dr hab. inż. Mikołaj Morzy, prof. PP

Instytut Informatyki Politechniki Poznańskiej

Recenzja rozprawy doktorskiej

Machine Learning Tools and Techniques Supporting News Media Bias Analysis

Autorka rozprawy: **Katarzyna Baraniak**

Jakie zagadnienie naukowe jest rozpatrzone w pracy (teza pracy) i czy zostało ono dostatecznie jasno sformułowane przez Autorkę? Jaki charakter ma praca (teoretyczny, doświadczalny, inny)?

Przedstawiona do recenzji rozprawa doktorska jest poświęcona zjawisku stronniczości w prezentowaniu wydarzeń politycznych, społecznych i gospodarczych w prasie internetowej. Stronniczość w przedstawianiu postaci, partii politycznych i innych bytów w mediach, zwłaszcza w gazetach i na portalach internetowych, jest problemem, który niepokoi wielu obserwatorów społecznych. Wynika to często z określonej agendy politycznej, którą reprezentuje dane źródło informacji. Ten brak obiektywizmu i polityczna demagogia mogą mieć bardzo negatywny wpływ na funkcjonowanie demokratycznych społeczeństw, wpływając na wybory wyborcze, politykę publiczną i relacje międzynarodowe.

Przykładowo, niektóre duże ośrodki informacyjne, takie jak Fox News w Stanach Zjednoczonych czy Russia Today w Rosji, są często krytykowane za nierzetelność i stronniczość w swoim raportowaniu. Ich wiadomości często skupiają się na promowaniu określonej agendy politycznej, zamiast dostarczania obiektywnych i zrównoważonych informacji. Stronniczość prasy może prowadzić do szeregu szkód społecznych. Może na przykład wzmacniać podziały polityczne i społeczne, wpływać na opinie publiczną w sposób, który nie odzwierciedla rzeczywistości, a także zniechęcać ludzi do zaangażowania się w procesy demokratyczne. W skrajnych przypadkach, może nawet prowadzić do konfliktów i przemocy. W związku z tym, istotne jest promowanie mediów, które są zobowiązane do obiektywności, precyzji i zrównoważonego przedstawiania różnych punktów widzenia. Recenzowana praca przedstawia wyniki różnorodnych badań prowadzonych przez Autorkę, dotyczących możliwości zastosowania nowoczesnych metod uczenia maszynowego i przetwarzania języka naturalnego do detekcji stronniczości w artykułach i komunikatach internetowych.

Głównym problemem w detekcji stronniczości w prasie jest wyzwanie związane z precyzyjnym zdefiniowaniem samego pojęcia stronniczości. Stronniczość jest zjawiskiem subiektywnym i wielowymiarowym, które może przybierać różne formy i manifestować się na różne sposoby, w

zależności od kontekstu. Ta niejednoznaczność i złożoność sprawiają, że budowa reprezentatywnych zbiorów danych, które mogłyby służyć do wytrenowania modeli klasyfikacyjnych, jest zadaniem niezwykle trudnym. Wymaga to uwzględnienia szerokiego zakresu czynników, takich jak ton, kontekst, wybór tematów, sposób przedstawiania faktów, a nawet to, co jest pomijane. Co więcej, ocena i etykietowanie tych danych byłyby prawdopodobnie obarczone własnymi formami stronniczości ze strony osób dokonujących oceny. Ta skomplikowana dynamika sprawia, że detekcja stronniczości w prasie jest trudnym wyzwaniem w dziedzinie analizy mediów i sztucznej inteligencji.

