Gdańsk, 10.03.2025 r.

Dr hab. Anna Bączkowska, prof. UG
Instytut Anglistyki i Amerykanistyki
Zakład Językoznawstwa Teoretycznego i Komputerowego
Uniwersytet Gdański
e-mail: anna.baczkowska@ug.edu.pl

Recenzja rozprawy doktorskiej pt.

„Exploring the Information Structure of Reddit

Through Natural Language Processing and Network Analysis"

na podstawie serii artykułów przedstawionych do recenzji przez

mgr inż. Jana Sawickiego

Uwagi ogólne

Doktorant przedstawił do oceny serię artykułów powiązanych tematycznie, eksplorujących wielorakie aspekty dotyczące danych tekstowych pozyskiwanych z platformy Reddit oraz sposobu analizy tych danych i ich wizualizacji. Siedem artykułów Autor uznał za podstawowe i dodatkowo pięć artykułów dołączył jako uzupełniające. W sumie do recenzji przedłożono zatem 12 artykułów, z czego jeden to publikacja jednoautorska, a jedenaście napisanych jest we współautorstwie; w sześciu z nich Doktorant jest pierwszym autorem.

Problem badawczy i metodyka badań

Głównym problemem badawczym jest proponowana metodyka identyfikacji typów interakcji zachodzących pomiędzy społecznościami internetowymi (ang. *communities*) i subredditami (subforami) na platformie społecznościowej Reddit. Opracowano kilka nowatorskich metod śledzenia rozwoju subredditów tworzonych przez społeczność internetową na przestrzeni lat oraz rozpoznawania i wizualizacji komunikacji pomiędzy użytkownikami i społecznościami na ww. platformie. Pozwoliło to na wgląd w dynamikę rozwoju społeczności na Reddicie oraz zaobserwowanie i zdefiniowanie istniejących

relacji tematycznych, a także na rozpoznanie ewoluujących zmian tematycznych subredditów i ich stopniową konsolidację. W opisanych badaniach modelowania struktury informacyjnej Reddita Doktorant korzysta z technik sieci neuronowych, w szczególności z transformerów BERT (kontekstowych/tekstowych osadzeń/zanurzeń, różnych technik używanych w modelowaniu tematycznym, w tym BERTopic) i jego uproszczonej, lecz bardziej efektywnej wersji DistilBERT, a także z algorytmu Leuven używanego do detekcji społeczności internetowych w mediach społecznościowych i identyfikacji relacji między nimi (w kontraście do metody Leiden).

Oryginalność badań

Oryginalność badań Doktoranta można krótko scharakteryzować w następujący sposób:

- zaproponowanie analizy interakcji między subredditami w oparciu o NER i sieci grafowe, a także tzw. crossposty (publikacje krzyżowe);

- identyfikacja crosspostów, jako zjawiska typowego tylko dla platformy Reddit, które we wcześniejszych badaniach były pomijane, a które wskazują nie tylko na interakcje zachodzące pomiędzy użytkownikami w obrębie subredditów, ale też mogą stanowić miarę siły połączeń między subredditami;

- zaproponowanie tzw. „autostrad" przepływu informacji (ang. *highways*) w miejsce znanych wcześniej „mostów" i „bram" (ang. *bridges*, *gateways*) w architekturze informacji na platformie Reddit, co wynika z zaobserwowanego zjawiska nakładania się „mostów" i „bram";

- zaproponowanie modelowania tematycznego w oparciu o połączenie 3 wcześniej używanych technik modelowania (LDA, NMF i BERTopic).

Trzeba podkreślić, że badanie interakcji między subredditami poprzez identyfikację nazw własnych (NER) nie było wcześniej standardową metodą używaną do tego typu analiz. Zastosowanie NER pozwala w znacznym stopniu zredukować dane wyjściowe, co ułatwia tworzenie sieci grafowej z uwagi na zmniejszoną liczbę węzłów. Ważne jest również, że zaproponowano nowatorskie rozwiązanie „autostrad" dla – jak zauważono – generujących redundantne informacje „mostów" i „bram". Z kolei wcześniej ignorowane zjawisko crosspostów zostało zbadane i doprecyzowane.

