Automated and Semi automated Architecture for Gathering Intelligence Using Twitter API

Govind kanav Singh
Department of Computer Science and I.T.
Central University of Jammu
Jammu, India
gksinghofficial@gmail.com

Dr.Bhavna Arora
Department of Computer Science and I.T.
Central University of Jammu
Jammu, India
bhavna.csit@cujammu.ac.in

***Abstract*—Intelligence plays a vital role in defending against rising security threats in our real and virtual worlds, Both can be protected easily only by exploiting sources available in public. In our research, we have found that cyber threat intelligence is useful against the increased risk of protecting cyber and real-world entities, one of the sources we have found best in the evaluation process of cyber threat intelligence sources is open-source intelligence. Our work will present the importance of cyber threat intelligence and the way its different sources play an important role in defending assets against cyber-attacks and maintenance of law and order in the present day scenario. We have developed and discussed different algorithms and techniques for intelligence gathering, and then we have previously implemented them in the evaluation process of best sources. This paper provides an in-depth review of varied cyber threat intelligence sources, prototypes, and evaluation among the best existing techniques. Open-source intelligence and social media intelligence could be a method of collecting data from openly available sources and will be discussed in detail with their relevant techniques available for gathering data. Here we will discuss different roles of social media intelligence and open source intelligence in global health security management, civil-military partnership in dealing with security, and role intelligence gathering in 5th generation warfare. Correlation between social media intelligence and open-source intelligence or different prototypes supported by these sources we have also discussed the difficulties and challenges faced in gathering cyber threat intelligence. Next, we have discussed algorithms using automatic and semi-automatic algorithms gathering information for analysis and the role of Twitter and its API in successfully gathering data in a legalized manner. At last, we will conclude with the evaluation of parameters we have specifically designed for the existing techniques and our prototype to find out how prototypes and their different stages perform.**

Keywords— Cyber threat intelligence, Twitter API, Open source intelligence, Social media intelligence, Security management, Civil-Military partnership, Security management, Automation, Semi automation.

I. INTRODUCTION

We are living in a world where a Somalia based terrorist group, AL-Shabab affiliate of AL-Quida having social media as a part of their modus operandi even, now, pirates operating in the Gulf of Aden also use social media, these types of groups operate, communicate, plans, and cooperates by using different internet platforms like Facebook, blogs, Twitter, and Instagram. Disturbing balance of civil society becomes easy with the help of fake news and attack bots, hitmen are being hired for target killing across the globe with the help of Facebook, all these things increased the problem for civil administration for maintaining law and order. This severe impact of crimes on the cyber world has left no choice for the governments and cyber security agencies to manage these crimes all across the globe by gathering intelligence for defending against virtual and real-world threats. CTI in what threat information becomes once it has been collected and evaluated for intelligence generation in the context of its reliability, sources, and analyzed through rigorous and structured tradecraft techniques by those with substantive expertise information which reduces uncertainty for the consumers. Cyber threat intelligence (CTI) gathering is a process that starts from planning to feedback process begins with the gathering of unstructured/raw data from CTI sources to transform it into intelligence is called the cyber threat intelligence cycle. CTI cycle is a process divided into five stages process starts with the planning of threat intelligence gathering.

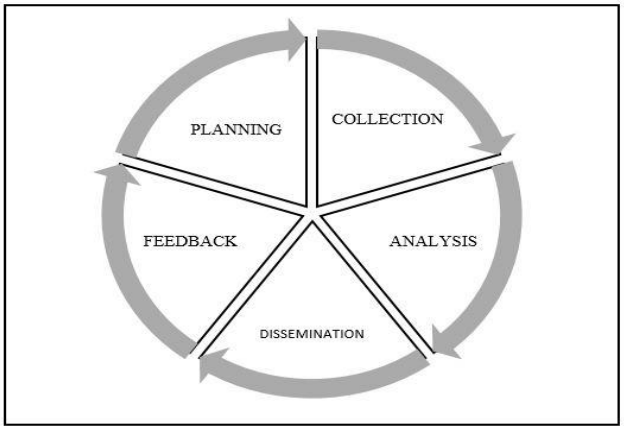


Fig.1 Cyber Threat Intelligence Cycle

This initial phase of the planning cycle for the execution of the process takes place that how data can be gathered and will be processed during the different stages cycle, the whole process is then analyzed in the final stage of feedback which we will discuss in the 5th stage of cyber threat intelligence cycle. The next stage is about data collection, data as we know a new oil of our virtual world. Data can be collected from the different sources of Intel gathering platforms in the raw form, and then it will be analyzed in the next stage. The collection of data plays an important role in the cyber threat intelligence cycle as well

as in our experiment. Collected data is in the raw form and unstructured, then processed in different stages of the process, and can be utilized after the filtration through different NLPs, and text pre-processing methods. Data after collection then moved to the next stage of analysis, data then visualized in a different state for the analysis. For the analysis of data, different graphs and tables can be used with the help of visualization techniques. Visualization of data can be useful in further accessing how data will be used for extracting vital information. Dissemination takes place when data analysis is over, and the dissemination is the last actionable stage in the CTI cycle processed. Intelligence through different stages collected and analyzed the execution of intelligence. The finished information in the form of intelligence then will be preceded to the stations where it is most needed. The final step in this process is feedback, feedback though not an actionable stage but an important stage because it plays a vital role in assessing the whole process. Feedback helps in finding the flaws in the process, measuring the quality of different evaluation parameters, the performance quality of the prototype how helpful it is in detecting cyber threat intelligence.

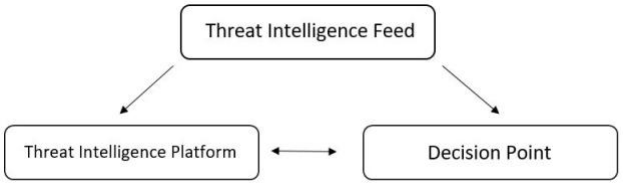


Fig2. Intelligence Feed Process

Threat intelligence feed, Threat intelligence platform, and decision point that may be delivered by vendors, products can be multiple or single depending on customer premises or in the cloud. Threat intelligence feed is the source point of the information, and the decision point is the destination of the information. Threat intelligence platform can act as an exchange point of the information between Threat intelligence feed and decision point.

Gathering intelligence can be tricky in deciding the source of extracting information that can be later utilized for averting harmful attacks on the cyber domain. There are sources like human intelligence, imagery intelligence, signal intelligence, measurements, and signature intelligence, and technical intelligence these are the effective source for gathering intelligence but not that successful for our cyber world. There is another source that we have here is a cyber intelligence source for gathering threat information and restricting threat actors from harming cyber assets. Cyber intelligence is further focusing on its sources are social media intelligence (SOCMINT), deep and dark web intelligence (DADWINT), and open-source intelligence (OSINT), Now we will focus on these CTI sources and discuss how these sources are useful in our experiment and which one will be best for gathering intelligence and most reliable one.

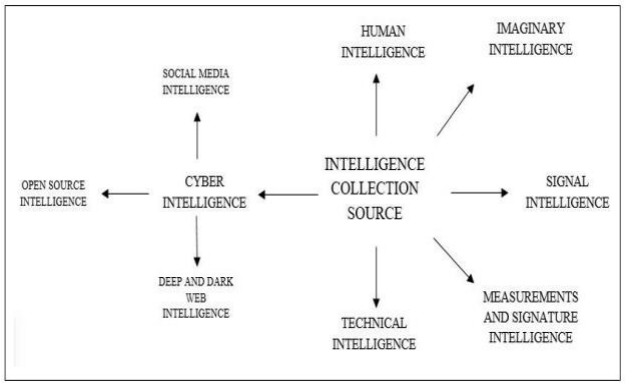


Fig3. Intelligence Collection sources

In this century, social media is an integral part of a human's life because of internet evolution. Social media platforms have billions of users in 2021, Facebook has 2.89 billion users, Instagram has around 1.074 billion, and Twitter has about 369.5 million users, but social media is not restricted to only these three platforms and holds a huge chunk of data in unstructured form. Social media can play a vital role in providing information and helpful in gathering actionable intelligence, not only intelligence for securing the world from threat actors but also helps in infectious disease outbreaks just like we are in a pandemic era these kinds of infectious outbreaks pose a disastrous effect on the international community. Social media monitoring helps in reducing its effect on humanity and how SOS messages can help the ones who need emergency assistance, and for more details, we have conducted in-depth research available on cyber threat intelligence "comparative analysis of its sources and parameters of evaluation". Now we will move to discuss the dark world of our virtual universe.

The deep and dark web has its effects on our lives and is a cause of concern for law enforcement agencies. Cyber agencies always try to keep their surveillance on high alert, all because of illegal activities from celling of illegally gathered banking data, personal, illegal auctions, human trafficking, smuggling of drugs, recruitment of terrorists, to the sales of radioactive substances and weapons. This whole dark world contains a large amount of raw data full of intelligence but, this type of data can only be extracted by cyber security professionals because it is quite dangerous for a researcher or any other individual to get data. The deep and dark web operators hide their users and their cyber footprints to protect them from security surveillance operators secure users with the high-end security encryption system to dodge new ways used by security agencies to penetrate deep web security encryption.

Open-source intelligence (OSINT) refers to the intelligence available in the public domain OSINT has a vast area for gathering intelligence like reports, articles, media, terrorist deliverables, and the internet. These public forums have further bifurcated into various streams, so the terrorist deliverables have leaflets, booklets, video, and audio, then media have news OPs and interviews, then articles have academic research and journals, now next we have reports it contains NGO, governments and international agencies,

in the last we have internet contains social media, websites, and forums.

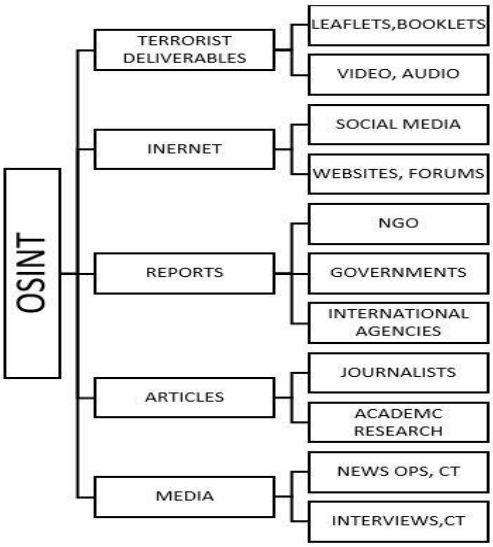


Fig4. Open Source Intelligence

Social media can alone on its own for providing information but, there is a catch when we try to extract data from platforms like Facebook and Instagram because they have a privacy policy for their users, and these accounts are highly secure by encryptions trying to extract data from these accounts will put you behind bars for violating privacy laws. Twitter is openly available without an encrypted system for everyone and is one of the most reliable sources of data for extracting information. In this paper, we will be discussing the extraction of data from Twitter by using Twitter API 'tweepy' and in-depth analysis of information with automatic and semi-automatic prototypes.

