INSTYTUT FIZYKI JĄDROWEJ
im. Henryka Niewodniczańskiego
POLSKIEJ AKADEMII NAUK

Dr hab. inż. Paweł Oświęcimka, prof. IFJ PAN Kraków, 20 listopada 2024 r.

Recenzja rozprawy doktorskiej

mgra inż. Artura Niewiarowskiego

pt. *Zastosowanie algorytmu odległości edycyjnej do ilościowej analizy danych tekstowych*

Uwagi wstępne

Rozprawa doktorska Pana mgra Artura Niewiarowskiego została przygotowana w Instytucie Podstawowych Problemów Techniki Polskiej Akademii Nauk w Warszawie. Promotorem rozprawy był dr hab. inż. Marek Stanuszek, profesor Politechniki Krakowskiej. Tematyka pracy dotyczy niezwykle aktualnego zagadnienia badawczego związanego z przetwarzaniem języka naturalnego, a dokładniej na propozycji nowego, efektywnego algorytmu porównywania tekstów literackich, w dużej mierze uniwersalnego ze względu na rodzaj języka, bez konieczności sprowadzania słów w tychże tekstach do form podstawowych lub rdzenia.

Aktualność podejmowanego tematu, ważna z punktu widzenia wyzwań technologicznych i społecznych, jest niejako oczywista w dobie cyfryzacji. Publikowanie cyfrowe z jednej strony zapewnia łatwy dostęp do zasobów, ale z drugiej niesie ryzyko nieuprawnionego ich kopiowania i wykorzystania. Problem plagiatu nie omija również środowiska akademickiego, stanowiąc istotne wyzwanie w kontekście opracowywania metod wykrywania takich działań, zwłaszcza w obszarze publikacji naukowych i prac dyplomowych. Szczególnego znaczenia nabiera to w dobie powszechnego dostępu do zaawansowanych modeli językowych, które potrafią generować i modyfikować tekst w sposób niemal niezauważalny.

W związku z tym prowadzone są intensywne badania nad rozwojem metod zdolnych do automatycznego wykrywania przypadków plagiatu. Systemy te opierają się na trzech kluczowych podejściach do analizy tekstu: leksykalnym, składniowym i semantycznym, angażując przy tym często zaawansowane algorytmy. Ze względu na złożoność problemu, nowoczesne

rozwiązania często łączą wspomniane podejścia, tworząc narzędzia, mogące identyfikować nadużycia bez konieczności ingerencji człowieka. Istotnym elementem tego typu badań jest rozwój metodologii przetwarzania języka naturalnego, co może znacząco wpłynąć na poprawę komunikacji między człowiekiem a komputerem, a także na bardziej efektywne wyszukiwanie i analizę określonych treści.

Nie mam zatem wątpliwości, że tematyka rozprawy, aktualna i niezwykle ciekawa, stanowi trudne wyzwanie dla badacza z punktu widzenia złożoności rozważanego problemu.

O treści rozprawy

Rozprawa doktorska Pana magistra Artura Niewiarowskiego ma charakter monografii i została oparta o wyniki badań przeprowadzonych przez Autora i przynajmniej w części opublikowanych w 7 czasopismach naukowych wymienionych w rozprawie.

Rozprawa została opisana na 167 stronach maszynopisu i podzielona na 7 rozdziałów: *Wstęp, Przegląd wybranych algorytmów analizy danych tekstowych, Implementacja algorytmu odległości edycyjnej w analizie podobieństwa zbiorów tekstowych, Weryfikacja przedstawionego mechanizmu analizy danych tekstowych, Podsumowanie, Literatura* oraz *Załączniki*.

Pierwszy rozdział rozprawy - *Wstęp* - zaczyna się od charakterystyki podjętego przez Doktoranta tematu badawczego, wraz z krótkim rysem historycznym i opisem aktualnego stanu badań, oraz jego znaczenia w przetwarzaniu języka naturalnego. Teza pracy, prezentowana w tym rozdziale, jest sformułowana jasno i jednoznacznie. Ten rozdział zakończony jest prezentacją zakresu prac wykonanych w celu obrony stawianej tezy. Brak mi jednak w tym rozdziale odniesienia do najbardziej aktualnych publikacji, co dawałoby pełniejszy obraz obecnego stanu wiedzy.

Drugi rozdział - *Przegląd wybranych algorytmów analizy danych tekstowych* - zawiera opis niektórych algorytmów wykorzystywanych do analizowania zbiorów tekstowych z opisem nie tylko ich zalet, ale i wad, co z punktu widzenia wyboru algorytmu może być szczególnie pomocne dla zainteresowanego czytelnika. W tym rozdziale Autor rozprawy wprowadza

również pojęcie odległości Levenshteina, wielkości która jest podstawą budowy proponowanego w następnym rozdziale algorytmu.

