Design of a Deep Learning Model for Cyberbullying and Cyberstalking Attack Mitigation via Online Social Media Analysis

Sandip A. Kahate
Research Scholar,
P. G. Computer Science Department,
SGBA University, Amravati,
Amravati, India-444602
sandip.kahate@gmail.com

Dr. Atul D. Raut,
Research Guide
Computer Department,
P. R. Pote COE and Management
Amravati, India-444601
atuldraut@gmail.com

***Abstract*—Identification of cyberbullying and cyberstalking for real-time use cases is a multi domain task that involves the design of social media data extraction, sentiment analysis, sentiment pattern evaluation, and regression models. To perform this task, researchers have proposed the use of high-density feature representation models that can extract social media sentiments, and combine them with user specific parameters like age, gender, time of post, etc. But existing models are either non-comprehensive or capable of achieving limited accuracy when used for real-time scenarios. Moreover, these models are not flexible to multimodal inputs, which further limits their scalability levels.**

To address these concerns, this paper proposes the development of a deep learning model for cyberbullying and cyberstalking attack mitigation via social media analysis. The proposed model initially collects tweets posted by users, extracts meta data, and analyses language features for training a Long-Short-Term Memory (LSTM) based Convolutional Neural Network (CNN), which assists in the pre-filtering of tweets. The filtered tweets are passed through a Natural Language Processing (NLP) engine that assists in sentiment identification for these texts. Sentiment data and Word Embedding capabilities are used to anticipate cyberbullying and cyberstalking attacks. This is done via CNN based pattern analysis, which assists in the efficient identification and mitigation of these attacks. Due to the integration of these models, the proposed method is able to improve attack detection accuracy by 3.5 %, while reducing the identification delay by 8.5 % in real-time scenarios.

***Index Terms*—Cyber-attacks, LSTM, CNN, Bullying, Social Media Analysis**

I. Introduction

Victims of harassment and stalking often face threats to their physical safety, false accusations, breaches of privacy, and even slander. Many online social networks designed for multimedia interaction have had their already sizable attack surfaces further enlarged as a result of advancements in information and communication technology (ICT). Text, images, sound, and moving graphics are all utilized to communicate with users in the context of Human Computer Interaction (HCI) techniques. Most often, this is accomplished with the use of various internet-accessible software programs and mobile application platforms. The research shows that ICT has a considerable effect on the intensity of cyberstalking. The Electronic Communication Harassment Observation (ECHO) project, for instance, has conducted research [1], [2], [3], [4] showing that many occurrences that began in cyberspace have migrated to the physical world. In extreme cases, victims have been forced to quit their usual routines, relocate, and/or change jobs, all of which have resulted in significant financial losses, instilled fear and concern in the victims, and disrupted the victims' daily lives. Because of this, terms like "cyberstalking" and "cyberbullying" have emerged to correctly define the environment in which victims and perpetrators operate [14] and to explain the problem more generally. Suicide and homicide have been linked to the mental anguish and physical injury caused by antisocial behaviours [5], [6]. Defining cyberstalking and its distinguishing characteristics has to be discussed at length. Cyberstalking is defined in this research as the following types of online interactions: In order to meet the definition of a crime, an act must meet four criteria: it must be unwanted or unwelcome; it must be transmitted by a known or unknown entity (perpetrator) that is determined and motivated; it must be conveyed with the intention of targeting one person (the victim); and it must be persistent. According to the UK's National Centre for Cyberstalking Research (NCCR),sending 10 or more unwanted messages over a time period equal to or less than four weeks is a third sign of persistent conduct. In this discussion, cyberstalking is distinguished from other forms of online harassment. Effective management of cyberstalking risks requires the use of technology for detection, event classification, automated responses, and incident reporting. Text (emails, SMS, IM, blog posts, Twitter tweets, etc.) seems to be the most widely used content type, highlighting the importance of text analysis and information retrieval (IR).

One of the first studies to focus on cyberbullying was [7], [8], [9], [10], which led to the creation of a system that makes use of Twitter's streaming API to collect and classify tweets on the topic, and Long Short-Term Memory (LSTM) based Gated Recurrent Neural Network (GRU) for efficient analysis

978-1-6654-9414-4/23/$31.00 ©2023 IEEE