Recenzja rozprawy doktorskiej mgr. inż. Jana Sawickiego „Exploring the Information Structure of Reddit Through Natural Language Processing and Network Analysis”

prof. UAM dr hab. Filip Graliński
Wydział Matematyki i Informatyki
Uniwersytetu im. Adama Mickiewicza w Poznaniu

10 lutego 2024

1 Wstęp

Celem niniejszej recenzji jest stwierdzenie, czy rozprawa doktorska mgr. Jana Sawickiego „Exploring the Information Structure of Reddit Through Natural Language Processing and Network Analysis” spełnia wymagania Ustawy Prawo o szkolnictwie wyższym i nauce (z dnia 20 lipca 2018 r. z późniejszymi zmianami).

Podstawowym zagadnieniem badawczym rozprawy są techniki analizy struktury informacyjnej serwisu Reddit.

2 Ocena formalnej strony rozprawy

2.1 Ocena układu pracy

Zasadniczą część pracy stanowi obszerny wstęp oraz cykl siedmiu artykułów naukowych (współ)autorstwa doktoranta. Układ tej, najistotniejszej, części nie budzi większych zastrzeżeń. Niestety całościowo rozprawa charakteryzuje się pewną chaotycznością: brak spisu treści i jednolitej numeracji stron (!, zestawienia na s. 22, 42 i 43 stanowią tylko częściowy substytut), dwa artykuły („Fusing individual choices into a group…" i „VisQualdex: a comprehensive…"), jak się domyślam, wyłącznie przez pomyłkę nie znalazły się w dodatku A.

Zastrzeżenia budzi też fizyczna forma pracy, różny format oryginalnych artykułów jest zrozumiały, ale w niektórych fragmentach sprawiał trudności w czytaniu (mały font).

2.2 Ocena stosowanego aparatu naukowego

Autor stosuje standardowy w pracach z przetwarzania języka naturalnego i uczenia maszynowego aparat naukowy, sprawnie odwołując się zarówno do teorii (uczenie nienadzorowane, teoria grafów), jak i narzędzi praktycznych (zanurzenia, modele języka Transformer *encoder-only* typu BERT).

Zastanawiać może brak odwołań do dużych modeli języka typu *decoder-only*, czy popularnych wcześniej modeli koder-dekoder (T5). We wcześniejszych pracach jest to zrozumiałe, duże modele języka dopiero powstawały, w późniejszych (rok 2024) ten brak może zastanawiać. Z drugiej strony można bronić autora tym, że w eksploracji danych tekstowych i uczeniu nienadzorowanym do niedawna modele typu BERT rzeczywiście stanowiły *state of the art*. Tak czy owak szkoda, że autor nie odniósł się do tego we wstępie.

2.3 Ocena spójności terminologicznej

Spójność terminologiczna nie budzi zastrzeżeń. Kluczowe pojęcia są w pracy zdefiniowane we właściwy sposób. Niefortunnym wyjątkiem jest pojęcie *wykrywania społeczności* (*community detection*), które nie zostało wprowadzone w klarowny sposób, brakuje omówienia różnic między wykrywaniem społeczności a analizą skupień (klastrowaniem).

2.4 Ocena stosowanego języka

Praca jest napisana klarownym i zrozumiałym językiem. Autor nie komplikuje niepotrzebnie języka, ale z drugiej strony unika zbytnich uproszczeń.

Autorowi zdarzają się błędy językowe czy niezręczności stylistyczne. Wymieńmy kilka przykładów: *emebddings* (s. 13), *it's* amiast *its* (dwa razy na s. 41), usterki interpunkcyjne (A3: *„analysis , ”*, s. 1131 *„score , ”*, s. 1132, niezamknięty nawias, s. 1132), *test* w algorytmie w pracy A4, *researchresearch* (A5, s. 1), *built from the word* (A5, s. 3), *Figure 2 of [?]* (A7, s. 69).

3 Ocena uzyskanych wyników

3.1 Uwagi ogólne

Spójność pracy nie budzi wątpliwości. Autor skupia praktycznie całą swoją uwagę na badaniu serwisu Reddit, jego własności w wymiarze „makro” dochodząc przy tym do interesujących ustaleń również w wymiarze „mikro”.

