Politechnika Wrocławska

Katedra Sztucznej Inteligencji

Wrocław, 21.04.2021 r.

Prof. dr hab. inż. Przemysław Kazienko
Katedra Sztucznej Inteligencji,
Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska
Tel. 71 320 36 09
Email: kazienko@pwr.edu.pl
http://kazienko.eu

RECENZJA

rozprawy doktorskiej mgr inż. Jana Chołoniewskiego
pt. *„Modelling Dynamics of News Media"*

Rozprawę napisano pod kierunkiem prof. dr hab. Janusza Hołysta oraz dr Juliana Sienkiewicza i złożono w styczniu 2022 r. Recenzję wykonano na zlecenie Rady Naukowej Dyscypliny Nauki Fizyczne Politechniki Warszawskiej.

I. Przedmiot, problematyka i charakter rozprawy

Tematyka rozprawy mieści się w stosunkowo wąskiej, ale dynamicznie się rozwijającej i ważnej społecznie dziedzinie zwanej *medioznawstwem obliczeniowym* (*quantitative media analysis*), w której za pomocą analiz danych próbuje się zmierzyć i zrozumieć różne zjawiska zachodzące w medialnej przestrzeni cyfrowej. Ma ona także bezpośredni związek z inną dziedziną - analizą sieci społecznych (*social network analysis*), będącą z kolei częścią szerszego obszaru nauki o sieciach złożonych (*complex networks*) oraz mediami społecznościowymi i cyfrowymi (*social media*, *digital media*).

Praca ma więc silnie interdyscyplinarny charakter łączący (1) fizykę (układy złożone, procesy dyfuzyjne, fizyka statystyczna, symulacje), (2) informatykę (eksperymentalna analiza danych z wykorzystaniem metod obliczeniowych – *soft computing*, danologia – *data science*, modelowanie wieloagentowe – *multi-agent systems*, przetwarzanie informacji – *information processing*, wyszukiwanie informacji – *information retrieval*), (3) statystykę matematyczną i oczywiście (4) nauki społeczne (medioznawstwo, media cyfrowe).

Ze względu na moje kompetencje związane z danologią oraz analizą sieci złożonych wywodzące się z nauk technicznych - z informatyki, niniejsza recenzja jest wykonana nieco bardziej z perspektywy metod obliczeniowych niż fizyki statystycznej i modelowania układów złożonych.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

Politechnika Wrocławska

Katedra Sztucznej Inteligencji

Jednocześnie warto zauważyć, że fizyka i obliczeniowe nauki techniczne bardzo się uzupełniają. Przywiązanie fizyki do zrozumienia i opisania zjawisk za pomocą stosunkowo prostych i łatwych do interpretacji zależności od niewielu zmiennych było dla mnie wielką inspiracją, którą odnalazłem kilka lat temu. Według mnie, jest to bardzo komplementarne z podejściem informatyki czyli danologii i sztucznej inteligencji, w których możliwość obliczenia zależności jest zwykle ważniejsza niż zrozumienie lub wytłumaczenie zjawisk. Owa – w pewnym sensie – *ułomność* informatyki skutkowała powstaniem w jej ramach nowej, eksplorowanej od kilku lat dziedziny zwanej wyjaśnialną sztuczną inteligencją (*XAI – explainable artificial intelligence*). Informatyka wychodzi więc od obliczeń, zaś fizyka od proponowanych i testowanych wzorów – zależności. Mocno upraszczając, można stwierdzić, że fizyka wierzy we wzory zaś informatyka w dane i obliczenia na nich wykonywane. Istota prawdy i poznania jest jednak gdzieś pomiędzy. Różne dyscypliny nauki poszukują jej używając właściwych sobie narzędzi. Z tej perspektywy niniejsza rozprawa idzie w bardzo pożądanym kierunku. Autor bierze więc wzory i prawa znane w fizyce, w tym przypadku prawo fluktuacji w układach złożonych, konkretniej prawo skalowania fluktuacji Taylora i weryfikuje je z danymi zwłaszcza pochodzącymi z rzeczywiście istniejących systemów, tj. mediów cyfrowych lub internetowych. Podejście danologii, uczenia maszynowego i ogólniej informatyki wystartowało by inaczej, tj. od zbudowania modelu uczonego na danych, a następnie próbowało go wyjaśnić, opisać za pomocą możliwie prostych zależności wyabstrahowanych z wytrenowanego modelu. Oba podejścia powinny się spotkać, ale jednak rzadko ma to miejsce. Między innymi dlatego, że bardzo trudno by było wydobyć wykładniczą zależność testowaną w pracy, np. z lasów losowych lub głębokiej sieci neuronowej. Owego spotkania nie ułatwia zbyt daleko idąc izolacja zespołów badawczych i mała liczba prac interdyscyplinarnych.

