Dr hab. Maria Ganzha

Instytut Badań Systemowych

Polskiej Akademii Nauk

Newelska 6,

01-447 Warszawa

Warszawa, 24.09.2023

Recenzja

Rozprawy doktorskiej mgr. Wojciecha Włodarczyka

zatytułowanej

„Modele ewaluacji poprawności danych pozyskanych metodą crowdsourcingu”

1. Tematyka rozprawy

Przedmiotem, przedstawionej do oceny, rozprawy doktorskiej Pana mgr. Włodarczyka są zagadnienia związane z crowdsourscingiem (określenie to nie ma dobrego polskiego odpowiednika, więc będzie stosowane w niniejszej recenzji), czyli procesem, który polega na angażowaniu się szerokiej grupy ludzi do wykonywania zadań, lub proponowania rozwiązań, w obszarach określonych przez „organizatora procesu crowdsourcingu”. W ostatnich latach crowdsourcing jest dosyć szeroko wykorzystywany przez firmy, instytucje publiczne oraz organizacje non-profit. W istniejących scenariuszach, crowdsourcing stanowi źródło finansowania przedsięwzięć lub zastępuje tradycyjnych pracowników. Sporą zaletą crowdsourcingu jest nie tylko obniżenie kosztów realizacji zadań, ale również możliwość pozyskiwania (bardziej) reprezentatywnych i różnorodnych danych, reprezentujących tak zwaną „wiedzę tłumu”.

Przykładami wdrożonego crowdsoursingu są znane aplikacje, takie jak Tripadvisor, Yanosik i ViVino. Niestety, zastosowanie crowdsourcingu nie zawsze kończy się sukcesem. W historii crowdsorcingu znane są przykłady niepowodzeń zastosowania tego procesu, takie jak wybór nazwy nowego napoju z grupy Mountain Dew lub pomysł na „Pepsi Refresh”, oba należące do firmy PepsiCo. Istnieje również wiele prac (naukowych i praktycznych) poświęconych tematyce crowdsourcingu. W większości dotyczą one oceny ryzyka stosowania crowdsourcingu, problemom zarządzania procesem, lub też oceną zebranych pomysłów i/lub rozwiązań.

Przedstawiona do recenzji rozprawa doktorska skupia się jednak na fundamentalnych i technicznych zagadnieniach związanych z jakością pozyskiwanych danych. W szczególności, autor zajął się problemem jakości pozyskiwanych danych lingwistycznych, sprawdzając wpływ informacji zwrotnej na jakość pozyskiwanych danych. W rozprawie zaproponowane zostały algorytmy, służące do tworzenia i dostarczania informacji zwrotnych. Skuteczność tych

algorytmów została oceniona empirycznie na podstawie przeprowadzonych eksperymentów. Tak więc rozprawa doktorska jednoznacznie należy do dziedziny Nauki Matematyczne, dyscyplina Informatyka.

2. Ocena treści rozprawy i wkładu oryginalnego

2.1. Treść rozprawy

Rozprawa została przygotowana w języku polskim, składa się z 5 rozdziałów oraz Wprowadzenia i ma objętość 195 stron, bez bibliografii.

Wprowadzenie poświęcone jest nieformalnemu wyjaśnieniu pojęcia crowdsourcingu, oraz celu przeprowadzonych badań.

Formalną definicję corwdsourcingu znajdujemy rozdziale pierwszym, gdzie zostały również zdefiniowane pojęcia potrzebne w dalszych częściach rozprawy, takie jak: uczestnicy procesu crowdsourcingu, platforma craowdsourcingowa, czy proces crowdsourcingu. Ponadto każdy z przedstawionych kroków oraz poszczególne elementy procesu crowdsourcingu, zostały formalnie opisane. Procedury, składające się na proces zostały przedstawione w postaci pseudokodów. Następnie, na podstawie obszernego przeglądu literatury, Doktorant opisał trzy podstawowe kierunki optymalizacji procesu crowdsourcingu:

- optymalizacja kosztu,

- optymalizacja czasu,

- optymalizacja jakości.

