of suspicious conversations in the training datset and G denotes the set of data generated by generator G when fed with input α. Since the initial training dataset had a lot more normal conversations than suspicious ones, the effectiveness of G in helping to train D' by generating feature vectors of fake suspicious conversations was investigated. To help with training, a few techniques were used. First, the input data was normalized to a range of $[-1, 1]$. Moreover, the labels of normal and suspicious conversations were flipped, random noise was added to real inputs (not generated by G) into D and the noise input for G was sampled randomly from $\mathcal{N}(\alpha; 0, 1)$. Furthermore, G was trained using the Adam Optimizer, while D and D' were trained using the Stochastic Gradient Descent technique. Hence, the feedback from D' guides G to produce fake conversations that appear as real suspicious conversations, while feedback from D guides G to produce fake conversations that are suspicious and not normal. On the other hand, outputs from G used to train D allows D to generalize better for suspicious conversations since there are relatively few suspicious conversations in the training dataset compared to normal conversations.

CNN

The CNN used in this paper consists of two stages, each containing a one-dimensional convolution layer, an activation function and random dropout, with two fully connected layers at the end. For the convolution layers, the first uses filters of width sixteen, one input channel and two output channels, while the second convolution layer uses filters of width eight, eight input channels and sixteen output channels; both layers had their stride hyperparameter set to two. It was trained using a cross-entropy loss function with the Adam Optimizer.

Evaluation

This section introduces the metrics used, the process of performance evaluation and the peformance of the three algorithms.

Performance Metrics

There are three performance metrics used to evaluate each of the three algorithms. They are accuracy, precision and recall.

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \end{aligned} \tag{3}$$

Where TP is the true positives, FP is the false positives, TN is the true negatives and FN is the false negatives. In addition to accuracy, precision and recall are used as well to get a better idea of how each algorithm performs since there is a huge imbalance in number of normal conversations over the number of suspicious conversations in the test dataset.

Evaluation Procedure

The training and testing dataset are provided directly by the data source. For the SVM, 10-fold cross validation was used to find the best set of hyperparameter. For the GAN-assisted MLP and CNN, 30% of training data was used as validation data to evaluate performance during training; random starts were also used to find the best model.

In addition, for the CNN, Algorithm 1 was used to evaluate the top 100 word tokens in the TF-IDF vectorizer by analyzing the weights assigned to each of the word feature and performing manual convolution using weights in the first convolution layer of the CNN.

Results

It can be seen that the CNN performs the best, followed closely by the linear SVM. The GAN-assisted MLP performed especially bad (see Table 1). This is possibly due to a generator network that is not adept at producing fake feature vectors of suspicious online conversations as it failed to learn the suspicious conversations' probability distribution well, which could be attributed to potential flaws in the pre-training phase.

Algorithm 1: Extraction of Most Important Word Tokens

```
Result: top 100 words
filters = get weights of filters in first convolution layer of CNN;
data = feature vectors of suspicious conversations from training data squashed into one row;
dict = empty dictionary;
for filter in filters do
    for (featureIndex = 0; featureIndex + len(filter) < len(data); featureIndex += 2) do
        for j in range(len(filter)) do
            // perform convolution with stride of two and accumulate products for each feature;
            dict[featureIndex] += data[featureIndex] * filter[j];
wordFeatures = get words used as features in vectorizer;
return wordFeatures[top 100 keys with highest values from dict];
```

Table 1: Performance of Algorithms

ALGORITHM	ACCURACY	PRECISION	RECALL
Linear SVM	98.51%	76.41%	55.20%
GAN-Assisted MLP	83.35%	6.09%	36.98%
CNN	98.60%	76.19%	61.54%

The poor recall results can be attributed to the way the test dataset was labelled. Specifically, all conversations that contain one or more authors who were identified as sexual predators were labelled as suspicious conversations. However, this leads to conversations involving sexual predators but not the expected interactions to be labelled as suspicious as well. For example, many test conversations containing only one message like ”Hey decoy name, I’m back in town... talk to you soon” written by a sexual predator was labelled as suspicious. However, in reality, the conversation does not seem suspicious based on the messages alone without prior knowledge that the author is a sexual predator. Similarly, many false negatives are conversations that seem normal but involve sexual predators. One main reason is these ”normal conversations” are often one of the many conversations between the sexual predators and the victims. While some has content expected of sexual predators, others do not. Other examples include ”hi hun” and ”wat up hey nothin much having fun on a saturday”(4 messages from 2 authors concatenated). The poor precision results can be associated to the many short conversations in the test data causing false positives like ”I’ll be back ..” and ”hi love you ttyl:)”. This may suggest that short conversations should be excluded from this classification task as not much information is present in them.

