Crna Gora
UNIVERZITET CRNE GORE
PRIRODNO-MATEMATIČKI FAKULTET
Broj 2024/01-2282
Podgorica, 11.09. 2024 god.

UNIVERZITET CRNE GORE
PRIRODNO-MATEMATIČKI FAKULTET, PODGORICA

Vijeću Prirodno-matematičkog fakulteta

Predmet: Izvještaj komisije o pregledu i ocjeni master rada Marka Đukanovića

Vijeće Prirodno-matematičkog fakulteta na sjednici održanoj 8. 7. 2022. godine, donijelo je Odluku o imenovanju komisije za ocjenu master rada **„Predstavljanje tekstualnih dokumenata strukturom grafa i njihova klasifikacija", kandidata Marka Đukanovića**, u sastavu:

1. Prof. dr Milenko Mosurović, redovni profesor PMF – član;
2. Prof. dr Aleksandar Popović, vanredni porfesor PMF - član;
3. Prof. dr Savo Tomović, redovni profesor PMF – mentor.

Kandidat Marko Đukanović je dana 18. 07. 2024. godine predao tekst master rada na uvid javnosti i ocjenu. Nakon uvida u podneseni materijal, a u vezi sa članom 22 Pravila studiranja na master studijama, podnosimo sljedeći

IZVJEŠTAJ

Master rad kandidata Marka Đukanovića, bečelora računarskih nauka, pod nazivom „Predstavljanje tekstualnih dokumenata strukturom grafa i njihova klasifikacija" ukupno ima 90 strana i ispunjava sve zahtjeve propisane Pravilima studiranja na master studijama.

Rad je iz oblasti računarskih nauka a uža oblast rada je obrada prirodnih jezika (engl. *Natural language processing*) primjenom metoda mašinskog učenja (engl. *machine learning)*. Glavni tekst rada je podijeljen u sedam osnovnih poglavlja: Uvod, Tradicionalni načini za predstavljanje dokumenata, Graf, Tradicionalni algoritmi za klasifikaciju dokumenata, Algoritam za klasifikaciju dokumenata predstavljenih grafom, Eksperimenti i Diskusija. Rad je napisan jasnim i elokventnim stilom, koherentan je i dobro strukturiran, poglavlja su podijeljena u potpoglavlja tako da svako od njih čini jednu logičku cjelinu.

Cilj ovog istraživanja je bio da ispita mogućnosti primjene strukture grafa u cilju efikasnijeg rješavanja problema klasifikacije dokumenata. Fokus istraživanja je razvoj novog algoritma za reprezentaciju tekstualnih dokumenata strukturom grafa, i detaljna analiza da li graf reprezentacija dokumenata obezbjeđuje bolje performanse u odnosu na tradicionalne i u literaturi opisane algoritme za klasifikaciju. Prethodno rečeno, posebno se odnosi na obradu i klasifikaciju dokumenata na crnogorskom jeziku, sa ciljem unapređenja trenutnog stanja i

doprinosa razvoju obrade tekstualnih podataka sa crnogorskog govornog područja. Dodatno, istraživanje doprinosi pripremom novog skupa podataka sa portala Vijesti, dostupan na https://drive.google.com/drive/folders/1RGVSohSpEZ8hrgKt8NgGe1_6UCtqvEEA?usp=sharing, koji se može koristiti za dalja istraživanja problema iz ove oblasti na crnogorskom i srodnim jezicima.

Za rješavanje problema klasifikacije tekstualnih dokumenata predlaže se pristup koji se zasniva na upotrebi *TextRank*-a, verzije *Google*-ovog *PageRank* algoritma. Prvi korak predloženog novog algoritma je transformacija tekstualnih podataka u strukturu grafa, koja opisuje složene odnose između riječi u tekstu. U ovom radu, formirani su grafovi su-pojava (engl. *co-occurrence)* na osnovu pojavljivanja parova riječi unutar određenog prozora. U ovom tipu grafa, čvorovi predstavljaju individualne riječi, dok ivice predstavljaju su-pojavu tih riječi. Nakon kreiranja grafa, upotrebom *TextRank*-a identifikuju se najuticajniji čvorovi, odnosno ključne riječi u tekstu. Ovaj postupak predstavlja fazu pre-procesiranja, nakon koje se ključne riječi kvantifikuju upotrebom *TF-IDF* vektorizacije, što ih priprema za dalju obradu od strane odabranih modela mašinskog učenja. *TF-IDF* vektori dobijeni od najuticajnijih riječi koriste se kao ulaz u različite modele mašinskog učenja. Odabrani modeli u ovom radu su *Naive Bayes*, *Logistic Regression*, *Support Vector Machine* i *Random Forest.* Ovi modeli se obučavaju na skupu podataka sa labelama koje označavaju kategorije i nakon faze obuke stiču sposobnost predviđanja kategorija za nove tekstove. Centralna ideja i naučni doprinos kandidata je razvoj novih tehnika pripreme podataka pomoću strukture grafa za modele mašinskog učenja. Radi objektivne procjene performansi i uporedne analize, istraživanje obuhvata analizu predloženog modela za klasifikaciju tekstova upotrebom strukture grafa u odnosu na tradicionalne modele, sprovedenu nad dva različita skupa podataka. Na ovaj način testirana je zavisnost performansi modela od različitih karakteristika skupa podataka. Jedan od skupova je na engleskom jeziku, *BBC news dataset*, koji je javno dostupan i u literaturi opisan, dok je drugi skup podataka sa portala „Vijesti" koji se sastoji od članaka na crnogorskom jeziku, posebno pripremljen za ovo istraživanje.

