COPE
Member since 2019
JM14475

W stronę nowej metodologii analizy treści. Podobieństwa i różnice pomiędzy modelowaniem tematycznym i jakościową analizą treści

Sławomir Mandes
Uniwersytet Warszawski

Agnieszka Karlińska
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy

https://doi.org/10.18778/1733-8069.20.4.06

Słowa kluczowe: jakościowa analiza treści, analiza tematyczna, modelowanie tematyczne, *text mining*, *mixed methods*

Abstrakt: Celem artykułu jest krytyczna refleksja nad relacją pomiędzy jakościową analizą tematyczną i modelowaniem tematycznym (ang. *topic modeling*), jedną z bardziej popularnych odmian automatycznego przetwarzania tekstu. Na podstawie wyników jakościowej i ilościowej analizy dokumentów Konferencji Episkopatu Polski autorzy pokazują wady i zalety modelowania tematycznego. Negatywnie weryfikują tezę o zastępowalności analizy tematycznej przez modelowanie tematyczne i wskazują na niezbędność połączenia podejścia jakościowego z ilościowym w ramach metodologii metod mieszanych (ang. *mixed methods*). W ostatniej części opracowania przedstawiają możliwe sposoby łączenia obu metod, za pomocą których badacze jakościowi i badaczki jakościowe na podstawie paradygmatu metod mieszanych mogą skorzystać z modelowania tematycznego i – ze świadomością jego wad i zalet – wzbogacić swój warsztat, rozszerzyć zakres badań oraz usprawnić proces analizy.

Sławomir Mandes
Doktor habilitowany, profesor UW, socjolog. Pracuje na Wydziale Socjologii Uniwersytetu Warszawskiego. Interesuje się socjologią religii, metodologią badań społecznych, problematyką życia lokalnego.
e-mail: mandess@is.uw.edu.pl

Agnieszka Karlińska
Doktor, socjolożka, literaturoznawczyni/językoznawczyni i kulturoznawczyni zainteresowana obliczeniowymi naukami społecznymi, data-centric AI, humanistyką cyfrową, legal NLP i socjolingwistyką. W Państwowym Instytucie Badawczym NASK prowadzi badania z zakresu automatycznej analizy tekstu i kieruje projektem PLLuM, którego celem jest stworzenie otwartego polskiego dużego modelu językowego.
e-mail: agnieszka.karlinska@nask.pl

Wprowadzenie

Rozwój internetu i cyfryzacja różnych obszarów życia społecznego powodują, że osoby badające społeczeństwo mają do dyspozycji coraz więcej danych tekstowych. Warto przypomnieć, że proces ten rozpoczął się jeszcze przed eksplozją internetu od digitalizacji gazet, a następnie książek. Internet umożliwił szeroki dostęp do tych treści. Z kolei rozwój mediów społecznościowych spowodował, że w zasięgu badaczek i badaczy znalazł się nowy rodzaj tekstu: zapis prowadzonych za pomocą klawiatury komputera lub telefonu interakcji społecznych. Współcześnie obserwujemy zwiększanie się efektywności modeli językowych służących do przetwarzania mowy na tekst, co pozwala sądzić, że niedługo będzie możliwe analizowanie automatycznie transkrybowanych dużych zbiorów komunikacji mówionej.

Wraz z przyrostem cyfrowych danych tekstowych postępował rozwój metod ich przetwarzania i analizy. W szczególności opracowywano algorytmy przetwarzania języka naturalnego (NLP), rozpowszechniały się łatwe w obsłudze narzędzia do analizy tekstów (np. WordSmith, AntConc, SketchEngine, QDA Miner, WordStat) oraz względnie łatwe w obsłudze pakiety do analizy tekstów za pomocą języków programowania, takich jak Python i R. Co jednak ważne, rozwój ten – motywowany w dużym stopniu potrzebami firm zainteresowanych komercyjnym wykorzystaniem danych tekstowych – zawdzięczamy przede wszystkim informatyce oraz językoznawstwu korpusowemu i komputerowemu, którym digitalizacja otworzyła zupełnie nowe możliwości badań. Nauki społeczne dołączyły do tego trendu względnie późno (zob. np. Grimmer, Stewart, 2013; Macanovic, 2022). Świadczy o tym choćby znaczna dysproporcja pomiędzy liczbą podręczników do NLP lub językoznawstwa korpusowego a liczbą podręczników z zakresu automatycznej analizy tekstu, opracowanych na potrzeby nauk społecznych.

Samo zapóźnienie nie jest jednak problemem. Osoby badające społeczeństwo mogą wykorzystać metody NLP czy językoznawstwa korpusowego. Trudność leży w tym, że powstały one w odpowiedzi na pytania badawcze i w kontekście metodologii właściwych tym dyscyplinom, a nierzadko po prostu w odpowiedzi na konkretne zapotrzebowanie technologiczne. Chociaż mechaniczne zastosowanie

tych metod przyniesie wyniki – są one skuteczne i efektywne – bez pogłębionej refleksji metodologicznej ich interpretacja jest dyskusyjna.

Celem artykułu jest krytyczna ocena jednej z metod automatycznego przetwarzania tekstu i pokazanie, w jaki sposób badacze jakościowi i badaczki jakościowe mogą, ze świadomością jej wad i zalet, wzbogacić swój warsztat, rozszerzyć zakres badań i usprawnić proces analizy. Jako przykład wybraliśmy modelowanie tematyczne (ang. *topic modeling* – MT), metodę coraz chętniej stosowaną w naukach społecznych, która uważana jest za komplementarną wobec klasycznej już analizy tematycznej (AT) wypracowanej w ramach analizy treści (Isoaho, Gritsenko, Mäkelä, 2021). Zweryfikujemy tezę o substytucyjności tych dwóch metod i na podstawie dwukrotnej analizy tego samego materiału – z wykorzystaniem jakościowej analizy treści i algorytmów MT – zaproponujemy cztery podejścia pozwalające na możliwie efektywne połączenie automatycznej analizy tekstu z metodami dobrze już ugruntowanymi w naukach społecznych. Nacisk położymy na pracę z tekstami w języku polskim, który – ze względu na poziom złożoności samego języka i mniejszą dostępność zasobów do jego automatycznego przetwarzania (zob. Vetulani, Vetulani, 2020) – stwarza nieco inne wyzwania niż praca z tekstami w języku angielskim.

Metodologia dla obliczeniowych metod analizy tekstu

Automatyczne metody analizy tekstu obejmują wiele technik. Można wśród nich wyróżnić metody słownikowe, polegające na określeniu częstości słów i fraz związanych z danymi konceptami (na przykład mową nienawiści), metody semantycznej i sieciowej analizy tekstu, umożliwiające identyfikację w dokumentach wzmianek o aktorach społecznych i ich działaniach, metody bazujące na numerycznej wektorowej reprezentacji tekstu i osiągnięciach semantyki dystrybucyjnej, które pozwalają na przeprowadzenie pogłębionej analizy znaczenia, na przykład w celu wykrycia różnic w języku używanym przez różnych aktorów, metody nienadzorowanego grupowania, takie jak MT, które automatycznie dzielą słowa i dokumenty na grupy cechujące się podobieństwem, pomagając w ten sposób uchwycić prawidłowości trudne niekiedy do zaobserwowania „gołym okiem" oraz metody klasyfikacji oparte na nadzorowanym uczeniu maszynowym, które umożliwiają automatyczne rozszerzenie wyników ręcznego znakowania na dużą liczbę tekstów (Evans, Aceves, 2016; Macanovic, 2022; pełen przegląd metod znaleźć można w Grimmer, Roberts, Stewart, 2022).

W początkach rozwoju automatycznych technik analizy tekstu wydawało się, że uda się za ich pomocą rozwiązać wiele problemów trapiących badania treści. Przede wszystkim eliminacja osoby badacza lub badaczki i zastąpienie jej algorytmem zdawały się wychodzić naprzeciw zarzutowi o subiektywizm badań jakościowych (Goldthorpe, 2012). Automatyzacja dużej części procesu analizy skracała także czas i koszty badania. Nowe metody umożliwiły ponadto zwiększenie skali badań, która nie sprowadza się wyłącznie do wzrostu objętości analizowanego materiału, ale polega także na stawianiu szerzej zakrojonych pytań badawczych, w tym takich, na które nie udałoby się znaleźć odpowiedzi bez użycia nowych technologii (Shah, Cappella, Neuman, 2015; Underwood, 2019). Szybko

jednak ujawniły się problemy. Wskazuje się, że tego rodzaju analizy, chociaż efektowne, często nie przynoszą żadnych interesujących obserwacji i sprowadzają się do potwierdzenia znanych ustaleń (Brennan, 2017). Co więcej, zdaniem krytyków zwiększenie objętości analizowanych danych ma się przekładać na powierzchowność obserwacji i sprzyjać ignorowaniu szerszego kontekstu i społeczno-kulturowych uwarunkowań wykrytych zjawisk (zob. Brosz, Bryda, Siuda, 2017).

Problem jednak tkwi nie w samych metodach, lecz w tym, w jaki sposób i w kontekście jakiej metodologii są stosowane. W literaturze coraz częściej dostrzega się rozdźwięk pomiędzy rozwojem metod i narzędzi z obszaru automatycznej analizy tekstu a wykorzystaniem ich do odpowiedzi na pytania badawcze czy odkrywania nowych zjawisk w obszarze nauk społecznych. Jedną z głównych wskazywanych przyczyn jest fakt, że większość prac skoncentrowana jest na technicznych możliwościach różnych modeli i narzędzi, a pomija istniejące metody badania tekstu i ignoruje pytanie, jak te modele wpisują się w tradycyjne, a bardziej właściwe dla nauk społecznych metody analizy treści. Zauważono, że zastosowanie procedur wypracowanych w ramach jednej dziedziny do badań prowadzonych w innej dziedzinie bywa problematyczne, kwantyfikacja (ang. *quantitization*) danych (Collingridge, 2013) nie jest ani epistemologicznie naturalnym procesem, ani przepisem na obiektywność, a sposoby pracy z danymi, w szczególności metody walidacji modeli za pomocą prostych wskaźników statystycznych, które zostały zaczerpnięte z informatyki, niekoniecznie sprawdzą się w humanistyce i naukach społecznych (Nzabonimpa, 2018; Da, 2019).

Na potrzebę metodologicznego namysłu nad wykorzystaniem metod obliczeniowych do analizy tekstu i oceny ich wyników w kontekście nauk społecznych wskazuje wielu badaczy (np. Grimmer, Roberts, Stewart, 2022; McLevey, 2022). Można w tym kontekście postawić pytanie, czy potrzebna nam jest jakaś nowa metodologia badania tekstu ponad to, co już mamy (np. Krippendorff, 2018). Czy nie wystarczy sięgnąć do istniejących opracowań i podręczników poświęconych analizie treści i skorzystać z gotowych rozwiązań? Problem w tym, że metody automatycznej analizy treści są niedopasowane do metodologii wypracowanych i sprawdzonych w naukach społecznych. Zwróćmy uwagę tylko na dwie kluczowe kwestie. Po pierwsze, w podejściach obliczeniowych z góry zakłada się, że teksty nie są czytane przez osobę badającą. Czasem określa się to jako „czytanie na dystans" (ang. *distant reading*; zob. Moretti, 2016), co jest mylące, ponieważ w przypadku tych metod nie ma mowy o czytaniu w żadnym sensie. Po drugie, rezultatem analizy są liczby, które kwantyfikują – w bardzo różny sposób, w zależności od przyjętej metody – wyodrębnione z tekstu znaczenia, inaczej więc niż w metodologii analizy tekstu opartej w większości przypadków na teorii interpretacji i sensu.

