studia medioznawcze
2022, tom 23, nr 2 (89)
ISSN 2451-1617

Big Data o mediach. Dominanty świata mediów

Włodzimierz Gogołek
Uniwersytet Warszawski
wlodzimierz@gogolek.pl
ORCID: 0000-0002-3073-3817

STRESZCZENIE

Cel: identyfikacja aktualnych dominant świata mediów (kontynentów i wysp) wskazujących determinanty biznesowej kondycji przemysłu mediowego. **Metodologia:** zidentyfikowano źródła informacji i automatycznie pobrano zgromadzone w nich treści (tekstowe dane źródłowe). Następnie przeprowadzono ilościową analizę tych danych i przygotowano wizualizację uzyskanych wyników. W tym celu zastosowano narzędzia rafinacji informacji – Big Data. Uzyskane w ten sposób informacje umożliwiły ocenę stanu i dynamiki zmian dominant mediów. **Wyniki i wnioski:** wyniki badań umożliwiły identyfikację zbioru najistotniejszych dominant świata mediów oraz ich atrybutów. Są propozycją swoistego paradygmatu parametrów biznesowego modelu inwestycji w przemyśle mediowym. **Ograniczenia badawcze:** brak doświadczeń i autorytatywnych opracowań/publikacji w zakresie korzystania z informacyjnego potencjału Big Data w badaniach przemysłu mediowego w obszarze poszukiwania dominant świata mediów. **Oryginalność:** autorowi nieznane są badania, w których wykorzystano zasoby Big Data (treści i metody/narzędzia) odnoszące się do poszukiwania dominant przemysłu mediowego. Ponadto brak jest teorii oceniającej wiarygodność wyników badań prowadzonych na dużych zasobach informacji – Big Data. Zalążkiem takiej teorii mogą być opisane dalej badania oraz wyniki kilkudziesięciu wcześniejszych badań empirycznych wymienionych w bibliografii. Dowodzą one, począwszy od 2010 roku (m.in. pierwsze na świecie zastosowania Big Data w predykcji wyborów prezydenckich), trafności ocen (dotychczas niekwestionowanych) stanu i predykcji badanych zjawisk, dokonywanych na podstawie analiz Big Data.

SŁOWA KLUCZOWE

Big Data, dominanty mediów, media, paradygmat przemysłu mediowego, przemysł mediowy, rafinacja informacji, sztuczna inteligencja

„Matematyka jest alfabetem,
przy pomocy którego Bóg opisał wszechświat"
– Galileusz

Media są jedną z najbardziej dochodowych branż przemysłowych, w których przedmiotem są informacja i narzędzia służące do jej identyfikacji, kolekcjonowania, przetwarzania i ostatecznie sprzedaży. Profitem w tej branży są nie tylko pieniądze, ale także – a może nawet przede wszystkim – efekty kształtowania informacyjnego konsumentów produktów mediowych. Troska o maksymalizację jej dochodowości decyduje o zasadności identyfikacji aktualnych dominant przemysłu mediowego. W tym znaczeniu termin „dominanta mediów" określa dające się wyróżnić czynniki odgrywającą zasadniczą rolę w funkcjonowaniu mediów. Determinują one efektywność całego cyklu produkcji, a zwłaszcza umownej sprzedaży mediowej. W niniejszym tekście przyjęto, iż dominanty te są lądami (kontynentami i wyspami) świata mediów. Tworzą one obraz świata mediów – sugerowaną podstawę biznesplanu każdego przedsiębiorstwa mediowego. Kondycja lądów determinuje dochodowość przemysłu mediowego. Celem jest zatem trafne określenie kierunków inwestycji materialnych, organizacyjnych i intelektualnych w utrzymanie i rozwój lądów świata mediów – kierunków będących determinantami sukcesów przemysłu mediowego.

O zasadności podjęcia badań nad identyfikacją biznesowych determinant mediów świadczą szybko rosnące przychody branży medialnej. Firma konsultingowa PricewaterhouseCoopers (2018, październik) szacuje, że rynek mediów i rozrywki na świecie w 2022 roku wart będzie 2,4 biliona dolarów i wzrastać będzie średniorocznie o 4,4%. W Polsce rynek ten osiągnie wartość 13,4 mld dolarów i będzie rosnąć o 3,5% rocznie. PwC wskazuje również najbardziej prawdopodobne kierunki jego rozwoju. Na pierwszym miejscu znajduje się VR (*virtual reality*) oraz OTT (*over the top*). Spadek zanotują takie media jak gazety i czasopisma. Telewizja i wideo osiągną znikomy wzrost (Pruchnik, 2019). Rzeczywistość 2022 roku dowodzi potrzeby aktualizacji tych przewidywań.

