


Michał Żytomirski

Katedra Informatologii i Bibliologii
Wydział Filologiczny Uniwersytetu Łódzkiego
e-mail: michal.zytomirski@gmail.com

 <https://orcid.org/0000-0003-1400-9949>

Określanie osobowości użytkowników Internetu poprzez analizowanie ich cyfrowych śladów w świetle wybranych badań dr. Michała Kosińskiego

DOI: <https://doi.org/10.18778/0860-7435.32.05>

Abstrakt: Według danych Internet Live Stats w kwietniu 2021 roku w czasie każdej sekundy zadawano około 92 tysięcy zapytań w wyszukiwarce Google. Każda czynność dokonywana przez użytkowników urządzeń cyfrowych jest indeksowana jako tak zwane cyfrowe ślady, dzięki którym możliwe jest, przy zastosowaniu odpowiednich technologii oraz metod, precyzyjne określanie cech osobowości, poglądów politycznych oraz orientacji seksualnych tychże użytkowników.

Artykuł został zainspirowany pracami dr. Michała Kosińskiego i stanowi opis problematyki związanej z analizowaniem cyfrowych śladów użytkowników Internetu (głównie mediów społecznościowych).

Nadrzędnym celem artykułu jest przedstawienie badań Michała Kosińskiego i zwrócenie uwagi środowiska informatologicznego na kwestie związane z analizowaniem cyfrowych śladów użytkowników Internetu. Praca ta ma charakter popularyzatorski, nie stanowi całościowego opisu dokonanych wskazanego naukowca. Artykuł nie przedstawia nowych informacji ani badań własnych – ma jednak zachęcić odbiorców do przeanalizowania literatury przedmiotu dotyczącej analizowania cyfrowych śladów oraz prywatności w dobie cyfrowej.

W pracy wykorzystano metodę analizy zawartości baz danych – do zebrania i przeanalizowania literatury przedmiotu. Skupiono się na zasobach udostępnianych przez dr. Kosińskiego poprzez prywatną stronę internetową – <https://www.michalkosinski.com>. Dodatkowo wykorzystano techniki związane z *data miningiem*, aby w podsumowaniu móc przedstawić archiwalne treści publikowane na stronie internetowej firmy Cambridge Analytica.

Zasięg chronologiczny odszukiwanych materiałów piśmienniczych zawężono do okresu od 2011 r. do pierwszego kwartału 2021 r., skupiając się głównie na latach 2013–2021, czyli od roku opublikowania artykułu *Private traits and attributes are predictable from digital records of human behavior* do roku publikacji tekstu *Facial recognition technology can expose political orientation from naturalistic facial images*. W artykule przedstawiono wybrane prace dr. Michała Kosińskiego, które stanowią „kamienie milowe” w badaniach nad określaniem osobowości użytkowników Internetu w obrębie publikacji wskazanego autora. Co za tym idzie, w artykule nie przedstawiono tych prac, które były opisem części badań lub były przyczynkiem do podjęcia większych, dalszych pomiarów.

Słowa kluczowe: cyfrowe ślady, mierzenie cech użytkowników, Five Factor Model, prywatność, Michał Kosiński

Wstęp

Cyfrowe ślady użytkowników Internetu to „zmiany w kodzie binarnym systemu teleinformatycznego, a także urządzenia cyfrowego zdolnego do przetwarzania, wysyłania, gromadzenia pakietów danych, będących wynikiem ingerencji zewnętrznej (fizycznej) bądź wewnętrznej (zdalnej)” (Kasprzak, 2015, s. 25). Oznacza to, że działania użytkowników są indeksowane zarówno wewnątrz wykorzystywanych urządzeń, jak i na zewnątrz, w ramach całych systemów teleinformatycznych (np. w obrębie serwerów dostawców usług). Również przez te zależności ukuto hipotezę o braku możliwości pełnego usunięcia swoich śladów z cyberprzestrzeni¹.

W skali globalnej coraz więcej działań dokonywanych jest przez ludzi w pełni lub częściowo przy wykorzystaniu narzędzi cyfrowych – dotyczy to zarówno interakcji międzyludzkich, rozrywki, zakupów, jak i pracy czy wyszukiwania informacji (Matz, Kosiński, 2017, s. 19). Analizowanie cyfrowych

¹ Def. 1.: Globalna domena środowiska informacyjnego składająca się z współzależnych sieci tworzonych przez infrastrukturę technologii informacyjnej (IT) oraz zawartych w nich danych, włączając Internet, sieci telekomunikacyjne, systemy komputerowe, a także osadzone w nich procesory oraz kontrolery (Wasilewski, 2013, s. 227). Def. 2.: Zbiór wszystkich urządzeń cyfrowych zdolnych do wzajemnej komunikacji i przesyłania, odbierania, gromadzenia pakietów danych (Kasprzak, 2015, s. 33).

