iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** XII **Month of publication:** December 2020

DOI: https://doi.org/10.22214/ijraset.2020.32564

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XII Dec 2020- Available at www.ijraset.com

Co-existence of Privacy and Security of Minors on Social Networks

Dhruv Mehta[1], Neelam Gujar[2], Preet R. Shah[3], Prof. Stevina Dias[4]
[1, 2, 3, 4]*Information Technology, D. J. Sanghvi College of Engineering,* Mumbai, India

Abstract: ***Currently, the world is in a technological era where everyone has easy access to the internet. Adding to this is the pandemic that has forced the schools to conduct classes online. This requires parents to give access to mobile phones and internet to children. The problem arises when children with access to internet don't use it responsibly. They can be easily exploited by people with malicious intent. They need to be protected from such people. In order to achieve this, we suggest a system which is a combination of artificial intelligence and human interference. A system which introduces human interference only when it is required. With the proposed system, we wish to ensure the safety of children by ensuring no abusive content reaches them and no unknown entity can contact them without their parent's permission. We wish to ensure their privacy while trying to protect them. In order to achieve this, we have placed a filtering system at the server side that filters the messages and only allows non-abusive content to reach the recipient. If any abusive content is detected, the message is sent to the parents of the sender as well as the receiver. This way, parents do not have to be bothered to check all the messages of their child and hinder their privacy.***

Keywords: Chat application, minors, privacy, protection, social networks, social media.

I. INTRODUCTION

We live in a technological era where everyone has access to the internet and all of its resources. Along with this, due to the ongoing pandemic, the schools are forced to shift classes to online medium. As a result, parents need to give access to mobile phones and internet to their children. In a world where children aren't taught about responsible online presence, their privacy and security become a big concern. While many companies have strong policies in place for security of minors, most of the time, it isn't enough. Most children don't know about privacy, they are not even aware of the risks they might face on SNS (Social Networking Site). This lack of knowledge among children is a pressing concern for parents. Children are exposed to various kinds of cybercrimes and privacy risks such as harassment, stalking and identity theft. Over time, the risk has only increased. In order to ensure their security, we need to transfer the control to the parents in a non-invasive manner. The goal is that children can easily have more granularity control over how and with whom to share their information.

II. LITERATURE SURVEY

A. *Results of a Survey*

The following are the results of a survey conducted by the authors of [5] in 2015. The survey included a total of 1800 participants. Students from primary school, middle school and high school were participants of the survey.620 students owned at least one account on a social networking site.

1) Out of the 1800 participants, 550 were between the age of 6 and 12.
2) In the survey, the participants were asked if the details shared by them on their profiles were true. About 85% of them shared their real names, 53% shared their actual phone number and 35% of them shared their actual home address on their public profile.
3) Out of the 35% students who shared their actual home address, 14% of them were primary school students.
4) Along with this, these primary school students accepted requests from strangers.

B. *Study of Various social Media Apps*

The results from the study of a few popular social media apps and the different methods they use to avoid abuse of minors have been tabulated in TABLE I. The following are a few of the methods used by various social media apps.

1) *Reporting:* Certain apps allows its users to report other users as abusive or as spam.
2) *Educating minors about privacy:* Certain social media apps try to educate their users who are minors regarding how to maintain their privacy on the app.
3) *Strong Policies:* Certain social media apps have strong policies against minor abuse.
4) *Moderation:* Certain social media apps employ people to moderate the content on their apps. These people check if the content that has been reported by people for abuse do actually abuse the company policy or not. If they do, the company takes action against the user, if required.

TABLE I. Different Social Media and their methods to avoid abuse of minors.

Parameters	Facebook	Instagram	Snapchat	WhatsApp
Reporting	Yes	Yes	Yes	Yes
Educating children about privacy	Yes	No	No	No
Strong Policies	Yes	Yes	Yes	No
Moderation	Yes	Yes	No	No

C. Existing Systems

1) *Crowd-Supported Detection:* Popular social media platforms such as Facebook, Twitter, YouTube, etc. use crowd-sourced reporting features to identify actions indicative of cyber bullying and/or abuse. [6]
2) *Natural Language Processing (NLP) Based Techniques:* There is a significant body of work in academia to detect cyber abuse from an NLP. Using a combination of words, sentiment analysis, and sequences of words (by carefully using a database of suspicious words), the idea is to design learning. [6]
3) *Approaches Taken by Parental Control Apps in the Market Today:* There is now a lot of demand for apps that parents can install on devices of their children that flag inappropriate content. Upon research, we find that existing apps primarily rely on only identifying suspicious keywords (e.g., “die", “hate", “drugs", “abuse", etc.) or inappropriate content in images (e.g., a bottle of alcohol, or a knife, or a gun) to flag them. Very minimal NLP or Image Processing is accomplished in this regard. Naturally, the false positive rate is too high, and after sometime, interests of parents wane out due to many false alarms. [6]
4) *Deep Learning:* There are some APIs having capability to detect abusive images using deep learning by building some complex models. These techniques use different classification algorithms having different accuracy on test data. Hence 100% accuracy is not possible. [6]

