dr hab. inż. Piotr A. Kowalski, prof. AGH
Katedra Informatyki Stosowanej i Fizyki Komputerowej,
Wydział Fizyki i Informatyki Stosowanej,
Akademia Górniczo-Hutnicza w Krakowie,
al. Mickiewicza 30, 30-059 Kraków
email: pkowal@agh.edu.pl

oraz

Centrum Informatycznych Metod Analizy Danych
Instytut Badań Systemowych
Polskiej Akademii Nauk

Kraków, 30.12.2024

**RECENZJA**

**rozprawy doktorskiej Pana mgra Jana Sawickiego pt.**
**„Exploring the Information Structure of Reddit Through Natural"**

## 1. Uwagi ogólne

Prawną podstawą przygotowania recenzji rozprawy doktorskiej Pana mgr. inż. Jana Swickiego jest Umowa z Politechniką Warszawską – Radą Naukową Dyscypliny Informatyka Techniczna i Telekomunikacja reprezentowaną przez Przewodniczącego Pana Profesora dr. hab. inż. Jarosława Arabasa, którą otrzymałem 25 listopada 2024 r.

Recenzja została przygotowana na podstawie rozprawy doktorskiej. Przedmiotowa praca została zrealizowana pod kierunkiem Pani prof. PW dr. hab. Marii Ganzhy. Recenzowana rozprawa doktorska przedstawiona została w postaci woluminu wydanego przez Politechnikę Warszawską, który zawiera artykuły naukowe stanowiące wkład naukowy do dysertacji. Rozprawa doktorska jest rozpatrywana w dziedzinie nauk inżynieryjno-technicznych w dyscyplinie informatyka techniczna i telekomunikacja.

Recenzowana praca doktorska podejmuje szczegółowe badania nad strukturą informacyjną Reddita, wykorzystując nowoczesne metody przetwarzania języka naturalnego (NLP) oraz analizy sieci grafowych. Opracowanie obejmuje siedem prac, z których każda dotyczy różnych aspektów analizy platformy. W początkowych etapach skupiono się na charakterystyce tematycznej Reddita, analizując tysiące subredditów. Zastosowano zaawansowane techniki NLP, takie jak modelowanie zanurzeń tekstu (tzw. embeddingów), by uchwycić strukturę informacyjną na podstawie treści postów.

Autor dysertacji w kolejnych etapach badań skoncentrował się na relacjach między subredditami, budując skierowane sieci grafowe na podstawie danych o krzyżowych publikacjach. Wykorzystano także innowacyjne podejście oparte na rozpoznawaniu nazwanych bytów (NER) i analizie grafów, co pozwoliło zrozumieć tematyczne powiązania między społecznościami. Dodatkowo wprowadzono nową koncepcję „autostrad" w analizie sieci, która rozszerza tradycyjne pojęcia mostów i bram. Badania te ujawniły dynamikę i zmienność tematyczną subredditów na przestrzeni lat 2015–2023, stosując metody wykrywania społeczności, takie jak algorytmy Louvain i Leiden.

Osiągnięcia pracy przyczyniły się do lepszego zrozumienia ekosystemu Reddita zarówno na poziomie mikro, jak i makro. Natomiast otrzymane wyniki dostarczają cennych informacji o dynamice społeczności internetowych oraz proponują nowe techniki analizy grafów i NLP. Te innowacyjne podejścia otwierają nowe perspektywy badawcze, oferując solidne podstawy dla przyszłych eksploracji interakcji społecznych w sieci.

Dysertacja napisana jest w języku angielskim, zawiera 18 punktów, które wypełniają 57 stron. Na samym końcu rozprawy, dołączone są w postaci załączników artykuły naukowe stanowiące integralną część pracy doktorskiej.

## 2. Ogólna charakterystyka rozprawy oraz aspektu badawczego

Rozdział 1 przedstawia motywację do analizy danych z mediów społecznościowych, w szczególności z platformy Reddit. Eksplozja ilości danych generowanych przez użytkowników na takich platformach stanowi wyjątkową okazję do zastosowania technik nauki o danych i przetwarzania języka naturalnego (NLP). Dzięki analizie tych danych badacze mogą odkrywać ukryte wzorce, przewidywać trendy oraz badać zjawiska społeczne w niespotykanej dotąd skali. Reddit, z jego bogatą strukturą tematyczną, interakcjami użytkowników oraz treściami tworzonymi przez społeczność, oferuje szczególnie wartościowe zasoby do badania dynamiki społeczności, analizy sentymentów oraz przepływu informacji.

