Cyber Bullying Detection from Facebook by NLP Classification Techniques

Shaharear Hossain Emon[1],Parom Guha Neogi[2],Allama Bakhtiyar Nafis[3],Rezwana Chaudhary Raka[4],
Md Mustakin Alam[5], Md Farhadul Islam[6] and Annajiat Alim Rasel[7]
[1,2,3,4,5,6,7]Department of Computer Science and Engineering, BRAC University, 66 Mohakhali, Dhaka, Bangladesh
[1]shaharear.hossain.emon@g.bracu.ac.bd, [2]parom.guha.neogi@g.bracu.ac.bd,
[3]allama.bakhtiyar.nafis@g.bracu.ac.bd, [4]rezwana.chaudhury.raka@g.bracu.ac.bd,
[5]md.mustakin.alam@g.bracu.ac.bd, [6]md.farhadul.islam@g.bracu.ac.bd, [7]annajiat@gmail.com

***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

Current critical infrastructure vulnerability assessment procedures rely heavily on the personnel of security operations centers (SOCs). (Feng et al., 2017). However, traditional manual and subjective audits fall short of the desired attack surface evaluation and dynamic CIs characteristics. For instance, certain commonly used tools for calculating the severity of vulnerabilities, such as the CVSS1 calculator, need user input and are based on qualitative assessments of vulnerability characteristics including exploitability, scope, and consequences. (Joh and Malaiya, 2011).However, textual sources including vendor announcements, blogs, whitepapers, and hacker forums contain enormous amounts of cybersecurity information [2]. The lengthy and ineffective process of manual labor is used in traditional threat analysis of information from textual sources. As a result, the security analysts are unable to fully utilize the verified cybersecurity information to promptly and accurately respond to cyber threats.

II. Existing Works

A. Literature Review

In the present age, cyberbullying is a severe issue, especially on social networking platforms like Facebook. The mental and emotional well-being of the target of technological harassment or intimidation may suffer. Researchers have investigated the use of natural language processing (NLP) techniques to automatically identify instances of cyberbullying on social media platforms like Facebook in an effort to stop the practice. Recent studies on Facebook's NLP for identifying cyberbullying are examined in this literature review.

As social media platforms continue to gain popularity, it is becoming more and more crucial to identify cyberbullying. This problem has been addressed in a number of research using machine learning methods. In one such study, the sentiment and context of phrases were determined using supervised machine learning. Unfortunately, this algorithm's accuracy rating was just 61.9%, suggesting that it might not be reliable for identifying cyberbullying [6].

On social networks, cyberbullying is a growing issue, making its detection and prevention crucial [7]. Provide a supervised machine learning approach for the detection of cyberbullying on Twitter-based networks in the paper the authors choose certain features that are derived from Twitter, like user behavior and tweet content, and incorporate them into a detection algorithm. The need for a strong and representative dataset is one of the difficulties encountered when creating such a model. They must balance the dataset to include an adequate number of tweets that demonstrate both cyberbullying and non-cyber bullying behavior.

To create their detection model, the authors employ a number of machine learning algorithms, such as decision trees, logistic regression, and support vector machines [8]. They experiment with various feature sets and discover that the best outcomes come from a blend of behavior, user, and tweet content aspects. The evaluation of the detection system reveals that it detects cyberbullying tweets with a high degree of accuracy, achieving an f-measure of 0.936 and an area under the receiver-operating characteristic curve of 0.943.

The automatic detection of cyberbullying on social media:

A deep learning strategy is presented in this work with the title ”Automatic Detection of Cyberbullying on Social Media: A Deep Learning Approach” uses a sizable dataset and a cutting-edge neural network architecture, the authors attempt to recognise and categorize various types of cyberbullying [9]. The diversity of cyberbullying material, the disparity between positive and negative samples, and the requirement to take into consideration the temporal and contextual character of social media posts are the primary issues addressed in this research.

