COPE
Member since 2019
JM14475

Modelowanie tematyczne w socjologii na przykładzie dobrobytu społecznego: wyzwania metodologiczne i komponent ludzki

Piotr Cichocki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Mariusz Baranowski
Uniwersytet im. Adama Mickiewicza w Poznaniu

https://doi.org/10.18778/1733-8069.20.4.05

Słowa kluczowe: modelowanie tematyczne, metodologia socjologii, dobrobyt społeczny, uczenie maszynowe, przetwarzanie języka naturalnego

Abstrakt: Biorąc pod uwagę dynamicznie rozwijające się obszary nauk społecznych uwarunkowanych technologiami sieciowymi oraz humanistyki cyfrowej (ang. *Digital Humanities*), warto przeanalizować adekwatność socjologicznych metodologii analizy danych w tych nowych warunkach. Dostępność dużych zbiorów zdigitalizowanych danych stanowi nie tylko wyzwanie dla „klasycznych" metod analizy, które opracowane zostały w innych warunkach i do innych celów. Jeszcze ważniejsza kwestia dotyczy tego, czy podział na metody ilościowe i jakościowe, między którymi istnieje wyraźna linia demarkacyjna, ma sens w obliczu *Big Data*. W niniejszym artykule, na podstawie modelowania tematycznego (ang. *topic modeling*), opartego na LDA (ang. *Latent Dirichlet Allocation*), autorzy stawiają tezę, że ilościowe metody (probabilistyczne modele statystyczne) nie stanowią uzupełnienia lub punktu wyjścia do analiz jakościowych (standardowe podejście), lecz ich integralną część. Teza ta zostanie zilustrowana przykładem wyznaczenia tematów w obrębie zbioru 17 278 artykułów na temat dobrobytu społecznego, opublikowanych w czasopismach indeksowanych w bazie Web of Science w latach 1992–2020. To empiryczne studium przypadku posłuży także do sformułowania uwag metateoretycznych na temat „kohezji" metod ilościowych i jakościowych w perspektywie uczenia maszynowego (ang. *machine learning*) i przetwarzania języka naturalnego (ang. *natural language processing* – NLP).

Piotr Cichocki
Doktor, socjolog, pracownik badawczo-dydaktyczny zatrudniony na Wydziale Socjologii Uniwersytetu im. Adama Mickiewicza w Poznaniu. Zainteresowania badawcze: monitorowanie postaw społecznych i politycznych w badaniach międzykrajowych, maszynowa analiza tekstu oraz metodologia badań sondażowych.
e-mail: piotr.cichocki@amu.edu.pl

Mariusz Baranowski
Doktor, socjolog, pracownik badawczo-dydaktyczny zatrudniony na Wydziale Socjologii Uniwersytetu im. Adama Mickiewicza w Poznaniu. Zainteresowania badawcze: socjologia ekonomiczna, socjologia polityki oraz zagadnienia związane z dobrobytem społecznym i transformacją energetyczną.
e-mail: mariusz.baranowski@amu.edu.pl

Wprowadzenie

Ilościowa analiza treści stanowi niezwykle istotną metodę badawczą w naukach społecznych, pozwalającą zredukować korpus danych różnorodnych komunikatów tekstowych do kluczowych składników znaczeniowych. W czasach powszechnej dostępności cyfrowej do różnorodnych treści ta forma analizy danych zastanych zyskuje na znaczeniu. Nie jest oczywiście techniką nową – znajdowała zastosowania od początku rozwoju akademickich nauk społecznych. Już w 1910 roku Max Weber zarysował w wystąpieniu na pierwszym zjeździe Niemieckiego Towarzystwa Socjologicznego rozległy projekt „Socjologii prasy", który miał stanowić narzędzie monitorowania „kulturowej temperatury" społeczeństwa (Jäger, Wiskind, 1991). Choć ten ambitny projekt nie doczekał się nawet próby realizacji, to już w latach dwudziestych ubiegłego wieku zaczęły pojawiać się pierwsze systematyczne wdrożenia analizy treści. Do znanych przykładów należą studia Harolda Lasswella nad propagandą (Lasswell, 1927), które opierały się na analizie częstotliwości i rodzajów słów używanych w artykułach prasowych w celu identyfikacji strategii wpływania na opinię publiczną. W latach pięćdziesiątych analiza treści doczekała się dojrzałej kodyfikacji w pracach Bernarda Berlesona, który zdefiniował ją jako technikę badawczą służącą do obiektywnego, systematycznego i ilościowego opisu jawnej treści przekazów informacyjnych (za Cartwright, 1965: 149). Niemniej jednak, choć znajduje się w repertuarze metod badawczych od dekad, na przestrzeni ostatnich kilkunastu lat możliwości ilościowej analizy treści wzrosły do tego stopnia, że nie tylko zyskała ona na znaczeniu, lecz zmieniła się wręcz nie do poznania.

Skokowy wzrost popularności ilościowej analizy treści mieści się w szerszym nurcie badań zorientowanych na wtórne wykorzystanie danych badawczych, takich jak na przykład harmonizacja danych sondażowych (Jabkowski, Cichocki, Kołczyńska, 2023). Digitalizacja archiwów danych ma szczególnie korzystny wpływ na możliwości hurtowego wykorzystania danych tekstowych (Lewis, Zamith, Hermida, 2013), czego dobrym przykładem są zasoby treściowe oraz narzędzia analityczne udostępniane w ramach klastrów European Strategy Forum on Research Infrastructures (ESFRI) lub Social Sciences and Humanities Open Cloud (SSHOC). Oprócz zwiększonej podaży łatwo dostępnych źródeł danych kluczowym czynnikiem okazał się także równoległy rewolucyjny postęp algorytmicznych technologii

przetwarzania języka naturalnego, dzięki którym badacze uzyskali dostęp do prostych w użyciu narzędzi analitycznych, pozwalających między innymi na określanie tematu, emocji i intencji wyrażanej w tekście (Hirschberg, Manning, 2015). Ważny wydaje się również czynnik popytowy – skoro znaleźliśmy się jako badacze i ludzie w „rwącym strumieniu" treści, to stosowanie ilościowych metod analizy treści stanowi wygodny sposób orientowania się w uniwersum dyskursu. Niniejszy artykuł poświęcony został możliwościom zastosowania tych metod do prowadzenia eksploracyjnych przeglądów literatury akademickiej. Podejmuje również bardziej ogólne kwestie metodologiczne dotyczące ich stosowania w badaniach socjologicznych oraz usytuowania względem metod tradycyjnych.

Przez eksploracyjny przegląd literatury przedmiotu rozumiemy tu takie scenariusze, w których badacze mają potrzebę uzyskania orientacji w dyskursie znajdującym się poza zakresem ich eksperckiej wiedzy odnoszącej się do danego obszaru badawczego. Tego typu przeglądy często znajdują zastosowanie na wstępnych etapach projektów badawczych (Koseoglu, Bozkurt 2018). W naszym przypadku będziemy wykorzystywali zawartość bazy publikacji Web of Science, przeszukanej pod kątem występowania hasła *social welfare*. Pozyskane dane zawierają komponent tekstowy: zawartość pól „Tytuł" i „Streszczenie"; na podstawie analizy tekstu przy użyciu algorytmu modelowania tematów każdemu artykułowi przypisano prawdopodobieństwo przynależności do wyróżnionych trzydziestu jeden odrębnych klastrów treści (tematów). Modelowanie tematyczne, stosunkowo niedawny postęp w przetwarzaniu języka naturalnego (NLP) (Hirschberg, Manning, 2015), wykorzystuje utajoną alokację Dirichleta (LDA) (Blei, Ng Jordan, 2003; Silge, Robinson, 2017: 89–108) w celu rozpoznania wspólnych tematów w korpusie tekstowym i oszacowania ich rozmieszczenia w dokumentach składowych (Baranowski, Cichocki, 2021). Nasza analiza wykorzystuje szczególne podejście LDA – *Structured Topic Modeling*, wykonane w środowisku R (R Core Team, 2022) poprzez pakiet STM (Roberts, Stewart, Tingley, 2019). Alternatywne implementacje podobnych algorytmów są oferowane w ramach takich projektów jak Common Language Resources and Technology Infrastructure (CLARIN-PL) i nie wymagają umiejętności programistycznych.

