

---

## *Recenzja rozprawy doktorskiej*

---

Tytuł:

**Neurokognitywne podejście do przetwarzania tekstów medycznych  
(Neurocognitive inspirations in medical text analysis)**

Autor rozprawy: **Mgr Paweł Matykiewicz**

Promotor: **prof. dr hab. Włodzisława Duch**

Rozprawa doktorska została wykonana na:

*Department of Biomedical Informatics  
Cincinnati Children's Hospital Medical Center  
3333 Burnet Avenue  
Cincinnati, Ohio, USA*

Rozprawa doktorska Pana mgra Pawła Matykiewicza dotyczy problemu wielokrotnego dziedziczenia w sieciach semantycznych. W ostatnich latach, sieci semantyczne stały się niezwykle ciekawą i dynamicznie rozwijającą się tematyką badawczą w dziedzinie neurokognitywistyki, a dotychczasowe próby zastosowania tych sieci w semantycznym przetwarzaniu języka naturalnego pokazują, że wciąż nie można otrzymać dobrego wyniku kategoryzacji tekstów w oparciu o metodach czysto semantycznych. Autor wykrył, że problem polega na zaistnieniu dużej liczby sprzeczności między pojęciami w bazie wiedzy, i jako główny temat rozprawy, postanowił zbadać niesemantyczne metody hamowania rozchodzenia aktywacji w sieciach semantycznych. W mojej ocenie, jest to właściwy wybór tematyki rozprawy.

W przedkładanej rozprawie Autor zaproponował dwa rozwiązania problemu wielokrotnego dziedziczenia w sieciach semantycznych. Oba rozwiązania wykorzystują dane treningowe (etykietowane) do definiowania kryterium przycinania sieci semantycznych. Proponowane metody zostały porównywane z istniejącymi metodami przycinania sieci semantycznych opartych o podejściu czysto semantycznym. Eksperymenty na korpusie dokumentów „Heart Diseases”, zwanym Ohsumed, przygotowanym przez grupę badawczą z School of Medicine at Oregon Health & Science University pokazują, że proponowane metody w połączeniu z klasyfikatorem SVM były w stanie uzyskać bardzo dobre skuteczności klasyfikacji (mierzone miarą  $F_1$ ), porównywalne z najlepszymi, istniejącymi metodami klasyfikacji.

Ogólnie, mimo pewnych niedociągnięć i wątpliwości, jakie nasuwa lektura przedkładanej rozprawy doktorskiej, oceniam ją pozytywnie. W pracy tej można znaleźć kilka ciekawych, autorskich opracowań, a same wyniki eksperymentów wyglądają na dość obiecujące. Dostrzeżone niedociągnięcia i wątpliwości, które mogłyby być tematem dyskusji w czasie obrony, podaję we fragmencie opinii omawiającym szczegółowo pracę.

## **OCENA ZAWARTOŚCI PRACY**

---

Przedstawiona rozprawa składa się z siedmiu rozdziałów, czterech dodatków oraz spisu literatury, w tym spis artykułów. W mojej ocenie, rozprawa została prawidłowo skonstruowana, a autor wykazał dobrą techniką pisanie pracy. Dotyczy między innymi podziału na rozdziały i podrozdziały, not biograficznych i ich cytowania, opisów rysunków oraz tabel. Na ogół, autor poprawnie posługuje się językiem naukowym i dobrze korzysta z materiałów źródłowych.

Poniżej opiszę i skomentuję zawartość każdego rozdziału.

Pierwszy rozdział ma charakter wprowadzenia do tematyki rozprawy. Autor omawia różne podejścia do problemu klasyfikacji tekstów oraz przewodnik po rozprawie.

Rozdział drugi przedstawia przegląd technik klasyfikacji tekstów opartych o bazę wiedzy (wyrażoną w formie ontologii lub taksonomii pojęć). W tym rozdziale, autor przedstawił również neurokognitywne wyjaśnienie mechanizmu dziedziczenia i jego zastosowania, między innymi w rozwoju neuro-kognitywnych systemów klasyfikacji tekstów.

Rozdział trzeci opisuje korpus dokumentów OHSUMED przygotowany przez grupę badawczą z School of Medicine at Oregon Health & Science University oraz zawartość i struktury bazy wiedzy Unified Medical Language System (UMLS).

W rozdziale czwartym, autor omawia problem transformacji między dwiema reprezentacjami dokumentów tekstowych, mianowicie między syntaktyczną reprezentacją, zwaną „workiem słów” (ang. bag-of-words), a semantyczną reprezentacją, zwaną „workiem pojęć” (ang. bag-of-concepts). Autor interpretuje takie transformacje jako zjawisko rozchodzenia aktywacji w sieciach semantycznych i reprezentuje je w postaci macierzy przejścia. Zasadnicza trudność w zastosowaniu reprezentacji worka pojęć w problemie klasyfikacji tekstów polega na



wieloznaczeniowości słów oraz hierarchiczności pojęć. Autor proponował dwie autorskie metody rozchodzenia aktywacji ze przycinaniem i nazwał te metody, jako „Eight edge types pruning” oraz „Twelve edge types pruning”.

Rozdział piąty opisuje znane techniki w dziedzinie klasyfikacji tekstów. Autor koncentrował się na opisie klasyfikatora SVM oraz na miarze  $F_1$  przy ocenie klasyfikatorów. Autor definiował również różne metody wizualizacji dokumentów.

