DETECTING CYBER GROOMING USING TEXT MINING, SUPPORT VECTOR MACHINE, NAÏVE BAYES AND RANDOM FOREST.

Bello Bilkisu Mohammed [1], Hashim Ibrahim Bisallah[2], Israel Musa[3] and Israel Okorie[4].

1. *Department of computer science, Kaduna Polytechnic, Kaduna state, Nigeria.*
candiebell2000@yahoo.com, bilkisbello09@gmail.com
2. *Kampala International University, Kampala, Uganda*
hashim.bisallah@kiu.ac.ug
3. *Institut Superieur de Genie-civil et de Gestion, Abomey-Calavi, Benin Republic.*
israelmusa9@gmail.com
4. *Department of computer, University of Abuja, Nigeria.*
okorieisrael15@gmail.com

ABSTRACT

The problem of online grooming has emerged as a notable apprehension in present-day society due to the increased use of the internet. This poses a threat to children, as they can be targeted by predators. In order to address this problem, we conducted a research study that utilized text analysis techniques to identify predatory messages. The goal of this research was to protect children from potential harm caused by paedophiles. We aimed to identify specific features and words that are indicative of predatory behaviour, which would enable us to accurately detect such messages in online conversations. By doing so, we aimed to enhance internet security for young children and eliminate grooming incidents. To classify adults who pretend to be children, we focused on identifying crucial features, with foreign words being particularly important. The dataset used in our research was collected from PAN, a well-known source for such data. In order to develop our model, we employed three different algorithms: Naïve Bayes, Random Forest, and Support Vector Machine. Through our findings, we were able to demonstrate that while it is challenging to distinguish between genuine children and adults posing as children within chat logs, we achieved an accuracy of 76.80% in identifying fake children using our best-performing model, SVM. This report discusses the accuracy of the methods we proposed and highlights the essential features that contributed to their success. The primary focus of our study was on detecting grooming conversations, but future research could involve identifying adults who pretend to be children or creating fake profiles. It is important to continue exploring and developing methods to protect children from online grooming and ensure their safety in the digital age.

Keyword: *Online grooming, Text analysis, Predatory messages, Child safety, Detection algorithms*

1. INTRODUCTION

The problem of Internet addiction is a growing concern in society [1]. By 2022, over 60% of the global population had Internet access, with even higher rates in developed countries like the United States and England, where nearly 90% of the population was connected [2] [3] . Many Internet users, especially those under 25, use various devices for activities like work, gaming, and socializing [4] [5].

However, frequent Internet usage among young individuals poses risks in today's society [6]. One concerning threat is the presence of online predators, who are harder to catch compared to real-life predators due to their hidden identities [7]. Shockingly, around one in seven children online still receives sexual solicitations [8]. Online predators often use grooming, a tactic where they build a relationship with a child to establish a sexual connection [9]. These predators gather information about their victims on social networking sites and exploit chat rooms designed for children [10] [11]. Young people, being vulnerable, may not fully grasp the risks of interacting with strangers or sharing personal information [10].

The Internet's accessibility and anonymity contribute to the exploitation and sexual abuse of children [12]. Online individuals can easily conceal their identity and provide false information, making it difficult for parents and law enforcement to identify and apprehend those with malicious intentions [13] [9]. Moreover, children may not report grooming incidents due to feelings of guilt, shame, or confusion about the abusive nature of the relationship [14]. Therefore, it is crucial to address the negative impacts of juvenile abuse in online spaces and ensure a safe environment for young individuals on social networks to promote public safety [14].

In order to address this form of societal wrongdoing, it is essential to analyze substantial amounts of unidentified chat records. Text analysis studies have shown the ability to estimate someone's age, gender, location, and level of education based on their online writings [15] [16]. Analyzing online content provides insights into individuals' behavior and personal information [17]. Such profiling techniques have applications in various fields, including forensics, plagiarism detection, and intelligence gathering [18]. Automated Online Predator Identification (OPI), alternatively referred to as Sexual Predator Identification (SPI) or Sexual Predator Detection (SPD, is an active method aimed at minimizing the negative consequences of such offenses. Although OPI encompasses both textual and visual information, focusing on textual data proves to be more practical for automation purposes. Therefore, this paper primarily focuses on the analysis of textual data.