III. PROPOSED SYSTEM

A. Ideology

We propose a system where we try to strike a balance between the privacy and security of minors by giving some control to the parents/guardians of these minors. By some control, we mean, the control related to the people who can get in touch with the minors, the ability to report a person/contact as well as to block them. Along with this, the proposed system system will be powered by Artificial Intelligence that will detect abusive messages and content and notify the parents/guardians of both, the sender as well as the receiver. This will ensure privacy of minors by ensuring that only abusive messages are sent to parents rather than all of the messages. In Fig. 1., depicts the architecture of the proposed system. The parent will use the Parent App to interact with the server. The app will allow the parent to check for any chat requests, group requests, any messages that have been identified as abusive by the algorithms. The child will be able to interact with the server using the Child App. They will be able to send messages, send message requests, send group requests, etc. The server will contain the database, message queue, etc. This will enable messaging and sending and receiving requests. Different machine learning/deep learning models will be deployed on the cloud to classify the messages as abusive or non-abusive. The server will interact with the deployed models to classify the messages and then divert them to their appropriate receivers.

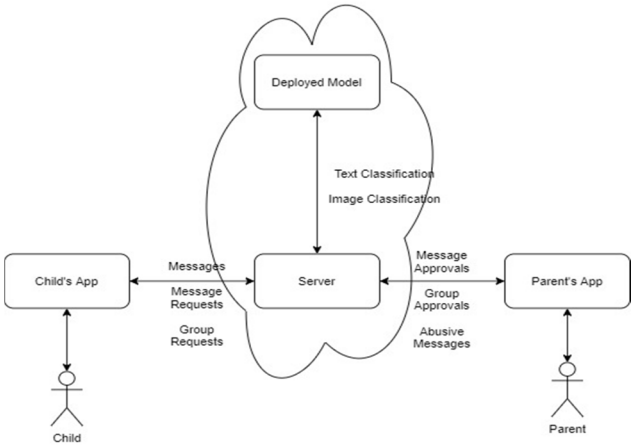


Fig. 1. System Architecture

B. Study Of Various Algorithms For Abusive Content Classification

1) *Text Classification Algorithms:* Text filters, also known as profanity filters are quite popular in gaming industry. As the intended users of games are mostly children, gaming companies have to implement profanity checks in their games, if they include online chatting. These profanity filters are industry standard and hence, very robust. The authors of [7] have did a comparative study between their system and other profanity systems of popular online games. From [7], we learn that online games have profanity filters that filter the profane (abusive) words from the in-game chat. However, people come up with new ways to use such words by slightly changing the words or using abbreviations. For example, the term "Son of a Bitch" can be used as "SoB" or as "50n of a 8itch" or the term "Shit" can be used as "Sh*t" or as "5hit". These words are coined so as to avoid them from getting filtered. The authors of [7] suggest that a static filter won't be able to keep up with the newly coined terms. So, we need a dynamic filter that can identify newly coined terms. They suggest an algorithm called "Approximate String Searching". For this purpose, they use an R* tree which is an extension of R-tree. Fig. 2., depicts the structure of such a system. The system will have a database of pre-identified vulgar (abusive) words and a database of normal (non-abusive) words. Any word that does not match the normal words database, it will be tested for being abusive as it may either be abusive or someone's name. A word will be considered as abusive if it is found in the abusive word database or if it has an alignment score closer to 1.

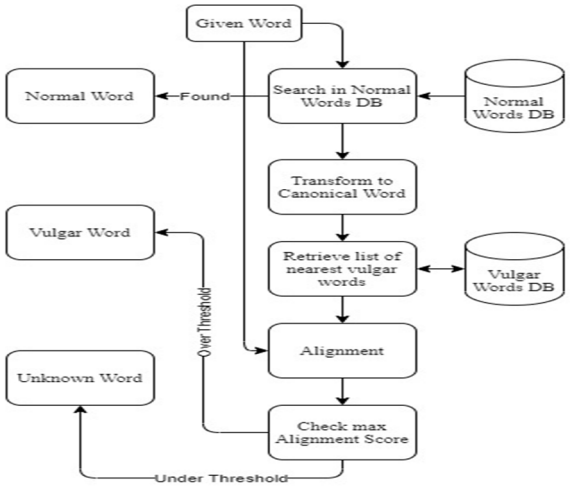


Fig. 2. Filtering System Overview. [7]

In TABLE II., the authors of [7] define the similarity score between certain letters and certain characters.