U prvom poglavlju rada detaljno je definisan problem koji će biti tretiran u radu, kao i motivacija za njegovo rješavanje. Takođe, prvo poglavlje daje pregled dosadašnjih istraživanja na ovu temu.

U drugom, trećem i četvrtom poglavlju detaljano su predstavljeni teorija i koncepati neophodni za razumijevanje rada. Drugo poglavlje opisuje tradicionalne tehnike reprezentacije dokumenata, treće poglavlje osnove teorije grafa i uvodi koncept centralnosti čvorova u grafu. Četvrto poglavlja upoznaje čitaoca sa osnovama teorije odabranih mašinskih modela.

Peto poglavlje pruža uvid u inovativnu ideju rada koja se sastoji u primjeni strukture grafa i *TextRank* algoritma za rješavanje problema klasifikacije tekstualnih dokumenata. U ovom poglavlju je detaljno opisan novi predloženi algoritam, uključujući *TextRank* algoritam i algoritam za kreiranje strukture grafa.

U šestom poglavlju daje se pregled i analiza skupova podataka korištenih u eksperimentima, kao i eksperimentalnog protokola, metrika i rezultata. Eksperimenti su sveobuhvatni. Urađeni su na skupovima različitih karakteristika i na različitim jezicima. Rezultati su predstavljeni i analizirani po različitim skupovima podataka i uz primjenu poznatih metrika za problem klasifikacije. Dobijeni rezultati pokazuju da je moguće razviti manje i efikasnije modele, čiji su rezultati isti ili približni tradicionalnim metodama za klasifikaciju dokumenata. Modeli koji

koriste pet ključnih riječi imaju veličinu koja iznosi samo 10% veličine tradicionalnih modela. Raznovrsnost testnih skupova podataka, kao i njihova veličina, potvrđuju da je algoritam generalan i da se može primjenjivati na tekstove na različitim jezicima.

Sedmo poglavlje je fokusirano na detaljnu analizu rezultata predstavljenih u prethodnom poglavlju, sa ciljem testiranja postavljenih hipoteza istraživanja. Analiza rezultata potvrđuje mogućnost konstrukcije efikasnog algoritma za reprezentaciju tekstualnih dokumenata strukturom grafa za probleme klasifikacije tekstova. Predloženi algoritam uspješno čuva ključne semantičke i sintaksne elemente teksta, omogućavajući efikasnu klasifikaciju. Dodatno, upotreba strukture grafa i *TextRank* algoritma omogućava bolju i intuitivniju vizuelizaciju odnosa unutar dokumenta, što može pomoći boljem razumijevanju strukturalnih odnosa unutar dokumenata. Detaljnom analizom rezultata u ovom poglavlju je utvrđeno da predloženi algoritam pokazuje konkurentne performanse u odnosu na tradicionalne metode klasifikacije. U zavisnosti od modela i skupa podataka rezultati ovog algoritma su slični ili bolji. Ovakvi rezultati sugerišu da predloženi algoritam može biti validna alternativa tradicionalnim metodama.

Zaključak i predlog

Na osnovu prethodno napisanog, Komisija smatra da je master rad kandidata Marka Đukanovića napisan jasno i u skladu je sa pravilima izrade naučnog rada i kriterijumima propisanim Pravilima studiranja na master studijama. Kandidat je kroz ovaj rad realizovao sve postavljene ciljeve master teze.

Kandidat je pokazao da odlično poznaje naučnu problematiku, kao i da posjeduje značajan nivo istraživačkih sposobnosti. Stoga komisija pozitivno ocjenjuje master rad **„Predstavljanje tekstualnih dokumenata strukturom grafa i njihova klasifikacija", kandidata Marka Đukanovića.**

Komisija predlaže Vijeću Prirodno-matematičkog fakulteta da rad pod naslovom **„Predstavljanje tekstualnih dokumenata strukturom grafa i njihova klasifikacija" kandidata Marka Đukanovića** prihvati kao master rad i odobri njegovu javnu usmenu odbranu.

U Podgorici, 9. 9. 2024. .godine

KOMISIJA

Prof. dr Milenko Mosurović, redovni profesor PMF – član;

Mosurović

Prof. dr Aleksandar Popović, vanredni profesor PMF– član;

Prof. dr Savo Tomović, redovni profesor PMF – mentor