W rozdziale tym szczegółowo opisano strukturę i funkcjonalność Reddita, który działa jako „sieć forów" skupionych w subreddity poświęcone różnym tematom. Treści są generowane przez użytkowników i oceniane za pomocą systemu głosów (upvote/downvote), co wpływa na ich widoczność. Unikalnym aspektem platformy jest jej elastyczność tematyczna, obejmująca zarówno popularne zagadnienia, jak i niszowe społeczności. Reddit wyróżnia się także wysokim stopniem anonimowości użytkowników oraz dostępnością większości treści bez potrzeby logowania, co różni go od innych platform społecznościowych.

Rozdział 1 podkreśla również specyfikę społeczności i kultury Reddita, w tym rolę moderatorów, którzy dbają o przestrzeganie zasad w subreddicie. Reddit stanowi centrum dla niszowych społeczności, które często są pomijane na innych platformach. Dzięki swojej strukturze i funkcjonalności, Reddit łączy w sobie cechy mediów społecznościowych, forów internetowych i społecznościowego zarządzania treścią, co czyni go nie tylko unikalnym w krajobrazie cyfrowym, ale także fascynującym obiektem badań naukowych.

W rozdziale 2 przedstawiony zostaje przegląd dotychczasowych badań dotyczących Reddita, wskazując na ich fragmentaryczny charakter. Dotychczasowe prace badawcze skupiały się głównie na analizie pojedynczych subredditów lub ich małych grup, obejmując różnorodne tematy, takie jak polityka, teorie spiskowe, zdrowie psychiczne czy moderacja treści. Autor zwraca uwagę, że wybór analizowanych subredditów był często intuicyjny, co prowadziło do pomijania powiązanych społeczności i ograniczało pełne zrozumienie struktury informacyjnej platformy.

Doktorant zauważa również, że część badań traktowała Reddit jako jednorodne źródło danych, ignorując jego wewnętrzną strukturę i skupiając się na ogólnych zagadnieniach, takich jak przetwarzanie języka naturalnego czy analiza obrazów. W podsumowaniu autor wskazuje na konieczność przeprowadzenia bardziej złożonych analiz uwzględniających specyfikę struktury Reddita, co pozwoliłoby na lepsze uchwycenie dynamiki informacyjnej i społecznościowej platformy.

Doktorant w rozdziale 3 koncentruje się na metodach stosowanych w badaniach nad Redditem, z uwzględnieniem specyficznych ograniczeń i decyzji badawczych. Ze względu na złożoność przetwarzania multimodalnego oraz problemy z dostępnością danych obrazowych, autor zdecydował się ograniczyć analizę do tekstowych postów, które są najbardziej dostępne i efektywne pod względem czasu oraz energii.

W podrozdziale 3.1 Doktorant omawia ewolucję dostępu do danych Reddita. Początkowo dane były pozyskiwane za pomocą oficjalnego API Reddita, które oferowało prosty dostęp, choć z pewnymi ograniczeniami dotyczącymi częstotliwości zapytań. Wraz z rosnącym zapotrzebowaniem na dane, popularność zyskało API Pushshift, które umożliwiało dostęp do obszernych archiwów postów i komentarzy, wspierając badania longitudinalne. Jednak w 2023 roku Reddit wprowadził istotne ograniczenia w dostępie do swojego API, co znacząco utrudniło korzystanie z danych przez społeczność naukową. Obecnie naukowcy korzystają z alternatywnych źródeł danych, takich jak archiwa dostępne na platformach typu Academic Torrents. Chociaż oferują one dostęp do dużych zbiorów danych historycznych, wiążą się również z wyzwaniami technicznymi i etycznymi, które doktorant podkreśla jako istotne aspekty w kontekście dalszych badań.

W rozdziale 3.2 omówiono proces filtracji danych wejściowych w kontekście analizy postów na Reddit, gdzie filtracja obejmuje usuwanie postów o minimalnej liczbie głosów (brak uwagi użytkowników), małej liczbie komentarzy, usuniętych przez użytkowników lub moderatorów oraz zduplikowanych postów. Te kroki filtracyjne zostały zastosowane podczas przygotowywania zbiorów danych używanych we wszystkich publikacjach stanowiących część pracy doktorskiej.

