CyberCops: Combating Facebook Cyberbullying using NLP Classification Techniques

***Abstract*—Unprecedented opportunities for connection and communication have been brought about by the growth of the internet and social media. Cyberbullying, a new type of bullying, has emerged as a result. With so many people using the internet, cyberbullying is becoming a widespread problem that affects individuals of all ages and backgrounds. Cyberbullying may have long-lasting impacts on victims, from painful messages to harmful posts. Today's methods for determining how vulnerable critical infrastructure mostly depend on the personnel of security operations centers. (SOCs). (2017) Feng and co. However, dynamic CIs and attack surface evaluation require certain properties that traditional manual and subjective audits lack. In this thesis paper, we propose a method for detecting cyberbullying on Facebook using Natural Language Processing (NLP) techniques. NLP is a subfield of artificial intelligence that focuses on the interaction between computers and human language. Our approach involves analyzing the content of Facebook posts and comments using NLP to identify patterns and markers of cyberbullying. By doing so, we hope to provide a more effective and efficient means of detecting and preventing cyberbullying on Facebook, which could ultimately lead to a safer and more positive online environment.**

***Index Terms*—Cyberbullying, NLP, Facebook, CIS,Victims**

I. Introduction

The employees of security operations centers (SOCs) are widely utilized in present vital infrastructure vulnerability assessment methods. (2017) Feng et al. Conventional manual and subjective audits, however, fall short of the necessary qualities of dynamic CIs and attack surface assessment. For example, certain frequently used tools for determining vulnerability severity, such the CVSS1 calculator, need input from users and depend on qualitative evaluations of vulnerability features, such as exploitability, scope, and repercussions. (2011) Joh and Malaiya. However, there is a wealth of textual information about cybersecurity, including in vendor declarations, blogs, whitepapers, and hacker forums. Traditional threat analysis of data from textual sources uses the time-consuming and unproductive human labor technique. The security analysts are therefore unable to properly utilize the validated cybersecurity information to react to cyber threats in a timely and correct manner.

II. Existing Works

A. Literature Review

Cyberbullying is a serious problem in the modern world, especially on social networking sites like Facebook. The victim of electronic intimidation or harassment may experience negative effects on their mental and emotional health. In an effort to put an end to the practice, researchers have looked into the use of natural language processing (NLP) techniques to automatically detect instances of cyberbullying on social networking sites like Facebook. This literature review looks at recent studies on Facebook's for detecting cyberbullying.

It is more important to recognize cyberbullying as social networking platforms continue to grow in popularity. Many studies have used machine learning techniques to solve this issue. In one of these studies, supervised machine learning was used to ascertain the sentiment and meaning of sentences. Unfortunately, the accuracy rating of this algorithm was only 61.9%, indicating that it may not be trustworthy for detecting cyberbullying. [1].

Cyberbullying on social media is an increasing problem, making its identification and prevention essential. To detect cyberbullying on Twitter-based networks, present a supervised machine learning method [2]. Users' actions and tweet content are two factors that the authors of the study choose from Twitter and put into an identification algorithm. One of the challenges in developing such a model is the requirement for a robust and representative dataset. A sufficient number of tweets must be included in the dataset to adequately represent both cyberbullying as well as non-cyberbullying behavior.

Dinakar et al. use a variety of machine learning methods, including logistic regression, decision trees, and assistance vector machines, to develop the detection model [3]. They experiment with several feature sets and find that a combination of actions, user, and tweet material components produces the greatest results. As shown by the assessment of the detection system, which achieved the f-measure of 0.936 as well as a region within the receiver-operating curves of 0.943, it effectively detects harassment tweets with a high level of accuracy.

Zhao et al. studied the automatic detection of cyberbullying on social media: A deep learning approach [4]. Using a significant dataset and cutting-edge neural network architecture, the authors try to identify and classify various types of cyberbullying. The key difficulties tackled by this research are the variety of cyberbullying content, the discrepancy among instances of positive and negative bullying, and the obligation to take into account the temporal and situational nature of social media posts.