śladów pozwala coraz precyzyjniej określać potrzeby, cechy i zachowania użytkowników. Globalna skala zjawiska oraz zakres przeprowadzanych analiz pozwala na określenie podobieństw i różnic między użytkownikami, dzięki czemu dane deklarowane (deklaratywne) przez użytkowników, np.: wiek, płeć, orientacja seksualna czy poglądy polityczne, mogą być porównywane do realnych zachowań (użytkowników), a następnie mogą podlegać sprawdzaniu pod kątem zgodności ze stanem faktycznym.

Autor pracy skupił się na przeanalizowaniu publikacji Michała Kosińskiego, ponieważ uznaje go za autorytet z zakresu analiz psychometrycznych użytkowników Internetu. Prace Kosińskiego są często cytowane i cieszą się znacznym uznaniem w środowisku naukowym. Z danych Google Scholar wynika, że w okresie od 2014 roku do 2021 roku były one cytowane 13455 razy, tym samym autor posiada h-index na poziomie 47 punktów (*Michał Kosinski, 2021*).

W celu weryfikacji literatury przedmiotu do artykułu, przeanalizowano bibliografię Michała Kosińskiego opublikowaną na stronie internetowej <https://www.michalkosinski.com/research> i zawężono ją do okresu od 2013 roku (od publikacji artykułu *Private traits and attributes are predictable from digital records of human behavior*) do kwietnia 2021 roku. Tym samym uzyskano łącznie 85 publikacji autorstwa M. Kosińskiego. W artykule przedstawiono publikacje, które stanowiły podsumowania badań², w ramach publikacji dr. Kosińskiego, nad analizowaniem cyfrowych śladów użytkowników Internetu w celu określania ich cech psychologicznych. Należy podkreślić, że tematyka dobranych publikacji opiera się na analizowaniu cyfrowych śladów behawioralnych, takich jak informacje o odwiedzanych stronach internetowych przez użytkowników lub informacje o obserwowanych profilach w mediach społecznościowych przez użytkowników. Dla przykładu artykuł *Personality, gender, and age in the language of social media: The open-vocabulary approach* nie został uwzględniony w badaniach, ponieważ opisuje on proces analizowania treści wiadomości użytkowników Facebooka.

Analiza literatury przedmiotu

W 2013 r. badacze z Centrum Psychometrii Uniwersytetu Cambridge stwierdzili, że można precyzyjniej określić cechy ludzi mierząc ich interakcje na portalu Facebook (FB) (wykorzystywanie reakcji „lubię to”) (Kosinski i in., 2013). Oznaczało to, że na podstawie analiz cyfrowych śladów użytkowników było możliwe określenie potrzeb oraz cech użytkowników. W ramach

² W pracy nie przedstawiano artykułów, które stanowiły wstęp do większych pomiarów lub zostały rozbudowane na dalszych etapach postępowania procesów naukowych.

chronologicznej dokładności należy zaznaczyć, że artykuł ten był poprzedzony latami innych badań, z których wnioski pozwoliły m.in., na stwierdzenie, że:

- możliwe jest kategoryzowanie użytkowników mediów społecznościowych ze względu na ich cechy osobowości (Quercia i in., 2011),
- poprzez określanie cech osobowości użytkowników mediów społecznościowych można wyraźnie określić ich podobieństwa i różnice (Tamże),
- określanie osobowości użytkowników może być dokonywane w prosty i mierzalny sposób przy wykorzystaniu danych dostępnych publicznie (na podstawie danych deklaracyjnych pozostawianych przez użytkowników – np. opublikowanie postu, polecenie piosenki, itp.) (Tamże),
- cechy osobowości użytkowników są skorelowane ze sposobami wykorzystywania przez nich mediów społecznościowych (Bachrach i in., 2012),
- cechy osobowości osób popularnych na Facebooku nie różnią się znacząco od cech osób popularnych poza Internetem, co pozwala sugerować, że część zachowań użytkowników odpowiada zachowaniom, które przedstawiają w relacjach społecznych poza Siecią (Quercia i in., 2012).

Należy dodatkowo odnotować, że od 2007 do 2012 r. badacze z Centrum Psychometrii Uniwersytetu Cambridge prowadzili badania w ramach projektu myPersonality (Stillwell & Kosinski, 2012). W projekcie wzięło udział, dobrowolnie poprzez aplikację, ponad 6 milionów użytkowników portalu społecznościowego Facebook (Stillwell, 2021). Prawdopodobnie to właśnie ten projekt pozwolił w następnych latach ustalić precyzyjne metody określania cech – w tym cech politycznych – użytkowników Internetu na podstawie ich cyfrowych śladów.