Rozdział 3.3 opisuje zaawansowane techniki przetwarzania języka naturalnego stosowane w analizie danych Reddit, w tym kroki oczyszczania danych, takie jak usuwanie duplikatów, nieistotnych elementów, czy stop słów, oraz tokenizację, lematyzację i stemming. Wśród metod wykorzystywanych do analizy tekstu wymieniono m.in. TF-IDF do wyodrębniania cech, LIWC do analizy emocjonalnych aspektów języka, oraz LDA, która umożliwia odkrywanie ukrytych tematów w dużych zbiorach tekstowych. Dodatkowo, stosowane są różnorodne metody analizy sentymentu oparte na słownikach emocji, takie jak NRC czy AFINN. W szczególności, współczesne badania nad Reddit wykorzystują model BERT (Bidirectional Encoder Representations from Transformers), który dzięki dwukierunkowej analizie tekstu osiąga wysoką skuteczność w zadaniach takich jak analiza sentymentu, detekcja złośliwych komentarzy, rozpoznawanie sarkazmu czy diagnoza zaburzeń psychicznych. Wspomniano także o efektywniejszej wersji BERT, DistilBERT, która zapewnia oszczędność zasobów obliczeniowych przy zachowaniu większości funkcji oryginału. Zastosowanie tych metod w badaniach Reddit otwiera

nowe możliwości analizy, w tym wykorzystanie embedowania tekstu do klasteryzacji czy analizy sieciowej.

W rozdziale 3.4 omówiono algorytmy klasteryzacji, które są metodami nadużywanego uczenia nienadzorowanego, mającymi na celu grupowanie punktów danych w klastry w taki sposób, aby punkty w obrębie jednego klastra były bardziej podobne do siebie niż do punktów w innych klastrach. Wśród popularnych metod klasteryzacji wymieniono K-means, który dzieli dane na K klastrów, minimalizując wariancję wewnątrz klastra, oraz klasteryzację hierarchiczną, która buduje drzewo klastrów przez łączenie lub dzielenie ich. Metody oparte na gęstości, takie jak DBSCAN, identyfikują klastry na podstawie gęstości punktów danych, co sprawia, że są skuteczne w odkrywaniu klastrów o dowolnym kształcie oraz radzeniu sobie z szumem w danych. W kontekście pracy doktorskiej szczególną uwagę poświęcono algorytmowi K-means, który wykazuje wysoką skuteczność w aplikacjach naukowych i został już zastosowany do analizy danych z Reddit przy użyciu embedowania tekstu, w tym np. w identyfikacji podobnych postów na subreddicie r/Advice. Choć dotychczas klasteryzacja z użyciem embedowania BERT była wykorzystywana w kontekście pojedynczych postów, to brak jest badań, które próbowałyby przeprowadzić klasteryzację embeddingów tekstów całych subreddytów, co stanowi nowatorską propozycję badawczą w pracy doktorskiej.

W rozdziale 3.5 omówiono zastosowanie sieci grafowych do analizy struktury Reddita, który jest analizowany za pomocą metod opartych na grafach, takich jak interakcje użytkowników, przepływ uwagi czy relacje między subreddytami. Grafy mogą przyjmować formę grafów skierowanych lub nieskierowanych, a także grafów ważonych, w zależności od tego, czy wierzchołki i krawędzie mają przypisane wagi. W pracy doktorskiej szczególną uwagę zwrócono na interkomunalne (między-subredditowe) relacje poprzez treść postów, co stanowi nowatorskie podejście, ponieważ dotychczas nie przeprowadzono analiz tego typu na dużą skalę. Dodatkowo, zastosowano klasyczne metody analizy grafów, takie jak liczba wierzchołków, liczba krawędzi, gęstość grafu, dystrybucja stopni wierzchołków, czy centralność stopni. Przeanalizowano również zaawansowane metody, takie jak osadzanie wierzchołków (node embeddings) za pomocą algorytmu Node2Vec, który zamienia wierzchołki grafu w reprezentacje wektorowe, a także analizę sieci czasowych, pozwalających na badanie rozwoju struktury Reddita w czasie.

Kolejnym istotnym zagadnieniem było wykrywanie społeczności w grafach, z naciskiem na porównanie dwóch metod: Louvain i Leiden. Metoda Louvain służy do wykrywania struktur społeczności poprzez optymalizację modularności, jednak jej wadą jest tworzenie niespójnych społeczności. Metoda Leiden, będąca udoskonaloną wersją Louvain, zapewnia lepszą spójność społeczności oraz wyższą modularność, co czyni ją bardziej odpowiednią dla dużych i złożonych sieci. Zajęto się także analizą mostów (bridge nodes) i bram (gateway nodes) w kontekście sieci subredditów, które pomagają w zrozumieniu interakcji pomiędzy różnymi społecznościami. Mosty stanowią elementy łączące różne społeczności, a bramy umożliwiają wejście do danej społeczności, co jest szczególnie przydatne w badaniach nad relacjami między subreddytami.