1.1 Literature

This section delves into different research studies and methodologies related to the detection of online predators. The primary objective of these studies is to create automated techniques capable of distinguishing between predators and victims during online conversations. A comprehensive overview of each study's key points and findings will be provided.

[19] research contributes valuable insights into the automated detection of online predators and the importance of safeguarding vulnerable individuals in online environments. The study relies on data from the Perverted Justice website to distinguish between sexual predators and potential victims. Pendars employs Support Vector Machines (SVM) and distance-weighted k-Nearest Neighbors (k-NN) classifiers with n-grams as features.

In their comprehensive study, [20] integrate principles from communication and computer science disciplines to create tools for safeguarding children from cyber predators. They introduce ChatCoder, a software program designed for content analysis of online conversations. ChatCoder demonstrates a moderate to high level of accuracy in identifying predators and distinguishing between predatory and non-predatory interactions. The research highlights the effectiveness of their multidisciplinary approach and the potential of ChatCoder as a proactive safeguarding mechanism against online child exploitation.

[21] Suggests employing automated text analysis, particularly the Linguistic Inquiry and Word Counting tool, to detect grooming stages in online interactions. While LIWC proves effective in revealing linguistic patterns, its inability to recognize misspelled words and internet language presents a limitation. Despite this, the study provides partial support for Wollis's hypothesis, highlighting the potential of automated text analysis in detecting grooming stages. The research paves the way for further advancements in leveraging technological tools to combat online predatory behavior and enhance the safety of online spaces.

[22] conduct content and data analysis to explore repetitive patterns indicating relationship with a minor. They utilize the NVivo software analysis tool to identify eight recurring themes associated with grooming behavior. The findings suggest that Offenders exhibit daring behavior, taking risks, and organize in-person gatherings with minimal vigilance.

[23] develop psycho-linguistic profiles to gain insights and identify patterns, dividing the grooming process into distinct phases. Their primary focus lies in creating profiles of offenders instead of performance metrics, aiming to create a live tool for identifying online conversations that may involve pedophiles.

[24] utilize Support Vector Machines (SVMs) to identify the behavioral characteristics of an online predator through textual chats. After training their model on chat logs containing predatory and non-predatory content, they attain a mean accuracy of 76.23% by employing SVMs with trigrams.

[25] examine the global competition focused on identifying sexual predators. Which involves identifying predators and detecting grooming behavior in chat logs. Participants utilize lexical and behavioral features, along with various classifiers like SVMs, neural networks, decision trees, and Naïve Bayes. The study highlights the importance of pre-filtering and explores different approaches for predator identification.

[26] devise a method to detect adults posing as minors. They classify authors as adults or children and subsequently determine if a child is genuine or an impostor. Although their results show promising performance, the authors express concerns about potential misclassifications involving law enforcement officers.

[27] conducts a comparative analysis of different text classification techniques and introduces a CNN-based method to recognize online predators. The research demonstrates that CNNs surpass pre-trained word vectors and SVMs in accurately identifying predators. Employing one-hot vectors and a single convolution layer produces superior outcomes when compared to alternative approaches, achieving an F-score of 0.8087.