W dalszej części rozdziału pierwszego, Doktorant po krótce przedstawił stan badań nad każdym z wymienionych typów optymalizacji, zaznaczając, że w przedstawionej rozprawie skupia się na problemach związanych z jakością danych. W związku z tym, w kolejnych sekcjach rozprawy, Doktorant omówił metody oraz techniki kontroli jakości danych, używając taksonomii zaproponowanej przez Daniel et al.[1] Zostały tam również opisane i przeanalizowane przykłady wykorzystania informacji zwrotnej w procesie zbierania danych metodą crowdsourcingu. W tym kontekście zostało zdefiniowane pojęcie informacji zwrotnej i omówiona została klasyfikacja informacji zwrotnej. Klasyfikacja ta została rozszerzona przez Doktoranta o dwa dodatkowe wymiary: *moment efektu* oraz *kanał komunikacji*. Natomiast, istniejący wymiar *format* został pominięty w prowadzonych badaniach, w związku z brakiem w literaturze przykładów użycia informacji zwrotnej w formacie innym niż tekstowy. Ostatecznie, wymiary informacji zwrotnej, zastosowane przez Doktoranta w badaniach to *czas, źródło treści, szczegółowość, liczba ocen, moment efektu oraz kanał komunikacyjny*. Następnie, Doktorant przeanalizował zawartość prac, dotyczących badania możliwości użycia informacji zwrotnej. Na podstawie tego stwierdzone

[1] Danile F., Kucherbaev P., Cappiello C., Benatallah B., Allahbakhsh M, Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions, ACM Comput. Surv., 51(1), 2018.

zostało, że brak jest prac, w których byłoby rozpatrzona kwestia skuteczności informacji zwrotnej dla różnego typu mikro-zadań, a w szczególności mikro-zadań dotyczących anotacji danych lingwistycznych. W wyniku analizy literatury stwierdzono również, że kwestie jakości przekazywanej informacji zwrotnej były rzadko brane pod uwagę. Warto podkreślić, że Doktorant zebrał i usystematyzował eksperymenty, które zostały przedstawione w przeanalizowanych pracach, dotyczących użycia informacji zwrotnej i przedstawił otrzymane wyniki w postaci tabeli, co znacząco zwiększyło czytelność przedstawionego stanu wiedzy, we wspomnianej dziedzinie.

Rozdział drugi, poświęcony jest opisowi mechanizmów, używanych do implementacji procesu generowania informacji zwrotnej w przypadku crowdsourcingu. Rozpatrywane są tutaj mechanizmy nauczania maszynowego (ang. machine teaching), oraz techniki jakie mogą być włączone w proces kontroli jakości danych, zbieranych przy pomocy crowdsourcingu. Opisane zostało również jak poszczególne elementy procesu nauczania maszynowego mogą realizować trzy główne cele kontroli jakości w metodzie crowdsourcingu, czyli model jakości (modele reprezentacji ucznia), ocena jakości (metody obliczania parametrów modelu ucznia), oraz zapewnienie jakości (metody wyboru sygnałów uczących). W dalszej części rozdziału Doktorant przedstawił metody i techniki oceny każdego z tych trzech modeli. Rozważania te stanowią podstawę formalną zaplanowania, wykonania i oceny wyników eksperymentów, opisanych w rozdziałach 3 i 4. Biorąc pod uwagę, że w analizowanych i cytowanych artykułach sporo elementów zostało niewyjaśnionych, czy też były one opisane na bardzo ogólnym poziomie, Doktorant wykonał istotną pracę, systematycznie uzupełniając nieoczywiste aspekty stosowania nauczania maszynowego o niezbędne elementy, takie jak model ucznia, algorytmy, niezbędne do obliczenia parametrów modelu ucznia, jak również algorytmy wyboru tak zwanego „sygnału nauczającego”.

