Ewa Niewiadomska-Szynkiewicz*
Martyna Różycka**
Katarzyna Staciwa***
Katarzyna Nyczka****

System wspomagający wykrywanie treści wizualnych i tekstowych zagrażających bezpieczeństwu dzieci w cyberprzestrzeni[1]

Streszczenie

W ostatnich latach obserwuje się istotny wzrost zagrożeń bezpieczeństwa dzieci w cyberprzestrzeni. Do tych o największym ciężarze gatunkowym należą angażowanie dzieci w nielegalne zachowania online (np. uwodzenie, nagabywanie czy szantaż na

* Prof. dr hab. inż. Ewa Niewiadomska-Szynkiewicz, kierownik Zespołu Złożonych Systemów, Instytut Automatyki i Informatyki Stosowanej, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska, e-mail: ewa.szynkiewicz@pw.edu.pl, ORCID: 0000-0003-4782-3816.
** Mgr Martyna Różycka, kierownik Działu Reagowania na Nielegalne Treści w Internecie Dyżurnet.pl w Centrum Cyberbezpieczeństwa i Infrastruktury, Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy, e-mail: martyna.rozyka@nask.pl.
*** Mgr Katarzyna Staciwa, kierownik Zespołu III Dyżurnet.pl, Dział Reagowania na Nielegalne Treści w Internecie Dyżurnet.pl, Centrum Cyberbezpieczeństwa i Infrastruktury, Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy, e-mail: katarzyna.staciwa@nask.pl, ORCID:0000-0003-0633-4696.
**** Mgr inż. Katarzyna Nyczka, kierownik Zespołu II Dyżurnet.pl, Dział Reagowania na Nielegalne Treści w Internecie Dyżurnet.pl, Centrum Cyberbezpieczeństwa i Infrastruktury, Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy, e-mail: katarzyna.nyczka@nask.pl.
1 Praca finansowana przez Narodowe Centrum Badań i Rozwoju w ramach projektu nr: CYBERSECIDENT/455132/ III/NCBR/2020.

tle seksualnym), a także wytwarzanie nacechowanych seksualnie treści z ich udziałem. W tej sytuacji podstawowego znaczenia nabiera budowanie wśród najmłodszych członków naszego społeczeństwa świadomości cyberzagrożeń oraz nabywanie przez nich umiejętności bezpiecznego korzystania z przypisanych cyberprzestrzeni produktów i usług. Podstawowym działaniem na rzecz skutecznej ochrony dzieci w tym środowisku jest także wczesne wykrywanie i zgłaszanie odpowiednim organom występujących w nim przypadków nielegalnych zachowań i treści. Ważną rolę odgrywają zespoły takie jak Dyżurnet.pl[2], do którego zadań należy obecnie reagowanie na zgłoszone przez użytkowników cyberprzestrzeni potencjalnie nielegalne treści, a w najbliższej przyszłości być może także prowadzenie proaktywnych działań w tym obszarze. Doświadczenia Dyżurnet.pl jednoznacznie pokazują, że skuteczne wykrywanie takich treści wymaga automatyzacji działań i odpowiednich narzędzi informatycznych. W artykule został prezentowany nowatorski system monitorowania sieci i wspomagania decyzji wykorzystujący metody sztucznej inteligencji, w tym uczenia głębokiego do automatycznego wykrywania potencjalnie szkodliwych materiałów takich, jak: treści przedstawiające wykorzystywanie seksualne dzieci (Child Sexual Abuse Material – CSAM), treści erotyczne z udziałem dzieci, treści pornograficzne z wytworzonym lub przetworzonym obrazem dziecka oraz stanowiące pornografię z udziałem dorosłych.

Słowa kluczowe: bezpieczeństwo cyberprzestrzeni, Child Sexual Abuse Material, CSAM, system wspomagania decyzji, metody sztucznej inteligencji, uczenie maszynowe, uczenie głębokie

Problem

Powstanie globalnej sieci komputerowej internet oraz towarzyszący mu rozwój technologii cyfrowych (Information and Communication Technologies – ICTs) spowodowały istotne zmiany kulturowe obejmujące zjawiska społeczne, ekonomiczne i polityczne. Zmienił się diametralnie sposób komunikowania z otoczeniem, zdobywania wiedzy, współdziałania członków społeczeństwa. Przestrzeń wirtualna, nazywana także cyberprzestrzenią, coraz silniej integruje się ze światem rzeczywistym, a produkty i usługi cyfrowe stały się codziennością. Dogłębne wyjaśnienie funkcjonowania cyberprzestrzeni nie jest celem tego artykułu. W celu uporządkowania terminologicznego warto przytoczyć definicję Janusza Wasilewskiego. Według niego „[...] istotę cyberprzestrzeni tworzy koncepcja powołania do życia swojego rodzaju równoległego środowiska, które jest nowym wymiarem dla ludzkich działań. Wymiar ten, z uwagi na sposób budowy, jest jednak obszarem wymykającym się opisowi za pomocą typowych, fizycznych miar, nie poddaje się zatem prostemu podziałowi geograficznemu pomiędzy państwa”[3].

2 Więcej zob. *O nas*, https://dyzurnet.pl/o-nas [dostęp: 17.04.2023].

3 J. Wasilewski, *Zarys definicyjny cyberprzestrzeni*, „Przegląd Bezpieczeństwa Wewnętrznego” 2013, nr 9, s. 231.

Do polskiego porządku prawnego definicję cyberprzestrzeni wprowadzono w art. 3 pkt 4 ustawy z 18 kwietnia 2002 roku o stanie klęski żywiołowej[4] oraz art. 2 ust. 1b ustawy z 29 sierpnia 2002 roku o stanie wojennym oraz o kompetencjach Naczelnego Dowódcy Sił Zbrojnych i zasadach jego podległości konstytucyjnym organom Rzeczypospolitej Polskiej[5]. Na ich podstawie cyberprzestrzeń powinna być rozumiana jako „[...] przestrzeń przetwarzania i wymiany informacji tworzoną przez systemy teleinformatyczne, określone w art. 3 pkt 3 ustawy z dnia 17 lutego 2005 roku o informatyzacji działalności podmiotów realizujących zadania publiczne[6], wraz z powiązaniami pomiędzy nimi oraz relacjami z użytkownikami".

