Prof. dr hab. Marek Walesiak
Uniwersytet Ekonomiczny we Wrocławiu
Wydział Ekonomii i Finansów
Katedra Ekonometrii i Informatyki
58-500 Jelenia Góra, ul. Nowowiejska 3

RECENZJA ROZPRAWY DOKTORSKIEJ MGR KATARZYNY WÓJCIK

pt. „Ocena miar podobieństwa dokumentów tekstowych na potrzeby automatycznej analizy opinii konsumenckich”

Uniwersytet Ekonomiczny w Krakowie

(maszynopis powielony, 292 strony plus aneksy załączone na pendrive)

Promotor: Prof. dr hab. Paweł Lula

Promotor pomocniczy: Dr hab. Grażyna Paliwoda-Pękosz, prof. UEK

Eksploracja zawartości dokumentów tekstowych (*text mining*) obejmuje odkrywanie i wykorzystanie wiedzy zawartej w zbiorze dokumentów. Jest to trudne zagadnienie badawcze. W modelowaniu zawartości dokumentów tekstowych informacje wejściowe cechuje brak ustrukturyzowania, a co za tym idzie nie ma tutaj struktury właściwej dla klasycznych metod analizy danych. Statystyczne modelowanie zawartości dokumentów tekstowych wymaga zastosowania w analizie informacji specjalnych technik, które wywodzą się z teorii informacji, systemów uczenia maszynowego, zarządzania wiedzą, baz danych, lingwistyki, matematyki czy statystyki. Podkreślić należy, że narzędzia zaprojektowane do analizy tekstów w jednym języku często nie mogą być bezpośrednio użyte w analizie tekstów napisanych w innym języku.

Recenzowana rozprawa doktorska poświęcona jest automatycznej analizie opinii konsumenckich na podstawie oceny podobieństwa dokumentów tekstowych. Potrzeba podjęcia badań związanych z automatyczną analizą opinii konsumenckich wynika m.in. z rosnącej w tempie wykładniczym liczba opinii dostępnych on-line. W punkcie 3.2.3 (s. 133) doktorantka zdefiniowała automatyczną analizę opinii konsumenckich, jako „ogół działań mających na celu zautomatyzowanie procesu wyszukiwania, ekstrakcji i analizy danych pochodzących ze specyficznych tekstów, jakimi są opinie użytkowników”. W ramach automatycznej analizy opinii konsumenckich wyróżniono trzy rodzaje analizy: klasyfikacja opinii, analiza ukierunkowana na cechy produktu, analiza porównawcza produktów. Dzięki automatycznej analizie opinii konsumenckich decydenci zarządzający przedsiębiorstwem mogą podjąć m.in. takie działania, jak reakcja na negatywne komentarze, usunięcie usterki czy wady produktu, dobór

grupy docelowej kampanii marketingowej, dobór mediów wykorzystywanych w kampanii marketingowej. Przedmiotem rozprawy doktorskiej są więc narzędzia statystyczne wspomagające proces podejmowania decyzji w zarządzaniu przedsiębiorstwem. Problematyka rozprawy mieści się w dyscyplinie nauki o zarządzaniu.

Recenzowana rozprawa doktorska składa się ze wstępu, czterech rozdziałów, zakończenia, wykazu literatury, spisu tabel i rysunków oraz aneksu zamieszczonego na nośniku elektronicznym. Pierwszy rozdział, mający charakter teoretyczny oraz wprowadzający w badaną problematykę, prezentuje charakterystykę opinii respondentów w kontekście społeczeństwa informacyjnego. Rozdziały 2-4 dotyczą oceny podobieństwa dokumentów tekstowych w świetle automatycznej analizy opinii konsumenckich. Rozdziały 2-3, mające charakter podstaw metodycznych, traktują o dokumentach tekstowych, pomiarze ich podobieństwa oraz automatycznej analizie opinii konsumentów. W rozdziale 4 przeprowadzono eksperymenty badawcze pozwalające ocenić jakość procedur badawczych w automatycznej analizie opinii konsumenckich. Układ rozprawy doktorskiej oraz jej strukturę z punktu widzenia jej celu oceniam pozytywnie.

