**INTRODUCTION**

R ECENTLY there has been an increasing reliance on the Internet for business, government, and social interactions as a result of a trend of hyper-connectivity and hypermobility. While the Internet has become an indispensable infrastructure for businesses, governments, and societies, there is also an increased risk of cyber attacks with different motivations and intentions. Preventing organizations from cyber exploits needs timely intelligence about cyber vulnerabilities and attacks, referred to as threats [1]. Threat intelligence is defined as “evidence-based knowledge, including context, mechanisms, indicators, implications, and actionable advice, about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject’s response to that menace or hazard” [2]. Threat intelligence in cyber security domain, or cyber threat intelligence, provides timely and relevant information, such as signatures of the attacks, that can help reduce the uncertainty in identifying potential security vulnerabilities and attacks. Cyber threat intelligence can generally be extracted from informal or formal sources, which officially release threat information in structured data format. Structured threat intelligence adheres to a well-defined data model, with a common format and structure. Structured cyber threat intelligence, therefore, can be easily parsed by security tools to analyze and respond to security threats accordingly. Examples of formal sources of cyber threat intelligence include the Common Vulnerabilities and Exposures (CVE) database 1 and the National Vulnerability Database (NVD) 2 . Cyber threat intelligence is also available on informal sources, such as public blogs, dark webs, forums, and social media platforms. Informal sources allow any person or entity on the Internet to publish, in real-time, the threat information in natural language, or unstructured data format. The unstructured and publicly available threat intelligence is also called Open Source Intelligence (OSINT) [3]. Cyber security-related OSINT are early warning sources for cyber security events such as security vulnerability exploits [4]. To conduct a cyber-attack, malicious actors typically have to 1) identify vulnerabilities, 2) acquire the necessary tools and tradecraft to successfully exploit them, 3) choose a target and recruit participants, 4) create or purchase the infrastructure needed, and 5) plan and execute the attack. Other actors— system administrators, security analysts, and even victims— may discuss vulnerabilities or coordinate a response to attacks [5]. These activities are often conducted online through social media, (open and dark) Web forums, and professional blogs, leaving digital traces behind. Collectively, these digital traces provide valuable insights into evolving cyber threats and can signal a pending or developing attack well before the malicious activity is noted on a target system. For example, exploits are discussed on Twitter before they are publicly disclosed [4] and on darkweb forums even before they are discussed on social media [6]. Open Source Intelligence (OSINT) is intelligence gathered from public-available sources such as social network sites, forums, wikis, blogs, and so on [7]. Malicious actors, system administrators, security analysts, and victims of cyber attacks usually use such platforms to discuss vulnerabilities, and exploits or to coordinate a response to attacks. Although more difficult to consume due to the volume and unstructured format of the content, data obtained from OSINT sources can complement intelligence obtained from structured intelligence sources, which usually provide malicious IP addresses and hashes, for example, as indicators of compromise (IOCs) that must be monitored or blocked by security platforms. Among OSINT sources available, we choose Twitter due to its ability to act as a natural aggregator of multiple sources [8] and its big data characteristics: a large volume of data, a highly diverse pool of users, high accessibility, and, mainly, timely production of new content [9]. The popularity of this medium in the cybersecurity community provides an environment for both offensive and defensive practitioners to discuss, report, and advertise timely indicators of vulnerabilities, attacks, malware, and other types of cyber events that are of interest to security analysts. In the past decade, Twitter has become an important source of intelligence. The real-time nature of information on Twitter has allowed researchers to use the microblog to extract intelligence about different areas such as terrorist attacks [10], earthquakes [11], forest fires [12] and so on. The value of Twitter with regards to security is well-demonstrated by the numerous initial reports of cyber events, examples of which include disclosures of multiple 0-day, user reports on DDoS attacks, and exposure of ransomware campaigns. For example, in June 2017, the global ransomware outbreak of ’Petya/NotPetya’ was discussed widely via Twitter before being reported by mainstream media [13]. Another more recent example of cyber threat initially discussed in Twitter was Log4Shell. Log4Shell was the name given to a 0-day exploit to a vulnerability in Log4j2 (CVE2021-44228), a popular Java logging library. The Log4j2 vulnerability along with a link to the exploit code, which means the code able to take advantage of a vulnerability in an easy way, was disclosed by the profile @P0rZ9 on December 9th, 2021, on Twitter. Following this post, hundreds of Twitter profiles, including independent researchers and journalists specialized in cyber security, started to post about the vulnerability. Given this strong and constant presence of the cyber security community in Twitter, over the recent years, the research on Twitter-based OSINT collection has led to the proposal of multiple frameworks [14], [15], [7], [16], [17] for detection and analysis of threat indicators in the Twitter stream. The shortness of tweets, which nowadays is a text of 280 maximum characters, is considered one of the main challenges when classifying tweets using machine learning algorithms [18]. In contrast with large document corpora, analyzing short documents such as tweets presents some specific semantic challenges towards extracting terms, relationships, patterns, and actionable insights in general [19].