Odpowiedzią na te wyzwania i problemy jest paradygmat (Onwuegbuzie, Johnson, Collins, 2009) kwestionujący epistemologiczną trafność binarnego podziału na metody jakościowe oraz ilościowe i wzywający do pragmatycznego podejścia poprzez rozwój metodologii mieszanej (ang. *mixed methods*) (Teddlie, Tashakkori, 2009). Wspomagana komputerowo analiza treści jest przykładem praktycznej realizacji takiego podejścia (Bazeley, 2010). Z tego punktu widzenia wykorzystanie automatycznych metod eksploracji tekstu jest następnym naturalnym krokiem. Gregor Wiedemann przekonuje, że metody te otwierają w naukach społecznych nowe możliwości wydobywania znaczenia (jawnego

i ukrytego) poprzez wyszukiwanie globalnych wzorców, zliczanie wystąpień i obliczanie ukrytych zmiennych wskazujących na różne aspekty semantyki, co sprawia, że dystans pomiędzy jakościową i ilościową analizą tekstu – „pomiędzy tym, jak badacze jakościowi postrzegają swój obiekt badań, a tym, co są w stanie zidentyfikować algorytmy komputerowe" – stale się zmniejsza (Wiedemann, 2013: 7). W rezultacie automatyczne metody analizy tekstu mogą stanowić pomost pomiędzy projektami badań jakościowych i ilościowych.

Biorąc pod uwagę zróżnicowanie jakościowych i automatycznych metod analizy treści, trudno przesądzić, czy założenia metodologii mieszanej są możliwe do przełożenia na praktykę badawczą. Dlatego, aby zejść z poziomu ogólnych rozważań i deklaracji, chcemy w pierwszej kolejności zbadać, czy MT – jedna z najbardziej popularnych obecnie technik automatycznej eksploracji tekstu – może zastąpić klasyczną metodę AT. Następnie pokażemy, w jaki sposób można połączyć MT z AT z pożytkiem dla badań jakościowych. Wyniki przeprowadzonej analizy posłużą nam do tego, aby sformułować konkretne zalecenia dla podejścia mieszanego, w którym wykorzystane będą obie metody.

Modelowanie tematyczne to zestaw metod obliczeniowych pozwalających na automatyczną identyfikację tematów w zbiorze tekstów (Blei, 2012; DiMaggio, Nag, Blei, 2013). Jest ono blisko spokrewnione z klasyczną analizą skupień, ale przyjmuje inne założenie dotyczące liczby kategorii, do których może być przypisany dany dokument. Zgodnie z podstawową intuicją, że teksty zawierają treści odnoszące się do wielu tematów, zamiast przypisywać każdy dokument do jednej grupy, w MT poszczególnym dokumentom przypisane jest prawdopodobieństwo przynależności do każdego z wyróżnionych tematów (Grimmer, Roberts, Stewart, 2022). Modelowanie tematyczne stosowano do analizy zróżnicowanego materiału, od tekstów prasowych, w tym dużych zbiorów publikacji historycznych, i artykułów naukowych, przez programy i komunikaty prasowe partii politycznych oraz przemówienia polityków, po dane z mediów społecznościowych czy korpusy literatury pięknej (przegląd zastosowań MT znajduje się w Boyd-Graber, Hu, Mimno, 2017). Umożliwiały one szerokie spojrzenie na zagadnienia podejmowane w naukach społecznych od lat, takie jak społeczna mobilizacja i rozwój ruchów społecznych (np. Perrin i in., 2014), dyskurs polityczny (np. Wiedemann, 2016), rozwój pól badawczych w ramach poszczególnych dyscyplin (np. Mann, Mimno, McCallum, 2006) i wiele więcej (zob. Chen i in., 2023). Na gruncie polskiej humanistyki i nauk społecznych nie prowadzono do tej pory wielu badań z wykorzystaniem MT. Jednym z nielicznych przykładów jest praca Agnieszki Kwiatkowskiej (2017) dotycząca skrajnej retoryki politycznej w debacie parlamentarnej. Autorka zastosowała algorytmy LDA oraz STM, uwzględniając wybrane metadane dokumentów (datę wypowiedzi i afiliację polityczną mówcy). Analogiczne podejście wykorzystała w analizie wystąpień parlamentarnych, w których mowa była o Stanach Zjednoczonych, prowadzonej wspólnie z Karolem Chwedczukiem-Szulcem i Bartoszem Bolechowem (2022). Mariusz Baranowski i Piotr Cichocki (2021) zastosowali LDA do metaanalizy anglojęzycznych artykułów dotyczących MT indeksowanych w bazie Scopus. Stwierdzili, że metoda ta sprawdza się przede wszystkim w badaniach eksploracyjnych, których celem jest mapowanie dyskursu i uzyskanie szybkiego wglądu w nieznane wcześniej teksty. Różni badacze i badaczki próbowali połączyć LDA z innymi cyfrowymi metodami analizy danych. Grzegorz Bryda (2020) wykorzystał MT, hierarchiczne grupowanie oraz metodę słownikową

do identyfikacji typów analiz narracyjnych w artykułach opublikowanych w anglojęzycznych czasopismach z zakresu metodologii badań jakościowych. Maciej Maryl i Maciej Eder (2017) połączyli LDA z analizą sieciową na potrzeby rekonstrukcji przemian w obrębie polskiego literaturoznawstwa na przykładzie artykułów opublikowanych w „Tekstach Drugich" w latach 1990–2014. Adam Pawłowski i Tomasz Walkowiak (2022) przeprowadzili analizę treści autopromocyjnych znajdujących się na stronach internetowych 391 uczelni wyższych w Polsce, z wykorzystaniem metod stylometrycznych, automatycznej taksonomii i MT. Z kolei Agnieszka Hess i Krzysztof Hwaszcz (2022) w eksploracyjnych badaniach dyskursu parlamentarnego pod kątem instytucjonalizacji dialogu obywatelskiego oraz w porównawczej analizie dyskursu rady miasta i dyskursu mediów w Krakowie zastosowali narzędzia do NLP dostępne w ramach infrastruktury naukowo-badawczej CLARIN-PL – obok MT wykorzystali również analizę terminologii dziedzinowej oraz analizę wydźwięku.

Polscy badacze i polskie badaczki skupiali się zwykle na pokazaniu zalet MT w analizach socjologicznych, politologicznych czy medioznawczych, przekonując, że może ono stanowić „szybszą i bardziej rzetelną alternatywę wobec słownikowych metod kategoryzowania danych tekstowych lub też służyć do weryfikacji kategorii uzyskanych w sposób tradycyjny" (Kwiatkowska, 2017: 106). Nie podejmowali natomiast kwestii ograniczeń tej metody i możliwości czy też, jak twierdzimy w tym tekście, niezbędności jej połączenia z metodami jakościowymi[1]. Przywołane prace były pozbawione głębszego namysłu metodologicznego – ich autorzy i autorki albo wykorzystywali MT jako samodzielną metodę, uznając ją za zamiennik analizy treści, albo stosowali MT obok innych metod cyfrowych, ale bez spójnej, jasno wyłożonej metodologii. Na wyzwania związane z implementacją MT w naukach społecznych jako jedyni zwrócili uwagę Baranowski i Cichocki (2021). Ich wnioski ograniczyły się jednak do kwestii odpowiedniego przetworzenia tekstów, w szczególności opracowania stop-listy[2] dostosowanej do analizowanego typu dokumentów. Co ważne, badania prowadzone były często na danych anglojęzycznych. Jeśli wykorzystywano teksty w języku polskim, nie podejmowano szerszej refleksji na temat dostosowania metod i narzędzi do specyfiki języka polskiego.

Materiał i metody

Dokumenty Konferencji Episkopatu Polski (KEP) są analizowane przez badaczy i badaczki z różnych dyscyplin nauk społecznych. Były między innymi przedmiotem analizy językoznawczej w pracach Katarzyny Skowronek (2006; 2007), badań prowadzonych z perspektywy analizy dyskursu (Leszczyńska, Zych, 2011; Szwed, 2018; 2019) oraz krytycznej analizy dyskursu (Kamasa, 2013a; 2013b). Były również przedmiotem systematycznej analizy prowadzonej z punktu widzenia problemów właściwych dla politologii (Delong, 2016; 2017a; 2017b). W większości wypadków analiza miała charakter

1 Nie jest to specyfika polskich nauk społecznych. Na brak krytycznej refleksji nad projektem badawczym i metodologicznymi wymiarami MT w pracach wykorzystujących metody automatycznej analizy tematycznej zwrócili uwagę Karoliina Isoaho, Daria Gritsenko i Eetu Mäkelä (2021).

2 Stop-lista to słownik zawierający wyrazy, które mają zostać wyłączone z analizy. Zazwyczaj znajdują się w nim słowa funkcyjne, liczebniki i często używane przysłówki.

jakościowy, jedynie Victoria Kamasa połączyła metody językoznawczej analizy korpusowej (frekwencje i konkordancje wybranych terminów) z analizą dyskursu. Nie jest to wyczerpująca lista – w wielu pracach analizie poddawane są poszczególne dokumenty w kontekście innych materiałów – dobrze jednak ilustruje fakt, że tekst może być analizowany w różnych dyscyplinach z użyciem różnych metodologii i metod.

Zebrana już wiedza na temat dokumentów KEP czyni z nich bardzo dobry materiał do badania komplementarności metod cyfrowych (MT) i klasycznej analizy treści (AT). Opracowany przez nas korpus dokumentów KEP składa się ze 181 tekstów przekazanych przez KEP do odczytania w kościołach w Polsce w okresie od stycznia 1990 do grudnia 2005 roku. Średnia długość dokumentu wynosi 1175 tokenów[3], najdłuższy ma 6994, a najkrótszy 95 tokenów. W zebranym korpusie można wyróżnić trzy typy tekstów:

1) listy pasterskie (inne tytuły: „słowo biskupów", „słowo pasterskie") – dokumenty o długości od 900 do 1300 tokenów, pisane z intencją odczytania w kościołach w czasie przeznaczonym na kazanie, poświęcone bardzo zróżnicowanej tematyce (np. środki masowego przekazu, wizyty Jana Pawła II w Polsce, beatyfikacja) i charakteryzujące się tym, że w obrębie jednego tekstu dane zagadnienie jest zazwyczaj omawiane z różnych perspektyw;
2) odezwy, stanowiska, zalecenia, memoriały – dokumenty krótsze, do 1000 tokenów, będące komunikatem do opinii publicznej, mające konkretny, zwykle jasno zdefiniowany w tytule temat (np. podział administracyjny w Kościele katolickim, wybory) i charakteryzujące się dużą spójnością tematyczną;
3) komunikaty – dokumenty krótsze, do 1000 tokenów, zawierające informacje o najważniejszych sprawach omawianych w czasie posiedzeń KEP.