W artykule *Personality and Patterns of Facebook Usage*, który poprzedzał pracę *Private traits and attributes are predictable from digital records of human behavior*, autorzy oparli swoje pomiary o dane pochodzące z projektu myPersonality, badając dane losowej grupy 250 użytkowników Facebooka. W pomiarach uwzględniono dane dotyczące aktywności (10 tysięcy statusów aktywności) oraz dane demograficzne (wiek i płeć) (Bachrach i in., 2012). W ramach tego badania naukowcy stwierdzili znaczne ograniczenia pomiarów, związane z błędami statystycznymi (błąd własnego wyboru, ang. *self-selection bias*) oraz z analizowaniem niepełnego zbioru danych (analizowanie samych polubień, a nie zawartości „lubianych” treści) (Tamże). Autorzy artykułu zostawili więc otwarte pytanie badawcze o możliwość ustalania cech osobowości w oparciu o szczegółowe cyfrowe ślady.

Artykuł *Private traits and attributes are predictable from digital records of human behavior* jest zatem kontynuacją serii poprzednich badań, badacze dopracowali metodykę badań (połączenie modeli psychometrycznych z pracą algorytmu) oraz rozszerzyli grupę badawczą. Przebadano grupę 58 tysięcy użytkowników portalu Facebook w ramach kwestionariusza FFM (Five Factor Model – badanie osobowości w ramach 5 czynników). Badania Five Factor Model³, cech zwanych także Big Five (wielka piątka) stanowiły podstawę metodyki tych, jak i poprzednich badań:

Wykorzystujemy model pięcioczynnikowy, który jest obecnie najbardziej rozpowszechnionym i ogólnie akceptowanym modelem osobowości, którego zdolność do przewidywania ludzkich zachowań została dobrze zbadana (Bachrach i in., 2012).

Model ten zakłada mierzenie cech respondentów w odniesieniu do pięciu czynników, w ramach których określa się odpowiednie cechy:

1. Ekstrawersję (np. skłonność do mówienia, poczucie humoru).
2. Ugodowość (np. umiejętność wybaczenia, zaufanie).
3. Sumiennność (np. umiejętność organizacji, produktywność).
4. Neurotyczność (np. skłonność do użalania się nad sobą, depresyjność).
5. Otwartość (np. szerokie spektrum zainteresowań, artystyczność) (McCrae & John, 1992).

Ad 1. Jednostki o wysokim współczynniku *ekstrawersji* zazwyczaj poszukują stymulacji w świecie zewnętrznym, potrzebują towarzystwa innych ludzi. Zazwyczaj są to osoby bardziej otwarte, doświadczające pozytywnych emocji, są też częściej aktywne społecznie. Przejawiają również łatwość w nawiązywaniu nowych kontaktów i nie przeszkadza im bycie w centrum uwagi (społecznej).

Ad 2. Ludzie o wysokim współczynniku *ugodowości* są zazwyczaj ugodowi i przyjaźni, przejawiają problemy w wyjawianiu tak zwanej trudnej prawdy. Osoby te są skłonne do współpracy, zazwyczaj ufają innym ludziom.

Ad 3. Osoby o wysokim poziomie *sumienności* są zorganizowane i rzadko przejawiają działania spontaniczne, przejawiają tendencje do planowania długoterminowego i cechują się konsekwencją w działaniu. Ludzie klasyfikowani do tej kategorii są zazwyczaj mniej tolerancyjni, są związani procedurami i zasadami.

Ad 4. Jednostki o wysokim współczynniku *neurotyczności* są niestabilne emocjonalnie, przejawiają predyspozycje do wahań nastrojów, negatywnych

³ Wykorzystywanie modelu Wielkiej Piątki do badań nad cyfrowymi śladami przedstawiono szerzej w artykule *Mining Facebook data for predictive personality modeling*, (Markovikj i in., 2013).

emocji, obwiniania się, jak i depresji. Osoby o tej cesze częściej doświadczają stresu i nerwowości.

Ad 5. Osoby indeksowane w tej kategorii przejawiają cechy kreatywne, interesują się kulturą, sztuką i przejawiają zainteresowanie estetyką. Cechują się wrażliwością, tolerancją i liberalizmem. Osoby z niskim współczynnikiem *otwartości* są zazwyczaj mniej kreatywne, są bardziej konwencjonalne i autorytarne. Dodatkowo unikają zmian i przejawiają cechy konserwatywne i tradycyjne (Kosinski i in., 2014).