Następny rozdział (Rozdział 3) poświęcony jest badaniu wpływu informacji zwrotnej na jakość danych lingwistycznych. Na początku rozdziału zostały przedstawione dwie hipotezy, które były sprawdzane w ramach przeprowadzonych eksperymentów:

H1. Zapewnienie synchronicznej informacji zwrotnej w pozytywny sposób wpływa na jakość pozyskiwanych danych lingwistycznych w procesie crowdsourcingu.

H2. Jakość przekazywanej informacji zwrotnej w procesie crowdsourcingu ma pozytywny wpływ na jakość pozyskiwanych danych.

Hipotezy te zostały zweryfikowane dla sześciu różnych zbiorów danych. Razem ze sformułowanymi do weryfikacji hipotezami, uzyskane zostały odpowiedzi na trzy pytania badawcze:

P1. Czy informacja zwrotna w procesie crowdsourcingu ma efekt długoterminowy?

P2. Czy informacja zwrotna ma wpływ na szybkość tworzenia anotacji w procesie crowdsourcingu?

P3. Czy informacja zwrotna w procesie crowdsourcingu pozytywnie wpływa na zaangażowanie anotatorów?

Przed omówieniem eksperymentów i ich wyników, opisane zostały wykorzystane w badaniach zbiory danych (w tym ich źródła), proces rekrutacji uczestników eksperymentu, oraz platformy użyte w eksperymentach. Warto zaznaczyć, że żadna z istniejących (ogólnodostępnych) platform do crowdsourcingu nie dostarcza mechanizmu, który pozwalałby na niezbędną modyfikację procesu anotacji, zmiany interfejsu, czy też na zebranie szczegółowych logów procesu, niezbędnych dla analizy wyników eksperymentów. W związku z tym Doktorant stworzył autorski system *Funcrowd*, zawierający dwie niezbędne funkcjonalności – mechanizm przekazywania dynamicznej informacji zwrotnej oraz mechanizm grupowania mikro-zadań. Integracja dwóch platform – autorskiej platformy *Funcrowd* oraz platformy *Amazon Mechanical Turk* – umożliwiła prowadzenie zaplanowanych eksperymentów. W tym kontekście należy pochwalić znaczącą ilość pracy koncepcyjnej włożonej w planowanie i przygotowanie eksperymentów. Dokładne opisy, krok po kroku, zawierające nie tylko diagram akcji, ale również szczegółową analizę przepływu informacji pomiędzy użytymi platformami sprawiają bardzo dobre wrażenie i świadczą o dojrzałości badawczej Doktoranta.

Następnie, w Rozdziale trzecim, omówiona została organizacja eksperymentów: dla grupy testowej, która otrzymywała informację zwrotną różnej jakości (wysokiej, średniej oraz niskiej) oraz dla grupy kontrolnej, która nie otrzymywała informacji zwrotnych. Zadania, przydzielane w grupie testowej były takie same, co w grupie kontrolnej.

Kolejnym elementem niezbędnym do przeprowadzenia rzetelnych eksperymentów są dane. Doktoranta przygotował 6 różnych zbiorów danych: skargi wobec usług bankowych, atrybuty produktów eBay, waga produktów eBay, wydźwięk opinii o hotelach, jednostki nazwane (ang. Named Entity) oraz wyrazy bliskoznaczne. Są to różnorodne zbiory, co w istotny sposób uwiarygadnia przeprowadzoną analizę i generalizowalność wyników przeprowadzonych eksperymentów. Każdy ze zbiorów danych zawierał po 2000 elementów z wyjątkiem zbioru „waga produktów eBay", który zawierał 1000 elementów. Dla każdego ze zbiorów przygotowany został specjalny (indywidualny) protokół informacji zwrotnej (zgodny z semantyką zbioru), oraz przygotowane informacje zwrotne średniej i niskiej jakości. Aby takowe otrzymać, zastosowane zostały algorytmy zniekształcenia informacji zwrotnej, które redukowały dokładność informacji (ang. accuracy) odpowiednio do 0.75 i 0.55.