Ogromną zaletą cyberprzestrzeni jest to, że jej użytkownicy mogą komunikować się ze sobą w dowolnym czasie, praktycznie z każdego miejsca na świecie. W czerwcu 2022 roku szacowana liczba użytkowników internetu wynosiła prawie 5,4 mld. Stanowili oni 67,9% globalnej populacji[7]. Urządzenia mobilne i media społecznościowe są stałym elementem codziennego życia ludzi na całym świecie, a dzieci i młodzież spędzają coraz więcej czasu online. Potwierdzają to wyniki ogólnopolskigo badania „Nastolatki 3.0"[8] przeprowadzonego przez NASK-PIB w grudniu 2020 roku. Stały wzrost liczby godzin przeznaczanych przez małoletnich respondentów na korzystanie z internetu jest widoczny od pierwszych edycji badania. Obecnie jest to nawet 5 godzin dziennie, a w dni wolne od zajęć szkolnych ponad 6 godzin[9].

Aktywność w cyberprzestrzeni oznacza także, że najmłodsi członkowie naszego społeczeństwa są narażeni na rozmaite zagrożenia. W tym kontekście zaprezentowane w raporcie „Nastolatki 3.0" ustalenia są co najmniej niepokojące. Wskazują one, że co piąty polski nastolatek doświadczył przemocy w cyberprzestrzeni polegającej najczęściej na: wyzywaniu (29,7%), ośmieszaniu

4 Ustawa z dnia 18 kwietnia 2002 roku o stanie klęski żywiołowej, t.j., Dz.U. 2017, poz. 1897.

5 Ustawa z dnia 29 sierpnia 2002 roku o stanie wojennym oraz o kompetencjach Naczelnego Dowódcy Sił Zbrojnych i zasadach jego podległości konstytucyjnym organom Rzeczypospolitej Polskiej, t.j., ibidem 2022, poz. 2091.

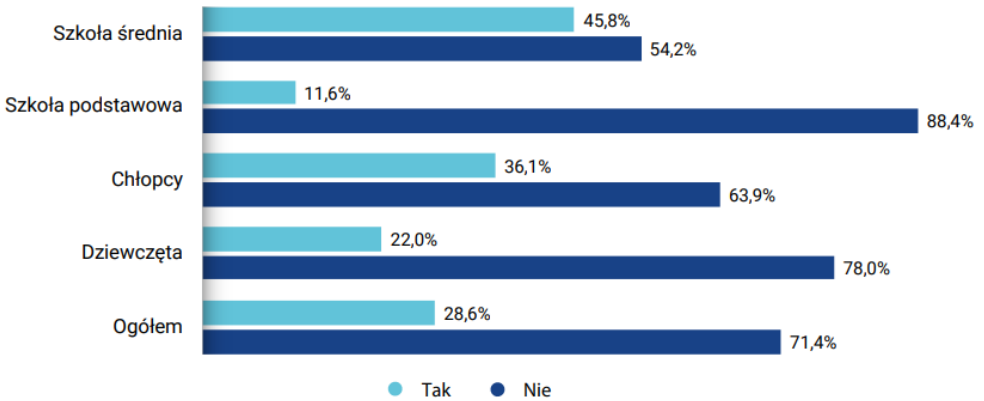
6 Ustawa z dnia 17 lutego 2005 roku o informatyzacji działalności podmiotów realizujących zadania publiczne, t.j., ibidem 2023, poz. 57.

7 *Internet Usage Statistics. The Internet Big Picture World Internet Users and 2023 Population Stats*, https://www.internetworldstats.com/stats.html [dostęp: 20.04.2023].

8 *Raport z badań „Nastolatki 3.0"*, 30.09.2021, https://www.nask.pl/pl/raporty/raporty/4295,RAPORT-Z-BADAN-NASTOLATKI-30-2021.html [dostęp: 20.04.2023].

9 Ibidem, s. 6.

(22,8%) czy poniżaniu (22%)[10]. Równie niepokojące są ustalenia dotyczące oglądania przez nastolatków treści pornograficznych dostępnych w tym środowisku, co zadeklarował co czwarty ankietowany (28,6%)[11]. Należy zwrócić uwagę na różnice występujące w tych deklaracjach ze względu na płeć. Wśród dziewcząt była to co piąta ankietowana (22,0%), a wśród chłopców co trzeci respondent (36,1%)[12]. Inną istotną obserwacją była ta, która dotyczyła przyrostu doświadczeń związanych z pornografią w cyberprzestrzeni. Ustalono, że lawinowy przyrost takich doświadczeń następuje w ciągu kilku lat, pomiędzy 11–12 a 16–17 rokiem życia (11,6% – szkoła podstawowa; 45,8% – szkoła średnia)[13]. Prezentuje to wykres 1.



Źródło: *Raport z badań „Nastolatki 3.0"...*

Wykres 1. Odpowiedzi na pytanie: „Czy zdarzyło Ci się oglądać jakieś treści pornograficzne w internecie?" według płci i typu szkoły

Do zagrożeń o największym ciężarze gatunkowym dla dzieci, charakterystycznych dla cyberprzestrzeni, należy angażowanie dzieci w nielegalne zachowania online (np. uwodzenie, nagabywanie czy szantaż na tle seksualnym) oraz wytwarzanie nacechowanych seksualnie treści z ich udziałem, takich jak CSAM. Przywołany tutaj podział jest istotny z uwagi na cel artykułu, którym jest prezentacja systemu wykorzystującego metody sztucznej inteligencji do automatycznego wykrywania treści należących do drugiej z wymienionych kategorii zagrożeń.

10 Ibidem.
11 Ibidem, s. 98.
12 Ibidem, s. 101.
13 Ibidem.

Dokładne określenie ilości treści z kategorii CSAM dostępnych obecnie w cyberprzestrzeni nie jest możliwe. Szacunkowe dane można pozyskać od wyspecjalizowanych w tym obszarze podmiotów. Dla przykładu, zdaniem ekspertów z organizacji WeProtect Global Alliance[14] zebrane przez nich dowody wskazują na wzrost od 2019 roku ilości dostępnych w cyberprzestrzeni materiałów przedstawiających wykorzystywanie seksualne dzieci[15]. Ten kierunek potwierdzają dane publikowane przez inną organizację, legitymującą się szczególnym mandatem, tj. National Center for Missing & Exploited Children (NCMEC)[16]. Zgodnie z amerykańskim prawem federalnym podmioty lokalnego sektora prywatnego mają obowiązek zgłaszania do infolinii CyberTipline (hotline) przypadków wykorzystywania seksualnego dzieci wykrytych w ich produktach i usługach. Warto podkreślić, że wśród podmiotów tego sektora są największe na świecie platformy społecznościowe, tj. Facebook (21,1 mln wysłanych raportów), Instagram (5 mln wysłanych raportów), Google (2,1 mln wysłanych raportów), Whatsapp (1,1 mln wysłanych raportów). Prezentowane dane dotyczą 2021 roku. CyberTipline to infolinia należąca do NCMEC i zrzeszona w stowarzyszeniu INHOPE[17]. Publikowane przez NCMEC dane są alarmujące: w 2021 roku nastąpił kolejny wzrost liczby zgłoszeń – do 29,3 mln, co stanowi 35% przyrost w stosunku do roku 2020, przy czym ponad 99% tych zgłoszeń dotyczyło incydentów z podejrzeniem CSAM[18]. Zgodnie z najnowszym raportem opublikowanym w 2022 roku liczba zgłoszeń wzrosła o kolejne 9% i wynosiła 32,1 mln raportów;

14 WeProtect Global Alliance to organizacja powstała z połączenia Global Alliance Against Child Sexual Abuse Online (powstała w 2012 roku na szczeblu międzynarodowym) oraz We Protect (utworzona w 2014 roku w Wielkiej Brytanii). Za jej pośrednictwem Komisja Europejska wspiera koordynację w obszarze zapobiegania i zwalczania wykorzystywania seksualnego dzieci na poziomie globalnym. Zob. *Who we are*, https://www.weprotect.org [dostęp: 20.04.2023].

15 Ch. Setter, N. Greene, N. Newman, J. Perry, *Global Threat Assessment 2021*, s. 5, https://www.weprotect.org/global-threat-assessment-21/#repor [dostęp: 20.04.2023].

16 National Center for Missing & Exploited Children – amerykańska organizacja non-profit, której misją jest niesienie pomocy w poszukiwaniu zaginionych dzieci oraz ograniczenie zjawiska ich seksualnego wykorzystywania. Zob. *Who we are*, https://www.missingkids.org/home [dostęp: 20.04.2023].

17 International Association of Internet Hotlines – działająca od 1999 roku organizacja zrzeszająca obecnie 50 infolinii działających w różnych częściach świata. Każda infolinia udostępnia mechanizm anonimowego zgłaszania przypadków ujawnienia materiałów z kategorii CSAM dostępnych w cyberprzestrzeni. Zob.: https://inhope.org/ [dostęp: 9.12.2022].

18 *CyberTipline 2021 Report*, https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata#overview [dostęp: 20.04.2023].

90% raportów wskazywało na regiony poza Stanami Zjednoczonymi, 235 tys. miało geolokalizację polską[19].

Brytyjska hotline również zrzeszona w INHOPE, tj. Internet Watch Foundation (IWF), wskazuje na inne, niepokojące zjawisko. Jest to zaobserwowany w latach 2019–2020 aż 77% wzrost treści o charakterze seksualnym, wytwarzanych i publikowanych online przez dzieci[20]. Na podstawie tych materiałów eksperci z IWF stwierdzili, że efekty niedawnego, przymusowego odizolowania dzieci i młodzieży, które miało miejsce podczas pandemii COVID-19, zaczynają mieć odzwierciedlenie w zgłaszanych tej infolinii treściach. Dane opublikowane przez tę infolinię są wstrząsające. W 2022 roku 63 050 raportów dotyczyło treści wytworzonych przez dzieci w wieku 7–10 lat, przy czym w wielu przypadkach miało to miejsce w wyniku uwodzenia przez sprawcę albo stosowania szantażu seksualnego[21].

Z kolei z raportów rocznych zespołu Dyżurnet.pl wynika, że w 2019 roku z ogólnej liczby 2157 zgłoszeń CSAM 9% zostało zaliczonych do kategorii materiałów intymnych wytworzonych samodzielnie przez osobę małoletnią. W 2020 roku materiały te stanowiły 14% z 2517 zgłoszeń CSAM, a w roku 2021 – 8% z 2069 zgłoszeń CSAM[22]. Zjawisku temu została poświęcona kampania „Nie na pokaz”[23] realizowana przez NASK-PIB i TikTok Polska w ramach działań Polskiego Centrum Programu Safer Internet (PCPSI).

Charakterystykę zagrożenia w postaci obecności materiałów z kategorii CSAM w cyberprzestrzeni należy zakończyć przywołaniem perspektywy uwidocznionych na nich osób. W ich przypadku można mówić o wtórnej wiktymizacji spowodowanej niekończącą się dystrybucją CSAM w cyberprzestrzeni. Zgodnie z wynikami ankiety przeprowadzonej w 2017 roku przez Canadian Centre for Child Protection[24] taka dystrybucja pogłębia wiktymizację dzieci

19 *CyberTipline 2022 Report*, https://www.missingkids.org/cybertiplinedata [dostęp: 20.04.2023].

20 *Trend: 'Self-generated' content*, https://annualreport2020.iwf.org.uk/trends/international/selfgenerated [dostęp: 21.03.2023].

21 *Sexual abuse imagery of primary school children 1,000 per cent worse since lockdown*, https://www.iwf.org.uk/news-media/news/sexual-abuse-imagery-of-primary-school-children-1-000-per-cent-worse-since-lockdown/ [dostęp: 21.03.2023].

22 *Analiza wyników badania dotyczącego treści intymnych publikowanych przez młodzież*, s. 9, https://dyzurnet.pl/publikacje [dostęp: 21.03.2023].

23 *Nie na pokaz. Mówimy „nie” publikowaniu treści intymnych*, https://www.saferinternet.pl/nie-na-pokaz/kampania.html [dostęp: 21.03.2023].

24 *International Survivor's Survey*, https://www.protectchildren.ca/en/resources-research/survivors-survey-results/ [dostęp: 17.04.2023].

i młodzieży i ma długotrwały, szkodliwy wpływ nawet wtedy, kiedy osiągną już dorosłość. Około 70% członków takiej populacji niezmiennie obawia się bycia rozpoznanym w życiu codziennym. Obowiązkiem dzisiejszego społeczeństwa jest nie tylko ochrona dzieci przed wykorzystywaniem seksualnym w rzeczywistym świecie, lecz także działanie polegające na usuwaniu z cyberprzestrzeni materiałów audiowizualnych dokumentujących przemoc seksualną, której dzieci i młodzież były ofiarami.