Uzyskane dzięki LDA klasyfikacje treści stanowią interesujący wgląd w strukturę dyskursu, tym bardziej że mogą być wykorzystywane w powiązaniu z innymi metadanymi charakteryzującymi rekordy publikacyjne – takie jak na przykład: rok publikacji, liczba cytowań czy też klasyfikacja dyscyplinarna. Takie podejście do analizy treści publikacji akademickich nie zastępuje „klasycznego" przeglądu literatury, lecz stanowi dla niego bardzo atrakcyjną alternatywę w sytuacjach, gdy przeglądu wymaga duży zbiór publikacji, którego „klasyczne" przeczytanie wymagałoby wielkich wysiłków techniczno-organizacyjnych. Wykorzystywany przez nas przykład 17 278 streszczeń artykułów stanowi egzemplifikację heterogenicznego korpusu znacznych rozmiarów. Analiza oparta na streszczeniach jest dobrze dopasowana do właściwości algorytmu LDA, ponieważ dokumenty takie są podobnej długości; w przypadku uwzględniania całych artykułów procedura analityczna wymagałaby sztucznego podziału na mniejsze fragmenty.

Modelowanie tematyczne a „klasyczny" przegląd literatury

Dzięki ogromnemu przyrostowi wolumenu publikacji naukowych badacze mają dziś stosunkowo swobodny dostęp do znacznych ilości zgromadzonej wiedzy, jednak śledzenie postępu badań przy użyciu tradycyjnych przeglądów literatury staje się coraz większym wyzwaniem. Szacuje się, że w samych tylko czasopismach akademickich może być publikowanych prawie dwa miliony artykułów rocznie (Altbach, de Wit, 2018). Implementacja modeli tematycznych daje efekty zbliżone do klasycznej ilościowej analizy treści przy zdecydowanie mniejszym nakładzie pracy, szczególnie w sytuacjach, gdy przedsięwzięcia badawcze wymagają stosunkowo szybkiego rozeznania w literaturze znacznych rozmiarów. Algorytmiczne podejście do analizy treści, oparte na LDA (Blei, Ng, Jordan, 2003; Silge, Robinson, 2017: 89–108) stanowi eksploracyjne narzędzie do oglądu tematycznej struktury dyskursu (Baranowski, Cichocki, McKinley, 2023). Metody eksploracyjne stosowane do rozpoznania dużych, luźno ustrukturyzowanych zbiorów danych tekstowych powinny być niedrogie, skalowalne i elastyczne, aby zapewnić ciągły ogląd dyskursu akademickiego (Thangaraj, Sivakami, 2018). Dzięki temu, że nie wymagają one poważnych wysiłków techniczno-organizacyjnych, można po nie sięgać za każdym razem, gdy tylko zachodzi taka potrzeba, uzyskując szybko pożądane rozpoznanie sytuacji.

Klasyczna ilościowa analiza treści ma długą tradycję oraz stabilną od dekad metodologię, obejmującą przeważnie sformułowanie problemu badawczego, dobór próby dokumentów, określenie jednostek obserwacji, ustalenie klucza kodowego, kodowanie oraz analizę statystyczną wyników kodowania (Mayntz, Holm, Hübner, 1976). Podejście klasyczne stanowi szczególne połączenie metody obserwacji i wywiadu, sprowadza się bowiem do swoistego „ankietowania" dokumentu przez kodującego. Uwikłane jest przy tym w liczne trudności metodologiczne i techniczno-organizacyjne, które sprawiają, iż nigdy nie odgrywało pierwszoplanowej roli w badaniach społecznych, a analizy treści przeważnie kojarzono z domeną badań jakościowych. Po pierwsze, dla dużych korpusów tekstu klasyczne podejście okazuje się przeważnie niewykonalne, co rodzi konieczność próbkowania materiału do czytania. Po drugie, choć klucz kodowy powstaje przeważnie po ograniczonej lekturze swobodnej oraz podlega pomiarom pilotażowym, to jednak pozostaje „sztywny" po rozpoczęciu właściwej fazy badania. Po trzecie, trenowanie i nadzorowanie osób kodujących stanowią ogromne wyzwanie, szczególnie w przedsięwzięciach wymagających współpracy bardzo wielu osób. Ograniczenia te sprawiają, że klasyczne podejście działa jedynie w stosunkowo niewielkiej skali, a pomimo najlepszych wysiłków badacza nad wynikami tych działań zawsze zalega cień subiektywności i arbitralności osób kodujących.

W przeciwieństwie do tradycyjnych podejść do przeglądu literatury eksploracja tekstu oparta na LDA nie wymaga narzucania restrykcyjnych kryteriów włączenia dokumentu do analizy (Snyder, 2019). Analiza jest możliwa na podstawie całego korpusu dostępnego tekstu, bez pobierania próbek lub innych form wstępnej selekcji. Co za tym idzie – modelowanie tematów okazuje się najbardziej przydatne w kontekście, w którym wolumen rozpatrywanego tekstu okazuje się zbyt duży, aby można było ją obsłużyć bez pomocy algorytmicznej klasyfikacji tekstu i narzędzi do ekstrakcji treści (DiMaggio, Nag, Blei, 2013: 577). Łatwa skalowalność i zdolność do wielu iteracji sprawiają, że dobrze nadaje się

również do analiz eksploracyjnych zróżnicowanych tematycznie i złożonych korpusów tekstowych (Ananiadou i in., 2009; Baranowski, Cichocki, 2021; Baranowski, 2022). Innymi słowy, nie stanowi realnej alternatywy dla systematycznych przeglądów literatury wykorzystujących ekspercką wiedzę domenową. Zapewnia natomiast atrakcyjne rozwiązanie dla ekstensywnych przeglądów eksploracyjnych; nawet jednak w ich przypadku przesadne wydaje się twierdzenie, jakoby przeglądy manualne stały się już przestarzałe (Asmussen, Møller, 2019: 1).

Przy analizie dużych zbiorów danych tekstowych zastosowanie LDA przypomina pod pewnymi względami jakościowe analizy treści prowadzone w paradygmacie teorii ugruntowanej (Nelson, 2020). Algorytm modelujący tematy automatycznie grupuje teksty na podstawie współwystępowania słów, co przypomina sposób, w jaki teoria ugruntowana kieruje się powtarzającymi się wzorcami i motywami w danych, pozwalając na indukcyjne wyłanianie kategorii i tematów (Carlsen, Ralund, 2022). Ta analogia nie oznacza bynajmniej sugestii, iż modelowanie tematyczne stanowi skomputeryzowaną wersję teorii ugruntowanej, która może potencjalnie zastąpić jakościowe badania tekstu prowadzone w tym paradygmacie. Algorytmy pozwalają szybko i łatwo identyfikować ukryte wzorce w skomplikowanych zbiorach danych, jednak mogą nie uwzględniać kontekstu, subtelności oraz interpretacyjnego podejścia do rzeczywistości społecznej. Podobnie jak w przypadku innych podejść do przeglądu literatury i badania tekstu techniki takie jak LDA należy traktować jako rozwiązania wspomagające, a nie zastępujące. Sensowne użycie metod obliczeniowych do analizy treści wymaga od badacza znajomości metod klasycznych (Isoaho, Gritsenko, Makela, 2021).

Problematyka dobrobytu społecznego (ang. *social welfare*) stanowi niezwykle heterogeniczny obszar badawczy między innymi dlatego, że podejmowana jest przez różne dyscypliny i subdyscypliny naukowe, które zainteresowane są odmiennymi aspektami tego procesu (Baranowski, 2022; Linares, Cabaña, 2022; Baranowski, Cichocki, McKinley, 2023; Nesterova, 2023; Ciziceno, 2024). To z kolei skutkuje „rozłącznymi" konceptualizacjami oraz operacjonalizacjami tego terminu, mającymi przełożenie na wybór podejść badawczych, a także samych czasopism, w obrębie których prezentowane są wyniki badań (Timms, 1980; Forder i in., 2019). Oznacza to w praktyce olbrzymie trudności, jeśli chodzi o próby „całościowych" analiz dobrobytu społecznego, które skutkowałyby np. charakterystyką podejmowanych wątków w czasie, mapowaniem korespondencji pomiędzy różnymi (sub) dyscyplinami naukowymi czy chociażby rozpoznaniem podejść teoretycznych wykorzystywanych w badaniach tego zjawiska. Do tego typu eksploracyjnych analiz przeznaczone są algorytmiczne podejścia do badania treści, takie jak LDA.

Wpływ badacza i intersubiektywna kontrolowalność

Niewątpliwą zaletą analizy treści przy użyciu modeli tematycznych jest możliwość wielokrotnego powtarzania procesu, dzięki czemu uzyskuje się coraz lepsze przybliżenia docelowego obrazu. W przypadku klasycznych analiz treści badacze pozostawali przeważnie związani strukturą klucza kodowego, który nie mógł podlegać znaczącym zmianom po rozpoczęciu właściwego etapu analizy.

