北海道大学
HOKKAIDO UNIVERSITY

Hokkaido University
Research Group of Information Media Science and Technology
Division of Media and Network Technologies
Language Media Laboratory
Rafal Rzepka (Dr inż.)
Kita-ku Kita 14 Nishi 9, 060-0814 Sapporo, Japan
TEL/FAX: +81-11-706-6535
rzepka@ist.hokudai.ac.jp
http://kabura.info

Sapporo, dnia 25.2.2019

**Recenzja rozprawy doktorskiej mgr Bartłomieja Balcerzaka pt. "Analysis and Automatic Recognition of Extremism in Online Texts"**

Mgr **Bartłomiej Balcerzak** przedłożył do oceny rozprawę, której część merytoryczna obejmuje ok. 140 stron standardowego tekstu. Na tekst główny składa się wstęp, przedstawienie danych wykorzystanych do badań, opis podobnych badań, opis cech (atrybutów) lingwistycznych użytych do wykrywania ekstremizmu w tekście, studium przypadku ISIS na Twitterze, opis procesu ulepszenia cech lingwistycznych dla poprawy ogólności zastosowanych metod oraz wnioski z całości badań. Rozprawa dotyczy zagadnienia automatycznego rozpoznawania ekstremizmu w tekście angielskim przy pomocy standardowych technik klasyfikacji. Głównym oryginalnym wkładem badań są: a) dodanie cech narracyjnych, b) zastosowanie ich (wraz ze standardowymi) do wykrywania różnego rodzaju ekstremizmu w tekstach pochodzących z Internetu, c) stworzenie zbiorów danych, które mogą być wykorzystane przez innych badaczy, oraz d) zaproponowanie i przetestowanie metody kategoryzacji (grupowania podobnych słów) przy pomocy reprezentacji wektorowej.

**Szczegółowe omówienie treści rozprawy**

Część zasadnicza rozprawy ma układ przemyślany i konsekwentny w stosunku do obranych celów, przedstawia również odpowiedzi na postawione hipotezy. Każdy z rozdziałów kończy się podsumowaniem i całość układa się w dokładny opis badań, które stanowić mogą praktyczny i fachowy poradnik dla przyszłych projektantów podobnych systemów mających na celu badanie i wykrywanie tekstu o specyficznym charakterze i tematyce. Przystępność i dokładność opisów sprawiają, mimo, iż zakres badań ograniczony jest do dość wąskiej tematyki, że rozprawa może pełnić funkcję edukacyjną nie tylko dla studentów zajmujących się przetwarzaniem języka naturalnego ale i bogatym wstępem dla lingwistów, psychologów, socjologów, a nawet organizacji, których celem jest zidentyfikowanie niebezpieczeństw związanych z grupami ekstremistycznymi, np. namawiających do zamachów terrorystycznych. Z drugiej strony, o czym niestety rozprawa nie wspomina, proponowane techniki mogłyby zostać użyte przez samych ekstremistów do zidentyfikowania swoich wrogów, którzy być może również posługują się specyficznym językiem.

**Rozdział 1. Wstęp**

Wstęp przystępnie przedstawia cele i hipotezy obrane przez Doktoranta, który anonsuje swoje zainteresowania problemem ekstremizmu w Internecie, opisuje (raczej pokrótce) jego wagę i podaje definicje, które wydają się problematyczne i nie do końca przekonywujące, być może niepełne.

Jednakże, ponieważ nie ma zgodności co do tych definicji, jest to jedynie subiektywne odczucie recenzenta. Mgr Balcerzak skrupulatnie podaje również definicje używanych w pracy terminów, parametry użyte w zastosowanych algorytmach, a na koniec opisuje co badania wnoszą do dziedziny.

**Rozdział 2. Dane wykorzystane do badań**

W rozdziale tym Doktorant opisuje dane tekstowe, które zebrał i przystosował do swoich badań - teksty zarówno umieszczane na sieci przez ekstremistów (co nie zawsze było łatwe, jak w przypadku skrajnej lewicy amerykańskiej), jak również te, które miały z nimi kontrastować (neutralne). Nie zawsze można było łatwo wywnioskować jaki zakres lematyzacji i stemmingu został zastosowany, oraz jakie stopwordy zostały wyeliminowane, gdyż zazwyczaj np. nie pozostawia się w tekście zaimków, które jednak w tych badaniach są akurat ważne i zapewne pozostały w pierwotnej postaci (teksty pisane przez ekstremistów z Bliskiego Wschodu mogły stosować wielkie litery w wyrazach pisanych małą w tekstach neutralnych, ale w rozprawie nie doszukałem się informacji o tym jak rozróżnienie liter zostało potraktowane). Tabele z przykładami najczęstszych słów zawierają czasowniki w czasie przeszłym, rzeczowniki w liczbie mnogiej, więc istnieje obawa, że bez dokładniejszego opisu zastosowanych technik przygotowania danych tekstowych trudno byłoby o dokładną replikację badań. Na szczęście doktorant umożliwia dostęp do kodu i części danych, więc wnioskuję, że nie jest to problem krytyczny dla rozprawy. Zabrakło również dokładniejszego opisu klucza, którym kierował się Doktorant przy zbieraniu zestawów danych - np. czy jakieś słowa kluczowe zostały użyte w wyszukiwaniach? Również jeśli chodzi o dane neutralne - czy dobór stron WWW był w zupełności subiektywny, losowy, czy proporcje są adekwatne do ilości typów stron znalezionych przez webcrawlera? Co dokładnie oznacza "upon review" i na czym polegał ten proces?. Jakie hashtagi były użyte dla znalezienia tweetów do zestawu "ISIS-Twitter"? Czy wszystkie hapaks legomena były usunięte, również nazwy własne? Jakie konkretnie zostały zastosowane progi podczas usuwania wyrażeń o niskiej częstotliwości występowania?

