Detecting Sexual Predatory Chats by Perturbed Data and Balanced Ensembles

Parisa Rezaee Borj, Kiran Raja, Patrick Bours
Norwegian University of Science and Technology (NTNU), Norway
E-mail: {parisa.rezaee;kiran.raja; patrick.bours} @ntnu.no

2021 International Conference of the Biometrics Special Interest Group (BIOSIG) | 978-1-6654-2693-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIOSIG52210.2021.9548303

***Abstract*—Securing the safety of the children on online platforms is critical to avoid the mishaps of them being abused for sexual favors, which usually happens through predatory conversations. A number of approaches have been proposed to analyze the content of the messages to identify predatory conversations. However, due to the non-availability of large-scale predatory data, the state-of-the-art works employ a standard dataset that has less than 10% predatory conversations. Dealing with such heavy class imbalance is a challenge to devise reliable predatory detection approaches. We present a new approach for dealing with class imbalance using a hybrid sampling and class re-distribution to obtain an augmented dataset. To further improve the diversity of classifiers and features in the ensembles, we also propose to perturb the data along with augmentation in an iterative manner. Through a set of experiments, we demonstrate an improvement of 3% over the best state-of-the-art approach and results in an F_1-score of 0.99 and an F_β of 0.94 from the proposed approach.**

***Index Terms*—Imbalanced dataset, Online Conversation, Predatory Detection, Sampling**

I. Introduction

Children are vulnerable due to the new sexual norms caused by advanced technology and increased time spending on online communities where chats with unknown persons are fully possible. The children can thus be targeted by sexual predators by convincing text messages [1]. Detecting and identifying the predatory chats has been a major problem for parents and law enforcement agencies. However, predatory conversation detection is a complex problem as the offenders apply many techniques to avoid disclosure. The predators may not necessarily discuss about sex in the conversations, but apply different strategies and variations in time, type, and intensity to keep the victim interested and eventually exploit them. The process of gaining the trust of victim is usually called grooming [2].

A common challenge in detecting online sexual predators is collecting the data as the chat providers do not make it publicly available, and accessing them requires legal permission. Of the few available datasets like PAN 2012 competition [3], one can observe the common problem encountered in most machine learning problems [3]. The datasets are heavily imbalanced due to normal conversations representing higher proportions than the predatory conversations. In reality, the percentage of sexual predatory data is 0.25% of the total online data that causes many problems for designing an automated machine learning driven detection models [4]. Such composition of dataset makes the predatory detection a challenging problem as handling the imbalanced dataset for the sexual predatory detection is critical.

In this work, we present a new approach for detecting predatory chat detection by providing a new strategy in handling the imbalance to provide a new approach. Specifically, we present an approach which first creates a balanced class distribution by increasing the minor class with a set of augmented and perturbed data. The balanced class distribution is increased until a 50% balance is obtained by simply augmenting and perturbing the data. With the refined class distribution, we create an ensemble of HistogramBoostedGradient classifiers which directly benefit from the augmented and perturbed data in selecting different set of features for creating ensembles. With the set of experimental validation, we evaluate the proposed approach on PAN 2012 [3] dataset where the proposed approach outperforms the existing approaches. The proposed approach results in a precision of 99%, a recall of 99% and a $F_{0.5}$ score of 94% with a gain of 3% over the recent work which reported a recall of 96%. In the rest of this paper, we first present briefly detail the dataset employed and discuss the imbalanced nature of the dataset in Section II. We then list out few related works which have tried to address the imbalanced nature of sexual predatory data for the convenience of the reader in Section III. We present the proposed approach in Section IV followed by the discussion on results in Section V. To the end, we make concluding remarks and list out few potential future works.

II. Database for Sexual Predatory Detection

A chat conversation typically is one of three types of conversations such as (a) a conversation without sexual topics, (b) a conversation between adults on sexual topics, or (c) a conversation between a predator and a minor victim which is considered as a predatory conversation. The PAN 2012 [3] competition dataset deals with the third category and the data contains the conversations between police officers who pretended to be minors and convicted predators extracted from the PJ website (http://www.perverted-justice.com/). The data also contains the ordinary chat without any sexual content extracted from http://www.irclog.org, and sexual conversation between consenting adults from Omegle (www.omegle.com). In addition to the conversation data, the data also consists of a unique conversation ID to distinguish between the conversations. Each message in a conversation further includes an author ID, a timestamp, and the text of the message. In training

978-1-6654-2693-0/21/$31.00 ©2021 IEEE

data, there are 951 predatory conversations and 8477 non-predatory conversations. The test data contains 1697 predatory samples and 19922 non-predatory conversations. More detail about the applied data and the pre-processing method can be found in [3] and [5].