TABLE II. Examples of Matching Value for Similar Characters. [7]

Source	Destination	Score
a	@	0.8
s	$	0.8
an	&	1.2
i	!	0.4

In TABLE III., the authors of [7] define how the algorithm will find the alignment score of different words.

TABLE III. Examples of global alignment. [7]

Source	Destination	Score	Alignment Result			
f!ll	fill	3.4	f	i	l	l
			f	!	l	l
			1.0	0.4	1.0	1.0
$uck	suck	3.8	s	u	c	k
			$	u	c	K
			0.8	1.0	1.0	1.0

The authors of [7] draw comparison between their system and the filtering system of different popular online games. The results suggest that even if the online games have a robust filtering mechanism, they do not work as well for new terms coined by the users. Their system is around 3-9 times accurate as that of the online games. However, one consideration maybe related to the speed of the system in filtering the sentences.

The authors of [1] use Aho-Corasick algorithm for offensive word detection. The Aho-Corasick algorithm builds a tree for all the keywords. Hence, we need an exhaustive set of abusive keywords for pattern matching using Aho-Corasick algorithm.

a) The algorithm works best if one has a set of words that is static and does not change over time.

b) With the help of this algorithm, we can match patterns from given text and find their occurrences in the text. If a keyword is found, it stores the keyword's occurrence in a separate array.

c) These indexes help us to mask those keywords with other characters such as '*'.

The algorithm works in two parts are as follows:

- To build a trie, it is an efficient information re***trie***val data structure, from the keywords. Fig. 3. is an example of a trie created for the keywords "he", "his", "she" and "hers".
- After building the trie, it searches the text from the given keywords.

The time complexity of the algorithm is O(n+m+z), where 'n' is the length of the text from which keywords are to be searched, 'm' is the total number of characters in all the keywords and 'z' is the total number of occurrences of keywords in the given test.

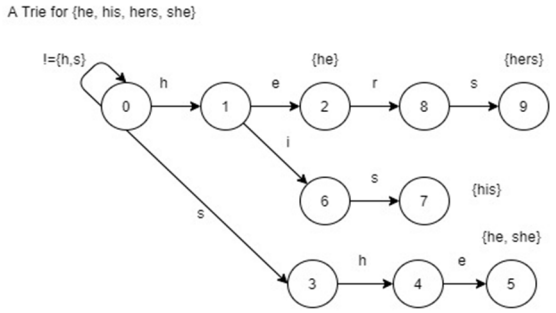


Fig. 3. Example of Trie for Aho-Corasick Algorithm

2) *Image Classification Algorithms:* For abusive image classification, the authors of [4] used ResNet50 architecture for the classifier which shows high accuracy up to 95% in the test set. Long Short-Term Memory (LSTM) networks to train the classifier. LSTMs model sentences as the chain of forget-remember decisions based on context. In TABLE V., we have compared various deep learning algorithms like Alexnet, Resnet-50, etc. based on various parameters such as accuracy, error and advantages.

TABLE IV. Comparison between different Deep Learning Algorithms.

Parameters	Alexnet	Resnet-50	VGG	InceptionV3
Top-1 Accuracy	57.1%	75.2%	70.5%	78.8%
Top-5 Accuracy	80.2%	93%	91.2%	94.4%
Advantages	Fixes the overfitting issue by using a dropout layer after every FC Layer.	Solves the vanishing gradient problem by using skip connections.	It learns more complex features at a lower cost as compared to Alexnet.	It is very fast as compared to VGG as it converts the dense neural network to a sparse one by eliminating unnecessary neurons.
Top-5 Error Rate	15.3%	3.6%	7.3%	6.67%

