IDENTIFYING ONLINE SEXUAL GROOMING CONTENT IN SOCIAL MEDIA USING NEURAL NETWORKS CLASSIFICATION TECHNIQUE

Luqman Hakim Mohd Nasir, Zurina Saaya[1] and Mohd Rizuan Baharon[2]

[1,2]Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

Corresponding Author's Email: [1]zurina@utem.edu.my

Article History: Received 8 September 2022; Revised 11 October 2022; Accepted 29 November 2022

ABSTRACT: Children nowadays are in danger more than ever. With the advent of technology, these kids are often the target of online sexual grooming activity. This threat is prevalent on social media, where innocent children are exposed to online predators. This paper uses a classification technique to identify online sexual grooming content in the form of text available on social media sites. This is because users are unaware of the contents lurking within any social media posts and comments, so a tool is needed to detect said grooming content to give awareness and censor it. The dataset used in the experiment is collected from different social media sites such as YouTube, Instagram, and Twitter. The inter-rater agreement is used to validate the process of manual data annotation. For the classification technique it is develop using neural networks algorithm based on n-gram and sequence features. In the end, the classification algorithm will be evaluated by comparing the accuracy between the algorithm and human perception. For the future works the developed detection model can be deployed for in a web-based system such as a browser extension. This can be as an approach for implementing awareness intervention sexual grooming conversation.

KEYWORDS: *online grooming, text classification, text mining, social media*

1.0 INTRODUCTION

Online grooming is depicted as a process in which an adult builds relationship with children through the internet [1]. This malicious intent towards children is often the cause of adults who have sexual desires towards them and will use any method they can to engage with them. With the advent of mobile phones, laptops and computers, online predators can interact with vulnerable children through social media sites such as Twitter, Instagram, TikTok and many more. Acknowledging that children use these social media sites and share their personal information without knowing the consequences is pretty concerning. Therefore, making them easy targets for these online predators as their next victims.

Also, they do not seem to comprehend that the content shared on social media contains grooming content considering their age. These problem statements are used to embark on the objectives of this research which are to develop a classification algorithm that can detect sexual grooming content on social media and evaluate the developed algorithm's accuracy.

This project will tackle this topic and create an approach that can counter online grooming by developing a simple algorithm that will be able to detect any potential grooming activities. With the assistance of such an algorithm, it is possible to produce an indication whenever potential grooming occurs and will further warn people of it. Of course, detecting any potential grooming can be challenging as the internet contains large amounts of information shared by people worldwide every day. So, this project will utilize a text mining approach to collect vast information and differentiate it between relevant and non-relevant information to detect potential grooming content [2]

One of the text mining methods is text classification and for this project, we will focus on this method as it sets predefined categories for open-ended text. It can be used to organize and categorize any kind of text. In this case, it will significantly assist in determining which content in social media is considered online grooming. The main benefit of this method is the capability to analyze millions of unstructured data, such as comments and posts on social media, within just a few minutes. Text classification can be performed in two ways, either manual or automatic. For manual, involves humans that interpret the content of the text and categorizes it accordingly to their perception. On the other hand, automatic text classification has multiple types of systems to be performed on, such as rule-based systems, machine learning-based systems, and hybrid systems.

2.0 RESEARCH BACKGROUND

2.1 Online Grooming

Internet solicitation, also known as grooming of children, has been described as a process by which an individual prepares a child and their environment for sexual abuse to occur, including gaining access to the child, creating compliance and trust, and ensuring secrecy to avoid detection. Individuals attempted to sexually exploit children by seducing their targets using attention, affection, kindness and gifts. Intrinsic characteristics of a genre define the discourse's function and purpose. It allows us to consider how the message functions in the sender and receiver discourse. In this topic alone, there are five intrinsic characteristics to consider, assessment, enticements, cyber exploitation, control and self-preservation.

2.2 Text Mining

Text mining, also known as text analytics, is an artificial intelligence (AI) technique that uses natural language processing (NLP) to convert the unstructured text in documents and databases into normalized and structured data to be analyzed. Text mining is a technique for identifying patterns and correlations in vast amounts of text. The application of text mining has been used in other research areas such as e-commerce, manufacturing, healthcare and more. Text mining has multiple techniques, including information extraction, information retrieval, natural language processing (NLP), categorization, clustering and summarization.