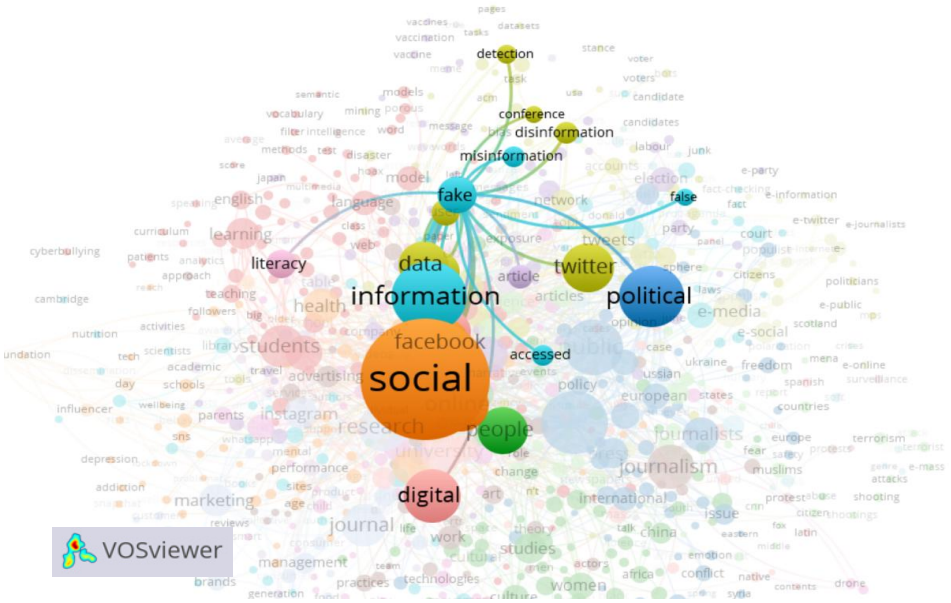
W 2015 roku, z inicjatywy Krajowej Rady Radiofonii i Telewizji (KRRiT), opracowany został dokument *Strategia rozwoju rynku medialnego w Polsce na lata 2015–2020* (Batorski et al., 2015). Za jego stworzenie odpowiadało grono autorytetów i ekspertów współpracujących z instytucjami wspierającymi rynek mediów w Polsce – KRRiT, PISF i ZAIKS. W dokumencie oprócz analizy otoczenia i wyzwań zawarto prognozy i postulaty rozwoju mediów – sugestie dotyczące poszukiwanych lądów. Okres, którego dotyczył dokument, już się zakończył. Na razie nie opublikowano podsumowania dotyczącego wykonania zapisanej w nim strategii ani podobnego dokumentu na kolejne lata.

Zasadność opisywanego tu istotnego statystycznie związku między dominantami świata mediów, zidentyfikowanymi w wyniku analiz dużych zasobów informacyjnych (Big Data), a procesem decydowania o kierunkach inwestowania potwierdziły wyniki 18 badań. Dotyczyły one determinant rozwoju kluczowego dla polskiej gospodarki świata nowych technologii – jego umownych kontynentów i wysp (m.in. energii, transportu, sztucznej inteligencji). Wyniki okazały się trafną predykcją inwestycji w tym zakresie[1]. Dotyczy to predykcji np. opłacalnych obecnie technologii/narzędzi związanych ze sztuczną inteligencją i kierunkami inwestycji w energetyce.

[1] Nieopublikowane materiały badawcze powstałe w ramach projektu badawczego *Oceny trendów nowych technologii. Eksploracja źródeł danych w zakresie działalności B+R+I. Projekt systemu rafinacji*, zleconego przez Narodowe Centrum Badań i Rozwoju i zrealizowanego przez Uniwersytet Warszawski (2018–2019).

O wiarygodności zastosowanych narzędzi badawczych Big Data świadczą także wyniki predykcji wyborów parlamentarnych i prezydenckich (Gogołek, 2013 & 2016). Były one dokładniejsze niż wyniki uzyskane przez polskie ośrodki badań opinii publicznej.

Mając na uwadze dynamiczne zmiany w świecie mediów, podjęto próbę identyfikacji wspomnianych lądów – kontynentów i otaczających ich wysp. Ich wielkości ilustrują średnice okręgów na rys. 1 i rys. 2. Odzwierciedlają one liczby cytowań nazw umownych lądów w analizowanych dokumentach źródłowych. Odległości między okręgami ukazują wzajemne powiązania (sąsiedztwo) wyróżnionych lądów. Przyjęto, iż miary te (wielkość i odległość/sąsiedztwo) wskazują poszukiwane dominanty świata mediów. Przykładami wysp w otoczeniu kontynentu *social* (media społecznościowe) (rys. 1) są m.in. wyspy: *information*, *political*, *digital*, *fake*, *twitter* i *facebook*.



Rys. 1. Wyspy kontynentu *social*.

Źródło: opracowanie własne.

