prof. dr hab. inż. Aleksander Byrski
Z-ca Dyrektora Instytutu Informatyki ds. Ogólnych
Wydział Informatyki, Elektroniki i Telekomunikacji
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Kraków 11.11.2021

Recenzja rozprawy doktorskiej mgr inż. Pavla Savova pt. Measuring the Novelty of Scientific Papers

W zalewie informacji i przy błyskawicznym wzroście dostępnych materiałów w sieci Internet obserwowanym na co dzień, automatyzacja analizy języka naturalnego jest jedyną rozsądną drogą, która może wspomagać użytkowników indywidualnych czy instytucje w pozyskiwaniu i weryfikacji różnego rodzaju informacji. Świat nauki jest nierozerwalnie związany z scjentometrią, która towarzyszy naukowcowi w zasadzie od samego początku pracy, a jej zastosowanie waha się od poszukiwania sprawdzonych publikacji w literaturze związanej z daną dyscypliną, do ewaluacji naukowców, którzy z przedmiotu stają się podmiotem zastosowania różnego rodzaju wskaźników na drodze awansu naukowego. Publikacje typu open-access, early-bird, mnóstwo pełnotekstowych treści dostępnych w serwisach takich jak Arxiv sprawia, że ich przeglądanie manualne w poszukiwaniu interesujących, wartościowych materiałów, powoli zaczyna być skazane na niepowodzenie.

Przedstawiona do recenzji praca próbuje dostarczyć metod i narzędzi, które mogą pomóc w radzeniu sobie z tą złożonością przez opracowanie dedykowanych rozwiązań koncentrujących się na predykcji daty publikacji i wykorzystaniu tego parametru do oceny innowacyjności, bazując wyłącznie na metodach przetwarzania języka naturalnego. Dlatego też proponowane podejście z założenia jest konstruowane jako metoda obiektywna, a właśnie obiektywizmu często brakuje w przypadku realizacji różnego typu procedur ewaluacyjnych w dzisiejszym świecie nauki.

Opracowana przez mgr inż. Pavla Savova rozprawa doktorska stanowi zbiór opublikowanych i powiązanych tematycznie artykułów naukowych:

- P. Savov, A. Jatowt, R. Nielek. "Identifying Breakthrough Scientific Papers." *Information Processing & Management* 57.2 (2020): 102168. (Impact factor: 4.787).

  W przedstawionym artykule autorzy przedstawiają wady klasycznych metod scjentometrii i proponują nową wykorzystującą klasyfikator predykujący lata publikacji bazując na analizy tematyki prac z wykorzystaniem metod z gatunku NLP. Autorzy proponują obliczanie współczynnika innowacyjności w celu identyfikacji przełomowych prac. Zaproponowane metody mają uzupełnić istniejące miary bazujące na liczbie cytowań. Propozycje zostają poddane testom w oparciu o dwa korpusy tekstu (Konferencja WWW oraz konferencja ADM SIGIR) zawierające ok.

3500 artykułów. Autorzy wskazują szczególnie ważne lata w badanych okresach oraz przedstawiają przykłady wyjątkowo wysoko cytowanych artykułów pozwalając na uzyskanie kompleksowego oglądu poziomu naukowego obydwu testowanych konferencji.

.

- P. Savov, A. Jatowt, R. Nielek. "Innovativeness Analysis of Scholarly Publications by Age Prediction using Ordinal Regression" *International Conference on Computational Science*, pp. 646-660. Springer, Cham, 2020, Konferencja CORE A.