Chciałbym w tym miejscu podkreślić wagę badań nad serwisami społecznościowymi, takimi jak Reddit, co, jak myślę, może być niedoceniane czy nawet, zupełnie niesłuszne, szufladkowane jako „przyczynkarstwo”. Wagę „redditologii” można uzasadnić dwojako. Po pierwsze Reddit jest olbrzymim źródłem danych i metadanych, co może zostać wykorzystane praktycznie. Wystarczy wspomnieć fakt, że sukces modelu języka GPT-2, bardziej niż na architekturze i sposobie trenowania, zasadzał się na sprawnym wyzyskaniu

metadanych z Reddita — twórcy modelu opracowali zbiór WebText pobierając strony wskazywane w postach serwisu o ocenie wyższej niż 3 (do tych ocen wielokrotnie odwołuje się też autor recenzowanej rozprawy). Swoją drogą wielka szkoda, że Jan Sawicki nie przywołał tego faktu w rozprawie.

Drugie uzasadnienie badań nad serwisem Reddit: jest to byt, chciałoby się nawet powiedzieć, *organizm*, z jednej strony fascynujący sam w sobie, z drugiej wywierający olbrzymi — pozytywny i negatywny — wpływ na życie jednostek i całych społeczeństw, wpływ porównywalny ze znaczeniem instytucji takich jak uniwersytety czy środki masowego przekazu, a zatem również powinien być przedmiotem refleksji naukowej. Rozprawa Jana Sawickiego znakomicie się wpisuje w ten imperatyw, szczególnie że autor wychodzi poza badanie serwisu Reddit jako „masy" tekstu, rozpatrując również aspekty strukturalne i diachroniczne.

Przejdę teraz do omówienia wyników badań opisanych w siedmiu artykułach cyklu, po czym dokonam podsumowania. Zaznaczę tylko w tym miejscu, że prace wyraźnie rozpadają się na trzy części: metaanaliza piśmiennictwa naukowego (dwie pierwsze prace), analiza struktury serwisu Reddit (cztery kolejne prace) i badanie ewolucji serwisu w czasie (ostatnia praca, zagadnienia ewolucji Reddita znajdują odzwierciedlenie również w poprzedniej). Ten wyraźny trójpodział absolutnie nie oznacza niespójności pracy, wręcz przeciwnie cała praca stanowi całość, spiętą w naturalny ciąg dociekań (i zdecydowanie wyróżnia się *in plus* w stosunku do wielu innych rozpraw budowanych na cyklach artykułów).

3.2 Ocena pracy „Exploring Usability of Reddit in Data Science and Knowledge Processing"

W tej części autor dokonuje starannej i wnikliwej (choć może niezbyt obszernej) metaanalizy literatury naukowej odwołującej się do serwisu Reddit. Widać, że praca ta stanowiła „preludium" do całokształtu badań podejmowanych w czasie studiów doktorskich i zainspirowała kierunki dalszych badań. Konkretne odkrycia dokonane przez autora są interesujące, choć może nie

mają dużego ciężaru gatunkowego: wzrost liczby artykułów w serwisie Reddit w czasie epidemii COVID-19, problemy z mierzeniem subiektywności, to, że Reddit często badany był razem z serwisem Twitter.

3.3 Ocena pracy „The state of the Art of Natural Language Processing — A Systematic Review of NLP Literature Using NLP Techniques"

Podobnie jak poprzednia praca, również i ta ma charakter metanalizy, tym razem obszerniejszej, bo dotyczącej całokształtu zagadnienia przetwarzania języka naturalnego. Należy zaznaczyć, że praca ta w najmniejszym stopniu dotyczy serwisu Reddit — nie stanowi to problemu z punktu widzenia spójności pracy, po prostu w tych dwóch pierwszych pracach autor przygotowuje się do zasadniczej części badań, idąc z dwóch kierunków: w pierwszej pracy wychodząc z literatury analizującej Reddita, w drugiej — idąc z pola problemowego przetwarzania języka naturalnego (przypomnijmy zasadniczą część tytułu rozprawy *Exploring the Information Structure of Reddit Through Natural Language Processing*). Dodajmy, że fakt, że tak bogaty korpus jak Reddit nie był w ścisłej czołówce zbiorów danych wykorzystywanych w NLP, wskazywał na lukę badawczą, wykorzystaną przez autora w całej rozprawie.

Metaanaliza wykonana w pracy jest staranna i wyczerpująca. Ważne wnioski to: zaskakująco duże znaczenie języka czeskiego, istotna obecność widzenia komputerowego w NLP, istnienie dużej grupy wzajemnie cytujących się prac.