Identyfikacja kontynentów i wysp świata mediów stanowiła cel badań. Pozwoliły one oznaczyć istotne ilościowo determinanty biznesowej kondycji mediów. Przyjęto, iż każda z nich, niezależnie od pozostałych, ma istotny wpływ na sukces mediów jako całości.

Metodologia. Narzędzia badawcze

Analiza i ocena dynamicznych zmian biznesowej funkcjonalności mediów wymaga rzetelnej informacji o ich stanie oraz perspektywach rozwoju w skali globalnej. Przyjęto, iż pozyskanie tej informacji jest możliwe dzięki analizie dużych zasobów informacyjnych Big Data. Big Data oznacza nie tylko duży zbiór informacji, ale także system powiązanych funkcjonalnie zbiorów metod i narzędzi odkrywania nowych informacji o obiekcie (np. o zjawiskach, procesach, organizacjach, osobach)[2].

[2] Formalną definicję Big Data proponują Międzynarodowa Organizacja Normalizacyjna oraz Międzynarodowa Komisja Elektrotechniczna („Information technology — Big data — Overview and vocabulary", n.d.).

System ten (element Big Data) nazwano rafinacją informacji. Obejmuje ona zarządzanie i analizę bogatych – głównie nieustrukturyzowanych – tekstowych materiałów źródłowych.

Rafinacja informacji wykorzystuje nowoczesne narzędzia służące identyfikacji i kolekcjonowaniu tematycznych materiałów źródłowych oraz ich analizie i wizualizacji wyników. Rafinacja informacji wraz z potencjałem informacyjnym Big Data to nowa dziedzina, z której rozwojem wiąże się duże nadzieje (Mayer-Schonberger & Cukier, 2017; Gogołek, 2017). Wiele autorytetów przewiduje, że zastosowanie metod statystyczno-analitycznych oraz sztucznej inteligencji do analiz dużych zasobów informacyjnych zrewolucjonizuje sposoby badania otaczającego nas świata (Surma, 2019). Siłę detekcyjną rafinacji potwierdzono także w kilkudziesięciu własnych badaniach, które m.in. pozwoliły przewidzieć wyniki wyborów parlamentarnych i prezydenckich w Polsce w latach 2011 i 2015 (Gogołek, 2016), a także zidentyfikować determinanty i trendy rozwoju technologii (Gogołek, 2017).

W artykule zaprezentowane zostały kluczowe elementy eksploracji tekstów (Silge & Robinson, 2017) służące rafinacji dużych zasobów informacyjnych w celu identyfikacji determinant biznesowych sukcesów świata mediów. Mając na uwadze wyniki wcześniej przeprowadzonych badań w tym zakresie (Gogołek, 2016 & 2017), przyjęto, iż unikalności/odrębności tych determinant (lądów) dowodzą miary frekwencji ich cytowań w tematycznych zasobach Big Data[3].

Prognozy rozwoju mediów

Wynikiem przeprowadzonych badań jest ocena biznesowej istotności kontynentów i wysp. Jej miarą jest wspomniana liczba cytowań nazw zidentyfikowanych dominant świata mediów. Tworzą one mapę świata mediów: lądów w formie kontynentów – dominujących, dużych skupień atrybutów mediów – oraz wysp – niewielkich, drugoplanowych skupień atrybutów. Atrybutami są najczęściej (w materiałach źródłowych) pojęcia występujące w sąsiedztwie wyróżnionych kontynentów i wysp – są to swoiste słowa kluczowe zidentyfikowanych dominant mediów.

Stan lądów determinuje wartość przemysłu mediowego. Są one proponowanym paradygmatem tworzącym podstawy budowy efektywnego modelu biznesowego mediów.

W celu budowy efektywnego (statystycznie istotnego) modelu konieczne jest wyróżnienie rzeczywistych/aktualnych odrębnych lądów i opisujących je atrybutów. Oznacza to konieczność poszukiwań ilościowo istotnych determinant wskazujących odrębność wysp i kontynentów. Odrębności oraz wielkości kontynentów i wysp wynikają z liczby cytowań opisujących je nazw w materiałach źródłowych oraz z miary ich bliskości (sąsiedztwa). Przykładem lądu są media społecznościowe (rys. 1) – największy obecnie kontynent świata mediów.