Recenzowana praca adresuje ten problem poprzez wprowadzenie nowego zbioru danych o nazwie "Sentiment concerning Entity in News headline" zawierającego dane w językach polskim i angielskim. Zbiór zawiera rzeczywiste nagłówki gazetowe pobrane z różnorodnych źródeł internetowych. Zbiór ten został poddany adnotacji eksperckiej i amatorskiej i udostępniony publicznie. Wewnątrz zbioru danych wyodrębniono listę ok. 30 nazwanych encji, reprezentujących przede wszystkim osoby publiczne, ale też partie lub nazwy krajów. Na bazie tego zbioru przeprowadzono eksperymenty przy użyciu popularnego modelu BERT nad możliwością automatycznej klasyfikacji sentymentu. W ramach eksperymentów opracowano też kilka modyfikacji oryginalnej architektury BERT-a pod kątem zwiększenia "uwagi" przypisywanej przez model analizowanym encjom. Biorąc pod uwagę zakres i tematykę przeprowadzonych prac można stwierdzić, że przedstawiona do recenzji rozprawa doktorska wpisuje się w aktualne trendy badawcze, starając się dostarczyć nowych narzędzi do analizy stronniczości w dyskursie publicznym. Poruszone w rozprawie tematy stanowią ważne cele badań naukowych i spełniają wymagania stawiane zwyczajowo rozprawom doktorskim. Rozprawa ma charakter wybitnie eksperymentalny.

Rozprawa doktorska jest napisana w języku angielskim i liczy 116 stron, wliczając w to bibliografię, listę rycin i listę tabel. Do rozprawy nie dołączono streszczenia w języku polskim, co stanowi naruszenie art. 187 ust. 4 ustawy Prawo o szkolnictwie wyższym i nauce z dnia 20 lipca 2018 r., w którym stwierdza się:

> Do rozprawy doktorskiej dołącza się streszczenie w języku angielskim, a do rozprawy doktorskiej przygotowanej w języku obcym również streszczenie w języku polskim. W przypadku gdy rozprawa doktorska nie jest pracą pisemną, dołącza się opis w językach polskim i angielskim.

Bibliografia zawiera 106 pozycji, spośród których cztery prace są współautorswa Doktorantki (jest ona pierwszą autorką). Rozprawa została podzielona na trzy części. Rozdział 1 otwiera część pierwszą i stanowi wprowadzenie do rozprawy, zdefiniowanie problemu adresowanego w rozprawie, oraz podsumowanie głównych kontrybucji Autorki. Rozdział 2 stanowi przegląd literatury naukowej związanej z poruszanym w pracy tematem. Rozdział 3 to bardzo obszerne przedstawienie szerokiego spektrum metod, algorytmów i narzędzi związanych z przetwarzaniem języka naturalnego. Część druga rozprawy to rozdziały 4-6. W rozdziale 4 Autorka przedstawia wyniki eksperymentów przeprowadzonych na danych pobranych z silnie spolaryzowanych źródeł,

a celem eksperymentów jest ilościowa analiza związków między walencją emocjonalną a zbiorem nazwanych encji reprezentujących osoby, partie i inne byty z domeny politycznej. W rozdziale 5 przedstawiono główną kontrybucję rozprawy jaką jest zbiór SEN oraz zaprezentowano wyniki benchmarków przeprowadzonych dla różnych modeli na tym zbiorze. Rozdział 6 przedstawia wyniki metody opracowanej na potrzeby międzynarodowego konkursu SemEval 2023. Na część trzecią rozprawy składa się pół strony zawierającej krótkie podsumowanie rozprawy.

Czy Autorka rozwiązała postawione zagadnienia, czy użyła właściwej do tego metody i czy przyjęte założenia są uzasadnione?

Główną kontrybucją rozprawy jest kompilacja unikalnego zbioru danych dotyczącego stronniczości emocjonalnej w przekazach prasowych. Ponieważ ewaluacja przedstawionych w rozprawie rozszerzeń modelu BERT jest uzależniona od jakości zbioru danych SEN, można przyjąć, że to właśnie jakość zbioru SEN determinuje istotność i uniwersalność wyników prezentowanych przez Autorkę. I tu kryją się moje dwie główne wątpliwości, pierwsza dotyczy samej procedury kompilacji zbioru SEN, druga ma charakter bardziej metodologiczny.