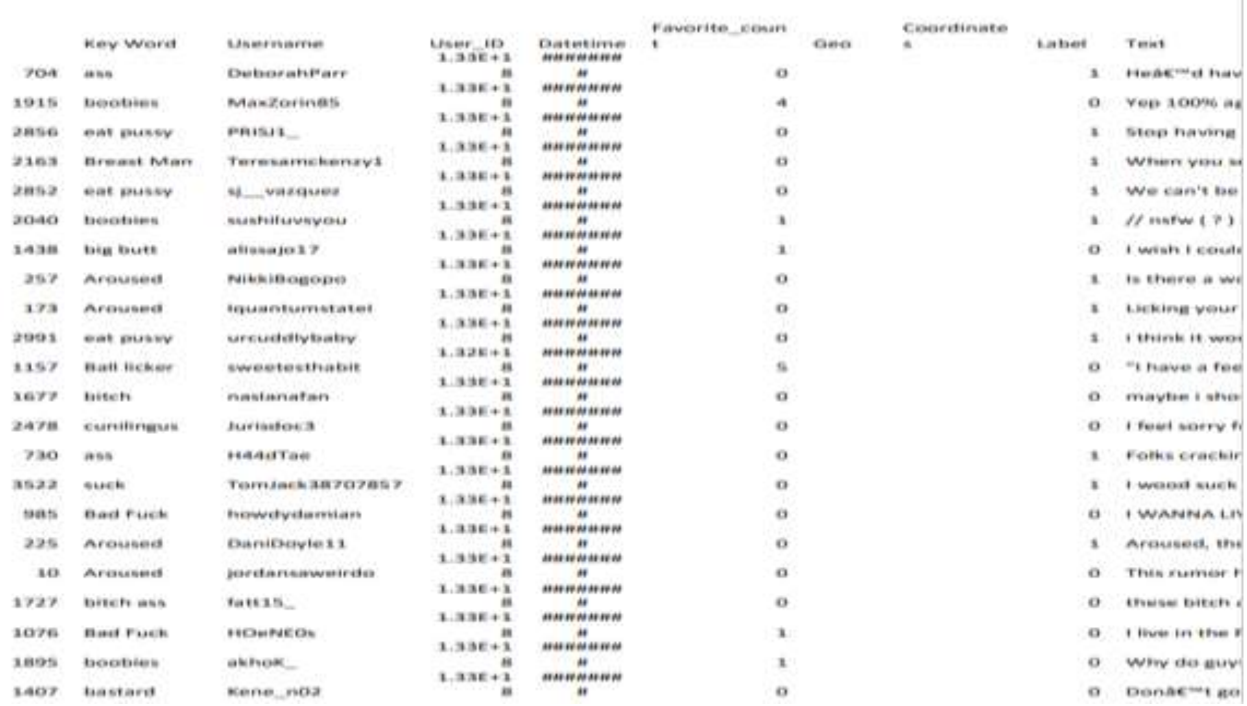
[28] tackle the societal problem of cyber grooming by examining current solutions that rely on lexical features and the theory of luring communication. They emphasize the prevalence of supervised learning approaches and suggest exploring semi-supervised and reinforcement learning methods.

These studies collectively demonstrate ongoing efforts to develop automated methods for identifying online predators, utilizing diverse approaches such as text categorization, content analysis, profiling, and machine learning algorithms. The results indicate that distinguishing between predators and victims with reasonable accuracy is attainable, but challenges persist, including robust data acquisition, pre-filtering strategies, and addressing the limitations of existing text analysis