Aplikacyjność badań w innych dyscyplinach

W przedłożonych do recenzji artykułach zaproponowano szereg interesujących rozwiązań z zakresu NLP, które mają bezpośrednie przełożenie na badania w innych dziedzinach nauki, w szczególności z zakresu językoznawstwa i analizy dyskursu, a także medioznawstwa, politologii i socjologii. Niektóre z potencjalnych zastosowań wymieniłam poniżej. Publikacje są wartościowe dla badaczy innych proweniencji również z uwagi na klarowność opisywanych technik komputerowych, co może zachęcać potencjalnych badaczy, w szczególności językoznawców i medioznawców, do rozszerzenia metodologii badań typowych w danej dyscyplinie o nowe, innowacyjne i mało znane w tych dyscyplinach techniki komputerowe.

Uwagi szczegółowe dotyczące poszczególnych artykułów

W artykule „Exploring usability of Reddit in data science and knowledge processing" Autorzy podejmują się ważnej kwestii, z punktu widzenia przetwarzania języka naturalnego, pozyskiwania danych tekstowych do analiz językowych z platformy Reddit. W szczególności istotne są narzędzia używane do przeglądania zasobów pod względem tematyki postów oraz do analizy danych. Interesującą obserwacją jest to, że tematyką dominującą w badaniach tekstów umieszczanych na platformie Reddit jest dyskurs obraźliwy i mowa nienawiści (tzw. *hate speech*), choć nie jest to obserwacja zaskakująca, biorąc pod uwagę ostatnie zainteresowania językoznawców badających dyskurs w mediach społecznościowych. Autorzy wymieniają co prawda jako najpopularniejszy temat „conversation analysis" (CA), „hate speech" jest drugi w rankingu popularności, jednak trzeba zauważyć, że CA to nie jest tematyka postów, a metodologia używana w językoznawstwie (w pragmatyce językowej i analizie dyskursu). Podobnie, „conspiracy theories" to popularna rama teoretyczna używana często w medioznawstwie. Zestawienie tych pojęć jako przykładów na popularne tematy poruszane na Reddicie nie ma więc uzasadnienia. Bardzo interesujące są metody komputerowe używane w badanych artykułach dotyczących Redditu, z których na pierwszy plan wysuwają się zanurzenia (*embeddings*). Jako metoda badawcza w językoznawstwie i przetwarzaniu języka naturalnego jest to dość nowe rozwiązanie dla analizy danych tekstowych. Interesujące jest, że do analizy sarkazmu w dyskursie mediów społecznościowych używane są głównie rekurencyjne sieci neuronowe (LSTM networks). Informacje na temat zastosowanych technologii w artykułach dotyczących platformy Reddit (rycina 7) zawierają zarówno metody pozyskiwania danych (np. Reddit API, Pushshift), jak i metody ich analizy (np. VADER używany do analizy sentymentu, GloVe czy word2vec używany do analizy podobieństw semantycznych, czy rekurencyjne sieci neuronowe), co nieco zamazuje obraz sytuacji, są to bowiem narzędzia informatyczne wykorzystywane do odmiennych celów. Wyszczególniony ranking jest zatem nieco mylący, bowiem obejmuje nie tyle (lub nie tylko) tematykę tekstów, ale też metodologie używane w językoznawstwie, medioznawstwie oraz w informatyce.