Pomiary wykonywane tym modelem często przedstawiane są w ramach skali przeciwstawieństw, np. ekstrawersja-introwersja, ugodowość-nieugodowość, itd. Dla przykładu – w ramach badań Apply Magic Sauce, prowadzonych przez Instytut Psychometrii Cambridge, będących rozwinięciem badań w ramach projektu myPersonality, skala pomiarów cech jest zmodyfikowaną skalą FFM i wygląda następująco:

1. Konserwatywny i tradycyjny – liberalny i artystyczny.
2. Impulsywny i spontaniczny – zorganizowany i ciężko pracujący.
3. Kontemplacyjny – zaangażowany w świat zewnętrzny.
4. Konkurencyjny (oparty na rywalizacji) – praca zespołowa i zaufanie.
5. Wyluzowany i zrelaksowany – popadający w stres i emocjonalny (Apply Magic Sauce, 2021).

Badania z wykorzystaniem Modelu Wielkiej Piątki są szeroko opisywane w literaturze przedmiotu, na potrzeby artykułu model ten (jak i pozostałe) został przedstawiony przyczynkowo, ponieważ opis metod badawczych nie stanowi celu pracy.

W badaniu Kosińskiego i innych z 2013 roku wykorzystano również kwestionariusze: IPIP (International Personality Item Pool), RSPM (Raven's Standard Progressive Matrices), SWL (Satisfaction With Life Scale). Dodatkowo, w celu ustalenia cech etnicznych, przebadano część zdjęć profilowych, którymi respondenci posługiwali się w ramach własnych kont na Facebooku (Kosinski i in., 2013).

Zebrany materiał badawczy przeanalizowano i przelożono na algorytm mierzący wskazane czynniki na podstawie cyfrowych śladów użytkowników. Określono, że analizy śladów cyfrowych są dokładne: w 95% w ramach określania cech etnicznych, w 93% w ramach określania płci, w 85% w ramach określania poglądów politycznych oraz w 82% w ramach określania wyznania (Tamże).

Reasumując, na podstawie działań użytkownika – ograniczonych do mierzenia klikania przycisku „lubię to” – możliwe było określenie jego poglądów

politycznych z dokładnością do 85%. Co za tym idzie, nowa metoda pomiarów ustalona przez Kosińskiego, Stillwella oraz Graepela pozwoliła na określanie cech jednostek oraz grup użytkowników bez konieczności wchodzenia z nimi w bezpośrednią reakcję i wykorzystywania metod deklaratywnych – np. ankiet. Oznaczało to, że badacze posiadający opracowany algorytm oraz dostęp do danych o użytkownikach portalu Facebook, mogli określać ich cechy – w tym ich poglądy polityczne.

Pokazujemy, że szeroka gama osobistych atrybutów ludzi, od orientacji seksualnej po inteligencję, może być automatycznie i precyzyjnie wnioskowana na podstawie ich polubień na Facebooku (Kosinski i in., 2013).

Badacze wyraźnie wskazali, że prowadzenie pomiarów cech/atrybutów ludzi poprzez wykorzystywanie danych cyfrowych (zarówno interakcji na portalu Facebook jak i innych) mogą mieć dobry wpływ na rozwój marketingu, personalizacji usług oraz rozwoju technologii. Niemniej jednak zwrócili też uwagę na fakt, że pomiary te (dokładne, jak i obarczone błędem) mogą stanowić bezpośrednie zagrożenia dla bezpieczeństwa ludzi:

Można sobie wyobrazić sytuacje, w których takie przewidywania, nawet błędne, mogłyby stanowić zagrożenie dla dobra, wolności, a nawet życia jednostki (Tamże).

W 2014 roku Kosiński i inni, wykorzystując wskazaną wcześniej metodykę, przebadali grupę 354 tysięcy użytkowników portalu Facebook, którzy posiadali aktywne konto minimum na 24 miesiące przed pobraniem danych. Badania zawężono do użytkowników pochodzących z USA (Kosinski i in., 2014). Badaniom podlegały: liczba znajomych, wydarzenia, zmiany statusu, zdjęcia, oznaczone zdjęcia oraz liczba członkostw w grupach. Celem badania było stwierdzenie, czy istnieją korelacje pomiędzy cechami (FFM) użytkowników a ich działaniami podejmowanymi na profilu Facebook (FB) – na to pytanie udzielono odpowiedzi twierdzącej, że jest to możliwe. Dla przykładu – stwierdzono, że osoby charakteryzujące się otwartością dokonują statystycznie więcej polubień materiałów na FB niż te, które przejawiają cechy kontemplacyjne. Dodatkowo stwierdzono, że osoby o cechach ekstrawertycznych znacznie częściej dokonują zmian statusów oraz dodają większą liczbę osób do grona znajomych (Tamże). Badacze podkreślili również, że ich analizy opierają się na rzeczywistych zachowaniach użytkowników, którzy dokonują ich w naturalnym środowisku – online.