Ahmed et al. used a deep learning-based technique that combines an LSTM network with a network of convolutional neural networks (CNN) to extract and contextual information from social media posts. The recommended method showed outstanding accuracy of 87 percent, outperforming several

state-of-the-art techniques. The paper [5] proposes a hybrid deep learning approach for the identification of cyberbullying in social media.The authors aim to overcome the challenges of finding and categorizing cyberbullying content in large-scale social media data by harnessing the strength of deep learning and natural language processing approaches.The difficulties in developing a highly accurate cyberbullying detection system are noted as being the complexity of social media data and the need for trustworthy feature extraction. By combining a convolutional neural network (CNN) and a bidirectional long short-term memory network (BiLSTM), the proposed hybrid technique categorizes cyberbullying content. The accuracy rate the authors achieved on the dataset they used was 95.2%, outperforming state-of-the-art models at the time.

To improve accuracy and reduce false positives, Iwendi et al. [6] propose a novel deep learning approach for recognizing cyberbullying via social media. The authors battled with a number of challenges, such as the dataset's asymmetries, the variety of comments that weren't about online harassment, and the vocabulary used in social media. The suggested method is a multi-layer convolutional neural network called the Cyberbullying Detecting Network (CDN), which combines character and word embeddings. With an accuracy of 95.73%, the CDN outperformed several other state-of-the-art algorithms when it came to of recall, accuracy, and F1-score.

B. Research Problem

Manual vulnerability assessment might produce erroneous results and call for time-consuming investigation. These security issues are also exacerbated by the diversity, incompleteness, and redundancy of security information in contemporary repositories. These issues are prevalent in both manufacturer and public security reports, making it difficult to find and fix security flaws by direct study. [7]

It is challenging to weed out important data from the massive flow of data. The amount of digital text information is fast expanding due to the expansion of social media and pervasive computers. One of the main duties of a safety center is to enable correlation. [8].By locating information on cyber threats, analysts may become more proactive in their risk assessment and monitoring. Due to the growing variety of systems of information used in security operations, it is essential to gather cyber threat-related data specific to a company to the public internet in order to achieve this.

In the text, it is discussed how cognitive cyber-physical networks of the Internet with Things (CPS-IoT) provide problems for the health care sector. Due to the increased cognitive complexity of these systems, both conventional CPS-IoT threats, as well as new threats arising from their innate cognitive functionalities, can affect them. Additionally, the widespread use of CPS-IoT broadens the attack surface, raising the risks to public safety for critical infrastructure. The growing interconnection of medical equipment and services, which exposes them to new cybersecurity vulnerabilities, makes hospitals and infrastructures even more susceptible to significant security hazards. According to reports, medical identity theft and cyberattacks are on the rise everywhere. The dynamics, complexity, unpredictability, and extensive connectivity of CPS-IoT-enabledconnected healthcare services and vital infrastructure are too much for the present security solutions to handle. The suggested solution calls for the development of cutting-edge methods for creating cognition cybersecurity of healthcare systems utilizing cognitive architecture as well as artificial intelligence to improve automated intelligence cybersecurity making decisions mechanisms with professional-level aptitude. [9]

It is difficult to identify online assaults within URLs, and there are a number of significant problems that must be solved. First off, since different assaults might conceal themselves in various ways inside their URLs, there is a demand for an efficient method to convert every type of URL into representation. Second, choosing features is difficult since various assaults have distinct signatures in their URLs. Thirdly, because many deep learning cyber security models only have a single algorithm for detection, it is challenging to maintain the system as new attacks are discovered. Finally, the centralization feature in the IoT cloud the environment can affect how distributed services, like IoT application network security mechanisms, are applied. [**?**]

The small number of items of particular categories in the labeled dataset can be explained in this way. It might not be immediately obvious who or what is behind an assault, and it can take some time to obtain the information required to pinpoint the parties responsible. Furthermore, some attacks might be executed by previously unidentified entities or categories, making it challenging to categorize them beforehand. As a result, it's possible that not all attack types and parties engaged were included in the annotated dataset, and more investigation and data gathering may be required to boost the effectiveness of autonomous detection systems. [10]

It is difficult to sort the vast amount of data into what is actually meaningful. Due to the growth of online communities and widely used computers, the volume of digital text material is rapidly increasing. To allow correlation is one of a security center's primary responsibilities. An analyst's situational awareness will improve with the identification of data related to cyber threats, enabling proactive detection and risk reduction [2]. The need to extract business-specific cyber threat-related information from the public internet is a result of the increasing amount of information systems used in security operations.