Artykuł „The State of the Art of NLP…" stanowi bardzo udany i łagodny pomost pomiędzy informatyką a humanistyką. Tekst ten w zrozumiały dla humanistów sposób przybliża istotne i nie zawsze łatwe zagadnienia informatyczne. Autorzy wskazują, które zasoby tekstowe należy wybierać i według jakich kryteriów (z punktu widzenia narzędzi i metod komputerowych). Dobór danych do analizy w tym artykule jest bardzo dobrze uzasadniony, a kryteria wykluczające ujęte w podsumowującej i przekonującej tabeli 1. Wyniki badania mogą stanowić też pewne sugestie dla badaczy humanistów, np. autorzy zauważyli, że dane tekstowe często analizowane są w parze dwóch mediów społecznościowych - Facebooka i Twittera - co może być wskazówką z jakich platform porównywać teksty, aby zaproponować analizy mniej standardowe pod względem doboru platform, a tym samym podjąć się bardziej oryginalnych badań. Równie cenną obserwacją jest najczęstszy układ par językowych poddawanych analizie w badaniach z zakresu NLP (ryc. 2). Z przedstawionej mapy ciepła wynika, że język polski jest zdecydowanie niedoreprezentowany. Niezwykle istotną informacją, w każdym razie dla językoznawców, którzy dopiero zaczynają badania w zakresie NLP, jest wskazanie (ryc. 7), które publikacje są kluczowe, tj. najczęściej cytowane. Dane te Autorzy uzyskali dzięki interesującym wizualizacjom. Jeśli chodzi o zagadnienie zwane w językoznawstwie stosowanym czytelnością (ang. *readability*), a omówione w artykule jako *text complexity* (punkt 3.8.1), to trzeba zauważyć, że Autorzy omawiają tzw. klasyczne testy określające stopień zrozumiałości tekstów w oparciu o mechaniczne liczenie liczby sylab w wyrazach i/lub liczby wyrazów w zdaniu. Tego typu testy nie odzwierciedlają prawidłowo stopnia złożenia tekstów i ich zrozumiałości.

Można by w dalszych badaniach uwzględnić też testy nowszej generacji, które oprócz odwoływania się do mechanicznego obliczania czytelności uwzględniają przede wszystkim semantyczne cechy tekstów (takie badania prowadzi m.in. Crossley, S. A., Skalicky, S., & Dascalu, M., 2019).

„Text embeddings and clustering for characterizing online communities on Reddit" to tekst dotyczący zastosowania osadzeń kontekstowych (ang. *text embeddings*), nazywanych też czasami zanurzeniami, w analizie danych pozyskanych z platformy Reddit w celu scharakteryzowania społeczności internetowych komunikujących się poprzez ww. platformę. Do badania zastosowano bardziej wydajną wersję transformera BERT zwaną DistilBERT. Artykuł wskazuje, które społeczności są najbardziej aktywne i jaka generalnie tematyka najczęściej się w nich przewija, co również jest cenną informacją dla osób pragnących analizować dyskurs internetowy typowy dla tej platformy. Metodologia opisanego badania nie wzbudza zastrzeżeń.

W kolejnym (czwartym) artykule zatytułowanym „Reddit CrosspostNet…" zaproponowano nową metodę badania powiązań między subreditami, która pozwala lepiej zrozumieć strukturę Reddita i interakcję zachodzącą między uczestnikami różnych społeczności. Metoda ta opiera się na identyfikacji postów, które pojawiły się na jednym subredicie, a następnie zostały udostępnione na innym (tzw. crossposts), co – jak się dowiadujemy z następnego artykułu – jest zjawiskiem typowym tylko dla Reddita, w każdym razie udostępniane w taki, a nie inny sposób, przez osoby, które sądzą, że dany temat może zainteresować uczestników innego subreddita. Autorzy do opracowania tej metody użyli 7 mln postów opublikowanych na 10 tysiącach subredditach, które pojawiły się w ciągu jednego roku. Zaproponowana metoda opiera się na algorytmie tworzenia sieci grafowej. Stwierdzono, że crossposting występuje najczęściej między trzema subreditami, wszystkie oparte są na strukturze pytanie-odpowiedź i polegają na wyrażaniu sądów określających moralność opisywanych zachowań (r/AmITheAngel, r/AmITheDevil, r/AmITheAsshole). Zauważono też, że w niemal stu procentach posty udostępniane są tylko jeden raz oraz że nie pojawiają się na subredditach zawierających negatywne treści, w szczególności na subforach dotyczących konfliktów. Opracowana metoda śledzenia interakcji między subredditami może być interesująca i niezwykle przydatna dla językoznawców (a także medioznawców i politologów) analizujących wyrażanie opinii i wartościowanie w języku, w szczególności wyrażanie sądów moralnych, ewaluacji oraz ich sentymentu.