W rozdziale trzecim – *Implementacja algorytmu odległości edycyjnej w analizie podobieństwa zbiorów tekstowych* – Autor rozprawy wprowadza algorytm bazujący na odległości Levenshteina do analizowania ciągów tekstowych. Według mnie jest to zatem najważniejsza część pracy ze szczegółową prezentacją proponowanej metodologii analizy tekstów. Autor dyskutuje różne warianty użycia tego algorytmu, rozważając połączenie go z technikami lematyzacji i stemmingu, lub wzbogacenie algorytmu o słownik wyrazów bliskoznacznych, co pokazuje pewną elastyczność proponowanej metodologii. Starając się zbudować łatwy w interpretacji algorytm Autor wprowadza pojęcie analizy macierzowej tekstu, gdzie wynik działania algorytmu przedstawiony jest w postaci graficznej na siatce o rozmiarach analizowanych danych. W tym miejscu algorytm został wzbogacony o metodę detekcji grup punktów (w rozprawie analizowane są dwie takie metody), co zapewnia automatyzację procesu analizy przy minimalnej ingerencji człowieka. Ponieważ algorytm zawiera zestaw parametrów, które mogą zwiększyć jego czułość Autor pokazuje efekt ich użycia na szeregu przykładach, co ułatwia wychwycenie ich znaczenia dla końcowego wyniku. Co ważne z praktycznego punktu widzenia, Autor przeprowadził również testy wydajności metody, sugerując rozwiązania poprawiające ten wskaźnik.

Rozdział czwarty - *Weryfikacja przedstawionego mechanizmu analizy danych tekstowych* - zawiera testy proponowanego algorytmu na szeregu przykładach w których oryginalne teksty zostały przetłumaczone na inny język za pomocą internetowych tłumaczy. Do testów użyto popularnych translatorów internetowych takich jak Google Translate, DeepL czy Microsoft Translator oraz bardzo popularnego ostatnio modelu językowego ChatGPT. Co istotne, Autor rozważa różne języki oryginalnego i przetłumaczonego tekstu pokazując uniwersalność proponowanej metodologii. Ponadto, w rozdziale zaprezentowano analizę tekstów wygenerowanych przez ChatGPT starając się odpowiedzieć na pytanie czy proponowana metoda jest w stanie zidentyfikować wspólne źródło pochodzenia tekstu. Wyniki prezentowane w tym rozdziale zostały wygenerowane przez program o nazwie N-DMS Antyplagius w którym proponowane podejście do analizy tekstu zostało zaimplementowane przez mgra

Niewiarowskiego. Niestety o autorstwie programu dowiadujemy się dopiero z *Podsumowania* , co niezbyt wytrwałemu czytelnikowi może ujść uwadze.

Rozdział piąty – *Podsumowanie* – zawiera zebrane konkluzje przeprowadzonych testów, które według Autora potwierdzają sformułowaną w rozdziale pierwszym tezę. Autor oprócz syntetycznego przedstawienia wniosków z analizy przedstawia również perspektywy rozwoju swojej metody oraz planuje dalsze analizy. Takie podejście do tematu charakteryzuje dojrzałego badacza z jasno zarysowaną tematyka naukową. Znaczna część tego rozdziału poświęcona została programowi N-DMS Antyplagius stworzonemu przez Autora rozprawy. Moim zdaniem, opis działania tej aplikacji powinien znaleźć się we wcześniejszych rozdziałach rozprawy.

Rozdział szósty – *Literatura* – zawiera 57 pozycji literaturowych w tym 7 pozycji autorstwa Doktoranta, w których jest On pierwszym autorem.

W tym miejscu należy podkreślić, że Doktorant podjął się w swojej pracy trudnego zadania jakim jest stworzenie metody zdolnej do oceny podobieństwa tekstów, a jednocześnie w dużej mierze uniwersalnej ze względu na rodzaj analizowanego języka. Warto podkreślić, że zaproponowany algorytm może być użyty w systemach antyplagiatowych, co w dobie współczesnej cyfryzacji wydaje się niezwykle istotne z punktu widzenia ochrony praw autorskich czy nadużyć przy generowaniu tekstów przez sztuczną inteligencję. W celu przetestowania proponowanej metodologii Autor rozprawy przeprowadził szczegółowe testy metody dla różnych języków o korzeniach europejskich oraz automatycznych systemów do generowania tekstów opartych na sztucznej inteligencji. Niewątpliwie, takie udokumentowanie skuteczności metody uwiarygadnia w znacznym stopniu proponowaną metodologię badań.

Pytania dotyczące rozprawy

Część uwag została już podniesiona we wcześniejszych rozdziałach mojej recenzji. Niemniej jednak chciałbym usłyszeć jak Autor ustosunkowuje się do poniższych zagadnień.