Additionally, it is discussed how the healthcare sector faces issues as a result of cognitive cyber-physical systems such as the Internet of Things (CPS-IoT). Due to these systems' increasing cognitive complexity, they are susceptible to both traditional CPS-IoT weaknesses and new threats resulting from their inborn cognitive functions. Furthermore, as a result of the increased attack surface brought on by the widespread adoption of CPS-IoT, critical infrastructure poses a greater risk to public safety. Healthcare infrastructures and services are increasingly more vulnerable to serious security risks as a result of the expanding interconnectivity of medical

equipment and goods or services, thus exposing businesses to new vulnerabilities in cybersecurity. Cyberattacks and medical identity theft are reportedly on the rise everywhere. Current security solutions are unable to handle the dynamic the natural world, complexity, unpredictability, and extensive connectivity to CPS-IoT enabled medical facilities and critical infrastructure. [3].

The study suggests a distributed approach for deep learning methods like CNNs as well as NLP models that uses URLs to identify online attacks. To improve system stability, the system may represent all types of URLs, discriminate between abnormal and regular requests, and use many concurrent models. The article also suggests a general distributed online assault detection method for cloud edge devices. The study is divided into many sections, including an overview of relevant works, the structure and method of the recommended system, data sets and experimental conditions, the experimental findings including discussion, with the result and future work. [4].

C. Research Objectives

NER has demonstrated success with BERT converters in named entity recognition and other natural language processing tasks. In situations when the annotation dataset is limited or lacks variety, the authors of this paper advise using BERT transformers to improve NER performance. In order to add keywords with automatically labeled named things to the training dataset, they also suggest an automated dataset augmentation approach. This paper examines the performance of different BERT models, incorporating an international model, a model modified on Russian data, as well as a model customized for cybersecurity literature. The authors also provide a novel method for information enhancement for NER tasks and investigate the influence of different factors on the performance of the NER system. [11].

The purpose of this study is to evaluate how well LSTM-based models of neural networks retrieve cybersecurity-related data. The paper assesses three distinct LSTM architectures on relation extraction (RE) tasks and contrasts the performance of an LSTM-based version with a CRF-based framework for recognizing named entities (NER) tasks. The National Risk Database (NVD) is used to construct a word embeddings model, which is then used to train the models on a corpus containing vulnerability descriptions. The models are judged according to their accuracy throughout training and testing, memory, F1 scores, and precision. [12].

The goal is to find cyber intelligence in natural language writings that correspond to terms that match the words in bold font in the table. One may not be able to translate the text if the program fails to differentiate between ”Windows” with ”operating system.” The attacker records the victim's computer's Windows version information. It is possible to enumerate all comparable formulations of a particular cyber idea, hence this is a difficult undertaking. [13].

This research aims to identify cybersecurity intelligence (CSI) in tweets. Gathering the most recent CSI is crucial to preventing or minimizing harm from hostile assaults, which is what cybersecurity refers to as defending digital systems and data against. OSINT professionals frequently discuss technical specifics and relate their personal experiences with cyber-attacks on Twitter. Positive CSI tweets include vital details about criminals, vulnerabilities, and targets that may help professionals respond to attacks in a productive way. Using CSI terms like typical exposures and vulnerabilities (CVE), the analysis divides tweet into two distinct groups: positive and negative. [14].

III. Methodology

Cyberbullying is on the rise in today's digital age, particularly on social media sites like Facebook. It is a type of harassment, coercion, or abuse directed toward those who utilize digital technology. Due of the massive amount of text that individuals post on Facebook every day, it could be difficult to see cyberbullying there. It has been investigated how methods based on natural language processing, or NLP, may be used to automatically detect cyberbullying on communication platforms. In this work, we use NLP techniques to build a Facebook detection of cyberbullying system.