Z racji na obszerność pakietów danych, które wymagają analiz, badacze byli zmuszeni do określenia, które z tych danych pozwalają na lepszym stopniu analizować cechy użytkowników. Lambiotte i Kosiński określili, że do pomiarów osobowości najlepiej nadaje się analizowanie grup społecznych sieci (np. ludzie

obserwujący się w mediach społecznościowych, członkostwo w grupach, itp.) oraz analizowanie interakcji „lubię to” (np. dodawanie reakcji „lubię to” pod treściami związanymi z muzyką) (Lambiotte & Kosinski, 2014). Dla przykładu, badacze sugerują, że za sprawą analizowania powiązań społecznych (dane z zakresu sieci społecznych) pozwalają określać orientację seksualną lub status związku badanych osób. Najważniejsze z perspektywy badaczy jest jednak określanie cech osobowości użytkowników poprzez analizowanie ich interakcji poprzez dodawanie reakcji „lubię to”. Stwierdzono, że istnieje duża korelacja pomiędzy gustem muzycznym (por. Ryc. 1) czy preferencjami w wyborze marki a cechami osobowości. Przykładowo osoby wchodzące w interakcję poprzez „like” z marką Hello Kitty przejawiali cechy wysokiej otwartości, niskiej sumienności i niskiej ugodowości.

Powyższe badania skłoniły naukowców do postawienia pytania, czy metody analityczne oparte na działaniach komputerowych są precyzyjniejsze od ludzi (np. grupy znajomych lub rodziny) w ocenianiu cech osobowości. W 2015 roku You You, Kosiński i Stillwell opublikowali artykuł *Computer-based personality judgments are more accurate than those made by humans*, w którym stwierdzili, że mierzenie cech osobowości ludzi jest znacznie dokładniejsze, szybsze i tańsze w przypadku wykorzystania metod komputerowych (You You i in., 2015). Badacze podkreślili, że ludzka osobowość może być określana precyzyjnie bez angażowania ludzkich umiejętności społeczno-poznawczych. Kolejny raz zwrócono również uwagę na fakt, że systemy komputerowe pozwalające na mierzenie cech ludzi mogą posłużyć do udoskonalenia technologii i poprawy jakości życia, ale ich zastosowanie może wpłynąć na rozwój technik manipulacyjnych i inwigilacyjnych.

Należy zwrócić uwagę, że znacznym ograniczeniem wszystkich przedstawionych metod badawczych jest fakt wykorzystywania obszernych pakietów danych pochodzących od użytkowników technologii cyfrowych. Co za tym idzie – dane te są trudno dostępne – wymagają zgody użytkownika lub dystrybutora treści (np. administracja Facebooka) oraz są trudne do analizowania ze względu na swoją obszerność (wymagają wykorzystywania metod z zakresu badań Big Data) (Kosinski i in., 2016).

Ryc. 1. Dendrogram autorstwa Lambiotte i Kosińskiego przedstawiający korelację pomiędzy zespołem/wykonawcą muzycznym, których „lubią” użytkownicy Facebooka a ich cechami osobowości. Im kolor jest bardziej cyjanowy tym cechy odbiorców są bardziej konserwatywne, natomiast im bardziej fioletowy tym cechy odbiorców są bardziej liberalne.



Źródło: Lambiotte & Kosinski, 2014.

W 2018 roku Wang i Kosiński opublikowali obszerny raport z badań dotyczących możliwości stwierdzenia orientacji seksualnej ludzi poprzez analizowanie ich zdjęć. Badacze wykorzystali w tym celu 35 326 zdjęć twarzy, które zostały przeanalizowane przez tak zwane głębokie sieci neuronowe⁴ do wskazania cech twarzy osób o orientacji homoseksualnej i heteroseksualnej. Dzięki tak zebranim danym i przeanalizowaniu jednego zdjęcia danej osoby możliwe

⁴ Przykładowo <https://www.faceplusplus.com>

było stwierdzenie czy osoba jest hetero czy homoseksualna w 81% przypadków w odniesieniu do mężczyzn oraz 74% przypadków w odniesieniu do kobiet. W analogicznym badaniu, w którym próbowano odgadnąć orientację seksualną na podstawie zdjęcia, wyniki te były znacząco niższe – 61% poprawności w odniesieniu do zdjęć mężczyzn oraz 54% w odniesieniu do zdjęć kobiet. Skuteczność algorytmu znacząco wzrosła – kolejno 91% i 83% poprawności, kiedy w badaniu uwzględniono po 5 zdjęć każdej osoby (Wang & Kosinski, 2018). Badacze przedstawili w swojej pracy narzędzie Face++, którego działanie można przetestować na stronie internetowej <https://www.faceplusplus.com>.