To gather the temporal and contextual details of social media posts, the authors utilized a deep learning-based method that combines a convolutional neural network (CNN) with an LSTM network. The suggested method outperformed numerous state-of-the-art methods and produced excellent accuracy of 87%. A hybrid deep learning strategy is suggested in the study titled ”A Hybrid Deep Learning Approach for Cyberbullying Detection in Social Media” for the detection of cyberbullying in social media [10]. By utilizing the strength of deep learning and natural language processing techniques, the authors hope to overcome the difficulties of identifying and categorizing cyberbullying content in large-scale social media data.The complexity of social media data and the requirement for reliable feature extraction are highlighted as challenges in creating a highly accurate cyberbullying detection algorithm. The proposed hybrid method classifies cyberbullying material by fusing a convolutional neural network (CNN) and a bidirectional long short-term memory network (BiLSTM). The authors outperformed current state-of-the-art models for accuracy, achieving 95.2% on the dataset they utilized.

In another paper, ”A Deep Learning Method for Forecasting Cyberbullying on Social Media” in order to increase accuracy and decrease false positives, the authors suggest an unique deep learning method for identifying cyberbullying via social media [11]. The authors struggled with issues including the asymmetry of the dataset, the diversity of messages that weren’t about cyberbullying, and the terminology employed in social media. The Cyberbullying Detection Network (CDN), the proposed technique, is a multi-layer convolutional neural network that integrates word embeddings with character embeddings. In terms of precision, recall, and F1-score, the CDN surpassed numerous other cutting-edge algorithms, with an accuracy of 95.73%.

B. Research Problem

The practice of manually assessing vulnerabilities can lead to inaccurate information and require complicated analysis. Additionally, the variety, lack of completeness, and duplication of security data in modern repositories contribute to these security concerns. These problems are common in both public and manufacturer vulnerability reports, making it challenging to identify and address security weaknesses through direct analysis. [1]

Extracting pertinent information from the overwhelming influx of data is a difficult task. With the proliferation of social networks and widespread computing, the volume of digital text content is growing rapidly. One of the primary responsibilities of a security centers, allowing for correlation. [2]analyst is to gain situational awareness by identifying cyber threat-related information, enabling proactive monitoring and risk mitigation. To accomplish this, it is imperative to extract cyber threat-related information specific to an organization from the public internet, due to the growing number of information systems used in security operation

The passage discusses the challenges posed by cognitive cyber-physical systems of the Internet of Things (CPS-IoT) in the healthcare sector. These systems have increased cognitive complexity, making them vulnerable to traditional CPS-IoT threats and new threats related to their inherent cognitive functionalities. Moreover, the ubiquity of CPS-IoT increases the attack surface, making the public safety risks higher for critical infrastructure. Healthcare services and infrastructures are particularly vulnerable to major security risks, and the situation is exacerbated by the increasing interconnectedness of medical devices and services that are exposed to new cybersecurity vulnerabilities. Reports indicate a rise in cyber-attacks and medical identity theft globally. The current security solutions are unable to tackle the dynamicity, complexity, uncertainty, and high connectivity of CPS-IoT enabled healthcare services and critical infrastructures. The solution proposed is to develop innovative techniques for building cognitive cybersecurity for CPS-IoT enabled healthcare ecosystems using cognitive architecture and artificial intelligence to enhance automated intelligent cybersecurity decision-making mechanisms with expert-level ability. [3]

Detecting web attacks within URLs is indeed a challenging task, and there are several major issues that need to be addressed. Firstly, there is a need for an effective way to transform all kinds of URLs into representations, as different attacks can hide in various ways within their URLs. Secondly, different attacks exhibit different signatures in their URLs, which makes feature selection a challenging task. Thirdly, many deep learning models used for cyber security only have one model for detection, making it difficult to update the system when new attacks emerge. Finally, in the IoT cloud environment, the centralization feature can impact the application of distributed services, such as network security mechanisms for IoT applications. To address these challenges, novel security models, controls, and decisions must be distributed at the edge of the cloud to support the new IoT paradigm, known as the EoT. [4]