2. MATERIALS AND METHODOLOGIES

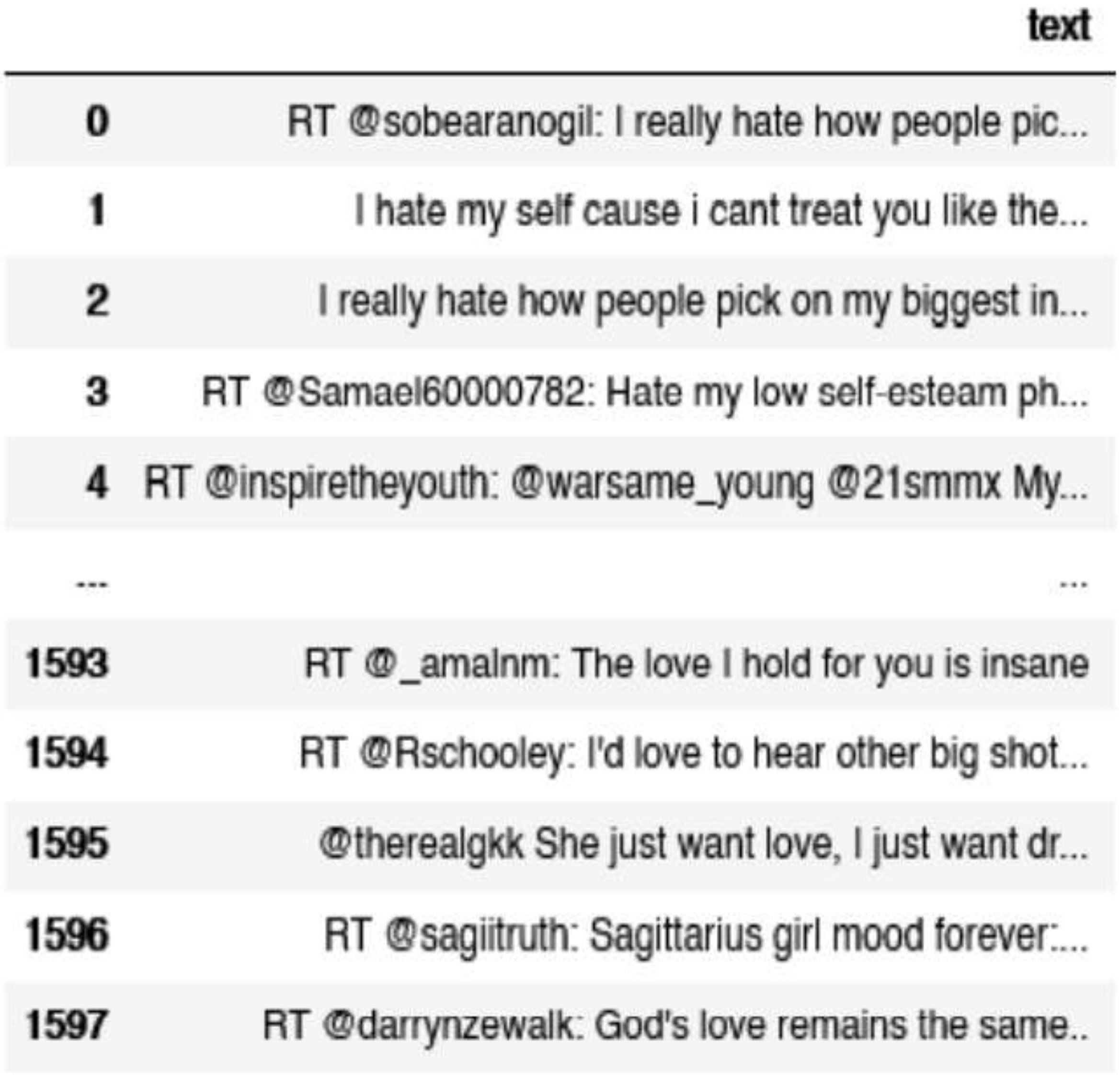
2.1 Data

The dataset utilized in this research was obtained from PAN, an organization that promotes research in digital text forensics through organized shared task evaluations. The dataset comprises two columns, namely "text" and "labels." The "labels" column classifies the data into two categories: predator and non-predator. The "text" column contains the actual text associated with the sentiment. The metadata associated with the dataset allows us to distinguish between texts written by adults and children. Figure 1 illustrates the raw dataset after data collection, while figures 2 present tabulated samples of the data after undergoing cleaning and preprocessing steps. Notably, the dataset contains entire conversations, making the classification of the Pan dataset more challenging.



	Key Word	Username	User_ID	Datetime	Favorite_count	Geo	Coordinates	Label	Text
704	ass	DeborahParr	1.33E+18	########	0			1	Heâ€™d hav
1915	boobies	MaxZorin85	1.33E+18	########	4			0	Yep 100% ag
2856	eat pussy	PRISJ1_	1.33E+18	########	0			1	Stop having
2163	Breast Man	Teresamckenzy1	1.33E+18	########	0			1	When you s
2852	eat pussy	sj__vazquez	1.33E+18	########	0			1	We can't be
2040	boobies	sushiluvsyou	1.33E+18	########	1			1	// nsfw (?)
1438	big butt	alissajo17	1.33E+18	########	1			0	I wish I coul
257	Aroused	NikkiBogopo	1.33E+18	########	0			1	Is there a w
173	Aroused	Iquantumstatel	1.33E+18	########	0			1	Licking your
2991	eat pussy	urcuddlybaby	1.33E+18	########	0			1	I think it wo
1157	Ball licker	sweetesthabit	1.32E+18	########	5			0	"I have a fee
1677	bitch	naslanafan	1.33E+18	########	0			0	maybe i sho
2478	cunilingus	Jurisdoc3	1.33E+18	########	0			0	I feel sorry f
730	ass	H44dTae	1.33E+18	########	0			1	Folks crackin
3522	suck	TomJack38707857	1.33E+18	########	0			1	I wood suck
985	Bad Fuck	howdydamian	1.33E+18	########	0			0	I WANNA LIV
225	Aroused	DaniDoyle11	1.33E+18	########	0			1	Aroused, the
10	Aroused	jordansaweirdo	1.33E+18	########	0			0	This rumor h
1727	bitch ass	fatt15_	1.33E+18	########	0			0	these bitch
1076	Bad Fuck	HOeNEOs	1.33E+18	########	1			0	I live in the
1895	boobies	akhoK_	1.33E+18	########	1			0	Why do guy
1407	bastard	Kene_n02	1.33E+18	########	0			0	Donâ€™t go