W 2021 roku Michał Kosiński na łamach „Nature Scientific Reports” opublikował podobne badanie, w ramach którego przeanalizowano ponad milion zdjęć osób w celu ustalenia ich poglądów politycznych. W trakcie badania stwierdzono, że twarze osób o poglądach liberalnych znacznie różnią się od osób o poglądach konserwatywnych. W przypadku tego pomiaru ustalono, że algorytm poprawnie odgadywał poglądy polityczne opierając się na analizie zdjęć – 72% poprawności. W podobnym teście ludzie określają poprawnie, czy dana osoba ma poglądy konserwatywne czy liberalne w 55% dokładności (Kosinski, 2021).

(...) nawet pobieżne oszacowanie cech psychologicznych odbiorców może drastycznie zwiększyć skuteczność masowej perswazji. Mamy nadzieję, że naukowcy, decydenci, inżynierowie i obywatele zwrócą na to uwagę (Tamże).

Przypadki wykorzystania cyfrowych śladów do nielegalnego analizowania danych cyfrowych użytkowników cyberprzestrzeni nie są szeroko opisywane w literaturze przedmiotu. Za przykład może posłużyć jednak komercyjna firma Cambridge Analytica (CA), która powstała w 2013 roku i została zamknięta w 2018 roku (Wikipedia, 2021). Nazwa firmy nie powinna być powiązywana z Centrum Psychometrii Uniwersytetu w Cambridge. Strona tej firmy pozostaje zamknięta – w celu dotarcia do tych materiałów, poniekąd dowodząc prawdziwości hipotezy o tym, że cyfrowe ślady nie są w pełni usuwalne, autor wykorzystał jedną z technik *data miningu* i przeanalizował bazy danych plików tymczasowych (cache)⁵ serwerów Google oraz Wayback Machine. Dzięki tym działaniom udało się dotrzeć do pełnej zawartości witryny CA z dnia 6 grudnia 2016. Na rycinie 2. przedstawiono nazwę modelu analitycznego wykorzystywanego przez CA w ramach analizowania cyfrowych śladów użytkowników.

⁵ Pamięć podręczna (ang. *cache*) – mechanizm, w którym część spośród danych zgromadzonych w źródłach o długim czasie dostępu i niższej przepustowości jest dodatkowo przechowywana w pamięci o lepszych parametrach. Ma to na celu poprawę szybkości dostępu do tych informacji, które przypuszczalnie będą potrzebne w najbliższej przyszłości (Pamięć podręczna..., 2021).

Akronim OCEAN, rozumiany jako O – otwartość, C – sumienność, E – ekstrawersja, A – ugodowość, N – neurotyczność, w literaturze przedmiotu przedstawiany jest częściej jako Five Factor Model.

Ryc. 2. Zrzut ekranu ze strony internetowej firmy Cambridge Analytica przedstawiający wykorzystywany model oceny użytkowników OCEAN zwany również Five Factor Model

At Cambridge Analytica we use data modeling and psychographic profiling to grow audiences, identify key influencers, and connect with people in ways that move them to action. Our unique data sets and unparalleled modeling techniques help organizations across America build better relationships with their target audience across all media platforms.

OCEAN Personality Model

We use the established scientific OCEAN scale of personality traits to understand what people care about, why they behave the way they do, and what really drives their decision making.

O	C	E	A	N
OPENNESS Do they enjoy new experiences?	CONSCIENTIOUSNESS Do they prefer plans and order?	EXTRAVERSION Do they like spending time with others?	AGREEABLENESS Do they put people's needs before theirs?	NEUROTICISM Do they tend to worry a lot?

Understand your unique personality type:

[Take our test](#)

Źródło: WayBack Machine,
<https://web.archive.org/web/20170113214511/https://cambridgeanalytica.org/>

W ramach odszukanych fragmentów stron internetowych CA przeczytać można było (por. ryc. 3):

Współpracowaliśmy z markami, organizacjami politycznymi i grupami rzeczowników na całym świecie, a nasza metodologia została zatwierdzona przez Ministerstwo Obrony Wielkiej Brytanii, Departament Stanu USA, Sandię i NATO.