Many of us heard about modern-day warfare, and maybe some of us studied 5th generation warfare, the war, which will not be fought with weapons of mass destruction or, we can say nonconventional warfare. In some reports, researchers have mentioned about the threats to us from china are not their hypersonic missiles but, the real threat is new tactics of cyber-warfare. This next-generation warfare mainly focuses on the OSINT domain where media, asymmetric warfare, and cyber-warfare.

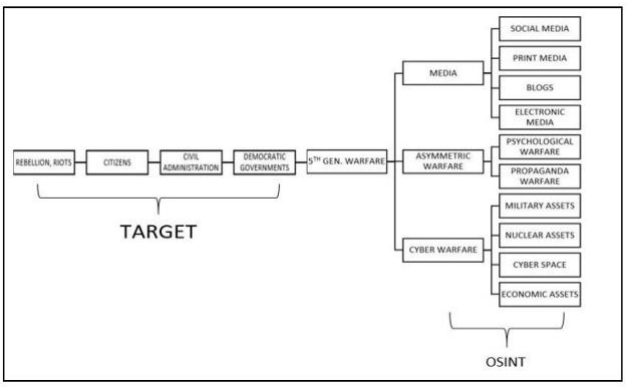


Fig5. 5th generation warfare and Open Source Intelligence

When these will be combined, influence social media, print media, blogs, electronic media, military assets, nuclear assets, cyberspace, and economic assets and can start psychological and propaganda warfare against the civilian population. These types of attacks are commonly faced by democratic nations in recent events, we have seen some of the initial experimental phases of modern-day warfare, from riots on capitol hill, Washington dc to Delhi riots, what we have common is that attackers have openly posted on social media before these incidents. 5th generation warfare directly hit democratic governments, their civil administration, citizens, and this will result in rebellion movements and riots. These types of incidents and attacks can be prevented by getting timely and accurate intelligence.

For developing a mechanism to protect and secure the cyber world, our little step is towards beginning a process of developing prototypes for guarding not only assets but also civilians. Here we have first worked on automated, and then semi-automated prototypes automation from the word we can understand it can operate freely without any human intervention now same goes with the semi-automation from the word we can understand it is partially automated and have human interference in it. We will see further in the implementation section how these prototypes will perform.

II. Literature Survey

A. Cyber Space and Intelligence Sharing

We have conducted in-depth research in different formats of cyber threat intelligence before conducting the experiment we have studied different sources and techniques in which CTI plays a major role in various situations like providing cyber intelligence in a life-threatening scenario, detection, and precautions against threats. Cyber threat detection and intelligence sharing are used to secure assets and organizations by the means of different prototypes and tools. There is a standard setup for the sharing procedure of threat detection in the cyber industry such standards are STIX, TAXII, and CyboX, data coexisted and supported by the both are same which we required for use cases of operations in cyber security and data supported by different formats and languages. The adoption and implementation of sharing standards performed poorly as per the author's analysis and suggestions. IDEA is the best format that has been practiced in cyber security cases because of its capabilities for qualifying cyber threat intelligence needs for reducing CTI distribution and standard issues can only be sorted out by introducing time stamp and origin in data feed this will be helping in completing the CTI data. There is a scope of improvement in research by improving quality access to the data feed because CTI has a vast variety of intelligence gathering [1]. The emerging defense concept of real-time actionable threat intelligence is focused on the suppression and detection of cyber threat intelligence. Information sharing between different organizations reduces efforts because one organization's efforts of detection is another's prevention. Threat intelligence sharing between different sectors, the cyber security industry has a significant cooperation attribute.

TABLE 1. LITERATURE SURVEY

Author	Title	Year	Source	Technique
Robert Baumgartner, Michal Ceresna, Gerald Ledermüller	Deep Web navigation in Web data extraction	2005	OSINT, DADWINT	Based on Lixto technology
Michael Glassman, Min Ju Kang	Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT)	2012	OSINT	
David Omand, Jamie Bartlett, Carl Miller	Introducing social media intelligence (SOCMINT)	2012	SOCMINT	
Hoepman, Jaap Henk	Privacy design strategies	2014		
Swati Agarwal, Ashish Sureka	Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats	2015	SOCMINT	Mining User Generated Content
Ahmed T.Zulkarnine, Richard Frank, Bryan Monk, Julianna Mitchell, Garth Davies	Surfacing collaborated networks in dark web to find illicit and criminal content	2016	DADWINT	Modified web crawler for Tor “Dark Crawler”
Xiaojing Liao, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, Raheem Beyah	Acing the IOC Game	2016		Indicators of Compromise (IOC)
Vasileios Mavroeidis, Siri Bromander	Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence	2017		Cyber Threat Intelligence (CTI) model
Elias Bou-Harb, Walter Lucia, Nicola Forti, Sean Weerakkody, Nasir Ghani, Bruno Sinopoli	Cyber meets control: A novel federated approach for resilient cps leveraging real cyber threat intelligence	2017		Cyber-physical systems (CPS)
Isuf Deliu, Carl Leichter, Katrin Franke	Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks	2017	OSINT	Convolutional Neural Network methods against more Traditional Machine Learning approaches
Paulo Shakarian	Dark-web cyber threat intelligence: From data to intelligence to prediction	2018	DADWINT	Predicting Cyber-Events by Leveraging Hacker Sentiment
Ghaith Husari, Xi Niu, Bill Chu, Ehab Al-Shaer	Using entropy and mutual information to extract threat actions from cyber threat intelligence	2018	OSINT	Action Miner
Satyanarayan Raju Vadapalli, George Hsieh, Kevin S. Nauer	TwitterOSINT	2018	OSINT	Twitter OSINT
Aviram Zrahia	Threat intelligence sharing between cybersecurity vendors: Network, dyadic, and agent views	2018		
Summer Lightfoot, Filip Pospisil,	Surveillance and privacy on the deep web	2018	DADWINT	web crawling browsers
Rose Bernard, G.Bowsher, C.Milner, P.Boyle, P.Patel, R.Sullivan	Intelligence and global health: assessing the role of open source and social media intelligence analysis in infectious disease outbreaks	2018	OSINT, SIGINT	Clandestine intelligence sector
Kira Vrist Rønn, Sille Obelitz Søe	Is social media intelligence private? Privacy in public and the nature of social media intelligence	2019	SOCMINT	
Ba Dung Le, Guanhua Wang, Mehwish Nasim, M.Ali Babar	Gathering cyber threat intelligence from twitter using novelty classification	2019	OSINT	Novelty Detection Model
Andrew Ramsdale, Stavros Shiaeles, Nicholas Kolokotronis	A comparative analysis of cyber-threat intelligence sources, formats and languages	2020		
Daniel Schlette, Fabian Böhm, Marco Caselli, Günther Pernul	Measuring and visualizing cyber threat intelligence quality	2021		Based on DQ and STIX data format

The cyber security industry has a small-world structure associated with communities that are suitable for cyber threat intelligence sharing. Industries collaboration can be featured by competition between loosely integrated complementary solutions. Cyber threat intelligence sharing relations are associated with the innovation level of a firm. There is a difference between publicly and privately held companies, the size of the effectiveness is three times in publicly held companies compared to the privately held companies. The author's analysis of datasets of CTI sharing environment between online protection sellers and recognizing its connected highlights, and experiences, imagining center points, network organizing structure, and networks. The analysis was done by the author based in the industry covering CTI sharing vendors [2].

The important aspect of practical applications is generating collaboration and exchanging cyber threat intelligence employing a shared platform. Only updated and accurate high-quality CTI can help to detect and defend against cyber-attacks because outdated, incomplete, and inaccurate information can pose a big threat to clients or the organization. As we know the cyber footprints are increasing and it is directly proportional to the amount of CTI and its availability but it doesn't guarantee the quality of information. So, with the increasing number of CTI, the stakeholders need to keep an eye on the quality of the CTI. The analysis is important and focused on improving cyber threat intelligence quality and making all the stakeholders aware of the facts of outdated and inaccurate information. The Platform Benefits Sharing of CTI helps broaden our knowledge of current cyber threats. The proposed metrics and dimensions help build a cyber threat intelligence bracing quality management method based on a structured threat intelligence representation data format. Factors that help customers understand how DQ measurement results were achieved were a visual indication of the object's overall quality, including scores for each of the individual quality dimensions. To finally build a bracing methodology for CTI quality assessment Implementing and extending the dimensions and matrices are necessary steps but also improving and assuring the quality of information on a sharing platform [3].

Tight integration among network-based totally, physical strategies, sensors, and software program incarnated by cyber-physical structures. Cyber technology when incorporated into the legacy structures will maximum certainly introduce improvements and possibilities now not but expected, however, it'll surely clean the way for misdeeds to make the most cyber belongings, sources, systems and will motive intense and drastic national damage. All works inside the literature nearly exclusively followed the safety of 1 impartial component of the cyber-physical device (i.e. physical or cyber), in this paper the author argued that these systems can't be decoupled. In this context, the author affords what he believes is a primary try to ever address the problems of cyber-bodily protection in a scientific and coupled manner. They proposed to derive tangible cyber-physical system attack models from empirical measurements as the core rationale behind the architecture, which can be deputed to attribute and infer real cyber-physical system attacks. Here the author talks about the innovation that CP threat detector combines attack signature from the cyber realm with the CP data flows from the physical realm [4].