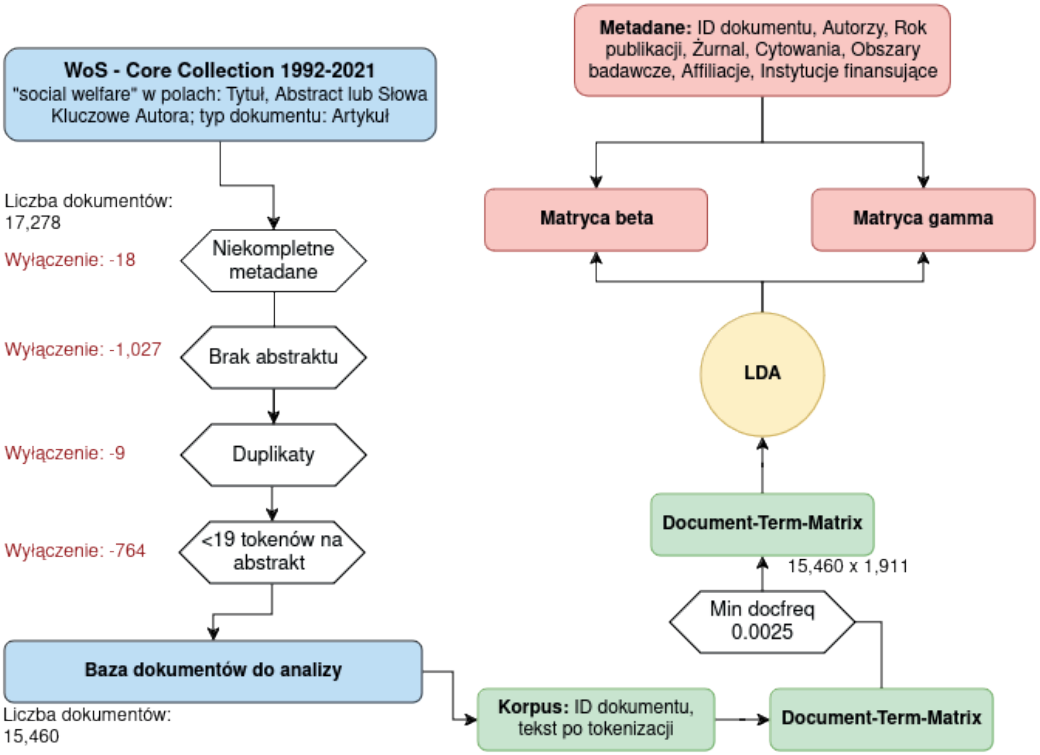
Co więcej, podejście algorytmiczne gwarantuje możliwość replikacji dzięki pełnej dokumentacji decyzji interpretacyjnych. Niemniej jednak nie należy mylić intersubiektywnej kontrolowalności z obiektywnością analizy. Pełne udokumentowanie i powtarzalność procesu badawczego nie są tożsame z uwolnieniem od arbitralnego wpływu decyzji badacza na uzyskiwany wynik. Każda analiza tematyczna wymaga podjęcia wielu decyzji na etapie przygotowania danych oraz wyznaczenia parametrów algorytmu poszukującego ukrytej struktury tematycznej w korpusie. Przede wszystkim wymaga ustalenia z góry liczby wygenerowanych tematów (Mohr, Bogdanov, 2013; Evans, Aceves, 2016; Syed, Spruit, 2018; Pääkkönen, Ylikoski, 2021), czemu poświęcimy zasadniczą część naszej analizy – pozostałe kwestie omawiając jedynie skrótowo.

Modelowanie tematów mieści się w ramach podejścia znanego jako worek słów (ang. *bag-of-words*), które polega na traktowaniu każdego dokumentu jako nieuporządkowanej listy tokenów. Tokeny powstają na bazie słów, które pozbawiane są końcówek fleksyjnych, a często również derywacyjnych. Tematy są definiowane na podstawie rozkładu prawdopodobieństwa występowania określonych tokenów. Choć dane pobierane z serwisu Web of Science są stosunkowo dobrze przygotowane, przez co nie wymagają daleko idących procesów czyszczenia, to ich dostosowanie do analizy przez algorytm LDA wymaga podjęcia kilku czynności przygotowawczych, takich jak tokenizacja (przekształcenie dokumentu w worek słów), lematyzacja (przekształcanie słów do ich podstawowej, słownikowej formy, z uwzględnieniem ich znaczenia w kontekście zdania), oznaczanie części mowy (ang. *POS-tagging*) czy wyodrębnienie często występujących fraz (n-gramizacja) oraz wykluczenie często występujących, lecz pozbawionych znaczenia słów (ang. *stop words*). Stopień agresywności procedur czyszczenia zależy od preferencji oraz potrzeb badacza i obciążony jest dużym ryzykiem popełnienia błędów zniekształcających korpus. W przypadku naszej analizy zastosowaliśmy dosyć standardowe zabiegi, przy czym najdonioślejsze konsekwencje ma decyzja o pozostawieniu w workach słów wyłącznie fraz rzeczownikowych. W połączeniu z wykluczeniem słów wysokiej częstotliwości powoduje to konieczność dodatkowego wyłączenia z analizy 746 dokumentów, co stanowi konsekwencję decyzji, że analizowane będą jedynie wynikowe worki słów zawierające nie mniej niż 19 tokenów. Stanowi to zasadnicze ograniczenie procedur analizy opartych na treści abstraktów, które są przeważnie bardzo krótkimi dokumentami. Po tokenizacji baza jest przekształcana w macierz dokumentów-terminów. Oryginalna macierz DTM (ang. *Document Term Matrix*) ma tyle wierszy, ile dokumentów, a tyle kolumn, ile tokenów. Jest przeważnie zbyt rzadka, tj. zawiera zbyt wiele terminów o niskiej częstotliwości. Redukcja rzadkości jest wymagana, aby wyeliminować elementy o niskiej częstotliwości. W naszym przypadku minimalna częstotliwość występowania została określona na 0,25% dokumentów korpusu, a maksymalna na 50% dokumentów korpusu. Wyłączenie terminów występujących powszechnie w większości dokumentów motywowane jest ich niską mocą rozdzielczą. Przyjęcie tych wartości granicznych wynikało z wcześniejszych doświadczeń oraz wielokrotnych iteracji modelu – konkretne ustawienia należy indywidualnie dopasowywać do charakterystyk analizowanego korpusu w kontekście celów badawczych.

Na rysunku 1 przedstawiono przeglądową ilustrację struktury procesu badawczego, który opiera się na wykorzystaniu wyników modelowania tematów w połączeniu z dostępnymi metadanymi.

W ramach wykorzystywanego przez nas modelu analizy algorytm LDA nie ma dostępu do metadanych – modelowanie wykonywane jest wyłącznie na podstawie zawartości tytułu i streszczenia. Natomiast wyniki modelowania w postaci matryc beta (temat-token) oraz gamma (dokument-temat) mogą być analizowane w kontekście innych znanych charakterystyk dokumentów. Dotyczy to w szczególności matrycy gamma, która przypisuje każdemu dokumentowi prawdopodobieństwo występowania każdego z wyróżnionych w modelowaniu tematów. Istotne znacznie dla wyników modelowania mają przyjęte wartości hiperparametrów alfa (jak prawdopodobne jest, że dokument będzie mieszanką więcej niż jednego tematu) oraz delta (jak prawdopodobne jest, że token będzie należał do więcej niż jednego tematu). W ramach przeprowadzonej przez nas analizy przyjęliśmy ustawienia zbliżone do domyślnych, ponieważ wstępne wyniki modelowania nie sugerowały konieczności ich zmieniania. Natomiast możliwość manipulowania tymi wartościami, jak również innymi parametrami algorytmu, takimi jak na przykład liczba dopuszczalnych iteracji, stanowi ścieżkę wpływu na wyniki analizy nie zawsze uświadamianą i zrozumiałą dla samych badaczy.

Rysunek 1. Modelowanie tematów: struktura procesu badawczego



Źródło: opracowanie własne.