Na s. 4-5 Wstępu uzasadniono podjęcie tematu, sformułowano cel główny pracy, sześć celów szczegółowych oraz dwie hipotezy badawcze.

Rozdział pierwszy (s. 8-77) poświęcony charakterystyce opinii konsumentów w świetle społeczeństwa informacyjnego składa się z wprowadzenia, trzech części oraz podsumowania. W podrozdziale 1.2 wyjaśniono koncepcję społeczeństwa informacyjnego oraz takie pojęcia, jak zarządzanie wiedzą i zarządzanie informacją, systemy informacyjne i informatyczne, informatyka i infologia. Podrozdział 1.3 przedstawia hierarchiczny model informacji (piramida wiedzy). W podrozdziale 1.4 scharakteryzowano zagadnienia dotyczące roli opinii konsumentów jako nośnika w społeczeństwie informacyjnym (opinie konsumentów jako czynnik zmniejszający ryzyko decyzji zakupowej, źródła i formaty opinii konsumentów, motywy dzielenia się opiniami przez konsumentów, badania marketingowe i badania rynku jako obszary badawcze zajmujące się opiniami konsumentów).

Rozdział drugi (s. 78-123), składający się z wprowadzenia, trzech części i podsumowania, traktuje o dokumentach tekstowych i pomiarze ich podobieństwa. W podrozdziale 2.2 wyjaśniono pojęcie dokumentu tekstowego, tekstu oraz pliku tekstowego, scharakteryzowano eksploracyjną analizę tekstów oraz zagadnienie przetwarzania języka naturalnego, który umożliwia komputerom rozumienie, interpretowanie i generowanie odpowiedzi w języku naturalnym w znaczący i użyteczny sposób. W podrozdziale 2.3 (s. 91-109) w szerokim zakresie przedstawiono zagadnienie podobieństwa i odległości w analizie danych. Podrozdział 2.4 (s.

110-122) odnosi się do pomiaru podobieństwa i odległości danych tekstowych. Stanowi więc kanwę dla dalszych badań podjętych w dysertacji.

Rozdział trzeci (s. 124-164) poświęcony automatycznej analizie opinii konsumenckich składa się z wprowadzenia, trzech części oraz podsumowania. W części pierwszej (podrozdział 3.2) określono oraz scharakteryzowano zagadnienie automatycznej analizy opinii konsumentów, w tym znaczenie tej problematyki w społeczeństwie informacyjnym, pojęcia, rodzaje oraz wiedzę dziedzinową związaną z automatyczną analizą opinii konsumentów. Zasadniczą część rozdziału trzeciego stanowi podrozdział 3.3 (s. 140-155), w którym doktorantka zawarła autorski model automatycznej analizy opinii konsumentów. Wyjaśniono tutaj cel opracowania takiego modelu, jego strukturę i charakterystykę (opis). Na koniec tej części zaprezentowano empiryczne podstawy konstrukcji tego modelu. Podrozdział 3.4 (s. 156-164) poświęcono sposobom oceny przydatności miar podobieństwa dokumentów tekstowych w automatycznej analizie opinii konsumentów.