Dominanty przemysłu mediowego

Wydaje się, że akceptowalną formą identyfikacji dominant przemysłu mediowego, poza wskazaniem lądów, jest analiza wartości ilościowych miar ich stanu oraz predykcji dynamiki ich zmian. Miary i predykcje są pochodną liczby cytowań wszystkich wyrazów w materiałach źródłowych. W pomiarze wartości ocen i predykcji ich zmian użyteczna była rafinacja tematycznych zasobów Big Data. Największe wartości wykazały nazwy (cytowane wyrazy) tych poszukiwanych lądów świata mediów, które charakteryzują się dominującą popularnością. Przyjęto, iż nazwy te wraz z frekwencjami ich cytowań i predykcjami zmian mogą być wiarygodnym źródłem odpowiedzi na pytanie o biznesowe dominanty przemysłu mediowego. Zidentyfikowane dominanty sygnali-

[3] Materiały źródłowe zebrali i ich ilościowe analizy i wizualizacje wykonali badacze z Katedry Technologii Informacyjnych Mediów WDIB UW: Katarzyna Jarzyńska, Aleksander Strzelczyk, Konrad Żukowski i Piotr Wierzbicki.

zują obecne i przewidywane najważniejsze uwarunkowania funkcjonowania mediów. Stanowią one proponowaną wskazówkę dla inwestorów przemysłu mediowego; tym samym można je uznać za ważne przedmioty badań świata mediów.

Materiały źródłowe

Szlak odkrywania kontynentów mediów znaczą dwa kamienie milowe: źródła informacji o mediach (źródła) oraz rafinacja zawartych w nich treści, obejmująca także ich wizualizacje i wynikające z nich wnioski. Wskazują one poszukiwane lądy i ich atrybuty.

Fundamentalne znaczenie dla powodzenia rafinacji ma identyfikacja i gromadzenie, w odpowiedniej formie, dużych tematycznych zbiorów danych źródłowych. Z dotychczasowych doświadczeń wynika, że ich wielkość determinuje rzeczywistą wartość informacyjną Big Data. Proces identyfikacji/poszukiwania źródeł poprzedziło opisanie przedmiotu badań – pojęcia mediów, w tym jego synonimów i pseudosynonimów („media", „telewizja", „television", „tv", „internet", „social media", „prasa", „gazeta", „newspaper", „magazine", „radio", „wideo", „video")[4]. Następnie, korzystając z odpowiednich narzędzi, zidentyfikowano źródła. Dominujące okazały się publikacje naukowe i profesjonalne/branżowe związane z mediami[5].

W celu zmniejszenia prawdopodobieństwa błędnych wyników rafinacji poszukiwanie materiałów źródłowych odbywało się bezstronnie (na wzór ślepej recenzji) i dwustopniowo. W pierwszym kroku zidentyfikowano źródła materiałów (kluczem był przyjęty opis pojęcia „media") za pomocą przeznaczonych do tego narzędzi i baz danych. W drugim zweryfikowano zgodność wyników kwerendy słów kluczowych, wyłonionych automatycznie z tekstów źródłowych (pełnej treści) wszystkich zebranych materiałów źródłowych, z przyjętym opisem pojęcia „media". Przekroczenie przyjętej miary zgodności wyniku kwerendy z opisem było kryterium kwalifikacji każdego materiału źródłowego do badań.

W ramach przedstawionej procedury wyboru materiałów źródłowych zgromadzono ponad 6 mln tekstów dotyczących problematyki mediów. Mając na uwadze aktualność badań, wybrano tylko 685 tys. materiałów opublikowanych między styczniem 2019 a czerwcem 2021 roku. Następnie przeprowadzono ich stosowną normalizację (czyszczenie, lematyzacja). Uzyskany w ten sposób zbiór poddany został ostatecznej weryfikacji pod kątem merytorycznego związku treści materiałów źródłowych z przedmiotem badań. Użyto do tego jednego z narzędzi automatycznego identyfikowania słów kluczowych dla każdego materiału źródłowego – Yake! (Campos, Mangaravite, Pasquali, Jorge, Nunes, & Jatowt, 2018). Owe klucze wykorzystano do kwerendy tylko tych treści, które dotyczyły aktualnej problematyki mediów. W ten sposób dokonano ostatecznej selekcji – spośród 6 mln tekstów – 3 tys. materiałów źródłowych, które są aktualne, opublikowane w języku angielskim i ściśle związane z przedmiotem badań – kontynentami i wyspami świata mediów.

Przedstawiona procedura wyszukiwania materiałów źródłowych z dużym prawdopodobieństwem obejmowała najistotniejsze tematycznie, otwarte publikacje w wyróżnionym okresie badawczym. Gwarancją kompletności (a nie wyliczonej statystycznie niewielkiej liczebnie próby reprezentatywnej) źródłowych materiałów jest powszechnie akceptowany profesjonalizm wy-