W rozprawie nie przedstawiono szczegółowo procedury kompilacji zbioru SEN i nie zaprezentowano jego dogłębnej analizy. Nie wiadomo, jak duża była zgodność między adnotatorami - rozumiem, że w przypadku przyjętego schematu adnotacji nie można było podać współczynników takich jak κ Cohena albo κ Fleissa, ale można było chociażby podać procent instancji, w których występowała niezgodność adnotacji i trzeba było odwoływać się do głosowania większościowego. Nie zaprezentowano też pełnego protokołu adnotacji (chyba, że sprowadzał się on do punktów przedstawionych na str. 61). Od kilku lat standardem jest wyposażanie publicznych zbiorów danych w szczegółowe opisy meta-danych, służą do tego *datasheets* [1], *nutrition labels* [2] czy *data cards* [3]. Szkoda, że wraz ze zbiorem SEN nie przygotowano jednego z tych opisów, niewątpliwie znacząco podniosłoby to jakość i użyteczność zbioru danych. Należy też podkreślić, że zadanie adnotacji było niezwykle trudne i jestem przekonany, że sam zbiór danych jest wewnętrznie niespójny. W rzeczywistości nie istnieje żaden "złoty standard", żaden "prawdziwy" rozkład walencji emocjonalnej, który jest przybliżany przez etykiety. Określenie walencji emocjonalnej jest nie tylko trudne, ale obarczone nieusuwalnym szumem wynikającym z indywidualnych doświadczeń osoby dokonującej adnotacji, jej subiektywnego postrzegania wypowiedzi pozytywnych i negatywnych, oraz osobistego stosunku do osób opisywanych w adnotowanych danych. Szkoda, że podczas przygotowywania zbioru danych nie udało się przeanalizować go pod kątem spójności i jakości etykiet, np. korzystając z metody Confident Learning [4], dla której istnieje gotowe środowisko `cleanLab`.

Moja druga wątpliwość związana jest z oceną przyjętego podejścia, zgodnie z którym skompilowany zbiór danych (nawet pomijając kwestie jego jakości) jest wystarczającym zasobem do analizowania tak złożonego zjawiska jak stronniczość. Mam wrażenie, że recenzowana rozprawa doktorska wpisuje

się w tradycję badań, w których pomija się wszystkie aspekty poza-cyfrowe, od razu przechodząc do próby zaaplikowania różnych modeli uczenia maszynowego. Szczególnie w przypadku NLP, w którym posiadamy dostęp jedynie do "formy powierzchniowej" tekstu, bez dostępu do głębszego kontekstu, jest to problematyczne. Cytując Emily Bender [5] "[…] *the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning*." Postawiony w rozprawie cel identyfikacji stronniczości w przekazach prasowych nie może ograniczać się jedynie do płytkiej analizy "formy powierzchniowej", tj. nagłówka prasowego. Zjawisko stronniczości jest bardzo złożone, z wieloma graczami reprezentującymi różne, przeciwstawne grupy interesów. Na konkretną instancję stronniczości wpływ mają: aktualna sytuacja polityczna, meta-narracje wokół których krystalizują się tożsamości (które odwzorowują się na konkretne wybory polityczne), geopolityka wpływająca na pojawianie się dezinformacji, interesy gospodarcze różnych podmiotów, itd. Istnieje bardzo bogata literatura dotycząca form, jakie przyjmuje stronniczość medialna, oraz długa tradycja badawcza w obszarze nauk politycznych, społecznych czy medioznawstwa, która w recenzowanym doktoracie została całkowicie pominięta. Proste wyszukanie wskazuje na liczne prace w tym obszarze, niekoniecznie publikowane w szeroko pojętej informatyce:

- Weatherly, Jeffrey N., et al. "Perceptions of political bias in the headlines of two major news organizations." Harvard International Journal of Press/Politics 12.2 (2007): 91-104.
- Konnikova, Maria. "How headlines change the way we think." The New Yorker 17 (2014).
- Lott, John R., and Kevin A. Hassett. "Is newspaper coverage of economic events politically biased?." Public choice 160.1 (2014): 65-108.
- Muddiman, Ashley, Jamie Pond-Cobb, and Jamie E. Matson. "Negativity bias or backlash: Interaction with civil and uncivil online political news content." Communication Research 47.6 (2020): 815-837.
- Morstatter, Fred, et al. "Identifying framing bias in online news." ACM Transactions on Social Computing 1.2 (2018): 1-18.