1. Autor w rozprawie przedstawia zalety proponowanego algorytmu, jednak brakuje podsumowywującej dyskusji na temat jego ograniczeń. W jakich warunkach proponowana metodologia mogłaby prowadzić do wyników sugerujących błędne wnioski? Jakie są potencjalne wady przedstawionego

algorytmu? Czy można określić krytyczne wartości parametrów, które mogłyby skutkować błędnymi wynikami? Co ciekawe, przy omawianiu popularnych algorytmów autor wymienia ich wady, co tym bardziej zwraca uwagę na brak analogicznej analizy w przypadku własnego rozwiązania.

2. Na stronie 37, Autor pisze „ *Reasumując, badania wykazały*,…". O jakie konkretnie badania tutaj chodzi? Czy o analizy przeprowadzone w ramach rozprawy?
3. Proszę o wyjaśnienie testu opisanego na stronie 57, a w szczególności: „…*Zastosowano w nim do analizy dwa teksty napisane w dwóch językach: portugalskim i hiszpańskim."* Czy do analiz użyto dwóch różnych tekstów, czy może były to tłumaczenia jednego na drugi?
4. Proszę o wyjaśnienie co znaczy z naukowego punktu widzenia określenie: teksty „niosące to samo przesłanie"? (str. 57).
5. Mam uwagi dotyczące opisu wzorów w rozprawie. Przy opisie podstawowego wzoru dotyczącego odległości Levenshteina (2.8) pojawia się pojęcie macierzy Levenshteina, które wyjaśnione jest dopiero niżej (razem z parametrami *m* i *n*) co budzi pewną trudność przy zrozumieniu prezentowanych formuł. Według powszechnie stosowanej praktyki wyjaśnienie użytych we wzorach wielkości powinno pojawić się bezpośrednio pod wzorem. To samo dotyczy podstawowego dla zrozumienia pracy wzoru (3.3) – czym jest *q*? Prawdopodobnie pomylono tutaj oznaczenia *q* z *bp*. Natomiast na stronie 62 widzę parametry *P* i *R* wyjaśnione niestety dopiero na stronie 63, tym razem bez symboli.
6. W spisie literatury znalazłem jedynie dwie pozycje z datą po 2020 roku (w tym jedną autorstwa Doktoranta) odnoszące się do aktualnych badań nad analizą tekstów, co sugerowałoby że tego typu działalność naukowa nie jest interesującą tematyką badawczą. Dlaczego nie uzupełniono literatury, a przez to nie zaprezentowano najbardziej bieżącego stanu wiedzy?
7. Bardzo ciekawym zagadnieniem jest analiza tekstów generowanych przez automatyczne systemy takie jak ChatGPT. Czy Autor rozważał porównanie tekstów generowanych przez różne systemy?
8. Wydaje mi się, że ocena jakości tłumaczenia byłaby również możliwa z użyciem proponowanej metodologii. Czy Autor rozważał przeprowadzenie takich analiz w przyszłości?

Uwagi redakcyjne

Rozprawa została napisana w języku polskim. Sama struktura pracy jest w mojej ocenie przejrzysta co pozwala dość łatwo zaznajomić się z tematyką pracy jak i jej myślą przewodnią. Niestety, z uwagi na obowiązek recenzenta muszę zwrócić uwagę na dużą liczbę błędów gramatycznych, stylistycznych oraz błędnego formatowania tekstu w rozprawie, które w mojej ocenie nie powinny były się zdarzyć. Poniżej wymieniam jedynie kilka z nich:
str. 11 – „pacy”,
str. 29 – brak odnośnika w tekście do tabeli 2.4, 2.5, 2.6,
str. 56 – „tymi języki”,
str. 65 – „zasttosowane”, oraz zły format zmiennych *gw* oraz *wv*,
str. 128 – „algorytmach ksploracji”.

Uwagi końcowe

Przedstawiona mi do recenzji rozprawa doktorska mimo krytycznych uwag, stanowi oryginalne rozwiązanie problemu naukowego, podejmowana w rozprawie tematyka jest interesująca, a uzyskane wyniki stanowią istotny wkład do rozwoju dziedziny. Na podstawie pracy można stwierdzić, że Autor rozprawy wykazał się dobrą znajomością podejmowanej tematyki badawczej, umiejętnością formułowania problemów badawczych, przeprowadzania samych badań oraz formułowania wniosków.

Stwierdzam zatem, że rozprawa doktorska mgra Artura Niewiarowskiego spełnia wymogi stawiane rozprawom przez ustawę z dnia 14 marca 2003 r. *o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki*, i wnoszę o dopuszczenie rozprawy do publicznej obrony i dalszych etapów przewodu doktorskiego w dyscyplinie Informatyka Techniczna i Telekomunikacja.