  W tym artykule autorzy dopracowują metodę pomiaru innowacyjności artykułów naukowych. W oparciu o diachroniczny zbiór artykułów z określonej dziedziny nauki, publikowanych w okresie kilku lat, wyodrębniają ukryte tematy i trenują model oparty na regresji porządkowej (ordinal regression), w celu predykcji lat publikacji bazując na różnych rozkładach tematów. Błąd predykcji służy do obliczenia wyniku innowacyjności, który może być wykorzystany do uzupełnienia informacji podawanej przez liczbę cytowań danego artykułu w procesie identyfikacji przełomowych publikacji. Autorzy poprawiają wcześniej zdefiniowaną miarę innowacyjności która aktualnie uwzględnia również rok publikacji, dzięki czemu wyniki uzyskane dla artykułów opublikowanych w różnych latach są bezpośrednio porównywalne. Poprawiono również dokładność przewidywania zastępując klasyfikację regresją porządkową a modele typu Latent Dirichlet Allocation modelami Correlated Topic Models. Autorzy używają do badań korpusu 3577 artykułów opublikowanych na Międzynarodowej Konferencji World Wide Web (WWW) w latach 1994-1029 oraz 835 artykułów opublikowanych w Journal of Artificial Societies and Social Simulation (JASSS) w latach 1998-2019.

- P. Savov, A. Jatowt, R. Nielek. "Predicting the Age of Scientific Papers". *International Conference on Computational Science*, 2021. Konferencja CORE A.

  W artykule zaprezentowano metodę przewidywania wieku prac naukowych na podstawie diachronicznego zbioru artykułów z danej dziedziny publikowanych w określonym czasie. Autorzy po opracowaniu modeli bazujących na regresji przechodzą do zastosowania modelu BERT i agregują wyniki predykcji otrzymane dla poszczególnych zdań. Autorzy bazują na dwóch korpusach publikacji (International World Wide Web Conference i Journal of Artificial Societies and Social Simulation), porównują różne metody agregacji wyników i dowodzą, że podejście wykorzystujące poszczególne zdaniach wypada lepiej niż bezpośrednia metoda bazująca na przetwarzaniu całego dokumentu.

Wyżej wymienione artykuły zostały zacytowane in extenso w rozdziałach przedstawionego manuskryptu, który na początku został opatrzony 18-stronicowym przeglądem literatury tematu. Poza tym manuskrypt posiada zwyczajowe części takie jak wstęp, streszczenia, spis treści, listy rysunków czy skrótów a także podsumowanie. W sumie manuskrypt liczy 75 stron, bibliografia zaś zawiera 186 pozycji.

W motywacji do proponowanych badań Doktorant odnosi się do wad klasycznych podejść scientometrycznych, np. przytaczając argumenty, że cytowanie wysoko cytowanych artykułów prowadzi do ,,podążania za tłumem", wspomina tzw. efekt Mateusza, zwiększoną widoczność artykułów które na samym początku zostały wielokrotnie zacytowane, cytowanie klasycznych pozycji które są zwracane na początku zapytań choćby w Google Scholar, czy autocytowania. Trudno jest się nie zgodzić z przytoczonymi argumentami, dlatego opracowanie kreatywnych algorytmów czy metod postępowania pozwalających na obiektywną scjentometrię należy niewątpliwie do celów wartych poświęcenia pracy naukowej.

Proponowane przez Doktoranta rozwiązanie w.w. problemów współczesnej scjentometrii rozpoczyna się od porównywania predykowanego i rzeczywistego roku opublikowania danego artykułu. Predykcje realizowane są za pomocą modelu trenowanego pod nadzorem na bazie korpusu tekstu pochodzącego z artykułów. Doktorant proponuje wykorzystanie błędu predykcji jako środka do oceny innowacyjności artykułu (błąd świadczy o braku podobieństwa artykułu do innych z danego okresu - albo inaczej, podobieństwa do artykułów opublikowanych później - czy też na poziomie słownictwa czy tematyki). Proponowany sposób postępowania w rozdziale 3 bazuje na wykorzystaniu modelu SVM przy założeniu że rozkład tematów zbudowane w oparciu o model LDA zostaje zakodowany jako wektory cech i poddany klasyfikacji. Doktorant proponuje rzeczywistoliczbowy parametr oceniający innowacyjność artykułu. W kolejnym rozdziale Doktorant uaktualnia algorytm zamieniając komponent SVM na model regresji porządkowej a LDA na Correlated Topic Models. W końcu w rozdziale 5 Doktorant opisuje w jaki sposób dokładność predykcji roku publikacji może być usprawniona z wykorzystaniem modelu BERT. Przedstawiony model postępowania przez Doktoranta nie budzi zastrzeżeń, a cytowane in extenso publikacje stanowią rzeczywisty cykl tematyczny, co spełnia formalne wymagania stosownej ustawy odnośnie warunków koniecznych do starania się o przyznanie stopnia doktora.