Jestem głęboko przekonany, że uwzględnienie i zamodelowanie tych zjawisk, nawet w formie bardzo uproszczonego grafu wiedzy, nie tylko podniosłoby jakość trenowanych modeli, ale przede wszystkim pozwoliłoby na znacznie dalej idącą generalizację wniosków płynących z przeprowadzonych eksperymentów.

Trzeci problem, który pojawił się podczas lektury rozprawy, związany jest z brakiem formalnej definicji stronniczości. Mimo, że cały rozdział 4 jest zatytułowany "Visibility and Agenda bias", termin *bias* nie doczekał się w rozprawie żadnej definicji. Żartobliwie można tu nawiązać do słynnej definicji obsceniczności sformułowanej przez sędziego Sądu Najwyższego USA Pottera Stewarta ("*I know it when I see it*"), jednak biorąc pod uwagę fakt, że wszystkie zaprezentowane w rozprawie analizy mają charakter ilościowy, spodziewałbym się mimo wszystko próby formalnego zdefiniowania zjawiska, które ma podlegać badaniu.

Czy Autorka wykazała umiejętność poprawnego i przekonującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

Poziom redakcyjny rozprawy jest poprawny, poza kilkoma drobnymi błędami literowymi lub gramatycznymi tekst rozprawy jest czytelny. W drugiej części rozprawy Autorka wykorzystuje dużą liczbę symboli matematycznych i pomiędzy poszczególnymi sekcjami nie zawsze zachowuje spójność oznaczeń. Umieszczenie w pracy listy wykorzystywanych oznaczeń matematycznych pomogłoby zachować pełną spójność. W rozprawie, szczególnie w częściach opisowych, występuje wiele bardzo krótkich paragrafów, co utrudnia lekturę rozprawy. Paragraf nie jest elementem wizualnym, to element struktury tekstu służący do przekazania jednej spójnej myśli. Najczęściej wymaga to zdania wprowadzenia, kilku zdań rozwinięcia, oraz zdania podsumowującego paragraf i stanowiącego przejście do następnego paragrafu. Dziwi również wydzielenie trzeciej części rozprawy składającej się z pół strony tekstu. Biorąc pod uwagę całą strukturę rozprawy można chyba było swobodnie zrezygnować z podziału na części i pozostawić jedynie rozdziały.

Poniżej zamieszczam listę bardziej szczegółowych uwag do tekstu (liczba na początku punktu to numer strony w manuskrypcie, której dotyczy uwaga):

- 7: niektóre adresy URL są zamieszone bezpośrednio w tekście, niektóre w postaci przypisów, nie jest dla mnie jasne jakim kryterium kierowała się Autorka wybierając jedną lub drugą formę
- 13-17: w rozdziale “Related work” poszczególne pozycje literaturowe zostały opisane bardzo zdawkowo, najczęściej jednym zdaniem. Po tym rozdziale spodziewałbym się mniej wyliczenia prac z rozważanego obszaru, a raczej nakreślenia głównych kierunków badań, obowiązujących założeń, przyjmowanych podejść, itd. Poza tym istotną rolą rozdziału “Related work” jest wykazanie, gdzie w stosunku do aktualnego stanu wiedzy plasuje się rozprawa doktorska, jakie kierunki rozszerza, co stanowi główną innowację w stosunku do obecnego stanu wiedzy.
- 20: wagi w metodzie TF-IDF są ograniczone do zakresu <0,1> tylko w przypadku zastosowania normalizacji, oryginalne sformułowanie nie zapewnia żadnego ograniczenia. Wystarczy wyobrazić sobie dwa dokumenty: pierwszy zawiera 100 wystąpień słowa “Ala”, drugi zawiera sto wystąpień słowa “Ola”. Zakładając logarytm o podstawie 2, oba tokeny będą miały wagę 100 ($100 \cdot \log_2 \frac{2}{1}$)
- 20: definicja członu IDF jest błędna. Istnieje wiele różnych sformułowań metryki *inverse document frequency* (np. wygładzona, maksymalna, probabilistyczna), ale w każdej wersji uwzględniany jest logarytm częstości występowania termu w korpusie i wynika to z przyczyn teorioinformacyjnych
- 20 i dalej: korzystając z systemu Latex warto jest używać poprawnego formatowania tekstu w trybie matematycznym aby uniknąć niepoprawnych ligatur i spacji między literami, zatem nie $BERT$ tylko BERT ($BERT$ vs $BERT$)
- 23: rysunek sugeruje, że instancje mieszczące się wewnątrz marginesu granicy decyzyjnej to