Wpływ badaczy na wyniki modelowania tematycznego zaznacza się najwyraźniej przy podejmowaniu decyzji o liczbie tematów, które ma wyznaczyć algorytm. Wartość ta musi zostać określona z góry i nie ma ustawienia domyślnego. Istnieją różne sposoby szacowania pożądanej liczby tematów, badacze mogą również mieć określone oczekiwania wynikające z ich wiedzy lub celów prowadzonej analizy. Typowy dylemat polega w tym wypadku na podjęciu decyzji, czy preferuje się mniejszą

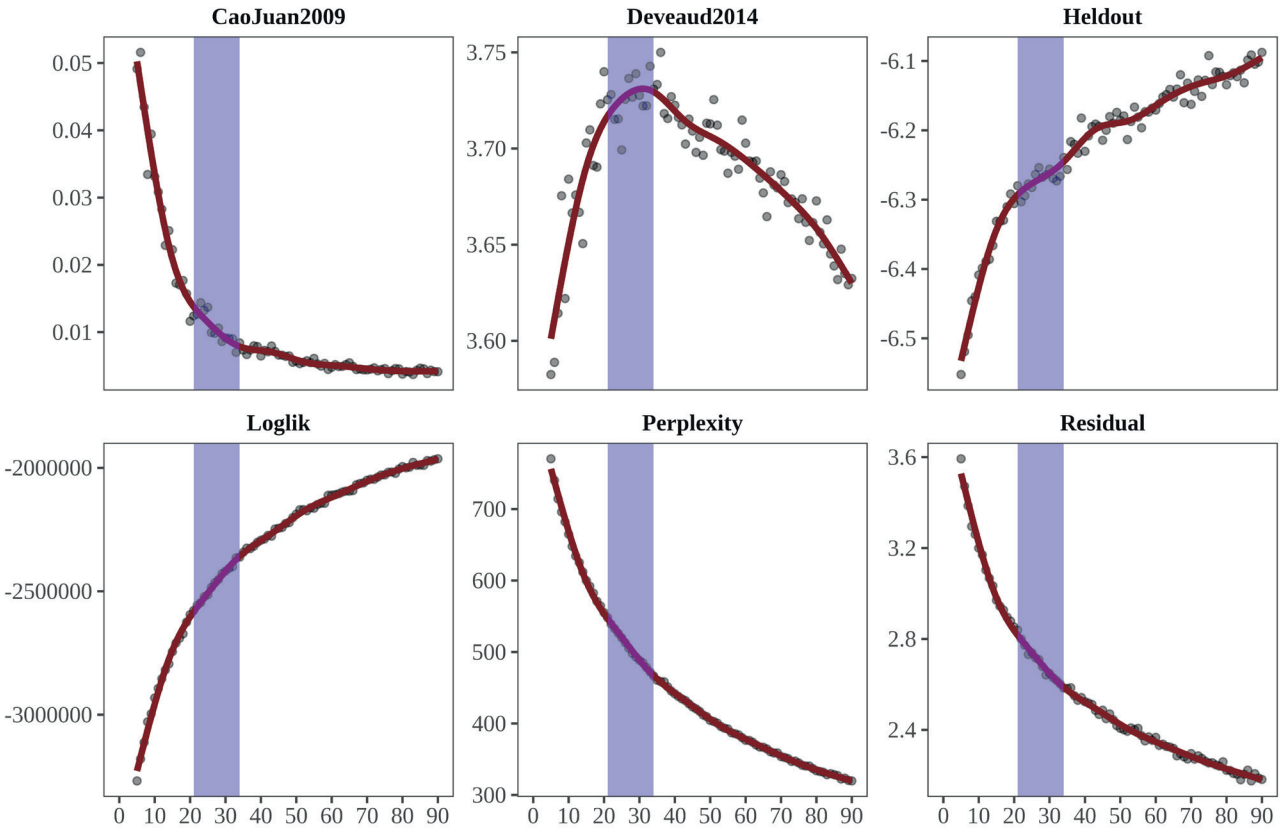
liczbę bardziej ogólnych tematów, czy też dużą liczbę stosunkowo małych i szczegółowo określonych tematów. W ramach przeprowadzonej przez nas analizy porównamy dwa możliwe warianty: modele wyznaczające 21 oraz 34 tematy.

Arbitralność liczby wyznaczanych tematów: 21 czy 34 tematy

Choć w nazwie procedury LDA pojawia się określenie „latentny" (ang. *latent*), to nie jest tak, że dla każdego korpusu istnieje jedna ukryta struktura tematyczna, która wyłania się poprzez modelowanie. Podejmując się analizy, badacze przeważnie wykorzystują procedury pozwalające na oszacowanie pożądanego rzędu wielkości, a następnie porównują kilka modeli i decydują się ostatecznie na jeden model końcowy, który przedstawiają jako wynik analizy. Decyzja o wyborze modelu pozostaje arbitralna, nawet jeśli może znaleźć oparcie w oszacowaniach statystycznych lub argumentach merytorycznych. W ramach stosowanego przez nas podejścia analitycznego punktem wyjścia do selekcji modelu jest przeprowadzenie modelowania dla różnych możliwych nastawień parametrów, a następnie obliczenie metryk statystycznych, które podlegają porównaniu wizualnemu. Na rysunku 2 przedstawiono wyniki takiego porównania dla sześciu wybranych metryk: CaoJuan2009 oraz Deveaud2014 z pakietu „ldatuning", Loglik oraz Perplexity z pakietu „topicmodels", Heldout likelihood oraz Residuals z pakietu „stm". Każda z tych metryk opiera się na określonych statystycznych miarach dopasowania, ich szczegółowa specyfikacja została opublikowana w dokumentacji technicznej pakietów. Możliwych do zastosowania metryk jest o wiele więcej – nie ma też żadnego rankingu ich jakości lub użyteczności. Ich zastosowanie nie daje jednoznacznego rozwiązania w postaci najlepszego ustawienia parametrów algorytmu, ale dostarcza badaczowi wskazań, w jakich zakresach należy tych odpowiednich dla prowadzonej analizy szukać. Oprócz manipulowania liczbą tematów można również w ten sposób porównywać różne nastawienia innych parametrów modelu, jednakże w przeprowadzonej przez nas analizie zmienną była wyłącznie liczba tematów.

W celu rozpoznania struktury tematycznej wykonano estymację kolejnych modeli dla nastawień od $K = 3$ do $K = 90$ oraz dla każdego modelu obliczono wartości sześciu statystyk, które porównano na rysunku 2. Niebieskim paskiem oznaczono przedział, w ramach którego mieści się docelowa liczba wyznaczanych tematów, opierając się na wizualnej interpretacji danych. Dla celów porównania wybraliśmy dwie skrajne wartości tego przedziału: 21 oraz 34 tematy, co jednak nie oznacza, że do analizy nie można by wybrać jakiejś wartości pośredniej. Takie podejście umożliwi eksplorację wchłaniania/rozpadu tematów pomiędzy dwoma różnolicznymi wariantami zastosowanego modelowania na poziomie ogólnym. Dodatkowo będziemy w stanie prześledzić dokładne „transfery" pomiędzy węższą i szerszą strategią szacowania liczby tematów, które same w sobie stanowią unikatową wiedzę o relacjach pomiędzy zawartością treści w publikacjach na temat dobrobytu społecznego. Jednak to, co najważniejsze w tych analizach, dotyczy unikatowego i innowacyjnego sposobu eksploracji rezultatów badań naukowych poświęconych dobrobytowi społecznemu w szerokim, gdyż obejmującym prawie trzy dekady okresie, oraz multi- czy transdyscyplinarnym sensie.

Rysunek 2. Metryki dopasowania modeli tematycznych



Źródło: opracowanie własne.

Tabele 1 i 2 prezentują zestawienie kluczowych tokenów dla 21 oraz 34 tematów. Do selekcji tokenów zastosowano miarę FREX (ang. *Frequency and Exclusivity*), która uwzględnia zarówno częstotliwość (tj. to, jak często dane słowo pojawia się w dokumentach przypisanych do konkretnego tematu w porównaniu do jego ogólnej częstotliwości we wszystkich tematach), jak i ekskluzywność (tj. to, jak specyficzne lub ekskluzywne jest słowo dla danego tematu w porównaniu z innymi tematami – słowo uznaje się za bardziej ekskluzywne, jeśli rzadko występuje w innych tematach). Tokeny pozwalają na identyfikację semantyki tematów, które zostały przez nas również opisowo zanalizowane w kontekście skojarzonych z nimi artykułów oraz czasopism. W celu uporządkowania prezentacji i dyskusji tematy zostały uszeregowane przy użyciu hierarchicznej analizy skupień, przeprowadzonej na matrycy gamma. Skupienia oznaczają zatem skłonność do współwystępowania w dokumentach. W przypadku K-21 wyróżniono pięć, a dla K-34 osiem skupień. Dla każdego tematu zachowano oryginalny numer tematu nadany przez model.