Najważniejszym aspektem każdej pracy naukowej, a w szczególności pracy doktorskiej, jest kontrybucja do aktualnego stanu wiedzy wraz z jej uzasadnieniem (porównaniami ilościowymi i/lub jakościowymi z innymi podobnymi, dotychczas udostępnionymi metodami). W sekcji 1.7 autor stwierdza, że recenzowana rozprawa zawiera następujące kontrybucje do Informatyki:

- Algorytm umożliwiający datowanie publikacji danego tekstu bazując na *latent topic modelling* oraz *Support Vector Machines* (rozdzialy 3 i 4).
- Metodę obliczania współczynnika innowacyjności artykułu (rozdziały 3 i 4).
- Ulepszoną metodę datowania artykułów naukowych przy użyciu metod *word embedding* oraz *ordinal regression* (rozdział 5).

Doktorant podaje następującą metodę walidacji zaproponowanych przez siebie innowacyjnych elementów rozprawy:

- Doktorant bazuje na korpusie tekstu pochodzącego z artykułów z różnych źródeł: WWW Conference, ACM SIGIR Conference i czasopisma JASSS.
- Doktorant analizuje popularność tematów w korpusie tekstu (szczególnie jeśli te tematy, szczególnie jeśli te tematy są popularne przez kilka lat.

- Na bazie wspomnianego korpusu Doktorant predykuje rok publikacji na podstawie jego zawartości (analizując tekst), przyznając opracowany przez siebie współczynnik innowacyjności. Następnie agreguje wspomniane współczynniki identyfikując lata, w których opublikowane zostały przełomowe artykuły, w których poruszono po raz pierwszy tematykę popularną w przyszłości.
- Doktorant porównuje opracowaną metodę z liczbą cytowań i oblicza współczynniki korelacji.

Pewien niedosyt wywołuje zapoznanie się z metodą walidacji innowacyjnych elementów rozprawy. Należy pamiętać, że Doktorant stara się o uzyskanie stopnia naukowego doktora w dyscyplinie Informatyki Technicznej i Telekomunikacji, czyli w dziedzinie Nauk Technicznych, a nie w dziedzinie Nauk Społecznych. Badania naukowe prowadzone np w Naukach Społecznych mogą wykorzystywać metody informatyczne, traktując je w sposób aplikacyjny i doprowadzać do zdobycia nowej wiedzy w zakresie badania prawideł rządzących społeczeństwem, w tym przypadku społecznością naukowców. W rzeczywistości w przedstawionej rozprawie Doktorant stosuje opracowane przez siebie metody do przebadania innowacyjności artykułów, identyfikacji lat przełomowych, określenia korelacji swoich wyników z innymi metodami scjentometrycznymi, w szczególności z klasycznym podejściem opartym na analizie cytowań.

Z punktu widzenia Informatyki, koncepcja opracowanego algorytmu predykcji przez Doktoranta (złożenie wybranych metod z obszaru ML i NLP) nie budzi zastrzeżeń. Wspomniana wcześniej walidacja modelu powinna doprowadzić do odpowiedzi na pytanie, na ile odpowiada on rzeczywistości, jednak korelację między proponowanymi algorytmami predykcji a wynikami analizy cytowań, w zasadzie można interpretować z dwóch przeciwstawnych punktów widzenia:

- gdyby korelacja była wysoka, oznaczałoby to, że algorytmy proponowane przez Doktoranta dublują istniejące modele, innowacyjność proponowanego podejścia byłaby niewielka,
- skoro korelacja jest niska, trudno jednak mówić o walidacji modelu per se, natomiast zgodnie z interpretacją przedstawioną przez Doktoranta, świadczy to o tym, że proponowana metoda może uzupełnić istniejące metody scjentometryczne, co w zasadzie można traktować jako krok w kierunku spełnienia wymogu innowacyjności.