Autor stwierdza, że „Transformers are connected strongly with attention", co jest raczej trywialnym wnioskiem (być może wynikającym z odniesień do pracy „Attention is all you need" wprowadzającej architekturę Transformer). Z drugiej strony brakuje czasami pogłębionej refleksji, na przykład dlaczego parsowanie zależnościowe jest bardziej popularne dla języków takich jak niemiecki, francuski czy czeski (tradycja badań? własności języków?)?

3.4 Ocena pracy „Text embeddings and clustering for characterizing online communities on Reddit”

Autor dokonuje w omawianej pracy automatycznego grupowania tzw. subredditów korzystając z zanurzeń tekstowych. Pod względem wypracowanych metod praca jest najmniej odkrywcza spośród czterech prac z drugiej grupy, z drugiej strony praca obfituje w interesujące konkretne wnioski. Na pewno doktorant wykazuje się znajomością tematu, co pokazuje przegląd literatury (sekcja II) i omówienie klastrowania (punkt III.C).

Najsłabszy punkt tego rozdziału to niejasne przedstawienie, jak powstają zanurzenia (tytuły? treści postów? osobno każdy post?) w autorskiej metodzie. Można mieć też zastrzeżenia metodologiczne do tego, jak powstały „ręcznie oznaczone kategorie” („manually annotated categories”).

3.5 Ocena pracy „Reddit CrosspostNet—Studying Reddit Communities with Large-Scale Crosspost Graph Networks

W tej pracy autor wprowadza nowatorską metodę wyznaczania wspólnot subredditów za pomocą analizy tzw. crosspostów. Zbudowany w ten sposób graf pozwolił autorowi dokonać interesujących obserwacji, szczególnie ważne są te z punktu 5.2, autor pokazuje, że uzyskana struktura pozwoliłoby rozszerzyć zakres wcześniejszych prac dotyczących serwisu Reddit. Mniej interesujące są punkty 4.3 i 5.1 (zwłaszcza jego początek). Odnoszę wrażenie, że autor nie dochodzi do żadnych wniosków godnych uwagi w tym fragmencie.

3.6 Ocena pracy „Applying Named Entity Recognition And Graph Networks to Extract Common Interests from Thematic Subfora on Reddit”

Autor dorzuca kolejne narzędzie do arsenału środków umożliwiających badanie serwisu Reddit, tym razem oparte na jednostkach nazwanych. Przegląd

istniejących metod (sekcja 2) stanowi mocny punkt pracy. Autor wykazuje się tutaj dogłębną znajomością tematu. Pewne zastrzeżenie należałoby poczynić przy omawianiu zbioru CoNLL-2003 (dla zadania NER), jest to zbiór już przestarzały, znane są jego słabe punktu. Pod koniec punktu 2.2 autor dobrze uzasadnia potrzebę sięgnięcia po nowe metody. Niestety nie do końca klarowny jest opis zaproponowanej metody (sekcja 4), zwłaszcza punkt 7. Zrozumienie przedstawionego podejścia wymaga dużego wysiłku.

Pozytywnie należy ocenić sposób ewaluacji przy użyciu crosspostów (co stanowi nawiązanie do poprzedniego artykułu). Pojawia się jednak pytanie, dlaczego autor nie sprawdził prostego punktu odniesienia (*baseline*), np. opartego na liczeniu przecięcia zbiorów jednostek nazwanych obu subredditów.

W każdym razie uzyskane konkretne wyniki, np. przedstawione w tabeli 3, pozwalają na nowatorskie spojrzenie na zależności między subredditami.

3.7 Ocena pracy „Exploring Reddit Community Structure: Bridges, Gateways and Highways”

Najistotniejszym elementem tej pracy jest dobrze przeprowadzona, i słuszna moim zdaniem, krytyka pojęć „mostów” i „wrót” (*bridges/gateways*) stosowanych w analizie struktury serwisu Reddit. Autor w ich miejsce wprowadza koncept „autostrady” (*highways*).

Według metody autora subreddity r/Sims4 i r/StardewValley (czyli dotyczące gier komputerowych) znalazły się we wspólnocie z subredditami o tematyce związanej ze stylem życia. Moim zdaniem, autor zbyt szybko przechodzi nad tym do porządku dziennego, można podejrzewać, że ten fakt jest efektem zastosowania takich a nie innych zanurzeń.