W pierwszej kolejności dokumenty KEP poddaliśmy AT (Braun, Clarke, 2006; 2021). Jest to metoda wypracowana w ramach jakościowej analizy treści. Spośród innych metod analizy treści wyróżnia ją to, że koncentruje się na identyfikacji wzorców lub tematów, które są obecne w danych, niekoniecznie definiując te tematy wcześniej. Dąży do uzyskania wglądu w subiektywne doświadczenia, przekonania i perspektywy jednostek lub grup, które są osobami autorskimi tekstu. Zazwyczaj wiąże się z bardziej indukcyjnym podejściem, gdzie badacz lub badaczka pozwala tematom wyłonić się z danych, ale nie wyklucza kodowania danych do kategorii lub podtematów na podstawie zidentyfikowanych wzorców. W przypadku analizy dokumentów KEP zastosowaliśmy podejście indukcyjne, kodowanie otwarte, bez wstępnych założeń w odniesieniu do treści. Przyjęto dwie jednostki kodowania:

1) cały dokument (dalej: tematy z poziomu pierwszego);
2) fragmenty dłuższe, od 1 do 4 akapitów (dalej: tematy szczegółowe z poziomu drugiego).

3 Token (segment) to podstawowy element tekstu, który zdefiniować można jako ciąg znaków interpretowany pod względem fleksyjnym, niezawierający odstępów i – zwykle – znaków interpunkcyjnych. W tym ujęciu token jest w większości przypadków tożsamy ze słowem tekstowym. Istnieją jednak odstępstwa od tej reguły (zob. Woliński, 2019).

Po zakodowaniu całego materiału zweryfikowaliśmy spójność pomiędzy kodami dotyczącymi całych dokumentów i kodami dotyczącymi fragmentów, tak aby kod dotyczący całych dokumentów odzwierciedlał dominujące tematy – zakodowane za pomocą kodów dotyczących fragmentów. Kodowanie zostało przeprowadzone z wykorzystaniem programu Atlas.ti (Friese, 2019).

Po przeprowadzeniu analizy jakościowej dokumenty KEP poddaliśmy MT. W analizie zastosowaliśmy popularny algorytm LDA (ang. *Latent Dirichlet Allocation*) wprowadzony w 2001 roku (Blei, Ng, Jordan, 2003)[4]. Jest on modelem probabilistycznym wykorzystującym dwie wartości prawdopodobieństwa: występowanie słów w tematach i występowanie tematów w dokumentach. W pewnym uproszczeniu proces analizy polega na tym, że osoba badająca na początku określa liczbę tematów[5], algorytm w procesie wielu powtórzeń (iteracji) tworzy model, w którym słowa są przypisywane do tematów, tematy zaś do dokumentów, a następnie osoba badająca sprawdza, czy i w jakim stopniu pasują one do analizowanego korpusu. Jako wynik otrzymuje zestawy słów (określane niekiedy jako „słowozbiory") wykazujących tendencję do współwystępowania ze sobą w pewnych dokumentach i słabo reprezentowane w innych – tendencja ta interpretowana jest jako powiązanie tematyczne (Eder, 2016) – oraz informacje o prawdopodobieństwie wystąpienia tych zestawów w poszczególnych tekstach (Blei, 2012). Proces ten powtarza się – zmieniając liczbę tematów oraz inne parametry – aż do uzyskania wyników zadowalających badacza czy badaczkę z punktu widzenia celów badania i jego/jej rozumienia tekstu.

Modelowanie tematyczne wykonaliśmy z użyciem biblioteki Gensim w języku Python oraz implementacji MT wykorzystującej próbkowanie Gibbsa (pakiet MALLET – *Mallet*…, b.r.). Przed przystąpieniem do analizy konieczne było odpowiednie przygotowanie danych. W pierwszej kolejności w sposób półautomatyczny połączyliśmy wyrazy podzielone przy przenoszeniu do kolejnego wiersza. Z powodu dysproporcji pod względem długości dokumenty najdłuższe (liczące powyżej 2000 tokenów) podzieliliśmy na fragmenty o objętości około 1500 tokenów (z zachowaniem granic akapitów). W ten sposób uzyskaliśmy łącznie 197 tekstów o średniej długości 1061 tokenów. W kolejnym etapie usunęliśmy z korpusu cyfry i znaki interpunkcyjne. Następnie przeprowadziliśmy tokenizację, oznaczanie części mowy i lematyzację (sprowadzanie słów do ich form podstawowych) przy użyciu biblioteki spaCy[6] (Honnibal i in., 2020). Do właściwej analizy zdecydowaliśmy się włączyć rzeczowniki, przymiotniki, czasowniki oraz nazwy własne jako części mowy niosące najwięcej znaczenia.

4 Algorytm ten był w kolejnych latach modyfikowany i rozszerzany. Warto tutaj wymienić Correlated Topic Model (CTM), który umożliwia uchwycenie związków między tematami (Blei, Lafferty, 2006), oraz Structural Topic Model (STM), pozwalający na włączenie do analizy informacji o metadanych (Roberts i in., 2014). Obecnie popularność zyskują nowe podejścia do MT, wykorzystujące tzw. wektory osadzeń dokumentów (ang. *embeddings*), generowane na podstawie wstępnie wytrenowanych modeli języka w technologii Transformer, takie jak BERTopic (Grootendorst, 2022).

5 Określenie odpowiedniej liczby tematów jest przedmiotem wielu kontrowersji i debat (Hoyle i in., 2021). Do tej kwestii wrócimy we wnioskach, tutaj warto tylko podkreślić, że nie ma jednoznacznych i obiektywnych metod określenia tej liczby i nie przypadkiem mówi się w tym kontekście o „wróżeniu z fusów" (Chang i in., 2009).

6 Zastosowaliśmy nowy model pl_nask-0.0.7, opracowany przez IPI PAN dla NASK PIB – Index of /~rtuora/spacy, b.r.

Analiza miała charakter iteracyjny. Przed wyborem optymalnego modelu przeprowadziliśmy wiele eksperymentów, w których modyfikowaliśmy poszczególne parametry (m.in. usuwanie najczęstszych i najrzadszych lematów, alpha, eta). Do wstępnego określenia liczby tematów skorzystaliśmy z miar spójności semantycznej i wyłączności, jednak ze świadomością, że bywają one zawodne (zob. Shadrova, 2021). Ostateczną decyzję podjęliśmy na podstawie jakościowej oceny uzyskanych wyników. Uwzględniliśmy w niej 20 słów o najwyższym prawdopodobieństwie przynależności do każdego z wyróżnionych tematów. Dodatkowo przyjrzeliśmy się dokumentom najbardziej reprezentatywnym dla danego tematu. W ten sposób wybraliśmy model cechujący się najwyższą interpretowalnością wyników[7].

Wyniki

Dokumenty KEP w świetle analizy tematycznej

Przedstawione poniżej wyniki nie stanowią wyczerpującej jakościowej analizy zebranych danych. Koncentrujemy się na poziomie tematów dla całego dokumentu, ponieważ dokument jest jednocześnie jednostką analityczną w MT. W zebranym materiale zidentyfikowaliśmy 94 tematy kodowane na poziomie całego dokumentu oraz 294 tematy kodowane na poziomie dłuższych fragmentów dokumentu. Biorąc pod uwagę, że przedmiotem analizy było 181 tekstów, już na pierwszy rzut oka widać, że tematy główne bardzo rzadko się powtarzały. Do wyjątków należały listy cykliczne – opracowywane co roku – w tym listy na uroczystości Świętej Rodziny (12 wystąpień) i dotyczące trzeźwości (6) oraz listy poświęcone sprawom ważnym w danym okresie: kwestii aborcji i ochrony życia (6) czy obchodom jubileuszu roku 2000 (5). Niska liczebność użycia kodów wynikała z otwartego sposobu kodowania, w którym nie brano pod uwagę bliskości niektórych tematów głównych. Na przykład temat wizyt Jana Pawła II w Polsce był omawiany w sumie w 7 listach, które zostały zakodowane oddzielnie z użyciem kodów „wizyta Jana Pawła II: znaczenie dla Polski i Kościoła", „wizyta Jana Pawła II: przed wizytą", „wizyta Jana Pawła II: po wizycie". Innym przykładem są dokumenty dotyczące wyborów, których w korpusie jest w sumie 10, a które były kodowane na trzy różne sposoby („wybory do parlamentu", „wybory do parlamentu: po wyborach" oraz „wybory prezydenta"). Oddzielną kategorię stanowią teksty oznaczone jako „komunikaty z posiedzeń KEP", których w korpusie znalazło się 21. Są one wyjątkiem od ogólnej reguły, ponieważ ich zawartość jest bardzo zróżnicowana. Wynika to z faktu, że podczas obrad poruszano różnorodną tematykę i treść komunikatu w punktach o tym informuje. Na przykład komunikat z 321. zebrania KEP: 1) informuje o początku

7 Osoby niemające kompetencji programistycznych mogą z powodzeniem wykonać podobną analizę, korzystając z narzędzi udostępnianych w ramach infrastruktury CLARIN-PL. Do MT służą obecnie usługi Topic – implementacja LDA (MALLET) i rzadko stosowanej metody BigARTM (CLARIN-PL, b.r., *Topiki*) – oraz shortextopic – implementacja techniki BERTopic (CLARIN-PL, b.r., *Shortextopic*). Na potrzeby badania przeprowadziliśmy analizę dokumentów KEP z wykorzystaniem narzędzia Topic i – po modyfikacji parametrów zgodnie z przyjętymi przez nas założeniami – otrzymaliśmy wyniki zbliżone do tych, które uzyskaliśmy z zastosowaniem narzędzi programistycznych. Zakres manipulacji parametrami w tym i podobnych narzędziach jest jednak ograniczony. Biblioteki do analizy tekstu i NLP w językach Python i R oferują więcej możliwości i pełną dowolność pod względem wstępnego przetwarzania tekstów, w związku z czym pozwalają na prowadzenie bardziej zaawansowanych analiz, dostosowanych do specyfiki konkretnego korpusu.

postu i związanych z tym działaniach Kościoła oraz zaleceniach dla wiernych, 2) daje wyraz troski o ubogich, 3) daje wyraz troski o życie publiczne i dotyczy potępienia korupcji, 4) odnosi się do zbliżającego się referendum akcesyjnego do UE, 5) informuje o planowanej ogólnopolskiej pielgrzymce do Rzymu. A wszystko to jest wyrażone zaledwie w 822 tokenach.

W celu bardziej syntetycznej reprezentacji uzyskanych wyników tematy z poziomu całego dokumentu zostały poddane dodatkowemu znakowaniu. Ponownie przyjęto założenie o otwartym kodowaniu, co doprowadziło do wypracowania trzech ogólnych grup tematów: religijnych, społecznych i politycznych. W tabeli 1 umieściliśmy przykładowe kategoryzacje tematów do poszczególnych grup na przestrzeni trzech wybranych lat.