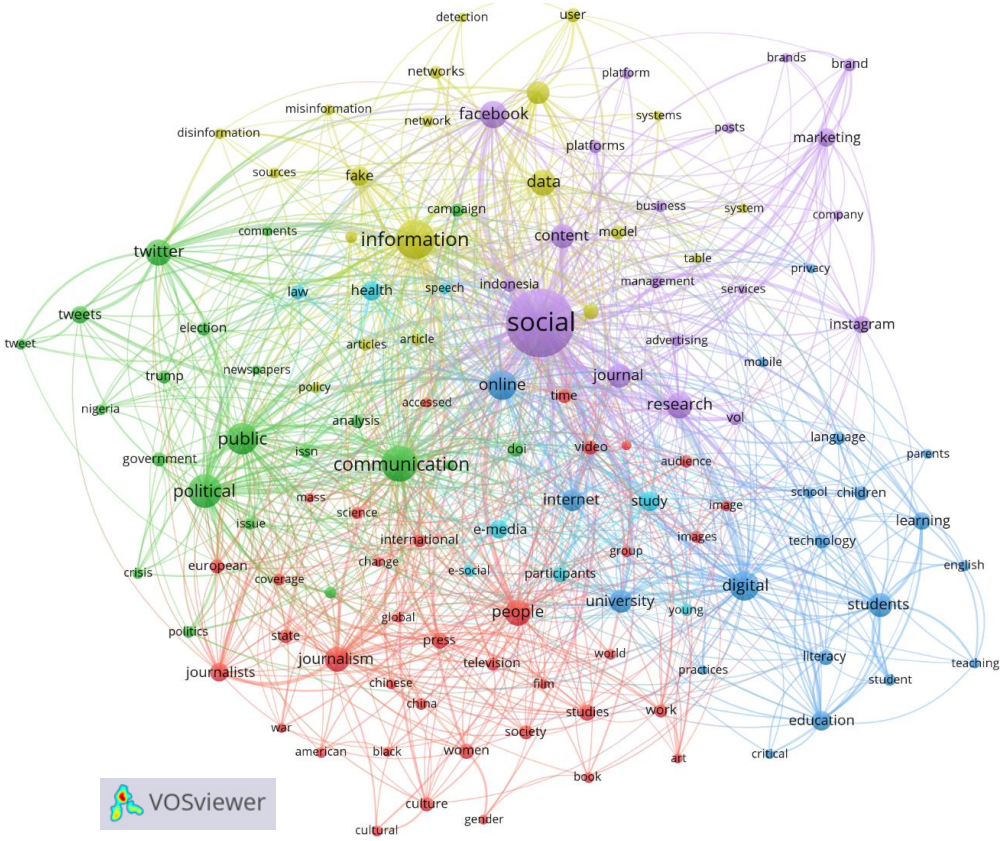
[4] Formalny zapis synonimów i pseudosynonimów: media|telewizja|television|tv|internet|(social.{0,30}media)|prasa|gazeta|newspaper|magazine|radio|wideo|videomedia. Nawias klamrowy {0,30} wskazuje dopuszczalną liczbę znaków między wyrazami (po lematyzacji) *social* i *media*.

[5] Pod adresem https://gogolek.pl/Referaty/zrodla-media.pdf dostępne są 24 przykładowe serwisy wyszukiwawcze źródeł zasobów poszukiwanych tematycznych publikacji oraz 35 326 zidentyfikowanych tematycznych źródeł materiałów źródłowych.

korzystanych baz i serwisów wyszukiwawczych (m.in. Web of Science, Scopus, OpenDOAR, Google Scholar, CORE, Base, Microsoft Academic Search, CiteSeerX, SciCentral, WolframAlfa, DOAJ). Należy podkreślić, iż obejmują one także zasoby głębokiej, niezaindeksowanej wyszukiwarkami (klasy Google czy Yahoo!) sieci. Dzięki temu zgromadzono swoistą populację generalną materiałów źródłowych, która gwarantuje jakościowe zróżnicowanie źródeł w zakresie ich wagi/znaczenia dla przedmiotu badań.

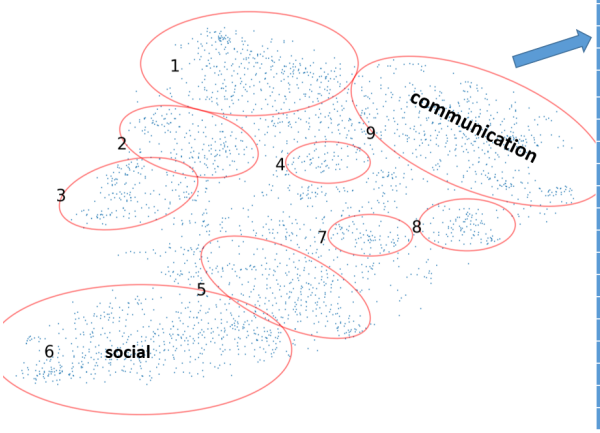
Rafinacja materiałów źródłowych

Podstawowym oczekiwanym wynikiem badania była identyfikacja kontynentów i wysp świata mediów. W tym celu zastosowano równocześnie dwa narzędzia – Yake! i VOSviewer. Wyniki, jakie dzięki temu uzyskano, ilustruje – na bazie materiałów źródłowych – rys. 2. Pokazuje on wielkości i siły powiązań nazw (wyrazów), które stosownie do wielkości (częstotliwości cytowania) uznano za poszukiwane kontynenty i wyspy. Ponadto na rys. 2 wyróżniono kolorami umowne (statystyczne) sąsiedztwo zidentyfikowanych lądów świata mediów.