That is a plausible explanation for the modest number of entities of certain types in the annotated dataset. During an attack, it may not be immediately clear who or what is responsible, and it may take some time to gather the necessary information to identify the entities involved. Additionally, some attacks may be carried out by previously unknown entities or groups, making it difficult to label them in advance. As a result, the annotated dataset may not be representative of all types of attacks and entities involved, and further research and data collection may be necessary to improve the performance of automated detection systems. [5]

The process of separating relevant data from the massive influx of data is challenging. The amount of digital text information is fast expanding due to the expansion of social networks and pervasive computers. One of the main duties of a security center is to enable correlation. Identifying information connected to cyber threats will help an analyst develop situational awareness, enabling proactive monitoring and risk mitigation [12]. Due to the increasing number of information systems utilized in security operations, it is essential to extract cyber threat-related information relevant to a business from the public internet to achieve this.

Moreover, it is discussed how cognitive cyber-physical systems of the Internet of Things (CPS-IoT) provide problems for the healthcare industry. Due to the growing cognitive complexity of these systems, both conventional CPS-IoT vulnerabilities and new threats arising from their innate cognitive functions can affect them. Additionally, the widespread use of CPS-IoT broadens the attack surface, raising the dangers to public safety for critical infrastructure. The growing interconnection of medical devices and services, which exposes them to new cybersecurity vulnerabilities, makes healthcare services and infrastructures even more susceptible to significant security hazards. According to reports, medical identity theft and cyberattacks are on the rise everywhere. The dynamic nature, complexity, uncertainty, and high connectivity of CPS-IoT enabled healthcare services and critical infrastructure cannot be addressed by current security solutions [13].

The paper proposes a distributed system for web attack detection from URLs using deep learning techniques, including CNNs and NLP models. The system can represent all kinds of URLs, distinguish anomalous requests from normal ones, and apply multiple concurrent models to enhance system stability. Additionally, the paper proposes a generic distributed web attack detection system on edge devices of the cloud. The paper is organized into different sections, including a brief review of related works, the architecture and methodology of the proposed system, the datasets and experiment settings, experimental results and discussion, and the conclusion and future work [14].

C. Research Objectives

Natural language processing activities such as named entity recognition have shown to be successful when using BERT transformers. (NER). The authors of this study suggest employing BERT transformers to enhance NER performance in cases where the annotated dataset is small or lacks diversity. They also recommend an automatic dataset augmentation technique to add words with automatically tagged named items to the training dataset. The performance of various BERT models, including a multilingual model, a model adjusted for Russian data, and a model tailored for cybersecurity texts, is examined in this work. The authors also explore the effects of various parameters on the effectiveness of the NER system and present a new technique for dataset augmentation for NER tasks [15].

This paper aims to assess the effectiveness of LSTM-based neural network models in extracting information related to cybersecurity. The study compares the performance of an LSTM-based model and a CRF-based model for named entity recognition (NER) tasks, and evaluates three different LSTM architectures for relation extraction (RE) tasks. The models are trained on a corpus of vulnerability descriptions using a bootstrapping algorithm, and a word embeddings model is generated from the National Vulnerability Database (NVD). The models are evaluated based on their precision, recall, F1 scores, and training/testing accuracy [16].

The objective is to locate cyber intelligence in texts written in natural language that map to words that resemble those in boldface in the table. If the application can't distinguish between ”Windows” and ”operating system,” one might not be able to map the text ”Adversary collects information about Windows version from the victim's machine.” This is a challenging task because it is impossible to list all equivalent formulations of a certain cyber notion [17].