Tabela 1. Kategoryzacja tematów całych dokumentów z lat 1991, 1995, 2004

Rok	Tematy społeczne	Tematy religijne	Tematy polityczne
1991	Sugestie różne na okres postu, Wezwanie do trzeźwości, Na niedzielę środków masowego przekazu, O miłości małżeńskiej i prawie do życia, Zalecenia w sprawie pomocy szkole	List przed pielgrzymką Papieża, List przed dniem młodzieży na Jasnej Górze, List przed kanonizacją bł. R. Kalinowskiego	W sprawie referendum na temat życia nienarodzonych, List w sprawie katechizacji, O zadaniach katolików wobec wyborów parlamentarnych, Stanowisko Episkopatu wobec wyborów parlamentarnych
1995	O roli katolików w procesie przekształceń na wsi i w małych miastach, O walorach turystyki, O Akcji Katolickiej, Wezwanie do trzeźwości, Na dzień Świętej Rodziny, Wspólne słowo biskupów niemieckich i polskich z okazji 30. rocznicy wymiany listów	Przekazanie katechizmu Kościoła katolickiego, O pielgrzymującej figurze Matki Boskiej Fatimskiej, Słowo przed jubileuszem 2000	O wyborach prezydenckich, O sprawach publicznych, O potrzebie dialogu i tolerancji w warunkach budowy demokracji
2004	W sprawie zarodków komórek macierzystych, Komunikat 327-9, Komunikat Biskupów Diecezjalnych, O sporcie, Apel o trzeźwość	O ślubach czystości, Słowo z okazji Dnia Papieskiego, List na rok Eucharystii	W sprawie rejestracji związków partnerskich, Słowo z okazji przyjęcia Polski do UE, Apel o odpowiedzialność za przyszłość Polski i Europy, Oświadczenie odnośnie ustawy o pomocy społecznej, Oświadczenie w sprawie Konstytucji UE (2 dokumenty), List w sprawie ZUS, Oświadczenie w sprawie ustawy o świadomym rodzicielstwie, W sprawie Funduszu Kościelnego

Źródło: opracowanie własne.

Przypisanie tematów (dokumentów) z poziomu pierwszego do grupy tematów religijnych nie stwarzało dużych trudności. Są to dokumenty o jednoznacznie określonym temacie, czasem religijnym w wąskim tego słowa znaczeniu (O ślubach czystości), najczęściej dotyczące uroczystości religijnych (List przed kanonizacją bł. R. Kalinowskiego) lub innych wydarzeń ogólnokościelnych (Słowo z okazji Dnia Papieskiego). Problematyczne było niekiedy przypisanie tematu albo do grupy tematów politycznych, albo do grupy tematów społecznych. O ile bowiem wątpliwości nie budziła kategoryzacja dokumentów poświęconych wyborom parlamentarnym lub prezydenckim, to już na przykład List w sprawie katechizacji może pasować i do grupy tematów społecznych, i do grupy tematów politycznych. Kryterium pozwalającym na dokonanie rozstrzygnięć w takich przypadkach było to, czy dokument dotyczy sprawy, która jest w okresie jego publikacji przedmiotem dyskusji w sferze publicznej, i czy jest to kwestia podlegająca decyzji politycznej. Zasada ta sprawdziła się w większości przypadków. Wyjątkiem był zbiór komunikatów z zebrań plenarnych KEP liczący 21 dokumentów. Jak już powiedziano, komunikaty były formułowane w postaci punktów poruszających bardzo różne sprawy, które albo były przedmiotem dyskusji w czasie posiedzenia, albo były na tyle ważne dla biskupów, aby wywołać ich reakcję. W związku z tym komunikatów KEP nie przypisaliśmy do żadnej z grup i w zestawieniu sumarycznym (tab. 2) prezentujemy je oddzielnie.

Tabela 2. Zestawienie sumaryczne grup tematów

Grupa ogólna	Typowe tematy	Liczba dokumentów	Liczba tokenów
Tematy społeczne	Rodzina, trzeźwość, media, sport, turystyka, praca, wolontariat, bezrobocie	59	81 423
Tematy religijne	Kanonizacja i beatyfikacja, prawdy wiary, na dzień: papieski, Świętej Rodziny, życie konsekrowane, uroczystości religijne, pielgrzymki Jana Pawła II	58	74 833
Tematy polityczne	Aborcja, katecheza, Konstytucja RP i UE, wybory prezydenckie i parlamentarne, związki partnerskie	22	51 339
Komunikaty KEP		21	15 516

Źródło: opracowanie własne.

W zestawieniu tym widać, że KEP poświęcała równie dużo uwagi tematom społecznym co religijnym. Liczba dokumentów w obu tych grupach jest niemal identyczna. Średnio na dokument poświęcony tematom społecznym przypada 1380 tokenów, a religijnym 1290. Różnica jest zatem nieznaczna. Dokumentów poświęconych tematom politycznym było mniej, ale były to teksty dłuższe (średnia liczba tokenów wynosiła 2334).

W omówieniu wyników analizy tematycznej skupiliśmy się na sumarycznej prezentacji wyróżnionych tematów. Powinniśmy jednak pamiętać, że liczby reprezentujące poszczególne grupy tematów są wynikiem dwuetapowego, jakościowego procesu kodowania. Po pierwsze, koder czytał wnikliwie dokument

KEP i dopiero na tej podstawie określał jego ogólny temat. Po drugie, przypisanie zidentyfikowanych w ten sposób tematów do grup również opierało się na pogłębionej lekturze dokumentu w kontekście całego korpusu. Warto podkreślić ten ostatni aspekt. Podstawą do wyodrębnienia dokumentów religijnych, społecznych i politycznych oraz osobnego potraktowania komunikatów KEP była znajomość wszystkich tekstów w korpusie, a kategoryzacja pojedynczego tekstu prowadzona była jednocześnie na podstawie lektury jego zawartości, jak również wiedzy, jak wypada on na tle pozostałych dokumentów. W obu wypadkach mamy zatem do czynienia z interpretacją znaczeń tekstów upublicznionych przez KEP.

Dokumenty KEP w świetle MT

Efektem analizy z użyciem algorytmów MT są zbiory słów wraz z informacją o prawdopodobieństwie ich wystąpienia w poszczególnych dokumentach. Wybrany przez nas model obejmuje 25 takich słowozbiorów. Każdy z nich – zgodnie z założeniami MT – reprezentuje temat obecny w dokumentach KEP. W tabeli 3 przedstawiliśmy po 15 słów o najwyższym prawdopodobieństwie przynależności do danego tematu.

Tabela 3. Zbiory słów (tematy) wyróżnione w toku modelowania tematycznego

	Temat 1	Temat 2	Temat 3	Temat 4	Temat 5
1	kościół	kongres	środek	prawo	święty
2	katolicki	wolność	przekaz	kościół	ojciec
3	nowy	Wojciech	społeczny	państwo	życie
4	ewangelizacja	eucharystyczny	masowy	sprawa	boży
5	misja	kraj	medium	polski	dzień
6	świecki	Polak	telewizja	konstytucja	rok
7	dzieło	Wrocław	prasa	mieć	wielki
8	misyjny	międzynarodowy	kultura	Polska	brat
9	wiara	Andrzej	dziedzina	osoba	czas
10	katechizm	wiek	informacja	religijny	nadzieja
11	nauka	patron	Polska	episkopat	siostra
12	ewangelia	śmierć	radio	ustawa	serce
13	watykański	Warszawa	odpowiedzialność	konkordat	cały
14	głosić	Bobola	program	prawny	świat
15	synod	męczeński	wierny	ochrona	miłość

	Temat 6	Temat 7	Temat 8	Temat 9	Temat 10
1	dobro	pokój	trzeźwość	Chrystus	matka
2	naród	dialog	alkohol	bóg	naród
3	ojczyzna	naród	alkoholowy	kościół	Maryja
4	państwo	pojednanie	nadużywać	Jezus	góra
5	polityczny	wojna	abstynencja	miłość	polski
6	Polska	przebaczenie	zagrożenie	słowo	Polska
7	społeczny	chrześcijanin	sierpień	duch	jasny
8	polski	Polak	naród	świat	bóg
9	społeczeństwo	wspólny	młodzież	boży	rok
10	system	wzajemny	środowisko	eucharystia	wielki
11	odpowiedzialność	sprawiedliwość	alkoholizm	nowy	boży
12	wspólny	wolność	choroba	wspólnota	modlitwa
13	stanowić	lud	trzeźwy	cały	sierpień
14	zasada	ziemia	miesiąc	ewangelia	jasnogórski
15	życie	Żyd	grupa	łaska	wolność
	Temat 11	**Temat 12**	**Temat 13**	**Temat 14**	**Temat 15**
1	społeczny	praca	rodzina	kościół	człowiek
2	nowy	bezrobocie	szkoła	jedność	móc
3	człowiek	społeczny	rodzic	unia	mieć
4	powinien	gospodarczy	dziecko	Rzym	sam
5	kościół	gospodarka	katecheza	diecezja	być
6	sytuacja	zmiana	pokolenie	chrześcijanin	bóg
7	problem	prawo	rok	prawosławny	inny
8	potrzeba	bezrobotny	wychowanie	stanowić	prawda
9	moralny	wieś	rodzinny	rzeczpospolita	Paweł
10	rzecz	pracownik	szkolny	jeden	Jan
11	troska	trzeba	mieć	greckokatolicki	ludzki

12	działanie	produkcja	nauczyciel	metropolia	musieć
13	należeć	własność	młody	Ukraina	trzeba
14	szczególny	problem	troska	kościelny	można
15	rodzina	zawodowy	młodzież	obrządek	życie
	Temat 16	**Temat 17**	**Temat 18**	**Temat 19**	**Temat 20**
1	wartość	Polska	rok	życie	rok
2	ludzki	biskup	Maksymilian	dziecko	jubileusz
3	życie	episkopat	Kolbe	miłość	święty
4	prawo	dzień	ojciec	kobieta	kościół
5	chrześcijański	Jan	Rafał	począć	tysiąclecie
6	osoba	polski	Marcelina	rodzina	Paweł
7	kultura	konferencja	rodzina	matka	ojciec
8	zasada	Paweł	Darowska	zabijać	Jan
9	podstawowy	kościół	męczennik	małżeństwo	wiara
10	godność	wierny	błogosławiony	prawo	Polska
11	stanowić	plenarny	religijny	obrona	pielgrzymka
12	postawa	rok	unita	małżeński	czas
13	ewangelia	zebranie	Rajmund	ludzki	wielki
14	moralny	modlitwa	święty	mężczyzna	dzieje
15	prawda	kardynał	syn	narodzić	papież
	Temat 21	**Temat 22**	**Temat 23**	**Temat 24**	**Temat 25**
1	święto	sumienie	życie	Europa	sport
2	niedziela	ludzki	konsekrować	europejski	ciało
3	praca	komórka	osoba	jedność	zdrowie
4	pański	macierzysty	czystość	duchowy	fizyczny
5	uroczystość	embrion	ubóstwo	kultura	turystyka
6	mieć	należeć	bóg	Polska	odpoczynek
7	misjonarz	cel	instytut	naród	wakacje

8	post	prawo	rada	unia	sportowy
9	nakazać	pierwszy	powołanie	wartość	sportowiec
10	świętować	zarodkowy	świadectwo	kontynent	zwracać
11	liturgiczny	formować	ślub	wspólnota	zagrożenie
12	rok	moment	ewangeliczny	biskup	seksualność
13	przyjąć	osoba	forma	tożsamość	obowiązek
14	dzień	istota	świecki	chrześcijański	świadomość
15	obchodzić	wykorzystywać	materialny	nadzieja	dobry

Źródło: opracowanie własne.

