# UNIVERZITET CRNE GORE
# PRIRODNO-MATEMATIČKI FAKULTET
## Vijeću Prirodno-matematičkog fakulteta

PREDMET: Izvještaj komisije o magistarskom radu pod nazivom „Novelty detekcija primjenom hibridnih algoritama nad tekstualnim podacima“, kandidata Aleksandra Plamenca, specijaliste Računarskih nauka.

Na LXIX sjednici Vijeća Prirodno-matematičkog fakulteta Univerziteta Crne Gore, imenovani smo u Komisiju za pregled i ocjenu magistarskog rada pod nazivom „**Novelty detekcija primjenom hibridnih algoritama nad tekstualnim podacima“ kandidata Aleksandra Plamenca**. Na osnovu pregledanog rada i prema uslovima utvrđenim Zakonom o visokom školstvu i Statutom Univerziteta Crne Gore, podnosimo sljedeći

## IZVJEŠTAJ

**Analiza magistarskog rada:**

Rad je napisan na 95 strana, i sadrži sljedeća poglavlja: Predgovor, Izvod rada, Abstract, Sadržaj, Uvod, Tradicionalni algoritmi, Hibridni algoritmi, Eksperimentalni protokol, Diskusija, Zaključak i Literatura. U radu se nalazi 31 slika, 18 tabela kao i 24 skripte napisane u programskom jeziku Python. Poglavlje Literatura sadrži 70 referenci.

**Postavljeni ciljevi rada:**

Problem novelty detekcije predstavlja aktuelan problem čija je primjena široka. Data tehnika se sa dobrim rezultatima može koristiti u medicinske svrhe, detekcije prevare iili virusa. Tehnika novelty detekcije se može koristiti nezavisno od formata podataka, sa jednakim značajem i uspjehom. Veliku ekspanziju doživljavaju sistemi obrade prirodnog jezika u realnom vremenu. Kod ove vrste aplikacija je ponovno treniranje važan zadatak. Ciljevi ovog rada su bili definicija hibridnog algoritma i specifične procedure za preprocesiranje teksta sa slovenskog govornog područja. Urađena je i studija slučaja i analiza govora mržnje sa lokalnog portala. Na osnovu analize je napravljen model korišćen za detekciju govora mržnje. Dodatni

cilj rada je definicija hibridnog algoritma kojim bi se koristeći datu tehniku preprocesiranja ubrzao proces treniranja, sa istom ili približnom preciznošću algoritma.

**Primijenjene metode:**

U cilju ostvarivanja prethodno navedenih zadataka, prikupljen je skup podataka od strane lokalnog portala. Skup je predstavljao kolekciju komentara korisnika sa portala. Komentari su podijeljeni u dvije grupe: dozvoljeni (*allowed*), i zabranjeni (*forbidden*), bazirano na tome da li su sadržali nedolični govor ili govor mržnje. U radu se kao hipoteza navodi da hibridni algoritmi sa sličnom preciznošću, ali bržim treningom mogu koristiti za problem novelty detekcije nad tekstualnim podacima. Konkretno, kao algoritam detekcije govora mržnje. Kako bi se utvrdila tačnost hipoteze, u radu je definisan eksperimentalni protokol po kome se podrazumijevalo da se kreira model sa podacima koji nisu preprocesirani, koristeći jednu od statističkih metoda, kao što je Naive Bayes. Radi adekvatne komparacije algoritama, isti podaci su preprocesirani kroz pipeline koji je uključivao nekoliko koraka:

- Prevod komentara na engleski jezik,
- Uklanjanje specijalnih karaktera iz riječi ovog govornog područja,
- Uklanjanje HTML entiteta iz komentara,
- Uklanjanje ćiriličnih komentara iz skupa podataka,
- Uklanjanje riječi koje imaju manje od 3 slova,
- Uklanjanje riječi baziranih na nekom regularnom izrazu,
- Uklanjanje interpunkcije iz komentara,
- Pretvaranje svih slova u mala,
- Transformacija riječi u njehov korijen

Formatirani skup podataka se koristio za trening modela zasnovanim na tehnikama kao što je mašina nosećih vektora (*support vector machine*). Kako bi kompletan hibridni algoritam bio korektno definisan, u radu se obraća pažnja na sljedeće principe koje algoritam treba da posjeduje:

- Princip robustnosti i razmjene - algoritam treba da maksimizuje isključenje novih instanci (*novelty*), a da minimizuje isključenje poznatih instanci,
- Princip uniformnog skaliranja podataka - Potrebno je da svi test i trening podaci budu uniformno prikazani, odnosno da "leže" u istom opsegu,

- Princip minimalizacije parametara - Broj parametara koje unosi korisnik kako bi trenirao algoritam treba biti što manji,
- Princip generalizacije - Sistem bi trebao biti sposoban za generalizaciju,
- Princip nezavisnosti - Sistem ne bi trebao da zavisi od broja karakteristika koje se posmatraju, kao i od broja i prirode klasa koje postoje u podacima,
- Princip prilagodljivosti - Sistem koji prepoznaje nove instance tokom testa bi trebao iskoristiti ove testove i za ponovno treniranje.
- Princip složenosti - S obzirom da su mnogi sistemi za *novelty* detekciju *online* i koriste se u realnom vremenu, jako je bitan segment vremenska složenost samog algoritma koji se koristi.

Dobijeni rezultati testiranja modela su posmatrani kroz parametre *precision* i *recall,* kao i vizualno prikazani kroz matrice konfuzije.

**Dobijeni rezultati:**

U poglavlju Diskusija, prethodno definisani protokol je sproveden za 6 karakterističnih eksperimentanih konfiguracija, različite balansiranosti skupa podataka. Nakon toga, posmatrani su i diskutovani rezultati kroz matricu konfuzije, *precision,* i *recall* parametre, kao i vrijeme potrebno za trening modela. Eksperimentalne konfiguracije su se razlikovale u broju procesiranih komentara, a jedna od eksperimentalnih konfiguracija je imala i nebalansirani skup podataka. Rezultati treniranja algoritma ukazuju na to da se preciznost algoritama ne mijenja značajno, i da se kreće u rasponu od 60%-70%. Takođe, i modeli koji su trenirani na preprocesiranim podacima daju istu preciznost, ili neznatno manju (2-3%). Kada je u pitanju vrijeme treniranja modela, u svakoj od eksperimentalnih konfiguracija se zaključuje poboljšanje i skraćenje vremena za 5-10%. Potvrdu ove činjenice daje i posljednja konfiguracija. U posljednjoj konfiguaciji je rađeno sa skupom podataka koji nije balansiran. Ovakav skup podataka je realniji u odnosu na bilo koji balansirani skup, s obzirom da je mnogo više dozvoljenih komentara na portalu, u odnosu na one koji nisu dozvoljeni. Iz prikazanog eksperimenta se zaključuje da je vrijeme procesiranja kraće, a da je preciznost slična, ili neznatno manja. Istraživanje koje je sprovedeno u radu upućuje u razvoj hibridnih algoritama za detekciju novelty instanci u tekstu.

**Zaključak i predlog komisije:**

Nakon pregledanog magistarskog rada, analize rezultata i značaja ostvarenih istraživanja, Komisija konstatuje da rad zadovoljava sve uslove naučno-istraživačkog rada. Zadata tema ovog rada je naučno aktuelna, s obzirom na veliku popularnost i primjenu jezičkih modela u današnjici. Tema je na adekvatan način obrazložena, a istraživanje je dalo rezultate koji sveobuhvatno prikazuju zadate ciljeve. Prikazano istraživanje je dalo značajne rezultate u domenu treniranja jezičkih modela i ubrzalo dati proces, što je za aplikacije iz ovog domena od izuzetnog značaja.

Na osnovu izloženog, Komisija predlaže Vijeću Prirodno-matematičkog fakulteta u Podgorici da rad kandidata Aleksandra Plamenca, pod naslovom: „**Novelty detekcija primjenom hibridnih algoritama nad tekstualnim podacima**“ prihvati kao magistarski rad i odobri javnu usmenu odbranu.

Podgorica, 21.09.2023. godine

**Komisija:**

1. **Dr Milenko Mosurović, redovni profesor, PMF, član**

   Mosurović

2. **Dr Aleksandar Popović, vanredni profesor, PMF, član**

3. **Dr Savo Tomović, redovni profesor, PMF, mentor**