The goal of this study is to detect cybersecurity intelligence (CSI) from tweets. Cybersecurity refers to protecting electronic systems and data from malicious attacks and collecting the latest CSI is important to prevent or minimize damage from these attacks. Twitter is a key source of OSINT where experts discuss technical details and share experiences about cyber attacks. Positive CSI tweets contain crucial information such as attackers, vulnerabilities, and targets, which can help experts respond to threats effectively. The study classifies tweets into positive and negative classes based on the relevance of CSI keywords, such as Common Vulnerabilities and Exposures (CVE) [18].

III. Methodology

In the modern digital era, cyberbullying is becoming a bigger issue, especially on social media platforms like Facebook. It is a form of abuse, pressure, or harassment aimed at those who use digital technology. It could be challenging to identify cyberbullying on Facebook due to the enormous volume of text data that users produce every day. The use of natural language processing (NLP) methods to automatically identify cyberbullying on social media sites has been studied. In this project, we create a Facebook cyberbullying detection system using NLP approaches.

A. Data Collection

A dataset of Facebook posts and comments is compiled as part of the initial phase of this investigation. We will compile the information using the Facebook API, which enables us to extract publicly accessible posts and comments about cyberbullying. We will filter the posts and comments based on specified hashtags and terms related to cyberbullying. The dataset will then undergo preprocessing to ensure that all of the data is text-based and to remove any irrelevant information.

B. Preprocessing

The dataset must then be prepared for analysis. Emojis, images, links, and any other non-text data from the dataset will be removed. After the text has been tokenized, or broken down into individual words, the data will be normalized using stemming and lemmatization techniques. Stopwords, or common words such as ”the” and ”and,” that do not advance the topic of the text, will be removed. To keep everything consistent, the remaining text will be changed to lowercase.

C. Extraction

In the third stage, relevant attributes are extracted from the preprocessed dataset. The text data will be converted into numerical features using NLP techniques such as bag-of-words, which represents the text as a collection of words, term frequency-inverse document frequency, which gauges the significance of each word in a document, and word embeddings, which represents words as vectors. The cyberbullying detection model will be trained using these features.

D. Model Development

The creation of a model for identifying cyberbullying is the fourth phase. The model will be developed using a range of machine learning approaches, including logistic regression, decision trees, and support vector machines. We will also look into the use of deep learning techniques like transformers and recurrent neural networks. The features of the preprocessed dataset will be used to train the models.

E. Model Evaluation

The effectiveness of the developed cyberbullying detection model is assessed in the final stage. Accuracy, precision, recall, and F1-score are a few of the measures that will be used to gauge the model’s efficacy. To ensure that the model generalizes effectively to fresh data, it will be assessed on a separate dataset. The evaluation will assist us in determining how effectively the model detects instances of cyberbullying on Facebook.

IV. Model Training

A. Data Processing

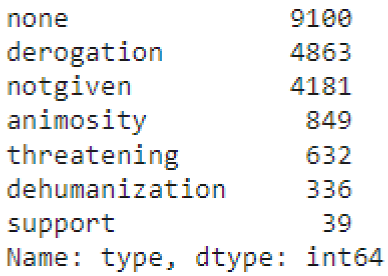
We first obtained the necessary data from a variety of sources, including kaggle, github, and other internet sites, due to a lack of data.The dataset was then mixed with some personally gathered data and modified before being ready for pre-processing. Next, we looked to see if any fields were null as well as the overall count. Later, we eliminated redundant information and combined all the words into a single string. We receive the finalized data for the primary work. Then we go through some Natural Language Processing techniques and focused on training the data.

B. Equation and Calculation

True Positive: Situation in which you accurately foresee a favorable outcome.

False Positive: When you expect a great outcome yet it turns out to be negative.