W rozdziale czwartym (s. 165-264) zaprezentowano eksperymenty symulacyjne związane z automatyczną klasyfikacją opinii konsumentów. Rozdział ten oprócz wprowadzenia (s. 165-167) i podsumowania (s. 259-264) zawiera dwa podrozdziały 4.2 (s. 167-212) i 4.3 (s. 212-259). Podrozdział 4.2 poświęcono pozyskaniu i przygotowaniu materiału badawczego obejmującego ekstrakcję opinii konsumentów, utworzenie korpusu opinii i jego wstępne przetworzenie, utworzenie i analizę macierzy częstości oraz charakterystykę pozyskanego materiału badawczego. W podrozdziale 4.3 doktorantka przedstawiła wyniki pięciu eksperymentów badawczych, których celem była ocena podobieństwa dokumentów tekstowych na potrzeby automatycznej analizy opinii konsumentów. W eksperymencie I (punkt 4.3.2, s. 226-230) zbadano wpływ 4 sposobów przekształcania elementów macierzy częstości (s. 226) na wyniki grupowania. W eksperymencie zbudowano 112 modeli (uwzględniając 4 sposoby przekształcania macierzy częstości, 4 miary odległości, 7 metod aglomeracyjnych). W tabeli 22 (s. 228) zaprezentowano 8 najlepszych modeli według wartości skorygowanego indeksu Randa. Celem eksperymentu II (punkt 4.3.3, s. 230-246) było zbadanie wpływu redukcji wymiarów macierzy częstości na wyniki grupowania. W ramach tego eksperymentu rozważono 3 metody redukcji wymiarów macierzy częstości: dolne i górne ograniczenie dla liczby dokumentów, w których musi wystąpić rozpatrywany wyraz (parametr *bounds*); analiza głównych składowych (uwzględniono 10 pierwszych głównych składowych); analizę ukrytych wymiarów semantycznych (uwzględniono w analizie 6 warstw ukrytych wymiarów semantycznych). W eksperymencie osobno dla tych trzech metod redukcji wymiarów macierzy częstości zbudowano modele. W tabelach zaprezentowano najlepsze modele według wartości skorygowanego in-

deksu Randa. W eksperymencie III (punkt 4.3.4, s. 246-252) zbadano wpływ ważenia elementów macierzy częstości wartościami określającymi nacechowanie wyrazów na jakość wyników grupowania. W ramach tego eksperymentu wyróżniono 3 rozwiązania (s. 248): pominięcie w analizie słów neutralnych, ważenie macierzy częstości słownikami nacechowania bez gradacji siły polaryzacji, ważenie macierzy częstości słownikami nacechowania ze zróżnicowaną siłą polaryzacji. W każdym przypadku zbudowano 112 modeli procedur analizy skupień (4 sposoby przekształcania elementów macierzy częstości, 4 odległości, 7 metod klasyfikacji). W trzech kolejnych tabelach zaprezentowano najlepsze modele według wartości skorygowanego indeksu Randa. Eksperyment IV (punkt 4.3.5, s. 253-256) przedstawia ocenę uporządkowania opinii konsumenckich zgodnie z metodą TOPSIS oraz wykorzystującą informację o wydźwięku wyrazów. Opinie konsumentów zostały potraktowane jako warianty wyboru a słowa jako kryteria. Wzorcem (antywzorcem) jest dokument złożony z wszystkich słów ze słownika wyrazów pozytywnych (negatywnych). Rozważono tutaj 16 modeli: 4 modele przekształcenia elementów macierzy częstości, 4 miary odległości. Jakość modeli oceniono na podstawie współczynnika korelacji liniowej Pearsona pomiędzy wartościami miary TOPSIS a liczbami gwiazdek przypisanych każdej opinii. W eksperymencie V (punkt 4.3.6, s. 257-259) Autorka według zapowiedzi porównała wyniki uzyskane w poprzednich eksperymentach z wynikami metody klasyfikacji opinii konsumentów opartej o analizę nacechowania poszczególnych słów w opinii. Niestety sposób prezentacji tego eksperymentu jest niejasny dla czytelnika.

Rozprawę doktorską wieńczy Zakończenie (s. 265-267), w którym w syntetyczny sposób przedstawiono jej podsumowanie dotyczące warstwy poznawczej, metodycznej i aplikacyjnej. Odniesiono się też do sformułowanych we Wstępie dwóch postawionych hipotez badawczych. Ponadto na s. 267 Autorka wskazała na możliwości zastosowania prezentowanej metodologii w obszarze zarządzania.

Rozprawa doktorska mgr Katarzyny Wójcik stanowi indywidualny i oryginalny wkład do problematyki automatycznej analizy opinii konsumenckich na podstawie badania podobieństwa dokumentów tekstowych. Recenzowana rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego oraz ukazuje umiejętność samodzielnego prowadzenia pracy naukowej. Do niewątpliwych walorów recenzowanej rozprawy doktorskiej należy zaliczyć:

1. Przyjęcie interesującej koncepcji w strukturze rozprawy. Każdy rozdział w końcowej części zawiera podsumowanie, w którym w syntetyczny sposób zebrano rozważania podjęte w danym rozdziale. Szczególnie cenne są wnioski końcowe zawarte w podrozdziale 4.4 (s. 259-264) a płynące z przeprowadzonych eksperymentów.