3.8 Ocena pracy „Application of Natural Language Processing and Temporal Networks to Analysis of Evolution of Reddit Communities”

Tematyka ewolucji serwisu Reddit znajdowała cząstkowe odzwierciedlenie w poprzednich pracach, w tej jednak pracy autor poświęca jej całość uwagi.

Tym razem, w przeciwieństwie do poprzednich prac, metody zastosowane przez autora (sekcja 3) zostały opisane w bardzo klarowny sposób. Najciekawszy wydaje mi się punkt 4.4, w którym autor przedstawia, można by rzec, efemerydalne połączenia między subredditami; szkoda że nie podążył dalej w tym kierunku. Wyniki uzyskane w punktach 4.2 i 4.3 są również frapujące, pojawia się jednak wątpliwość, czy nie są one prostym efektem powiększania się serwisu Reddit?

3.9 Podsumowanie

Rozprawa zasługuje na pozytywną opinię, autor bada strukturę serwisu Reddit, sprawnie używając różnych narzędzi zarówno teoretycznych, jak i praktycznych. W swoich badaniach autor doszedł do wielu ciekawych i oryginalnych wniosków, począwszy od bardzo ogólnych, a skończywszy na wielu interesujących mikroobserwacjach, które mogą być inspirujące dla kolejnych badaczy tego serwisu. W pełni zgadzam się ze stwierdzeniem autora, że „[the] presented Dissertation contains a comprehensive analysis of reddit information structure”.

Szkoda, że autor nie pokusił się o głębszą refleksję i bardziej zasadnicze porównanie różnych metod identyfikacji skupień i wspólnot. Należało podjąć pytanie, co to znaczy, że taki a nie inny klastrów/wspólnot jest „dobry” (czy choćby użyteczny). Cząstkowe analizy idące w tym kierunku przedstawione w pracy „Applying Named Entity Recognition And…” pozostawiają niedosyt.

W szczególności większego przemyślenia wymagałaby też kwestia, jak bardzo połączenia „intencjonalne” w grafie Reddita (np. crossposty) odróżniają się od połączeń „nieintencjonalnych” (podobna treść, użycie tych sa-

mych jednostek nazwanych) i czy uzyskane w ten sposób grafy są zasadniczo odmienne.

Mimo tych zastrzeżeń pracę oceniam pozytywnie.

We wszystkich artykułach naukowych, które składają się na rdzeń pracy, doktorant jest pierwszym autorem. W jednej z prac mgr inż Sawicki jest jedynym autorem, w pozostałych — jednym z 2-4 współautorów, zakres wkładu doktoranta i jego kluczowa rola nie budzą wątpliwości, tym bardziej że cykl artykułów spaja spójna tematyka. Artykuły ukazały się w punktowanych czasopismach i materiałach konferencyjnych, wprawdzie nie tych pierwszorzędnych, ale znowu należy podkreślić, że spójność artykułów wzmacnia ich wagę.

4 Wiedza kandydata

Mgr inż. Jan Sawicki wykazał się w rozprawie ogólną i szczegółową wiedzą teoretyczną w przedmiotowej dyscyplinie naukowej.

Badania przeprowadzone przez autora zdecydowanie potwierdzają umiejętność samodzielnego prowadzenia pracy naukowej.

Po przeczytaniu rozprawy, z całym przekonaniem stwierdzam, że kandydat prezentuje ogólną wiedzę w dyscyplinie informatyka techniczna i telekomunikacja.

5 Wniosek

Rozprawa nie jest pozbawiona pewnych słabości. Mimo wszystko badania mgr. inż. Jana Sawickiego stanowią istotny wkład w zakresie zagadnień na pograniczu przetwarzania języka naturalnego, uczenia nienadzorowanego i nauki o sieciach. Rozprawa stanowi oryginalne rozwiązanie problemu naukowego.

Stwierdzam, że recenzowana rozprawa mgr. inż. Jana Sawickiego spełnia wymagania stawiane pracom doktorskim przez Ustawę Prawo o Szkolnictwie

Wyższym i Nauce z dnia 20 lipca 2018 r. (z późniejszymi zm.).

Z przekonaniem wnioskuję o dopuszczenie Doktoranta, pana mgr. inż. Jana Sawickiego, do publicznej obrony jego rozprawy doktorskiej.