Fig -1: Dataset highlighting keywords



	text
0	RT @sobearanogil: I really hate how people pic...
1	I hate my self cause i cant treat you like the...
2	I really hate how people pick on my biggest in...
3	RT @Samael60000782: Hate my low self-esteam ph...
4	RT @inspiretheyouth: @warsame_young @21smmx My...
...	...
1593	RT @_amalnm: The love I hold for you is insane
1594	RT @Rschooley: I'd love to hear other big shot...
1595	@therealgkk She just want love, I just want dr...
1596	RT @sagiitruth: Sagittarius girl mood forever:...
1597	RT @darrynzewalk: God's love remains the same..

Fig -2: Preprocessed data

2.2 Data Preprocessing

The first step in data cleaning was the removal of unwanted characters from the text, including HTML/XML tags, punctuation marks, non-alphabetic characters, extra whitespace, and language corpus-specific characters. Python's "re" method was employed for this purpose. Subsequently, the entire sentences were split into tokens using the "word_tokenize" method from the NLTK toolkit in sklearn. Additionally, stopwords were removed from the text using the NLTK corpus of stopwords. Finally, to facilitate the training of the model, the text was vectorized using the TfidfVectorizer from sklearn's feature_extraction method.
For the model training process, sklearn's "train_test_split" method was used to split the data into training and test sets. The training set consisted of 70% of the entire dataset, while the remaining 30% was used for testing the model.

2.3 Analytical Techniques

Machine learning tools were employed to analyze the collected data, and interpretations were drawn from the obtained results, as shown in figure 3.