**Rozdział 3. Opis podobnych badań**

Bibliografia zdaje się być opracowana skrupulatnie jeśli chodzi o samą kwestię wykrywania ekstremizmu, choć zapewne wiele prac pokrewnych wykrywaniu tzw. "hejtu" jak np. "cyberbullying detection" (Ptaszyński et al.) zostało pominiętych. Interesująca i godna uwagi jest część opisująca ekstremizm w świetle teorii psychologii czy języka (np. modelu Jakobsona), świadczy ona o wnikliwości Doktoranta, który nie ogranicza się do porównania technik NLP, ale wykazuje znajomość tematyki na wielu płaszczyznach.

**Rozdział 4. Opis cech (atrybutów) lingwistycznych użytych do wykrywania ekstremizmu w tekście**

Doktorant poprawnie opisał w tym rozdziale cechy wykorzystane do przeprowadzenia eksperymentów potwierdzających skuteczność zaproponowanych cech narracyjnych. Rozdział układa się w logiczną całość i przy pomocy nie tylko tekstu ale również tabel, wzorów oraz rysunku przejrzyście opisuje cele oraz składniki algorytmu. Na koniec wykazana zostaje skuteczność cech oraz ich kombinacji.

Lektura rozdziału nasuwa jednak wiele pytań. Jakiego narzędzia do anotacji użyto? Czy "annotation" (użyte zresztą na rysunku 4.3) nie byłoby dokładniejszym określeniem oznaczania tekstu przez anotatorów niż "coding"? Jakie dokładnie dostali oni instrukcje (annotation guidelines)? Czy różnice rzetelności między współczynnikami kappa mogły być spowodowane zróżnicowaną interpretacją wytycznych? "Intake of substance" w przykładzie oznaczone zostało jako "emotion", co sugeruje dużą swobodę daną anotatorom i można się zastanawiać czy aby nie zbyt dużą. Jak zróżnicowane były ilościowo oznaczone kategorie?

**Rozdział 5. Studium przypadku (ISIS na Twitterze)**
Jedna z najciekawszych części rozprawy obfitująca nie tylko w techniczne opisy ale i dokładną analizę prawdopodobnych przyczyn zróżnicowania wyników, które nie zawsze były zadawalające. Należy podkreślić, iż Doktorant starannie opisuje pracę klasyfikatora w zadaniu rozpoznawania ekstremizmu, jednak odczuwalny jest brak konkretnych przykładów zdań dla lepszego zilustrowania analizy błędów, co umożliwiłoby czytającym lepszą weryfikację wyciągniętych wniosków. Po części rolę te spełniają tabele korelacji, więc nie jest to mankament krytyczny. Potwierdzone i obalone prawdziwości postawionych hipotez stanowią ważną część rozprawy, a cały rozdział wnosi dużo materiału i obserwacji o skuteczności poszczególnych cech oraz o specyfice poszczególnych danych.

**Rozdział 6. Propozycja ulepszenia atrybutów językowych**
Niespodziewanie niskie wyniki dla wektorowej reprezentacji odkryte podczas studium przypadku skłoniły mgr. Balcerzaka do głębszej analizy i chęci poprawy generalizacji zastosowanych modeli. Obrana metoda (kategoryzacja wektorowa w celu odnalezienie słów o podobnym znaczeniu) okazała się po części słuszna i usprawniła działania klasyfikatorów, dając najczęściej lepsze wyniki niż metoda BoW (choć nie zawsze), szczególnie w odpowiednich kombinacjach cech. Nie do końca można być przekonanym co do głębi zbadania innych prób kategoryzacji w celu porównania z nimi zaproponowanej metody Word2Category, ale jest to temat na tyle szeroki, iż można potraktować pomysł Doktoranta jako nowy i oryginalny przykład zastosowania automatycznego grupowania słów znaczeniowo podobnych.
Na rysunku 6.2 imiona Mary oraz Tony nie zostały sklasyfikowane jako należące do tej samej kategorii, trudno domyślić się czy chodzi o pokazanie niedoskonałości algorytmu czy też autor pomylił się stosując nieprawidłowy kolor. Natomiast na poprzednim rysunku (6.1) kolor fioletowy jest zbyt ciemny, co drastycznie obniża jego czytelność, przynajmniej na papierze.