Idąc jednak dalej w głąb, recenzowana praca dotyczy rozprzestrzeniania się wiadomości a ogólniej informacji, w szczególności ich powielania i przekazywania dalej. Jest to proces istniejący od zawsze w ludzkich społecznościach – patrz procesy gromadzenia wiedzy poprzez przekazywanie informacji z ust-do-ust. Obecnie, wciąż stosunkowo nowe medium jakim jest Internet oraz media społecznościowe spowodowały, że dokonuje się to na masową i globalną skalę, zaś duża dostępność ułatwia gromadzenie i analizę danych. Do modelowania procesu rozprzestrzeniania wykorzystano w pracy probabilistyczny model niezależnych kaskad.

Idąc nieco w bok, Autor dotyka przy tym jeszcze jednego aspektu, tj. *zdarzeń* opisywanych przez wiadomości medialne. Identyfikacja tego, czy dana wiadomość jest związana z tym samym zdarzeniem jest blisko związana z dziedziną zwaną analizą tematyczną (*topic analysis*) lub ekstrakcją tematyki (*topic extraction*). W monografii zostały wykorzystane w tym celu dobre i

Politechnika Wrocławska

Wydział Informatyki i Telekomunikacji

Katedra Sztucznej Inteligencji

Wybrzeże Wyspiańskiego 27
50-370 Wrocław

T: +48 71 320 24 54

www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl

REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

narzędzia opracowane w Josef Stefan Institute w Słowenii w ramach dużego projektu i serwisu EventRegistry.

Prócz silnie interdyscyplinarnego, rozprawa posiada także charakter międzynarodowy, gdyż była realizowana we współpracy z naukowcami ze Słowenii i Wielkiej Brytanii, na danych dostarczonych przez Słoweńską Agencję Prasową. Było to możliwe dzięki udziałowi Doktoranta w europejskim projekcie RENOIR, co wprost unaocznia zyski jakie płyną z projektów międzynarodowych.

Podsumowując, praca dotyczy fluktuacji wiadomości w mediach online i praw rządzących tym zjawiskiem.

II. Hipotezy i cele pracy

Trzy hipotezy, trzy główne cele oraz siedem najważniejszych osiągnięć opisanych w monografii zostało wymienionych w pkt. 1.2. Dotyczą one zwłaszcza prawa temporalnego skalowania fluktuacji w złożonym systemie (*Temporal Fluctuation Scaling Law*). Prawo to inaczej jest nazywane przez Autora prawem skalowania fluktuacji Taylora, zaś owym złożonym systemem są wiadomości publikowane w cyfrowej przestrzeni medialnej przez ich wydawców. Wraz ze sfomułowaniem badanego prawa zaproponowano w pracy system wieloagentowy, dzięki któremu możliwe było przeprowadzenie symulacji działania tego prawa a w konsekwencji porównanie symulacji z rzeczywistością. W efekcie zbadano wielkości błędów pomiędzy wartościami wynikającymi z testowanego modelu a rzeczywistymi danymi dla różnych parametrów, w szczególności dla różnych wielkości okien czasowych oraz zakładanego i zdefiniowanego w pracy poziomu reaktywności, korzystając z analizy czynnikowej PCA. Autor poszedł tutaj dalej i zbadał ową reaktywność dla różnych kontekstów: krajów pochodzenia, politycznych preferencji i tematyk, tj. słów kluczowych. Jest to szczególne wartościowe, gdyż mieści się w ważnym nurcie badań, w którym bada się nie tylko ogólne zjawiska, ale także ich zróżnicowanie zależne od kontekstu. W uczeniu maszynowym takie analizy wyników czasami nazywa się *ablation study*.

III. Zawartość rozprawy

Rozprawa została napisana w języku angielskim a jego poziom jest dobry a tekst zrozumiały.

Pierwszy rozdział to wprowadzenie, hipotezy i cele rozprawy.