W ostatniej części rozdziału Doktorant opisał przeprowadzone eksperymenty i dokonał analizy otrzymanych wyników. Najpierw opisana została metodologia, zastosowana w celu weryfikacji badanych hipotez. Dodatkowo przedstawione zostały zbiory metryk jakości, wyspecyfikowane dla każdego zbioru danych. Podsumowaniem rozdziału są generalizacje wyników analizy rezultatów. Otrzymane wnioski są dosyć ciekawe, czasem nawet nieoczekiwane, podpowiadające pewne „oszczędne" strategie, możliwe przy stosowaniu metody crowdsourcing (na przykład w przypadku przekazywania informacji zwrotnej o obniżonej jakości). Pozytywnym wnioskiem (choć ten wniosek jest w jakiś sposób oczekiwany) jest to, że informacja zwrotna poprawia jakość

zbieranych danych. Dosyć interesującym wynikiem jest natomiast to, że poprawa jakości ma miejsce natychmiastowo, a ponadto utrzymuje się ona również długoterminowo.

W związku z obiecującymi wynikami eksperymentów, w Rozdziale czwartym zaproponowany został autorski model *Dynamicznej Informacji Zwrotnej (DIZ)*. Model ten bazowany jest na wprowadzonym przez Doktoranta modelu ucznia. Na początku rozdziału czwartego Doktorant, zdefiniował formalnie model DIZ oraz przedstawił, w postaci diagramu i pseudokodu, proces jego stosowania.

Następnie, Doktorant skupił się na weryfikacji skuteczności zastosowania modelu DIZ w procesie crowdsorcingu. W tym kontekście sformułowane zostały dwa pytania badawcze, dotyczące zaproponowanego modelu DIZ, a mianowicie:

P1: Czy zastosowanie modelu DIZ pozwala na wygenerowanie informacji zwrotnej o wyższej jakości, w porównaniu do modelu referencyjnego?

P2: Czy jakość oznaczeń anotatorów wpływa na jakość informacji zwrotnej, generowanej przez model DIZ?

Następnie bardzo szczegółowo opisane zostały wszystkie elementy tworzonego modelu DIZ, wybrane metryki, wybrane elementy użytej sieci neuronowej (ogólna architektura takiej sieci został przedstawiona w postaci diagramu), które zostały użyte do implementacji modelu DIZ w celu przeprowadzenia eksperymentów weryfikujących. Eksperymenty te, ponownie, zorganizowane zostały dla dwóch wariantów – z informacją zwrotną i bez niej (grupa kontrolna) – dla opisanych wcześniej zbiorów danych. Ponadto zdefiniowane zostały dwa dodatkowe parametry – liczba treningowych mikro-zadań S oraz jakości anotacji q, które mają poważny wpływ na przebieg eksperymentów.

Analiza wyników eksperymentu, przeprowadzonego dla danych empirycznych, nie wykazała skuteczności modelu DIZ, rozumianej jako poprawy jakości generowanej informacji zwrotnej. Natomiast, dla danych symulacyjnych już tak, zwłaszcza przy wysokiej jakości anotacji.

W pewnym sensie jest to wynik negatywny, ale warty uzyskania dlatego, że jest on empirycznie zweryfikowaną odpowiedzią na istotne pytania badawcze.

Ostatni rozdział rozprawy doktorskiej – Podsumowanie – jest bardzo krótki, w zasadzie dwu stronicowy. Zawiera on podsumowania działań badawczych Doktoranta oraz plany na przyszłe badania. W planach tych centralne miejsce zajmuje system Funcrowd, który ma stać się centralnym elementem rozwijanej przez Doktoranta platformy crowdsourcingowej, pozwalającej na zastosowanie informacji zwrotnej. Jednym z kroków już dokonanych przez Doktoranta – jest wdrożenie systemu w projekcie *„Sprawdzamy jak jest”*, w którym anotatorzy nieodpłatnie weryfikują dokumenty przesłane przez polskie instytucje publiczne.