Piąty artykuł („Applying Named Entity Recognition and Graph Networks…") kontynuuje tematykę nowatorskiej metody opisanej w poprzednim artykule dotyczącej tzw. crosspostów. Jak już wiadomo z poprzedniego artykułu, crossposty świadczą o interakcji użytkowników Reddita skupionych w odmiennych społecznościach internetowych. Z tego artykułu można dowiedzieć się, jak automatycznie wykrywać podobieństwa tematyczne między subredittami za pomocą NER, które pozwalają identyfikować elementy (słowa oznaczające nazwy własne) wykorzystywane następnie do tworzenia sieci grafowych. Technika NER nie jest zwykle używana do modelowania tematycznego, jednak autorzy tego tekstu proponują, aby stosować NER ww celu, co jest nietypową i innowacyjną metodą identyfikacji podobieństw w przestrzeni internetowej. Zaletą tej metody jest znacznie zredukowana liczba danych wyjściowych (*output*) oraz ich jakość – słowa oznaczające konkretne rzeczywiste byty, co ułatwia w dalszych krokach analizy tworzenie sieci grafowych. Metoda ta jest pomocna przy analizie postów i komentarzy na określony temat publikowanych w subredditach o odmiennych tytułach. Innymi słowy, metoda ta pozwala badać ukryte tematy w ramach innych, dominujących treści wynikających z tytułu danego subreddita. Podobne badania prowadzone są przez medioznawców i językoznawców analizujących tzw. „trzecią przestrzeń" (pojęcie zapożyczone z socjologii) na forach internetowych (np. Vochocova i Rosenfeldova, 2019). Jednakże analizy te przeprowadzane są ręcznie przez tych badaczy, tymczasem opracowane narzędzie opisane w ww.

artykule umożliwia dokonanie takiej analizy automatycznie i na zdecydowanie większych danych tekstowych.

Szósty artykuł („Exploring Reddit Community Structure") przedstawia architekturę informacji na platformie Reddit pogłębiając wcześniejsze rozważania dotyczące analizy interakcji i przepływu informacji między subredditami opartej na sieciach grafowych. Autorzy stosują technikę osadzeń/zanurzeń tekstowych (*text embeddings*) bazującą na podobieństwie kosinusowym (czyli na mierze dystansu ogólnie przyjętej i szeroko stosowanej w NLP) oraz aplikują algorytm Louvain w celu detekcji społeczności internetowych (odwołując się do znanych z wcześniejszych badań pojęć, tzw. „mostów" i „bram"; ang. *bridges* i *gateways*). „Mosty" i „bramy" stanowią węzły wskazujące na przepływ informacji pomiędzy subredditami zlokalizowanymi w różnych społecznościach („mosty") lub węzły pozwalające na „wejście" do nowego subredditu, tj. przemieszczenie się z jednej społeczności do innej („bramy"). Autorzy zauważyli, że „mosty" i „bramy" generują redundancje (mają tendencje do nakładania się), w wyniku czego zaproponowali inne rozwiązanie, które nazwali „autostradami" (ang. *highways*). Pojęcie to opisuje ukryte, najczęściej używane ścieżki połączeń w architekturze społeczności na Reddicie. W opisanym badaniu Autorzy rozszerzają zakres tematyczny (wcześniej ograniczający się do 4 orientacji politycznych) i zwiększają liczbę subredditów (powyżej 4), a także zmieniają przedział czasowy na rok 2022 (wcześniej badano przedział od 2017 do 2020). Wyeliminowali też ze swojego badania subreddity dotyczące treści dla dorosłych (tzw. subreddity NFSW). W wyniku tych i innych jeszcze drobniejszych ograniczeń w aktualnym badaniu powtarzają się subreddity z badania wcześniejszego, na które powołują się Autorzy jedynie w ok. 45%. Doktorant bada modularność społeczności korzystając z aktualnie zalecanej metody Louven, która porównuje gęstość połączeń w obrębie jednej społeczności z połączeniami pomiędzy różnymi społecznościami. Wydaje się, że badanie dotyczące tylko jednego roku (2022) może nie być reprezentatywne dla całości społeczności na Reddicie. Warto byłoby porównać te badania z innym przedziałem lat. W punkcie 3 ww. artykułu podano ważne informacje dotyczące różnicy między poprzednim badaniem (wprowadzającym metodykę „mostów" i „bram") z tym proponowanym przez Autorów, jednak dane te są nieklarowne, bowiem użyto odmiennych kryteriów porównawczych: w oryginalnym badaniu opierano się na 492 subredditach, a w obecnym na 100 tysiącach subskrybentów. Nie jest zatem jasne, na ilu subskrybentach opierało się oryginalne badanie i/lub na ilu subredditach obecne (można się jedynie domyślać, że jest ich w aktualnym badaniu znacznie więcej).