Poszczególne tematy przypisaliśmy do trzech kategorii wyróżnionych wcześniej w analizie jakościowej: religijnej, społecznej i politycznej. Opisy poszczególnych tematów są hasłowe – poprzez wskazanie ramy spajającej dany zbiór słów – i tylko na przykładzie tematu 1 zademonstrujemy bardziej rozbudowaną interpretację i sposób jej przeprowadzenia. Spośród słów składających się na ten temat szczególną uwagę zwracają – jako specyficzne, czyli niewystępujące w innych tematach – „nowy" i „ewangelizacja", co odsyła do projektu „nowej ewangelizacji", czyli programu pontyfikatu Jana Pawła II, którego celem było dostosowanie sposobów promowania wiary katolickiej do nowoczesnego społeczeństwa. O metodach „nowej ewangelizacji" dyskutowano na synodach. Ten temat poruszano również w kontekście integracji UE, która w latach dziewięćdziesiątych ubiegłego wieku była postrzegana przez przedstawicieli Kościoła katolickiego zarazem jako wyzwanie i szansa dla wzmocnienia religijności (list z 1992 r. *O nowej ewangelizacji przez synod plenarny*). Wypowiedzi dotyczące ewangelizacji UE współwystępowały z wątkiem misji, którą Kościół miał podjąć ze zdwojonym wysiłkiem na terenie Europy Zachodniej i – szerzej – całego świata (list z 1999 r. *Misje odnawiają Kościół*). Przyjmując, że projekt ewangelizacji wyznacza główną oś interpretacyjną tematu 1, to słowa – pamiętajmy, że mamy do czynienia z lematami, czyli formami podstawowymi wyrazów – takie jak „wiara", „nauka", „głoszenie" wyznaczają oś pomocniczą w interpretacji tematu, wskazując na działanie – głoszenie nauki lub wiary, która wynika z projektu nowej ewangelizacji – oraz całość zadania – „dzieło", które ma podjąć i zrealizować Kościół katolicki wraz z wiernymi (świeckimi). Należy też zwrócić uwagę, że nie wszystkie słowa dają się zinterpretować na wymiarze osi głównej lub pomocniczej. Słowem odstającym jest „Watykan". W tym wypadku nie chodzi o siedzibę papieża, lecz o Sobór Watykański II, o którym mowa jest w tekstach dotyczących nowej ewangelizacji. Słowo to w dokumentach KEP występuje dość rzadko, co spowodowało, że algorytm je wychwycił i uwzględnił, ale jego merytoryczne powiązanie z tematem jest co najwyżej pośrednie. W temacie 1 pojawia się także wyraz „ewangelia", który występuje równocześnie w dwóch innych tematach. Jest to słowo niespecyficzne, czyli takie, które jest dość typowe dla całego korpusu dokumentów i jednocześnie zrozumiałe w kontekście tego konkretnego słowozbioru, choć jego wystąpienie nie pomaga w interpretacji tematu.

Nie wchodząc już w szczegółową interpretację każdego tematu, wyniki MT po analizie pogrupowaliśmy następujący w sposób:

1. Tematy religijne: T1 – nowa ewangelizacja, T2 – kongres eucharystyczny/proces beatyfikacji lub kanonizacji, T5 – pielgrzymki Jana Pawła II, T9 – nauczanie o Jezusie, T10 – rok maryjny, nauczanie o Matce Boskiej, T14 – relacje z Kościołem grekokatolickim, T16 – osoba ludzka w ujęciu religijnym, T17 – informacje o posiedzeniach KEP, T18 – procesy beatyfikacji i kanonizacji, T20 – obchody milenijne, T23 – uroczystości dnia życia konsekrowanego.
2. Tematy społeczne: T3 – środki masowego przekazu, T7 – pojednanie między religiami i narodami, T8 – problem alkoholizmu, T11 – transformacja i jej społeczne następstwa, T12 – bezrobocie, T13 – nauczanie religii w szkole, edukacja szkolna, T25 – sport i zdrowie.
3. Tematy polityczne: T4 – prawo (konstytucja, konkordat), T6 – opinia o wyborach lub, szerzej, o sprawach publicznych, T19 – aborcja w kontekście rodziny, T22 – zapłodnienie in vitro, T24 – integracja europejska.

Identyfikacja głównej osi interpretacyjnej jest w większości przypadków bezproblemowa. Na przykład w temacie 22 wyrażenia „sumienie ludzkie" oraz „komórka", „macierzysty" i „embrion" na czele listy jednoznacznie wskazują na temat zapłodnienia in vitro, poruszany w trzech dokumentach z 2003 i 2004 roku. Na marginesie warto zauważyć, że termin „in vitro" nie pojawia się na liście słów, bo był używany przez biskupów bardzo rzadko (zaledwie 6 razy), podobnie zresztą jak słowo i temat „eutanazja", obecne w jednym dokumencie. Czasem jest to mniej oczywiste, jak to widać w przypadku tematu 6, w którym słowa „system" i „społeczny" wyznaczają główną oś interpretacji. Pomocne w takich przypadkach jest zapoznanie się z dokumentami, które w największym stopniu zawierają dany temat. W tym wypadku są to odpowiednio: *Słowo biskupów polskich w sprawie wyborów do parlamentu* (1993), *Stanowisko episkopatu polski w sprawie wyborów parlamentarnych* (1991) i *Słowo biskupów polskich w sprawach publicznych* (1995).

Nie zawsze jednak zapoznanie się z listami jest wystarczające. Nie bez powodu tematy 15 i 21 zostały pominięte w przedstawionym wyżej zestawieniu. Temat 21 wydaje się dotyczyć inauguracji nowego roku liturgicznego, ale słowa „praca" czy „misjonarz" nie pasują do tej interpretacji. Teksty najbardziej reprezentatywne dla tego tematu – *List Episkopatu Polski na temat przykazań kościelnych* (2003), *List otwarty Przewodniczącego Komisji Misyjnej EP do Posłów Sejmu III RP* (2004), *W sprawie powieści „Kod Leonarda da Vinci"* (2005) – dotyczą zupełnie różnych zagadnień. Podobne pomieszanie występuje w przypadku tematu 15. Oba te przypadki ilustrują zjawisko charakterystyczne dla MT, polegające na tym, że modele często przesuwają „zaszumione" dane do nieinterpretowalnych słowozbiorów w taki sposób, aby wzmocnić spójność pozostałych tematów. W związku z tym właściwym testem modelu jako całości jest jego zdolność do identyfikacji pewnej liczby merytorycznie znaczących i analitycznie użytecznych tematów, nie zaś możliwość optymalizacji wszystkich tematów (DiMaggio, Nag, Blei, 2013).

Wyniki MT nie są zatem ani bezproblemowe, ani jednoznaczne. Metoda ta wcale też nie automatyzuje całości procesu analizy: opracowanie wygenerowanych list słów to dopiero połowa pracy. Zwróćmy uwagę, że na etapie identyfikacji tematu i przypisania go do jednej z wyróżnionych kategorii ogólnych mieliśmy do czynienia z analizą jakościową. Polegała ona na lekturze listy słów i ich interpretacji.

Zadanie to wykonała osoba, która przeprowadziła jakościową analizę tematyczną, bazując na swojej wiedzy o treści dokumentów. Kiedy tego samego zadania podjęła się osoba, która nie czytała wcześniej dokumentów KEP, miała ona trudności z nadaniem jednoznacznych etykiet niektórym słowozbiorom. W tym przypadku konieczna była lektura dokumentów najbardziej reprezentatywnych dla danego tematu, zgodnie z podejściem zaproponowanym przez Justina Grimmera (2010). Dopiero taka dwustopniowa analiza pozwoliła na rekonstrukcję głównych osi tematów.

Porównanie wyników modelowania matematycznego i analizy tematycznej

Na pierwszy rzut oka wyniki AT i MT są do siebie podobne. Z obu analiz wynika, że biskupi zasiadający w KEP dużo uwagi poświęcają sprawom religijnym i społecznym, nieco rzadziej wypowiadają się natomiast w kwestiach politycznych. Również kiedy dokładniej przyjrzymy się tematom w każdym z tych trzech obszarów, wyniki analiz okażą się zbliżone. Wśród wyodrębnionych tematów społecznych bez problemu dostrzeżemy zagadnienia takie jak alkoholizm, media lub sport, którym poświęcone były odrębne listy, czy wątki dotyczące pracy, bezrobocia lub rodziny, poruszane w różnych dokumentach.

Podobieństwa AT i MT wynikają z tego, że tak jak osoba kodująca teksty pod kątem pojawiających się w nich tematów zwraca uwagę na charakterystyczne słowa, ich synonimy i kontekst, w jakim się pojawiają, tak podobnie algorytm identyfikuje powtarzające się wzorce współwystępowania słów. Można zatem zasadnie twierdzić, że algorytm przetwarzania tekstu w MT do pewnego stopnia naśladuje pracę osoby czytającej tekst. Sięgając głębiej, można zauważyć, że obie metody łączy założenie, zgodnie z którym znaczenie tekstu da się sprowadzić do „faktycznej inwariancji semantycznej" (ang. *effective semantic invariance*) (Enfield, 2014), czyli słów, których podstawowy sens wyraża się w znaczeniu względnie jednoznacznym (leksykalnym), dostępnym dla wszystkich odbiorców. Z tego punktu widzenia temat można ujmować jako grupę słów powiązanych wspólnym znaczeniem wyrażonym we współwystępowaniu. Co więcej, MT pozwala na uchwycenie polisemii tekstu i rozróżnienie zastosowań danego terminu na podstawie kontekstu, w którym się on pojawia. W ten sposób metoda ta realizuje założenie o relacyjności znaczenia, przyjmowane na gruncie lingwistyki i socjologii kultury (DiMaggio, Nag, Blei, 2013). Można więc dojść do wniosku, że różnice pomiędzy lekturą tekstu przez człowieka i lekturą tekstu przez komputer są stosunkowo niewielkie, a MT i AT to metody w dużym stopniu ekwiwalentne.

Wniosek taki byłby jednak pochopny. Modelowanie tematyczne nie może zastąpić analizy treści. Co więcej, jak postaramy się poniżej uzasadnić, wraz z rozwojem obliczeniowych nauk społecznych badania jakościowe wydają się bardziej potrzebne niż kiedykolwiek. Zdając sobie sprawę z podobieństw MT do AT, trzeba wiedzieć, jakie są między nimi różnice – dopiero w ten sposób będzie można w pełni wykorzystać potencjał leżący w połączeniu obu metod.