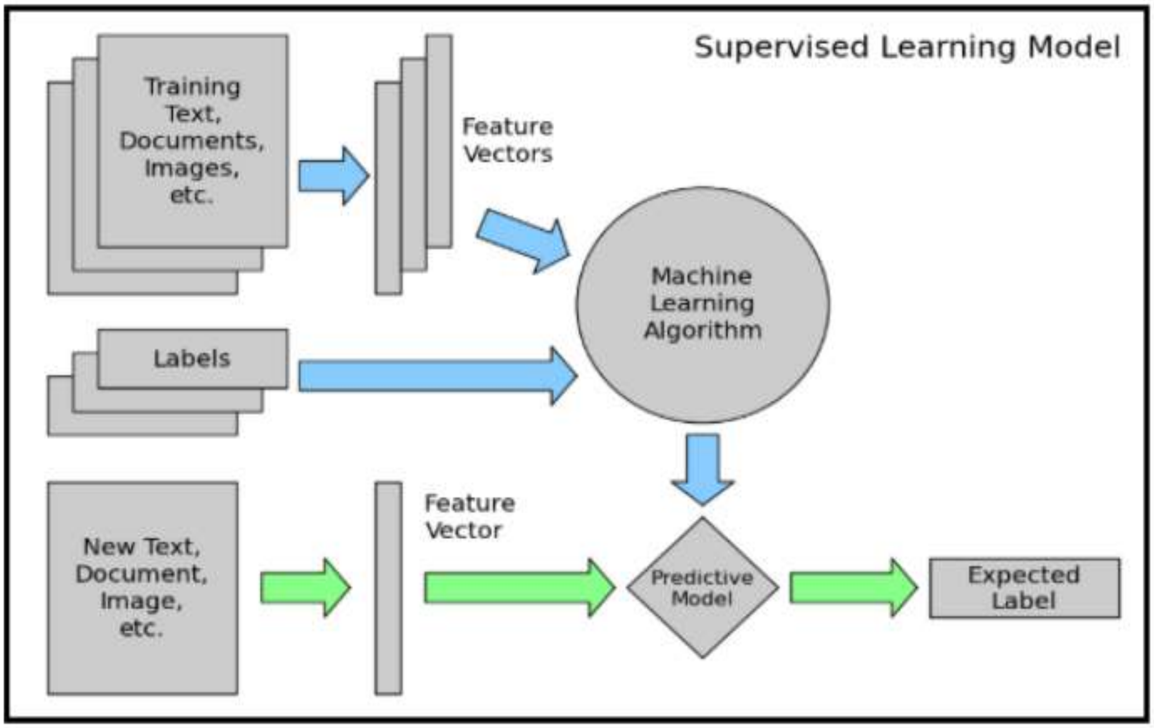

Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Fig -3: Machine Learning Approach for Sexual Predator Detection

2.3.1 Naïve Bayes

In predicting cyber grooming, Naïve Bayes involves calculating the posterior probability of a class given a text document using Bayes' theorem. The equation is as follows, where C represents the class label (e.g., sexually predating or non-sexually predating) and X is the input text document:

$P(C|x_1, x_2, ..., x_n) = (P(C) * P(x_1|C) * P(x_2|C) * ... * P(x_n|C)) / P(x_1, x_2, ..., x_n)$

2.3.2 Random Forest

Random Forest is applied by transforming text data into numerical features using techniques like TF-IDF. The process involves creating multiple decision trees and aggregating their predictions for the final output.

2.3.3 Support Vector Machine (SVM)

SVM aims to find a hyperplane in a high-dimensional feature space that maximally separates different classes of data for binary classification problems. The decision boundary is defined by the equation w^T * x + b = 0, where w is a vector perpendicular to the hyperplane, x is the feature vector of a data point, and b is the bias term.

2.4 Evaluation Metrics

Various evaluation metrics were employed to assess the performance of the models:

2.4.1 Accuracy

Accuracy measures the proportion of correct predictions out of all predictions made. It is calculated as:

Accuracy = (Number of correct predictions) / (Total number of predictions)

2.4.2 Recall

Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify positive cases. The formula is:

Recall = (True Positives) / (True Positives + False Negatives)

2.4.3 Precision

Precision quantifies the accuracy of positive predictions made by the model. It is computed as:

Precision = (True Positives) / (True Positives + False Positives)

2.4.4 F1 Score

The F1 score is the harmonic mean of precision and recall and is calculated as 2 * (Precision * Recall) / (Precision + Recall).

3. RESULT

We opted to utilize precision, recall, and the F-measure as assessment criteria instead of relying solely on accuracy, the result of the models is presented at Table 1. This choice was influenced by the imbalanced nature of the data. Specifically, we selected precision due to its sensitivity to data distribution, in contrast to accuracy. Conversely, recall remains uninfluenced by data distribution. To strike equilibrium and consider the varying importance assigned to precision and recall, we employed the F-measure, which combines both measures. This decision is well-justified within the context of the particular problem we are addressing. For the SVM model, it attained a precision of 76.67%, recall of 76.53%, F1 score of 76.60, and accuracy of 76.84%. The RF model showcased a precision of 71.48%, recall of 71.65%, F1 score of 71.48, and accuracy of 76.84%. Regarding the NB model, its precision was gauged at 63.51%, accompanied by a recall of 61.37%, F1 score of 60.68, and accuracy of 63.11%.