Kolejny, siódmy artykuł („Application of natural language processing and temporal networks…") poświęcony jest śledzeniu rozwoju relacji pomiędzy subredditami na przestrzeni 8 lat (2015-2023) i analizie nie tyle węzłów (ang. *nodes*) ile połączeń między nimi (tzw. *edges*), a także dystrybucji tych połączeń. W analizie tej porównano ponadto dwa konkurencyjne podejścia (algorytmy) służące do detekcji społeczności internetowych znanych jako metoda Louven i metoda Leiden. Wnioski sugerują większą tendencję do identyfikacji modularnych społeczności przez algorytm Louven i tworzenie bardziej uporządkowanej struktury wielu społeczności oraz ujawnianie połączeń między nimi przez algorytm Leiden. W sieci grafowej główne węzły określane były wcześniej na podstawie m.in. liczby użytkowników, ich aktywności, komentarzy do postów i podobieństw między subredditami. Niewiele studiów uwzględniało treści postów w modelowaniu relacji między nimi, co stanowi właśnie cel badań opisanych w tym artykule. Treści analizuje się używając technik NLP, jednak wcześniej w opisywanym badaniu utworzono zanurzenia tekstowe za pomocą transformera DistilBERT. W efekcie tych badań zauważono, że na przestrzeni lat liczba społeczności na platformie Reddit zmniejsza się, a te społeczności, które pozostają ewoluują.

Kolejnych 5 publikacji uzupełnia, w mniejszym lub większym stopniu, zakres tematyczny wcześniej opisanych badań. Rozszerzenie zagadnienia modelowania tematycznego z nowym rozwiązaniem łączącym trzy znane i wcześniej używane techniki modelowania (LDA, NMF i BERTopic) zaproponowano i zwalidowano oraz wyniki opisano w artykule „Topic modeling applied to the Reddit posts". Bezpośrednie przełożenie badań komputerowych zakresu NLP na badania językoznawcze (pragmatyczne) wyraźnie widać w artykule „Decoding gender bias …", napisanym wspólnie z językoznawczynią. Tekst zatytułowany „Agents assembly…", z kolei, opisuje utworzony przez autorów ekosystem opracowany dla niespecjalistów z metod komputerowych, który składa się z języka specyficznego dla danej dziedziny (AASM), środowiska, które tłumaczy ten język na język Python, działając na platformie SPADE. Dzięki tej metodzie modelowania (wykorzystującej jednostki autonomiczne), możne symulować zachowania i interakcje w systemach złożonych, badając przepływ danych (np. *car traffic simulation*). Inny artykuł („VisQualdex…") przedstawia zestaw zaleceń do tworzenia wizualizacji statycznych (tj. wykluczających wizualizacje interaktywne i trójwymiarowe) opartych na znanej np. Zaproponowane rozwiązania mogą być przydatne dla badaczy z różnych proweniencji, w tym reprezentujących nauki humanistyczne i społeczne. Artykuł ten w moim odczuciu powinien być ujęty w serii 7 artykułów podstawowych opisanych powyżej. Publikacja pt. „Fusing individual choices into a group decision…" opisuje aplikację służącą do podejmowania decyzji grupowych, co zilustrowano na przykładzie wyboru menu posiłków. Rozwiązanie to może mieć zastosowanie w asystentach osobistych (np. Alexa czy Cortana) czy innych popularnych aplikacjach (np. Grupon). Tekst ten wydaje się być najmniej związany z pozostałymi artykułami i ogólnym celem badań Doktoranta; nie wskazuje on na możliwości zastosowania opisanych rozwiązań w badaniach naukowych dotyczących analizy mediów społecznościowych, lecz dotyczy praktycznych rozwiązań ułatwiających funkcjonowanie w świecie rzeczywistym.