B. Dark Web of Intelligence

The number of internet users is increasing at a rapid pace because of the digitalization process across the globe, and this situation has come up with the flaws of exposing vulnerable assets and systems to threat actors. This evolution of the internet has been done by different means like socializing, business, learning, and organizing, cyber security and anonymity is vital to individuals, I.T. sectors, state and non-state actors. Web crawlers are not the usual browsers that we commonly used for web surfing. Web crawling is mainly used for validating hypertext mark-up language code and auto maintenance of tasks on the web known for undercover illegal activities and cannot be accessed by the regular web browsers commonly known as deep web. The deep web is a data treasure source because it can only be accused by special browsers, and the dark web contains a huge chunk of data from secretly operative organizations, criminals, individuals, and smugglers to the terrorist outfits. The dark web has created a web wall to secure its user from legal scrutiny with the help of highly encrypted software. A large amount of unstructured data is held by the dark web, and extracting it from there is quite an impossible task because of its huge size and lack of modern infrastructure, big data is present there in the rarest form, untapped, and not possible for the analyst to analyse [5].

The way we have decided to discuss and implement two types of approaches towards the implementation of prototypes based on automatic and semi-automatic, here the author have discussed the same in their literature and the extraction of data is just one-half of the game. The data extraction from web forums is quite a task for the security agencies and organizations responsible for the management of assets have to go through various obstacles like non-

hypertext mark-up language formats, dynamic changes on websites, java scripts, session ids, and web forum iteration. In the experiment, the author has embedded a Mozilla web browser into the visual wrapper and has used an open source intelligence API for data extraction and two application domains where Dark Web navigation capabilities play a critical role, which is automotive B2B cyber forum and Business Intelligence scenarios [6].

Behind the white screen of the web a whole black universe dealing in child sex abuse, drug marketing, human trafficking, and animal and artifacts illegal auction market. Conventional methods in law enforcement and investigation are now unsuccessful and time-consuming in the investigation process. The author has explored various aspects of the dark web mentioned by renowned researchers. To create a Tor's network map author previously develop a web crawler based on keywords provided for the automated searching on the web, this crawler has successfully performed previous operations in searching and extracting various proofs on child exploitation linkages in cyberspace. The author has discussed in detail the working on the dark side of the internet and has crucial findings on how the terrorist and their support arms interconnected working [7].

C. Social Media World of Intelligence

Social media platforms and hacker forums contain a huge amount of information related to cyber threat intelligence. Manual analysis for the extraction of cyber threats from sources like social media and hacker forums is time-consuming and requires the allocation of resources due to its error-prone nature. The use of records in textual content from a hacker discussion board, the writer in comparison the text category performance of Convolutional Neural community methods against greater conventional ML-based techniques. The authors have discovered that conventional ML techniques, along with SVM can produce an overall performance of excessive degrees that is on par with Convolutional Neural community algorithms. The authors show the subservience of supervised ML algorithms for classifying hacker forum posts. In assessment to hacker forum publish-type demonstration, the authors have determined that an aid vector mechanism produced effects that have been on par with extra present-day Convolutional Neural Networks. With the computational complexity of CNN architectures, this paper's effects suggest that the guide vector mechanism is advanced for the functions of real-time, realistic Cyber hazard Intelligence packages [8].

Cyber intelligence covers a large amount of variety of sources used for gathering intelligence are social media intelligence (SOCMINT), deep and dark web intelligence (DADWINT), and open-source intelligence (OSINT). There is a need and demand raised from time to time in different parts of the globe for the inclusion of social media intelligence in national security frameworks for helping cyber security agencies in securing rapidly growing internet users. There is a setup of two crucial tests before the inclusion of SOCMINT, firstly it should be based on methodological evidence, verification, collection, and understanding. The second one demands the privacy of users should be secured with the help of constitutional law and legal means. A structured process has been introduced by the author on how to carry out the inclusion of threat intelligence in the framework. Social media intelligence is capable of becoming a branch of law enforcement structure for gathering internal and external intelligence information's validity is a major problem that creates rucksack between intelligence agencies and users, and can be solved by making netizens aware of the day to day scenario so they can understand why, when and what type of restrictions have been applied under SOCMINT are undertaken those government who are willing to implement such framework inclusion must uphold democratic values like human rights, privacy, and accountability [9].

A simple prototype can save millions in this situation only by tracking SOS messages broadcasted on social media platforms and microblogging applications. This pandemic era has taught us a lot, but one of the things that can be broadly discussed and should be included in the framework of handling natural disaster situations with the civil-military partnership. The civil-military partnership is a time-tested mechanism that can play a vital role in gathering intelligence from open and social media intelligence because these two sources can play a crucial role in emerging situations whether we are tackling an infectious disease outbreak or some man made illegal actions. Threats are posing from different directions in the cyber world converting it into a warzone because of the increasing use of social media platforms. There is a need for a threat defense mechanism against increasing information warfare, psychological warfare, and propaganda warfare [10].

D. Open Source Intelligence and Twitter

There is an uninterrupted boom within the growth of brand new cyber-assaults makes cyber risk intelligence sharing important for imparting advanced cyber risk observation and enabling in time reaction to cyber-assaults. The writer's goal in

this paper is to broaden an automated machine to extract cyber risk intelligence from publicly available cyber hazard intelligence assets (OSINT) for taking timely motion towards cyber-assaults. Successfully, especially, and innovatively used the metrics modern-day entropy and mutual records, from facts theory so that text to be had within the cybersecurity area may be analyzed. Combining it with primary natural language processing techniques, the writer's framework called action miner that framework has achieved higher accuracy and remember than the state of the art Stanford typed dependency parser, no matter being a state of the art library it simplest works with trendy English but not with cyber specific domain texts. The authors after the use of action miner are capable of conducting analytic research series on mutual records and cybersecurity word entropies and the latest cyber danger movements in a particular period, to confer pernicious in-depth knowledge about the cyber risk landscape state-of-the-art a selected period and the common malicious behaviour present-day threat actors. The authors had been presenting an automated cyber chance intelligence extraction for hazard movement expression and extracting records as the important thing statistics trendy a report for in addition evaluation via machine [11].

The intelligence sector in cyber security is evolving at a high pace, professional's use mediums like Twitter, blogs, forums, etc. for exchanging IOC (indicators of compromise) because all these mediums are publically available IOC are botnets IPs, malware signatures are available on open source and their information published in the white paper, article, posts, etc. all this information can be converted in to format easily understood by the machine. The mechanism for automated analysis of open IOC (indicators of compromise) their rapid development to the intrusion detection system, the fact that making management difficult for professionals is data generated by the multiple sources in cyber domain at a high pace and large amount. The limitations in NLP (natural language processing) have obstructed the efforts of generating intelligence from raw data automatically. The IOC demands higher standards of accuracy, and vast coverage cannot be met easily for direct input for the mechanism of defense [12].

Cyberspace is a vast area for extracting data and one of the best places for gathering intelligence because netizens post cyber-crime data, information statics, and details on internet platforms. Getting timely inputs about cyber-crime and attacks can help us in defending assets and organizations from threat actors. Collecting information from different intelligence sources is quite a task for professional analysts, so time plays a very crucial role because it delays in extracting intelligence also delays the response time for countering cyber threats. The author has proposed a detection prototype for gathering intelligence with the help of an automated machine called a novelty detection model. This proposed prototype works on the architecture of extracting intelligence information post on public exchequers like common vulnerabilities and exposure (CVE) and categorized a post on microblogging platform on Twitter as either suspected for threat or normal. This proposed architecture prototype focused on posts generated from 50 Twitter accounts containing suspicious data over 12 months. In 2018 author have created this dataset with the help of 50 Twitter accounts, and the research conducted on this prototype expose its vulnerabilities, the collected data from these Twitter posts show that they do not often contain the CVE identifiers related to cyber threats. Information gathered with the help of this proposed model from Twitter has CVE identifiers further provides vital information about the cyber threat intelligence [13].

CTI analysis is a crucial aspect of gathering intelligence in the current time frame for securing organizational assets against rapidly increasing sophisticated cyber threats. Social media intelligence is not purely open for gathering intelligence but gathering data without any restrictions there are only limited resources that can come into play like the microblogging platform Twitter and these types of resources belong to open source intelligence which is one the best source for gathering actionable intelligence. The author after research and experiment come up with a prototype based on a Twitter API in which twitter a microblogging platform used for gathering data the author has discussed in detail the implementation of a prototype and analysis of Twitter data. The author has come up with solutions for how can two different NLP works because using a state of the art NLP Stan core does not fit for text processing cyber entities data for, this problem this research proposed a solution of using oak ridge national lab (ORNL) NLP for the smooth working of this prototype. This prototype does have some flaws like while processing the data mechanism removing some useful data while filtration, and also some irrelevant data pass through its filtration process [14].

I. RESEARCH METHODOLOGY

In our proposed methodology of gathering information from Twitter and using it for generation intelligence for saving human lives and for other different purposes, we are using a Twitter API "tweepy" for live streaming of Twitter data then we will analyze the live streaming data and visualize

data with the help of an animation graph which will help us in providing a live graph, then we will evaluate the result by doing this automation of Twitter analysis. Then we will move to the other end where we will first download the tweets and create datasets for two different types of evaluation one is for gathering information about the disaster, and another one is for suicidal intentions. Our proposed model will work in different steps. In our first step we'll do text processing and text visualization in the second, we will do deep learning for big datasets and data visualization for the smaller data set. In the third section, we'll evaluate the accuracy of our model, we will apply the BERT accuracy model if we are not getting the desired accuracy. Then at the end of the fourth section, we show an ML prediction model for the prediction of events joints to happen or happing right now.