2.3 Text Classification

Several issues must be addressed when attempting to identify the type of content, for example, which features of the content should be used to determine its category or type. The task for content identification could be formulated as a machine learning problem, specifically as a text-based classification task. Classification is the task of mapping items to a predefined set of labels. Classification algorithms are a branch of machine learning. They are typically categorized as supervised learning algorithms, in which a model is learned from a set of correctly classified or labeled examples (training set). Once the model is learned, it can be applied to unlabeled items (test set). Classification has been used in many areas, such as medical applications [3], [4] and financial applications [5]. Classification algorithms such as Naive Bayes [6], Decision Trees [7], Support Vector Machines [8] and Neural Networks [9] are commonplace solutions for document or text classification.

2.4 Proposed Solution

The proposed solution for this project is to develop an algorithm that would be able to detect online sexual grooming content on social media sites. Text classification using neural network algorithms is chosen because of its usefulness in extracting relevant and valuable information based on text-based content on social media sites[10]. Any form of written sources on social media, such as comments and posts, could be focused on instead of images and videos, as text-based content are easier to capture and analyze.

3.0 METHODOLOGY

The project methodology is vital to ensure the research is carried out according to the right and smooth procedure so the project can be completed within the given time frame. Four processes are involved in the methodology: Data Collection, Pattern Identification, Algorithm Development and Algorithm Evaluation. Figure 1 shows all the processes involved in detecting online grooming content on social media.

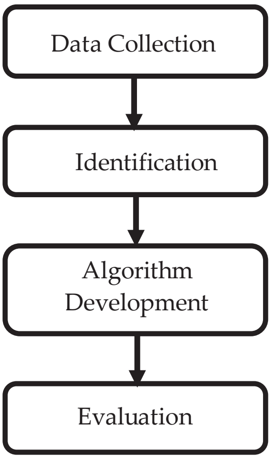


Figure 1: Project Methodology

Based on Figure 3, the first stage that is carried out for this project is the data collection of online grooming content. This stage is crucial for the rest of the stages because real data is needed so that proper research on this topic and algorithm development can be done.

The next stage is content annotation. During this stage, two human annotators have been employed to manually classify 1260 posts extracted from social media platforms, according to whether the text contained grooming message or not. This stage holds the key to the next stage, which is algorithm development.

After that, stages for classification algorithm development and evaluation will take place. For algorithm development, it will be done based on a set of training data that will be used as the basis for model creation. This stage is then followed by algorithm evaluation, which is a process of evaluating the effectiveness of the algorithm that has been developed by comparing the detection rate from the algorithm and test data.

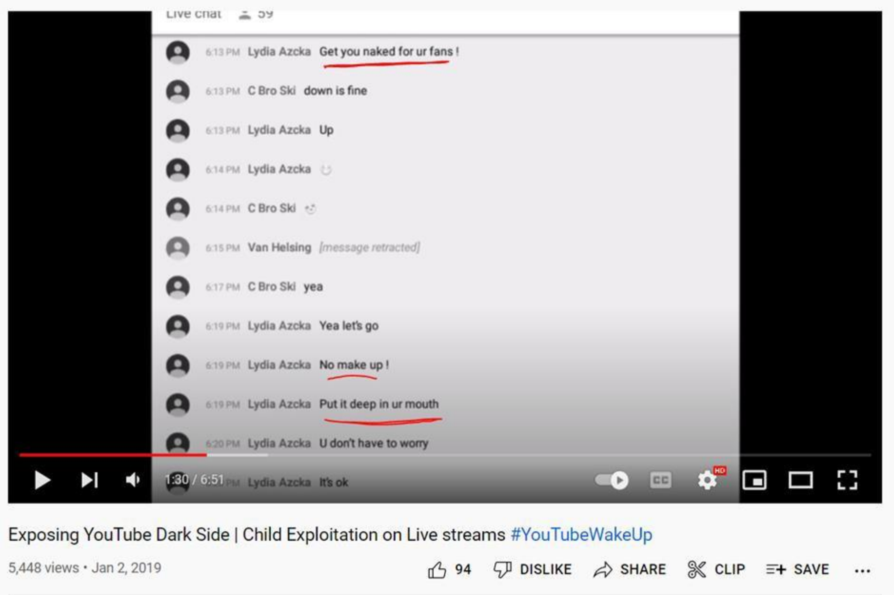
4.0 DATASETS

The data set for this research is collected from different social media sites such as YouTube, Instagram, and Twitter. Table 1 shows the sample of data.