2. Przedstawienie w podrozdziale 3.3 autorskiego modelu automatycznej analizy opinii konsumentów a następnie przeprowadzenie w rozdziale 4 badań zgodnie z tym modelem. Na s. 143 w postaci rys. 38 przedstawiono autorski schemat budowy kompleksowego modelu automatycznej analizy opinii konsumentów (*Complex Opinion Mining Framework* – COMF). W punkcie 3.3.3 szczegółowo zaprezentowano charakterystykę modelu ujętego na rys. 38 oraz rys. 39. Przy prezentacji modelu COMF w punkcie 3.3.3 wyodrębniono 5 poziomów. Szkoda, że nie wskazano ich na rys. 38.

3. Zaprezentowanie w klarowny i precyzyjny sposób na rys. 41 (s. 166) struktury badań empirycznych przeprowadzonych w rozdziale 4 obejmujących dwie fazy:

– faza I – pozyskanie i przygotowanie materiału badawczego obejmującego 3 etapy: ekstrakcja opinii konsumentów, utworzenie korpusu opinii i jego wstępne przetworzenie, utworzenie i analiza macierzy częstości. Poszczególne etapy fazy I zobrazowano w postaci schematów blokowych na rysunkach o numerach 42 (s. 169), 50 (s. 177) i 53 (s. 188) oraz scharakteryzowano w podrozdziale 4.2,

– faza II – pięć eksperymentów badawczych. Celem eksperymentów badawczych w podrozdziale 4.3 była ocena podobieństwa dokumentów tekstowych w automatycznej analizie opinii konsumentów. Pierwsze trzy eksperymenty polegały na przeprowadzeniu analizy skupień badanego zbioru opinii i porównaniu z wzorcami grupowania (zob. s. 213) za pomocą indeksów zgodności podziałów (zob. s. 216). W procedurze analizy skupień uwzględniono 4 miary odległości oraz 7 metod hierarchicznej klasyfikacji aglomeracyjnej (s. 214). W eksperymencie IV dokonano oceny uporządkowania opinii konsumentów zgodnie z metodą TOPSIS wykorzystując informację o wydźwięku wyrazów. Najmniej czytelny jest jednak eksperyment V ujęty na s. 257-259.

4. Zaprezentowanie, na podstawie literatury przedmiotu, szerokiej listy czynników utrudniających automatyczną analizę języka naturalnego (s. 127-128). Uzmysławia to czytelnikowi, że automatyczna analiza opinii konsumentów jest zagadnieniem złożonym i wymagającym zastosowania zaawansowanych metod przetwarzania języka naturalnego.

5. Zestawienie w tabeli 10 na s. 153-155 współautorskich badań doktorantki w formie publikacji prezentujących procesy, narzędzia, zasoby wejściowe i wyjściowe. Przeprowadzone tam badania empiryczne stworzyły solidne podstawy konstrukcji autorskiego modelu automatycznej analizy opinii konsumentów.

6. Automatyzacja procesu analizy opinii konsumenckich wymagała od Doktorantki opracowania wiele narzędzi programistycznych, w tym pozwalających na: automatyczne pozyskiwanie opinii z zasobów internetowych, przeprowadzenie analizy tekstów opinii na pozio-

mie morfologicznym, dokonanie różnorodnych transformacji macierzy częstości, analizę opinii konsumenckich za pomocą metod klasyfikacji oraz metod analizy wielokryterialnej. Cenne skrypty w programie R zawarto w aneksie 3.

7. Recenzowana rozprawa doktorska jest przykładem rzetelnie przeprowadzonych analiz i badań, dociekliwości Autorki, dbałości o wiarygodność i obiektywność rezultatów.