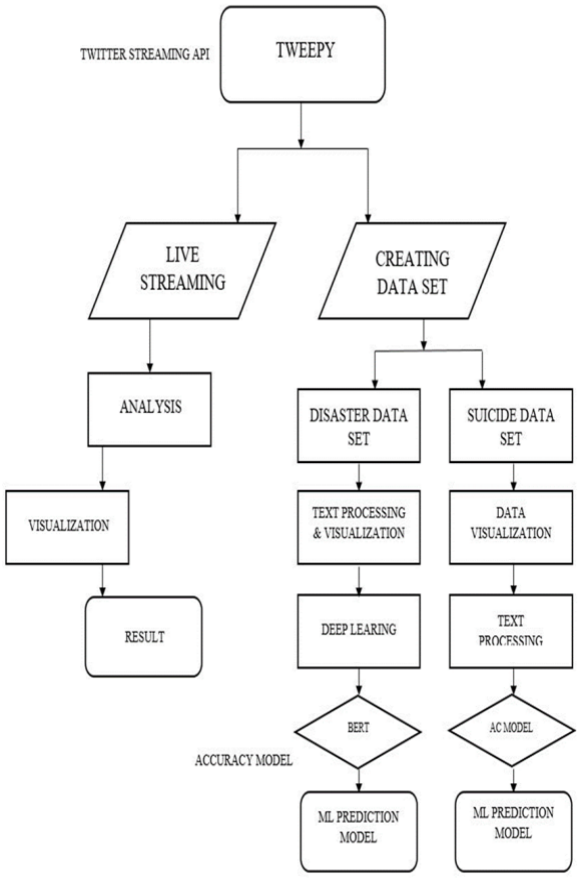


Fig6. Automated and Semi-Automated Architecture Flowchart

A. Twitter API Working

We have done rigorous research on CTI its sources and different techniques, we have studied ways of gathering information for proposes like disaster prediction, pandemic outbreak, human psychology, elections, law and order management, etc. But before we start any analysis, we need a Twitter developer account for that you have to apply then in response, Twitter will ask you to elaborate the purpose of your request for a developer account, and then if Twitter is satisfied with your answer, they will allow accessing the data through their API. We have tried different types of analysis to get appropriate results to take on different prototypes with which we will evaluate our proposed work. So in this chapter, we will elaborate on our proposed work, research methodology for gathering intelligence using Twitter API tweepy, different ways of downloading tweets, live streaming, creating datasets, gathering intelligence, evaluating the proposed model's accuracy. In the next section, we will discuss the research methodology, and further, we'll elaborate our work on live streaming of Twitter data and the creation of Twitter data sets, and the proposed model.

II. IMPLEMENTATION

A. Automation

Our implementation starts with the automation process of gathering intelligence, in this process we will try to evaluate the working of automation of threat intelligence gathering, its performance, and reliability. As we have studied various prototypes and techniques, there we have found that automation comes with some flaws, flaws that will lead your research and analysis in unwanted directions. Automation may lead to the collection of junk and irrelevant data that will decrease the accuracy rate of the proposed model. Text-processing will be difficult if there is a large amount of junk available in the data and automation of collecting data can attract irrelevant data and this will also affect the performance of the prototype and also affect its run time.

a) Live Streaming

Process starts with the import of tweepy, Tweepy is a Twitter API tool for legally download Twitter data but this will not work until few keys and specific types of token access we have because tweepy and Twitter need you to authenticate your identity so that twitter keep a record of downloaded data and this will also work as a bridge between your command ide and Twitter developer account. So, then we have provided our customer_key, customer secret, access_token, and access_token_secret. All these four credentials have been provided by Twitter when we sign in to our account for the first time in the next step user authentication is starts and Twitter has given access to Twitter data with the help of tweepy. Now after completing the authentication process, we start running a command so that we can check tweepy actually working and working properly, we run a command for fetching the data from my own Twitter home page and you can see it in the below

image where my whole timeline is present, it happened only because tweeter developer account is linked with your Twitter account and both share same password and username. Now we will start our command line and write commands for fetching the data and we will use lib. Like stream listener, because stream listener will help us in doing live streaming of tweets by this we are fetching the whole information live and the whole tweets will come in java script format for that we have to import JSON library as well.

we want to download our tweets according to our own specific keywords, in this pandemic we have seen a surge in the graph of cyberattacks because most of the world moving online for shopping, studying, official work, paying bills, this situation creates a lot of opportunities for hackers and cyberbullies and we see a rapid increase in cyber-attacks so want to do an analysis and visualize this whole scenario with the help of this Twitter analysis with the help of tweepy. Our two keywords will be "covid19" and "cyberattack" so will have the tweets containing these two keywords. So we are going to create a class and then we will have two functions stream listener function and error function so that it will not show any other data. In our function well load JSON because we are going to get tweets in JSON format. Then we'll create a Twitter stream and this stream will go to ask about the listener. And we are going to call Twitter stream and we are going to drive the covid and cyber-attack. Here is the full code with we are going to perform the twitter streaming. After performing the algorithm we will see our output like this in fig. the tweets start coming and the count will increase and how the count is increasing and how we are analysing this will we'll show you this in the visualization section where we are going to perform the live data visualization and analysis.

a) Live Animated Visualization

Now for performing visualization we are going to import some important for plotting the graph and we need an animation function for performing live graph for live streaming of tweets. We will import matplotlib and then matplotlib. Animation these both will together perform and helps in the creation of a live graph.

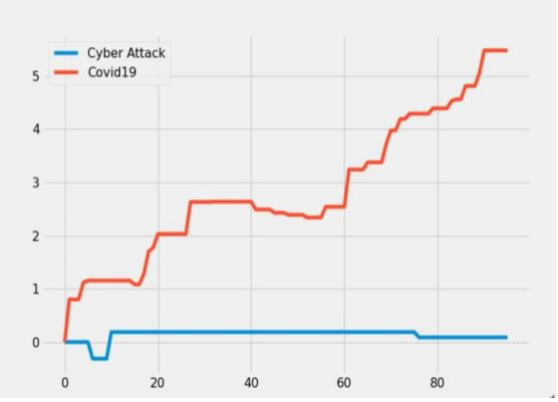


Fig7. Animated Visualization Graph 1

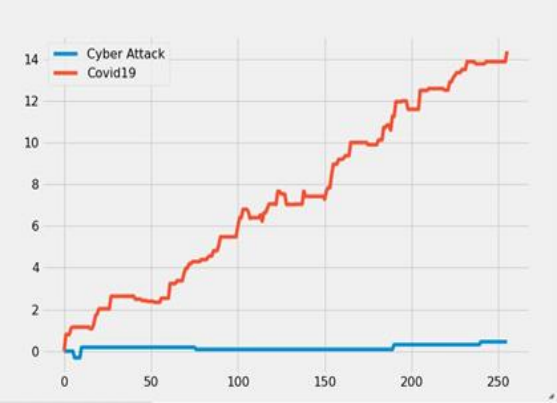


Fig8. Animated Visualization Graph 2

After performing the visualization algorithm here we have the result in the form of the graph we cannot present here live graph. So we have the images of the different numbers of tweets being analysed, first image was taken at above 80 tweets one is above 800, the third one is above 1000, and the last one at 2000. In the graph blue line represent "cyber-attack" and the orange line represents "the covid 19".

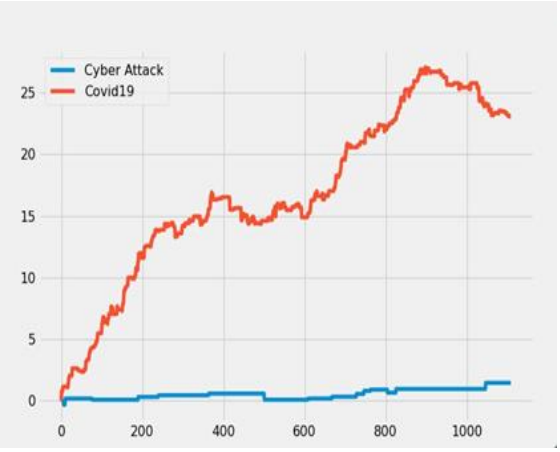


Fig9. Animated Visualization Graph 3

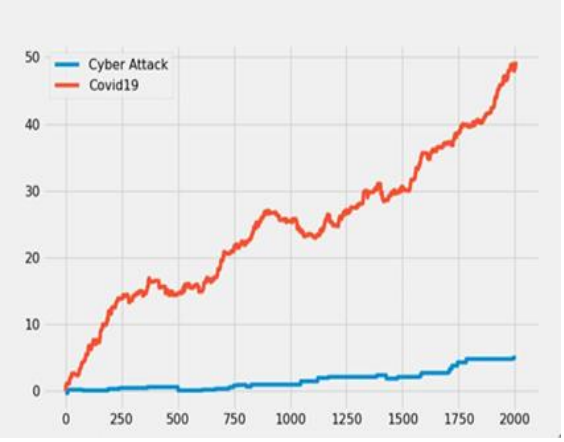


Fig10. Animated Visualization Graph 4

As we were discussing above the rising case of cyber-attacks but our analysis shows everything going great so we come to a conclusion the automation of gathering intelligence is not reliable and has very less accuracy than the semi-automated models, the gap between both lines clearly show the automation of gathering information works only in some limited events like elections and comparison of two personalities but not fit gathering vital information.

B. Semi-Automation

Semi automation as you know from the word itself tells about prototype does not support or operate on the fully automated system. The prototype is partially automated so that it can be manually operated by the operator for getting the desired result and controlling the whole process. We start with the process of creating datasets so that we can utilize those datasets for driving information and then that information will be converted into intelligence. To make the process easier we download all the required libraries that will help in the creation of data sets and as well as we'll need them in our program to derive all the important information from the Twitter datasets. So we'll first create a data set for suicidal tendency and then we'll create a dataset for disaster prediction. Our experimental prototype for semi-automation has two model-1 (suicidal tendency) and model-2 (disaster management) both the prototypes have different sizes of datasets. Model-1 contains a smaller dataset as compared to model-2. So we'll further discuss the model in the increasing order the whole discussion is carried out in this manner so that we can explain the whole reason behind having two data sets of different sizes.

model-1 small dataset

Model -1 was specifically selected for the suicidal tendency because it has a smaller data set than the second model which has a larger dataset. We have two datasets of different types and with different types of difficulties so that we can test our prototypes rigorously. Model -1 was specifically selected for the suicidal tendency because it has a smaller data set than the second model which has a larger dataset. We have two datasets of different types and with different types of difficulties so that we can test our prototypes rigorously. Suicidal tendency data is about the person having a suicidal tendency or not/ faking and having different intentions related to committing suicide. Our prototype will assign "1" to those tweets in which the intentions of a person reflects higher chances of committing suicide and "0" to those who do not reflect such intention of committing suicide. All the libraries which we will need for performing the algorithm right from pandas to NumPy and then tweepy our main API tool for gathering tweets, then matplotlib to seaborn for the visualization, tqdm which will be helpful in the life tracking and rest of the libraries we'll import like TensorFlow, ktrain, etc so that we make you understand why we are using these libraries. Next in the algorithm, we'll give our keys and token access for the authentication process, and then we'll apply query because it will download the tweets according to our provided keywords. Then apply the query in our function so we can download our tweets and create a dataset. And our dataset will be in excel format it means we have to download it in CSV extension for that we have also imported CSV and now we'll start our download and create datasets for the semi-automated collection of twitter's information for intelligence gathering. So when we choose to do suicidal tendency prediction the question is what is a suicidal tendency and how we'll able predict it, people often get depressed, and then they post their psychologically depressing views on social media platforms like Facebook, Instagram, Twitter, etc. so we choose twitter for the prediction because as we have proved that the OSINT is one of the best sources of gathering information and Twitter is the best source for collecting data, so Twitter is our obvious choice for this work.