2.2. Wkład oryginalny

Najważniejsze samodzielne i oryginalne osiągnięcia Doktoranta to:

1. Całościowa analiza procesu crowdsourcingu, w tym formalne definiowanie składowych tego procesu, sposobów optymalizacji procesu crowdsourcingu, zaczynając od analizy elementów wpływających na jakość, kończąc strategiami usprawniającymi proces crowdsourcingu. Zostali również formalnie zdefiniowani uczestnicy procesu crowdsourcingu (*Zleceniodawca* oraz *Anotator*). Opracowany został także ogólny schemat procesu corwdsourcingu. Każdy krok tego procesu został następnie opisany za pomocą pseudokodów oraz uzupełniony o wyjaśniające przykłady.
2. Zaproponowana modyfikacja taksonomii cech informacji zwrotnej. Po pierwsze, zostało zdefiniowane pojęcie informacji zwrotnej. Po drugie, zostało zaproponowane rozszerzenie taksonomii cech informacji zwrotnej o dwa nowe wymiary: *moment efektu* (natychmiastowy / długoterminowy) oraz *kanał komunikacji (bezpośredni / pośredni)*. Wymiary te nie były uwzględnione w istniejącej dotychczas taksonomii. Jednakże, badania literaturowe jak i wykonana praca badawcza Doktoranta, potwierdzają istotny wpływ tych dwóch charakterystyk na jakość danych, pozyskiwanych w procesie crowdsourcingu.
3. Gruntowna analiza istniejących implementacji informacji zwrotnej w systemach crowdsourcingowych. Analiza została uzupełniona podsumowującą tabelą, w której między innymi można zaobserwować sensowność wprowadzenia dwóch dodatkowych charakterystyk informacji zwrotnej, zaproponowanych przez Doktoranta. Analiza ta jest poprzedzona bardzo szczegółowym zbadaniem wymiarów modeli jakości – Dane, Zadania, Uczestnicy, oraz rozmaitymi metodami poprawy jakości łącznie z takimi jak wybór anotatorów, zapewnienie motywacji, czy szkolenie anotatorów.
4. Ciekawe i pomysłowe zastosowanie mechanizmu „nauczania maszynowego" do implementacji procesu przekazywania informacji zwrotnej, jako mechanizmu kontroli jakości. W tym celu proces nauczania maszynowego został po pierwsze, zmapowany na poprzednio przeanalizowany i rozłożony na składowe proces informacji zwrotnej: model reprezentacji ucznia – model jakości; model obliczania parametrów ucznia – komponent oceny jakości, metody wyboru sygnałów nauczających – zapewnienie jakości.
5. Zaprojektowanie i realizacja środowiska do przeprowadzenia eksperymentów z wykorzystaniem rzeczywistych anotatorów rejestrujących się na rzeczywistej platformie crowdsourcingowej, czyli stworzenie zintegrowanego systemu składającego z platformy *Amazon Mechanical Turk* oraz autorskiej platformy *FunCrowd*. Trzeba podkreślić jako element bardzo pozytywny, że platforma FunCrowd nie została stworzona tylko i wyłącznie do pracy doktorskiej – jest ona również używana w projekcie *„Sprawdzamy jak jest"*, w którym anotatorzy nieodpłatnie weryfikują dokumenty przesłane przez polskie instytucje publiczne. Bardzo pozytywnie świadczy to o jakości stworzonej platformy.
6. Zaproponowanie autorskiego modelu *Dynamicznej Informacji Zwrotnej (DIZ).* Zadaniem modelu jest generowanie informacji zwrotnej w sposób automatyczny. Główna cecha, która odróżnia DIZ od istniejących rozwiązań jest to, że model DIZ został oparty na modelu ucznia. Dodatkowo, w celu podwyższenia jakości przekazywanej informacji zwrotnej, DIZ używa aktualnej anotacji danych.