Po pierwsze, zastosowanie algorytmów MT nie wymaga wcześniejszego zakodowania materiału – tematy wyłaniają się z analizy oryginalnych, nieoznakowanych tekstów. Punktem wyjścia nie jest intuicja badaczy i badaczek, lecz identyfikacja powtarzających się wzorców użycia słów. To pozwala zająć się

problemem stronniczości (ang. *bias*) i subiektywizmu procesu analizy, ograniczającego możliwość powtórzenia badania w celu weryfikacji uzyskanych wyników, oraz umożliwia dostrzeżenie w korpusie nieoczywistych wzorców i zależności, trudnych do uchwycenia podczas analizy prowadzonej blisko tekstu. Nie bez powodu wprowadzenie elektronicznych korpusów porównywano do wynalezienia mikroskopu i lunety, które pozwoliły badaczom obserwować obiekty, jakich nigdy wcześniej nie widzieli (Stubbs, 1996: 231–232). Jednak narzędzia do MT nie mają wiedzy na temat wątków poruszanych w analizowanych dokumentach. Badania tekstu nie da się zredukować do odczytania słów, przypisania im leksykalnego znaczenia i wyodrębnienia powiązań między nimi na podstawie współwystępowania. Powyżej opisaliśmy szczegółowo jeden temat, aby pokazać, jak można zinterpretować jego sens, ale co ważniejsze, chcieliśmy w ten sposób zwrócić uwagę na fakt, że w interpretację tematu, w praktyce zbioru słów w ich podstawowej formie, włączona jest rozbudowana wiedza o analizowanym materiale. Wyodrębnienie tematu jest prowadzone nie tylko na podstawie podobieństw grup słów, ale również różnic pomiędzy nimi. Co ważniejsze, do interpretacji niezbędna jest wiedza o podmiocie i przedmiocie tekstu. W analizie tematu wykorzystywaliśmy między innymi wiedzę o działalności Kościoła w analizowanym okresie, miejscu tej instytucji w społeczeństwie, wydarzeniach politycznych i społecznych, religijności społeczeństwa polskiego. Nie bez znaczenia były również założenia wstępne określające kontekst interpretacji, z których tylko część była zapewne uświadomiona (Aspers, Corte, 2019).

Po drugie, MT umożliwia pracę z dużo większymi zbiorami danych. Przejście od jakościowej analizy treści do MT często oznacza przejście od *Small* do *Big Data*. Jest to zazwyczaj jedyne rozwiązanie w sytuacjach, kiedy osoba badająca ma do czynienia z dużymi korpusami tekstów, których nie byłaby w stanie przeanalizować ręcznie ze względu na ograniczenia czasowe i finansowe. W przypadku MT wielkość zbioru jest w zasadzie nieograniczona, a czas analizy znacznie krótszy. Dane nie muszą mieć też formy ustrukturyzowanej, mogą na przykład pochodzić z mediów społecznościowych. Jednak nie każdy zbiór tekstów można z równym powodzeniem poddać MT. Nie powinny to być teksty zbyt długie (dlatego na potrzeby analizy podzieliliśmy część dokumentów KEP na mniejsze części), a, głównie w przypadku LDA, także zbyt krótkie (Tang i in., 2014). Poszczególne dokumenty nie powinny być też nadmiernie różnorodne (tj. poruszać zbyt wielu tematów w zbyt małej liczbie słów) (zob. Boyd-Graber, Hu, Mimno, 2017).

Po trzecie, wprowadzenie obliczeniowych metod analizy tekstu zmienia spojrzenie na tekst. Modelowanie tematyczne zostało pomyślane jako rodzaj rozszerzonego czytania i teoretycznie pozwala wykryć zależności między tekstami bez konieczności zapoznania się z ich treścią. W praktyce jednak choćby ogólna znajomość analizowanego materiału jest niezbędna, szczególnie na etapie podejmowania decyzji dotyczącej liczby wyodrębnianych tematów. Modele o różnych parametrach mogą prowadzić do odmiennych wyników (Grimmer, Roberts, Stewart, 2022). Sporym rozczarowaniem może być fakt, że algorytmy nie wskażą, ile tematów znajduje się w korpusie. Określenie optymalnej liczby tematów, bez choćby pobieżnej znajomości analizowanych tekstów, może być bardzo problematyczne. Powstało kilka metod, które pozwalają zautomatyzować ten krok, ułatwiając podjęcie decyzji i walidację wyników, ale są one zawodne (Hoyle i in., 2021). Co jednak ważniejsze, podejście jakościowe pokazuje, że temat nie jest obiektywnym bytem zamkniętym w tekście niczym muszka w bursztynie. Nie wystarczy go rozłupać i wydobyć. Temat – jak w naszej analizie dokumentów KEP – może zostać zdefiniowany na

różnym poziomie ogólności. Kluczowe są pytania badawcze i intuicja wynikająca ze znajomości materiału, a analiza ma charakter iteracyjny – tworzymy kilka modeli z różną liczbą tematów, odmiennymi parametrami i wybieramy ten, który charakteryzuje się najwyższą interpretowalnością wyników i wysokim dopasowaniem z punktu widzenia celów badania oraz wiedzy o przedmiocie tekstu. Miary statystyczne mają charakter pomocniczy i należy je traktować raczej jako wskazówki niż jako główną metodę walidacji modelu i podstawę wyboru parametrów (Isoaho, Gritsenko, Mäkelä, 2021). Jak wskazuje się w literaturze, pełna walidacja wymaga kombinacji miar statystycznych (zob. Mimno, Blei, 2011), semantycznych (ręczne kodowanie wybranych tekstów i weryfikacja, czy model rozróżnia poszczególne znaczenia tych samych lub podobnych terminów) i predykcyjnych lub zewnętrznych (np. sprawdzenie, czy częstość występowania poszczególnych tematów w kolejnych tekstach zmieniała się zgodnie z oczekiwaniami badaczy) (DiMaggio, Nag, Blei, 2013; DiMaggio, 2015).

Po czwarte wreszcie, MT i AT różnią się pod względem jednostek analizy. W AT są to cytaty, kody i tematy, a w MT dokumenty i słowa. Przekłada się to na odmienne rozumienie tego, czym są tematy. Jak już zostało powiedziane, w MT tematy to w istocie zbiory współwystępujących i – z założenia – podobnych do siebie słów i również teksty reprezentowane są jako zbiory słów. Dokumenty i słowa są czymś obserwowalnym, natomiast tematy i ich rozkład w dokumencie oraz rozkład słów w tematach (słowozbiorach) nie są obserwowane i reprezentują, jak to określił David Blei (2012), „ukrytą strukturę". Założenie o istnieniu takiej „ukrytej struktury" – jako pewnej niezmiennej i możliwej do zdefiniowania przestrzeni semantycznej – jest problematyczne. Tematy czy kategorie treści są trudne do zdefiniowania, a ich granice są zazwyczaj nieostre i w dużej mierze wywodzą się raczej ze znaczenia ukrytego niż ze współwystępowania słów (Shadrova, 2021). Z faktu, że słowa zgrupowane na podstawie współwystępowania wykazują podobieństwa, nie wynika koniecznie, że niosą one ze sobą odrębne tematy (Klein i in., 2015).

Wykorzystanie modelowania tematycznego na potrzeby badań jakościowych

Wniosek, który wynika z naszych dotychczasowych analiz i rozważań, znany jest od dawna, ale warto go przypominać, szczególnie w sytuacji, kiedy nowe podejście zyskuje ogólne zainteresowanie: nie ma metod idealnych, wszystkie metody „są hybrydami, emergentnymi, interaktywnymi produkcjami, produktywnie rozszerzającymi dyskurs mieszanych metod-paradygmatów" (Denzin, 2010: 423). Podobnie jak analiza treści ma znane od dawna ograniczenia, swoje ograniczenia ma również MT. Możemy zrównoważyć problemy każdej z metod, stosując je razem. Prowadzić to będzie do zwiększenia możliwości przeprowadzenia pogłębionej analizy danych i przyczyni się do jej nasycenia (Flick, 1992). Szkopuł tkwi w połączeniu metod jakościowych i ilościowych. Wyzwaniem jest nie tylko odmienność ich epistemologicznych i ontologicznych założeń. Jak zauważył Uwe Flick (2010: 30): „po obu stronach tego podziału istnieje nurt jednoznacznie odrzucający drugą stronę". Apelował on o przekroczenie tych uprzedzeń, w łączeniu obu podejść upatrując „przyszłości badań społecznych w ogóle" (Flick, 2010: 34).

Pod sztandarem koncyliacyjnych apeli Flicka chcielibyśmy w podsumowaniu artykułu przedstawić konkretne propozycje sprzęgnięcia AT z MT, widząc w tym przyczynek do szerszej debaty nad łączeniem

podejść obu zwaśnionych obozów. Wbrew technooptymizmowi niektórych zwolenników automatycznych metod analizy tekstu nie twierdzimy, że wyprą one jakościową analizę treści. Wręcz przeciwnie. Niezaprzeczalnie mają one wiele zalet, jednak na każdym etapie realizacji badania niezbędne jest zaangażowanie człowieka, zgodnie z perspektywą określaną jako *human-in-the-loop* (Rahman, Kandogan, 2022). Przypomnijmy tylko to, co najważniejsze: po stronie osób badających leży decyzja o wyborze danych do analizy i ich przygotowaniu oraz ocena wartości wykrytych wzorców i odróżnienie wśród nich tych, które nie wnoszą nic nowego, od tych, które mają charakter nieznanej, odkrywanej wiedzy (Bryda, 2014). Tak więc, jak to określili Grimmer, Roberts i Stewart (2022), jeśli chcemy, aby automatyczna analiza dużych zbiorów danych prowadziła do interesujących ustaleń empirycznych i rozwoju teorii, niezbędna jest pogłębiona praca interpretacyjna, której nie da się zautomatyzować.

Biorąc pod uwagę powyższe ustalenia, proponujemy cztery sposoby wykorzystania MT w jakościowej analizie treści. Po pierwsze, wykorzystanie MT do walidacji. Modelowanie tematyczne jako narzędzie walidacji jakościowej analizy treści pozwoli na weryfikację, czy została ona prawidłowo przeprowadzona, i sprawdzenie, czy w toku analizy coś nam nie umknęło. Rozwiązanie takie warto stosować szczególnie w przypadku, kiedy w badanie zaangażowanych jest kilka osób kodujących. Wykorzystanie wskaźników zgodności (ang. *reliability tests*), takich jak Cohena kappa lub Krippendorffa alpha, jest czasochłonne i, co ważniejsze, możliwe do zastosowania do niewielkiej części materiału. Modelowanie tematyczne nie zastąpi tych wskaźników – mają one również inne funkcje – ale wyobraźmy sobie sytuację, w której chcemy sprawdzić, czy materiał oznakowany danym kodem lub grupą kodów jest tematycznie spójny. Możemy wyodrębnić zakodowany materiał i przeprowadzić na nim MT. Jeśli efekty będą bardzo podobne, możemy przyjąć, że kodowanie jest spójne.

Drugi sposób odnosi się do etapu konceptualizacji. Sprawdzi się on szczególnie w badaniach eksploracyjnych, kiedy wiedza na temat badanego tekstu jest niewielka lub pytania badawcze sformułowane zostały w sposób ogólny. W takich przypadkach stosujemy zwykle technikę kodowania otwartego, często w połączeniu z teorią ugruntowaną. Eksploracja materiału za pomocą MT może wspomóc ten proces, stanowiąc albo punkt wyjścia do opracowania klucza kodowego, albo metodę weryfikacji klucza tworzonego podczas lektury. Procedura ta wpisuje się w podejście indukcyjne (ang. *bottom-up*) w ramach AT, zgodnie z którym tematy są silnie związane z samymi danymi i wyłaniają się w toku analizy, pozwala jednak usprawnić proces badawczy.