W pracy brakuje rzetelnego porównania i uzasadnienia wybrania tych a nie innych komponentów algorytmu proponowanego przez Doktoranta, miejscami jedynie zdawkowo Doktorant odnosię do poprawy skutecznosci podejścia (np w porównaniu do poprzedniej publikacji [176] str 71). Doktorant powinien przedstawić dokładną analizę skuteczności różnych podejść (choćby bazując na różnych wariantach algorytmu), wyniki porównać z punktu widzenia statystycznego, np. przeprowadzając testowanie hipotez statystycznych, tak aby przekonać czytelnika, że zestaw takich a nie innych komponentów jest zweryfikowany, przemyślany i zwalidowany, a nie np. bazujący na intuicji.

Dlatego też pozwalam sobie postawić następujące pytania, o odniesienie do których uprzejmie proszę Doktoranta w trakcie obrony:

- Dlaczego zostały wybrane akurat te a nie inne komponenty swoich algorytmów do datowania artykułów? Dlaczego akurat SVM, LDA, Ordinal Regression, Correlated Topic Models, BERT? Czy były przeprowadzone wstępne testy innych komponentów? Istnieje mnóstwo metod klasyfikacji i predykcji, streszczania tekstu czy identyfikacji tematyki (topic modelling), taką analizę należałoby przeprowadzić wcześniej dla szeregu metod, porównać ich efektywność i skuteczność, złożoność czasową i pamięciową.... Taka analiza mogłaby wnieść zdecydowanie bardziej wyraźną wartość dodaną do Informatyki, aktualnie praca ma charakter aplikacyjny a walidacja proponowanego modelu nie jest przeprowadzona do końca zgodnie z utartymi metodami znanymi w dziedzinie Nauk Technicznych, a w szczególności w Informatyce.
- Doktorant nie zamieszcza porównania z istniejącymi podobnymi algorytmami predykującymi rok publikacji (co byłoby bardzo pożądane z punktu widzenia określenia innowacyjności i walidacji), być może rzeczywiście nie ma dostępnych szeroko podobnych rozwiązań, kodów i wyników, aczkolwiek proponuję spróbować się odnieść np. do: Guo, S., Edelblute, T., Dai, B., Chen, M., & Liu, X. (2015). Toward Enhanced Metadata Quality of Large-Scale Digital Libraries: Estimating Volume Time Range. iConference 2015 Proc,, czy też dokumentu: https://www.hathitrust.org/files/PlaleChen-HTRC-Analytics.pdf. Może nie są to publikacje takiej rangi, jak zaprezentowane przez Doktoranta, ale wiążą się one bezpośrednio z dyskutowanym tematem i nawet jeśli nie ma tam szczegółów implementacji, czy takie szczegóły są niedostępne, można spróbować się odnieść jakościowo do samej koncepcji algorytmu.
- Opracowanie współczynnika innowacji - ta kontrybucja jest bezpośrednio związana z zastosowaniem metody w obszarze Nauk Społecznych, gdyby oceniać ją z punktu widzenia Informatyki należałoby znowuż zaproponować szereg różnych metod wyliczania współczynnika innowacji i przeprowadzić porównanie wraz z analizą statystyczną.
- Doktorant przedstawił różne warianty algorytmu predykcji daty artykułu - warto byłoby w sposób spójny zaprezentować porównanie ich efektywności i skuteczności, a nie zdawkowe odwołanie się do wspomnianego artykułu [176]. Skoro praca bazuje na trzech algorytmach, które z zasady nie powinny się pokrywać tematycznie, część wstępna pracy powinna zawierać spójną analizę porównawczą różnych wariantów algorytmu proponowanych przez Doktoranta i określanych jako główne kontrybucje do stanu wiedzy.

Przedstawienie odpowiedzi na w.w. pytania wymaga poparcie ich analizą choćby jakościową, bazując na przeglądzie i zacytowaniu relewantnych artykułów naukowych (swoich i obcych) i przedstawiając wyciągnięte z tego wnioski, natomiast nie wymagam przedstawienia wyników ilościowych gdyż zdaję sobie sprawę, że na przeprowadzenie nowych wyników badań eksperymentalnych może nie starczyć czasu, natomiast zdecydowanie tego typu podejście należałoby przewidzieć w przyszłych pracach.