2.3. Ocena zawartości pracy oraz uwagi polemiczne

Jak już zostało wspomniane, w rozprawie Doktorant podjął próbę stworzenie gruntownej analizy procesu crowdsourcingu, istniejących modeli kontroli jakości, oraz próbę usprawnienia procesu zbierania danych za pomocą przekazywania informacji zwrotnej jak i automatyzacji tegoż procesu.

Główna wątpliwość wzbudza sposób zastosowanie nauczania maszynowego. W pewnym sensie, całe nauczanie maszynowe zostało sprowadzone do tak zwanego semi-supervised learning. Z mojego punktu widzenia próby podkreślenia różnicy między nauczaniem maszynowym a uczeniem częściowo nadzorowanym nie wykazały jednak takowej.

To podobieństwo również nasuwa następujące pytanie: czy Doktorant w swoich badaniach brał pod uwagę Uczenie ze Wzmocnieniem (ang. Reinforcement Learning) w celu wytwarzania jak najlepszej jakości informacji zwrotnej. Podejście to – uczenie ze wzmocnieniem – jeszcze bardziej naturalnie można byłoby zmapować z procesem Nauczania maszynowego. W posiadaniu Doktoranta był (i wciąż jest) unikalny materiał – dane, zawierające również informację o wpływie informacji zwrotnej na anotacje. Oczywiście wymagałoby to pogłębionej pracy nad rozważaniem jaką metodę zastosować oraz nad stworzeniem funkcji nagrody. Warto jednak zauważyć, że tworzenie modelu nauczania maszynowego z zastosowaniem uczenia głębokiego do tworzenia informacji zwrotnej to tak naprawdę już prawie element Uczenia ze Wzmocnieniem, stosujące uczenie głębokie (ang. Deep Reinforcement Learning). Niestety, brak wspomnienia o takim podejściu nawet w rozważeniu o możliwych (przyszłych) podejściach do rozwiązania problemu.

Kolejny problem pracy to spora ilość literówek i to nie tylko „tekstowych", ale również we wzorach. Wprowadza to spore utrudnienia w zrozumieniu tego, co dokładnie zostało zmodyfikowane (przykładowo: Procedura 4 vs Rysunek 2.7, wzór (2.3), wzór (2.4), w którym z nieznanych powodów pojawia się „do kwadratu" lub też wzór opisujący problem ucznia).

Konkluzja końcowa

Oceniając przedstawioną do recenzji Rozprawę Doktorską, należy zwrócić uwagę na kilka ciekawych pomysłów usprawniających proces zbierania danych metodą crowdsourcingu, co jest bardzo cennym osiągnieciem biorąc pod uwagę szybko rosnącą popularność tej metody. W rozprawie zostały również zaproponowane interesujące algorytmy oraz przeprowadzone wszechstronne badania zaproponowanych usprawnień, które mają bardzo wysoką wartość praktyczną. Tym bardziej, że jednym z wyników tej pracy jest działająca i wciąż rozwijana platforma *FunCrowd*, używana do rzeczywistych zastosowań. Dodatkowo, wszystkie eksperymenty były wykonane na danych rzeczywistych, czyli zebranych z rzeczywistych procesów anotacyjnych, wykonywanych przez rzeczywiste osoby. Takie dane znacznie uwiarygodniają wyniki eksperymentów.

Podsumowując, uważam, że hipotezy, postawione przez autora rozprawy, zostały jednoznacznie potwierdzone.

Wobec powyższego stwierdzam, że rozprawa doktorska mgr. Wojciecha Włodarczyka spełnia warunki stawiane przez ustawę o tytule naukowym i stopniach naukowych w odniesienie do rozpraw doktorskich, a zatem powinna być dopuszczona do publicznej obrony o co wnoszę do Rady Dyscypliny Naukowej Matematyka i Informatyka Uniwersytetu Adama Mickiewicza.

Elektronicznie
podpisany
przez
Maria Ganzha
2023.09.24
12:28:13