Trzeci sposób dotyczy kwestii selekcji materiału do analizy. Problemem spotykanym w analizie dokumentów publikowanych w dłuższym okresie jest identyfikacja i dobranie tekstów do badania. Modelowanie tematyczne może wspomóc istniejące już metodologie i oszczędzić wiele czasu. Przyjmijmy, że – posługując się przykładem dokumentów KEP – interesują nas wszystkie wypowiedzi KEP na temat aborcji (a możemy sięgnąć aż do XIX wieku). Możemy dokonać selekcji na podstawie słów kluczowych, takich jak „aborcja" czy „życie poczęte". Wady tego podejścia to jednak braki: hierarchii tematycznej, możliwości skupienia się na konkretnej tematyce i analizy powiązań między dokumentami. Modelowanie tematyczne pozwala na identyfikację wątków dotyczących aborcji (w przeprowadzonej przez nas analizie jest to temat 19) i wskazanie dokumentów, w których są one w największym

stopniu reprezentowane. Te dokumenty można następnie poddać analizie jakościowej. Co więcej, model wyuczony na wyjściowym korpusie można wykorzystać do selekcji nowego materiału (np. najnowszych dokumentów KEP), zapewniając powtarzalność procedury. Innym przypadkiem, w którym sprawdzi się takie rozwiązanie, jest analiza prasy. Modelowanie tematyczne pozwala na efektywne przeszukiwanie archiwów prasowych pod jasno określonym kątem.

Ostatni, czwarty sposób dotyczy pięty achillesowej badań jakościowych – generalizacji. Zarzutami często stawianymi analizie treści – pozostawmy na boku kwestię, na ile dobrze uzasadnionymi – są możliwość uogólniania wyników badań i wiarygodność budowanej na ich podstawie teorii. Modelowanie tematyczne nie rozwiąże wszystkich problemów z tym związanych, ale może pomóc w wykryciu zmian w ciągu wielu lat i sprawdzeniu powszechności czy też specyficzności występowania określonych tematów. W ramach proponowanego podejścia punktem wyjścia jest AT przeprowadzona na próbce tekstów z większego zbioru. Celem analizy jest określenie, ilu i jakich tematów należy spodziewać się w całym korpusie. Wiedza z etapu AT jest następnie wykorzystana w MT – służy do budowy optymalnego modelu i pogrupowania dokumentów z całego korpusu zgodnie z wytycznymi[8]. W ten sposób jakościowa analiza treści wspomaga budowę teorii, a MT – poprzez płynne przejście od *Small* do *Big Data* – pozwala na generalizację i weryfikację wyników na dużych zbiorach danych. Konkretne rozwiązania metodologiczne w tym przypadku będą się różnić w zależności od problemu i dostępnych danych.

W artykule tym pokazaliśmy, że MT nie zastąpi pogłębionej jakościowej analizy treści. Metoda ta, choć z pozoru ściśle automatyczna, zakłada domyślnie uprzednią wiedzę osoby badającej, która, jeśli nie zostanie poddana urefleksyjnieniu w ramach spójnej metodologii, prowadzi do niepewnych, a nawet arbitralnych wniosków. Z drugiej strony MT może w istotny sposób wzbogacić warsztat osoby prowadzącej badania jakościowe. Bez względu na to, z jakim badaniem mamy do czynienia, wykorzystanie MT i – szerzej – obliczeniowych metod analizy tekstu w naukach społecznych wymaga dobrze przemyślanej metodologii. Narzędzi do analizy jest coraz więcej, rozwijane są programy i aplikacje internetowe, z których można korzystać bez znajomości języków programowania, uzyskując wyniki za pomocą kilku kliknięć. Jednak łatwość stosowania nie idzie w parze z rozwojem standardów dotyczących przygotowania tekstów, dopasowania algorytmów do typu dokumentów oraz oceny wyników. Dodatkowymi wyzwaniami są specyfika języka polskiego i specyfika nauk społecznych jako dziedziny. Sprawiają one, że nie można mechanicznie implementować rozwiązań dostosowanych do innych języków, danych i problemów badawczych – taka prosta implementacja prowadzi bowiem zwykle do mało pogłębionych interpretacji. Potrzeba zatem więcej prac, które – korzystając z metodologicznie ugruntowanych podejść badawczych – krytycznie zweryfikują efekty metod automatycznych.

8 Jednym z wariantów tego podejścia może być częściowo nadzorowane MT, w którym osoba badająca – mając wiedzę na temat korpusu i określone oczekiwania dotyczące zakresu rozpoznawanych tematów – nie tylko określa liczbę słowozbiorów, lecz także dostarcza dla poszczególnych tematów słowa „zalążkowe" (ang. *seed words*). Następnie – w dużym uproszczeniu – algorytm buduje tematy wokół dostarczonych słów. Wariant ten z jednej strony pozwala na nakierowanie modelu w stronę identyfikowania oczekiwanych tematów, określonych na przykład w ramach analizy jakościowej, z drugiej pozostawia miejsce na odkrycie „nieznanych" tematów (zob. np. Venugopalan, Gupta, 2022).

Bibliografia

Aspers Patrik, Corte Ugo (2019), *What is Qualitative in Qualitative Research*, „Qualitative Sociology", vol. 42(2), s. 139–160, https://doi.org/10.1007/s11133-019-9413-7

Baranowski Mariusz, Cichocki Piotr (2021), *Good and bad sociology: Does topic modelling make a difference?*, „Society Register", vol. 5(4), s. 7–22.

Bazeley Pat (2010), *Computer assisted integration of mixed methods data sources and analyses*, [w:] Abbas Tashakkori, Charles Teddlie (red.), *Handbook of mixed methods in social and behavioral research*, Los Angeles: Sage Publications, s. 431–468.

Blei David M. (2012), *Probabilistic topic models*, „Communications of the ACM", vol. 55(4), s. 77–84, https://doi.org/10.1145/2133806.2133826

Blei David M., Lafferty John D. (2006), *A correlated topic model of Science*, „Advances in Neural Information Processing Systems", vol. 18, s. 147–154, https://doi.org/10.1214/07-AOAS114

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

Boyd-Graber Jordan, Hu Yuening, Mimno David (2017), *Applications of Topic Models*, „Foundations and Trends in Information Retrieval", vol. 11(2–3), s. 143–296, https://doi.org/10.1561/1500000030

Braun Virginia, Clarke Victoria (2006), *Using Thematic Analysis in Psychology*, „Qualitative Research in Psychology", vol. 3(2), s. 77–101.

Braun Virginia, Clarke Victoria (2022), *Thematic analysis: a practical guide*, Los Angeles: Sage Publications.

Brennan Timothy (2017), *The digital-humanities bust: After a decade of investment and hype, what has the field accomplished? Not much*, „Chronicle of Higher Education", vol. 64(8).

Brosz Maciej, Bryda Grzegorz, Siuda Piotr (2017), *Big Data i CAQDAS a procedury badawcze w polu socjologii jakościowej*, „Przegląd Socjologii Jakościowej", vol. XIII, nr 2, s. 6–23.

Bryda Grzegorz (2014), *CAQDAS, Data Mining i odkrywanie wiedzy w danych jakościowych*, [w:] Jakub Niedbalski (red.), *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 13–40.

Bryda Grzegorz (2020), *Whats and Hows? The Practice-Based Typology of Narrative Analyses*, „Przegląd Socjologii Jakościowej", vol. XVI, nr 3, s. 120–142.

Chang Jonathan, Boyd-Graber Jordan L., Gerrish Sean, Wang Chong, Blei David M. (2009), *Reading tea leaves: How humans interpret topic models*, „Advances in Neural Information Processing Systems", vol. 22, s. 1–9.

Chen Yingying, Zhao Peng, Sei-Hill Kim, Chang Won Choi (2023), *What We Can Do and Cannot Do with Topic Modeling: A Systematic Review*, „Communication Methods and Measures", vol. 17(2), s. 1–20, https://doi.org/10.1080/19312458.2023.2167965

CLARIN-PL (b.r.), *Shortextopic*, https://ws.clarin-pl.eu/shortextopic [dostęp: 14.03.2023].

CLARIN-PL (b.r.), *Topiki*, https://ws.clarin-pl.eu/topic [dostęp: 14.03.2023].

Collingridge Dave S. (2013), *A Primer on Quantitized Data Analysis and Permutation Testing*, „Journal of Mixed Methods Research", vol. 7(1), s. 81–97, https://doi.org/10.1177/1558689812454457

Da Nan Z. (2019), *The Computational Case against Computational Literary Studies*, „Critical Inquiry", vol. 45(3), s. 601–639, https://doi.org/10.1086/702594

Delong Marek (2016), *Konferencja Episkopatu Polski wobec wybranych kwestii politycznych i społecznych w Polsce w latach 1989–2014*, Rzeszów: Wydawnictwo Uniwersytetu Rzeszowskiego.

Delong Marek (2017a), *Problem prawnej ochrony życia w enuncjacjach Konferencji Episkopatu Polski w latach 1989–2011*, „UR Journal of Humanities and Social Sciences", vol. 2(1), s. 84–97, https://doi.org/10.15584/johass.2017.1.5

Delong Marek (2017b), *Wybrane kwestie wychowania młodego pokolenia w enuncjacjach Konferencji Episkopatu Polski w latach 1989–2013*, „Studia Sandomierskie. Teologia – Filozofia – Historia", vol. 24(1), s. 249–260.

Denzin Norman K. (2010), *Moments, Mixed Methods, and Paradigm Dialogs*, „Qualitative Inquiry", vol. 16(6), s. 419–427, https://doi.org/10.1177/1077800410364608

DiMaggio Paul (2015), *Adapting computational text analysis to social science (and vice versa)*, „Big Data & Society", vol. 2(2), s. 1–5, https://doi.org/10.1177/2053951715602908

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

Eder Maciej (2016), *Słowa znaczące, słowa kluczowe, słowozbiory – o statystycznych metodach wyszukiwania wyrazów istotnych*, „Przegląd Humanistyczny", vol. 60(3), s. 31–44.

Enfield N.J. (2014), *The Utility of Meaning: What Words Mean and Why*, Oxford: Oxford University Press.

Evans James A., Aceves Pedro (2016), *Machine Translation: Mining Text for Social Theory*, „Annual Review of Sociology", vol. 42(1), s. 21–50, https://doi.org/10.1146/annurev-soc-081715-074206

Flick Uwe (1992), *Triangulation Revisited: Strategy of Validation or Alternative?*, „Journal for the Theory of Social Behavior", vol. 22(2), s. 175–197, https://doi.org/10.1111/j.1468-5914.1992.tb00215.x

Flick Uwe (2010), *Projektowanie badania jakościowego*, przełożył Paweł Tomanek, Warszawa: Wydawnictwo Naukowe PWN.