W wyniku przeglądu stanu wiedzy na temat Reddita i metod badawczych stosowanych do jego analizy, w Rozdziale 4 zidentyfikowano kilka istotnych luk badawczych, które stanowią fundament dla prac zawartych w niniejszej rozprawie. Kluczowe luki badawcze to: brak pełnej charakterystyki struktury informacji Reddita na dużą skalę, nieprzeprowadzona analiza struktury Reddita z uwzględnieniem specyficznych subredditów, brak dużej analizy Reddita na podstawie treści, nieprzeanalizowana ewolucja czasowa relacji między subreddytami oraz niewykorzystanie nowoczesnych metod NLP, takich jak osadzanie tekstów oparte na BERT, do modelowania struktury informacji Reddita. Dodatkowo, brak badań nad wykorzystaniem crosspostów w aplikacjach naukowych, niewykorzystanie podobieństw między subreddytami z różnych dziedzin tematycznych oraz brak testów teorii „mostów" i „bram" stanowiły istotne luki.

W Rozdziale 5 Autor przedstawił zestawienie 7 artykułów naukowych wchodzących w skład pracy doktorskiej wraz z rozdziałami, w których zostały one omówione.

Artykuły:

[A1] Sawicki, Jan, Maria Ganzha, Marcin Paprzycki, and Amelia Bădică. *"Exploring Usability of Reddit in Data Science and Knowledge Processing."* Scalable Computing: Practice and Experience 23, no. 1 (2022): 9-22;

[A2] Sawicki, Jan, Maria Ganzha, and Marcin Paprzycki. *"The State of the Art of Natural Language Processing—A Systematic Automated Review of NLP Literature Using NLP Techniques."* Data Intelligence 5, no. 3 (2023): 707-749;

Zostały omówione w rozdziale 6. Artykuł [A1] bada potencjał Reddita jako źródła danych dla nauki o danych i przetwarzania wiedzy. Autorzy analizują 180 publikacji, identyfikując kluczowe tematy, takie jak analiza rozmów i wpływ pandemii COVID-19. Zastosowano metody, takie jak osadzania tekstu i analiza sieci, a także przetestowano narzędzia do pozyskiwania danych, w tym API Reddita i Pushshift. Wzrost liczby badań naukowych opartych na danych z Reddita wskazuje na jego rosnące znaczenie w analizie współczesnych trendów i sentymentów. Artykuł [A2] przedstawia systematyczny przegląd literatury dotyczącej przetwarzania języka naturalnego (NLP), oparty na analizie 4712 publikacji z repozytorium arXiv. Badanie identyfikuje popularne zestawy danych (np. Wikipedia, Twitter) i języki (np. chiński, niemiecki), a także najczęściej badane zadania, takie jak analiza sentymentu i tłumaczenie maszynowe. Wykorzystano zaawansowane techniki NLP, w tym osadzania tekstów i klasteryzację, aby systematycznie przedstawić rozwój dziedziny.

Oba artykuły stanowią solidne podstawy teoretyczne dla dalszych badań. [A1] koncentruje się na możliwościach wykorzystania Reddita w nauce, podkreślając jego znaczenie jako źródła danych, natomiast [A2] dostarcza kompleksowego obrazu stanu wiedzy w dziedzinie NLP. Razem te prace przeglądowe wyznaczają kierunki badań, łącząc eksplorację nowoczesnych metod z praktycznym wykorzystaniem danych.

Kolejna praca:

[A3] Sawicki, Jan. *"Text embeddings and clustering for characterizing online communities on Reddit."* In 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS), pp. 1131-1136. IEEE, 2023;

dotyczy analizy struktury i ewolucji społeczności na Reddit w latach 2019–2022 przy użyciu zaawansowanych technik NLP i klasteryzacji danych. Z bazy Pushshift wybrano dane z 3090 największych subredditów, ograniczając się do 1000 najlepiej ocenianych postów na każdy z nich, co dało zbiór ponad 12 milionów postów. Teksty przetwarzano za pomocą modelu DistilBERT, generując osadzenia o 768 wymiarach, a następnie grupowano subreddit w 200 klastrów przy użyciu algorytmu K-means. Do oceny dynamiki klastrów w czasie zastosowano indeks Jaccarda, co pozwoliło śledzić stabilność i migracje tematów. Badanie wykazało, że najliczniejsze klastry obejmowały takie tematy jak pornografia, memy, technologia, polityka i sport, a stabilność w czasie wykazały m.in. klastry dotyczące ogrodnictwa i motoryzacji. Wpływ pandemii COVID-19 zaobserwowano w klastrach naukowych i zdrowotnych, które włączyły tematy związane z pandemią. Porównanie wyników z wcześniejszymi badaniami (2015) potwierdziło pewną spójność, ale także ujawniło ograniczenia danych. Praca podkreśla dynamiczny i różnorodny charakter społeczności Reddit oraz wskazuje kierunki dalszych badań, szczególnie w modelowaniu relacji tematycznych między subredditami.