Table -1: Comparative Model Performance

S/N	Classifier	Precision	Recall	F1	Accuracy
I	Support vector machine	76.67%	76.53%	76.61%	76.84%
III	Random Forest	71.48%	71.65%	71.48%	71.57%
III	Naïve Bayes	63.51%	61.37%	60.68%	63.11%

3.1 Discussion

To grasp the significance of our findings, we have juxtaposed our results against relevant studies in the existing literature. Our Support Vector Machine (SVM) model achieved an accuracy of 76.70%, surpassing the accuracy of **[21]**'s SVM model at 20%, as well as **[24]**'s model with an accuracy of 76.23%. It's noteworthy that **[24]** utilized a limited dataset for their model. Nevertheless, our model's performance was marginally below that of **[25]**, whose model achieved an accuracy of 80%. It's plausible that the **[25]** model contained noise due to the absence of pre-filtered text for the classification process. As illustrated in Table 1, all performance metrics for our SVM model exceed 70%. This signifies that our SVM model is capable of predicting cyber grooming with an accuracy surpassing 76%, and it maintains precision of over 70% in positive instance prediction (sexual predation). The recall value from our model implies its capacity to classify potential cases of sexual grooming. Among the metrics, recall is of particular significance since our aim is to identify all occurrences of sexual predation.

The RF model exhibited slightly lower performance than the SVM model, but it can be deemed satisfactory when compared to prior literature utilizing Random Forest for sexual grooming prediction. Conversely, our Naïve Bayes model displayed the least effectiveness among the three models, relative to the others. It only achieved an overall accuracy of 63.11%. Furthermore, its accuracy in predicting instances of sexual predation was limited, with only 63.50% precision and a recall rate of 61.37% for the complete dataset.

However, our model showcased commendable performance in forecasting instances of sexual predation based on chat logs. It's vital to acknowledge that the dataset used for training the model was modest in size, thus warranting a more comprehensive scrutiny and additional examination of the results.

4. CONCLUSIONS

The study investigates the issue of grooming, which involves individuals preparing children for exploitation. The research strategy involves distinguishing between adults and children by analyzing their writing styles. The goal is to verify if a person claiming to be a child is indeed genuine. This differentiation is accurate when analyzing formal language such as in book reviews, but it becomes difficult when dealing with text from blogs and chat conversations. Nonetheless, the research generally succeeds in telling apart real children from adults pretending to be children.

The research focuses on identifying conversations that involve grooming behavior. To achieve this, a two-step process is used: first, individual messages are classified, and then entire conversations are categorized based on the classifications of the messages. Traditional machine learning models are utilized for both of these steps. These models are trained on messages with known labels and conversations that have been pre-labeled. This approach improves the ability to identify grooming conversations compared to previous methods, specifically by enhancing the recall rate.

Future studies should gather a diverse range of datasets for training purposes. It's important to include data from different forms of communication and to explore challenges related to identification beyond just chat conversations. This broader approach would likely contribute to the advancement of the field. Overall, utilizing psycho-linguistic profiles to detect groomers shows promise and could be further investigated in the future.

5. ACKNOWLEDGEMENT

The authors did not receive any financial support for this study, and no external entity or individual provided technical assistance during the course of this research.

6. REFERENCES

[1]. Mihajlov, M., & Vejmelka, L. (2017). Internet addiction: A review of the first twenty years. Psychiatria Danubina, 29(3), 260–272. https://doi.org/10.24869/psyd.2017.260\

[2]. ITU. (2023, January 31). Statistics. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

[3]. ITU. (2022, June 6). Global Connectivity Report 2022. https://www.itu.int/en/mediacentre/Pages/PR-2022-06-06-Global-potential-of-internet-remains-untapped.aspx

[4]. Petrosyan, A. (2023, January 27). U.S. internet usage penetration by age and device 2022. Statista. https://www.statista.com/statistics/1360723/us-internet-usage-penetration-by-age-group-and-device/

[5]. Ritchie, H., Mathieu, E., Roser, M., & Ortiz-Ospina, E. (2023, April 13). Internet. Our World in Data. https://ourworldindata.org/internet

[6]. Bryant, A. (2018). The Effect of Social Media on the Physical, Social Emotional, and Cognitive Development ofAdolescents.MerrimackCollege.https://scholarworks.merrimack.edu/cgi/viewcontent.cgi?article=1036&context=honors_capstones

[7]. Woodhams, J., Kloess, J. A., Jose, B., & Hamilton- Giachritsis, C. E. (2021, March 12). Characteristics and behaviors of anonymous users of dark web platforms suspected of child sexual offenses. Frontiers. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.623668/full

[8]. Adorjan, M., & Ricciardelli, R. (2018). Gender, sexting, and the teenaged years. In Cyber-Risk and Youth (pp. 111–131). Routledge. https://doi.org/10.4324/9781315158686-7