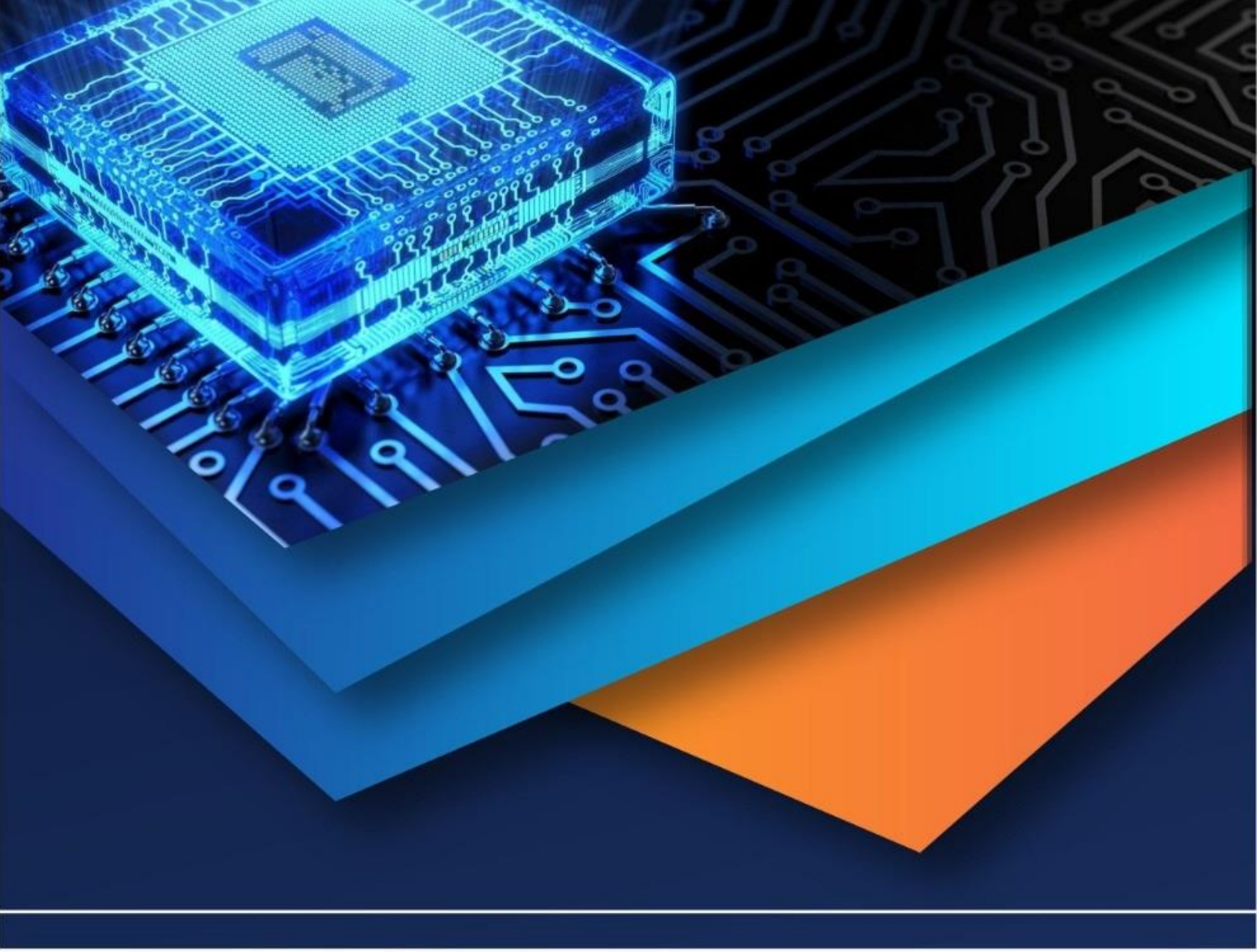
For the proposed system, we will use Approximate String Searching as the text classification algorithm, and ResNet50 as image classification algorithm. However, the use of the different algorithms may depend on their speed as speed is an important factor with respect to any chat application.

IV. CONCLUSION

From the study of the various papers, it has been observed that a number of different systems have tried various methods to try and protect minors on the internet. Some of these methods include moderation, reporting, educating minors about privacy and strong policies. Even though, content moderation works it cannot be used with respect to private chats because it hinders privacy of the users. This calls for an autonomous system that can determine whether the messages sent by one user to another are abusive or not. Such a system maybe slower for its main purpose like chatting. However, it will ensure protection of minors from abusive content.

REFERENCES

[1] Mr. Shashank H. Yadav and Mr. Pratik M. Manwatkar, “An Approach for Offensive Text Detection and Prevention in Social Networks”, IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICECS 15
[2] Nasriah Zakaria, Lau Keng Yewl, Nik Mohd Asrol Alias, Wahidah Husain, “Protecting Privacy of Children in Social Networking Sites with Rule-based Privacy Tool”
[3] Adrienne F. and David E. (2008). “Privacy Protection for Social Networking APIs”. The Web 2.0 Security and Privacy 2008 (in conjunction with 2008 IEEE Symposium on Security and Privacy).
[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. (2015). “Deep Residual Learning for Image Recognition”. Microsoft Research.
[5] P. Santisarun and S. Boonkrong, "Social network monitoring application for parents with children under thirteen," 2015 7th International Conference on Knowledge and Smart Technology (KST), Chonburi, 2015, pp. 75-80.
[6] Atanu Shom, Md. Mizanur Rahman,Sriram Chellappan, A. B. M. Alim Al Islam, “A Generalized Mechanism beyond NLP for Real-Time Detection of Cyber Abuse through Facial Expression Analytics” 2019 Association for Computing Machinery. ACM ISBN 978-1-4503-7283-1/19/11
[7] Taijin Yoon, Sun-Young Park and Hwan-Gue Cho, “A Smart Filtering System for newly Coined Profanities by Using Approximate String Alignment” 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010.129)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY