Karolina Kuligowska[1], Agnieszka Repka[2], Nursulu Sagimbayeva[3]

Text Mining w przeciwdziałaniu dezinformacji: bezpieczeństwo poznawcze człowieka

1. Wprowadzenie

Złożony, globalny, szybko ewoluujący ekosystem informacyjny posiada wiele luk – miejsc narażonych na działania dezinformacyjne. W celu wspierania społeczeństwa w docieraniu do prawdziwych, wiarygodnych i dokładnych informacji, należy zatem propagować podejście oparte na kwestionowaniu i analizowaniu otrzymywanych wiadomości oraz na stosowaniu najlepszych metod i technik pozwalających na weryfikację informacji. Ekosystem ten jest obecnie tak zanieczyszczony, że każdy użytkownik internetu musi wziąć odpowiedzialność za samodzielne sprawdzanie i weryfikowanie tego, co czyta w sieci.

Istnieje wiele dziedzin badań, które koncentrują się na projektowaniu i wdrażaniu systemów, służących do przechowywania, zarządzania, wyszukiwania i komunikowania informacji. W odpowiedzi na ogromną ilość i znaczną złożoność informacji, które napływają do każdego człowieka, część z tych dziedzin zaczęła koncentrować się na zmniejszaniu obciążenia poznawczego [1]. Rozwiązania w tym zakresie można dziś spotkać na każdym kroku: 1) literatura popularno-naukowa w przystępny, zilustrowany graficznie i semantycznie sposób, tłumaczy zawiłości nauki szerszemu gronu odbiorców, 2) ekonomiści próbują wyjaśniać przedsiębiorcom przepisy podatkowe przy pomocy użytecznych, wziętych z życia przykładów ich zastosowania, 3) korporacje, w komunikacji wewnętrznej z pracownikami, jak i zewnętrznej z klientami i inwestorami, stosują m.in. storytelling w celu przybliżenia wyników analizowanych przez siebie danych i prezentowania dalszych strategii rynkowych.

Powyższe rozwiązania bazują m.in. na mechanizmie tworzenia sensu, jako rodzaju wnioskowania narracyjnego, w którym jednostka jest w stanie działać odpowiednio w takim zakresie, w jakim utożsamiła się z historią i odgrywaną w niej rolą. Rozwiązania te są jednak pewnego rodzaju socjotechniką. Zatem dodatkowymi kategoriami, które należy tu wprowadzić, są: intencje nadawców oraz rzetelność naukowa w odniesieniu do interpretacji surowych danych wejściowych. Rzetelność, wiarygodność i prawdziwość stanowią te cechy informacji, których sprawdzenie stanowi kolejne wyzwanie dla świata nauki.

Nie ulega zatem wątpliwości, że reakcje rządów, podmiotów odpowiedzialnych za naukę i oświatę, bezpieczeństwo publiczne, w tym szczególnie cyberbezpieczeństwo, powinny być szybkie, adekwatne i scentralizowane. Widoczne są oczywiście starania w tym względzie, zarówno podmiotów prywatnych, jak i państwowych – w mediach

[1] kkuligowska@wne.uw.edu.pl, Wydział Nauk Ekonomicznych, Uniwersytet Warszawski, www.wne.uw.edu.pl.
[2] a.repka@student.uw.edu.pl, Wydział Nauk Politycznych i Studiów Międzynarodowych, Uniwersytet Warszawski, www.wnpism.uw.edu.pl.
[3] nurs.sagimbayeva@gmail.com, Department of Foreign Philology and Translation Studies, Al-Farabi Kazakh National University, www.kaznu.kz.

popularyzowany jest tzw. fact checking, czyli zestaw praktyk potrzebnych do weryfikacji danego tekstu: jego autora, stopnia obiektywności, odwołań do innych treści itp. Na stronach rządowych publikowane są infografiki ostrzegające o zagrożeniach w sieci oraz informujące o bieżących trendach dezinformacji. Do szkół zapraszani są eksperci cyberbezpieczeństwa, również z ramienia Policji, tłumaczący mechanizmy działania scamerów, fraudsterów i hackerów.

Wciąż jednak brakuje pewnych wyprzedzających kroków, wzmacniających prewencję, nastawionych na systemowe podejście i interdyscyplinarne działanie w długiej perspektywie. Z jednej strony cyberbezpieczeństwo skupia się na ochronie informacji w danej infrastrukturze. Z drugiej strony psychologia i kognitywistyka badają podatność i odpowiedź człowieka na nowe zjawiska i zagrożenia, a także motywy twórców tych zagrożeń. Z kolei kryminologia i nauki o bezpieczeństwie badają nowe trendy przestępstw i zagrożeń w internecie oraz ich wpływ na bezpieczeństwo jednostek i grup. Wreszcie dyscyplina Text Mining oferuje liczne narzędzia mogące istotnie wesprzeć filtry poznawcze człowieka. Dopiero połączenie wyżej wymienionych dziedzin i objęcie wspólną kategorią: 1) wspomnianych zjawisk i zagrożeń, 2) mechanizmów ich powstawania, 3) modus operandi sprawców, 4) odpowiedzi poznawczej człowieka oraz 5) wszelkich działań zaradczych, mogłoby się przyczynić do większej systematyzacji, szerszego zrozumienia wzajemnych powiązań i optymalizacji w kwestii zwalczania oraz prewencji.

Takim wspólnym interdyscyplinarnym mianownikiem, w naszej opinii, może się stać rodząca się pośród naukowych dysput dyscyplina bezpieczeństwa kognitywnego. Jedną z pierwszych wzmianek o tej dziedzinie była wypowiedź Randa Waltzmana. Ten wybitny ekspert w zakresie stosowania sztucznej inteligencji w środowiskach informacyjnych, apelował w 2015 roku przed Podkomisją ds. Cyberbezpieczeństwa Senackiej Komisji Sił Zbrojnych USA o utworzenie Cetrum Bezpieczeństwa Kognitywnego w odpowiedzi na nowe zagrożenia informacyjne [2]. Centrum to zostało powołane i działa na niewielką skalę pod nazwą COGSEC Cognitive Security & Education Forum jako platforma łącząca ekspertów i inicjująca projekty badawcze [3]. Dyscyplina bezpieczeństwa poznawczego człowieka jest w fazie formowania się i bywa różnorako rozumiana. W swoim sednie może odwoływać się do licznych podejść psychologicznych, psychokognitywnych i społeczno-kulturowych. Dziedziny te, rozpatrywane odrębnie, nie wykorzystują w pełni podejścia interdyscyplinarnego w zakresie aktualnie rozwijanych technologii, a zwłaszcza narzędzi informatycznych. Natomiast dziedzina Text Mining, silnie powiązana z systemami informatycznymi przetwarzającymi dane, chociaż nie jest stricte nastawiona na rozwiązywanie problemów dotyczących dzisiejszych zagrożeń informacyjnych, to poprzez swoje metody i techniki coraz częściej pojawia się w badaniach w kontekście przeciwdziałania dezinformacji (przykłady przytoczono w dalszej części artykułu, przypisy [36-49]). Z uwagi na powyższe, propagujemy połączenie tych dziedzin we wspólnych badaniach, mając na uwadze stworzenie nowych możliwości zwiększenia bezpieczeństwa osobistego człowieka.

Niniejszy rozdział składa się z pięciu części. Wprowadzenie stanowi część 1. W części 2 przybliżono pojęcia obciążenia poznawczego i umiejscowiono człowieka w epicentrum zagrożeń takich jak: szum informacyjny, dezinformacja, trolling, fałszywe wiadomości – fake newsy. W części 3 omówiono do jakich mechanizmów odpowiedzi kognitywnej użytkowników odnoszą się media i aktorzy w nich działający, aby posiąść

i utrzymać ich uwagę. Wymieniono w niej definicje wpływu społecznego i opinii publicznej, a także popularne heurystyki i błędy logiczne, którymi posługują się media i agenci wpływu. W części 4 przedstawiono dziedzinę Text Mining oraz przybliżono możliwości wykorzystania jej narzędzi w celu przeciwdziałania dezinformacji i pokrewnym zagrożeniom. W części 5 zwrócono uwagę na potrzebę interdyscyplinarnego podejścia i zaczerpnięcia z dziedzin związanych z bezpieczeństwem, kognitywistyką oraz nowymi technologiami w celu systemowego zwalczania analizowanych zjawisk i zagrożeń.

2. Mechanizmy poznawcze człowieka w obliczu dezinformacji i pokrewnych zjawisk

W dobie XXI wieku ludzkość, w życiu zawodowym oraz prywatnym, operuje w cyfrowym świecie. Rzeczywistość oparta na danych wiąże się z problemem ogromnej ilości informacji napływającej nieustannym strumieniem. Rozwój technologii cyfrowych, zwłaszcza w zakresie szybkości przesyłu danych i ich niemal nieograniczonych możliwości przechowywania, spowodował zmiany w dostępności, wytwarzaniu, strukturze i komunikowaniu informacji w historycznie niespotykanym tempie [4]. W 2020 roku zostało wytworzonych prawie 60 zettabajtów (tj. 60 bilionów gigabajtów) danych i oczekuje się, że ilość danych cyfrowych utworzonych w latach 2021-2025 znacznie przekroczy skumulowaną ich ilość stworzoną od lat 70. XX wieku, kiedy to zaczęto używać masowej pamięci cyfrowej [5]. Te ogromne ilości danych każdego dnia zwiększają swój wolumen, stąd wypływa konieczność ich filtracji i selekcji. Jednak opanowanie zalewu ogromną ilością danych różnymi narzędziami z zakresu Big Data, Data Mining i Text Mining to tylko jeden aspekt ery informacji. Wraz z dalszym, nieuniknionym rozwojem nowych technologii, coraz bardziej paląca staje się potrzeba zmniejszenia obciążenia poznawczego człowieka, spowodowanego przeładowaniem informacyjnym.

Objętość, gęstość i strukturalna złożoność informacji ma niewątpliwy wpływ na procesy poznawcze. Chociaż mogłoby się wydawać, że dostarczenie większej ilości informacji sprzyja lepszemu wnioskowaniu i szybszej stabilizacji sensu odbieranych treści, to często bywa przeciwnie. Bowiem kiedy mózg nie może zredukować złożoności problemu, zmniejsza złożoność strategii użytej do nadania mu sensu [6]. Gdy dana osoba jest narażona na potencjalnie istotne, ale sprzeczne informacje, podawane jej w tempie niezgodnym z czasem i wysiłkiem wymaganym do ich integracji oraz nie ma dostępu do odpowiednich narzędzi, zaufanej sieci ekspertów lub szkoleń specyficznych dla danej dziedziny – umysł tej osoby może odpowiedzieć nie tylko zmniejszoną zdolnością do radzenia sobie ze stresem, ustalania priorytetów, czy skutecznego podejmowania decyzji w zakresie kategoryzacji tych informacji, ale przede wszystkim ograniczeniem zdolności wykrywania logicznej niespójności pomiędzy otrzymanymi informacjami [4]. Wspomniany mechanizm tworzenia sensu, jako jeden spośród wielu innych w spektrum odpowiedzi kognitywnej, behawioralnej i emocjonalnej człowieka na informacje, stanowi pożywkę dla nadawców zainteresowanych wszelkiego rodzaju manipulacją i dezinformacją. Przy czym rozróżnić tutaj można dwa typy informacji, którymi społeczeństwo XXI wieku jest zalewane: szeroko pojęty szum informacyjny oraz celowa dezinformacja.

Szum informacyjny, nazwany także metaforycznie mgłą informacyjną lub smogiem medialnym [7], dotyczy zjawiska przeładowania informacjami, które dodatkowo można scharakteryzować jako chaotyczne, niespójne, nieaktualne oraz częściowo lub całkowicie nieprawdziwe [8]. Dezinformacja natomiast jest działaniem związanym z podawaniem nieprawdziwych informacji, które są tworzone z zamiarem wyrządzenia szkody grupom lub osobom i które mają za zadanie celowo wprowadzić odbiorcę w błąd [9].

Skala masowej dezinformacji w cyberprzestrzeni rośnie z dnia na dzień. Jej klasycznym i wyrazistym przykładem były informacje generowane od pierwszych dni pandemii COVID-19 w 2020 roku [10]. Wówczas dotyczyły one m.in. rozmaitych opisów genezy wirusa, często wykluczających się wytycznych związanych z ochroną przed zachorowaniem, a także późniejszych różnorodnych informacji odnośnie opracowanych szczepionek, ich zawartości oraz stopnia ich skuteczności. Z kolei jeżeli chodzi o celową masową dezinformację w zakresie bezpieczeństwa publicznego, to przykładem może być tutaj ta związana z rosyjską agresją przeciwko Ukrainie w 2022 r. Intensywne działania szerzące strach, niepewność i uprzedzenia społeczne na różnorodnym tle przybierają na sile w miarę rozwoju wydarzeń. Procesy te, dotąd realizmwane przez tzw. agentów wpływu, stają się obecnie zautomatyzowane z powodu zaangażowania w nie technologii. Dziś dezinformacja jest rozpowszechniana w mediach społecznościowych przez boty i tzw. trolle, a więc sztuczne produkty potężnych farm serwerowych, które wraz z fałszywymi, ale obsługiwanymi przez ludzi kontami, służą jako mnożniki siły dla operacji dezinformacyjnych, a także hejterskich, tworząc pozory autentycznego zaangażowania i debaty [11]. Z tego powodu – w przywołanym przykładzie masowej dezinformacji w zakresie bezpieczeństwa publicznego, wiele mediów zdecydowało się wyłączyć możliwość komentowania zamieszczanych treści pod tekstami dotyczących inwazji Rosji na Ukrainę, przykładowo Grupa Wirtualna Polska [12], Interia oraz Polska Press [13] i inne [14]. Również Rządowe Centrum Bezpieczeństwa RP raportowało postępujące prowokacje informacyjne i w związku z tym zdecydowało się przygotować adekwatny materiał, a mianowicie krótki poradnik pt. „Bądź gotowy – poradnik na czas kryzysu i wojny", w którym podaje m.in. jak chronić się przed dezinformacją, aby nie ulec np. panice [15].

Pojęcie fake newsów dotyczy publikowania w mediach, a zwłaszcza w mediach społecznościowych, nieprawdziwych, wprowadzających w błąd, częstokroć absurdalnych w swym wydźwięku, treści [16]. Geneza celowego rozpowszechniania propagandowej nieprawdy sięga czasów starożytnych. Intencją tworzenia i rozpowszechniania fałszywych wiadomości – w przeszłości i w dzisiejszych czasach – są najczęściej korzyści finansowe, polityczne lub wizerunkowe osoby lub grupy osób tworzących fake newsy. Relacjonowanie wydarzeń zgodnie z obraną perspektywą, najczęściej władcy zlecającego napisanie kroniki, jest zgodne ze znanym powiedzeniem „historię piszą zwycięzcy" – nawiasem mówiąc ten cytat jest błędnie przypisywany Winstonowi Churchillowi, gdyż powiedzenie to znane było już w XVIII wieku, ale oryginalny autor nie jest znany. Natomiast w XIX w. posługiwano się terminem „żółta prasa", który nawiązywał do dziennikarstwa bazującego na publikowaniu sensacyjnych, niepotwierdzonych, skandalizujących historii o morderstwach i krwawych wypadkach. Wtedy to dziennikarze Joseph Pulitzer i Wiliam Hearst odkryli, że „żółta prasa" wywoływała większe zainteresowanie i entuzjazm czytelników niż codzienne, zwykłe i „nudne" dla społeczeństwa wiadomości [17].

Historia preparowania fake newsów jest zatem dość długa, ale dopiero w erze informacji, w rezultacie powstania i dynamicznego rozwoju mediów społecznościowych, przybrała postać cyfrową. W świecie fake newsów informacje nie opierają się na faktach, ale na ich interpretacji i odpowiednim modelowaniu wykonywanym przez piszących dziś alternatywne historie „zwycięzców", jakimi można nazwać tych, którzy chcą i z powodzeniem wpływają na zawartość treści w mediach. Kreowane przez nich tak zwane operacje wpływu, których celem jest między innymi rozpowszechnianie propagandy w celu zdobycia przewagi nad przeciwnikiem, stosowane są z powodzeniem zarówno w kontekście cywilnym, jak i militarnym. Ich celem może być ośmieszenie przeciwnika politycznego, zdyskredytowanie konkurencyjnego twórcy lub celebryty, a nawet destabilizacja całego społeczeństwa danego kraju lub regionu, jego polaryzacja, radykalizacja i w końcu realizacja znanego od zarania dziejów mechanizmu „dziel i rządź". Wielu odbiorców fake newsów, uznając je za w pełni prawdziwe i wiarygodne, z pełnym zaangażowaniem oddaje się propagowanej przez nich sprawie. W ten sposób powielają oni szkodliwe treści bez wiedzy o rzeczywistych beneficjentach tych działań. Do tak zmanipulowanych odbiorców przylgnął termin znany jeszcze z czasów analogowej propagandy, mianowicie „pożyteczni idioci" [18].

3. Wpływ mediów na odpowiedź kognitywną

Media prześcigają się w używaniu mechanizmów wpływu, by przyciągnąć widza i osiągnąć najwyższą oglądalność lub tzw. klikalność w przypadku mediów społecznościowych. Wiodące media stają się coraz bardziej stronnicze, w zależności od grupy nacisku, finansowania przez daną opcję polityczną, czy politykę wydawniczą, a z powodu opłacalności tego działania – stronią od wywoływania dysonansu poznawczego, utrwalając odbiorcę w jemu właściwym zestawie poglądów i wniosków. Z kolei prezentacja treści przeciwnych jest jednocześnie ukierunkowana na ich zbagatelizowanie, zdeprecjonowanie lub wyśmianie oraz na wywołanie u odbiorcy efektu obcowania ze zrównoważoną i obiektywną relacją obu stron [19].

Jednym z mechanizmów, na znajomości którego bazują media, jest mechanizm kreowania postaw i jej komponentów: komponentu poznawczego (stereotypu), afektywnego (uprzedzenia) oraz behawioralnego (dyskryminacji). Postawa oparta na poznaniu wywodzi się z myśli i przekonań na temat właściwości obiektu postawy. Postawa oparta na afekcie wynika z emocji i uczuć wobec przedmiotu postawy, bardziej niż na racjonalnym myśleniu. Natomiast postawa oparta na zachowaniu jest obrazem działania wobec obiektu postawy [20]. Apelując do wszystkich komponentów postaw, manipulacje mediów dotyczą zjawiska wybiórczego poszukiwania informacji (tzw. cherry picking), potrzeby potwierdzania słuszności własnych poglądów i selektywności pamięci, w ramach której lepiej zapamiętujemy treści zgodne z naszymi przekonaniami.

Media społecznościowe idą jeszcze dalej w celu utrzymania uwagi widza: po zauważeniu, poprzez targetowanie behawioralne i kontekstowe, jaki zestaw preferencji wobec odbieranych treści reprezentuje dany użytkownik, algorytmy podsuwają kolejne – takie, które odpowiadają temu kluczowi. O ile w przypadku celów reklamowych zjawisko to jest raczej neutralne i niegroźne, o tyle w kontekście poglądów prowadzi do tworzenia się tak zwanych „baniek informacyjnych", w których zamyka się dany użytkownik. Pojęcie bańki informacyjnej (inaczej filtrującej) opisał po raz pierwszy

Elie Pariser w 2011 roku [21]. Zgodnie z jego teorią algorytmy wyszukują informacje najbardziej odpowiadające gustom odbiorców i tylko takie kolejno im prezentują. W momencie, gdy odbiorca nie weryfikuje wiedzy w innych źródłach, zatraca on pełny obraz rzeczywistości, ponieważ kształtuje poglądy i podejmuje decyzje jedynie na podstawie jednolitego i jednostronnego przekazu zawężającego obraz danego zjawiska. Zradykalizowane w ten sposób grupy stają się potężną lokomotywą dezinformacji i katalizatorem agresji wobec ludzi o odmiennych poglądach.

Badania pokazują, że osoba, która rozpoczyna korzystanie z platformy społecznościowej Facebook, może zradykalizować się w czasie 4-6 miesięcy, nawet do poziomu przynależności do grupy terrorystycznej [22]. Zjawisko to opisywane jest jako tzw. rabbit hole effect, tłumaczonym na język polski jako efekt króliczej nory. W jej obszerny wlot, w kontekście prezentowanych treści, wpada dany użytkownik, po czym spiralnie podąża w dół, lejkiem coraz bardziej wąskich i radykalnych poglądów [23]. Nie należy tutaj jednak zrzucać całej winy na algorytmy targetujące, czy upatrywać w ich działaniu szerszej negatywnej socjotechniki zaimplementowanej przez ich twórców. Zjawisko króliczej nory jest realne, ale aktywną stroną inicjującą radykalizację pozostaje odbiorca. Badania wewnętrzne twórców platformy YouTube, zaniepokojonych doniesieniami o tym efekcie, pokazały, że algorytmy nie rozróżniają poziomu „ekstremalności" treści danego filmu, dlatego też nie może być mowy o celowym podsuwaniu coraz bardziej radykalnych treści [24]. Z katalogu podobnych materiałów to właśnie użytkownik wybiera kolejny do odtworzenia, a jak pokazały badania Woolley i Sharif, efekt króliczej nory występuje wtedy, gdy zwiększona jest dostępność treści w danej kategorii i poczucie zanurzenia w niej. Użytkownicy przewidują wtedy, że kolejne materiały w tej samej kategorii będą przyjemniejsze w odbiorze niż pozostałe [25].

Możliwości rozprzestrzeniania fake newsów, dezinformacji i wszelkiego rodzaju manipulacji dokonywanych przez jednostki, grupy, a także rządy i media są ogromne – odwołują się do różnych sfer percepcji i mechanizmów reakcji człowieka oraz korzystają z kanałów indywidualnych, jak i zbiorowych. Najszerszymi z nich zdają się tu być zjawiska wpływu społecznego i opinii publicznej, których wykorzystanie do kreowania postaw stanowi podstawę komunikacji politycznej i medialnej.

Wywieranie i uleganie wpływowi społecznemu (tj. zmianie w zachowaniu, spowodowanej przez rzeczywisty lub wyobrażony nacisk ze strony innych ludzi [26]) jest zjawiskiem powszechnym. Zachodzi poprzez liczne mechanizmy. Trzy wyróżniające się spośród nich to: 1) naśladownictwo (kopiowanie cudzych zachowań), 2) konformizm (uleganie naciskowi, czasem jedynie wyobrażonemu, ze strony danej większości ludzi) oraz 3) posłuszeństwo wobec autorytetu (podporządkowywanie się ludziom mającym władzę czy autorytet) [20]. Zaś sześć najważniejszych reguł wpływu stanowią: reguła wzajemności, konsekwencji, społecznego dowodu słuszności, lubienia, autorytetu i niedostępności [27].

Opinia publiczna to zbiór dominujących w społeczności reakcji na ważne w danym czasie kwestie publiczne (m.in. polityczne, społeczne) [28]. Zjawisko to uświadamiane było już za czasów debat publicznych w starożytnych agorach, jako „pogłoska" (łac. feme), wyrażane także w starożytnym Rzymie, jako „vox populi, vox dei" (powiązanie powszechnej opinii ludu z autorytetem boskim). Aktualnie opinia publiczna prezentuje charakter legitymizacyjny, jako że racje popiera się pojęciem „rozstrzygającej więk-

szości". Deutsch i Gerard opisują dwie potrzeby psychiczne, które doprowadzają ludzi do dostosowywania się do oczekiwań innych, w tym dopasowywania się do opinii większości. Jest to naturalna potrzeba zachowania pozytywnego obrazu siebie oraz potrzeba bycia akceptowanym społecznie [29].

Opinia publiczna może się dynamicznie zmieniać w krótkim czasie, podlega bowiem wpływom: edukacji, wychowania, tradycji kulturowej, norm grupowych, a także podziałów społecznych (klasowości, warstwowości, stowarzyszeń). Jednym z kluczowych narzędzi wpływu są jednak media. Goban-Klas uważa, że „dla powstania opinii publicznej nie trzeba mediów masowych, wystarczają środki prostsze, interpersonalne, jednak połączenie środków masowych i interpersonalnych jest dla powstania opinii publicznej najskuteczniejsze" [30]. Pośród mechanizmów wpływu mediów Lepa wymienia: propagandę polityczną (a w niej propagandę szeptaną), działanie przywódców opinii, spiralę milczenia, poprawność polityczną i manipulowane sondaże. Narzędziami zaś mogą być odpowiednio nagłośnione deklaracje, komunikaty, oświadczenia czy orędzia przywódców opinii, polityków, ekspertów [31]. Kuśmierski wyodrębnia dodatkowe składniki uczestniczące w formowaniu opinii publicznej, tj.: wyobrażenia o interesach grupowych, wzory kulturowe, stereotypy, mity, przesądy i uprzedzenia, pogłoski i plotki oraz wiedzę o faktach [32]. Znając te składniki i mechanizmy można stwierdzić, że media wywierają wpływ nie tylko poprzez przekazy w zakresie polityki, czy opinii społecznych. Robią to również epatując treściami nastawionymi – mniej lub bardziej świadomie – na utrwalanie stereotypów, a więc swoistego kompasu społecznego w zakresie naszego postrzegania innych ludzi. Wg teorii kategoryzacji społecznych są one ubocznym skutkiem normalnego funkcjonowania ludzkiego umysłu, czyli dzielenia napotykanych ludzi na kategorie i gromadzenia o nich wiedzy [20].

Analizując opinię publiczną należy jednak wziąć pod uwagę nie tylko stereotypy, przez pryzmat których patrzą ludzie, ale również heurystyki wypowiadania sądów, które aplikują do swoich wypowiedzi, a więc „drogi na skróty", uproszczone reguły myślenia, które pozwalają na szybkie formułowanie opinii bez wnikliwej analizy treści. Najważniejsze z nich to: heurystyka dostępności (szacujemy prawdopodobieństwo wystąpienia danego zjawiska na podstawie łatwości znalezienia jego przykładów w swoim otoczeniu), heurystyka zakotwiczenia/dostosowania (zakładamy jakąś liczbę na podstawie już usłyszanej wartości, oscylując wokół niej, jak wokół kotwicy), heurystyka symulacji (prognozujemy przebieg zdarzeń na podstawie łatwości wyobrażenia) i heurystyka reprezentatywności (opiniujemy przynależność danej jednostki na podstawie jej podobieństwa do innych jednostek w danej kategorii) [20]. Błędy logiczne, jak również podstawowe chwyty retoryczne używane często nieświadomie przez uczestników jakichkolwiek dyskusji, w manipulowaniu opinią publiczną są używane z pełną premedytacją, np. w reklamie czy w debacie politycznej. Do najpopularniejszych kategorii „manowców umysłowych" zaliczyć można błędy rozumowania (np. odwołania ad personam, fałszywa dychotomia, demagogia) logiczne (np. rozumowanie nie na temat – ignoratio elenchi, pozorne uzasadnienia), metodologiczne (odstępstwa od metod procesów naukotwórczych, argumenty anegdotyczne), teoriopoznawcze (np. bezkrytyczność, efekt Dunninga-Krugera) i semiotyczne (np. niewystarczająca zrozumiałość przekazu, homonimia, amfibolia, elipsa) [33].

4. Podejście interdyscyplinarne: zastosowania Text Mining w bezpieczeństwie poznawczym

Tekst jest cennym źródłem informacji i nośnikiem wiedzy, który daje wgląd w różne dziedziny, od marketingu i finansów po opiekę zdrowotną, psychologię i wreszcie dziedzinę bezpieczeństwa. Niestety również, jak zostało to już wielokrotnie wspomniane, tekst może być nośnikiem zmanipulowanych lub fałszywych treści. W ostatnich latach rozwinęły się i upowszechniły zupełnie nowe źródła tekstów, takie jak platformy mediów społecznościowych, fora internetowe oraz szeroko pojęte media cyfrowe. Równocześnie ogromne ilości danych tekstowych powodują, iż ich dokładna, mozolna i „ręczna" analiza jest dziś praktycznie niemożliwa. Z pomocą przychodzi automatycznie wykonywana eksploracja tekstu przy pomocy metod Text Mining.

Text Mining to dyscyplina, która wyrasta z dziedzin Data Mining i przetwarzania języka naturalnego, lingwistyki komputerowej oraz sztucznej inteligencji. Jest to również zestaw metod analitycznych, technik statystycznych i narzędzi informatycznych, które pomagają wydobyć znaczącą, przydatną oraz nieznaną wcześniej wiedzę z danych tekstowych [34]. Główna różnica między Data Mining (eksploracją danych liczbowych) a Text Mining (eksploracją danych tekstowych) polega na tym, że Data Mining zwykle analizuje ustrukturyzowane dane, będące liczbami i wartościami, natomiast Text Mining polega na przetwarzaniu danych, które są nieustrukturyzowanym tekstem.

Najważniejsze metody Text Minigu obejmują: klasyfikację, klastrowanie, wykrywanie tematów i analizę sentymentu. Zadaniem klasyfikacji tekstu jest zaklasyfikowanie dokumentów tekstowych do z góry określonego zestawu kategorii, przy czym każdy dokument tekstowy reprezentowany jest jako tzw. worek słów, w którym znajdują się słowa wycięte z ich wcześniejszego kontekstu, czyli z ich położenia w zdaniu lub w dokumencie. Z kolei metody klastrowania tekstu służą do grupowania w kolekcji tekstowej podobnych dokumentów poprzez analizę ich zawartości [35]. Różnica między klasyfikacją tekstu a klastrowaniem dokumentów tekstowych polega na tym, że modele klasyfikacji wykorzystują dane uprzednio oznaczone etykietami na predefiniowane przez analityka kategorie, podczas gdy w przypadku klastrowania uzyskane kategorie są rezultatem analizy podobieństwa dokumentów tekstowych wykonanej przez algorytm klastrujący. Natomiast wykrywanie tematów (modelowanie tematów) jest używane do wykrywania motywów przewodnich w kolekcji tekstowej. Ostatnia metoda – analiza sentymentu – służy do wykrywania wydźwięku emocjonalnego danego tekstu. Najbardziej podstawowy podział wykrytego wydźwięku dotyczy pozytywnego, negatywnego i neutralnego sentymentu badanej treści.

Wszystkie wymienione metody Text Mining mogą być nieocenionym wsparciem filtrów poznawczych człowieka. W szczególności – człowieka narażonego na szum informacyjny, celową dezinformację i fake newsy. Zatem oparcie tych metod o głębsze poznanie zagrożeń bezpieczeństwa kognitywnego człowieka może zamienić dotychczas podejmowane działania naprawcze wobec już wyrządzonych szkód w wyprzedzające kroki prewencyjne.

Dla zrozumienia potencjału, jaki kryje się za zastosowaniem takiego podejścia, dla każdej z metod Text Mining przytaczamy bieżące możliwości ich zastosowań w przezwyciężaniu dezinformacji i fake newsów, badane na przestrzeni ostatnich lat:

1. Klasyfikacja

Prosta metoda klasyfikacji nie wystarczy do poprawnego wyodrębnienia fake newsów z setek tysięcy komentarzy i postów, ponieważ metody klasyfikacji po pierwsze nie są wyspecjalizowane w rozróżnianiu między fałszywymi a niefałszywymi wiadomościami, a po drugie ich stosowanie wymaga ogromnej ilości mocy obliczeniowych i czasu w kontekście stałego napływu kolejnych treści. Aby przyspieszyć te prace naukowcy sięgają po rozwiązania sztucznej inteligencji (ang. Artificial Intelligence, AI). Sztuczna inteligencja na aktualnym etapie zaawansowania odwołuje się do zestawów instrukcji zwanych algorytmami oraz architektur opartych o sztuczne sieci neuronowe, symulujących działanie neuronów biologicznych mózgu. Istnieje wiele algorytmów klasyfikacji dostępnych w uczeniu maszynowym i tzw. uczeniu głębokim, które mogą nauczyć się dostrzegać wzorce w tekście i na tej podstawie kategoryzować treści. Dzięki integracji sztucznej inteligencji i przetwarzania tekstowego można budować zaawansowane klasyfikatory, które wyodrębniają różnorodne cechy tekstu, a następnie włączają te cechy do klasyfikacji. Aby opracować klasyfikatory, które faktycznie uczą się czym są fake newsy, należy zmusić je do rozważenia kontekstów, w których pojawiają się zaprezentowane im dane tekstowe.

Przykładowo, analiza poziomu dokładości klasyfikacji tekstu przeprowadzona przez Ahmeda, Hinkelmanna i Corradiniego dotyczyła trzech różnych klasyfikatorów uczenia maszynowego (Naïve Bayes, SVM, Passive Aggressive Classification) trenowanych na dwóch publicznie dostępnych zbiorach danych [36]. Pierwszy zbiór zawierał 18000 newsów zebranych z platform informacyjnych New York Timesa, Guardiana i Bloomberga w gorącym okresie przedwyborczym w USA w 2016 roku. Newsy te były uprzednio oznaczone przez etykiety binarne (0 lub 1), oznaczające odpowiednio prawdziwy news oraz fake news. Drugi zbiór danych zawierał 5000 artykułów prasowych zebranych z Signal Media News, w którym newsy również były oznaczone przez te same etykiety binarne. Wyniki badania Ahmeda, Hinkelmanna i Corradiniego sugerują, że zastosowane podejście pomaga w klasyfikowaniu fałszywych wiadomości: zaobserwowano, że proponowane modele klasyfikacyjne osiągnęły najwyższą dokładność klasyfikacji do 93% z klasyfikatorem Passive Aggressive Classification, 85% z Naïve Bayes i 84% z SVM. Z kolei Ott i in. zastosowali klasyfikator SVM z funkcjami LIWC+ Biggrams, który osiągnął poziom dokładności do 89%, łącząc w sobie również cechy stylometryczne, takie jak częstości występowania słów mało istotnych, powtarzalność słów oraz występowanie specyficznych językowych struktur [37]. Podkreślić przy tym należy, iż każdy fake news ma zazwyczaj inną charakterystykę, co utrudnia zbudowanie klasyfikatora, który dogłębnie sprawdziłby treść każdej wiadomości.

Kolejnym przykładem może być badanie wykonane przez Kim i in. dotyczące błędnych informacji na temat żywności i suplementów diety, w tym zwłaszcza czosnku, które były szeroko rozpowszechniane u progu epidemii COVID-19. Badanie to miało na celu zastosowanie modelu sieci neuronowych BERT do klasyfikacji dezinformacji na temat czosnku w kontekście COVID-19 na Twitterze [38]. Badacze wykorzystali 5929 oryginalnych tweetów wzmiankujących jednocześnie czosnek i COVID-19 (4151 tweetów do nauczenia klasyfikatora i 1778 do testowania). Tweety były ręcznie oznaczane jako „dezinformacja" i „inne". Model BERTweet-large wykazał się poziomem 91% skuteczności w klasyfikowaniu dezinformacji w tweetach. Dzięki takim

wynikom można wskazać, iż sieci neuronowe mogą wspomagać wykrywanie fałszywych informacji i są przydatne do klasyfikowania dezinformacji na Twitterze w odniesieniu do różnorodnych tematów. Równocześnie, ponieważ BERTweet-large jest modelem BERT specyficznym dla tweetów, trudno byłoby go zastosować na innych platformach społecznościowych. Dodatkowo, wszystkie algorytmy klasyfikacji sztucznej inteligencji wymagają trenowania na ogromnych korpusach tekstowych, które muszą zostać uprzednio oznaczone etykietami „ręcznie" przez badaczy. Jest to właściwie największe wyzwanie w tym obszarze: niedostępność dużych, oznaczonych korpusów tekstowych, które mogłyby służyć jako zbiory treningowe ukierunkowane na wykrywanie fałszywych wiadomości i flagowanie dezinformacji. Główne platformy mediów społecznościowych stosują już co prawda narzędzia sztucznej inteligencji do oznaczania fałszywych informacji lub hejtu i mowy nienawiści, ale klasyfikacja treści w internecie nadal nie jest prosta.

2. Klastrowanie

Poprzez zastosowanie technik klastrowania tekstu można m.in. wyizolować i ujawnić podstawowe czynniki, cechy lub podmioty dezinformacji – a zatem np. zidentyfikować i scharakteryzować klastry użytkowników mediów społecznościowych, którzy rozprzestrzeniają fake newsy, tym samym docierając do głównych węzłów dezinformacji. Dalej można przeprowadzić analizę takich skupisk użytkowników pod kątem wykrywania sieci powiązanych kont, ich historii w portalach społecznościowych oraz położenia geograficznego i tym samym zadziałać w zarodku, poprzez usuwanie głównych punktów dystrybucji fałszywych treści.

Niewątlipwym wyzwaniem w przeciwdziałaniu dezinformacji jest automatyczne identyfikowanie samej treści fałszywych wiadomości poprzez ich klastrowanie (grupowanie w skupiska) i powiadamianie użytkowników o charakterze newsów, wpisów, artykułów, z którymi mają do czynienia online. W wyniku analizy takich klastrów można na przykład ujawnić różne wspólne cechy stosowanego słownictwa ich twórców i powtarzalne schematy w konstrukcji wypowiedzi [39]. To jednak nie jedyne możliwości jakie niesie ze sobą użycie klastrowania. Następnie można bowiem przeprowadzić analizę takich skupisk pod kątem wykrytych słów kluczowych, opisujących każdy klaster, by finalnie zidentyfikować tematykę każdego klastra.

Przykładowo, Hosseinimotlagh i inni postanowili poprzez klastrowanie podzielić na różne rodzaje i zidentyfikować ponad 12000 fałszywych wiadomości [40]. Zaproponowali oni schemat tensorowy, który wykorzystuje kontekst pojęciowy poprzez uchwycenie relacji przestrzennych między treścią każdej wiadomości. Wyniki Hosseinimotlagha i jego zespołu pokazały, że zaproponowany algorytm tensorowy był w stanie podzielić wszystkie, różniące się od siebie, fałszywe wiadomości w zebranym korpusie tekstu na następujące klastry tematyczne: stronniczość, nienawiść, teorie spiskowe, satyra, śmieciowa nauka, wiadomości państwowe. Stronniczość dotyczyła wiadomości, które promują propagandę polityczną i rażące zniekształcanie faktów; nienawiść grupowała wiadomości aktywnie promujące wszelkiego rodzaju dyskryminację; teorie spiskowe to grupa wiadomości będących promotorami szeroko pojętego konspiracjonizmu; satyra to źródła, które dostarczają humorystycznych komentarzy do bieżących wydarzeń w formie fake newsów; śmieciowa nauka skupia wiadomości promujące pseudonaukę, metafizykę, błędy naturalistyczne i inne wątpliwe z naukowego punktu widzenia twierdzenia; wreszcie klaster wiadomości państwowe grupuje wiado-

mości dystrybuowane w represyjnych reżimach przez oficjalne media działające pod presją sankcji rządowych. Wykonane przez Hosseinimotlagha i innych klastrowanie odznaczało się wysoką jednorodnością poszczególnych grup (średnia jednorodność na grupę do 80%), a różnorodność wartości odstających były zredukowane do 2,5%.

3. Wykrywanie tematów

Treść fałszywych wiadomości bywa znacznie zróżnicowana pod względem tematów, stylów pisania i używanego słownictwa – w tym wypadku metody klasyfikacji i klastrowania raczej nie dadzą istotnych wyników. Z pomocą przychodzi jednak metoda wykrywania tematów. Przykładowo, analiza text miningowa przeprowadzona przez NATO StratCom w 2018 r. próbowała ocenić wpływ rosyjskiego trollingu na rozpowszechnianie dezinformacji w skandynawsko-bałtyckim środowisku informacyjnym [41]. Badaniu poddano 3671 treści artykułów na temat aneksji Krymu we wschodniej Ukrainie, zamieszczone na różnych portalach internetowych w języku rosyjskim, litewskim, łotewskim, estońskim i polskim. Przebadano także wszystkie komentarze pod tymi artykułami. Analiza wykazała, że komentarze tzw. trolli do danego artykułu były zwykle związane ze wzrostem ilości komentarzy publikowanych przez użytkowników nie będących trollami, jeśli ci ostatni inicjowali specyficzne wątki dyskusji, wykryte również w toku analizy wykrywania tematów. Na przykład, gdy poprzez dezinformację dany tekst podsycał hejt na tle sporów narodowych, etnicznych i wyznaniowych – właśnie takie artykuły częściej komentowały trolle, w efekcie eskalując mowę nienawiści. Wskazuje to na tendencję do wyszukiwania przez agentów wpływu swoistych zarodków konfliktów, oscylujących wokół danego tematu, na których można oprzeć całą strategię dalszej dezinformacji.

Z kolei Gautam i inni wykorzytali zbiór danych fałszywych wiadomości dotyczących COVID-19 [42]. Był to ręcznie oznaczony zbiór danych zawierający 10700 postów opublikowanych w mediach społecznościowych oraz artykułów z prawdziwymi i fałszywymi wiadomościami dotyczących COVID-19. Artykuły z fałszywymi wiadomościami były zbierane z kilku witryn i narzędzi do sprawdzania faktów, podczas gdy prawdziwe i wiarygodne artykuły były zbierane z Twittera za pomocą zweryfikowanych narzędzi samego Twittera. Używając architektury sieci neuronowej XLNet z probabilistycznym modelowaniem tematycznym LDA badacze wykazali, że wykrywanie tematów (w tym wypadku związanych z COVID-19) w zbiorze treningowym przyczyniło się do zwiększenia możliwości wykrycia fałszywych wiadomości. Rozkłady prawdopodobieństwa tematów zwiększały moc dyskryminacyjną modelu, dzięki czemu ta metoda okazała się bardzo wydajna do dalszego wnioskowania.

Różnice lub podobieństwa tematów między wiadomościami oznaczonymi jako fałszywe i prawdziwe były badane także przez Xu i in. [43]. Zbiór danych składał się z czterdziestu – prawdziwych i fałszywych – bardzo popularnych trendów wiadomości zebranych w 2016 roku w okresie trzech miesięcy z dziesiątek serwisów informacyjnych: z tych znanych ze swojego dezinformacyjnego charakteru, ale także z szanowanych głównych serwisów informacyjnych, takich jak New York Times, Washington Post, NBC News, USA Today i Wall Street Journal. Analiza Xu i in., poza zastosowaniem modelowania LDA, była uzupełniona o zbadanie podobieństwa dokumentów spośród fałszywych, prawdziwych i hybrydowych artykułów informacyjnych. Tematy wykryte w fałszywych wiadomościach (w badanym zbiorze danych),

skupiały się głównie na trzech wątkach: nazwiskach rządzących (np. prezydentów, członków parlamentu), określeniach dotyczących pojęć abstrakcyjnych (np. czas, informacja, stan) oraz wykonywanych czynnościach (np. powiedział, będzie, kandyduje). Tematy wykryte w wiadomościach prawdziwych skupiały się na wątkach takich jak: ważne osoby (zarówno nazwiska, jak i pełnione funkcje np. prezydent, oraz podmiot zbiorowy „ludzie"), źródła informacji (np. facebook, ludzie, nazwiska polityków) oraz przymiotniki dotyczące popieranej partii (np. republikanie, demokraci). Wreszcie hybrydowy zbiór tekstów pozwolił na wyłonienie następujących tematów: dotyczących osób (ludzie, politycy, Amerykanie), nazwisk prezydentów i partii z której się wywodzą, informacji związanych z konkretną kontrowersyjną osobą (np. Donald Trump). Badanie to pokazuje, jak trudno jest interpretować wyniki wykrywania tematów w celu skutecznego identyfikowania fałszywych wiadomości – w tym zakresie istnieje niewielka, bardzo subtelna różnica między fałszywymi a prawdziwymi wiadomościami. Jednocześnie odsłania ono obiecujący aspekt wykorzystania podobieństwa dokumentów do rozróżniania fałszywych i prawdziwych wiadomości poprzez pomiar podobieństwa testowanych dokumentów do wcześniej opracowanego korpusu fałszywych i prawdziwych wiadomości.

4. Analiza sentymentu

Posty i komentarze użytkowników mediów społecznościowych, a także wpisy na blogach, są bogatym źródłem informacji o samych użytkownikach. Czytając ich treści nie tylko zapoznajemy się z opinią autorów w danej kwestii, ale również poznajemy ich postawy wobec danego zjawiska wyrażone w wydźwięku emocjonalnym, poziomie zaangażowania emocjonalnego i nastrojach, które są przydatne w wykrywaniu fałszywych wiadomości i dezinformacji. Twórcy fake newsów często stosują różne chwyty stylistyczne, by zapewnić jak najszersze rozpowszechnianie swoich wpisów, a jednym z takich zabiegów jest właśnie apelowanie do emocji odbiorców. Zauważenie tej prawidłowości skłoniło niektórych naukowców do użycia metody analizy sentymentu do wykrywania dezinformacji i fałszywych wiadomości.

W badaniach nad dezinformacją w mediach społecznościowych Shu i in. wykazali, iż silna polaryzacja sentymentu postów i komentarzy odznacza się wysokim prawdopodobieństwem ich fałszywości [44]. Również Bhutani i in. udowodnili w swoich badaniach, że sentyment wyrażony podczas pisania wiadomości lub artykułu informacyjnego może służyć jako kluczowy czynnik decydujący w procesie identyfikowania wiadomości jako fałszywych lub prawdziwych [45]. Zastosowali oni klasyfikator Naive-Bayes, aby określić sentyment tekstów, a następnie wykorzystali go jako główną cechę klasyfikatorów Multinomial Naive-Bayes i Random Forest do wykrywania fałszywych wiadomości, przy czym ten ostatni uzyskiwał najlepsze wyniki na poziomie dokładności 84,3%. Podobnie, analizując wydźwięk emocjonalny w treściach mikroblogów, wiadomości ze sprzecznymi punktami widzenia najczęściej okazywały się fake newsami według badań Jin i in. [46]. Cui i in. Badali także, czy ukryty (niejawny) wydźwięk emocjonalny zaszyty w komentarzach użytkowników w mediach społecznościowych wspiera odróżnianie fałszywych wiadomości od wiarygodnych treści [47]. Walidacja ich wielomodalnego modelu klasyfikacyjnego Sentiment-Aware Multi-Modal Embedding przy użyciu dwóch rzeczywistych zbiorów danych, PolitiFact (serwis dziennikarski weryfikujący fakty, sprawdzając konkretne wypowiedzi polityków)

i GossipCop (plotki dotyczące celebrytów), pokazała skuteczność na poziomie ok. 90% w wykrywaniu fałszywych wiadomości.

Sprzeczne informacje lub sprzeczne punkty widzenia mają kluczowe znaczenie do weryfikacji wiadomości, jednak wciąż trudno je wykryć, zwłaszcza w przypadku analiz dziesiątek lub setek stron zagregowanych treści. Teksty takie są często podparte wypowiedziami naukowców: autorami dezinformacyjnych artykułów nierzadko są ludzie z tytułami naukowymi, jednak uprzednio odsunięci od pracy na uniwersytetach, pozbawieni możliwości publikacji lub wcześniej demaskowani jako propagatorzy dezinformacji, wypowiadający się poza swoją utytułowaną dziedziną, ekstrapolując sam tytuł naukowy do podniesienia wiarygodności treści. Naukowcy ci przedstawiani są często jako autorytety, ponieważ wpasowują się nie tylko w jedną z zasad wywierania wpływu (odwołanie do autorytetu), ale również w popularną pułapkę logiczną, jaką jest błąd konfirmacji, a więc przyjęcie za fakt takiej informacji, która potwierdza preferowane przez nas założenie. Zastosowana w takim przypadku analiza sentymentu pozwala wyłonić ukryty – w pozornie neutralnym tekście – negatywny wydźwięk emocjonalny, manipulujący odbiorem poznawczym czytelnika.

Różnice między fałszywymi a prawdziwymi wiadomościami w ramach analizy sentymentu badali także Horne i Adali na zbiorze danych Silverman's Buzzfeed Political News Data [48]. Analizowali oni wiadomości polityczne pod kątem polaryzacji sentymentu dla każdego zdania, jego cechy stylistyczne (częstość każdej części mowy w artykule, interpunkcję, cytaty, negacje, słowa nieformalne i przekleństwa, pytania, słowa pisane z wielkiej litery), a także złożoność słowa (liczba sylab w słowach, stosunek unikalnych słów oraz liczba popularnych słów i specjalistycznego słownictwa) oraz złożoność zdania (liczba słów w zdaniu, głębokość drzewa składni zdania dla fraz rzeczownikowych i czasownikowych). Odkryli oni, że pozytywne i negatywne nastroje były statystycznie istotnymi cechami odróżniającymi treść prawdziwych i fałszywych wiadomości w zestawie danych dotyczących wiadomości politycznych. Rezultaty te nie powtórzyły się, gdy brano pod uwagę tylko nagłówki wiadomości – prawdopodobnie były one zbyt krótkie dla rozważenia analizy sentymentu[49].

5. Podsumowanie

Pomimo zdolności obronnych człowieka wobec wpływu na zachowania i postawy, często nie mamy szans na utrzymanie homeostazy w dzisiejszym cyfrowym świecie, zdominowanym przez skrajnie stronnicze treści, szum informacyjny, dezinformację i fake newsy. W najlepszym razie – stajemy się ofiarami mniej lub bardziej wyrafinowanej manipulacji, zamykając się i radykalizując w informacyjnych bańkach. W najgorszym – możemy stać się narzędziem w rękach skrajnych ugrupowań, drastycznie odcinając się od zróżnicowanych treści, prezentując mowę nienawiści, powielając szkodliwe i fałszywe treści, lub posunąć się jeszcze dalej, do czynów o charakterze terrorystycznym.

Kroki zaradcze podejmowane przez instytucje państwowe, badaczy wielu dziedzin, decydentów w sektorze prywatnym oraz pracowników oświaty i aktywistów – edukatorów są widoczne w przestrzeni i debacie publicznej oraz docierają do coraz szerszego grona odbiorców w każdej grupie wiekowej. Dalszy postęp w wykrywaniu dezinformacji i fake newsów, a także rozwój coraz lepszych metod i narzędzi do pokonywania dezinformacji oraz prewencja i edukacja w tych zakresach jest jednak

nadal palącą kwestią społeczną, ze względu na rosnącą złożoność, różnorodność i multimodalność fałszywych informacji. Dostępne są już częściowo zautomatyzowane rozwiązania Text Mining, związane z przetwarzaniem języka naturalnego, uczeniem maszynowym i różnymi technikami wykrywania fake newsów i przeciwdziałania dezinformacji. Zautomatyzowane rozwiązania, choć nie zastępują ludzkiej ekspertyzy, pomagają w ocenie prawdziwości i wiarygodności napotkanych w internecie treści. Narzędzia te służą jednak bardziej zaawansowanym użytkownikom, biegle posługującym się wiedzą programistyczną i są nastawione na wsparcie mechanizmów wykrywania dezinformacji i fake newsów przez platformy społecznościowe, fact-checkingowe, czy na potrzeby mediów. Brakuje powszechnych, darmowych rozwiązań, z których mógłby korzystać każdy użytkownik internetu na co dzień. Dostępne obecnie narzędzia, a więc uproszczone interfejsy zaawansowanych technik Text Miningu są często płatne lub wciąż wymagają choćby podstawowej wiedzy programistycznej.

Połączenie wysiłków ekspertów z dziedzin bezpieczeństwa, kryminologii, kognitywistyki, psychologii, socjologii, cyberbezpieczeństwa, informatyki, sztucznej inteligencji, Data Science i Text Mining w ramach wschodzącej dziedziny bezpieczeństwa poznawczego człowieka, może stanowić przełom w podejściu do przeciwdziałania dezinformacji, fake newsom, a także szeroko pojętym przestępstwom internetowym, których wspólnym mianownikiem jest manipulacja i znajomość mechanizmów poznawczych człowieka. Tak zaimplementowana interdyscyplinarność jest w naszej opinii nie tylko kluczem do dalszego rozwoju metod przeciwdziałania nowym zagrożeniom i dogłębnej ich znajomości, ale również do tak potrzebnego przejścia z technik defensywnych i naprawczych do ofensywnych i prewencyjnych.

Literatura

1. Friedman D.A., Cordes R.J., *Knowledge Management Archipelago*, Zenodo 2021, s. 2-3.
2. Waltzman R., *Proposal for a Center for Cognitive Security*, Information Professional Association 2015.
3. https://www.cogsec.org/ [data dostępu: 04.2022].
4. Cordes R.J., Applegate-Swanson S., Friedman D.A., Knight V.B., Mikhailova A., *Narrative Information Management*, [w:] Cordes R.J., Friedman D.A., *Narrative Information Ecosystems. Conflict and Trust on the Endless Frontier*, Cognitive Security & Education Forum (COGSEC), 2021, s. 23.
5. Rydning J., Reinsel D., *Worldwide Global StorageSphere Forecast, 2021–2025: To Save or Not to Save Data, That Is the Questio*n, IDC, 2021, s. 1-29.
6. Fox J.R., Park B., Lang A., *When Available Resources Become Negative Resources: The Effects of Cognitive Overload on Memory Sensitivity and Criterion Bias*, Communic Res., 34, 2007, s. 277-296.
7. Konopka B., *Szum informacyjny i jego rola w kształtowaniu warunków medialnych i kulturowych*, Transformacje, nr 1-2, 2020, s. 167-184.
8. Szynkiewicz M., *Metafora smogu informacyjnego a procesy informacyjne*, Studia metodologiczne 2014, nr 32, Wydawnictwo Naukowe UAM, 2014, s. 65-77.
9. Fallis D., *What Is Disinformation?*, Library Trends, vol. 63 no. 3, 2015, s. 401-426.
10. Bangyal W.H., Qasim R., Rehman N., Ahmad Z., Dar H., Rukhsar L., Aman Z., Ahmad J., *Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches*, Computational and Mathematical Methods in Medicine, vol. 2021, s. 1-14.

11. Helmus T.C., Bodine-Baron E., Radin A., Magnuson M., Mendelsohn J., Marcellino W., Bega A., Winkelman Z., *Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe*, Santa Monica, CA: RAND Corporation, 2018.
12. https://wiadomosci.wp.pl/oto-dlaczego-wylaczylismy-komentarze-pod-tekstami-o-inwazji-rosji-na-ukraine-6740816216714208a [data dostępu: 04.2022]
13. https://www.wirtualnemedia.pl/artykul/agresja-rosja-na-ukraina-wirtualna-polska-wylaczyla-komentarze-newsy-artykuly-interia-blokada [data dostępu: 04.2022]
14. https://www.walbrzych24.com/artykul/35368/wylaczylismy-komentarze-pod-artykulami-o-ataku-rosji-na-ukraine [data dostępu: 04.2022]
15. https://www.gov.pl/web/rcb/badz-gotowy--poradnik-na-czas-kryzysu-i-wojny [data dostępu: 04.2022]
16. Łódzki B., *Fake news – dezinformacja w mediach internetowych i formy jej zwalczania w przestrzeni międzynarodowej*, Polityka i Społeczeństwo, nr 4 (15), 2017, s. 19-30.
17. Niklewicz K., *Weeding out fake news: An Approach to Social Media Regulation*, Wilfried Martens Centre for European Studies, 2017, s.15-16.
18. https://wsjp.pl/haslo/podglad/40953/pozyteczny-idiota [data dostępu: 04.2022]
19. Puzio M., *Przekazy medialne jako mechanizmy kształtowania społecznej oceny polityków na przykładzie polskich wyborów prezydenckich w 2015 roku*, Studia Wyborcze, tom 28, Łódź 2019.
20. Wojciszke B., *Człowiek wśród ludzi*, Wydawnictwo Naukowe „Scholar" , Warszawa 2003.
21. Pariser E., *The Filter Bubble: What the Internet is Hiding from You*, Penguin Press, New York 2011.
22. Klausen J., *A Behavioral Study of the Radicalization Trajectories of American "Homegrown" Al Qaeda-Inspired Terrorist Offenders*, U.S. Department of Justice, Office of Justice Programs, Brandeis 2016.
23. Ledwich M., Zaitsev A., *Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization*, arXiv preprint arXiv:1912.11211, 2019.
24. https://www.nytimes.com/2019/03/29/technology/youtube-online-extremism.html [data dostępu: 04.2022]
25. Woolley K., Sharif M.A., *Down a Rabbit Hole: How Prior Media Consumption Shapes Subsequent Media Consumption*, Journal of Marketing Research 2022.
26. Kenrick D., Neuberg S., Cialdini R., *Psychologia społeczna*, GWP, Gdańsk 2002.
27. Cialdini R., *Wywieranie wpływu na ludzi. Teoria i praktyka*, GWP, Gdańsk 2000.
28. Opinia publiczna, [w:] Nowotny S., *Encyklopedia socjologii*, Warszawa 2000, t. 3; Opinia publiczna, [w:] Marshall G. (red.), *Słownik socjologii i nauk społecznych*, Warszawa 2004.
29. Deutsch M., Gerard H., *A study of normative and informational social influences upon individual judgment, Journal of Abnormal and Social Psychology*, 1955.
30. Goban-Klas T., *Media i komunikowanie*, PWN, Warszawa 2004.
31. Lepa A., *Etyczne i wychowawcze aspekty opinii publicznej*, Etyka w Życiu Gospodarczym, 2009.
32. Kuśmierski S., *Świadomość społeczna – opinia publiczna – propaganda*, Warszawa 1987.
33. Jaroszyński Cz., *Podstawy retoryki klasycznej*, Warszawa 1998.
34. Gentzkow M., Kelly B., Taddy M., *Text as Data*, Journal of Economic Literature, 57 (3), 2019, s. 535.
35. Allahyari M., Pouriyeh S., Assefi M., Safaei S., Trippe E.D., Gutierrez J.B., Kochut K., *A brief survey of text mining: Classification, clustering and extraction techniques*, arXiv preprint arXiv:1707.02919, 2017, s.4.
36. Ahmed S., Hinkelmann K. , Corradini F., *Development of Fake News Model Using Machine Learning through Natural Language Processing*, World Academy of Science,

Engineering and Technology, Open Science Index 168, International Journal of Computer and Information Engineering, 14(12), 2020, s.454-460.
37. Ott M., Choi Y., Cardie C., Hancock J.T., *Finding deceptive opinion spam by any stretch of the imagination*, w: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, Association for Computational Linguistics, 2011, s. 309-319.
38. Kim M.G., Kim M., Kim J.H., Kim K., *Fine-Tuning BERT Models to Classify Misinformation on Garlic and COVID-19 on Twitter*, Int. J. Environ. Res. Public Health 19, 5126, 2022, https://doi.org/10.3390/ijerph19095126.
39. Yazdi K., Yazdi A., Khodayi S., Hou J., Zhou W., Saedy S., *Improving Fake News Detection Using K-means and Support Vector Machine Approaches*, World Academy of Science, Engineering and Technology, Open Science Index 158, International Journal of Electronics and Communication Engineering, 14(2), 2020, s. 38-42.
40. Hosseinimotlagh S., Papalexakis E.E., *Unsupervised Content-Based Identification of Fake News. Articles with Tensor Decomposition Ensembles*, University of California Riverside 2018.
41. Bērziņa I., Cepurītis M., Kaljula D., Juurvee I., *Russia's Footprint in the Nordic-Baltic Information Environment. Executive Summary*, Riga: NATO Strategic Communications Centre of Excellence 2018.
42. Gautam A., Venktesh V., Masud S., *Fake news detection system using XLNet model with topic distributions: CONSTRAINT@AAAI2021 shared task*, [w:] Chakraborty T., Shu K., Bernard H.R., Liu H., Akhtar M.S., *Combating online hostile posts in regional languages during emergency situation*, vol 1402, Springer, Cham, 2021, s.189-200.
43. Xu K., Wang F., Wang H., Yang B., *Detecting fake news over online social media via domain reputations and content understanding*, Tsinghua Science and Technology, vol. 25, no. 1, 2020, s. 20-27, doi: 10.26599/TST.2018.9010139.
44. Shu K., Sliva A., Wang S., Tang J., Liu H., *Fake news detection on social media: A data mining perspective*, KDD exploration newsletter, 2017.
45. Bhutani B., Rastogi N., Sehgal P., Purwar A., *Fake News Detection Using Sentiment Analysis*, w: *Proceedings of the 2019 Twelfth International Conference on Contemporary Computing*, IC3 2019, Noida, India, 8–10 August 2019; IEEE: Piscataway, NJ, USA 2019.
46. Jin Z., Cao J., Zhang Y., Luo J., *News verification by exploiting conflicting social viewpoints in microblogs*, AAAI 2016.
47. Cui L., Wang S., Lee D., *SAME: Sentiment-aware multi-modal embedding for detecting fake news*, 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2019.
48. Horne B.D., Adali S., *This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News*, w: *Proceedings of the Workshops of the Eleventh International AAAI Conference on Web and Social Media* (ICWSM 2017), Montreal, QC, Canada, 15–18 May 2017; An J., Kwak H., Benevenuto F., AAAI Press: Palo Alto, CA, USA, 2017, Volume AAAI Technical Report WS-17-17: News and Public Opinion, pp. 759–766.
49. Alonso M.A., Vilares D., Gómez-Rodríguez C., Vilares J., *Sentiment Analysis for Fake News Detection*, Electronics 10-1348, 2021, https://doi.org/10.3390/electronics10111348.

Text Mining w przeciwdziałaniu dezinformacji: bezpieczeństwo poznawcze człowieka

Streszczenie

Pomimo niezaprzeczalnie pozytywnej roli internetu w przekształcaniu świata w globalną wioskę, internet jest także kanałem, poprzez który każdego dnia ludzkość zalewana jest informacjami, w tym także tymi nieprawdziwymi. Zmanipulowane treści dezinformacyjne wykorzystują mechanizmy uprzedzeń poznawczych, wzorców behawioralnych lub emocjonalnych oraz nawyki użytkowników, dlatego są tak trudne do rozróżnienia od faktów. Celem niniejszego rozdziału jest analiza tych kwestii zarówno z perspektywy bezpieczeństwa poznawczego, jak i dziedziny Text Mining, wskazująca na potrzebę interdyscyplinarnego podejścia do przeciwdziałania dezinformacji. Wiedza o metodach Text Mining jest nadal mało znana szerokiemu gronu odbiorców, tak samo jak rodząca się dopiero dziedzina bezpieczeństwa poznawczego. Rezultat mariażu tych dziedzin w kontekście przeciwdziałania dezinformacji może przyczynić się do przejścia z technik defensywnych i naprawczych do ofensywnych i prewencyjnych.

Słowa kluczowe: bezpieczeństwo poznawcze, text mining, przeciwdziałanie dezinformacji, wykrywanie fake newsów

Text Mining in overcoming disinformation: human cognitive security

Abstract

Despite the undeniably positive role of the internet in transforming the world into a global village, the internet is also a channel through which humanity is flooded with information every day, including misleading and fake one. Manipulated disinformation content exploits human cognitive biases, behavioral or emotional patterns, and user habits, which is why it is so difficult to distinguish it from the facts. The aim of this chapter is to analyze these issues both from the perspective of cognitive security and the field of Text Mining, pointing to the need for an interdisciplinary approach to counteracting disinformation. The knowledge of Text Mining methods is still little known to a wider audience, as is the emerging field of cognitive security. The result of merging these domains in the context of counteracting disinformation may contribute to the shift from defensive and corrective techniques to offensive and preventive ones.

Keywords: human cognitive security, text mining, overcoming disinformation, fake news detection