False Negative: Suppose you forecast a negative outcome, but it turns out to be a positive one.



```
none             9100
derogation       4863
notgiven         4181
animosity         849
threatening       632
dehumanization    336
support            39
Name: type, dtype: int64
```

Fig. 1: Data frequency

True Negative: Scenario where you accurately foresee a disastrous outcome

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)

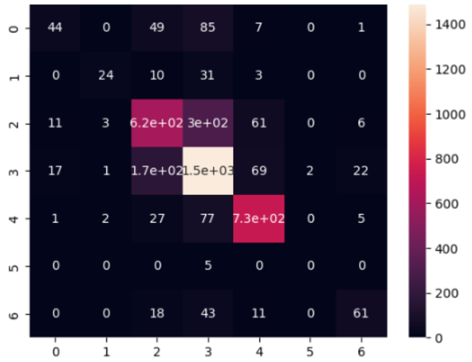
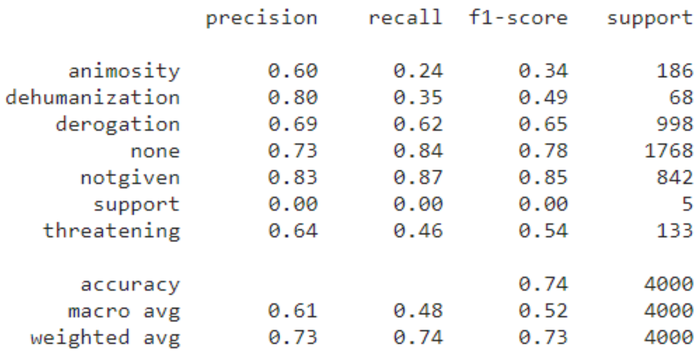