Tabela 1. Zestawienie kluczowych tokenów dla 21 tematów

K	t.21	Kluczowe tokeny (top-7 FREX)
1	9	consumer, advertising, price, retailer, producer, profit, surplus
	12	firm, merger, competition, oligopoly, vertical, trade, private_firm
	1	emission, carbon, green, supply_chain, carbon_emission, investment, port
	2	contract, cost, externality, incentive, fix, credit, sharing
2	5	service, customer, digital, queue, service_provider, provider, delivery
	10	transport, vehicle, travel, road, rail, passenger, evs
	20	load, electricity_market, demand_response, electricity, wind, energy, grid
	3	agent, fairness, coalition, negotiation, approximation, player, game
	13	spectrum, auction, wireless, cloud, resource_allocation, peer, task
3	17	inequality, redistribution, equity, wealth, preference, tax, taxation
	14	social_policy, democracy, political, ideology, discourse, neoliberal, religious
	18	law, corporate, legal, ethical, enforcement, court, moral
	4	innovation, economy, growth, patent, economic_growth, capital, technology
	8	bank, finance, financial, spending, fiscal, public, federal
4	6	patient, hospital, disease, health, covid, medical, health_insurance
	19	disability, student, university, life, professional, person, social_worker
	7	child, parent, abuse, mother, family, violence, school
	15	unemployment, immigrant, employment, poverty, suicide, status, labor_market
5	11	water, agricultural, sustainable, farmer, forest, management, crop
	16	housing, rural, migrant, urban, city, parking, resident
	21	community, region, local, tourism, initiative, social_capital, territory

Źródło: opracowanie własne.

W ramach pierwszego klastra znajdują się cztery tematy (T1, T2, T9 oraz T12) pokrywające obszar zainteresowania ekonomii dobrobytu wzbogaconej o kwestie wąsko rozumianych zmian klimatu (T1). Zawężone ujęcie zmian klimatu jest konsekwencją koncentracji głównie na problematyce polityki węglowej, podatku węglowym, systemie handlu emisjami dwutlenku węgla oraz na wycenie kontroli zanieczyszczeń. Trzy pozostałe tematy są ściśle związane z rozwijaną od początku XX wieku

ekonomią dobrobytu, co widać wyraźnie po tytułach czasopism, z których pochodzą dokumenty (np. „Journal of Economics", „Applied Economics", „Journal of Economic Theory", „MIS Quarterly").

Drugi klaster zawiera pięć tematów (T3, T5, T10, T13, T20) reprezentujących mniej spójne obszary zagadnień. Temat T3 dotyczy teorii gier i zachowań odnoszących się do optimum w sensie Pareto lub równowagi Nasha, a T5 skupia się na wąskim obszarze matematycznej teorii prawdopodobieństwa określanej jako łańcuchy lub procesy Markowa, aplikowanej do zachowań w ramach optymalnego dobrobytu społecznego. W T10 poruszane są kwestie różnych form transportu indywidualnego i zbiorowego w kontekście optymalizacji w postaci maksymalizacji budżetu dobrobytu społecznego. Natomiast T13 również stanowi wyspecjalizowany obszar technologii teleinformatycznych, ujmowanych w ramach perspektywy *optimal social welfare* (por. Duan i in., 2022), a T20 poświęcony jest energetyce, w tym odnawialnym źródłom energii, w świetle dotacji rządowych i perspektywy łańcuchów dostaw.

Do trzeciego klastra zaklasyfikowane zostały T4, T8, T14, T17 i T18. W przypadku T4 mamy do czynienia tematyką zogniskowaną wokół kwestii wzrostu i rozwoju gospodarczego, z dużym naciskiem na problematykę innowacyjności i prac badawczo-rozwojowych. Finansowe i fiskalne zagadnienia wydatków publicznych związanych z utrzymaniem m.in. systemu ubezpieczeń społecznych czy administracji różnych poziomów podejmuje T8. Natomiast T14 skupia się na polityce społecznej oraz społeczeństwie obywatelskim, jak również na powiązanym z nimi zagadnieniu sprawiedliwości społecznej, a T17 dotyczy problematyki nierówności ujmowanej od strony redystrybucyjnej funkcji państwa oraz podejścia utylitarnego (dominują tu publikacje z czasopism „Social Choice and Welfare" oraz „Economics Letters"). Ostatni w tym klastrze jest T18, czyli prawny i etyczny wymiar instytucji dobrobytu społecznego.

Czwarty klaster składa się z T6, T7, T15 i T19. W ramach T6 mieszczą się zagadnienia dotyczące zdrowia publicznego, łącznie z ubóstwem żywnościowym. Artykuły poświęcone tej problematyce publikowane są w „International Journal of Health Services", „Public Health", „Food Security", „Journal of Poverty" czy „British Food Journal". Zjawiska związane z dobrobytem społecznym (i jego brakiem) w kontekście rodzicielstwa oraz usług społecznych skierowanych do dzieci i młodzieży obejmuje T7 („Child & Family Social Work", „Journal of Child and Family Studies", „Children and Youth Services Review", „Journal of Social Welfare and Family Law"). Obszar problemów społecznych w wymiarze bezrobocia, bezdomności, uzależnień, migracji, czyli grup defaworyzowanych, wyznacza T15. Natomiast T19 rozszerza poprzednie kategorie o osoby z niepełnosprawnościami, studentów, seniorów („International Journal of Special Education", „Disability & Society", „Aging – Clinical and Experimental Research", „Social Work").

Ostatni klaster tworzą tematy T11, T16 oraz T21, które są najbliższe geografii społeczno-ekonomicznej i regionalnej. I tak T11 dotyczy kwestii środowiskowych związanych z ekosystemami, bioróżnorodnością, polityką wodną i rozwojem zrównoważonym („Ecological Economics", „International Journal of Agricultural Sustainability", „Reviews in Fisheries Science", „Water"), T16 obejmuje tematykę

miejską i wiejską, z kwestiami mieszkaniowymi, parkami, parkingami i transportem włącznie, a T21 poszerza poprzedni temat o zagadnienia turystyki, kapitału społecznego oraz koncepcję *empowerment* („International Regional Science Review", „Annals of Tourism and Cultural Change", „Journal of Urban Planning").

Tabela 2. Zestawienie kluczowych tokenów dla 34 tematów

K	T.34	Kluczowe tokeny (top-7 FREX)
1	4	innovation, patent, technology, software, licensing, development, knowledge
	29	platform, adoption, payment, peer, blockchain, compatibility, transaction
	1	emission, carbon, investment, port, carbon_emission, carbon_tax, tariff
	20	manufacturer, retailer, supply_chain, subsidy, green, brand, producer
	2	contract, sharing, toll, regulation, airport, transfer, demand
	9	advertising, search, price, purchase, price_discrimination, seller, buyer
	12	merger, firm, competition, vertical, entry, oligopoly, public_firm
2	17	preference, voting, alternative, utilitarian, equity, aggregate, rank
	3	cooperation, coalition, game, player, nash_equilibrium, equilibria, stable
	25	fairness, agent, negotiation, approximation, allocation, valuation, mechanism_design
3	5	customer, service, queue, service_provider, delivery, provider, tier
	10	transition, rail, transport, passenger, infrastructure, speed, fare
	16	city, urban, parking, location, driver, ride, spatial
4	6	poverty, household, health, food, inequality, health_insurance, poor
	21	housing, region, economic_development, rural, tourism, village, local
	24	corporate, organizational, islamic, business, social_enterprise, entrepreneurship, ethical
	8	public, local_government, corruption, spending, federal, official, government
	23	immigrant, migrant, european, migration, europe, social_assistance, attitude
	32	discourse, neoliberal, religious, liberal, social_justice, ideology, advocacy

5	19	disability, person, employment, life, worker, social_security, retirement
	15	status, mortality, depression, suicide, socioeconomic, unemployment, neighborhood
	22	patient, hospital, medical, cancer, nursing, disease, therapy
	14	student, volunteer, university, education, sport, voluntary, language
	7	child, mother, parent, family, maternal, child_welfare, father
	28	violence, abuse, drug, sexual, emotional, crime, intervention
6	11	water, farmer, agricultural, crop, land, farm, conservation
	27	sustainable, industry, sustainable_development, environmental, mining, waste
7	31	demand_response, load, electricity, wind, electricity_market, energy, grid
	13	spectrum, cloud, auction, resource_allocation, virtual, truthful, computational
	33	security, network, transmission, device, mobile, expansion, node
8	26	tax, taxis, capital, taxation, redistribution, financial, wage
	34	bank, inflation, exchange, shock, money, monetary_policy, finance
	18	law, legal, court, enforcement, intellectual_property, agreement, constitutional
	30	insurance, oil, risk, liability, compensation, loss, aversion

Źródło: opracowanie własne.