Rozdziały czwarty i piąty to kluczowa część dla całej rozprawy, dlatego Autor starał się szczegółowo opisać swoje metody. W mojej opinii zbyt duża szczegółowość w dużym stopniu zmniejszyła czytelność tego rozdziału. Wszystkie definicje (jest ich 7) w tym rozdziale były oznaczone zmiennymi, których dotyczy dana definicja. Dla przykładu, Definicja 4.3 dotyczy sześciu zmiennych, we wszystkie zawierają literę C, ale różniących się jedynie krojem czcionki:

**Definition 4.3** ( $C, \mathcal{C}, \mathcal{C}^A, \mathcal{C}^B, |\mathcal{C}^A|, |\mathcal{C}^B|$ ).

z tym, że pierwsza litera to wektor etykiet dokumentów, druga litera to stały zbiór  $\{A,B\}$  (czyli nie trzeba było ująć w definicji), kolejne litery to znowu wektory, a ostatnie zmienne to liczby. Ten styl pisania definicji spowoduje ogromną trudność dla czytelnika.

Rozdział szósty przedstawia obszerne wyniki eksperymentów z różnymi metodami rozchodzenia aktywacji. Według mnie to jest główna część rozprawy i bardzo wysoko cenię wyniki przeprowadzonych eksperymentów. Autor wykazał wiele ciekawych interpretacji swoich wyników z punktu widzenia neurokognitywistyki. Autor porównywał jakości klasyfikacji za pomocą klasyfikatora SVM z użyciem trzech metod konstrukcji reprezentacji worka pojęć: metody bez przycinania oraz dwóch proponowanych przez siebie metod przycinania (rozdział 6.5). Okazało się, że metody Autor rozprawy generują znacznie mniejszą sieć pojęć związanych z dokumentem, ale zachowując taką samą lub lepszą skuteczność klasyfikacji. W porównaniu jakości klasyfikacji z innymi metodami klasyfikacji tekstów, proponowane metody przez Pana mgra Pawła Matykiewicza też wypadły bardzo dobrze. Wyniki eksperymentów na wybranym korpusie tekstów pokazują, że proponowane rozwiązania znajdują na drugim i trzecim miejscu tablicy rankingowej. Autor podał również górne oszacowanie złożoności obliczeniowej dla swoich metod. Ze względu na reprezentacji macierzowej, wszystkie operacje są bardzo kosztowne. Niestety autor nie podjął próby szukania bardziej efektywnych struktur danych, lub algorytmów, lub innych, bardziej skalowanych implementacji dla proponowanych metod.

Rozdział siódmy przedstawia wnioski końcowe i podsumuje rozprawę. Autor stwierdził, że przedstawione metody przycinania dziedziczenia spełniają swój oczekiwany efekt, czyli usuwały niepotrzebne fragment bazy wiedzy jednocześnie były w stanie rekonstruować sugerowaną przez Lamb (1999) semantyczną ścieżkę mózgu od obserwacji do podejmowania decyzji. Na zakończeniu, Autor przedstawił również plan badawczy na przyszłość.

Pan mgr Paweł Matykiewicz jest autorem bądź współautorem ponad 25 publikacji, z których 9 artykułów dotyczy tematyki rozprawy.



## UWAGI KRYTYCZNE

---

W mojej ocenie, praca nie została dobrze opracowana pod względem edytorskim. Oprócz wspomnianych wcześniej zarzutów dotyczących stylu pisania definicji, można zauważyć, że wszystkie definicje są bardzo długie, czasem zajmują więcej niż jedna strona formatu A4. To samo dotyczy opisów tabel i figur. Wszystkie te niedociągnięcia spowodują, że praca jest bardzo nieczytelna.

Mam również uwagę dotyczącą wyboru korpusu danych oraz plan eksperymentów. Wszystkie eksperymenty były dokonane na jednym korpusie, więc możliwe jest zjawisko przeuczenia. W rozdziale 6.6, Autor powinien przeprowadzić dodatkowe eksperymenty na innych zbiorach danych tekstowych, aby wyniki były bardziej obiektywne. Poza tym, autor korzystał z swoich metod w procesie wstępnego przetwarzania danych do klasyfikatora SVM. Dlatego powstaje wątpliwość, czy dobre wyniki eksperymentów świadczą o dobrym efekcie działania metod proponowanych przez Autora, czy zawdzięczają wysokiej skuteczności samego klasyfikatora SVM. W mojej opinii, aby uwiarygodnić tezę rozprawy, Autor powinien przetestować kilka innych klasyfikatorów.

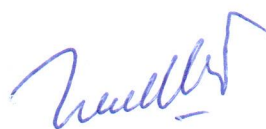
Autor zupełnie pominął dyskusje na temat efektywnej implementacji, a bezpośrednia implementacja proponowanych metod ma dość wysoką złożoność obliczeniową, co obniża wartość praktyczną tych metod. Gdyby autor podjął próbę poprawiania swoich metod pod względem obliczeniowym, to rozprawa byłaby o wiele bardziej wartościowa.

## UWAGI KOŃCOWE

---

Mimo drobnych wątpliwości i niedociągnięć natury edytorskiej, pozytywnie oceniam osiągnięte wyniki badań przedstawionych w rozprawie. Wyniki eksperymentów są dość obiecujące i mogą stanowić bazę dla dalszych badań.

Podsumowując stwierdzam, że recenzowana przeze mnie praca spełnia wymagania stawiane pracom doktorskim i wnoszę o dopuszczenie Pana mgra Pawła Matykiewicza do dalszych etapów przewodu doktorskiego.

  
Hung Sou Nguyen