A. Constraints of Dataset

As with any other type of data investigation, predatory detection requires pertinent data. The amount of predatory data is much lower than the normal chatlogs, making it challenging to find appropriate subset of data. Further, analyzing the data, we note the heavily imbalance in the data where predatory data is less than 0.25% of the total data [4]. Such imbalance leads to sub-optimal classifiers favoring one class over the other resulting either in underfitting or overfitting. When the training data is highly imbalanced, it becomes more critical as the class with fewer samples is severely under-sampled and causes to not capturing the complete information of the given data. If one does not consider the class imbalance problem, the learning techniques can be overwhelmed by the majority class, and the minority class will be easily ignored. An imbalanced classification problem is a problem where the datasets have skewed distributions. It has several characteristics, including class overlapping, small sample size, and small disjuncts [6]. A predatory dataset can suffer from all these characteristics as there are many overlaps and disjuncts between a predatory conversation and a non-predatory one. In addition, the number of predatory conversation samples is much lower than the non-predatory ones. Also, a chat conversation might contain some sentences or topics that are common in both predatory and non-predatory talks.

III. Related Works

Earlier research works mainly have used conventional methods for sexual predatory detection disregarding the imbalanced dataset [3], [5], [7]–[11]. However, we restrict our focus to few sample works and focus on works that deal with data imbalance problem within predator detection. Cardei et al. [12] tested several techniques for coping with the imbalanced data in sexual predatory detection, such as cost-sensitive technique and sampling techniques including BalanceCascade [13], and CBO - a clustering-based method using k-means [14]. The authors found that the cost-sensitive model where a cost matrix gave a penalty for misclassifying gained the best performance experimenting on PAN 2012 dataset [3]. Their proposed model had two stages where it investigated behavioural features that cover the users' behaviour on the online platform. It considered the ratio of questions, underage expressions, slang words, and the bag of word feature vectors and obtained an $F_{0.5}$ score of 0.95 [12]. Zuo et al. [15] presented an adaptive fuzzy method for artificial neural networks (ANNs) to address the imbalanced data in sexual predatory detection. They used conventional fuzzy inference based on dense rule and fuzzy rule interpolation to handle the imbalanced dataset in the sexual predatory detection problem. Their method was a combination of an adaptive fuzzy inference-based activation function with the artificial neural networks (ANNs) that extracted BoW and TFIDF as feature sets, classified the data sets, and gained an accuracy of 0.766.

IV. Proposed Approach

The overall pipeline of the proposed approach is illustrated in the Figure 1. The proposed approach starts with the pre-processing of the data, followed by feature extraction using Word2Vec [16] and the proposed strategy of learning the ensemble classifiers as detailed below in this section. As the data is heavily noisy, we first preprocess the data to eliminate the irrelevant entries from PAN 2012 dataset. Based on the common properties of the predator victim contacts, we assert that these kinds of conversations have only two authors. We therefore eliminate all the other conversations that involve multi-parties or have only one author. Further, we discard all the conversations with less than seven messages as such conversations contain too little information to be classified as either predatory or non-predatory. Further, as another refinement, we analyze chat messages to eliminate non-English words that did not provide any special meaning or do not follow standard grammar in a remote manner, have many slang and emoticons. To keep the information as much as it is possible, no stemming or lemmatization in the preprocessing of the data was performed. We then extract the features the from the chat logs that cover the word relationships in different contexts with a low dimensional feature vector. In order to fully exploit the word analogies, we extract features using the Word2Vec embedding model with pre-trained networks with a 300 dimensional vectors. Word2Vec provides distributed representation of the text data which we further use to design the classifier.

Algorithm 1: Pseudocode for Proposed Approach

```
initialization : T iterations, Number of base estimators,
  Number of bins for HistogramBoostedGradient;
procedure CLASS DISTRIBUTION BALANCE;
t ←∈ Titerations
while K do
    Compute class distributions;
    Compute balanced hybrid sampling;
    Compute the expected class distribution (number
    of samples from each class);
    Compute the intra-class balanced sampling weights
    by inversing prediction error distribution;
    Undersample or oversample the features;
end
procedure AUGMENT DATA;
Perturb and augment data;
  procedure CREATE ENSEMBLE;
For each set of augmented data, create classifier -
HistogramBoostedGradient;
Fit HistogramBoostedGradient estimator;
Choose features if the loss is less than iteration t-1;
```