W pracy:

[A4] Sawicki, Jan, Maria Ganzha, Marcin Paprzycki, and Yutaka Watanobe. *"Reddit CrosspostNet—studying Reddit communities with large-scale Crosspost graph networks."* Algorithms 16, no. 9 (2023): 424.

Doktorant skupił się na analizie dynamiki strukturalnej i relacyjnej społeczności Reddit poprzez badanie zjawiska crosspostów, czyli przypadków udostępniania treści między różnymi subredditami. Crossposty posłużyły jako unikalna perspektywa do zrozumienia interakcji między społecznościami oraz przepływu informacji. W tym celu autor skonstruował skierowany ważony graf, w którym węzły reprezentowały subreddity, a krawędzie - liczbę crosspostów między nimi. Analiza obejmowała rozkład stopni w grafie, współczynniki klasteryzacji, wykrywanie społeczności za pomocą algorytmu Louvain, analizę połączeń, a także osadzenia węzłów przy użyciu Node2Vec. Praca ujawniła, że sieć subredditów charakteryzuje się rozkładem stopni zgodnym z prawem potęgowym, co wskazuje na istnienie kilku silnie połączonych subredditów (tzw. hubów) oraz wielu mniej powiązanych. Algorytm Louvain zidentyfikował 796 społeczności, z czego większość składała się z 2-3 subredditów, choć znaleziono także większe społeczności tematyczne, takie jak polityka, memy czy pornografia. Szczególną uwagę zwrócono na subreddity skupione na osądach moralnych, jak r/AmITheAsshole (r/AITA) i jego pokrewne. Wyniki pracy wskazują, że crossposty są skutecznym narzędziem w badaniu relacji między społecznościami, dostarczając nowych wniosków uzupełniających dotychczasowe analizy interakcji użytkowników i treści tekstowych. Autor zasugerował dalsze badania nad automatyzacją wykrywania podobnych subredditów oraz analizą temporalną sieci subredditów, by zrozumieć ich ewolucję w czasie.

W publikacji:

[A5] Sawicki, Jan, Maria Ganzha, Marcin Paprzycki, and Yutaka Watanobe. *"Applying Named Entity Recognition and Graph Networks to Extract Common Interests from Thematic Subfora on Reddit."* Applied Sciences 14, no. 5 (2024): 1696

autorzy skupili się na opracowaniu metody odkrywania podobieństw tematycznych między subredditami oraz na roli crosspostów w tym procesie. Badanie, oparte na technikach rozpoznawania nazwanych encji (NER) i analizie grafów, miało na celu zrozumienie, w jaki sposób informacje są strukturyzowane i udostępniane między różnymi społecznościami Reddit. Wykorzystano obszerny zestaw danych obejmujący ponad 32 miliony postów z 3189 subredditów, z których około 2,3% stanowiły crossposty. Analiza obejmowała czyszczenie danych, ekstrakcję encji nazwanych, a następnie budowę grafów, w których encje stanowiły węzły, a współwystępowania w postach tworzyły krawędzie. Dalsze kroki obejmowały porównanie charakterystyk grafów poszczególnych subredditów oraz identyfikację wspólnych encji. Rezultaty badania wskazały, że zaproponowana metoda skutecznie ujawnia podobieństwa między subredditami, osiągając wyniki porównywalne z innymi nowoczesnymi systemami rekomendacyjnymi. Crossposty okazały się wiarygodnym wskaźnikiem wspólnych zainteresowań, ujawniając zarówno oczekiwane, jak i mniej oczywiste relacje między społecznościami. Metryki oparte na metadanych, takich jak liczba komentarzy czy oceny postów, skutecznie wspomagały analizę, a charakterystyki grafów, takie jak stopień węzłów czy współczynnik klasteryzacji, dodały głębi w interpretacji wyników. Badanie potwierdziło również obecność „zasady 1%" w strukturach subredditów, wskazując na dominację węzłów peryferyjnych.

Autorzy podkreślili, że zaproponowana metoda stanowi solidne ramy do analizy danych w sieciach społecznościowych na dużą skalę, oferując praktyczne zastosowania w identyfikacji tematycznie powiązanych subredditów. W przyszłości planują rozszerzyć badania o analizę treści multimedialnych oraz wieloletnich danych, aby zbadać ewolucję tematów subredditów.