Ponieważ liczba wygenerowanych tematów zależy ostatecznie od decyzji badaczek/badaczy, choć posiłkują się one/oni określonymi parametrami (por. wykres 1), proponujemy przeanalizować wariant z 34 wątkami. Miary przedstawione na wykresie 1 mają charakter pomocniczy i nie wskazują jednej precyzyjnej i optymalnej liczby obszarów tematycznych. Ponadto LDA jest metodą eksploracji danych, a więc liczba wyeksponowanych tematów zależy od wyartykułowanego celu badawczego. Z perspektywy metodologicznego oglądu modelowania tematycznego uważamy, że taki „eksperyment" jest niezwykle użyteczny w celu zrozumienia mechanizmu działania LDA.

Wizualizację studium relokacji 21 tematów w 34 przedstawia rysunek 3, w którym za pomocą różnego rodzaju strzałek (por. legenda do rysunku 3) określono odsetek dokumentów przemieszczonych w ramach dwóch badanych wariantów. Jeśli chodzi o samą alokację dokumentów w ramach różnolicznych zbiorów tematów, to przyjmuje ona postać rozwarstwienia na 2, 3, 4 lub 5 w poszerzonej, tj. 34-tematycznej opcji.

Wyjdźmy od sytuacji rozszczepienia jednego tematu w dwa bardziej szczegółowe. Opisany powyżej T21 (K5), który dotyczy kwestii miejsko-wiejskich z rozbudowanym komponentem turystyki i kapitału społecznego, rozpada się na T21 i T32 (oba w K4) w poszerzonej wersji 34 *topics*. Szczególnie

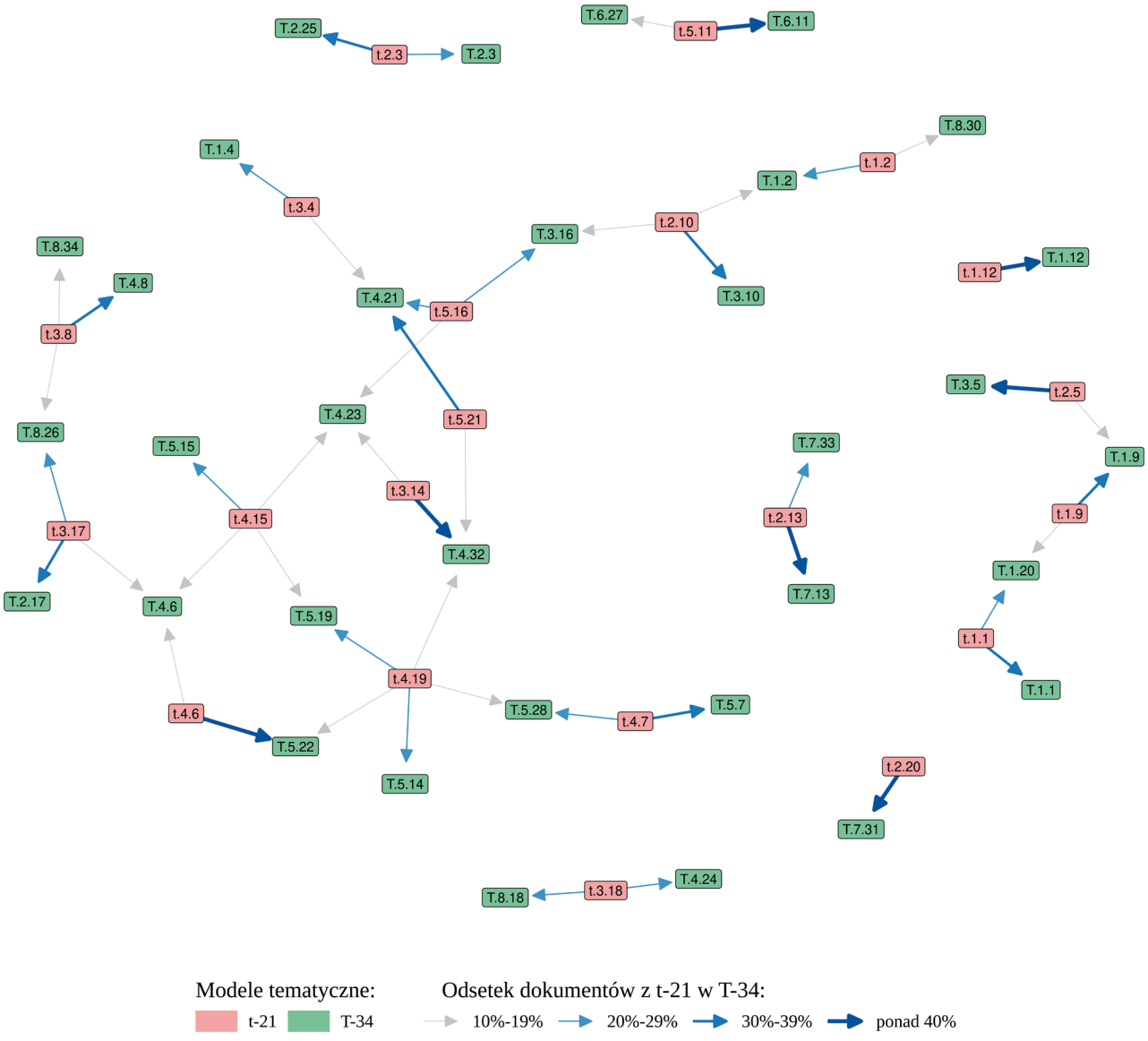
będzie nas interesował „nowy" T21, ponieważ transfer dokumentów między węższym i szerszym T21 mieści się w przedziale od 30% do 39%. Na przykładzie tokenów wyróżnionych w tabeli 2 widać wyraźną koherencję pomiędzy obszarami tematycznymi (co jest konsekwencją przypisania części tych samych dokumentów do obu tematów). Jednak zawartość wąskiego T21 znajduje także reprezentację (od 10% do 19% dokumentów) w szerokim T32. Ten ostatni natomiast wiąże się z problematyką polityczno-kulturową oraz kwestiami sprawiedliwości społecznej, publikowanymi w takich czasopismach, jak „Journal of Policy Practice", „Critical Social Policy", „Development and Change", „Critical and Radical Social Work".

Weźmy za kolejny przykład T10 (K2), dotyczący różnorodnych form transportu indywidualnego i zbiorowego, który rozpadł się na T10 (K3), T2 (K3) i T16 (K3). W poszerzonym wariancie T10 dotyczy bezpośrednio transportu oraz opłat. Największy odsetek dokumentów wyjściowego tematu (30–39%) zawiera się w tym poszerzonym obszarze treści. Dwa pozostałe (T2 i T16) konweniują na niższym poziomie 10–19%. Pierwszy obszar obejmuje zakres różnych form transportu z mocno eksponowanym komponentem opłat i regulacji, drugi (T16) akcentuje zaś problematykę komunikacji w ujęciu miejskim i przestrzennym.

Przyjrzyjmy się jeszcze jednej sytuacji, w której temat T19 rozpada się na pięć mniejszych, z czego dwa są silniej wiążące (T14 i T19) niż pozostałe (T22, T28, T32). Przypomnijmy, że T19 dotyczył problemów społecznych z akcentem na osoby z niepełnosprawnościami, studentów oraz seniorów. Dokumenty z tego tematu formułują T14, czyli obszar poświęcony edukacji na poziomie wyższym, oraz T19, czyli kwestie niepełnosprawności w kontekście zatrudnienia i pomocy społecznej. Jeśli chodzi o pozostałe trzy tematy, to dotyczą kolejno: (T22) systemu opieki zdrowotnej, (T28) przestępstw i uzależnień oraz (T32) wymiaru kulturowo-politycznego z akcentami neoliberalnej ideologii. Cała paleta pięciu wyodrębnionych w poszerzonym wariancie tematów reprezentuje odmienne, ale wewnętrznie spójne obszary dobrobytu społecznego.

Rozważając warianty wyboru liczby obszarów tematycznych, które podporządkowane są przede wszystkim celom i problemom badawczym, warto mieć na uwadze następstwa określonych decyzji. Otóż, jak wynika z dokonanych powyżej analiz, mniejsza liczba tematów wiąże się z możliwością „wchłonięcia" obszarów tematycznych relatywnie rzadziej występujących, ale jednak istotnych dla problematyki *social welfare* (przykład T10 w ramach 21 tematów). Poszerzona paleta tematów pozwala zatem zniuansować duże obszary tematyczne, które są – co widać przy 34 tematach – bardziej zróżnicowane wewnętrznie (perspektywa mniejszej liczby tematów nie oddaje tej różnorodności).

Rysunek 3. Przesunięcia dokumentów między modelami



Źródło: opracowanie własne.