A. Data Collection

In the first stage of this research, a dataset of Social remarks and posts is gathered. Using the Facebook Application Programming Interface (API), which enables us to obtain publicly available posts and comments concerning cyberbullying, we will gather the data. Based on predetermined hashtags and phrases connected to cyberbullying, we will narrow down the postings and comments. The dataset will next go through preprocessing to make sure that it is entirely text-based and to get rid of any extraneous data.

B. Preprocessing

After that, the dataset has to be ready for analysis. The dataset's emojis, photos, links, and other non-text data will all be eliminated. Following tokenization, or the breakdown of the text into individual words, lemmatization and stemming techniques will be used to normalize the data. Stopwords, or everyday words like ”the” and ”and,” which do not further the text's theme, will be eliminated. The remaining content will be altered to lowercase to maintain consistency.

C. Extraction

The third stage involves extracting pertinent characteristics from a preprocessed dataset. Using methods from natural language processing (NLP), such as bag-of-words, that depicts the text as a collection of phrases, term recurrence-inverse document frequency, which assesses the importance of every word in a paperwork, and word encoding, which represents words as vectors, the text data will be transformed into numerical features. These characteristics will be used in the training of the cyberbullying detection model.

D. Model Development

The fourth stage involves developing a model for recognizing cyberbullying. A variety of machine learning techniques, such as decision trees, logistic regression, and support vector machines, will be used to create the model. We'll also investigate the application of deep learning methods like recurrent neural networks and transformers. The models will be trained using the preprocessed dataset's features.

E. Model Evaluation

In the final stage, the developed cyberbullying identification model's efficacy is evaluated. Several metrics, including precision, recall, precision, and F1-score, will be used to evaluate the model's effectiveness. The model will be tested on a different dataset to see how well it generalizes to brand-new data. The evaluation will help us figure out how well the algorithm can spot Facebook cyberbullying incidents.

IV. Model Training

A. Data Processing

Due to a shortage of data, we initially acquired the necessary information from a number of sources such as Kaggle, GitHub, and additional websites.Before being prepared for pre-processing, the dataset was changed and combined with some information I had personally collected. Next, we checked the total count and any fields that were null. The words were later concatenated into just one string after we removed any extraneous material. The completed data for the main job is delivered to us. Following that, we go over some NLP techniques with a focus on data training.

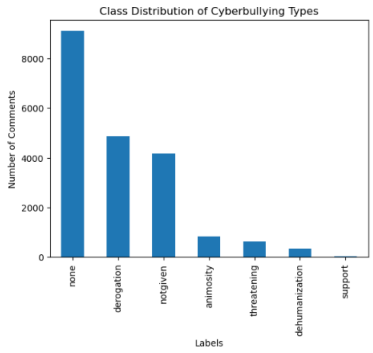

Fig. 1: Data frequency

B. Equation and Calculation

True Positive: A circumstance in which you can predict with certainty a beneficial outcome.

False Positive: If you anticipate a positive result but it is really negative.

False Negative: Let's say you predict a bad result, but it comes out to be a good one.

True Negative: A scenario in which you can anticipate that something bad would happen.

a) Accuracy = (TP+TN)/(TP+TN+FP+FN)
b) Recall = TP/(TP+FN)
c) Precision = TP/(TP+FP)
d) F1 Score = (2 * Precision * Recall) / (Precision + Recall)