Rys. 2. Kontynenty i wyspy świata mediów w 2021 roku.
Źródło: opracowanie własne.

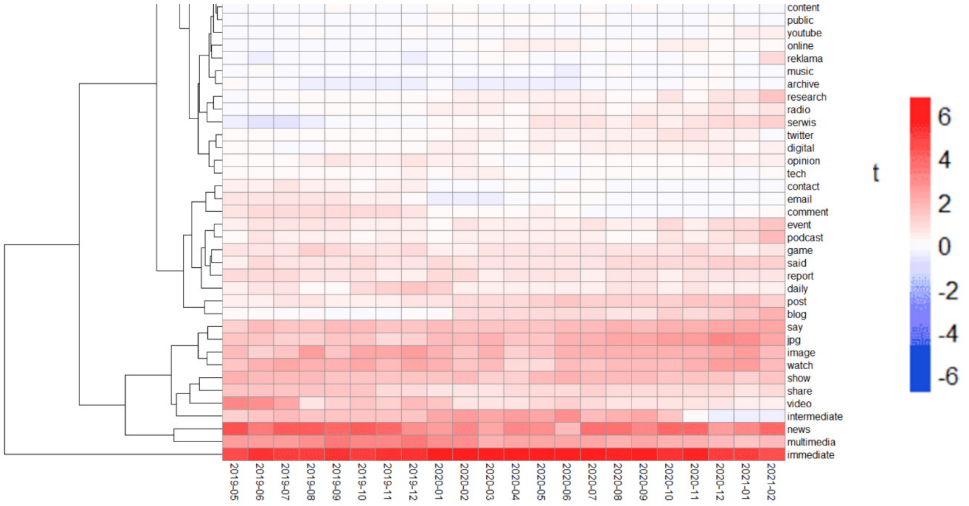
Uzyskane w ten sposób dane ilościowe wzbogacono wynikami zastosowania narzędzia uczenia się maszynowego (*machine learning toolkit*) – TSNE (*t-distributed Stochastic Neighbor Embedding*). Pozwala ono ocenić podobieństwo między sąsiadującymi wyrazami i ich wzajemną umowną odległość w analizowanych tekstach źródłowych. Teksty o podobnej treści tworzą odrębne klastry/skupienia. Wyniki są uzupełnieniem rezultatów uzyskanych z Yake!/VOSviewera i wskazują poszukiwane kontynenty mediów.



social	182292
study	160229
digital	160165
public	126701
research	113710
political	107793
news	104858
information	100501
issue	088985
content	085793
online	079821
people	078687
journal	074751
technology	074245
society	071285
time	068223
analysis	063842
effect	062656
role	062435
data	054365
platform	039047
user	037093

Rys. 3. Skupienia (TSNE) treści związanych z problematyką mediów.

Źródło: opracowanie własne.

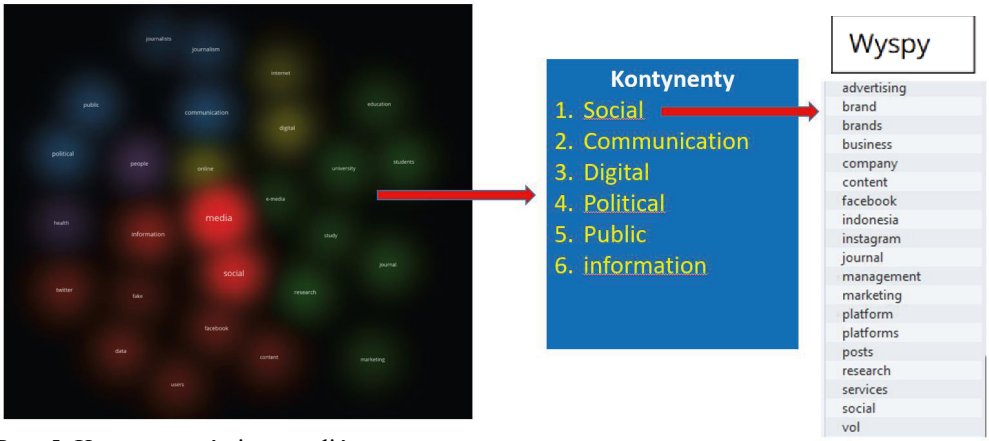


Rys. 4. Fragment mapy cieplnej atrybutów mediów.

Źródło: opracowanie własne. Mapa cieplna została opracowana na podstawie 1 477 211 źródeł, w których klucz „media" wystąpił 16 762 167 razy.