Dyskusja

Przeprowadzona analiza skupiała się na możliwościach zastosowania modelowania tematycznego do eksploracyjnych przeglądów literatury. Poza ogólną prezentacją na przykładzie streszczeń artykułów dotyczących problematyki dobrobytu społecznego szczególną uwagę zwrócono na konsekwencje decyzji badacza odnośnie do poszukiwanej przez algorytm LDA liczby tematów. Publikacje wykorzystujące modelowanie tematyczne przeważnie prezentują jeden wynikowy model, który uznano za najlepszą odpowiedź na pytania badawcze. Arbitralny wpływ badacza na wynik zostaje w ten sposób „schowany pod płaszczem" obiektywności algorytmu. W przypadku modelowania tematycznego

wybór liczby tematów pozostaje decyzją najsilniej wpływającą na uzyskany wynik, co zademonstrowano, porównując dwa modele T21 oraz T34, które znajdowały się na dwóch skrajach wyznaczonego pasma poszukiwań. Są też możliwe rozwiązania „eklektyczne", w których wyodrębniamy niewielką liczbę tematów w celu zmapowania danej problematyki, by następnie pogłębić tę wiedzę za pomocą ponownego zastosowania LDA do „dużych" tematów (por. Baranowski, Cichocki, 2021; Baranowski, Cichocki, McKinley, 2023). Bez względu na to, jakie podejście zostanie zastosowane, otrzymujemy wartościowy materiał badawczy, który – biorąc pod uwagę choćby wolumin uwzględnionych dokumentów – pozwala eksplorować dotąd „trudno analizowalne" bazy danych (a na dodatek cała metoda oparta jest na przejrzystej i intersubiektywnie kontrolowalnej procedurze).

Klasyfikacja tekstu do tematu może być również analizowana w kontekście innych znanych charakterystyk, takich jak łatwo dostępne w przypadku publikacji akademickich metadane. Dobrym przykładem są w tym wypadku analizy wykorzystujące czas publikacji lub kategoryzację dyscyplinarną (Jakubowska, Cichocki, Jabkowski, 2023). Atrakcyjne wydaje się również wykorzystanie modelowania tematycznego w powiązaniu z klasycznymi metodami analizy tekstu. Biorąc pod uwagę pewne podobieństwa podejścia LDA do analiz prowadzonych w paradygmacie teorii ugruntowanej (Nelson, 2020; Carlsen, Ralund, 2022), szczególnie interesujące wydają się możliwości wykorzystania modeli tematycznych jako metody wspierającej lub metodologicznie triangulującej jakościowe analizy tekstu (Jacobs, Tschötschel, 2019). Mając na uwadze rewolucyjne postępy w komputerowym przetwarzaniu języka naturalnego, uczeniu maszynowym i sztucznej inteligencji (Battista, 2024), konieczne wydaje się przemyślenie klasycznych metod socjologicznych. Doświadczenia badawcze z modelowaniem tematycznym wskazują, iż wraz z zanikaniem klasycznej dystynkcji między danymi ustrukturyzowanymi i nieustrukturyzowanymi nie do utrzymania może się okazać linia demarkacyjna tradycyjnie rozdzielająca metody ilościowe od jakościowych.

Zakończenie

W obliczu dostępu do niespotykanej wcześniej ilości zdigitalizowanych danych dotyczących funkcjonowania jednostek i społeczeństw dla rozwoju nauk społecznych coraz ważniejsze staje się wykorzystywanie metod implementujących rozwiązania z zakresu sztucznej inteligencji. Szczególnie w przypadku analizy danych tekstowych metody algorytmiczne są nie tylko tańsze w realizacji i wolne od obciążeń wynikających z konieczności próbkowania korpusów w podejściach klasycznych, ale umożliwiają także eksplorację ogromnych zbiorów tekstu w krótkim czasie. Modelowanie tematów stanowi jedno z wielu obiecujących podejść analitycznych, które pozwalają na eksploracyjne rozpoznanie różnorodnych treści. Wykorzystanie modelowania tematycznego sugeruje konieczność przemyślenia na nowo klasycznego podziału na metody ilościowe i jakościowe analizy treści, ze względu na łatwość rozpoznawania struktury w danych do niedawna traktowanych jako nieustrukturyzowane i wymagające ludzkiej interpretacji. W niniejszym artykule skupiliśmy się przede wszystkim na modelach tematycznych w publikacjach dotyczących dobrobytu społecznego,

co stanowi jedynie wycinek możliwych zastosowań procedur analizy języka naturalnego do analizy treści (por. Naskar i in., 2016; Akhmedov i in., 2021).

Zastosowanie modelowania tematycznego do zjawiska tak zróżnicowanego i podejmowanego przez różne dyscypliny naukowe jak dobrobyt społeczny (Titmuss, 1967; van Praag, 1989; Midgley, 1997; Adler, 2019; Ciziceno, 2024) potraktowaliśmy jako swego rodzaju poligon doświadczalny dla rozpatrzenia atutów oraz słabości uwikłanych w stosowanie modeli LDA. Ich niewątpliwą zaletą pozostaje zdolność do generowania tematów w obrębie znacznej liczby artykułów opublikowanych w czasopismach indeksowanych w Web of Science, co pozwala na uzyskanie przeglądu problematyki podejmowanej w badaniach nad dobrobytem społecznym. Podejście to pozwoliło również pozyskać wiedzę na temat tendencji do współwystępowania tematów w określonych dokumentach. Konkretny przykład zastosowania tej metody do analizy piśmiennictwa poświęconego dobrobytowi społecznemu pozwolił na wyodrębnienie „ukrytych" tematów, które mogą nie być uświadamiane przez badaczy pracujących w ramach tego zróżnicowanego teoretycznie i metodologicznie podejścia. Przejrzystość metody oraz możliwość replikacji wyników sugerują, że maszynowe, zautomatyzowane algorytmy będą coraz częściej „zastępować" opracowane wiele lat temu metody analizy zawartości treści, oparte na badaczach i opracowanych przez nich kluczach kodowych. „Zastępowanie" oznacza w tym przypadku integrację probabilistycznych modeli statystycznych z „klasycznie" jakościowymi metodami analizy treści, co z jednej strony oznacza wykraczanie poza tradycyjne podziały na ilościowe i jakościowe podejścia, a z drugiej – ze względu na zautomatyzowany charakter tych modeli – na objęcie analizą dużych wolumenów danych.

Źródło finansowania

Dr Mariusz Baranowski otrzymał finansowanie z Narodowego Centrum Nauki, nr grantu: 2021/05/X/HS6/00067.

Bibliografia

Adler Matthew D. (2019), *Measuring Social Welfare: An Introduction*, Oxford: Oxford University Press.

Akhmedov Farkhod, Abdusalomov Akmalbek, Makhmudov Fazliddin, Cho Young I. (2021), *LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model*, „Applied Sciences", vol. 11(23), 11091, https://doi.org/10.3390/app112311091

Altbach Philip G., Wit Hans de (2018), *Too much academic research is being published*, „University World News", 7 September, https://www.universityworldnews.com/post.php?story=20180905095203579 [dostęp: 24.09.2024].

Ananiadou Sophia, Rea Brian, Okazaki Naoaki, Procter Rob, Thomas James (2009), *Supporting Systematic Reviews Using Text Mining*, „Social Science Computer Review", vol. 27(4), s. 509–523, https://doi.org/10.1177/0894439309332293

Asmussen Claus Boye, Møller Charles (2019), *Smart literature review: a practical topic modelling approach to exploratory literature review*, „Journal of Big Data", vol. 6(93), s. 1–18, https://doi.org/10.1186/s40537-019-0255-7

Baranowski Mariusz (2022), *Epistemological aspect of topic modelling in the social sciences: Latent Dirichlet Allocation*, „Przegląd Krytyczny", vol. 4(1), s. 7–16, https://doi.org/10.14746/pk.2022.4.1.1

Baranowski Mariusz, Cichocki Piotr (2021), *Good and bad sociology: does topic modelling make a difference?*, „Society Register", vol. 5(4), s. 7–22, https://doi.org/10.14746/sr.2021.5.4.01

Baranowski Mariusz, Cichocki Piotr, McKinley Jim (2023), *Social welfare in the light of topic modelling*, „Sociology Compass", vol. 17(8), e13086, https://doi.org/10.1111/soc4.13086

Battista Daniele (2024), *Political communication in the age of artificial intelligence: an overview of deepfakes and their implications*, „Society Register", vol. 8(2), s. 7–24, https://doi.org/10.14746/sr.2024.8.2.01

Blei David M., Ng Andrew Y., Jordan Michael I. (2003), *Latent Dirichlet Allocation*, „Journal of Machine Learning Research", vol. 3, s. 993–1022.