Table 1: Sample of messages from social media platform

MessageID	Message
1	You look amazing. Do you have one we can see you in the full dress?
46	hi girl sweety love you where you from im philipines
49	All your yoga poses is so beautiful and you are really gorgeous.
77	Do you want to have sex?
84	Can I see your booooobs? I'm the only one watching
128	take panties off ha ha
131	wow yummy p ssy
183	r u shy? U got great body
185	You are hot af

Most of the messages collected from social media sites that most likely will inappropriately use language and words in them. For example, most of the comments are collected from YouTube live chat, where live streams of children exposing themselves on camera are shown in their live stream see Figure 2. Sexual predators then target these children into pressuring them to perform sexual acts during the Livestream.



6:13 PM Lydia Azcka Get you naked for ur fans !
6:13 PM C Bro Ski down is fine
6:13 PM Lydia Azcka Up
6:14 PM Lydia Azcka
6:14 PM C Bro Ski
6:15 PM Van Helsing [message retracted]
6:17 PM C Bro Ski yea
6:19 PM Lydia Azcka Yea let's go
6:19 PM Lydia Azcka No make up !
6:19 PM Lydia Azcka Put it deep in ur mouth
6:20 PM Lydia Azcka U don't have to worry
PM Lydia Azcka It's ok

Exposing YouTube Dark Side | Child Exploitation on Live streams #YouTubeWakeUp

5,448 views • Jan 2, 2019

Figure 2: Comments from a YouTube live chat of minor live streaming herself (Source: YouTube)

The grooming activity does not only occur on YouTube live streams. It is also prevalent on other social media sites such as Instagram. Sexual predators will aim for Instagram accounts featuring children and leave nasty and inappropriate comments in one of their posts to exploit them.

4.1 Human Annotator

The dataset is then rated whether is contain any sexual grooming content or not. This evaluation is done by two respondents that have been chosen based on their gender and age. A male respondent at 18 and a female respondent at 25 are chosen for this purpose. The reason behind this is that people from different genders or ages may have different perceptions of this topic; for example, an adult woman may consider most of the text to contain excessive grooming content, while a young male teenager may consider only some of the text reader to contain said extreme grooming content. Also, there might be examples where some comments are evaluated as non-grooming content where the other respondent thinks it is. Data listing will then be labeled and filtered based on their criteria. For labeling, the datasets are classified as 0 (no sexual content) or 1 (with sexual content), see Table 2.

Table 2: The datasets with their appropriate rating

Message ID	Message	Message Rate
1	You look amazing. Do you have one we can see you in the full dress?	1
60	please get undressed	1
157	How old are youu lol	0

4.2 Inter-rater Agreement

The Kappa statistic is frequently used to measure interrater reliability. This is important because it represents the extent to which the data collected in the study are correct representations of the variables measured. Interrater reliability is a concern to one degree or another in most large studies because multiple people collecting data may experience and interpret the phenomena of interest differently. For this project, percent agreement is used for measuring human rater reliability [11]. The agreement of has Inter-rater agreement between the two human rater is 94.3%. This proves that the manual rating method of data collection for this project possesses good reliability for algorithm development.

4.3 Classification Algorithm

Choosing a particular classification algorithm, such as neural networks, nearest neighbor classifiers, decision trees, naïve Bayes classifiers, or support vector machines, will impact the types of patterns a classifier can recognize. However, this strongly depends on the nature of the input data [2]. The neural network techniques have been chosen for this research as a text classification algorithm. This algorithm is widely used in natural language processing tasks aiming at automatically classifying text documents into one or more defined categories [9] [12]. Specifically, we will apply sequence and n-grams models for evaluation purposes [13].

5.0 EXPERIMENT SETTING

Before unstructured texts in the raw data can be used as datasets in the classification algorithm, they must be preprocessed. We removed all punctuation marks, numbers, and additional white and converted all uppercase letters to lowercase letters. Table 3 summarizes the important metrics for the datasets used in this experiment. The dataset is partitioned into a training set (85% of the comments) and a test set (25% of the comment).

Table 3: Information about the dataset

Metrics	Total
Number of samples	1240
Number of classes	2
Number of samples per class	620
Average number of words per sample	6

The experiment is done using the Keras framework with Tensorflow [14]. The number of epochs used in training is 10. The optimizer is Adam with default parameters. The activation function is traditional sigmoid.

6.0 RESULT AND FINDINGS

We used accuracy as the metric in the experiments to measure the performance of the classification algorithm, which is built using our datasets. Accuracy refers to the number of correct classification predictions divided by the total number of predictions. The result of the classification algorithm is depicted in Figure 3. As shown in the figure, the performance using the n-gram model is better than the sequential with a score of 81.6% accuracy while accuracy score for sequential is 69.8%. This result indicated that n-gram model is suitable for small datasets is it also mentioned in [15].