Rysunek 3 ilustruje uzyskane skupienia, w tym nazwy dwóch oznaczonych klastrów-kontynentów: *social* i *communication*. Ponadto pokazuje zidentyfikowane przez TSNE, dla skupienia *communication*, atrybuty z ich umownymi wartościami wielkości. Wyróżnione atrybuty mogą być jednakowe dla różnych skupień. Podobnie jak wspólne przedmioty na różnych kierunkach studiów.

Uzupełniającym źródłem informacji o wadze/istotności atrybutów mediów jest narzędzie umożliwiające tworzenie globalnej mapy cieplnej (widma) (rys. 4). Widoczne są na niej krotności występowania w czasie (dynamika zmian) i wzajemnych (jednostkowych) powiązań atrybutów. Na przykład ostatnie trzy wiersze pokazują największą wagę atrybutów *news* i *multimedia*, które wchodzą w skład pojęcia *immediate*, a także to, że liczba cytowań wyrazu *intermediate* wyraźnie maleje w czasie.



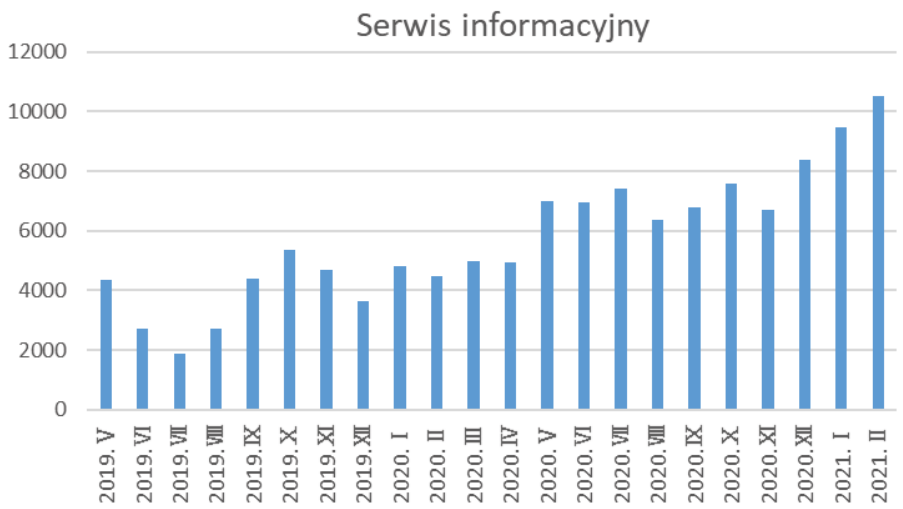
Rys. 5. Kontynenty świata mediów.

Źródło: opracowanie własne.

Rysunek 5 pokazuje wielkości umownych sześciu lądów świata mediów i – w swojej centralnej części – pojęcia *media*. Największe skupienia (część rozmytych okręgów) to poszukiwane kontynenty; każdy z nich otaczają wyspy. Na rys. 5 widoczne są (w malejącej kolejności frekwencji występowania) wyspy sąsiadujące z kontynentem *social*. Największą wyspą przy kontynencie mediów społecznościowych jest reklama. Sugeruje to trafność inwestycji w dominantę mediów – media społecznościowe – a głównie w jej wyspę – reklamę. Wyraźnym tego potwierdzeniem są roczne przychody Google z reklamy (głównie w YouTubie), które w 2021 roku wyniosły 61 239 mld USD, co oznacza wzrost o 30% względem 2020 roku i stanowi 81% sumy wszystkich przychodów tej firmy („Alphabet Announces Fourth Quarter and Fiscal Year 2021 Results", 2022, February 1).

Przykładem detekcyjnej siły zastosowanych narzędzi są zidentyfikowane atrybuty wyspy *multimedia*. Kolejność ich prezentacji nie jest przypadkowa – podkreśla różne wartości wag każdego z atrybutów multimediów. Najważniejsze to *video*; potem: *share*, *show*, *watch*, *image*, *jpg*, *say*, *music*, *stream*, *service*, *photo*, *article*, *picture*, *layout*, *audio*, *camera*; ostatni, ale wyróżniający się atrybut to *copyright*.

Ostatnią z grup uzyskanych wyników badań są predykcje – trendy rozwoju poszczególnych lądów. Rysunek 6 ukazuje dynamikę zmian częstotliwości cytowania w materiałach źródłowych jednej z wysp – serwisów informacyjnych. Wartości tych zmian są później wykorzystywane w powszechnie stosowanych modelach predykcyjnych.



Rys. 6. Dynamika zmian zainteresowania problematyką serwisów informacyjnych.
Źródło: opracowanie własne.