Ryc. 3. Zrzut ekranu ze strony internetowej firmy Cambridge Analytica przedstawiający stopkę zawierającą informację o organizacjach zatwierdzających prawidłowość działania metod CA



Źródło: Way Back Machine,
<https://web.archive.org/web/20170113214511/https://cambridgeanalytica.org>

Zakończenie

Badania zainicjowane w 2007 roku przez badaczy z Centrum Psychometrii Uniwersytetu w Cambridge – myPersonality, dały do tej pory niespotykane możliwości analizowania cech osobowości ludzi na podstawie ich cyfrowych śladów. Dzięki tym badaniom możliwe było ustalenie sprawdzalnych, powtarzalnych metod analizowania cech użytkowników urządzeń cyfrowych. Metody te początkowo opierały się na mierzeniu użytkowników poprzez kwestionariusze, a następnie za sprawą systemów uczących się przygotowywano analizy komputerowe. Współcześnie dzięki opisaniu i stwierdzeniu wyraźnej korelacji między zachowaniem ludzi w środowisku cyfrowym a ich cechami osobowości możliwe jest przeprowadzanie badań w oparciu o wskazane metody, wykorzystując jedynie ślady cyfrowe. Oznacza to, że badacze nie muszą przeprowadzać dodatkowych pomiarów z wykorzystaniem tradycyjnych metod badawczych (np. kwestionariusza modelu pięciu czynników). Co więcej, jak udowodniono w artykułach *Deep neural networks are more accurate than humans at detecting sexual orientation from facial images* oraz *Facial recognition technology can expose political orientation from naturalistic facial images* możliwe jest określanie cech takich jak orientacja seksualna i poglądy polityczne w oparciu o analizy zdjęć ludzi. Pomiarzy te nie są jeszcze doskonale precyzyjne, ale ze względu na historyczny rozwój tego typu badań można stwierdzić, że ich dokładność będzie rosła wraz z postępem technologii i rozwojem metodyki.

Zasób danych o użytkownikach Sieci pozwala określać cechy ludzi, informacje te w bezpośredni sposób mogą być wykorzystywane w celu rozwoju

technologii, rozwoju społecznego, niwelowania barier cyfrowych jak i technik manipulacyjnych, inwigilacyjnych i polaryzacyjnych – współcześnie⁶ dostępność tych danych jest prostsza niż kiedykolwiek w historii.

W świetle przedstawionych analiz literatury przedmiotu stwierdzono, że współcześnie należy zwrócić uwagę publiczną na kwestie związane z analizowaniem cyfrowych śladów w celu ustalania cech osobowościowych, poglądów politycznych czy orientacji seksualnej ludzi. Budowanie świadomości wśród użytkowników powinno skutkować obniżeniem negatywnego wpływu tych zjawisk na społeczeństwo.

Bibliografia

- 1 Second – Internet Live Stats. ([2021]). Internet Live Stats [online]. Pobrane 2 maja 2021, z: <https://www.internetlivestats.com/one-second/#traffic-band>
- Bachrach, Yoram, Kosinski, Michal, Graepel, Thore, Kohli, Pushmeet, & Stillwell, David J. (2012). Personality and patterns of Facebook usage. W: *Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12* (s. 24–32) [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1145/2380718.2380722>
- Cambridge Analytica. (2021). W: Wayback Machine [online]. Pobrane 2 maja 2021 r., z: <https://web.archive.org/web/20170113214511/https://cambridgeanalytica.org/>
- Cambridge Analytica. (2021). W: Wikipedia. The Free Encyclopedia [online]. Pobrane 30 czerwca 2021 r., z: https://en.wikipedia.org/wiki/Cambridge_Analytica
- Kasprzak, Wojciech A. (2015). *Ślady cyfrowe. Studium prawnokryminalistyczne*. Warszawa: Difin.
- Kosinski, Michal (2021). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports*, **11**(1), Article number: 100 [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1038/s41598-020-79310-1>
- Kosinski, Michal, Bachrach, Yoram, Kohli, Pushmeet, Stillwell, David J., & Graepel, Thore (2014). Manifestations of user personality in website choice and behavior on online social networks. *Machine Learning*, **95**, 357–380 [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1007/s10994-013-5415-y>
- Kosinski, Michal, Stillwell, David J., & Graepel, Thore (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(15), 5802–5805 [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1073/pnas.1218772110>
- Kosinski, Michal, Wang, Yilun, Lakkaraju, Himabindu, & Leskovec, Jure (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, **21**(4), 493–506 [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1037/met0000105>

⁶ Na początku drugiej dekady XXI wieku do stwierdzenia celów użytkowników przy wykorzystaniu analiz śladów cyfrowych były niezbędne m.in.: technologia, metodyka oraz trudny do zdobycia zasób danych. Współcześnie zasób tych danych jest często publicznie dostępny – zdjęcia profilowe na portalach społecznościowych, zatem do ich zebrania jak i analizowania potrzebna jest przede wszystkim technologia.