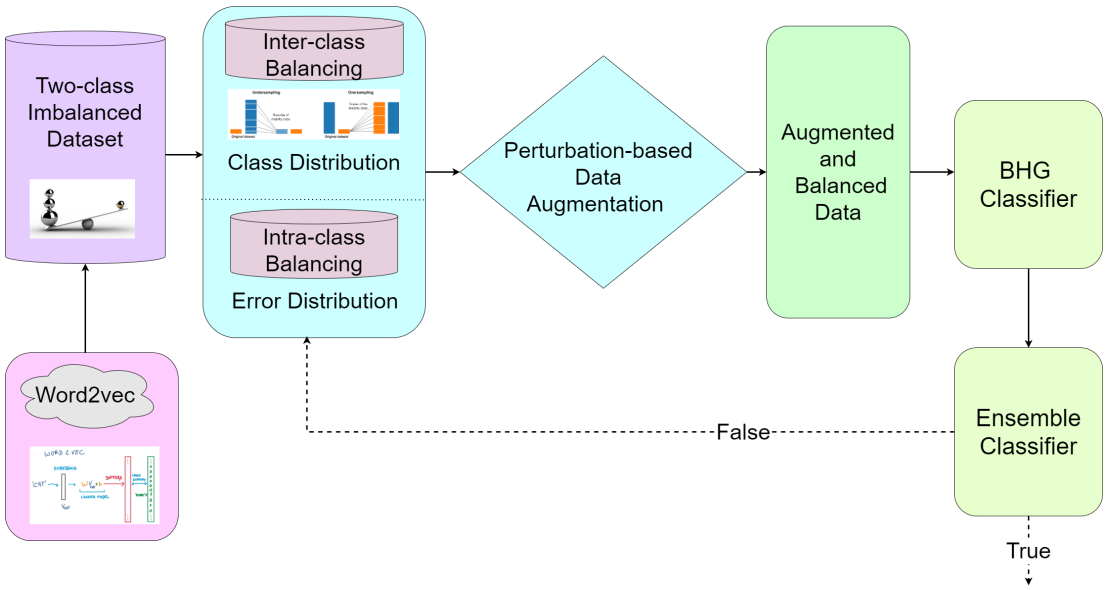

Fig. 1. Proposed approach for predatory chat detection

A. Balanced and Augmented Dataset

Given the dataset $\mathcal{D}$ with n classes and m features, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, each class can consist of k samples. When all the classes have equal number of samples, i.e., $k \approx k_{ave}$ one can effectively learn a classifier. However, when the number of samples k_i for a chosen class i is significantly lower than average number of samples from all other classes, the classifier is challenged with skewed data distribution. As it happens, the number of predatory samples is much lower than the number of non-predatory conversation with a sample distribution ratio $1.00 : 8.91$ for predatory to non-predatory samples in our case. Thus, irrespective of the sampling approaches to be used, the minor class will contribute to imbalance for learning a classifier. Thus, we first propose to create an augmented dataset $\mathcal{D}'$ for T number of iterations.

For each class C_i in the n classes, we employ two kind of sampling such that class with higher samples is under-sampled and the class with lower samples is over-sampled. In order to achieve this, we simply resort to progressively balanced hybrid sampling using the class distribution. The balanced classes for each iteration is then used to compute the error distribution for the true class distribution and inverse error distribution. Further, as the number of samples in one class can be much higher than the other class, for instance, in our case non-predatory conversations are much higher than the predatory samples, we augment the features in both classes such that the minor class is represented equally with a set of perturbation. The perturbation factor p therefore leads to new samples of the minor class which can be represented as $x' \rightarrow x + \alpha.x^p$ where α is a linear scaling factor. Thus, the new augmented samples lead to creation of $\mathcal{D}'$. For each of these samples obtained, we obtain new class distribution C' for a given iteration t in total number of iterations T. Using the newly augmented dataset $\mathcal{D}'$ with new class distribution C' with balanced, augmented and perturbed data, we learn a classifier Histogram Gradient Boosted Decision Trees as detailed in the next section. In every iteration t, the class distribution and inverse class distribution is used to balance the samples chosen to learn the classifier.

The Algorithm 1 represents the pseudocode of the approach.