zmienne tolerancyjne (?) (nie znalazłem polskiego tłumaczenia terminu *slack variable*), to może być nieco mylące bo przecież te zmienne to element algorytmu wyznaczania granicy decyzyjnej. Oczywiście punkty danych dla których zmienne przyjmują wartości dodatnie faktycznie trafiają do wewnątrz marginesu, więc rozumiem zamysł, może to jednak być źródłem nieporozumienia.

- 24: dwa lata temu nie miałbym problemu z określeniem modelu BERT mianem *large language model*, ale w roku 2023, w którym modele z 2 miliardami parametrów to skwantyzowane lub destylowane wersje znacznie większych modeli, BERT z 340 milionami parametrów jawi się jako bardzo mały model językowy.
- 29: precess the sequence $\rightarrow$ process the sequence
- 31-32: opis konwolucyjnych sieci neuronowych jest bardzo powierzchowny i w żaden sposób nie jest nawet nakierowany na zastosowanie CNN-ów do tekstu, poza tym w eksperymentach takie sieci nie są wykorzystywane stąd ten podrozdział mógłby być spokojnie pominięty bo niczego do rozprawy nie wnosi.
- 35: states of this two $\rightarrow$ states of these two
- 36: which is s dot $\rightarrow$ which is the dot
- 37: it relays $\rightarrow$ it relies
- 43-55: trochę dziwi umieszczenie w pierwszej kolejności raportu z analizy wystąpień encji w artykułach z dwóch portali, opisane w tym rozdziale wyniki nie są szczególnie odkrywcze i interesujące, na pewno stanowią znacznie mniej istotną kontrybucję niż zbiór SEN, trudno też stwierdzić w jakim stopniu stanowią badania wstępne dla wyników prezentowanych w rozdziale 5. W mojej opinii rozdziały 4 i 6 stanowią uzupełnienie rozprawy i prezentację dodatkowych, niekoniecznie kluczowych wyników, więc naturalniej byłoby zamienić rozdziały 4 i 5 kolejnością.
- 53: nie znalazłem w pracy informacji na temat tego, jak są zdefiniowane słowa kluczowe, na podstawie których wyznaczane jest podobieństwo między artykułami.
- 54: Tabela 4.9: zwykłe głosowanie większościowe daje dokładność na poziomie 0.95 więc wyniki dla poszczególnych klasyfikatorów należałoby zaprezentować względem takiej wartości referencyjnej, w przeciwnym wypadku czytelnik odnosi błędne wrażenie że modele sobie dobrze radzą z zadaniem (co sugeruje Autorka pisząc “All algorithms have quite good results”, podczas gdy w rzeczywistości algorytmy radzą sobie fatalnie o czym świadczą metryki dla klasy mniejszościowej)
- 54: articles was $\rightarrow$ articles were
- 55: dentoed $\rightarrow$ denoted
- 62-64: przyjęta metoda oceny stronniczości osób adnotujących jest w moim przekonaniu błędna, wyznaczana jest liczba wskazująca na relatywny stosunek nagłówków pozytywnych i negatywnych dla każdej encji, a następnie ta liczba jest agregowana dla wszystkich encji adnotowanych przez daną osobę. Dana osoba nie ma wpływu na to, czy do adnotacji otrzymuje nagłówki pozytywne czy negatywne ani z jakiego źródła pochodzą adnotowane nagłówki. W przypadku określenia stronniczości samych modeli językowych względem poszczególnych

encji należało posłużyć się testami behawioralnymi, dla których istnieją wygodne narzędzia, choćby `CheckList` czy `langtest`. W przypadku analizy obciążenia wprowadzanego przez poszczególne osoby dokonujące adnotacji można było skorzystać z podejść podobnych do Label Quality Model [6]