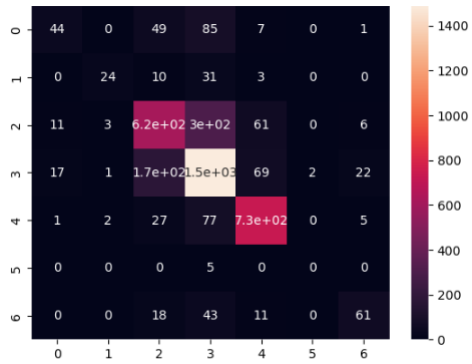

Fig. 2: Confusion Matrix

C. Classification Result

Here, we train a logistic regression model on our dataset.Although this model's performance is only marginally satisfactory, we are interested in developing another model in order to increase accuracy.The following is the analysis of the logistic regression model:

	precision	recall	f1-score	support
animosity	0.60	0.24	0.34	186
dehumanization	0.80	0.35	0.49	68
derogation	0.69	0.62	0.65	998
none	0.73	0.84	0.78	1768
notgiven	0.83	0.87	0.85	842
support	0.00	0.00	0.00	5
threatening	0.64	0.46	0.54	133
accuracy			0.74	4000
macro avg	0.61	0.48	0.52	4000
weighted avg	0.73	0.74	0.73	4000

Fig. 3: Classification Report

Conclusion

In this study, we proposed a system architecture that might potentially extract information about cyberthreats automatically to aid human operators. We tested the applicability of the Natural Language Filter component of the system by using the neural embedding method doc2vec.According to our findings, a doc2vec-based natural language models trained with text data specialized to cybersecurity and special preprocessing might be used as a 74 percent accurate Natural Language Filter within the proposed autonomous system.We will continue to work on it using a variety of ways to ensure that the system remains as accurate as possible.

ACKNOWLEDGMENT

Praise be given to the Almighty, whose help made it possible for us to finish the research's composition. After that, we would want to express our gratitude to our family and friends for their support and their insightful comments on the study. Last but not least, we would like to thank Annajiat Alim Rasel sir for serving as our supervisor and helping me with this research.

REFERENCES

[1] D. Yin, Z. Xue, L. Hong, B. Davison, A. Edwards, and L. Edwards, "Detection of harassment on web 2.0," 01 2009.

[2] O. Mendsaikhan, H. Hasegawa, Y. Yamaguchi, and H. Shimada, "Identification of cybersecurity specific content using the doc2vec language model," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1, 2019, pp. 396–401.

[3] H. Abie, "Cognitive cybersecurity for cps-iot enabled healthcare ecosystems," in *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*, 2019, pp. 1–6.

[4] Z. Tian, C. Luo, J. Qiu, X. Du, and M. Guizani, "A distributed deep learning system for web attack detection on edge devices," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 1963–1971, 2020.

[5] M. Tikhomirov, N. Loukachevitch, A. Sirotina, and B. Dobrov, "Using bert and augmentation in named entity recognition for cybersecurity domain," in *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, H. Horacek, and P. Cimiano, Eds. Cham: Springer International Publishing, 2020, pp. 16–24.

[6] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," pp. 433–443, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563216303788

[7] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," 01 2011.

[8] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proceedings of the 17th International Conference on Distributed Computing and Networking*, ser. ICDCN '16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2833312.2849567

[9] M. F. Ahmed, Z. Mahmud, Z. T. Biash, A. A. N. Ryen, A. Hossain, and F. B. Ashraf, "Cyberbullying detection using deep neural network from social media comments in bangla language," 2021.

[10] C. Iwendi, G. Srivastava, S. Khan, and P. Reddy, "Cyberbullying detection solutions based on deep learning architectures," 10 2020.

[11] M. Tikhomirov, N. Loukachevitch, A. Sirotina, and B. Dobrov, *Using BERT and Augmentation in Named Entity Recognition for Cybersecurity Domain*, 06 2020, pp. 16–24.

[12] H. Gasmi, J. Laval, and A. Bouras, "Information extraction of cybersecurity concepts: An lstm approach," p. 3945, 09 2019.

[13] M. Das Purba, B. Chu, and E. Al-Shaer, "From word embedding to cyber-phrase embedding: Comparison of processing cybersecurity texts," in *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2020, pp. 1–6.

[14] H.-S. Shin, H.-Y. Kwon, and S.-J. Ryu, "A new text classification model based on contrastive word embedding for detecting cybersecurity intelligence in twitter," p. 1527, Sep 2020. [Online]. Available: http://dx.doi.org/10.3390/electronics9091527