Zauważ i usuń

Jeżeli chodzi o krajową odpowiedź na problem wykorzystywania seksualnego dzieci w cyberprzestrzeni, to w pierwszej kolejności należy nawiązać do doświadczeń Dyżurnet.pl. Jest to jedyny w Polsce, działający od blisko 18 lat, punkt kontaktowy do zgłaszania nielegalnych treści w internecie. Uwaga koncentruje się na treściach przedstawiających wykorzystywanie seksualne dzieci. W 2018 roku działanie tego zespołu zostało umocowane w ustawie o krajowym systemie cyberbezpieczeństwa[25]. Zapisano w niej szczególne zadanie przypisane instytutowi NASK-PIB, który został ponadto wskazany jako jeden z zespołów reagowania na incydenty komputerowe (Computer Security Incident Response Team – CSIRT). To szczególne zadanie polega na „[...] zapewnieniu obsługi linii telefonicznej lub serwisu internetowego prowadzących działalność w zakresie zgłaszania i analizy przypadków dystrybucji, rozpowszechniania lub przesyłania pornografii dziecięcej za pośrednictwem technologii informacyjno-komunikacyjnych”[26]. Procedurą reagowania przez Dyżurnet.pl są objęte następujące treści:

– przedstawiające seksualne wykorzystywanie dziecka: art. 202 par. 3, 4, 4a, 4b ustawy z 6 czerwca 1997 r. – Kodeks karny (k.k.);
– przedstawiające tzw. twardą pornografię: art. 202 par. 3 k.k.;
– propagujące rasizm i ksenofobię: art. 256 k.k.;
– inne nielegalne, tj. te, które nie należą do żadnej z powyższych kategorii, ale zagrażają bezpieczeństwu dzieci, np. propagowanie lub pochwalanie zachowań o charakterze pedofilskim (art. 200b k.k.), uwodzenie małoletniego

25 Ustawa z dnia 5 lipca 2018 r. o krajowym systemie cyberbezpieczeństwa, t.j., Dz.U. 2022, poz.1863.
26 Ibidem, art. 26, ust. 6, pkt 3.

poniżej 15 roku życia przez internet (art. 200a k.k.), zjawisko szantażu na tle seksualnym (określane również jako sextortion).

Użytkownicy cyberprzestrzeni mogą dokonywać zgłoszeń na kilka sposobów, tj. przez formularz znajdujący się na stronie internetowej: www.dyzurnet.pl, skrzynkę poczty elektronicznej: dyzurnet@dyzurnet.pl, automatyczną infolinię: 801 615 005, a od 2020 roku również przez wtyczkę do przeglądarek Firefox i Chrome.

Lokalizacja serwera, na którym znajdują się treści z kategorii CSAM, jest warunkiem przyjęcia jednego z dwóch scenariuszy postępowania. Jeżeli treści znajdują się na serwerze zlokalizowanym w Polsce albo poza jej terytorium, ale w kraju, w którym nie działa infolinia zrzeszona w INHOPE, to informacja o nich jest przekazywana do Komendy Głównej Policji w Warszawie na adres: cyber-kgp@policja.gov.pl, oraz do Interpolu[27]. Jeżeli serwer został zlokalizowany poza terytorium Polski, ale na terenie kraju, w którym działa infolinia zrzeszona w INHOPE, to właśnie do niej i do Interpolu trafia stosowna informacja.

Powiadomienie Interpolu odbywa się za pośrednictwem systemu ICCAM udostępnionego członkom stowarzyszenia IHOPE w postaci portalu internetowego – I see Child Abuse Material, który został sfinansowany z funduszy unijnych i uruchomiony w 2015 roku przez INHOPE, we współpracy z firmą prywatną Ziuz Forensics. Do najważniejszych zadań systemu należy umożliwianie klasyfikacji treści ze względu na ich cechy, np. płeć oraz przybliżony wiek uwiecznionej na nich osoby, oraz przesłanie sklasyfikowanego pliku do bazy prowadzonej przez Interpol – ICSE DB. Na podstawie wykonanej analizy do ICSE DB[28] przekazywane są treści sklasyfikowane jako uznawane za nielegalne we wszystkich państwach współpracujących z Interpolem (baseline), a także te uznawane za nielegalne w kraju działania infolinii otrzymującej zgłoszenie (national)[29]. Kryterium zaklasyfikowania treści do kategorii baseline jest następujące: zdjęcie lub film bez jakichkolwiek wątpliwości przedstawia obraz prawdziwego dziecka, w okresie przedpokwitaniowym lub na bardzo wczesnym etapie dojrzewania płciowego, jeżeli wiek dający się określić na podstawie oceny wizualnej nie przekracza 13 roku życia, to uczestniczącego

27 Interpol – Międzynarodowa Organizacja Policji Kryminalnych. Zob. https://www.interpol.int/

28 ICSE DB – International Child Sexual Exploitation database, międzynarodowa baza zawierająca treści przedstawiające seksualne wykorzystywanie dzieci.

29 The Association of Internet Hotline Providers, Międzynarodowe Stowarzyszenie Internetowych Zespołów Reagujących. Zob. https://www.inhope.org [dostęp: 23.04.2023].

lub będącego świadkiem seksualnej aktywności albo obraz jest zogniskowany na rejon genitalny lub analny tego dziecka. Do pełnej klasyfikacji treści jako baseline niezbędna jest niezależna, pozytywna ocena trzech wyspecjalizowanych funkcjonariuszy.

Jeżeli na podstawie wstępnej klasyfikacji analityka infolinii treści znajdujące się na zgłoszonej stronie internetowej zostały uznane ze nielegalne, to adres URL takiej strony jest przekazywany do systemu ICCAM. Następnie jest dokonywane automatyczne przeszukanie wszystkich informacji znajdujących się pod tym adresem, nadanie wartości hash każdemu zdjęciu lub filmowi wideo, a także ustalenie lokalizacji serwera. Wartość hash jest następnie porównywana z listami innych wartości hash będącymi częścią systemu ICCAM. Jeżeli wartości hash nowo zgłoszonych treści nie pasują do żadnej z tych list, to podlegają one indywidualnej klasyfikacji przez analityka, który nadaje im jedną z trzech kategorii: baseline, nielegalne w kraju pracy analityka (national) lub legalne w tymże kraju. W Polsce analitycy Dyżurnet.pl posługują się w swoich działaniach podziałem na treści definiowane jako „treści pornograficzne z udziałem małoletniego" (art. 202 par. 3, 4, 4a, 4b k.k.) oraz „treści prezentujące dziecko w kontekście seksualnym" takie jak nacechowane seksualnie pozowanie.

Informatyczny system wspomagania decyzji APAKT

Podstawowym działaniem w celu skutecznej ochrony dzieci przed zagrożeniami w cyberprzestrzeni jest wczesne wykrywanie i zgłaszanie przypadków zamieszczania materiałów zawierających nielegalne zachowania i treści. Zespół Dyżurnet.pl otrzymuje coraz więcej zgłoszeń o zaistnieniu tego typu zdarzeń. Zgłoszenia te wymagają weryfikacji, co istotnie angażuje pracowników. Ze względu na łatwość umieszczania danych w internecie i lawinowy wręcz przyrost niepożądanych treści manualny monitoring przestaje być skuteczny. Konieczna jest automatyzacja działań i wyposażenie zespołów hotline w nowoczesne narzędzia informatyczne do bieżącego przeglądania i śledzenia zasobów internetu oraz wspomagania klasyfikacji treści na legalne i niepożądane. Dlatego w 2019 roku powstał pomysł budowy informatycznego systemu wspomagania decyzji wykorzystującego najnowsze rozwiązania z zakresu oprogramowania oraz algorytmy uczenia maszynowego do analizowania zasobów cyberprzestrzeni i wykrywania oraz weryfikacji prób propagowania w sieci nielegalnych i wrażliwych treści.

System APAKT (Automatyczne Przeszukiwanie, Analiza i Klasyfikacja Treści) jest tworzony w ramach projektu badawczo-rozwojowego współfinansowanego przez Narodowe Centrum Badań i Rozwoju. W realizację projektu są zaangażowane trzy instytucje: NASK – Państwowy Instytut Badawczy pełniący funkcje koordynatora, Politechnika Warszawska (PW) oraz firma Enamor International Sp. z o.o.

Głównym zadaniem systemu APAKT jest analiza treści multimedialnych zamieszczonych w internecie oraz wskazanie materiałów przedstawiających wykorzystywanie seksualne dzieci zarówno tych sklasyfikowanych w przeszłości, jak i zupełnie nowych. Zakłada się, że system, działając na serwerach zespołu Dyżurnet.pl, będzie wspomagał pracowników zespołu w szybkiej identyfikacji nielegalnych i szkodliwych treści. Preklasyfikacja i priorytetyzacja treści będzie wykonywana w sposób automatyczny z wykorzystaniem odpowiednich algorytmów, m.in. uczenia maszynowego. Wynikiem systemu będzie ocena analizowanego materiału i przypisanie go do jednej z klas reprezentujących dane o różnym stopniu szkodliwości. Ostateczną decyzję o tym, czy dany materiał jest szkodliwy i nielegalny będzie podejmował człowiek.

Schemat klasyfikacji w systemie powstał w wyniku identyfikacji charakterystycznych cech rozpoznanych w klasyfikowanych materiałach. Przed przystąpieniem do procesu pozyskiwania i gromadzenia danych dokonano gruntownej analizy obrazów pod kątem niezbędnych elementów decydujących o ich nielegalnym charakterze. Próbowano również zidentyfikować grupy obrazów posiadające wspólne cechy. W rezultacie oprócz głównego podziału treści CSAM w systemie na baseline i national wyodrębniono kilka podklas. Przyjęto nomenklaturę angielskojęzyczną ze względu na międzynarodowy charakter klasyfikacji i ciągłą konieczność wymiany informacji z pozostałymi zespołami hotline zrzeszonymi w INHOPE. Wspomniane podklasy to:

1. Minors & Adults (małoletni i dorośli) – czynności seksualne wykonywane przez małoletniego z dorosłym lub w obecności dorosłego;

2. Minors Only (wyłącznie małoletni) – czynności seksualne podejmowane przez małoletniego bez udziału i obecności dorosłego;

3. In Presence of a Minor (w obecności małoletniego) – obecność małoletniego (bez jakiejkolwiek aktywności seksualnej) w trakcie czynności seksualnych między dorosłymi;

4. Focus (fokus) – brak czynności seksualnych małoletniego, materiał przedstawiający tzw. fokus na jego genitalia lub obszar analny – fokus oznacza, że uwaga osoby patrzącej na zdjęcie koncentruje się na genitaliach lub obszarze analnym małoletniego;

5. Sexual Posing (pozowanie seksualne) – pozowanie seksualne małoletniego – brak czynności seksualnych podejmowanych przez małoletniego, przyjęta pozycja ciała lub sposób wykonania zdjęcia/filmu ma na celu ekspozycję genitaliów małoletniego;

6. Exploitative Nudity (nagość z elementami wykorzystywania) – widoczne genitalia dziecka (bez pozowania seksualnego) oraz możliwe pozowanie erotyczne lub widoczny seksualny kontekst zdjęcia;

7. Other CSAM (pozostały, niesklasyfikowany CSAM) – treści wytworzone lub przetworzone cyfrowo oraz pozostałe treści niedające się sklasyfikować w wyżej wymienionych kategoriach.

Oprócz danych typu CSAM, uznanych w polskim systemie prawnym za nielegalne, wyodrębniono grupy treści mających charakter legalny (Not CSAM), posiadających jednak cechy pokrewne do treści będących przedmiotem prac w projekcie. Należą do nich:

- Child Erotism (erotyka dziecięca);
- Child Nudity (nagość dzieci);
- Adult Pornography (pornografia dorosłych).

Oprócz schematu klasyfikacji treści wyodrębniono wiele charakterystycznych cech kategoryzujących dane gromadzone w systemie, przydatnych z punktu widzenia zarówno analizy treści, jak i późniejszego postępowania identyfikującego sprawcę oraz pokrzywdzonego. Niektóre z cech zapisywanych w systemie to:

- wiek/poziom dojrzałości seksualnej małoletniego;
- płeć najmłodszych osób widocznych na zdjęciu;
- rodzaj wykonywanej przez małoletniego czynności seksualnej;
- rodzaj interakcji seksualnej widocznej na zdjęciu (dorosły-małoletni, małoletni-małoletni, małoletni samodzielnie);
- widoczność genitaliów;
- pozowanie erotyczne;
- treść wytworzona samodzielnie przez małoletniego (tzw. *self-generated*).

Powyższa klasyfikacja i kategoryzacja jest kompatybilna z nowym systemem klasyfikacji treści wypracowanym w ramach projektu „The Global Standard Project" realizowanego przez Stowarzyszenie INHOPE ze środków End Violence Against Children. Celem projektu było stworzenie globalnego systemu klasyfikacji i kategoryzacji treści CSAM umożliwiającego wymianę informacji i automatyczne tłumaczenie oraz mapowanie schematów klasyfikacyjnych funkcjonujących w poszczególnych systemach legislacyjnych lub w organizacjach zajmujących się walką z CSAM. Efektem prac grupy roboczej

zaangażowanej w tworzenie uniwersalnej klasyfikacji, w której udział brał również zespół Dyżurnet.pl, jest dokument „Universal Classification Schema". Ma on stanowić bazę do stworzenia skuteczniejszego systemu wymiany informacji między instytucjami oraz organizacjami na świecie.

Architektura systemu APAKT

APAKT to nowatorski system oprogramowania wykorzystujący skutecznie metody sztucznej inteligencji do analizy danych różnego typu. Składa się z trzech podstawowych komponentów:

– repozytorium danych zawierające zdjęcia, materiały wideo (docelowo audio-wideo) oraz teksty podejrzane o zawieranie szkodliwych i niedozwolonych treści dotyczących wykorzystania seksualnego nieletnich;

– biblioteki algorytmów do klasyfikacji zdjęć, materiałów wideo oraz tekstów na klasy zdefiniowane przez zespoły hotline;

– środowisko oprogramowania oferujące usługi zarządzania modułami i wymianą danych między nimi oraz usługi raportowania i komunikacji z użytkownikiem.

Najważniejszym komponentem systemu jest biblioteka klasyfikatorów stosujących metody sztucznej inteligencji, w tym uczenia głębokiego do automatycznego wykrywania potencjalnie szkodliwych materiałów, m.in.: zdjęć, filmów wideo i tekstów. Opracowując i projektując algorytmy rozpoznawania wzorców i modeli sieci neuronowych[30], naukowcy z NASK i PW przeprowadzili dogłębną analizę rozwiązań omawianych w literaturze. Wykonane w ramach projektu i oferowane w systemie APAKT autorskie narzędzia wykorzystują rozbudowane, wstępnie wytrenowane modele sztucznych sieci neuronowych dostępne w internecie oraz biblioteki modeli sieci dostępne w bibliotekach TensorFlow[31] i PyTorch[32]. Ostatecznie system APAKT oferuje cztery moduły do klasyfikacji materiałów pobieranych z internetu i wspomagania procesu podejmowania decyzji o kwalifikacji danych: dwa alternatywne moduły do analizy i klasyfikacji zdjęć, moduł analizy i klasyfikacji materiałów wideo oraz moduł klasyfikacji tekstów. W przyszłości planuje się dołączenie dodatkowego modułu do klasyfikacji materiałów audio.

30 Ch.C. Aggarwal, *Neural Networks and Deep Learning*, Cham 2018.
31 TensorFlow zob. https://www.tensorflow.org/?hl=pl [dostęp: 23.04.2023].
32 PyTorch zob. https://pytorch.org/ [dostęp: 23.04.2023].

Analiza i klasyfikacja zdjęć

Zadaniem klasyfikatora obrazu jest wykrywanie seksualnego wykorzystania dzieci na podstawie analizy zdjęć zamieszczanych w internecie. System APAKT udostępnia dwa alternatywne moduły przetwarzania i klasyfikacji obrazów. Wykorzystują one wstępnie wytrenowane, dostępne w internecie modele sieci neuronowych (np. model NudeNet[33]) oraz popularne algorytmy uczenia maszynowego, w tym binarną i wieloklasową maszynę wektorów nośnych C-SVM (C-Support Vector Machine)[34]. Do klasyfikacji obrazów na zdefiniowane wcześniej klasy są wykorzystywane następujące cechy: liczba osób na obrazie, różne ujęcia twarzy, części ciała oraz relacje między nimi, proporcje różnych części ciała, proporcje sylwetek. Oba wspomniane moduły klasyfikacji obrazów różnią się architekturą oraz zastosowanymi modelami sieci. Pierwszy zakłada fuzję wyjść modeli i klasyfikację do zdefiniowanych klas, w drugim rozważana jest hybrydowa architektura inspirowana pracą[35], w której decyzje podejmowane są w dwóch etapach – generowane są propozycje obiektów, które następnie są doprecyzowane, i wskazywana jest przynależność do klas. W przypadku obu rozwiązań pierwszym etapem procesu klasyfikacji jest estymacja widoczności obiektów związanych z poszczególnymi cechami sylwetki oraz detekcja części ciała osób występujących na analizowanych obrazach. Na podstawie uzyskanych wyników jest prowadzona dalsza, pogłębiona analiza. Ważnym klasyfikatorem wykorzystywanym przez oba wspomniane moduły oraz klasyfikator wideo jest model sieci określający wiek osoby na zdjęciu na podstawie widocznych narządów płciowych, proporcji sylwetki i ostatecznie obrazu twarzy. Ustalenie wieku, szczególnie rozróżnienie nastolatków od osób dorosłych, jest poważnym wyzwaniem.

Analiza i klasyfikacja materiałów wideo

Zadaniem klasyfikatora jest rozpoznanie aktywności człowieka w analizowanym strumieniu wideo. Autorzy rozwiązania opracowali dwa, operujące na

33 *NudeNet: Neural Nets for Nudity Classification, Detection and selective censoring*, https://pypi.org/project/NudeNet/ [dostęp: 23.04.2023].

34 C. Cortes, V. Vapnik, *Support-vector networks*, „Mach Learn" 1995, nr 20, s. 273–297.

35 K. He, G. Gkioxari, P. Dollar, R.B. Girshick, *Mask R-CNN*, arXiv:1703.06870 [cs.CV], 2018, https://doi.org/10.48550/arXiv.1703.06870 [dostęp: 20.04.2023].

różnych danych, algorytmy. Połączenie ich wyników pozwala na wykrywanie materiałów zawierających treści wskazujące na wykorzystanie seksualne dzieci. Wspomniane algorytmy to:

– klasyfikator niskopoziomowy analizujący pełne klatki obrazu;
– klasyfikator danych szkieletowych analizujący sylwetki osób i grup osób występujących w filmie.

Pierwszy algorytm identyfikuje zachowanie osób oraz ich wzajemne relacje na podstawie analizy sekwencji pełnych klatek filmu[36]. Dokonuje ekstrakcji i bada niskopoziomowe cechy obrazu takie, jak: surowe kanały RGB, cechy przepływu optycznego – surowe wektory oraz wektorowe operatory różniczkowe pola przepływu optycznego (m.in. rotacja, dywergencja). Ze względu na specyfikę rozpoznawanych sekwencji istotne jest również wykrywanie czynności o charakterze powtarzalnym. W tym celu badane są również cechy częstotliwościowe dla krótkich podsekwencji obrazów. Algorytm zwraca zagregowaną informację o klatce obrazu, która stanowi wejście do modelu sieci neuronowej.

W przypadku drugiego algorytmu uwaga koncentruje się na sylwetce osoby lub grupie sylwetek osób[37]. Analizowane są tylko wybrane części obrazu, te zawierające sylwetki. Sylwetki są następnie reprezentowane przez ich uproszczone szkielety[38]. O możliwych wzajemnych interakcjach osób decydują odległości wyznaczane między odpowiadającymi sobie punktami (węzłami) różnych sylwetek. Wynikiem działania algorytmu jest próba określenia aktywności obserwowanych elementów w krótkich fragmentach sekwencji. W tym przypadku dane o odległościach między sylwetkami stanowią również wejście do modelu sieci neuronowej.

Ostateczna klasyfikacja materiału wideo na CSAM i nie CSAM jest wynikiem fuzji danych wyjściowych modeli neuronowych o różnych architekturach i parametrach, w tym modelu określania wieku analizowanych osób.

36 A. Wilkowski, M. Kamola, W. Kasprzak, P. Piwowarski, B. Laskowska, *Human Activity Classification in Video*, PP-RAI, Łódź 2023.

37 W. Kasprzak, B. Jankowski, *Light-Weight Classification of Human Actions in Video with Skeleton-Based Features*, „Electronics" 2022, t. 11, nr 14.

38 W. Kasprzak, S. Puchała, P. Piwowarski, *On Multi-stream Classification of Two Person Interactions in Video with Skeleton-Based Features* [w:] *Computer Vision and Graphics. ICCVG 2022. Lecture Notes in Networks and Systems*, red. L.J. Chmielewski, A. Orłowski, t. 598, Cham 2023.

Analiza i klasyfikacja tekstów

Klasyfikacja materiałów tekstowych opiera się na analizie zebranych materiałów pod kątem rodzaju oraz funkcji klasyfikowanego tekstu. Kwalifikacja tekstów do klasy CSAM wymaga obecności opisów czynności seksualnych podejmowanych z małoletnimi. Klasyfikacja danych tekstowych jest zdecydowanie trudniejsza ze względu na możliwość niejednoznacznych opisów sytuacji, z których wiek osoby małoletniej możemy jedynie wnioskować na podstawie przytaczanych okoliczności, np.: szkoła, przedszkole, córka, dziewczynka, chłopczyk.

Przyjęty w projekcie schemat kategoryzacji treści tekstowych wyglądał następująco:

– rodzaj tekstu: rozmowa (chat), narracja, post, komentarz, podpis pod zdjęciem, inne, nieznane;

– funkcja tekstu dla treści CSAM: uwodzenie (grooming), narracja, promowanie pedofilii, inne;

– funkcja tekstu dla treści nie CSAM: seksualna lub brak.

W trakcie projektu udało się zbudować narzędzie do automatycznej klasyfikacji treści narracyjnych pod kątem obecności elementów seksualnych wraz ze wskazaniem prawdopodobnego wieku osób małoletnich.

Zastosowano dwuetapowy klasyfikator do wykrywania tego typu treści. W pierwszej fazie każde zdanie było klasyfikowane jako erotyczne lub neutralne. Jeżeli stosunek zdań erotycznych do wszystkich zdań w tekście był większy od empirycznie ustalonej granicy, to tekst był klasyfikowany jako erotyczny. W następnym kroku dokonywano rozróżnienia tekstu erotycznego od noszącego znamiona nielegalności z użyciem modelu opartego na cechach stylometrycznych. W rozwiązaniu zastosowano hybrydowy model semantyczno-gramatyczny z użyciem wektorów StyloMetrix. Taka metoda zapewnia wyjaśnialność otrzymanych predykcji.

Repozytorium danych trenujących

Baza danych jest ulokowana w środowisku produkcyjnym Dyżurnet.pl. Zawarte w niej dane są wykorzystywane do wytrenowania klasyfikatorów, dlatego muszą być one odpowiednio przygotowane. Klasyfikatory zaprojektowane w ramach projektu wykorzystują podejście z uczeniem nadzorowanym, czyli

tzw. uczenie z nauczycielem (supervised learning)[39]. Stąd potrzeba oznaczania każdego wykorzystywanego materiału, czyli oznaczania na nim interesujących dla klasyfikatora obiektów. Wstępnego przetworzenia i adnotacji danych dokonują specjaliści, tj. pracownicy hotline, często ze wsparciem biegłych sądowych. Wyodrębniają i oznaczają interesujące, z punktu widzenia przyszłej klasyfikacji, obiekty na zdjęciach lub filmach, części tekstów, w tym opowiadań z udziałem dzieci o charakterze erotycznym czy pornograficznym. Efektem prac są zestawy danych uczących z przypisanymi odpowiednio etykietami. Obecnie baza danych zawiera ponad 14,6 tys. zdjęć, 1,4 tys. materiałów wideo oraz ponad 421 różnej długości tekstów sklasyfikowanych jako CSAM[40]. Ponad 8 tys. obrazów (zarówno CSAM, jak i nie CSAM) przeszło już proces adnotacji.

Infrastruktura i środowisko programistyczne

Ze względu na charakter zbieranych i analizowanych danych (dane wrażliwe) w środowisku produkcyjnym zespołu Dyżurnet.pl utworzono specjalnie zaprojektowaną i zabezpieczoną infrastrukturę sprzętowo-programistyczną, w której działa system APAKT i jest ulokowana omówiona powyżej baza danych. Uruchomione środowisko programistyczne oferuje narzędzia do monitorowania internetu, klasyfikacji materiałów wideo, zdjęć i tekstów oraz adnotacji treści z całego arsenału typów i rozszerzeń. Głównym celem projektu APAKT było zbudowanie systemu na potrzeby Dyżurnet.pl, niemniej jednak jego otwarta i modułowa architektura pozwala na łatwą rozbudowę i powiększanie o kolejne funkcje. Zakłada się stały rozwój systemu, wzbogacanie go o nowe, coraz skuteczniejsze klasyfikatory, zwiększanie listy atrybutów przypisywanych treściom i dostosowywanie do trendów w Europie i na świecie. System APAKT może z powodzeniem być wykorzystywany przez inne zespoły hotline.

Zakończenie

Tematyką artykułu są zagrożenia bezpieczeństwa dzieci w cyberprzestrzeni. Uwaga koncentruje się na nasilających się w ostatnich latach zagrożeniach

39 Ch.C. Aggarwal, op. cit.
40 Stan bazy APAKT na 22 maja 2023 roku.

o największym ciężarze gatunkowym, tj. próbach seksualnego wykorzystania nieletnich. Zostały omówione regulacje prawne, których celem jest podniesienie skuteczności ochrony dzieci, oraz role, jakie w systemie ochrony odgrywają tworzone w różnych krajach organizacje, w szczególności zespoły wykrywania i reagowania na nielegalne treści, tj. zespoły hotline.