Text Pre-processing

Now the next thing why we are calling it a semi-automated collection is because we are performing algorithms in which we are doing both automation and semi automation, automation where we feel it is necessary and we also perform no automated function for downloading datasets we do use automation and also for text processing we use a pre-process tool we have created and the imported and we are directly applying it to our text. So here we are using "import preprocess_GKStool as gks "for our text processing to clean our data, we have used spacy for our NLP here in this tool for text filtering and cleaning of data. So now how we are able to perform

our model here we have given 1 to those sentences we feel have the true intentions of suicide, and 0 to those tweets which are not threatening to human life. Further, we download our data set so that we can visualize our dataset in a data frame form, then we'll perform text pre-processing with the help of our tool.

TEXT VISUALIZATION

Data visualization helps us in viewing how our text pre-processing has performed whether filtered data is ready for further process of training and testing or not because raw data is full of junks like U.R.L, #tags, special characters, e.t.c. so a better way to deal with the problem is text pre-processing and then visualize the data. In the process of text filtration, raw text allotted with the intention of "1" and "0", "1" shows the suicidal tendency of the person in the text, and "0" indicates the fake intent of committing suicide. After the filtration process of raw text has been completed and after that, it will be presented in a table format where we have serial number, tweet, and intentions.

	tweet	intention
0	my life is meaningless i just want to end my l...	1
1	muttering i wanna die to myself daily for a fe...	1
2	work slave i really feel like my only purpose ...	1
3	i did something on the 2 of october i overdose...	1
4	i feel like no one cares i just want to die ma...	1

Fig11. Intentional Tweets

ML PREDICTION

After the training and testing of our model, we will calculate our accuracy. We have successfully performed our model and achieved an accuracy of 93%. After that, we have performed our ml prediction model successfully. So far precision of our prototype for the smaller data set is the best among the best we have reviewed during our research.

	precision	recall	f1-score	support
0	0.94	0.94	0.94	1060
1	0.91	0.91	0.91	764
accuracy			0.93	1824
macro avg	0.93	0.93	0.93	1824
weighted avg	0.93	0.93	0.93	1824

Fig12. Prototype Prediction

Further, we check the prediction accuracy of our prototype then we assign two sentences for the prediction so, according to the prototype "1" will be assigned to the sentence having higher chances of committing suicide and "0" to the one having different intent of posting the tweet. The first sentence is "no one care about me, I will die alone" the prediction result is 1 for this phrase because it shows the clear intent. The second sentence we have given is "today I am so happy, thanks a lot for making it special" "0" is the outcome of this statement because this phrase does not show the intention of committing suicide.

2 Model-2 Big Dataset

Prediction on Twitter data is a game of words any prediction we did and we will do is depend on how we use proper coding on text, so now we are going to predict whether there is a disaster that happened or not. We'll be using the BERT model on Twitter data, there are keywords that indicate that there is something happing or not, this is going to be a very complex procedure, in this procedure we will train our model with Tfidf, linear SVM, EDA, word to vector representation, deep learning methods, in the end, we will use BERT model for increasing our accuracy, BERT model has outperformed other models. BERT model not only uses the current layer it uses backward and forwards layer also for the implementation before there is no model which can perform like this. BERT has achieved state-of-the-art and accurate performance. So what do people do if something happened they start tweeting whether it is a wildfire, earthquake, thunderstorm, an accident, so for the prediction we will simplify the text, do text pre-processing and visualization to understand it better then well apply some deep learning methods, then we'll apply BERT model so that we can get the desired accuracy and then we'll perform our ML prediction model. BERT performs two-stage processing supervised and unsupervised learning. And we will use ktrain, ktrain is a library that we used as a wrapper library for tensor flow, we use it for our deep learning models. We'll do deep learning tokenization and also perform deep learning word embedding. We'll perform programs for text processing and then we'll visualize data so that we can apply our model to it. For visualization, we have imported different libraries like matplotlib and seaborn.

a) Text Pre-Processing

A special tool is used for the text pre-processing that we are using here is programming with the different packages that have different steps for the filtration of the raw text because when we download data, it is in the raw form, but that raw form is not useable for gathering intelligence for that we have to make

data fit for programming process. This tool helps in removing junk like URLs, emails, retweets, HTML tags, #tags, accented characters, special characters, etc, and not only these types of characters but also helps and works with different NLP forms. We have applied 'gks' tool kit for text processing and have separated data according to word_count, char_count, avg_wordlength, stopword_count, so in the next phase of the process to analyze it according to these characteristics.

0	our deeds are the reason of this earthquake ma...
1	forest fire near la ronge sask canada
2	all residents asked to shelter in place are be...
3	13000 people receive wildfires evacuation orde...
4	just got sent this photo from ruby alaska as s...

Fig13. Filtered Dataset

So, here you can see in the above table we have filtered the raw text and made it fit for the process, and now if you can notice all the junk has been removed, we can visualize it for further analysis. Only after getting filtered data, we can move forward with the process of test and train.

b) Visualization

During our visualization, we have found that there is a rise in the graph where we find tweets are genuine and genuine are have around 120 character count, that is also clearly indicated in violin chart also in word count graph there is a rise in genuine disaster tweets from 10 to 20, average word length appeared to be same in the graph, so further we have evaluated and found that clearly on word cloud that we have successfully separate the genuine and fake tweets are real word cloud shows the words which indicate disaster, and vice versa with the false one. The pie chart has clearly shown the amount of real and false tweets. The real one indicates “1” and the false indicates “0”.

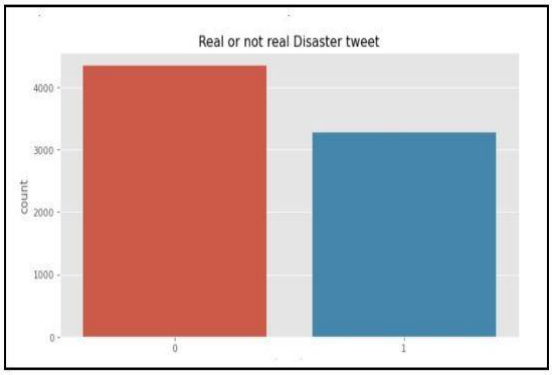


Fig14. Real or Not Real Disaster Graph

At the start of the visualization process here, we have a bar graph of real and non-real tweets, here we have 3000 tweets that are real and the amount of non_real tweets is above 4000. Based on this information we can plan further steps to identify and precise our processing.

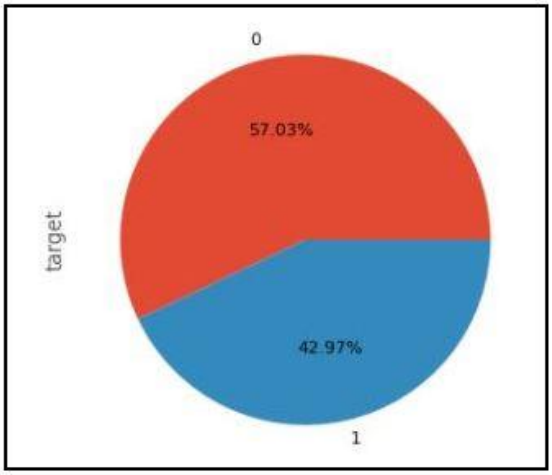


Fig15. Real or Not Real Disaster Pie Chart

Next, we have the pie chart with target "0" and "1" which shows the percentage of tweets that we want to focus on and target for further assessment. Real tweets stands at 42.97% and non-real are at 57.03%. This 42.97% of tweets that have the intent and are real can contain useful information which can be converted into threat intelligence. Now we will take this data analysis through visualization into depth so that we can know how this analysis can help drive vital information. Right from word count to word length, character count e.t.c.

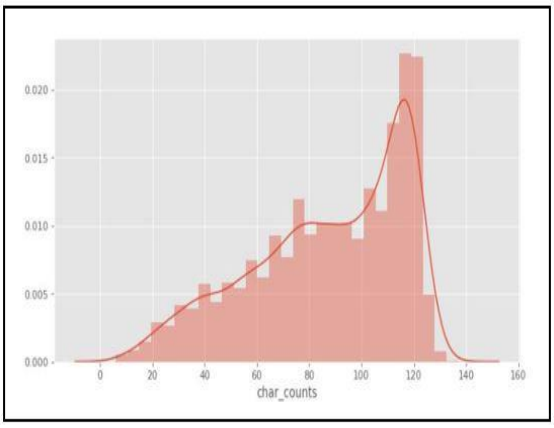


Fig16. Character Counts Line graph and Manhattan of Original Tweets

The bar chart above shows the character count (char_counts) for visualizing the output of the pretext processing of the text and the out of the positive tweets in terms of the tweets indicating disaster happening to have the code 1 in the form of

Manhattan here we can visualize the clear indication of how to know exactly what real tweets indicates. We can see a high rise in Manhattan near 120 this clearly shows that tweets with intentions of posting disaster tweets have a total character count of about 120.

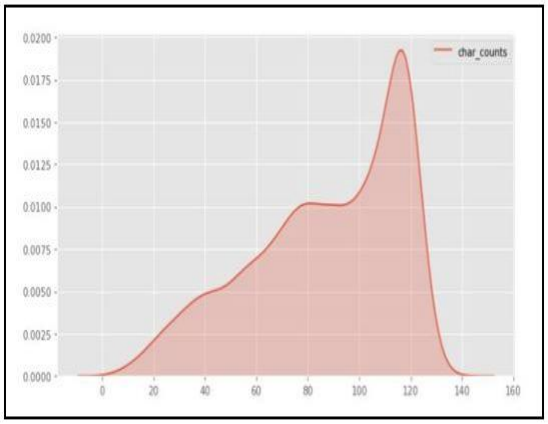


Fig17. Character Counts Line Graph of Original Tweets

Here we have the line graph indicated the same as the previous manhattan graph this graph will help further in the comparison to the other graphs and especially those who carry no real tweets.