Ta niewątpliwie wartościowa praca doktorska skłania do kilku uwag, choć niektóre z nich mają z pewnością zabarwienie dyskusyjne:

1. Doktorantka przedstawiając na rys. 38 (s. 143) kompleksowy model automatycznej analizy opinii konsumentów (COMF) wskazała na s. 141 trzy cele opracowania takiego modelu:

- identyfikacja kluczowych etapów całego procesu automatycznej analizy opinii konsumentów,
- wskazanie potencjalnych punktów wyjścia do dodatkowych analiz wspierających lub rozszerzających automatyczną analizę opinii konsumentów,
- porównanie różnych schematów / procedur automatycznej analizy opinii konsumentów opisanych w literaturze i stosowanych w praktyce.

Moim zdaniem z tymi celami budowy modelu COMF nie koresponduje w pełni tytuł rozprawy doktorskiej. Dlaczego w tytule jest mowa tylko o ocenie miar podobieństwa dokumentów tekstowych? Z przeprowadzonych eksperymentów w rozdziale 4 wynika, że ocena dotyczy procedur.

2. W ostatnim punkcie podrozdziału 4.2 (punkt 4.2.4) przedstawiono charakterystykę pozyskanego z Ceneo.pl materiału badawczego (pobrano 5675 opinii). Moim zdaniem punkt 4.2.4 powinien być na początku rozdziału 4, zaraz po wprowadzeniu lub na początku podrozdziału 4.2. Najpierw trzeba pozyskać materiał badawczy a następnie przygotować ten materiał zgodnie z trzema etapami przedstawionymi w punktach 4.2.1-4.2.3.

3. W tabeli 7 na s. 98 występują błędy i niedociągnięcia: błędy w dzieleniu wyrazów, brakuje przykładu dla zmiennej ilorazowej, na skali ilorazowej powinno być $y \in R_+$, w opisie skali nominalnej zamiast „brak uporządkowania zmiennych" powinno być „brak uporządkowania kategorii zmiennych". Dodatkowo pod tabelą 7 na s. 98 (3 w. od dołu) nie rozumiem zdania „Skale nominalna i porządkowa nie posiadają początku".

4. Wzór (18) oraz oznaczenia dla miary GDM2 na s. 103 zawierają usterki (cyfra 2 powinna być w mianowniku przed pierwiastkiem, wagi przypisane dla zmiennych oznaczone są raz jako w_i a innym razem jako $w_j - j = 1, \dots, m$ oznacza w pracy numer obiektu).

Do recenzowanej pracy zgłaszam też uwagi szczegółowe mające głównie charakter uchybień natury redakcyjnej, stylistycznej i językowej:

– przy odwoływaniu się w tekście do rysunków i tabel brakuje odmiany słów zgodnych z językiem polskim (np. na s. 28 jest „Jak wynika z Rysunek 6 …"; na s. 151 jest „… w Tabela 10 …"),

– błędy lub niejasności (w tym językowe i redakcyjne): na s. 28 na rys. 6 zamiast „LUDZI" powinno być „LUDZIE"; na s. 33 (17 w. od dołu) powinno być „trzecią fundamentalną wartością"; w tytule punktu 1.3.4 (s. 38) brakuje przecinka po słowie „Dane"; błędy z dzieleniem wyrazów na rys. 14 (s. 47); tytuł tabeli 2 powinien być na s. 52 nad tabelą a jest na s. 51; symbole na rys. 24 (s. 85) należy objaśnić; na rys. 25 (s. 87) zamiast „zanczeń" powinno być „znaczeń"; na rys. 26 (s. 88) są literówki w słowach „pojęcie, symbol, komputer"; na s. 88 (10 w. od dołu) zamiast „istotną role" powinno być „istotną rolę"; na s. 94 brakuje wyróżników dla wyodrębnionych trzech podejść; na s. 100 zamiast „symetry" powinno być „symmetry"; na s. 105 w zdaniu pod rysunkiem 29 dwa razy jest „Rysunek 28" oraz zamiast „rysunek 2929" powinno być „rysunek 29"; na s. 112 (10 w. od dołu) zamiast „z problemami charakterystycznych dla …" powinno być „z problemami charakterystycznymi dla …"; na s. 113 (12 w. od dołu) zamiast „modele niekorzystające wiedzy zewnętrznej …" powinno być „modele niekorzystające z wiedzy zewnętrznej …"; odnośnik 15 na s. 148 wygląda na niedokończony; na rys. 40 (s. 165) jest sporo literówek („macierz odległoście", „poszcególncyh", „reprezentujacych"); na s. 180 (3 w. od dołu) zamiast „Zostało ona napisany …" powinno być „Został on napisany …"; s. 181 (3 w. od góry) powinno być „… na Uniwersytecie Ekonomicznym w Krakowie …"; na s. 194 widać tylko część numeru strony; na s. 253 (12 w. od góry) zamiast „sów" powinno być „słów"; na s. 257 (11 w. od góry) nie „ilość" a „liczba"; na s. 267 zamiast „przygotowano programu niezbędne do analizy …" powinno być „przygotowano program niezbędny do analizy …",