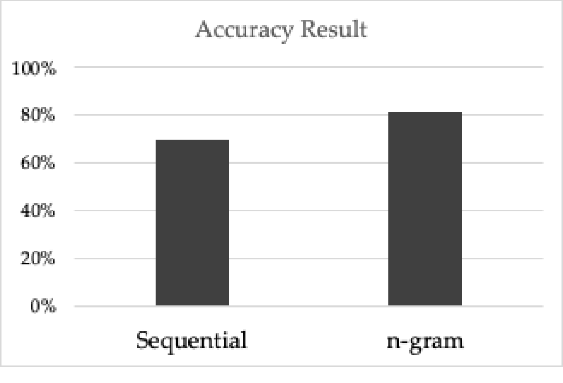


Figure 3: Accuracy result

7.0 CONCLUSION

In this paper, we demonstrated how the raw dataset could be created from social media content and apply it in classification algorithms development. For example, this algorithm can be deployed for future use in a web-based system such as a browser extension. This can be as an approach for implementing awareness intervention sexual grooming conversation. This can be a part of education as a means of prevention and protection from cybercrime.

REFERENCES

[1] H. Whittle, C. Hamilton-Giachritsis, A. Beech, and G. Collings, “A review of young people’s vulnerabilities to online grooming,” *Aggression and Violent Behavior*, vol. 18, no. 1. 2013. doi: 10.1016/j.avb.2012.11.008.

[2] K. Christensen, S. Nørskov, L. Frederiksen, and J. Scholderer, “In Search of New Product Ideas: Identifying Ideas in Online Communities by Machine Learning and Text Mining,” *Creativity and Innovation Management*, vol. 26, no. 1, 2017, doi: 10.1111/caim.12202.

[3] W. Lehnert, S. Soderland, D. Aronow, F. Feng, and A. Shmueli, “Inductive text classification for medical applications,” *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 7, no. 1, 1995, doi: 10.1080/09528139508953800.

[4] O. Kaurova, M. Alexandrov, and O. Koshulko, “Classifiers of Medical Records Presented in Free Text Form (GMDH Shell application),” International Conference in Inductive Modelling ICIM, 2013.

[5] F. Assef, M. T. Steiner, P. J. Steiner Neto, and D. G. D. B. Franco, “Classification Algorithms in Financial Application: Credit Risk Analysis on Legal Entities,” *IEEE Latin America Transactions*, vol. 17, no. 10, 2019, doi: 10.1109/TLA.2019.8986452.

[6] S. Xu, “Bayesian Naïve Bayes classifiers to text classification,” *J Inf Sci*, vol. 44, no. 1, 2018, doi: 10.1177/0165551516677946.

[7] J. R. Quinlan, “Induction of Decision Trees,” *Mach Learn*, vol. 1, no. 1, 1986, doi: 10.1023/A:1022643204877.

[8] C. C. Chang and C. J. Lin, “LIBSVM: A Library for support vector machines,” *ACM Trans Intell Syst Technol*, vol. 2, no. 3, 2011, doi: 10.1145/1961189.1961199.

[9] S. Lyu and J. Liu, “Convolutional recurrent neural networks for text classification,” *Journal of Database Management*, vol. 32, no. 4, 2021, doi: 10.4018/JDM.2021100105.

[10] M. Malekzadeh, P. Hajibabaee, M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, “Review of Graph Neural Network in Text Classification,” in 2021 IEEE 12th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2021, 2021. doi: 10.1109/UEMCON53757.2021.9666633.

[11] J. Belur, L. Tompson, A. Thornton, and M. Simon, “Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making,” *Sociol Methods Res*, vol. 50, no. 2, 2021, doi: 10.1177/0049124118799372.

[12] M. H. Goldani, S. Momtazi, and R. Safabakhsh, “Detecting fake news with capsule neural networks,” *Appl Soft Comput*, vol. 101, 2021, doi: 10.1016/j.asoc.2020.106991.

[13] T. Georgieva-Trifonova and M. Duraku, “Research on N-grams feature selection methods for text classification,” in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1031, no. 1. doi: 10.1088/1757-899X/1031/1/012048.

[14] F. Ertam, “Data classification with deep learning using tensorflow,” in *2nd International Conference on Computer Science and Engineering, UBMK 2017*, 2017. doi: 10.1109/UBMK.2017.8093521.

[15] M. García, S. Maldonado, and C. Vairetti, “Efficient n-gram construction for text categorization using feature selection techniques,” *Intelligent Data Analysis*, vol. 25, no. 3, 2021, doi: 10.3233/IDA-205154.