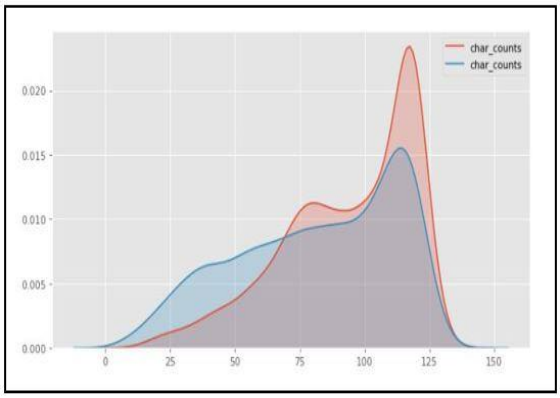


Fig18. Comparison Character Counts of Fake and Real Tweets

Now we focus on the comparisons of the different entities so the first one in our list is character count (char_count) and we'll differentiate here how real and fake one works when it comes to the character count. As we know if we have a piece of information about some disaster so we usually try to put all the vital information in one go so, it indicates that the word count for the real tweet carrying the disaster information having characters counts at a higher level than the one pretending to have a real tweet. Here the fake tweet with improper intentions has a raise around 0.015 so that's the maxing point of the tweets. Now, if focus on original tweets we see character count at a higher level than fake ones indicating the difference between the two.

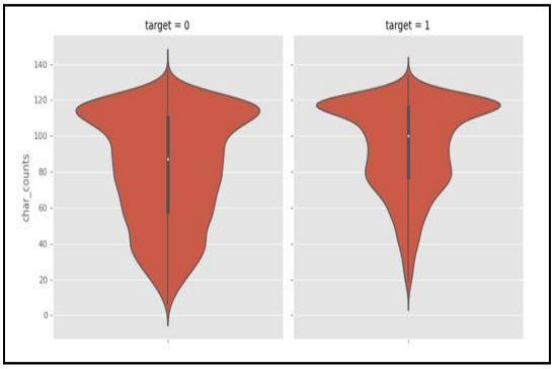


Fig19. Comparison Character Counts Violin Graph

The Violin graph shows the rise in both sides in a two-dimensional manner in the shape of the musical instrument violin we have target = 0 indicates tweets with wrong information and 1 with the information of disaster that happened. we see the violin graph of target =0 tweets having a stretched out shape and 1 with the stretched in the shape so it clearly shows the original tweets are more precise and in fitted form.

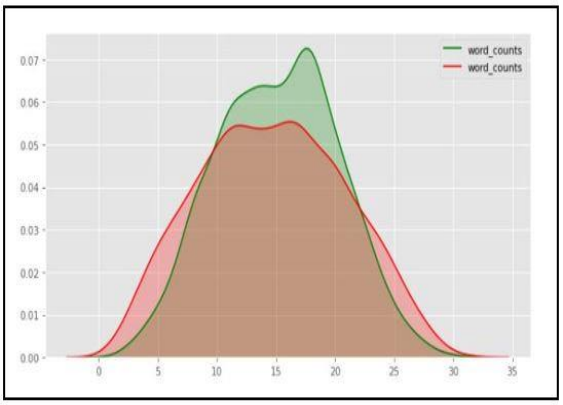


Fig20. Comparison Word Counts of Fake and Real Tweets

Word count (word_counts) graph sheds light on the count of words in target and non-target tweets from the above character count graph we have observed that original tweets that contain relevant information have a higher amount of characters it is obvious now they will also have a higher word count. Now, if we observe the word count graph there is a rise from 15 to 20 in tweets containing the original content, and those who are with non-target tweets have a lower word count, and the result is quite obvious according to our first analysis of graphs.

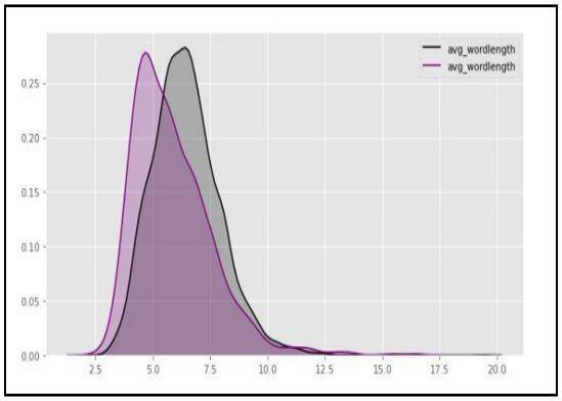


Fig21. Comparison Average word length of Fake and Real Tweets

As we have observed in the graphs above that differentiating between the original tweets and the fake ones is easy, but here we are not getting a clear picture because both types of tweets have the same average word length (avg_wordlength). Using the average word length (avg_wordlength) graph for visualization is disastrous here because it cannot clear the process and make it complex for further study of the prototype.

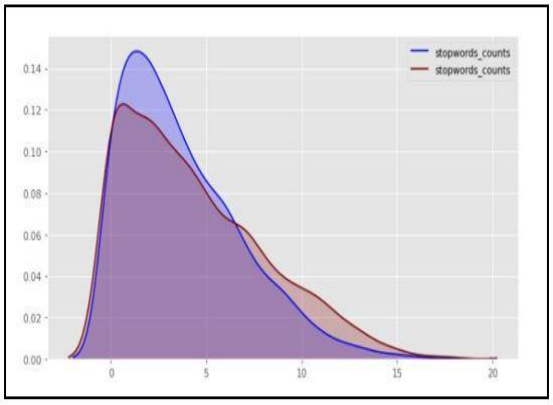


Fig22. Comparison Stop word Counts of Fake and Real Tweets

Stop word count (stopwords_counts) graph is helpful just like we have analysed the character count graph, violin graph, word count graph above. In the stop word count graph (stopwords_counts) we can differentiate the target tweets and the tweets which are ignored by the prototype and consider a junk difference is seen in this graph. From 0-5 and 0.12-0.14, the difference is visible between tweets of original content and junk tweets, the original tweets contain a slightly higher portion than the others, and here we can easily identify the disaster tweets. Now, we'll focus on the particular words max and min also the on positive and negative words, positive here means tweets with having disaster information and vice versa.

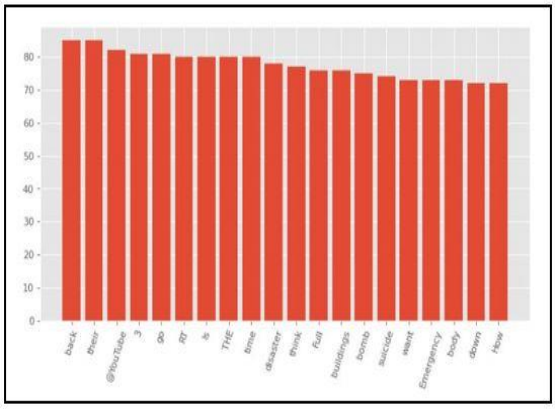


Fig23. Top 20 Most Used Words

Bar chart for showing top 20 words that have been used in our data set so that we can recognize our prototype is on right track and have a record of words we have found according to target. Target words we have found here clearly show that a disaster has happened, so words are the bomb, suicide, disaster, and emergency. the target words that show us disaster happened to lie between 70 - 80, so this means that these have been used from 70 t0 80 times in the targeted tweets, but we have also found some correlated words like building and down if we can use these with the words show some disaster content then we will see this can make sense like some disaster is happening some were to a building that has gone down.

https://t.co/AlnV51d95x	1
ravioliâĚwith	1
http://t.co/tt4kVmvuJq	1
MANAGER	1
Responders-	1
@MarioMaraczi	1
#09	1
@cncpts	1
Oliver.	1
http://t.co/C0SlAbBP7j	1
dotish	1
CENTER	1
front.	1
barracks	1
Metal)	1
http://t.co/sEquWmvFx4	1
waste...noxious...	1
figures	1
knowledge:	1
Tita	1

Fig24. Top 20 Least Used Words

Table for the words used at least 20 times in our data for disaster management, and table clearly shows words lack intent of becoming able to be an intelligence-gathering indicator. This table also shows junk content used here, which will be clear during text pre-processing.

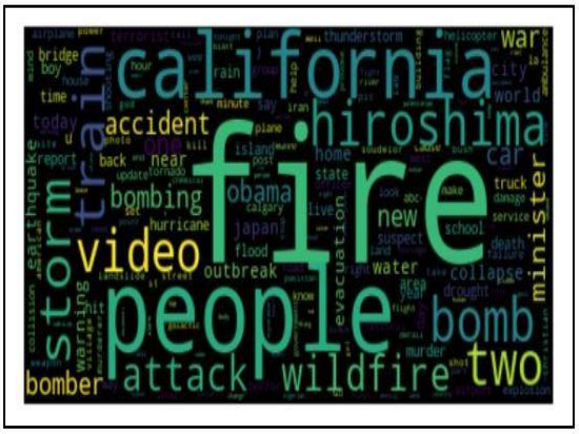


Fig25. Word Cloud of Real Tweets

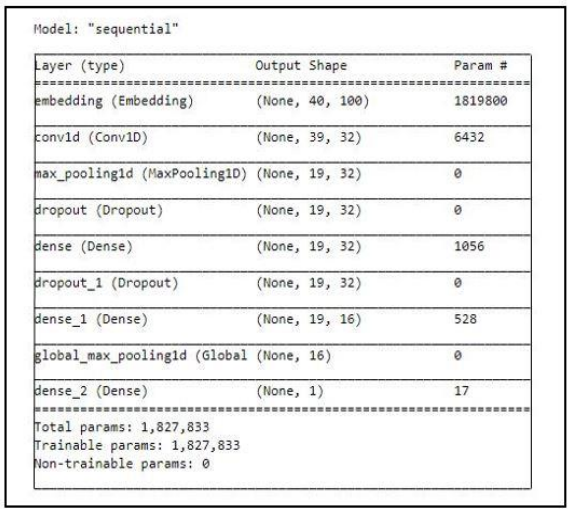
We'll form here a cloud of words (word cloud) for the target = 1, tweets, and it will combine all the target words and combine them in one frame. The words are fire, bomb, attack, wildfire, bomber, war, accident, storm, and bombing are visible here as the indication that a disaster has happened that's why the word cloud is very useful in understanding the process have successfully picked the write tweets for gaining intelligence.