B. Histogram Gradient Boosted Decision Trees

Given the augmented, balanced and perturbed dataset $\mathcal{D}'$ with n samples and m features, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$, the predictions from the boosted decision tree model, $\hat{y}_i$, is defined as a tree-based additive ensemble model, $\phi(x_i)$, comprising of K additive functions, f_k, defined as:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \;\; f_k \in \mathcal{F}$$

where $\mathcal{F} = \{f(x) = w_{q(x)}\}$ is a collection of Classification and Regression Trees, such that $q(x)$ maps each input feature x to one of T leaves in the tree by a weight vector, $w \in \mathbb{R}^T$.

Given the function defined above, the Gradient Boosted algorithm minimizes the following regularized objective function:

$$\tilde{\mathcal{L}} = \sum_{i} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k)$$

where $l(y_i, \hat{y}_i)$ is the loss function of the ith sample between the prediction $\hat{y}_i$ and the target value y_i, and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ is the regularization component to penalize k^{th} tree in growing additional leaves by λ - a regularization parameter and a weight vector w. We approximate the loss function using a second-order Taylor expansion [17], and we omit the details for the brevity of the paper considering the page limit.

C. Ensemble Construction

Based on the augmented features selected in each iteration, a classifier is chosen if the loss $l(y_i, \hat{y}_i)$ is the loss function of the ith sample between the prediction $\hat{y}_i$ and the target value y_i, and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ is the regularization component to penalize k^{th} tree in growing additional leaves by λ - a regularization parameter and a weight vector w.

TABLE I
PERFORMANCE OF VARIOUS APPROACHES AGAINST PROPOSED APPROACH. THE BLOCKS IN GRAY COLOR INDICATE THE APPROACHES THAT HANDLE DATA IMBALANCE AND CAN BE DIRECTLY COMPARED TO OUR PROPOSED APPROACH.

Ref	**Accuracy**	F_1	F_β
Bogdanova et al. [8]	0.97	-	-
Villatoro et al. [9]	0.92	0.87	0.93
Borj & Bours [18]	0.98	0.86	-
Fauzi & Bours [19]	0.95	0.90	0.93
Bours & Kulsrud [20]	-	0.94	0.97
Borj et al. [5]	0.99	0.96	-
Ebrahimi et al. [21]	-	0.80	-
Ebrahimi et al. [11]	0.99	0.77	-
Imbalance based approaches			
Cardei et al. [12]	-	-	0.95
Zuo et al. [15]	0.76	-	-
Proposed Model	**0.99**	**0.99**	**0.94**

V. EXPERIMENTAL RESULTS

For detection of predatory conversation, all the messages of a single conversation were merged into a single text block. Then, we extracted the Word2Vec feature vector for each of the merged texts. The main focus of this analysis is to handle the imbalanced nature of the dataset applying the proposed method. Thus, we select two state-of-the-art approaches which are close to our work to provide a comparison. Specifically, we compare our results against Cardei et al. [12] and Zuo et al. [15] who propose strategies to handle the imbalance in the predatory data. Further, we also compare our results against other state-of-the-art approaches to give a broader comparison.

Predatory detection techniques have been evaluated using different metrics such as accuracy, precision, recall, and F_1-score. Further, to avoid many false-positive detection F_β is also recommended as another primary metric for analyzing the performance [3] with $\beta = 0.5$. Table I demonstrates the obtained results and compares them with the baseline of various works. The proposed approach obtains a gain of 3% over the best benchmark, while it gains more than 23% more accuracy compared to the earlier approach [15] in a similar category of using balancing strategies.

Further, we also analyze the effect of perturbation factor in augmenting the dataset, and the obtained accuracy is presented in Figure 2. As noted from Figure 2, the performance changes slightly when the perturbation factor is increased to more than 20%. Despite the slight drop in performance, one can note the superiority of the proposed approach as compared to the accuracy reported in Table I. Thus, we deduce that the perturbation factor should not be more than 50% to obtain a reliable classification accuracy.

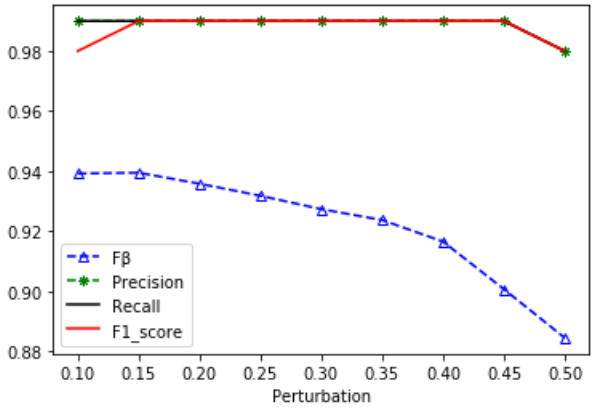

Fig. 2. Performance variation to perturbation factor in data.