Carlsen Hjalmar, Ralund Snore (2022), *Computational grounded theory revisited: From computer-led to computer-assisted text analysis*, „Big Data & Society", vol. 9(1), https://doi.org/10.1177/20539517221080146

Cartwright Dorwin P. (1965), *Zastosowania analizy treści*, [w:] Stefan Nowak (red.), *Metody badań socjologicznych*, Warszawa: Państwowe Wydawnictwo Naukowe, s. 149–161.

Ciziceno Marco (2024), *Who will take care of them? A reflection on Southern European welfare regimes*, „Society Register", vol. 8(1), s. 27–42, https://doi.org/10.14746/sr.2024.8.1.02

DiMaggio Paul, Nag Manish, Blei David (2013), *Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding*, „Poetics", vol. 41(6), s. 570–606, https://doi.org/10.1016/j.poetic.2013.08.004

Duan Jingyuan, Tian Ling, Mao Jianqiao, Li Jiaxin (2022), *Optimal social welfare: A many-to-many data transaction mechanism based on double auctions*, „Digital Communications and Networks", vol. 9(5), s. 1230–1241, https://doi.org/10.1016/j.dcan.2022.04.020

Evans James A., Aceves Pedro (2016), *Machine Translation: Mining Text for Social Theory*, „Annual Review of Sociology", vol. 42, s. 21–50, https://doi.org/10.1146/annurev-soc-081715-074206

Forder Anthony, Caslin Terry, Ponton Geoffrey, Walklate Sandra (2019), *Theories of welfare*, London: Routledge.

Hirschberg Julia, Manning Christopher D. (2015), *Advances in natural language processing*, „Science", vol. 349(6245), s. 261–266, https://doi.org/10.1126/science.aaa8685

Isoaho Karoliina, Gritsenko Daria, Mäkelä Eetu (2021), *Topic Modeling and Text Analysis for Qualitative Policy Research*, „Policy Studies Journal", vol. 49, s. 300–324, https://doi.org/10.1111/psj.12343

Jabkowski Piotr, Cichocki Piotr, Kołczyńska Marta (2023), *Multi-Project Assessments of Sample Quality in Cross-National Surveys: The Role of Weights in Applying External and Internal Measures of Sample Bias*, „Journal of Survey Statistics and Methodology", vol. 11(2), s. 316–339, https://doi.org/10.1093/jssam/smab027

Jacobs Thomas, Tschötschel Robin (2019), *Topic models meet discourse analysis: a quantitative tool for a qualitative approach*, „International Journal of Social Research Methodology", vol. 22(5), s. 469–485, https://doi.org/10.1080/13645579.2019.1576317

Jakubowska Honorata, Cichocki Piotr, Jabkowski Piotr (2023), *References to sex and gender differences in the social sciences: analysis of journal publication records (1971–2021)*, „Ruch Prawniczy, Ekonomiczny i Socjologiczny", vol. 85(4), s. 275–297, https://doi.org/10.14746/rpeis.2023.85.4.14

Jäger Friedrich, Wiskind Ora (1991), *Culture or Society? The Significance of Max Weber's Thought for Modern Cultural History*, „History and Memory", vol. 3(2), s. 115–140, http://www.jstor.org/stable/25618620

Koseoglu Suzan, Bozkurt Aras (2018), *An exploratory literature review on open educational practices*, „Distance Education", vol. 39(4), s. 441–461, https://doi.org/10.1080/01587919.2018.1520042

Lasswell Harold D. (1927), *The Theory of Political Propaganda*, „The American Political Science Review", vol. 21(3), s. 627–631, https://doi.org/10.2307/1945515

Lewis Seth C., Zamith Rodrigo, Hermida Alfred (2013), *Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods*, „Journal of Broadcasting & Electronic Media", vol. 57(1), s. 34–52, https://doi.org/10.1080/08838151.2012.761702

Linares Julio, Cabaña Gabriela (2022), *Towards an ecology of care: basic income after the nation-state*, „Society Register", vol. 6(3), s. 29–56, https://doi.org/10.14746/sr.2022.6.3.03

Mayntz Renate, Holm Kurt, Hübner Peter (1976), *Wprowadzenie do metod socjologii empirycznej*, Warszawa: Państwowe Wydawnictwo Naukowe.

Midgley James (1997), *Social Welfare in Global Context*, London: Sage Publications.

Mohr John W., Bogdanov Petko (2013), *Introduction – Topic models: What they are and why they matter*, „Poetics", vol. 41(6), s. 545–569, https://doi.org/10.1016/j.poetic.2013.10.001

Naskar Debashis, Mokaddem Sidahmed, Rebollo Miguel, Onaindia Eva (2016), *Sentiment analysis in social networks through topic modeling*, [w:] Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož: European Language Resources Association, s. 46–53.

Nelson Laura (2020), *Computational Grounded Theory: A Methodological Framework*, „Sociological Methods & Research", vol. 49(1), s. 3–42, https://doi.org/10.1177/0049124117729703

Nesterova Iana (2023), *Responsibilities towards places in a degrowth society: How firms can become more responsible via embracing deep ecology*, „Society Register", vol. 7(1), s. 53–74, https://doi.org/10.14746/sr.2023.7.1.03

Pääkkönen Juho, Ylikoski Petri (2021), *Humanistic interpretation and machine learning*, „Synthese", vol. 199, s. 1461–1497, https://doi.org/10.1007/s11229-020-02806-w

Praag Bernard M.S. van (1989), *The Relativity of the Welfare Concept*, „World Institute for Development Research of the United Nations University, Working Paper", no. 69, s. 1–43.

R Core Team (2022), *_R: A Language and Environment for Statistical Computing_*, „R Foundation for Statistical Computing", Vienna, https://www.R-project.org/ [dostęp: 24.09.2024].

Roberts Margaret E., Stewart Brandon M., Tingley Dustin (2019), *stm: An R Package for Structural Topic Models*, „Journal of Statistical Software", vol. 91(2), s. 1–40, https://doi.org/10.18637/jss.v091.i02

Silge Julia, Robinson David (2017), *Text Mining with R: A Tidy Approach*, Sebastopol: O'Reilly.

Snyder Hannah (2019), *Literature review as a research methodology: An overview and guidelines*, „Journal of Business Research", vol. 104, s. 333–339, https://doi.org/10.1016/j.jbusres.2019.07.039

Syed Shaheen, Spruit Marco (2018), *Selecting Priors for Latent Dirichlet Allocation*, [w:] *IEEE 12th International Conference on Semantic Computing (ICSC)*, Laguna Hills: IEEE s. 194–202, https://doi.org/10.1109/ICSC.2018.00035

Thangaraj Muthuraman, Sivakami Muthusamy (2018), *Text Classification Techniques: A Literature Review*, „Interdisciplinary Journal of Information, Knowledge, and Management", vol. 13, s. 117–135, https://doi.org/10.28945/4066

Timms Noel (1980), *Social welfare: Why and how?*, London: Routledge.

Titmuss Richard M. (1967), *The Welfare Complex in a Changing Society*, „The Milbank Memorial Fund Quarterly", vol. 45(1), s. 9–23, https://doi.org/10.2307/3349045

Cytowanie

Piotr Cichocki, Mariusz Baranowski (2024), *Modelowanie tematyczne w socjologii na przykładzie dobrobytu społecznego: wyzwania metodologiczne i komponent ludzki*, „Przegląd Socjologii Jakościowej", t. XX, nr 4, s. 98–117, https://doi.org/10.18778/1733-8069.20.4.05

Topic Modeling in Sociology Using Social Welfare as an Example: Methodological Challenges and the Human Component

Abstract: Considering the dynamically evolving realms of social sciences influenced by network technologies and digital humanities, it is crucial to examine the adequacy of sociological data analysis methodologies in these new conditions. The availability of extensive digitized datasets not only poses a challenge to "classical" analysis methods developed under different circumstances and for different purposes, but also raises the question of whether the traditional demarcation between quantitative and qualitative methods, marked by a clear boundary, remains relevant in the era of Big Data. In this paper, based on topic modeling utilising Latent Dirichlet Allocation (LDA), we argue that quantitative methods (probabilistic statistical models) are not merely complementary or a starting point for qualitative analyses (the standard approach), but, rather, constitute an integral part of them. This thesis is illustrated through a case study involving the identification of themes within a dataset of 17,278 articles published in Web-of-Science-indexed journals between 1992 and 2020, focusing on social welfare. This empirical case study also serves to formulate meta-theoretical observations regarding the "cohesion" of quantitative and qualitative methods in the context of machine learning and natural language processing.

Keywords: topic modeling, methodology of sociology, social welfare, machine learning, Natural Language Processing