[A6] Sawicki, Jan, and Maria Ganzha. *"Exploring Reddit Community Structure: Bridges, Gateways and Highways."* Electronics 13, no. 10 (2024): 1935;

To artykuł, który podejmuje tematykę związaną z analizą struktury informacji w serwisie Reddit poprzez modelowanie relacji między subredditami oraz interakcji użytkowników w formie grafu. Autorzy nawiązali do wcześniejszych badań skoncentrowanych na politycznych subredditach sprzed 2020 roku, rozszerzając analizę na kompleksowy zbiór danych obejmujący rok 2022. Dzięki zastosowaniu technik przetwarzania języka naturalnego, takich jak tekstowe osadzanie i podobieństwo kosinusowe, badanie umożliwiło bezpośrednie modelowanie treści subredditów, budowę grafu, wykrywanie społeczności metodą Louvain, a także analizę tych społeczności w kontekście teorii mostów i bram. Wprowadzono również nowatorskie pojęcie „autostrad" – najczęściej używanych ścieżek między subredditami – w celu odkrycia dodatkowych zależności.

Przeprowadzona analiza wykazała istnienie 170 społeczności tematycznych, z największymi dotyczącymi stylu życia, gier, lokalizacji, komedii, zwierząt, polityki, popkultury, technologii i finansów. Wyniki wskazują, że większość węzłów pełniących funkcję mostów pomiędzy społecznościami jednocześnie działa jako bramy, co sugeruje, że skupienie się wyłącznie na bramach może wystarczyć do zrozumienia połączeń między społecznościami. Dodatkowo, koncepcja autostrad pozwoliła na głębszą analizę relacji między subredditami, uwzględniając ścieżki o największym natężeniu. Badanie podkreśliło znaczenie zarówno węzłów, jak i połączeń w badaniach nad społecznościami internetowymi, oferując nowe spojrzenie na strukturę sieci społecznych Reddit.

Ostatni artykuł:

[A7] Sawicki, Jan, Maria Ganzha and Marcin Paprzycki. *"Application Of Natural Language Processing And Temporal Networks To Analysis Of Evolution Of Reddit Communities"* Journal of Automation, Mobile Robotics and Intelligent Systems (2024);

koncentruje się na analizie ewolucji czasowej struktury informacji w serwisie Reddit. Głównym celem badania było zbadanie interakcji pomiędzy subredditami, identyfikacja kluczowych i zmieniających się węzłów oraz uchwycenie ścieżek tematycznych w czasie. W tym celu autorzy porównali skuteczność dwóch algorytmów detekcji społeczności – Louvain i Leiden – w kontekście dynamiki subredditowych społeczności. Dane, obejmujące okres 2015–2023, zawierały posty z subredditów mających co najmniej 100,000 subskrybentów, a do analizy czasowej utworzono miesięczne migawki sieci. Badania wykazały, że podobieństwo tematyczne subredditów ustabilizowało się w 2020 roku, co sugeruje spowolnienie w ewolucji relacji między ich treściami. Zaobserwowano istotne przesunięcia tematyczne, takie jak wzrost popularności politycznych subredditów po 2019 roku oraz spadek znaczenia tych związanych z tematyką zwierząt. Dynamika społeczności wskazała na konsolidację treści – liczba społeczności zmniejszała się, lecz ich rozmiary rosły, co utrudniało identyfikację trwałych podspołeczności w większych grupach. Porównanie metod Louvain i Leiden pokazało, że ta druga lepiej radzi sobie z dużymi i złożonymi sieciami, co podkreśla jej przewagę w takich analizach.

Podsumowując, badania zawarte w [A7] dostarczają wgląd w zmieniającą się strukturę informacji Reddita, ukazując kluczowe trendy w interakcjach i tematyce subredditów. Połączenie osadzeń tekstowych z analizą temporalnych grafów okazało się skuteczne, a uzyskane wyniki mogą stanowić punkt wyjścia dla przyszłych badań nad zaawansowanymi motywami temporalnymi w społecznościach online.

## 3. Ocena rozprawy

a. Uwagi krytyczno-polemiczne:

W niniejszej cześć pokrótce przedstawione zostaną główne mankamenty dysertacji. Ze względu na istotność tejże części recenzji niniejsze uwagi zostaną przedstawione w punktach, do których łatwiej będzie się odnieść Doktorantowi.