Kolejny, drugi rozdział zaczyna się od opisu wykorzystanych danych (dwa zbiory po kilkadziesiąt milionów wiadomości), z których wydzielono podzbiory dotyczące łącznie 31 tematów oraz serie czasowe z nimi związane. Wykorzystane dane zawierają metadane, zwłaszcza kraj i źródło pochodzenia (serwis). Serie dla zadanego serwisu i tematyki są podstawowym źródłem danych dla analizy prawa temporalnego skalowania fluktuacji.

Politechnika Wrocławska

Wydział Informatyki i Telekomunikacji

Katedra Sztucznej Inteligencji

Wybrzeże Wyspiańskiego 27
50-370 Wrocław

T: +48 71 320 24 54

www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl

REGON: 000001614
NIP: 896-000-58-51

Nr konta:
37 1090 2402 0000 0006 1000 0434

Trzeci rozdział to badania z wykorzystaniem wieloagentowego modelu, w którym pojedynczy agent odpowiada cyfrowemu serwisowi-wydawcy, tj. medium publikującemu wiadomości. Z dużego zbioru danych wydzielono prawie 500 artykułów, które zostały ręcznie oznaczone jedną z czterech klas przez czterech pracowników Słoweńskiej Agencji Prasowej. Ze względu na niezgodności między anototatorami oraz małe liczności niektórych klas rozważano także różne połączenia klas. Owe klasy odpowiadają różnym relacjom względem źródła pochodzenia dwóch rozważanych tekstów. Zaproponowany model podobieństwa tekstów bazujący na worku n-gramów o różnych długościach i zadanym progu w istocie został potraktowany jako klasyfikator konfrontowany względem klasyfikacji dokonywanych przez czterech anotatorów (*ground truth*). Do walidacji zastosowano typowe miary jakości predykcji, w szczególności miarę F1, która jest często stosowana dla niezbalansowanych klas a z takimi mamy tutaj do czynienia. Dodatkowym, cennym wkładem tego rozdziału są badania nad identyfikacją sieci powiązań pomiędzy serwisami (wydawcami wiadomości) i kaskadami przepływów – zależności pomiędzy wiadomościami, korzystając z hierarchicznego grupowania aglomeracyjnego dokumentów poprzedzających w czasie zadany tekst. W efekcie możliwe było zbudowanie sieci społecznej wydawców mającej prawie 6 tys. węzłów. Na jej największej składowej z tysiącem węzłów przeprowadzono symulacje niezależnych kaskad rozprzestrzeniania się informacji. Do podstawowego modelu, w którym prawdopodobieństwo aktywacji jest określone przez wagę krawędzi, dodano nowy autorski czynnik atrakcyjności (nośności tematu) nazwany *hype*, który określa jak bardzo dana wiadomość zwiększa zainteresowanie wydawców lub inaczej ile razy bardziej wszyscy wydawcy są zmuszeni do 'zajęcia się zadanym tematem'. Porównano to z wynikami dla sieci losowych oraz sieci Barabasiego-Alberty o podobnych wymiarach i gęstości a także tych sieci z przetasowanymi krawędziami. Dla prawie wszystkich przypadków uzyskane kaskady były zgodne z prawem temporalnego skalowania fluktuacji. Wartościową obserwacją z tej części badań jest powiązanie wykładnika potęgowego owego skalowania z wielkością zastosowanego okna czasowego i parametru kaskad (str. 56).

Największy objętościowo rozdział czwarty zawiera analizę rezyduów – błędów pomiędzy wartościami wynikającymi z modelu skalowania fluktuacji a danymi rzeczywistymi. Łącząc średnie wartości różnic związanych z dynamicznym procesem oraz ich wariancje wykorzystano autorską modyfikację miary POLAR (*Power Law Residuals*) jako opisująca stabilność modelu zagregowaną po różnych rozmiarach okna czasowego. W efekcie przeprowadzonych eksperymentów stwierdzono większa zależność owej miary od wielkości okien niż od rodzaju rozpatrywanej tematyki. Stosując analizę czynnikową PCA wyodrębniono m.in. pierwszą składową główną, która została zdefiniowana przez Autora jako *reaktywność* wydawcy w ramach zadanej tematyki. Innymi słowy –w pewnym uproszczeniu – reaktywność oznacza jak bardzo dany wydawca odstaje od innych w dynamice