- 70: in th same $\rightarrow$ in the same
- 71: `spaCy` to biblioteka a nie baza danych
- 70: szkoda, że nie udało się uwzględnić w pracach badawczych modeli generatywnych, przeprowadziłem szybki eksperyment na zbiorze SEN-pl z wykorzystaniem modelu `gpt-3.5-turbo` i po pół godzinie pracy, praktycznie bez używania zaawansowanej inżynierii zachęt, dostrajania zachęty czy starannego wyboru przykładów typu *few shot* uzyskałem wyniki niewiele gorsze od najlepszych zaprezentowanych w Tabeli 5.7 ($F_1 = 0.54$, $acc = 0.58$). Pokazuje to też, jak krótkotrwałe są osiągnięcia typu SOTA, gdzie z miesiąca na miesiąc pojawiają się coraz bardziej kompetentne modele językowe. Osobiście jestem przekonany, że biorąc pod uwagę trudność i nieprecyzyjność zadania adnotacji danych, użycie modeli generatywnych ma większą szansę powodzenia niż próba opracowywania dedykowanych modyfikacji architektury BERT czy dostrajanie modelu pre-trenowanego na danych domenowych.
- 72: nie rozumiem zdziwienia Autorki, że ocena sentymentu wyrażenia zmienia się w wyniku podmienienia instancji encji (np. Trump $\rightarrow$ Biden). Wydaje się oczywiste, że modele językowe niosą ze sobą bardzo silny komponent emocjonalny zakodowany w reprezentacji kontrowersyjnych postaci. Właśnie dlatego tak wartościowe byłoby przeprowadzenie testów behawioralnych na zbiorze danych SEN aby wykazać, jaka część sentymentu może być przypisana do kontekstu, a jaka jest nierozerwalnie związana z postacią, której nagłówek dotyczy.

Jaka jest przydatność rozprawy dla nauk inżynieryjno-technicznych?

Przedstawiona do recenzji rozprawa doktorska adresuje ciekawy, trudny i społecznie bardzo ważny problem. Co prawda trudno jest uznać, że przedstawione badania stanowią definitywne rozwiązanie problemu detekcji stronniczości, cieszy fakt przygotowania dedykowanego zbioru danych dla języka polskiego i podjęcie dużego trudu związanego z adnotacją tego zbioru. Nie sposób określić, na ile użyteczne są wyniki eksperymentalne (trzeba przyznać, że model BERT nie jest dziś modelem SOTA), choć sugestie dotyczące możliwych modyfikacji ogólnej architektury mogą być inspirujące.

Lektura rozprawy przekonuje mnie o kompetencjach naukowych Autorki i opanowaniu przez Nią warsztatu programistycznego. Zgłoszone przeze mnie uwagi mają charakter dość subiektywny (szczególnie odnośnie fundamentalnej niemożności rozwiązania problemu przy wykorzystaniu jedynie formy powierzchniowej tekstu) i w pełni akceptuję, jeśli Autorka jest odmiennego zdania, zapraszając ją do dyskusji na ten temat. Biorąc pod uwagę brak spełnienia ustawowego wymogu zamieszczenia streszczenia w języku polskim, wnoszę o warunkowe dopuszczenie rozprawy doktorskiej mgr inż. Katarzyny Baraniak "*Machine Learning Tools and Techniques Supporting News Media Bias Analysis*" do publicznej obrony po uzupełnieniu streszczenia.

[1] Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64.12 (2021): 86-92.

[2] Stoyanovich, Julia, and Bill Howe. "Nutritional labels for data and models." A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering 42.3 (2019).

[3] Pushkarna, Mahima, Andrew Zaldivar, and Oddur Kjartansson. "Data cards: Purposeful and transparent dataset documentation for responsible ai." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.

[4] Northcutt, Curtis, Lu Jiang, and Isaac Chuang. "Confident learning: Estimating uncertainty in dataset labels." Journal of Artificial Intelligence Research 70 (2021): 1373-1411.

[5] Bender, Emily M., and Alexander Koller. "Climbing towards NLU: On meaning, form, and understanding in the age of data." Proceedings of the 58th annual meeting of the association for computational linguistics. 2020.

[6] Gu, Keren, et al. "An instance-dependent simulation framework for learning with label noise." Machine Learning 112.6 (2023): 1871-1896.