Fig26. Word Cloud of Non Real Tweets

Target = 0 here in the word cloud indicates the fake tweets having words like new, love, one, body, time, know, think, and day shows that our text pre-processing has performed exceptionally well and worked as per the process.

DEEP LEARNING

Now we'll perform linear SVM on our model and get the classification report and it shows an accuracy rate of 80%, we move forward and call our NLP "spacy" and also import its English large library, and then we imply text to vector model on our data set and then we again check our accuracy rate, it is same we have mentioned above.



Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 40, 100)	1819800
conv1d (Conv1D)	(None, 39, 32)	6432
max_pooling1d (MaxPooling1D)	(None, 19, 32)	0
dropout (Dropout)	(None, 19, 32)	0
dense (Dense)	(None, 19, 32)	1056
dropout_1 (Dropout)	(None, 19, 32)	0
dense_1 (Dense)	(None, 19, 16)	528
global_max_pooling1d (Global	(None, 16)	0
dense_2 (Dense)	(None, 1)	17

Total params: 1,827,833
Trainable params: 1,827,833
Non-trainable params: 0

Fig27. Deep Learning Model

Then we have started applying tokenization to our word text dataset. And then we apply linearSVC to check our train and test model and shocking to our accuracy have dropped to just 52%. They apply the sequential model of deep learning after that we have trained and tested the model so that we can move on BERT model for our accuracy improvement.

TABLE 2. Keras Model Runnning

EPOCH	TIME	TIME/STEP	ACCURACY	VAL_LOSS	VAL_ACCURACY
1/5	7s	28ms/step	0.6266	0.5417	0.7571
2/5	5s	27ms/step	0.8066	0.4591	0.7958
3/5	5s	27ms/step	0.8750	0.4684	0.7919
4/5	5s	27ms/step	0.9176	0.5089	0.7741
5/5	5s	29ms/step	0.9401	0.5560	0.7669

After calling the BERT model have tested and trained our data for the last time on which BERT has helped us to improve our accuracy to 86%. Word embedding and classification with deep learning will create vectors internally using a one-dimensional conventional neural network 50% of the neural network will be dropped while training the prototype, the sigmoid says that we need binary output.

BERT MODEL

Here we have to perform an accuracy matrix that will exactly tell the process and condition of the current phase of the process. Hereafter the implementation of the matrix in the result we have found that validation accuracy (val_accuracy)

overfitting and considerably low because prototype working on a smaller dataset for employing proper neural network programing we need data in millions. Still, we can improve our accuracy if we change the "dense layer" dropout in the above neural network model. BERT - bidirectional encoder representation from transformers is a model used here to fine-tune our text or the processing of our prototype for getting the desired result for that first has to install k-train. K-train is a wrapper it will download all the dependencies and the necessary trained python packages.

TABLE 3. Ktrain Cycle 1

Ktrain Cycle 1						
Max lr	**TIME**	**TIME/STEP**	**LOSS**	**ACCURACY**	**VAL_LOSS**	**VAL_ACCURACY**
0.0002	5638s	52s/step	0.4726	0.7803	0.3939	0.8320

Ktrain cycle-1(TABLE 3. Ktrain Cycle 1) process fine-tunes and uplifts the prototype working it enhances the overall text processing and accuracy of the model. Cycle-1's detail has mentioned above in the table accuracy stands at 78, and val_accuracy stands at 83, next well run another cycle and try to improve the accuracy. After the test and train completion of the first cycle, the process of the second cycle takes place, and the details have been mentioned above.

TABLE 4. Ktrain Cycle 2

Ktrain Cycle 2						
Max lr	**TIME**	**TIME/STEP**	**LOSS**	**ACCURACY**	**VAL_LOSS**	**VAL_ACCURACY**
0.0002	2059s	19s/step	0.3386	0.8624	0.3843	0.8412

Ktrain cycle-2(TABLE 3. Ktrain Cycle 2) performs better than the 1st cycle, as we can see here in the results above accuracy have increased and raised to 86, and val_accuracy raised around at 84. We can get results as per our desires, just need to play with the learning field (lr=2e-4, epochs=1) by changing the field we can get the accuracy up to what we want.

ML PREDICTION

ML prediction helps in getting the information about the working of the prototype, and it helps in testing the accuracy of the model. We will assign two sentences to our model to predict target or not target, and if it is working up to the level of its accuracy, it correctly predicts the target as '1' and not target as '2'. So here we have given two sentences, the first one is (i met you by an accident) and the second one is (I am hit by a car, I am injured). We have given these sentences to our prototype to predict the target sentence that indicates the disaster, and we have got our result as 1.

III. RESULTS AND DISCUSSION

Existing Techniques

The two existing techniques which we are going to look at are Twitter OSINT and Action miner, both of them are intelligence gathering prototypes and have a high level of accuracy. So, we have chosen these two for the comparison and analysis based on parameters developed previously mentioned in cyber threat intelligence comparative analysis of its sources and parameters of evaluation.

Twitter OSINT (open-source intelligence) is a sophisticated automated system for cyber threat intelligence gathering which collects information from openly available sources for analysis. Twitter is an openly available source for gathering actionable information people post different types of informative tweets about recent or upcoming activities. It is a java based system backed by the Twitter streaming API for downloading tweets in a legalized manner based on the keywords provided by the author. Downloaded tweets were then processed and analyzed through NLP and used different libraries for English literature as well as words used specifically for cyber entities. JSON is used to store the processed data after the initial process prototype uses the elastic stack for built-in data analytics and ML capabilities. A smaller data set was used for the preliminary experiment, and the process also did remove some of the relevant tweets at around 12% during the execution of the prototype.

Rapid growth in attacks on the virtual and real-world left the world vulnerable to upcoming threats, so the requirement of a cyber threat intelligence model was felt by the author. For countering these unwanted threats, they have developed a prototype called Action Miner for countering cyber threats by analyzing text available in the cyber domain combined with some NLP techniques. This prototype has achieved very high precision and has denied the use of the state-of-the-art Stanford library due to its flaws against text related to cyber security.

Back Ground

In [9] had given concluding remarks on the development of a methodology and ethical use of social media intelligence, Next in [15] analysis of the SOCMINT elaborates the enlighten the knowledge in all important directions on how intentionally target Facebook users can affect elections of world's oldest democracy. Various key issues like these undoubtedly affect the reliability of sources like social media. [11] One framework that outperforms the state-of-the-art tools is action

miner, but we have to remember that it is for only low-level threats, therefore it covers only a minor portion of the cyber threat world. [4] Research has come up with a unique idea of filling the essential difference between the practical and theoretical approaches for a flexible cyber-physical system by reliable use of cyber threat intelligence. [16] This model exposes the CTI systems approach lacks an ontology covering the complete spectrum of CTI.

Our rigorous study on various sources, techniques, approaches, and models has led us to develop necessary parameters and evaluation processes for sources so that we know the most trusted sources and on which source we can rely for the cyber threat intelligence gathering. We have wisely decided there will be 4 parameters on which our 3 important sources will be evaluated 4 parameters are Sources, CTI Model, Reliability, and Performance. Our first parameter is the source itself that tells us precisely about sources on which we will evaluate that source, the second parameter is the CTI model, which elaborates the how many models is important for a source for intelligence gathering, reliability is our third parameter in which we will explain briefly about the reliability of a source for cyber threat intelligence gathering, fourth and the last parameter is performance this parameter is one of the most significant ones for the evaluation because it will depend on the result of the physical approach of a source. We will evaluate the sources based on parameters, there are two tables, and their brief description will further elaborate the whole process. This is the process through which we have found out parameters on which we are evaluated and then the grading system for getting output.

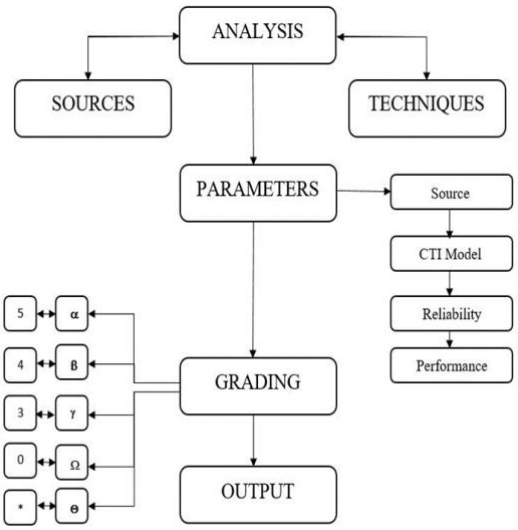


Fig28. Flow Chart of Parameters of Evaluation

We have successfully implemented our semi-automated collection and analysis of the Twitter data model we have a quite good response from the implementation of our proposed work one of our models gets 93% accuracy and the other one is having an accuracy of 86%, now in the chapter, our goal is to compare the proposed algorithm with the existing ones on different parameters. we have done a rigorous study on various aspects of CTI and have found that accuracy is one of the important features that every prototype or model should acquire by performing well, so accuracy will be our first feature, the second one will be reliability like we have performed automated live streaming gathering data and then after analyzing it we have found have we can't rely on this type of model because it is on reliable so we moved on the semi-automated system for better prediction of events and gathering of intelligence, reliability will our second parameter for the evaluation process. So for the better performance of the evaluation process, our third parameter will be the CTI source that is we guess is important for driving appropriate information because CTI plays a vital role in the gathering of informational data. Our third parameter will be the CTI source.

Our parameters evaluation process is very simple and precise and also going to be published in Scopus indexed journal .The only difference we have here is there we have evaluated the sources of CTI and here we are going to evaluate the existing model with the previous technologies. in various technologies we will have TwitterOSINT and ActionMiner, TwitterOSINT is just like our prototype but it is an automated prototype for the collection of CTI and worked on OSINT, having an accuracy of 78% and ActionMiner is state-of-the-art technology for gathering cyber threat intelligence but only useful against small threats, gathering information from SOCMINT having a high accuracy rate of 92%. Nowhere our model will be going to perform is semi-automated threat intelligence gathering model, we are not going to use our disaster prediction model because we are using their BERT Accuracy Model and with that, we can achieve 100% accuracy so for a fair evaluation we are going use here our suicidal tendency model.

