Dr hab. Jerzy Korzeniewski prof. UŁ

Instytut Statystyki i Demografii
Uniwersytet Łódzki

Łódź, 27.11.2024

Recenzja

rozprawy doktorskiej mgr Katarzyny Wójcik pt. „ Ocena miar podobieństwa dokumentów tekstowych na potrzeby automatycznej analizy opinii konsumenckich”

promotor: Prof. dr hab. Paweł Lula,
promotor pomocniczy: dr hab. Grażyna Paliwoda-Pękosz, prof. UEK

1. Wprowadzenie i ocena formalna

Przyjmując obowiązki formalne recenzenta odwołuję się do ustawy z dnia 14 marca 2003 roku o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki ustanawiającej elementy podlegające ocenie tj.:
- zaproponowany temat i oryginalność rozwiązania problemu naukowego;
- ogólną wiedzę teoretyczną zaprezentowaną w rozważaniach objętych przedmiotem dysertacji;
- umiejętność samodzielnego prowadzenia pracy naukowej poprzez dokonanie analizy i syntezy materiału źródłowego oraz zaprojektowanie, przeprowadzenie i wnioskowanie z przeprowadzonego procesu badań empirycznych.

Od strony formalnej praca liczy 292 strony, przy czym cztery rozdziały liczą 264 strony. Oprócz czterech rozdziałów są jeszcze zakończenie, bibliografia, spis tabel, spis rysunków i aneksy. Objętość pracy należy uznać za poprawną, może nawet nieco za dużą. Kolejność rozdziałów jest prawidłowa z punktu widzenia prowadzonego wywodu. Doktorantka rozpoczyna od charakterystyki opinii konsumenckich i ich znaczenia, następnie charakteryzuje dokumenty tekstowe ze szczególnym uwzględnieniem miar podobieństwa po czym przechodzi do automatycznej analizy opinii konsumenckich, proponuje swój własny model takiej analizy i bada go w kilku eksperymentach.

Rozprawa jest poprawna od strony formalnej. Napisana jest językiem właściwym dla prac naukowych z odpowiednią terminologią statystyczną oraz z zakresu eksploracyjnej analizy tekstu. W całej pracy znalazłem około 30 literówek. Prawidłowe jest sporządzanie przypisów, opisu rysunków, tabel i wykresów, przejrzyste są reguły dotyczące przygotowania literatury.

2. Ocena wyboru tematu badawczego i oryginalności rozwiązania problemu naukowego

Problem badawczy tj. ocena przydatności różnych miar podobieństwa pomiędzy tekstami w kontekście ich wykorzystania w automatycznej analizie opinii konsumentów uważam za godny uwagi pomimo to, że literatura poświęcona problemowi pomiaru wspomnianego podobieństwa jest bardzo obszerna. Moim zdaniem problem jest bardzo złożony, wieloaspektowy i zapewne nie ma najlepszej miary podobieństwa tekstów, ale pomimo to wart badania, gdyż szybko zmieniająca się rzeczywistość ekonomiczna i handlowa pociągają za sobą zmieniające się formy komunikacji pomiędzy sprzedawcą a klientem. Zadaniem specjalistów od zarządzania jest tworzenie narzędzi informatycznych, które mogłyby sprostać coraz nowszym wyzwaniom, ewentualnie współpraca przy tworzeniu takich narzędzi.

Doktorantka cel zasadniczy pracy sformułowała następująco:
Zasadniczym celem niniejszej rozprawy jest prezentacja, klasyfikacja i ocena przydatności różnych miar podobieństwa pomiędzy tekstami w kontekście ich wykorzystania w automatycznej analizie opinii konsumentów.

Następnie, Doktorantka sformułowała kilka celów szczegółowych, z których za najciekawsze uważam dwa ostatnie tj.

- *Analiza procesu automatycznej analizy opinii konsumentów pod kątem zastosowania pomiaru podobieństwa tekstów na różnych jego etapach.*
- *Przeprowadzenie badań empirycznych zmierzających do oceny przydatności miar podobieństwa dokumentów tekstowych w automatycznej analizie opinii konsumentów.*

Doktorantka sformułowała również następujące dwie hipotezy:

- *Wykorzystanie wiedzy dotyczącej wydźwięku (nacechowania) wyrazów występujących w tekście opinii wpływa pozytywnie na jakość wyników automatycznej analizy opinii konsumentów.*
- *Za szczególnie przydatne podejście w analizie opinii konsumenckich należy uznać rozwiązanie bazujące na podejściu wzorowanym na metodach analizy wielokryterialnej opartych na koncepcji wzorca rozwoju (np. metoda TOPSIS), w których wzorzec i antywzorzec rozwoju budowany będzie w oparciu o informacje o wydźwięku wyrazów.*

W ten sposób określone cele rozprawy oraz hipotezy badawcze są według mnie wystarczająco oryginalne by można je było uznać za mające na celu poszerzenie wiedzy w przedmiocie badań. Ponadto, metody pracy nad problemem badawczym są w dużym stopniu oryginalne, zwłaszcza te związane z oprogramowaniem metod eksploracji tekstu polskojęzycznego oraz projektowaniem modelu automatycznej analizy opinii klientów o czym w dalszym ciągu recenzji.