Przedstawiona do recenzji seria artykułów uwidacznia ewolucję badań nad platformą Reddit prowadzonych przez Doktoranta, od analiz interakcji na poziomie micro (crossposty, analiza treści postów) do analizy na poziomie makro (interakcja pomiędzy społecznościami i zmiany ewolucyjne w czasie). W przedstawionych do recenzji tekstach widoczne jest też zainteresowanie Doktoranta potencjalnym zastosowaniem różnych technik komputerowych w innych dziedzinach/dyscyplinach.

Kompetencje Doktoranta i znaczenie badań

Na podstawie list bibliograficznych występujących w artykułach oraz proponowanych nowatorskich rozwiązań, a także opisu znanych już wcześniejszych algorytmów i technik komputerowych można stwierdzić, że Doktorant ma rozległą i szczegółową wiedzę z zakresu uczenia maszynowego i architektury sieci neuronowych, w tym z zakresu tzw. transformerów. Wiedzę tę można z powodzeniem zastosować w prowadzeniu wielorakich badań w różnych dyscyplinach wiedzy, w szczególności w dziedzinie nauk społecznych i humanistycznych, co sugeruje Doktorant w niektórych tekstach.

Prace mgr inż. Jana Sawickiego mają ogromną wartość w szczególności dla językoznawców i innych badaczy zajmujących się analizą danych tekstowych ekstrahowanych z mediów społecznościowych (politologów, medioznawców, socjologów). Badania dotyczące pozyskiwania i analizy danych pobieranych z platformy Reddit systematycznie zyskują na popularności, w szczególności wśród językoznawców analizujących język mowy nienawiści. Zaproponowane w publikacjach przedłożonych do recenzji rozwiązania informatyczne (modelowania tematycznego, ekstrakcji danych czy w szczególności zaproponowane nowatorskie wykrywanie podobieństw między subredditami z wykorzystaniem NER i sieci grafowych oraz crosspostów)

stanowią nieocenioną wskazówkę dla badaczy analizujących dane tekstowe publikowane na platformie Reddit. Publikacje wyjaśniają i definiują pojęcia z zakresu narzędzi informatycznych i ewaluują opisywane narzędzia i techniki komputerowe, co stanowi cenną podpowiedź dla językoznawców. Za mocną stronę prac (współ)autorstwa pana Jana Sawickiego uważam uściślenia w doborze prób oraz rygor metodologiczny. Opis metod jest zawsze dokładny, a wybór narzędzi do badania jest dobrze uzasadniony i przekonujący.

W przedłożonych tekstach zdarzają się niekiedy niezgrabności językowe, nieścisłości interpunkcyjne lub brak przedimków i inne przeoczenia, w tym gramatyczne (w szczególności w artykule 7), jednak te pomyłki i niedoskonałości nie mają zasadniczo wpływu na merytoryczną wartość prac oraz zaproponowane innowacyjne rozwiązania metodologiczne. Pragnę też podkreślić, jako anglistka, że ogólnie stylistyka tekstów (które napisane są w języku angielskim), nie wzbudza zastrzeżeń; jest ona zbieżna z innymi tekstami publikowanymi w języku angielskim z zakresu PJA w czasopismach czy punktowanych artykułach pokonferencyjnych. Styl jest zwięzły, precyzyjny i klarowny, co czyni przedłożone publikacje przydatne i zrozumiałe dla nauk humanistycznych i społecznych.

Konkluzje

Stwierdzam, że praca doktorska pt. „Exploring the Information Structure of Reddit Through Natural Language Processing and Network Analysis" przedstawiona przez pana mgr inż. Jana Sawickiego **spełnia wymagania ustawy o stopniach naukowych i tytule naukowym**, w związku z tym **wnoszę o jej przyjęcie i dopuszczenie do publicznej obrony**.