Friese Susanne (2019), *Qualitative Data Analysis with Atlas.ti*, Los Angeles: Sage Publications.

Goldthorpe John H. (2012), *Współczesna etnografia społeczna: problemy i perspektywy*, [w:] John H. Goldthorpe, *O socjologii: integracja badań i teorii*, przełożyła Jerzyna Słomczyńska, Warszawa: Wydawnictwo IFiS PAN, s. 103–136.

Grimmer Justin (2010), *A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases*, „Political Analysis", vol. 18(1), s. 1–35, https://doi.org/10.1093/pan/mpp034

Grimmer Justin, Stewart Brandon M. (2013), *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*, „Political Analysis", vol. 21(3), s. 267–297, https://doi.org/10.1093/pan/mps028

Grimmer Justin, Roberts Margaret E., Stewart Brandon M. (2022), *Text as Data: A New Framework for Machine Learning and the Social Sciences*, Princeton: Princeton University Press.

Grootendorst Maarten (2022), *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*, https://doi.org/10.48550/arXiv.2203.05794

Hess Agnieszka, Hwaszcz Krzysztof (2022), *Językoznawstwo korpusowe w badaniach medioznawczych – ujęcie historyczne i praktyczne*, „Journal of Humanities and Social Sciences", vol. 4(25), s. 118–132.

Honnibal Matthew, Montani Ines, Van Landeghem Sofie, Boyd Adriane (2020), *spaCy: Industrial-strength Natural Language Processing in Python*, https://doi.org/10.5281/zenodo.1212303

Hoyle Alexander, Goel Pranav, Hian-Cheong Andrew, Peskov Denis, Boyd-Graber Jordan, Resnik Philip (2021), *Is automated topic model evaluation broken? The incoherence of coherence*, „Advances in Neural Information Processing Systems", vol. 34, s. 2018–2033.

Index of /~rtuora/spacy (b.r.), http://mozart.ipipan.waw.pl/~rtuora/spacy/ [dostęp: 11.03.2023].

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

Kamasa Victoria (2013a), *Rodzina w dyskursie polskiego Kościoła katolickiego. Badania korpusowe z perspektywy krytycznej analizy dyskursu*, „Socjolingwistyka", vol. 27, s. 139–152.

Kamasa Victoria (2013b), *Naming "In Vitro Fertilization": Critical Discourse Analysis of the Polish Catholic Church's Official Documents*, „Procedia – Social and Behavioral Sciences", vol. 95, s. 154–159.

Klein Lauren F., Eisenstein Jacob, Sun Iris, Jacko J.A. (2015), *Exploratory Thematic Analysis for Digitized Archival Collections*, „Digital Scholarship in the Humanities", vol. 30, s. 30–41.

Krippendorff Klaus (2018), *Content analysis: an introduction to its methodology*, Los Angeles: Sage Publications.

Kwiatkowska Agnieszka (2017), *„Hańba w Sejmie" – zastosowanie modeli generatywnych do analizy debat parlamentarnych*, „Przegląd Socjologii Jakościowej", t. XIII, nr 2, s. 82–109.

Kwiatkowska Agnieszka, Chwedczuk-Szulc Karol, Bolechów Bartosz (2022), *Disentangling the Moral Rightness of Securitization: Data Mining of the Process of Framing and Shaping of Poland-United States Relations*, „Polish Political Science Review", vol. 10(1), s. 35–58, https://doi.org/10.2478/ppsr-2022-0003

Leszczyńska Katarzyna, Zych Łukasz (2011), *Wzory kobiecości w dyskursie Kościoła rzymskokatolickiego w Polsce*, [w:] Krystyna Slany (red.), *Kalejdoskop genderowy. W drodze do poznania płci społeczno-kulturowej w Polsce*, Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego, s. 201–216.

Macanovic Ana (2022), *Text mining for social science – The state and the future of computational text analysis in sociology*, „Social Science Research", vol. 49(1), 102784, https://doi.org/10.1016/j.ssresearch.2022.102784

Mallet: MAchine Learning for LanguagE Toolkit (b.r.), https://mimno.github.io/Mallet/index [dostęp: 11.03.2023].

Mann Gideon S., Mimno David, McCallum Andrew (2006), *Bibliometric impact measures leveraging topic analysis*, [w:] *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL '06)*, New York: Association for Computing Machinery, s. 65–74, https://doi.org/10.1145/1141753.1141765

Maryl Maciej, Eder Maciej (2017), *Topic Patterns in an Academic Literary Journal: The Case of "Teksty Drugie"*, https://dh-abstracts.library.virginia.edu/works/4012 [dostęp: 15.12.2022].

McLevey John (2022), *Doing computational social science: a practical introduction*, Los Angeles: Sage Publications.

Mimno David, Blei David M. (2011), *Bayesian Checking for Topic Models*, [w:] *EMNLP'11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh–Stroudsburg: Association for Computational Linguistics, s. 227–237.

Moretti Franco (2016), *Wykresy, mapy, drzewa. Abstrakcyjne modele na potrzeby literatury*, przełożyli Tomasz Bilczewski, Anna Kowalcze-Pawlik, Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Nzabonimpa Jean Providence (2018), *Quantitizing and qualitizing (im-)possibilities in mixed methods research*, „Methodological Innovations", vol. 11(2), s. 1–16, https://doi.org/10.1177/2059799118789021

Onwuegbuzie J. Anthony, Johnson R. Burke, Collins M. Kathleen (2009), *Call for mixed analysis: A philosophical framework for combining qualitative and quantitative approaches*, „International Journal of Multiple Research Approaches", vol. 3(2), s. 114–139, https://doi.org/10.5172/mra.3.2.114

Pawłowski Adam, Walkowiak Tomasz (2022), *Statistical tools, automatic taxonomies, and topic modelling in the study of self-promotional mission and vision texts of Polish universities*, [w:] Makoto Yamazaki, Haruko Sanada, Reinhard Köhler, Sheila Embleton, Relja Vulanović, Eric S. Wheeler (red.), *Quantitative Approaches to Universality and Individuality in Language*, Berlin–Boston: De Gruyter Mouton, s. 131–145.

Perrin Andrew J., Tepper Steven J., Caren Neal, Morris Sally (2014), *Political and Cultural Dimensions of Tea Party Support, 2009–2012*, „The Sociological Quarterly", vol. 55(4), s. 625–652, https://doi.org/10.1111/tsq.12069

Rahman Sajjadur, Kandogan Eser (2022), *Characterizing Practices, Limitations, and Opportunities Related to Text Information Extraction Workflows: A Human-in-the-Loop Perspective*, [w:] *CHI Conference on Human Factors in Computing Systems*, New Orleans: ACM, s. 1–15, https://doi.org/10.1145/3491102.3502068

Roberts Margaret E., Stewart Brandon M., Tingley Dustin, Lucas Christopher, Leder-Luis Jetson, Kushner Gadarian Shana, Albertson Bethany, Rand David G. (2014), *Structural Topic Models for Open-Ended Survey Responses*, „American Journal of Political Science", vol. 58(4), s. 1064–1082, https://doi.org/10.1111/ajps.12103

Shadrova Anna (2021), *Topic models do not model topics: epistemological remarks and steps towards best practices*, „Journal of Data Mining & Digital Humanities", 7595, https://doi.org/10.46298/jdmdh.7595

Shah Dhavan V., Cappella Joseph N., Neuman W. Russell (2015), *Big Data, Digital Media, and Computational Social Science: Possibilities and Perils*, „The ANNALS of the American Academy of Political and Social Science", vol. 659(1), s. 6–13, https://doi.org/10.1177/0002716215572084

Skowronek Katarzyna (2006), *Między sacrum a profanum: studium językoznawcze listów pasterskich Konferencji Episkopatu Polski (1945–2005)*, Kraków: Wydawnictwo Lexis.

Skowronek Katarzyna (2007), *Między sacrum a profanum*, „Zeszyty Prasoznawcze", nr 50(3–4), s. 191–192.

Stubbs Michael (1996), *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*, Oxford: Blackwell.

Szwed Anna (2018), *„Przyszłość ludzkości idzie przez rodzinę!". Konstruowanie kryzysu rodziny w wypowiedziach Kościoła rzymskokatolickiego w Polsce – treści i funkcje*, „Przegląd Religioznawczy", t. 2, s. 81–96.

Szwed Anna (2019), *Typy legitymizacji w wypowiedziach hierarchów Kościoła rzymskokatolickiego w Polsce na temat gender i praw reprodukcyjnych*, „Studia Socjologiczne", t. 3, s. 81–108.

Tang Jian, Meng Zhaoshi, Nguyen Xuan Long, Mei Qiaozhu, Zhang Ming (2014), *Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis*, [w:] *Proceedings of the 31st International Conference on Machine Learning*, vol. 32, New York: ACM, s. 190–198.

Teddlie Charles, Tashakkori Abbas (2009), *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*, Los Angeles: Sage Publications.

Underwood Ted (2019), *Distant Horizons: Digital Evidence and Literary Change*, Chicago: University of Chicago Press.

Venugopalan Manju, Gupta Deepa (2022), *An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis*, "Knowledge-Based Systems", vol. 246, 108668.

Vetulani Zygmunt, Vetulani Grażyna (2020), *The case of Polish on its Way to Become a WellResourced-Language*, [w:] Adda Gilles (red.), *International conference on language technologies for all: enabling linguistic diversity and multilingualism worldwide. Proceedings of LT4All*, Paris: UNESCO Headquarters, European Language Resources Association, s. 388–392.

Wiedemann Gregor (2013), *Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences*, „Forum Qualitative Sozialforschung/Forum: Qualitative Social Research", vol. 14(2), s. 332–357.

Wiedemann Gregor (2016), *Text mining for qualitative data analysis in the social sciences*, New York–Berlin–Heidelberg: Springer.

Woliński Marcin (2019), *Morfeusz 2. Dokumentacja techniczna i użytkowa*, http://download.sgjp.pl/morfeusz/Morfeusz2.pdf [dostęp: 21.01.2023].

Cytowanie

Sławomir Mandes, Agnieszka Karlińska (2024), *W stronę nowej metodologii analizy treści. Podobieństwa i różnice pomiędzy modelowaniem tematycznym i jakościową analizą treści*, „Przegląd Socjologii Jakościowej", t. XX, nr 4, s. 118–143, https://doi.org/10.18778/1733-8069.20.4.06

Toward a New Methodology for Content Analysis: Similarities and Differences Between Topic Modeling and Qualitative Content Analysis

Abstract: The aim of the paper is to critically reflect on the relationship between qualitative thematic analysis and topic modeling, one of the most popular variants of automatic text mining. Based on the results of a qualitative and quantitative analysis of the documents of the Polish Bishops' Conference, we show the advantages and disadvantages of topic modeling. We negatively verify the thesis of the substitutability of thematic analysis by topic modeling and point to the necessity of combining qualitative and quantitative approaches within the mixed methods methodology. In the final section, we present possible ways of combining the two methods so that qualitative researchers, based on the mixed methods paradigm, can benefit from the advantages of topic modeling and, with the awareness of its advantages and disadvantages, enrich their workshop, broaden the scope of research, and enhance the process of analysis.

Keywords: qualitative content analysis, thematic analysis, topic modeling, text mining, mixed methods