Parameters of Evaluation

TwitterOSINT

- Accuracy: 78%
- Reliability: It is not a reliable technique because using may have flaws and can lead whole information gathering in the wrong direction, like we have performed above how automation of gathering information can be disastrous in some situations.
- CTI Source:TwitterOSINT is based on gathering information OSINT, and OSINT is

one of the best sources for gathering information.

B) ACTIONMINER

- Accuracy: 92%
- Reliability: ActionMiner is a good technique for gathering CTI but still is only good for small threats, in a different scenario where we are facing something big it will be useless.
- CTI Source: ActionMiner is based on SOCMINT, SOCMINT has its pros and cons like accounts having privacy and gathering information from a private is illegal and SOCMINT can provide crucial information at the time of need but privacy issues make it less effective than OSINT.

C) SEMI-AUTOMATED: CTI TWITTER

- Accuracy: 93%
- Reliability: CTI Twitter is a good model for performing data collection and driving information from collected data and its ml precision model's performance we have seen above.
- CTI Source: From the evaluation process of chapter 4. We know that OSINT is one of the best sources for gathering data.

Comparison

For comparison, we are considering here one of our research papers, cyber threat intelligence 'comparative analysis of its sources and parameters of evaluation'. We have conducted rigorous research on how to evaluate cyber threat intelligence sources and the prototypes on different parameters.

TABLE 5. Grading Scale for Parameters

Performance	Grade
Best	α [5]
Good	β [4]
Average	γ [3]
Poor	Ω [0]
Underdevelopment	Θ [*]

Here are the parameters on which we are going to evaluate our prototype for comparison, and the parameters are Accuracy, Reliability, and cyber threat intelligence model (CTI Model). The grading system we are using here will give ratings from best to poor and underdevelopment for those prototypes that are not fully functional, the best = alpha (5), good = beta (4), average = gamma (3), poor = omega (0), and underdevelopment = theta (*).

TABLE 6. Evaluation on the Basis of Parameters

Parameters	TwitterOSINT	ActionMiner	CTI:Twitter
Accuracy	0	3	4
Reliability	3	3	4
CTI Model	4	3	4

Our model has outnumbered every other technique in the evaluation process and performed well but still, there is scope for improvement and our next is making our model achieve 98% with all sorts of threats and making it about 70% automated system.

IV. CONCLUSION

In our research, we've found that correlation analysis shows that some social media platforms that contain similarities like Facebook and Instagram come under the same ownership, so there is a possibility that soon we can develop a prototype that can gather all the information which is obtainable without privacy restrictions from the platforms together. Here we will gather information from social media platforms but through an open-source intelligence method. We can develop a similar method like we have mentioned above for platforms where we can see extensive use of hashtags "#" Like Twitter, Instagram, and even some platforms used for streaming videos like YouTube, we can use these hashtags for gathering information from different social media sources but again we will use an open-source intelligence method for generating cyber threat intelligence. Rigorous analysis of multiple sources resulted in the development of parameters for the evolution of different CTI sources which indicates the performance of different sources based on different parameters. Social Media Intelligence [SOCMINT] & Open Source Intelligence [OSINT] will play a vital role in containing cyber threats and for the management of law and order in the real world. Our evaluation shows SOCMINT and OSINT perform better than deep and dark web intelligence. According to our evaluation process, we have seen open source intelligence perform better than any other source but also have scope for improvement.

We have also successfully implemented our proposed prototype and have achieved an accuracy of 93% with a semi-automated analysis of Twitter data, we have also performed automated collection of data and live streaming and visualization of Twitter data, but that performed drastically bad so considered using the semi-automated model and have introduced BERT state-of-the-art model for improving accuracy and have achieved 86% with that, BERT can also achieve 100% accuracy if we want but have use suicidal tendency model for the

fair evaluation of process and then we have conducted an evaluation of proposed work with the different existing techniques, and we have proved our model have performed better than the other two models which are TwitterOSINT and Action Miner. Need more prototypes that cover vast areas of social media despite limiting them to one platform that should embrace multiple platforms with a high level of accuracy. A fruitful collaboration of modern researchers and government organizations should be promptly formed to gather reliable information from complex and large sources.[3] There are frequently specific issues that should be addressed, a few of them are maintaining the quality of the threat generation, structuring of data, sharing of information between different organizations, analysis of developed prototypes at frequent intervals.

V. REFERENCES

[1] A. Ramsdale, S. Shiaeles, and N. Kolokotronis, "A comparative analysis of cyber-threat intelligence sources, formats and languages," Electron., vol. 9, no. 5, 2020, doi: 10.3390/electronics9050824.

[2] A. Zrahia, "Threat intelligence sharing between cybersecurity vendors: Network, dyadic, and agent views," J. Cybersecurity, vol. 4, no. 1, pp. 1–16, 2018, doi: 10.1093/cybsec/tyy008.

[3] D. Schlette, F. Böhm, M. Caselli, and G. Pernul, "Measuring and visualizing cyber threat intelligence quality," International Journal of Information Security, vol. 20, no. 1. pp. 21–38, 2021, doi: 10.1007/s10207-020-00490-y.

[4] E. Bou-Harb, W. Lucia, N. Forti, S. Weerakkody, N. Ghani, and B. Sinopoli, "Cyber meets control: A novel federated approach for resilient cps leveraging real cyber threat intelligence," IEEE Commun. Mag., vol. 55, no. 5, pp. 198–204, 2017, doi: 10.1109/MCOM.2017.1600292CM.

[5] S. Lightfoot and F. Pospisil, "Surveillance and privacy on the deep web," no. March, pp. 0–27, 2018, doi: 10.13140/RG.2.2.21692.74889.

[6] R. Baumgartner, M. Ceresna, and G. Ledermüller, "Deep Web navigation in Web data extraction," Proc. - Int. Conf. Comput. Intell. Model. Control Autom. CIMCA 2005 Int. Conf. Intell. Agents, Web Technol. Internet, vol. 2, no. May, pp. 698–702, 2005, doi: 10.1109/cimca.2005.1631550.

[7] A. T. Zulkarnine, R. Frank, B. Monk, J. Mitchell, and G. Davies, "Surfacing collaborated networks in dark web to find illicit and criminal content," IEEE Int. Conf. Intell. Secur. Informatics Cybersecurity Big Data, ISI 2016, no. August 2017, pp. 109–114, 2016, doi: 10.1109/ISI.2016.7745452.

[8] I. Deliu, C. Leichter, and K. Franke, "Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018-January, no. December 2017, pp. 3648–3656, 2017, doi: 10.1109/BigData.2017.8258359.

[9] D. Omand, J. Bartlett, and C. Miller, "Introducing social media intelligence (SOCMINT)," Intell. Natl. Secur., vol. 27, no. 6, pp. 801–823, 2012, doi: 10.1080/02684527.2012.716965.

[10] R. Bernard, G. Bowsher, C. Milner, P. Boyle, P. Patel, and R. Sullivan, "Intelligence and global health: assessing the role of open source and social media intelligence analysis in infectious disease outbreaks," J. Public Heal., vol. 26, no. 5, pp. 509–514, 2018, doi: 10.1007/s10389-018-0899-3.

[11] G. Husari, X. Niu, B. Chu, and E. Al-Shaer, "Using entropy and mutual information to extract threat actions from cyber threat intelligence," 2018 IEEE Int. Conf. Intell. Secur. Informatics, ISI 2018, no. November, pp. 1–6, 2018, doi: 10.1109/ISI.2018.8587343.

[12] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the IOC Game," pp. 755–766, 2016, doi: 10.1145/2976749.2978315.

[13] B. D. Le, G. Wang, M. Nasim, and M. Ali Babar, "Gathering cyber threat intelligence from twitter using novelty classification," arXiv. 2019.

[14] S. R. Vadapalli, G. Hsieh, and K. S. Nauer, "TwitterOSINT," Proc. Int. Conf. Secur. Manag., pp. 220–226, 2018, [Online]. Available: https://csce.ucmss.com/cr/books/2018/LFS/CSREA2018/SAM9750.pdf.

[15] K. V. Rønn and S. O. Søe, "Is social media intelligence private? Privacy in public and the nature of social media intelligence," Intell. Natl. Secur., vol. 34, no. 3, pp. 362–378, 2019, doi: 10.1080/02684527.2019.1553701.

[16] V. Mavroeidis and S. Bromander, "Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence," Proc. - 2017 Eur. Intell. Secur. Informatics Conf. EISIC 2017, vol. 2017-Janua, pp. 91–98, 2017, doi: 10.1109/EISIC.2017.20.

[17] M. Macdonald, R. Frank, J. Mei, and B. Monk, "Identifying digital threats in a hacker web forum," in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15. Association for Computing Machinery (ACM), Aug 2015, pp. 926–933.

[18] D. Shackleford, "The SANS state of cyber threat intelligence survey: CTI important and maturing," SANS Institute, Tech. Rep., 2016. [Online]. Available: https://www.sans.org/reading-room/whitepapers/bestprac/state-cyber-threat-intelligence-survey-cti-important-maturing-37177

[19] A. Abbasi, W. Li, V. Benjamin, S. Hu, and H. Chen, "Descriptive analytics: Examining expert hackers in web forums," in 2014 IEEE Joint Intelligence and Security Informatics Conference. Institute of Electrical and Electronics Engineers (IEEE), Sep 2014, pp. 56–63.

[20] R. McMillan, "Open threat intelligence," https://www.gartner.com/doc/ 2487216/definition-threat-intelligence, 2013.

[21] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, and P. Shakarian, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," in 2016 IEEE Conference on Intelligence and Security Informatics (ISI). Institute of Electrical and Electronics Engineers (IEEE), Sept 2016, pp. 7–12.

[22] J. H. Hoepman, "Privacy design strategies," IFIP Adv. Inf. Commun. Technol., vol. 428, pp. 446–459, 2014, doi: 10.1007/978-3-642-55415-5_38.

[23] M. Glassman and M. J. Kang, "Intelligence in the internet age: The emergence and evolution of Open Source

Intelligence (OSINT)," Comput. Human Behav., vol. 28, no. 2, pp. 673–682, 2012, doi: 10.1016/j.chb.2011.11.014.

[24] S. Agarwal and A. Sureka, "Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats," pp. 1–18, 2015, [Online]. Available: http://arxiv.org/abs/1511.06858.