– nie wiadomo co ilustruje każdy z 3 rysunków oznaczonych wspólnym numerem rys. 31 (s. 109). Moim zdaniem są to trzy identyczne rysunki,

– często w tekście stosowany jest termin „zamodelowanie" (np. s. 111, 115). Nie jest to zbyt dobre sformułowanie,

– wszystkie wzory oraz oznaczenia do nich powinny być zapisane w edytorze równań. Niestety nie zawsze tak jest. Te same oznaczenia są raz pochylone a innym razem nie są pochylone (np. s. 117-118),

– we wzorze (29) na s. 117 oraz w jego opisie D błędnie oznacza macierz częstości oraz liczbę dokumentów. Ponadto lewą stronę wzoru (29) należało pogrubić lub zapisać w formie $[D]$,

– do zaprezentowanych wzorów w punkcie 2.4.3 (s. 117-119) nie ma powołania na literaturę,

– dlaczego na rys. 87 (s. 254) mowa jest o utworzeniu macierzy odległości? Na s. 253 (14-15 w. od góry) mowa jest o wyznaczaniu odległości od antywzorca z wykorzystaniem miary TOPSIS (39),

– w doktoracie spójnik „iż" jest nadużywany. Jak podaje „Nowy słownik poprawnej polszczyzny" pod red. A. Markowskiego (Warszawa, PWN, 1999), spójnika iż można używać, żeby uniknąć kilkakrotnego powtórzenia spójnika że w jednym zdaniu (np. Wiedział, że go śledzą i że nie dadzą mu spokoju za to, iż wydał wtedy ich wspólnika). W innych wypadkach używanie iż zamiast że jest manieryczne i pretensjonalne.

Recenzowana rozprawa doktorska została obudowana wieloma pozycjami literaturowymi (216 pozycji, w tym 98 w języku angielskim). Wykaz literatury został przygotowany starannie. Podkreślić należy, że wśród cytowanych są dwie pozycje autorskie i 13 współautorskich doktorantki.

Biorąc pod uwagę walory naukowe i poznawcze recenzowanej rozprawy doktorskiej, jej wysoki poziom merytoryczny (w tym: samodzielne rozwiązanie przez Doktorantkę oryginalnego problemu naukowego ujętego w tytule rozprawy, wiedzę teoretyczną w dyscyplinie naukowej – nauki o zarządzaniu, umiejętność samodzielnego prowadzenia pracy naukowej poprzez m.in. zaprezentowanie w rozdziałach 3 i 4 wyników własnych badań, w tym rozwiązań metodycznych i przeprowadzonych badań empirycznych, ujęcia syntetyzujące i porządkujące wiedzę) stwierdzam, że dysertacja mgr Katarzyny Wójcik pt. „Ocena miar podobieństwa dokumentów tekstowych na potrzeby automatycznej analizy opinii konsumenckich w pełni odpowiada wymogom stawianym rozprawom doktorskim określonym w art. 13 ustęp 1 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. 2017 r., poz. 1789, z późn. zm.) i wnoszę o dopuszczenie do publicznej jej obrony.

Ponadto biorąc pod uwagę walory naukowe oraz praktyczne recenzowanej rozprawy doktorskiej umieszczone w recenzji wnioskuję o jej wyróżnienie.

Jelenia Góra, 26 listopada 2024 r. Marek Walesiak