Autorki zwróciły uwaga na konieczność automatyzacji działań związanych z monitorowaniem internetu na rzecz zwiększenia poziomu wykrywalności nielegalnych treści stanowiących poważne zagrożenie młodych użytkowników. Prezentowany jest system oprogramowania, który może być istotnym wsparciem dla zespołów hotline. Aktualnie rozwiązanie jest testowane na rzeczywistych danych zbieranych z sieci. Wykrywanie materiałów seksualnego wykorzystania dzieci jest ogromnym wyzwaniem. Przeprowadzone badania i eksperymenty z wykorzystaniem różnych rozwiązań wykorzystujących algorytmy uczenia maszynowego, w tym uczenia głębokiego, pokazują jak ważny jest dobór i ilość danych do trenowania klasyfikatorów. Uczenie modeli tylko na danych legalnych, w tym pornograficznych, jest nieskuteczne. Z drugiej strony, tworzone modele sieci neuronowych, szczególnie w przypadku analizy wideo, są niezwykle złożone i wymagające pod względem infrastruktury obliczeniowej, dlatego zastosowane podejście – wykorzystanie modeli wstępnie wytrenowanych na danych legalnych i ich dotrenowanie na danych CSAM.

Omawiany system APAKT jest rozwiązaniem unikatowym w skali kraju. Jego realizacja była dużym wyzwaniem zarówno dla specjalistów z zespołu Dyżurnet.pl, jak i wykonujących go naukowców i inżynierów. Konieczne było opracowanie szczegółowego systemu klasyfikacji treści CSAM i utworzenie bibliotek odpowiednio anotowanych danych. Ze względu na specyfikę przetwarzanych treści konieczne było przygotowanie bezpiecznego środowiska sprzętowo-programowego. Prace utrudniał brak możliwości bezpośredniej analizy skuteczności klasyfikatorów przez opracowujących je naukowców i inżynierów.

System APAKT został już wdrożony i funkcjonuje w środowisku produkcyjnym Dyżurnet.pl. Planuje się, że będzie on na bieżąco wspierał pracowników zespołu. Dzięki otwartej architekturze systemu będzie mógł być on rozbudowywany. Przewiduje się, że będzie mógł być również wykorzystywany przez inne zespoły działające na rzecz ochrony dzieci w cyberprzestrzeni funkcjonujące w kraju i na świecie.

Bibliografia

Aggarwal Ch.C., *Neural Networks and Deep Learning*, Cham 2018.
Analiza wyników badania dotyczącego treści intymnych publikowanych przez młodzież, https://dyzurnet.pl/publikacje [dostęp: 21.03.2023].
Cortes C., Vapnik V., *Support-vector networks*, „Mach Learn" 1995, nr 20.
CyberTipline 2021 Report, https://www.missingkids.org/gethelpnow/cybertipline/cybertipline-data#overview [dostęp: 20.04.2023].
He K., Gkioxari G., Dollar P., Girshick R.B., *Mask R-CNN*, arXiv:1703.06870 [cs.CV], 2018, https://doi.org/10.48550/arXiv.1703.06870 [dostęp: 20.04.2023].
International Survivor's Survey, https://www.protectchildren.ca/en/resources-research/survivors-survey-results/ [dostęp: 17.04.2023].
Internet Usage Statistics. The Internet Big Picture World Internet Users and 2023 Population Stats, https://www.internetworldstats.com/stats.html [dostęp: 20.04.2023].
Kasprzak W., Jankowski B., *Light-Weight Classification of Human Actions in Video with Skeleton-Based Features*, „Electronics" 2022, t. 11, nr 14.
Kasprzak W., Puchała S., Piwowarski P., *On Multi-stream Classification of Two Person Interactions in Video with Skeleton-Based Features* [w:] *Computer Vision and Graphics. ICCVG 2022. Lecture Notes in Networks and Systems*, red. L.J. Chmielewski, A. Orłowski, t. 598, Cham 2023.
Nastolatki 3.0. Raport z ogólnopolskiego badania uczniów, 2022, https://www.nask.pl/pl/raporty/raporty/4295, Raport-z-badań-nastolatki-3.0-2021.html [dostęp: 20.04.2023].
Nie na pokaz. Mówimy „nie" publikowaniu treści intymnych, https://www.saferinternet.pl/nie-na-pokaz/kampania.html [dostęp: 21.03.2023].
NudeNet: Neural Nets for Nudity Classification, Detection and selective censoring, https://pypi.org/project/NudeNet/ [dostęp: 23.04.2023].
Setter Ch., Greene N., Newman N., Perry J., *Global Threat Assessment 2021*, https://www.weprotect.org/global-threat-assessment-21/#report [dostęp: 20.04.2023].
Sexual abuse imagery of primary school children 1,000 per cent worse since lockdown, https://www.iwf.org.uk/news-media/news/sexual-abuse-imagery-of-primary-school-children-1-000-per-cent-worse-since-lockdown/ [dostęp: 21.03.2023].
Trend: 'Self-generated' content, https://annualreport2020.iwf.org.uk/trends/international/selfgenerated [dostęp: 21.03.2023].
Wasilewski J., *Zarys definicyjny cyberprzestrzeni*, „Przegląd Bezpieczeństwa Wewnętrznego" 2013, nr 9, s. 231.
Who we are, https://www.missingkids.org/home [dostęp: 20.04.2023].
Who we are, https://www.weprotect.org [dostęp: 20.04.2023].

Decision support system for detecting child abuse content in cyberspace

Abstract

In recent years, there has been a significant increase in threats to children's safety in cyberspace. The most serious of these include children's participation in illegal online activities and the production of sexually explicit content involving them. Therefore, it is of fundamental importance to build awareness of cyber threats among our society's youngest members and teach them skills for the safe use of products and services assigned to cyberspace. A key action for effectively protecting children in this environment is the early detection and reporting to the relevant authorities of illegal behavior and child

abuse content. Teams such as Dyżurnet.pl, whose tasks currently include responding to potentially illegal content reported by cyberspace users, and in the near future, possibly also conducting proactive activities in this area, play an important role here. The experience of Dyżurnet.pl clearly shows that effective detection of such content requires automation of activities and appropriate IT tools. This paper presents a novel network monitoring and decision support system using artificial intelligence methods, including deep learning, to automatically detect potentially harmful material, such as Child Sexual Abuse Material (CSAM), erotic content involving children, pornographic content with a created or processed image of a child and pornography involving adults.

Key words: cybersecurity, Child Sexual Abuse Material, CSAM, decision support system, artificial intelligence, machine learning, deep learning