[9]. McManus, M. A., Almond, L., Cubbon, B., Boulton, L., & Mears, I. (2015). Exploring the online communicative themes of Child sex offenders. Journal of Investigative Psychology and Offender Profiling, 13(2), 166–179. https://doi.org/10.1002/jip.1450

[10]. Lim, Y. Y., Wahab, S., Kumar, J., Ibrahim, F., & Kamaluddin, M. R. (2021). Typologies and psychological profiles of Child sexual abusers: An extensive review. Children, 8(5), 333. https://doi.org/10.3390/children8050333

[11]. Mitchell, K. J., Finkelhor, D., Jones, L. M., & Wolak, J. (2010). Use of social networking sites in online sex crimes against minors: An examination of national incidence and means of utilization. Journal of Adolescent Health, 47(2), 183–190. https://doi.org/10.1016/j.jadohealth.2010.01.007

[12]. Briggs, P., Simon, W. T., & Simonsen, S. (2010). An exploratory study of internet-initiated sexual offenses and the chat room sex offender: Has the internet enabled a new typology of sex offender? Sexual Abuse, 23(1), 72–91. https://doi.org/10.1177/1079063210384275

[13]. Reaves, M. (2016). Pedophilia: Prevention or paternalism?. Voices in Bioethics. https://doi.org/10.7916/vib.v2i.5991

[14]. Chassiakos, Y., Radesky, J., Christakis, D., Moreno, M. A., Cross, C., Hill, D., Ameenuddin, N., Hutchinson, J., Levine, A., Boyd, R., Mendelson, R., & Swanson, W. S. (2016). Children and adolescents and Digital Media. Pediatrics, 138(5). https://doi.org/10.1542/peds.2016-2593

[15]. Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2017). Gender, genre, and writing style in formal written texts. Text, 27(4), 437-454.

[16]. Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.

[17]. McGuire, J. (2022). Politifact - why it's not "grooming": What research says about gender and sexuality in schools. @politifact. https://www.politifact.com/article/2022/may/11/why-its-not-grooming-what-research-says-about-gend/

[18]. Smith, J., & Johnson, A. (2018). Text analysis for demographic profiling. Journal of Computational Linguistics, 25(3), 123-145.

[19]. Pendar, N. (2017, September). Toward spotting the paedophiles telling victim from predator in text chats. In International Conference on Semantic Computing (ICSC 2007) (pp. 235-241). IEEE.

[20]. Edwards, Y., & Schapire, R. (2016). Experiments with a new boosting algorithm. In Proceedings of the thirteenth International Conference on Machine Learning (pp. 148–156).

[21]. Wollis, M. (2015). A Linguistic Analysis of Online Predator Grooming. Cornell University, Honors thesis, Jan 2015. https://core.ac.uk/download/pdf/4918216.pdf

[22]. Egan, V., et al. (2017). Perverted justice: A content analysis of the language used by offenders detected attempting to solicit children for sex. Antisocial Behaviour: Causes, Correlations and Treatments, 20(3), 273-297.

[23]. Guapta, A., et al. (2016). Characterizing paedophiles conversations on the internet using online grooming. arXiv preprint arXiv:1208.4324.

[24]. Pandey, S. J., et al. (2017, May). Detecting predatory behavior from online textual chats. In International Conference on Multimedia Communications, Services and Security (pp. 270-281). Springer, Berlin, Heidelberg.

[25]. Inches, G., & Crestani, F. (2018). Overview of the International Sexual Predator Identification Competition at PAN-2012.

[26]. Ashcroft, M., et al. (2015). A Step towards Detecting Online Grooming - Identifying Adults Pretending to be Children. In 2015 European Intelligence and Security Informatics Conference (pp. 98-104). IEEE.

[27]. Meyer, M. (2019). Machine learning to detect online grooming. Uppsala University, Master thesis, July 2019. http://uu.diva-portal.org/smash/get/diva2:846981/FULLTEXT01.pdf

[28]. Mabuza, C. (2018). Conversation Level Constraints on Paedophile Detection in Chartrooms. In CLEF 2012 Conference and Labs of the Evaluation Forum, 01, S. 1–13.