- Lambiotte, Renaud, & Kosinski, Michal (2014). Tracking the digital footprints of personality. *Proceedings of the IEEE*, **102**(12), 1934–1939 [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1109/JPROC.2014.2359054>
- Markovikj, Dejan, Gievska, Sonja, Kosinski, Michal & Stillwell, David J. (2013). Mining Facebook data for predictive personality modeling. *Proceedings of the International AAAI Conference on Web and Social Media*, **7**(1), 23–26. Pobrane 30 czerwca 2021 r., z: <https://ojs.aaai.org/index.php/ICWSM/article/view/14466>
- Matz, Sandra C., Kosinski, Michal, Nave, Gideon, & Stillwell, David J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America*, **114**(48), 12714–12719 [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1073/pnas.1710966114>
- McCrae, Robert R., & John, Oliver P. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, **60**(2), 175–215 [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1111/j.1467-6494.1992.tb00970.x>
- Michal Kosinski (2021, maj 2). W: Google Scholar [online]. Pobrane 2 maja 2021 r., z: <https://scholar.google.com/citations?user=01-XV0YAAAA&hl=en>
- Pamięć podręczna (2021). W: Wikipedia. Wolna encyklopedia [online]. Pobrane 30 czerwca 2021 r., z: https://pl.wikipedia.org/wiki/Pamięć_podręczna
- Quercia, Daniele, Kosinski, Michal, Stillwell, David J., & Crowcroft, Jon (2011). Our twitter profiles, our selves: Predicting personality with twitter. W: *Proceedings – 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011* (s. 180–185) [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1109/PASSAT/SocialCom.2011.26>
- Quercia, Daniele, Lambiotte, Renaud, Stillwell, David J., Kosinski, Michal, & Crowcroft, Jon (2012). The personality of popular facebook users. W: *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW'12* (s. 955–964) [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1145/2145204.2145346>
- Stillwell, David J. (2021). myPersonality database [online]. Pobrane 30 czerwca 2021 r., z: <https://www.psychometrics.cam.ac.uk/productsservices/mypersonality>
- Stillwell, David J. & Kosinski, Michal (2012). myPersonality project Example of successful utilization of online social networks for large-scale social research. W: *Proceedings of the ACM Workshop on Mobile Systems for Computational Social Science (MobiSys)* [online]. The Psychometric Centre, University of Cambridge. Pobrane 30 czerwca 2021 r. z: <https://drive.google.com/file/d/1KzTSyd2coi0oTprgHRRYwWScQ7-TaWIs/view>
- Wang, Yilun, & Kosinski, Michal (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, **114**(2), 246–257 [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1037/pspa0000098>
- Wasilewski, Janusz (2013). Zarys definicyjny cyberprzestrzeni. *Przegląd Bezpieczeństwa Wewnętrznego*, **9**(5), 225–234 [online]. Agencja Bezpieczeństwa Wewnętrznego. Pobrane 30 czerwca 2021 r., z: <http://www.abw.gov.pl/download/1/1284/Segregator13.pdf>
- Youyou, Wu, Kosinski, Michal, & Stillwell, David J. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(4) 1036–1040 [online]. Pobrane 30 czerwca 2021 r., z: <https://doi.org/10.1073/pnas.1418680112>

Determining the personality of Internet users by analyzing their digital footprints in the light of selected research by Michał Kosiński

ABSTRACT: According to data from Internet Live Stats, in April 2021, about 92,000 queries were made, every second, in the Google search engine. Each activity performed by users of digital devices is indexed as so-called digital footprints [also called digital traces], thanks to which it is possible, using appropriate technologies and methods, to precisely define personality traits, political views and sexual orientations of these users.

The article was inspired by the works of PhD Michał Kosiński and describes the issues related to analyzing digital traces of Internet users (mainly social media).

The main aim of the article is to present the research by PhD Michał Kosiński and to draw the attention of the IT community to the issues related to analyzing the digital traces of Internet users. This work is of a popularizing nature and does not constitute a comprehensive description of the achievements of the indicated scientist. The article does not present new information or own research – however, it is intended to encourage recipients to analyze the literature on the subject of analyzing digital traces and privacy in the digital age.

The work uses the method of database content analysis – to collect and analyze the literature on the subject. The focus was on the resources made available by PhD Kosinski through his private website – <https://www.michalkosinski.com>. Additionally, datamining techniques were used to summarize the archival content published on the Cambridge Analytica website.

The chronological range of the searched writing materials was narrowed from 2011 to the first quarter of 2021, focusing mainly on the years 2013–2021, i.e. from the year of publishing the article Private traits and attributes are predictable from digital records of human behavior until the year of publication of the text Facial recognition technology can expose political orientation from naturalistic facial images.

The article presents selected works by PhD Michał Kosiński, which constitute "milestones" in research on determining the personality of Internet users within the publication of the indicated author. Consequently, the article does not present the works that described part of the research or contributed to taking larger, further measurements.

KEYWORDS: digital traces, digital footprints, measuring user personality, Five Factor Model, privacy, Michał Kosiński