

МОЛБА ЗА ОДОБРАВАЊЕ ТЕМЕ МАСТЕР РАДА

Молим да се одобри израда мастер рада под насловом:

**Екстракција информација из докумената на српском језику применом БЕРТ и великих језичких модела у условима оскудних ресурса: поређење стратегија аугментације и ано-
тације података**

Значај теме и области:

Посебно важан проблем у области обраде природних језика јесте екстракција информација из неструктурисаних текстуалних докумената, која омогућава аутоматско издвајање релевантних чињеница и њихово претварање у структурисане податке погодне за даљу анализу и доношење одлука. Савремени приступи засновани на БЕРТ архитектури и великим језичким моделима донели су значајан напредак у задацима као што су препознавање именованих ентитета, екстракција информација и разумевање текста. Међутим, примена ових метода на српском језику и другим језицима са ограниченим ресурсима и даље је отежана недостатком квалитетно анотираних скупова података. Због тога је од посебног научног и практичног значаја истраживање метода које омогућавају ефикасно обучавање модела у оваквим условима. Тема овог рада припада области обраде природних језика и управо има за циљ да допринесе развоју поузданијих система за екстракцију информација на српском језику у различитим доменама (јавна управа, правни системи, пословна аналитика, осигурање, итд.) кроз анализу различитих стратегија припреме података.

Специфични циљ рада:

Циљ рада је развој и евалуација система за екстракцију информација из докумената на српском језику применом БЕРТ архитектуре и великих језичких модела у условима ограничене количине анотираних података.

Посебан фокус рада биће на анализи и поређењу различитих стратегија аугментације и аотације података са циљем превазилажења проблема недостатка означених података за српски језик. У оквиру рада биће испитане и експериментално евалуиране одабране технике припреме података, укључујући и приступе потпомогнуте великим језичким моделима, као и њихов утицај на перформансе система за екстракцију информација.

Користио би се скуп формиран од јавно доступних докумената на српском језику. Као примарни извор података планирана су саопштења Министарства унутрашњих послова Републике Србије, која садрже релевантне информације о догађајима, лицима, институцијама и локацијама и погодна су за задатке екстракције информација. По потреби, корпус би могао бити допуњен документима из јавно доступних база судске праксе.

Очекује се да резултати рада допринесу бољем разумевању могућности примене великих језичких модела у условима оскудних ресурса и понуде смернице за развој система за обраду природних језика на српском језику.

Остале битне информације:

Поред академског доприноса, истраживање има и потенцијалну практичну примену у аутоматизованој обради докумената, на пример у домену осигурања, што би могло бити предмет даљег рада након развоја и евалуације методологије.

(својеручни потпис студента)

<датум>
(датум подношења молбе)

(својеручни потпис ментора)

Чланови комисије

1. Александар Картељ
2. Младен Николић

Катедра за рачунарство и информатику је сагласна са предложеном темом.

(шеф катедре)

(датум одобравања молбе)