Other than that, important insights were gained by extracting the most important word features in the TF-IDF vectorizer using the trained CNN (see Algorithm 1). The top 100 words can be broadly grouped into several categories. The first is sexual words like *chest*, *squirt* and *masturbating*. There are also common misspellings of words used online or internet slang like *nooo*, *ut* (means ”you there?”), *frnd* and *prety*. However, more interesting words include *petaluma*, which is a place in California, USA. Although its crime rates are lower than the national crime rates, Petaluma has more rape and assault cases than the national average as reported in (AreaVibes 2018), suggesting a possible correlation. Moreover, words that might reflect the themes of suspicious online conversations have been extracted as well, including *secret*, *younger*, *hangout*, *control*, *power*, *rough*, *juicy* and *fantasize*. Interestingly, numbers like 40 and 48 are extracted; they have been used in expressing an author’s age and phone numbers in suspicious conversations of the training dataset. Lastly, there are (to the authors’ knowledge) irrelevant words like *iam*, *trade* and *square*. The full list of the top 100 word features is also provided (see Figure 2).

```
hoping          shouldnt        listening
secret          evening         forwards
younger         ure             poppit
nooo            square          ann
neways          watched         dating
power           trade           pass
cloths          suposed         pays
struck          48              station
living          masturbating    prety
handsome        surprise        perhaps
hangout         positive        battry
iam             obviously       fact
town            goofy           fuked
stayin          40              ouch
sarahs          feed            200
fav             ut              fantasize
speak           schools         petaluma
1230            swear           ment
july            ti              goto
sposed          forgive         alrite
control         distance        rid
musta           shoot           highway
def             meen            gah
dogs            juicy           potty
ofcourse        paid            privacy
breakfast       oo              frnd
otherwise       rough           drank
chest           sed             shots
freeway         thur            allway
arbor           tehn            swetie
yelled          killing         eats
coz             known           definitely
squirt          gfs
lest            soooooo
```

Figure 2: Top 100 word features

Discussion

In this paper, it is shown that machine learning is highly suitable for the task of identifying sexual predators in online conversations. While SVM has once again shown to be very effective, the CNN, which is not usually used for textual classification tasks, has proven to be appropriate as well.

However, the same cannot be said for the GAN-assisted MLP.

Furthermore, the CNN filters provides valuable insights about word features that are assigned the highest priorities when detecting predators in online conversations. Applying zero pre-processing on the dataset proved useful as many misspellings and internet slang were seen as important features.

Conclusion

Online predators pose an urgent important social problem that machine learning is well-suited to resolve. Not only is machine learning a very suitable tool for classifying potential suspicious conversations, this paper has also shown that it can serve other purposes like generating a relevant list of keywords to watch out for by analyzing the CNN. Future efforts will be devoted to improving and tuning the various algorithms, especially the generator network of the GAN-assisted MLP, for better performances. The analysis of the top word features also suggests that more precise information about predators like their location and themes of conversations could be extracted as well with more research. Moreover, if deployed for a real-world chat room or chat application, modifications could be made such that the model can continuously learn from live conversations.

It is the sincere hope of the authors that these findings will encourage the use of machine learning by authorities or relevant parties for the task of identifying online predators, allowing the youth of today to have a safer and improved Internet browsing experience.

References

Anderson, M., and Jiang, J. 2018. Teens, social media and techonology 2018. *Pew Research Center*.

AreaVibes. 2018. Reported annual crime in petaluma.

authors, A. 2018. Anomaly detection with generative adversarial networks. *International Conference on Learning Representations 2018*.

Dan Li, Dacheng Chen, J. G., and Ng, S.-K. 2019. Anomaly detection with generative adversarial networks for multivariate time series. *arXiv preprint ArXiv:1809.04758v3*.

Darnes Vilario, Esteban Castillo, D. P. I. O., and Len, S. 2012. Information retrieval and classification based approaches for the sexual predator identification. *International Conference of the Cross-Language Evaluation Forum,2012*.

Esa Villatoro-Tello, Antonio Jurez-Gonzlez, H. J. E. M. M.-y.-G., and Villaseor-Pineda, L. 2012. A two-step approach for effective detection of misbehaving users in chats. *PAN at CLEF 2012*.

for Missing & Exploited Children, N. C. 2017. Key facts.

Hojjat Aghakhani, Aravind Machiry, S. N. C. K., and Vigna, G. 2018. Detecting deceptive reviews using generative adversarial networks. *arXiv preprint ArXiv:1805.10364*.

Houssam Zenati, Chuan-Sheng Foo, B. L. G. M., and Chandrasekhar, V. R. 2018. Efficient gan-based anomaly detection. *arXiv preprint ArXiv:1802.06222*.

Hughes M, Li I, K. S., and T, S. 2017. Medical text classification using convolutional neural networks. *arXiv preprint ArXiv:1704.06841*.

Hugo Jair Escalante, Esau Villatoro-Tello, A. J. L. V., and y Gomez, M. M. 2013. Sexual predator detection in chats with chained classifiers. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Javier Parapar, D. E. L., and Barreiro, A. 2012. A learning-based approach for the identification of sexual predators in chat logs. *PAN at CLEF 2012*.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint ArXiv:1408.5882v2*.

Meyer, M. 2015. Machine learning to detect online grooming.

Morris, C., and Hirst, G. 2012. Identifying sexual predators by svm classification with lexical and behavioral features. *PAN at CLEF 2012*.

of Justice, T. U. D. 2018. Raising awareness about sexual abuse facts and statistics.

PAN. 2012. Author identification. *PAN at CLEF 2012*.

Wolak, J.; Mitchell, K.; and Finkelhor, D. 2006. Online victimization of youth: Five years later.

Yang Liu, Lixin Ji, R. H. T. M. C. G., and Zhang, J. 2018. An attention-gated convolutional neural network for sentence classification. *arXiv preprint ArXiv:1808.07325*.