1. **Reprezentatywność danych [A4, A5, A6, A7].** W kilku badaniach Doktorant stosuje rygorystyczne kryteria filtrowania danych, takie jak wykluczanie subredditów NSFW [A6] czy ograniczanie analizy do subredditów z co najmniej 100,000 subskrybentów [A7]. O ile takie podejście pomaga w uproszczeniu analizy, może prowadzić do pominięcia ważnych społeczności, które, choć mniej popularne, mają znaczący wpływ na strukturę sieci Reddita. Czy było rozważane włączenie mniejszych subredditów lub zastosowanie metod pozwalających na analizę zróżnicowanych grup użytkowników.
2. **Brak analizy jakościowej wyników [A5, A6].** Prace te koncentrują się głównie na analizie ilościowej, wykorzystując metryki sieciowe i algorytmy, takie jak NER czy cosine similarity. Brakuje jednak głębszej analizy jakościowej wyników, np. weryfikacji tematycznych podobieństw identyfikowanych przez algorytmy. Włączenie analizy jakościowej, przeprowadzonej przez ekspertów w dziedzinie lingwistyki czy socjologii, mogłoby zwiększyć wiarygodność i interpretacyjność wyników.
3. **Nieadekwatność do dynamicznej natury Reddita [A7].** Doktorant w [A7] zauważa stabilizację tematyczną subredditów po 2020 roku, ale nie analizuje, w jaki sposób zmiany w polityce platformy, w algorytmach czy migracje użytkowników wpłynęły na tę stabilizację. Reddit jest platformą dynamiczną, a takie czynniki mogą mieć kluczowe znaczenie dla interpretacji wyników. Uwzględnienie dodatkowych danych kontekstowych (np. zmian w regulaminie Reddita) mogłoby pogłębić analizę.
4. **Redundancja koncepcji "bram" i "mostów" [A6].** W artykule [A6] zostaje stwierdzone, że 80% mostów pełni jednocześnie rolę bram, co sugeruje znaczną redundancję tych pojęć. Może to wskazywać na potrzebę przeformułowania tych koncepcji lub bardziej krytycznego spojrzenia na ich zastosowanie w analizie sieciowej. Wprowadzenie bardziej wyrafinowanych miar lub wskaźników mogłoby zwiększyć wartość poznawczą wyników.
5. **Ograniczenie do pojedynczych lat lub wybranych okresów [A4, A5, A6].** Analizy w kilku pracach są ograniczone do danych z jednego roku (np. 2022 w A4, A5 i A6), co może zniekształcać wnioski dotyczące długoterminowych trendów i dynamiki sieci. Taki wybór, choć uzasadniony ograniczeniami obliczeniowymi, może prowadzić do utraty informacji o ewolucji subredditów w szerszym kontekście. Uwzględnienie danych z wielu lat (np. w analizie podobnej do [A7]) mogłoby znacząco wzbogacić badania.

Podkreślając wysoką wartość merytoryczną pracy oraz ciekawy dobór tematyki, należy zauważyć, że sam układ pracy wydaje się mniej przemyślany. Struktura sprawia wrażenie przygotowanej bez należytej staranności, co nieco utrudnia płynność odbioru całości. Być może jest to pokłosie hybrydowego trybu postępowania doktorskiego, w którym rezygnuje się z pełnej i kompletnej dysertacji na rzecz skrótowego opisu artykułów będących podstawą dorobku Doktoranta. Niemniej jednak, staranniejsze zaplanowanie i uporządkowanie poszczególnych części znacząco podniosłoby odbiór tak wartościowego opracowania.

W szczególności z racji na obowiązki recenzenta muszę w tym miejscu wskazać jeszcze kilka uwag związanych z pojedynczymi niedociągnięciami stylistycznymi, czy dość skąpym opisem pewnych części pracy. Doktorant również nie ustrzegł się nielicznych mankamentów natury technicznej, takich jak błędy interpunkcyjne czy też pomyłki w gramatyce języka polskiego itp. jednak w dużej mierze nie rzutują one na relatywnie wysoką ocenę pracy.

b. Ocena ogólna.

Doktorant bardzo dobrze rozumie zagadnienia oraz zakres procedur analizy danych i sieci społecznościowych, zarówno w ujęciu ogólnym, jak i w kontekście specyficznych zastosowań opisywanych w poszczególnych publikacjach. W szczególności wykazał umiejętność: precyzyjnego opisu kluczowych cech i własności badanych struktur, syntetyzowania algorytmów pozwalających na pogłębioną analizę, prezentacji wyników w sposób przejrzysty i uporządkowany, jak również doboru i omówienia właściwych metod analitycznych.

Sposób sformułowania problemu badawczego oraz jego analiza przedstawiona w poszczególnych częściach pracy świadczy o dojrzałości naukowej Doktoranta. Problematyka poruszona w rozprawie została niezwykle precyzyjnie określona, co znacząco podnosi jakość merytoryczną pracy. Bardzo imponujące jest także to, że podczas pracy nad doktoratem Pan mgr inż. Jan Sawicki opublikował znaczną liczbę wysoko ocenianych artykułów naukowych, z których większość jest wynikiem międzynarodowej współpracy naukowej, co dodatkowo podkreśla szeroką perspektywę badawczą Kandydata.