3. Ocena ogólnej wiedzy teoretycznej zaprezentowanej w rozważaniach objętych przedmiotem dysertacji

Wiedzę teoretyczną Doktorantki zaprezentowaną w rozważaniach objętych przedmiotem dysertacji oceniam jako dobrą. Opinia ta odnosi się zarówno do części związanej z zarządzaniem, społeczeństwem informacyjnym, znaczeniem informacji wydobytej z opinii klientów jak i do części statystyczno-informatycznej. Należy zaznaczyć, że najlepszą orientację Doktorantka posiada w temacie narzędzi informatycznych, które mogą być wykorzystywane w analizie eksploracyjnej opinii konsumentów. Doktorantka sięgnęła po najnowsze opracowania naukowe pozwalające na prawidłową lematyzację słów z języka polskiego, co jest godne podkreślenia. W dalszej kolejności wymieniłbym wiedzę Doktorantki w temacie metod statystycznych obejmujących analizę skupień, na końcu zaś wiedzę z zarządzania i o społeczeństwie informacyjnym. Co prawda pozycje literatury ściśle związane z tematem dysertacji nie są najświeższe, najnowsze sprzed około ośmiu lat, związane z pracami statystyków ze środowiska krakowskiego, ale istotne jest to, że dysertacja jest poświęcona analizie eksploracyjnej tekstów w języku polskim. W literaturze przedmiotu można znaleźć wiele metod statystycznych i algorytmów komputerowych poświęconych grupowaniu dokumentów tekstowych, które, siłą rzeczy, nawet jeśli nie podkreślają swojego uzależnienia od mierzenia podobieństwa tekstów, to intensywnie takimi miarami posługują się. To są jednak metody pracujące głównie w języku angielskim, rzadko innym. Nie znam innych propozycji badania sentymentu tekstów polskojęzycznych poza tymi wymienianymi przez Doktorantkę.

Niektóre definicje oraz wywody teoretyczne z rozdziału pierwszego i drugiego można było pominąć, gdyż nie są one w żaden sposób wykorzystywane w dalszym ciągu dysertacji i są zbyt podstawowe. Doktorantka przytacza wiele definicji, często niekompatybilnych ze sobą i trochę brak jej własnego stosunku do nich. Niektóre fragmenty dotyczące charakterystyk liczbowych stosowanych w analizie skupień, np. kilka indeksów liczby skupień można było pominąć. Te indeksy nie są później używane, ponadto, w rozważanym przez Doktorantkę problemie liczba skupień jest na ogół znana, bo równa 2 lub 3. W przypadku grupowania dokumentów tekstowych mamy ten komfort, że można przerwać swoiste błędne koło analizy skupień i dostosować inne etapy tej analizy do znanej liczby skupień.

4. Ocena umiejętności samodzielnego prowadzenia pracy naukowej

Doktorantka podjęła się zanalizować, moim zdaniem, bardzo złożony i trudny problem analizy opinii klientów oraz zaproponować autorski model automatycznej analizy tych opinii. Problem jest złożony gdyż obejmuje zagadnienia natury lingwistycznej, w kontekście języka

polskiego bardzo trudne, o wiele bardziej zawiłe niż w przypadku języka np. angielskiego, zagadnienia z dziedziny zarządzania (interakcje pomiędzy sprzedawcą a klientem), zagadnienia ze statystyki oraz informatyki. W gąszczu tej złożonej problematyki Doktorantka umiejętnie prowadziła prace naukową i badawczą. Rozpoczęła od charakterystyk teoretycznych najpierw problemów zależności pomiędzy efektywnością sprzedaży a czytaniem i publikowaniem opinii klientów, następnie problemów związanych z matematycznymi modelami dokumentów tekstowych, przechodząc dalej do miar podobieństwa dokumentów tekstowych i badania znaczenia tych miar dla grupowania opinii klientów ze względu na ich sentyment. Moi zdaniem Doktorantka wykazała się zrozumieniem tych problemów co pozwoliło jej na ukoronowanie swoich badań zaproponowaniem modelu automatycznej analizy opinii klientów.

Można mieć uwagi do niektórych szczegółów procesu naukowo-badawczego.

Z pewnością uwagę zwraca olbrzymia praca włożona przez Doktorantkę na etapie lematyzacji polskich tekstów i wykorzystanie najnowszych zdobyczy nauki w tym temacie.