VI. CONCLUSION

Predatory conversation detection based on text messages is a crucial problem to avoid exploiting under-aged or minors for sexual favors. Owing to the limited real datasets available, current works employ a standard dataset with less than 10% predatory data leading to a heavy imbalance in the dataset resulting in a classifier that may be sub-optimal. This work has proposed a new approach for handling the imbalanced nature of predatory data by hybrid sampling and class re-distribution to obtain an augmented dataset. Further, to improve the diversity of classifiers and features in the ensembles, this work also proposes to perturb the data along with augmentation in an iterative manner. With the set of experiments on the state-of-the-art dataset, we demonstrate that the proposed approach obtains an improvement over the best state-of-the-art approach by 3% and results in a F_1-score of 0.99 and a F_β of 0.94 Unlike this work, in future works, we also intend to explore different feature extraction approaches to validate the scalability of the proposed approach for predatory detection. Further, this work can also be extended by generating the predatory data through advanced approaches, including Generative Adversarial Networks.

REFERENCES

[1] H. Bentley, O. O'Hagan, A. Raff, and I. Bhatti, "How safe are our children," *The most comprehensive overview of child protection in the UK*, 2016.

[2] S. Craven, S. Brown, and E. Gilchrist, "Sexual grooming of children: Review of literature and theoretical considerations," *Journal of sexual aggression*, vol. 12, no. 3, pp. 287–299, 2006.

[3] G. Inches and F. Crestani, "Overview of the international sexual predator identification competition at pan-2012." in *CLEF (Online working notes/labs/workshop)*, vol. 30, 2012.

[4] M. Latapy, C. Magnien, and R. Fournier, "Quantifying paedophile queries in a large p2p system," in *2011 Proceedings IEEE INFOCOM*. IEEE, 2011, pp. 401–405.

[5] P. R. Borj, K. Raja, and P. Bours, "On preprocessing the data for improving sexual predator detection: Anonymous for review," in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*. IEEE, 2020, pp. 1–6.

[6] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17–26, 2017.
[7] V. Egan, J. Hoskinson, and D. Shewan, "Perverted justice: A content analysis of the language used by offenders detected attempting to solicit children for sex," *Antisocial behavior: Causes, correlations and treatments*, vol. 20, no. 3, p. 273297, 2011.
[8] D. Bogdanova, P. Rosso, and T. Solorio, "On the impact of sentiment and emotion based features in detecting online sexual predators," in *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, 2012, pp. 110–118.
[9] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y Gómez, and L. V. Pineda, "A two-step approach for effective detection of misbehaving users in chats." in *CLEF (Online Working Notes/Labs/Workshop)*, vol. 1178, 2012.
[10] D. Bogdanova, P. Rosso, and T. Solorio, "Exploring high-level features for detecting cyberpedophilia," *Computer speech & language*, vol. 28, no. 1, pp. 108–120, 2014.
[11] M. Ebrahimi, C. Y. Suen, O. Ormandjieva, and A. Krzyzak, "Recognizing predatory chat documents using semi-supervised anomaly detection," *Electronic Imaging*, vol. 2016, no. 17, pp. 1–9, 2016.
[12] C. Cardei and T. Rebedea, "Detecting sexual predators in chats using behavioral features and imbalanced learning." *Nat. Lang. Eng.*, vol. 23, no. 4, pp. 589–616, 2017.
[13] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2008.
[14] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 1–6, 2004.
[15] Z. Zuo, J. Li, B. Wei, L. Yang, F. Chao, and N. Naik, "Adaptive activation function generation for artificial neural networks through fuzzy inference with application in grooming text categorisation," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2019, pp. 1–6.
[16] X. Rong, "word2vec parameter learning explained," *arXiv preprint arXiv:1411.2738*, 2014.
[17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[18] P. R. Borj and P. Bours, "Predatory conversation detection," in *2019 International Conference on Cyber Security for Emerging Technologies (CSET)*. IEEE, 2019, pp. 1–6.
[19] M. A. Fauzi and P. Bours, "Ensemble method for sexual predators identification in online chats," in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2020, pp. 1–6.
[20] P. Bours and H. Kulsrud, "Detection of cyber grooming in online conversation," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2019, pp. 1–6.
[21] M. Ebrahimi, C. Y. Suen, and O. Ormandjieva, "Detecting predatory conversations in social media by deep convolutional neural networks," *Digital Investigation*, vol. 18, pp. 33–49, 2016.