Fig. 2: Confusion Matrix

C. Classification Result

Here we train our dataset with the Logistic regression model.The performance of this model is a little satisfactory but we want to train some other model in our future work to improve the accuracy.The anayzing report of Logistic regression model is given below:



```
                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 work, we suggested a system design that might automatically extract data about cyber threats to help human operators. Using the neural embedding technique doc2vec, we conducted tests to see if the Natural Language Filter portion of the system could be implemented in practice.Our analysis shows that a doc2vec-based natural language model trained with cybersecurity-specific text data and unique preprocessing might be employed as a Natural Language Filter for the proposed autonomous system with 74% accuracy.We will further work on it with different techniques so that we can get maximum accuracy among the existing system.

Acknowledgment

Praise be to the Great Allah, who's assistance enabled us to successfully complete the composition of our research. Then, we would want to thank our loved ones for their encouragement and their enlightening remarks on the study. Finally, we would like to express our gratitude to Dr. Md. Ashraful Alam, who served as our supervisor and assisted me with this study. Last but not least, we would like to thank Shakib Mahmud Dipto sir for his assistance with the project.

References

[1] Journal of Network and Computer Applications — ScienceDirect.com by Elsevier. (n.d.).
https://www.elsevier.com/locate/jnca

[2] Mendsaikhan, O., Hasegawa, H., Yamaguchi, Y., & Shimada, H. (2019). Identification of Cybersecurity Specific Content Using the Doc2Vec Language Model. Computer Software and Applications Conference.
https://doi.org/10.1109/compsac.2019.00064

[3] Abie, H. (2019). Cognitive Cybersecurity for CPS-IoT Enabled Healthcare Ecosystems. International Symposium on Medical Information and Communication Technology.
https://doi.org/10.1109/ismict.2019.8743670

[4] Tian, Z., Luo, C., Zhao, J., Du, X., & Guizani, M. (2020). A Distributed Deep Learning System for Web Attack Detection on Edge Devices. IEEE Transactions on Industrial Informatics, 16(3), 1963–1971.
https://doi.org/10.1109/tii.2019.2938778

[5] Tikhomirov, M. N., Loukachevitch, N. V., Sirotina, A., & Dobrov, B. V. (2020). Using BERT and Augmentation in Named Entity Recognition for Cybersecurity Domain. Lecture Notes in Computer Science, 16–24.
https://doi.org/10.1007/978-3-030-51310-8_2

[6] Yin, D. (2009, January). Detection of harassment on web 2 - researchgate. Retrieved April 6, 2023
https://www.researchgate.net/profile/Brian-Davison/publication/
228978102_Detection_of_harassment_on_Web_20/links/
00b4951e732d50743a000000/Detection-of-harassment-on-Web-20.pdf

[7] Wang, W., Royen, K. V., Tokunaga, R. S., Salmivalli, C., Räbiger, S., Liu, Y., Li, Q., Kavanaugh, A. L., Hosseini, M., González-Bailón, S., Fawcett, T., Connolly, I., Calvete, E., Bollen, J., Bellmore, A., Bauman, S., Balakrishnan, V., Adali, S., Aggarwal, A., . . . Kohavi, R. (2016, May 31). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior. Retrieved April 6, 2023, // https://www.sciencedirect.com/science/article/abs/pii/S0747563216303788

[8] Dinakar, K., Reichart, R., Lieberman, H. (2021, August 3). Modeling the detection of textual cyberbullying. Proceedings of the International AAAI Conference on Web and Social Media.// https://ojs.aaai.org/index.php/ICWSM/article/view/14209

[9] University, R. Z. N. T., Zhao, R., University, N. T., University, A. Z. N. T., Zhou, A., University, K. M. N. T., Mao, K., Metrics, O. M. V. A. (2016, January 1). Automatic detection of cyberbullying on social networks based on bullying features.
https://dl.acm.org/doi/10.1145/2833312.2849567

[10] Ahmed, M. F., Mahmud, Z., Biash, Z. T., Ryen, A. A. N., Hossain, A., Ashraf, F. B. (2021, June 8). Cyberbullying detection using deep neural networks from social media comments in the Bangla language.
https://arxiv.org/abs/2106.04506

[11] Iwendi, C., Srivastava, G., Khan, S., Maddikunta, P. K. R. (2020, October 13). Cyberbullying detection solutions based on Deep Learning Architectures - multimedia systems.
https://link.springer.com/article/10.1007/s00530-020-00701-5

[12] Mendsaikhan, O., Hasegawa, H., Yamaguchi, Y., Shimada, H. (2019). Identification of Cybersecurity Specific Content Using the Doc2Vec Language Model. Computer Software and Applications Conference.
https://doi.org/10.1109/compsac.2019.00064

[13] Abie, H. (2019). Cognitive Cybersecurity for CPS-IoT Enabled Healthcare Ecosystems. International Symposium on Medical Information and Communication Technology.
https://doi.org/10.1109/ismict.2019.8743670

[14] Tian, Z., Luo, C., Zhao, J., Du, X., Guizani, M. (2020). A Distributed Deep Learning System for Web Attack Detection on Edge Devices. IEEE Transactions on Industrial Informatics, 16(3), 1963–1971.
https://doi.org/10.1109/tii.2019.2938778

[15] Tikhomirov, M. N., Loukachevitch, N. V., Sirotina, A., Dobrov, B. V. (2020). Using BERT and Augmentation in Named Entity Recognition for Cybersecurity Domain. Lecture Notes in Computer Science, 16–24.
https://doi.org/10.1007/978-3-030-51310-8_2

[16] Gasmi, H., Laval, J., Bouras, A. (2019). Information Extraction of Cybersecurity Concepts: An LSTM Approach. Applied Sciences, 9(19), 3945.
https://doi.org/10.3390/app9193945

[17] Purba, M. D., Chu, B., Al-Shaer, E. (2020). From Word Embedding to Cyber-Phrase Embedding: Comparison of Processing Cybersecurity Texts. Intelligence and Security Informatics.
https://doi.org/10.1109/isi49825.2020.9280541

[18] Shin, H., Kwon, H., Ryu, S. (2020). A New Text Classification Model Based on Contrastive Word Embedding for Detecting Cybersecurity Intelligence in Twitter. Electronics, 9(9), 1527.
https://doi.org/10.3390/electronics9091527