Uważam, że decyzja o skoncentrowaniu się na klasyfikacji bezwzorcowej tekstów czyli grupowania tekstów była słuszna. W dzisiejszej, niekiedy, bardzo szybko zmieniającej się rzeczywistości ekonomicznej nie ma uzasadnienia dla gromadzenia obszernych korpusów tekstów anotowanych. Przyczyn jest kilka. Jedna to wysoki koszt takich operacji, druga to brak pewności bezbłędnej anotacji pomimo wysokiego kosztu, trzecia to konieczność ciągłego uaktualniania korpusów. Konieczność ciągłego uaktualniania jest aż nadto oczywista w przypadku tematyki omawianej w pracy – każda nowinka techniczna w sprzęcie komputerowym czy fotograficznym powinna być wprowadzona do zbiorów uczących, co może oznaczać nawet kompletną wymianę korpusów tekstów.

Można spierać się o to czy wybór metody grupowania hierarchicznego był trafny. Ta metoda rozpoczyna grupowanie od możliwie największego rozproszenia korpusu tekstów (pojedynczych dokumentów) łącząc je następnie w większe skupienia. Grupowanie dokumentów tekstowych jest na tyle specyficznym zadaniem, że można wykorzystać bardzo wysokie podobieństwa pomiędzy niektórymi rodzajami dokumentów i od razu, w pierwszym kroku, tworzyć z nich większe skupienia, które będą w późniejszych fazach grupowania bardziej rozpoznawalne i stabilne. Grupowanie hierarchiczne po części spełnia ten postulat. Uważam, że aglomeracja hierarchiczna jest z pewnością lepsza od metod podziałowych (np. *k*-średnich) z zadaną z góry właściwą i małą liczbą skupień tj. 2 lub 3. Alternatywą do grupowania aglomeracyjnego mogłoby być grupowanie podziałowe, ale na pewno nie z tak małą liczbą skupień, następnie zaś aglomeracja wyznaczonych w pierwszym etapie niewielkich skupień. Częściowo tę uciążliwość łagodzi jeden ze wzorców grupowania rozpatrywany przez

Doktorantkę rozpoczynający się od podziału na rodzaje produktów.

Pewne zastrzeżenia może budzić optymalizacja macierzy częstości – usuwanie słów zbyt częstych, zbyt krótkich oraz, wcześniej, słów ze stoplisty. Jeżeli np. w zwrotach: *to jest do niczego* lub *to jest na nic* usuniemy przyimki *do, na,* to czy pozostanie jakakolwiek informacja? Czy takie zwroty tj. *do niczego, na nic,* znajdują się w słowniku wyrazów negatywnych?

Niewątpliwą zaletą dysertacji jest to, że wszystkie eksperymenty badawcze zostały dokładnie opisane i zilustrowane, co należy podkreślić.

Wymienione uwagi nie zmieniają oczywiście całościowej pozytywnej oceny pracy.

Pozwolę sobie sformułować następujące pytania do Doktorantki.

- Jak w procesie ETL wydobywane są zalety i wady produktu?
- Co, zdaniem Pani, było czy mogło być, przyczyną słabego podobieństwa podziałów na klasy dokumentów pozytywnych i negatywnych uzyskanych metodami analizy skupień oraz innymi, do podziałów opartych na informacjach obiektywnych?
- Czy, zdaniem Pani, eksploracyjna analiza tekstu pod kątem badania jego sentymentu oparta na metodzie bag-of-words ma przyszłość?

5. Ocena kwalifikacyjna i konkluzja

Zaprezentowane w pracy wyniki badań dotyczących eksploracyjnej analizy tekstu ze szczególnym uwzględnieniem wpływu miar podobieństwa tekstów na wyniki analizy sentymentu mają charakter oryginalny i dostarczają nowej wiedzy w tym zakresie oraz nowych narzędzi informatycznych. Doktorantka wykazała się dobrą wiedzą teoretyczną w tej dyscyplinie, umiejętnie wykorzystała rezultaty innych badań do zbudowania autorskiego modelu automatycznej analizy opinii klientów i ocenienia go we własnych badaniach empirycznych. W sposób przekonujący przedstawiła ich wyniki weryfikując sformułowane hipotezy i osiągając cel pracy. To świadczy pozytywnie o tym, że Doktorantka opanowała warsztat naukowy i posiada niewątpliwe zdolności samodzielnej pracy naukowej i badawczej.

Rozprawa doktorska mgr Katarzyny Wójcik pt. „Ocena miar podobieństwa dokumentów tekstowych na potrzeby automatycznej analizy opinii konsumenckich" stanowi oryginalne rozwiązanie problemu naukowego i spełnia w mojej ocenie wymagania określone w art. 13 ust. 1 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz stopniach i tytule w zakresie sztuki (Dz. U. 2016 r. poz. 882, 1311 ze zm.). Na tej podstawie występuję do Rady Dyscypliny Nauki o Zarządzaniu i Jakości Uniwersytetu Ekonomicznego w Krakowie o jej

przyjęcie jako pracy doktorskiej w dziedzinie nauk ekonomicznych, w dyscyplinie nauk o zarządzaniu oraz dopuszczenie do publicznej obrony.

dr hab. Jerzy Korzeniewski prof. UŁ