**Rozdział 7. Główny wniosek**
Ostatnia część rozprawy zawiera podsumowanie pracy. Doktorant raz jeszcze opisuje pokrótce cele, obserwacje, problemy dotyczące danych, definicji i samych wyników. Raczej pobieżnie przedstawione zostały również ograniczenia i możliwe kolejne kroki na drodze udoskonalenia metod wykrywania ekstremizmu w tekście, np. poprzez zastosowanie techniki LSTM dla powtarzających się fraz.

**Silne strony rozprawy**

Praca stanowi przyjemną lekturę, w której jasno postawione zostały cele badań, przedstawiono problemy dotychczasowych metod bogato ilustrowanych bibliografią, wytłumaczono zastosowaną metodykę i zaproponowano udoskonalenia będące bardziej zbliżone do aktualnie popularnych metod (reprezentacje wektorowe). Wnioski zostały elokwentnie opisane i całość (napisana przystępnym językiem) może stanowić zestaw wskazówek do przeprowadzenia podobnych badań również dla naukowców dopiero rozpoczynających stosowanie przetwarzania języka dla wzbogacenia swojego wachlarza narzędzi naukowych. Skrupulatnie wykazana została istotność statystyczna wyników oraz korelacji.
Sama tematyka rozprawy jest ważna i rodzi cenne pytania, na które mogliby odpowiedzieć kolejni badacze jak i sam Doktorant w kolejnych etapach badań.

**Słabe strony rozprawy**

Algorytmy zastosowane w porównaniach klasyfikatorów są wprawdzie standardowymi metodami do podobnych badań, ale nie są to może najlepiej sprawujące się metody na dzień dzisiejszy. Domyślam się, że Doktorant nie dysponował wystarczającą ilością danych do przeprowadzenia dodatkowych eksperymentów, ale ta kwestia nasuwa kolejne pytanie - czy nie można było pokusić się o większe zestawy przykładów, by lepiej odpowiedzieć na pytanie: jak wielkość danych wpływa na precyzję klasyfikacji tekstów wyrażających poglądy ekstremistyczne przy zastosowaniu najnowszych trendów w klasyfikacji tekstu?
W pracy nie zawarto wszystkich szczegółów potrzebnych przy replikacji badań, wielu trzeba się domyślać, jednak nie są to braki krytyczne. Rozprawie w wielu miejscach brak ostatecznej ogłady od strony formalnej, o czym poniżej.

**Formalna strona pracy**

Praca sformatowana jest poprawnie i przejrzyście, jednakże zawiera tak wiele niewielkich niedociągnięć, iż po parunastu stronach czytania ma się wrażenie, iż częściowo pisana była ona w pospiechu i bez ponownego przeczytania. Już w spisie treści, tabel, itp. rzuca się w oczy brak spacji między dłuższymi numerami a tekstem, brak konsekwencji w używaniu spacji przed numerem bibliografii w tekście (na stronie 39 numer bibliografii zastąpiony jest tytułem, a na stronie 29 bibliografia przytoczona jest w przypisie), wyrażenia w tabelach potrafią wychodzić poza zakres (str. 59). W wielu miejscach ma się wrażenie, że autor zapomniał o dokończeniu dodawania bibliografii w tekście lub w tabelach (np. Tabela 3.2) a nawet o rozwinięciu wątku, jak sugeruje np. brak kropki na końcu str. 79. Mimo, iż język angielski, w którym napisano pracę, czyta się świetnie, nie obyło się bez wielu minimalnych błędów, które podobnie do braku konsekwencji w spacjowaniu, niejednolitym użyciu separatorów w dużych liczbach, itd., proszą się o edycje (np. "it's use" zamiast "its use", "vs" zamiast "vs.", "twitter" z małej litery, "Us based" zamiąst "US based", albo raczej "US-based", "word2Category" zamiast "Word2Category", "reprezented" zamiast "represented" (rysunek 6.2), itd.).
Nie zawsze kolejność w tabelach jest jasna - czasem porównywane wyniki uporządkowane są w kolejności od najgorszych do najlepszych, innymi razy - nie są uporządkowane w ogóle, co utrudnia czytającemu szybki wgląd w rezultaty.
Rzecz jasna powyższe detale nie wpływają bezpośrednio na ocenę merytoryczną pracy, ale doktorant powinien zwrócić uwagę na detale gdyż mogą one sugerować niewystarczający stopień wykończenia rozprawy.

**Konkluzja**

Biorąc pod uwagę sumę poczynionych obserwacji stwierdzam, iż przedłożona rozprawa spełnia wymogi stawiane przez Ustawę o Stopniach i Tytułach Naukowych w stopniu umożliwiającym przejście do dalszych etapów przewodu doktorskiego. Wnioskuję zatem o przyjęcie rozprawy oraz o dopuszczenie mgr. Balcerzaka do kolejnych etapów tegoż przewodu.

Sapporo, dnia 25.2.2019

Dr. inż. Rafał Rzepka

〒060-0814 札幌市北区北14条西9丁目
北海道大学情報科学研究科メディアネットワーク専攻
情報メディア学講座メディア知能研究室