Wnioski

Fundamentalnym rezultatem badań jest oczekiwana identyfikacja kontynentów i wysp świata mediów oraz ich atrybutów. Kontynenty te to, według wielkości (liczby cytowań): 1. Media społecznościowe (*Social*); 2. Komunikacja (*Communication*); 3. Cyfryzacja (*Digital*); 4. Polityczne zaangażowanie (*Political*); 5. Publiczne zaangażowanie (*Public*), 6. Informacja (*Information*). Kontynenty wskazują poszukiwane dominanty potencjału biznesowego mediów. Wydaje się, że mogą one być także propozycją zarysu współczesnego paradygmatu mediów. Tworzy go zbiór dominujących – i wartych dyskusji – problemów dzisiejszych mediów. Ułatwieniem w debacie może być znajomość atrybutów definiujących wskazane dominanty. Na przykład jedną z wyróżniających się (ilościowo) dominant (wysp w sąsiedztwie mediów społecznościowych) jest problematyka fejków. Trafnie opisują ją zidentyfikowane atrybuty, m.in. *misinformation*, *disinformation*, *sources*, *campaign*, *detection*, *network*[6].

Przedstawione wyniki badań uzyskano za pomocą rafinacji dużych zasobów informacji – Big Data. Rafinacja – poza przetwarzaniem danych – kompleksowo wspomaga tradycyjną analizę literatury przedmiotu. Dzięki automatyzacji można w krótkim czasie zidentyfikować miliony materiałów źródłowych i przeprowadzić – nieosiągalną dla człowieka – ich kompleksową jakościową oraz ilościową analizę.

Dzięki zastosowaniu rafinacji informacji zarysowano aktualny obraz świata mediów, uzyskano także wyniki wskazujące na dynamikę, kierunek i wielkość zmian zainteresowania tym, co jest najistotniejsze w mediach. Są one propozycją wersji beta paradygmatu przemysłu mediowego.

[6] O wadze problematyki fejków, zidentyfikowanej jako wyróżniająca się dominanta świata mediów, świadczy projekt IKONA (*Identyfikacja, KOlekcjonowanie i oceNA nieprzyjaznych operacji dezinformacyjnych w cyberprzestrzeni*), współfinansowany przez Narodowe Centrum Badań i Rozwoju w ramach Programu Badań Naukowych i Prac Rozwojowych „CyberSecIdent – Cyberbezpieczeństwo i e-Tożsamość", którego beneficjentami są Centrum Rafinacji Informacji Sp. z o.o. i Uniwersytet Warszawski.

Bibliografia

Alphabet Announces Fourth Quarter and Fiscal Year 2021 Results. (2022, February 1). Retrieved on 2021, April 26, from https://www.sec.gov/Archives/edgar/data/1652044/000165204422000015/googexhibit991q42021.htm

Batorski, D., Dziomdziora, W., Gackowski, T., Garapich, A., Kowalski, T., Miczka, T., Ogrodowczyk, A., & Piątek, S. (2015). *Strategia rozwoju rynku medialnego w Polsce 2015–2020*. Warszawa: Fundacja Sztuka Media Film. Pobrano z http://sztukamediafilm.pl/wp-content/uploads/2014/09/SMF-Strategia-rozwoju-rynku-medialnego-w-Polsce-2015-2020.pdf

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C., & Jatowt, A. (2018). YAKE! Collection-Independent Automatic Keyword Extractor. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in Information Retrieval. ECIR 2018. Lecture Notes in Computer Science*, vol. 10772. Cham: Springer. https://doi.org/10.1007/978-3-319-76941-7_80

Gogołek, W. (2017). Refining Big Data. *Bulletin of Science, Technology & Society*, *37*(4), 212–217. https://doi.org/10.1177/0270467619864012

Gogołek, W., & Kuczma, P. (2013). Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych. Część 1. Blogi, fora, analiza sentymentów. *Studia Medioznawcze*, *2*(53), 89–105.

Information technology — Big data — Overview and vocabulary. (n.d.). Retrieved on 2022, April, from https://www.iso.org/obp/ui/#iso:std:iso-iec:20546:ed-1:v1:en

Mayer-Schonberger, V., & Cukier, K. (2017). *Big Data. Rewolucja, która zmieni nasze myślenie, pracę i życie*. Warszawa: MT Biznes.

PricewaterhouseCoopers. (2018, październik). *Perspektywy rozwoju branży rozrywki i mediów w Polsce 2018–2022*. Pobrane z https://www.pwc.pl/pl/pdf/publikacje/2018/media-i-rozrywka-2018-raport-pwc.pdf

Pruchnik, P. (2019). Identyfikacja trendów w polskich mediach na przykładzie kwartalnika „Studia Medioznawcze". Wykorzystanie narzędzi Big Data. *Studia Medioznawcze*, *21*(1), 412–428. https://doi.org/10.33077/uw.24511617.ms.2020.1.113

Silge, J., & Robinson, D. (2019). *Text Mining with R. A Tidy Approach*. O'Reilly Media. Retrieved in 2022 from https://www.tidytextmining.com