Warto podkreślić, że Pan mgr inż. Jan Sawicki wykazał się solidnym warsztatem analitycznym i informatycznym, co znalazło odzwierciedlenie w samodzielnym opracowaniu głównych algorytmów stosowanych w badaniach. Umiejętności te świadczą o jego niezależności jako naukowca. Ponadto, liczba wykonanych eksperymentów oraz jakość uzyskanych wyników jednoznacznie wskazują na szczególne predyspozycje Kandydata do prowadzenia badań naukowych. Doktorant potrafi dobrać odpowiednie wskaźniki oceny jakości modeli, zilustrować wyniki w formie wykresów, przedstawić kluczowe rezultaty w tabelach oraz wyciągnąć trafne i istotne wnioski.

Dodane do rozprawy elementy analizy matematycznej i formalnego opisu powodują, że praca jest kompletna i w pełni wartościowa, stanowiąc ważny wkład w rozwój nauki w obszarze analizy sieci społecznościowych oraz tematyki związanej z analizą danych w skali globalnej.

**4. Podsumowanie**

Rozprawa doktorska Pana mgr. inż. Jana Sawickiego stanowi ambitne i wartościowe opracowanie problematyki analizy struktur sieciowych oraz dynamicznych relacji w ekosystemie Reddit. Praca podejmuje tematykę wykorzystania nowoczesnych metod przetwarzania języka naturalnego, analizy grafów oraz zaawansowanych metod eksploracyjnych do modelowania i interpretowania relacji między społecznościami internetowymi. Głównym celem pracy było zrozumienie struktur informacyjnych, identyfikacja podobieństw tematycznych pomiędzy społecznościami oraz uchwycenie ewolucji interakcji i tematów w czasie.

Główne osiągnięcie rozprawy to opracowanie spójnej metodologii integrującej techniki NLP i analizy sieci, umożliwiającej systematyczną analizę społeczności online na niespotykaną dotąd skalę. W ramach badań Doktorant zaproponował kilka innowacyjnych podejść, w tym:

- Modelowanie relacji tematycznych między subredditami z wykorzystaniem metod grafowych, które łączą tekstowe podobieństwa i charakterystyki sieciowe.
- Wprowadzenie pojęcia "autostrad" (highways) w analizie grafów, które pozwalają na identyfikację najczęściej wykorzystywanych ścieżek między społecznościami, uzupełniając klasyczne koncepcje "mostów" i "bram".
- Analiza ewolucji tematycznej w długim okresie czasu, która uwzględnia zmiany w dominujących narracjach, konsolidację społeczności oraz wpływ wydarzeń zewnętrznych na dynamikę interakcji.
- Porównanie skuteczności różnych metod detekcji społeczności (Louvain i Leiden), co stanowi istotny wkład w rozwój metodologii analizy dużych sieci.

Badania przedstawione w rozprawie są wyjątkowe nie tylko ze względu na zastosowane innowacyjne metody, ale również poprzez skalę przeanalizowanych danych – obejmujących dużą liczbę postów, oraz zróżnicowane podejście, które łączy analizę tekstu z analizą strukturalną sieci.

Praca wnosi istotny wkład do dziedziny analizy sieci społecznościowych, oferując nowe narzędzia i perspektywy badawcze, które mogą znaleźć zastosowanie w różnych obszarach, od socjologii i komunikacji po marketing i analizę danych wielkoskalowych. Główne osiągnięcie rozprawy, jakim jest nowatorskie podejście do integracji technik NLP z analizą sieciową, może również posłużyć jako podstawa do dalszych badań naukowych w przyszłości.

W rezultacie rozprawa Pana mgr inż. Jana Sawickiego jest nie tylko wyrazem jego dojrzałości naukowej i umiejętności badawczych, ale także istotnym wkładem w rozwój interdyscyplinarnej nauki o danych. Można zatem uznać, że recenzowana rozprawa doktorska ma charakter oryginalnej pracy projektowo-naukowej, o której mówią bieżące przepisy.

**Konkludując uważam, że rozprawa doktorska mgr inż. Jana Sawickiego zdecydowanie spełnia wymagania stawiane w odpowiednich przepisach rozprawom doktorskim i wobec tego stawiam wniosek o jej dopuszczenie do dalszych, przewidzianych Ustawą, etapów postępowania o nadanie stopnia doktora.**

Piotr Kowalski

dr hab. inż. Piotr A. Kowalski, prof. AGH