Politechnika Wrocławska

Wydział Informatyki i Telekomunikacji

Katedra Sztucznej Inteligencji

Wybrzeże Wyspiańskiego 27
50-370 Wrocław

T: +48 71 320 24 54

www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl

REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

publikowania wiadomości na zadany temat na przestrzeni dłuższego okresu podzielonego na wiele okien, przy czym odstępstwa określane są względem modelu temporalnego skalowania fluktuacji wyznaczającego wzorcową zmienność (dynamikę). Oznacza to, że bez zastosowania owego modelu zaproponowanego przez Autora, wyznaczenie tej miary i dalsze analizy byłyby niemożliwe. Spora część tego rozdziału to badania statystyczne zwłaszcza różnego rodzaju korelacji, w tym dla dodatkowego procesu oczyszczania danych. Interesujące są tutaj analizy zależności błędów dla wydawców z różnych krajów, co określa jak bardzo wydawcy danego kraju są różni od innych w ramach zadanej tematyki. Podobne podejście zastosowano do poglądów politycznych, grupując je ze względu na podobną stronniczość polityczną, wykorzystując w tym celu metadane z serwisu EventRegistry, który z kolei korzysta z serwisu mediabiasfactcheck.com. W efekcie możliwe było porównanie wzorców zachowań mediów cyfrowych o różnych orientacjach politycznych. Zaproponowana miara reaktywności ma pewne cenne właściwości wynikające bezpośrednio z jej definicji. W szczególności wymieniłbym tutaj niezależność od samej aktywności. Dzięki temu można porównywać małych wydawców z dużymi, publikującymi wiele wiadomości wydawcami oraz tematy bardzo i mało popularne. Cennym dokonaniem jest tutaj także porównanie miary reaktywności do innych miar, zwłaszcza do miary zaproponowanej przez Fano, dla której owa zależność od poziomu aktywności jest znacząco większa. Wydaje się także, że reaktywność ma cechy lepiej dyskryminujące wielkość rozważanego okna czasowego, przynajmniej jest tak dla analizowanego zbioru danych i wybranych kontekstów.

IV. Oryginalne osiągnięcia

Praca zawiera wiele wartościowych osiągnięć, w szczególności:

1. Wykazano, że procesy publikacji wiadomości w przestrzeni internetowej można modelować za pomocą temporalnego skalowania fluktuacji.
2. Zaproponowano modyfikację modelu niezależnych kaskad poprzez dodanie wolnozmiennego mnożnika *hype* określającego zewnętrzną atrakcyjność wiadomości.
3. Utworzono i zbadano sieć społeczną wydawców korzystając z miary odległości pomiędzy publikowanymi przez wydawców wiadomościami. Dla tak zdefiniowanej sieci przeprowadzono symulacje procesów rozprzestrzeniania się informacji bazując na zmodyfikowanym modelu niezależnych kaskad, do czego także wykorzystano czasowy porządek ukazywania się podobnych wiadomości, potencjalnie oznaczających kolejne źródła (inspiracje do publikowania) albo posiadających inne ale wspólne źródło.
4. Zidentyfikowano trzy zakresy rozmiarów okien czasowych (<15min, >1 dnia i pomiędzy), które dają istotnie różne spojrzenie na zachodzące zjawiska fluktuacji. Zwłaszcza dotyczy to

Politechnika Wrocławska

Wydział Informatyki i Telekomunikacji

Katedra Sztucznej Inteligencji

Wybrzeże Wyspiańskiego 27
50-370 Wrocław

T: +48 71 320 24 54

www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl

REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

wkład poznawczy, metodologiczny a także praktyczny, gdyż w wydzielonym zakresie identyfikacji źródeł jest stosowana przez agencje prasowe. Treść jest metodologiczne poprawna i mieści się w aktualnych kierunkach badań na świecie.

W związku z powyższym stwierdzam, że opiniowana rozprawa doktorska mgr inż. Jana Chołoniewskiego spełnia wymagania stawiane w obowiązujących przepisach ustawy o stopniu naukowym doktora i wnoszę o dopuszczenie jej Autora do publicznej obrony.

Biorąc pod uwagę wysoki poziom treści rozprawy a także weryfikację osiągniętych rezultatów w dobrych czasopismach naukowych proponuję rozważenie jej wyróżnienia.

Politechnika Wrocławska

Wydział Informatyki
i Telekomunikacji

Katedra Sztucznej Inteligencji

Wybrzeże Wyspiańskiego 27
50-370 Wrocław

T: +48 71 320 24 54

www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl

REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434