УНИВЕРЗИТЕТ У НОВОМ САДУ
ПРИРОДНО-МАТЕМАТИЧКИ ФАКУЛТЕТ

ИЗВЕШТАЈ О ОЦЕНИ МАСТЕР РАДА

I ПОДАЦИ О КОМИСИЈИ

1. **Датум и орган који је именовао Комисију**
   10.01.2023. Веће Департмана за математику и информатику Природно-математичког факултета Универзитета у Новом Саду
2. **Састав Комисије са назнаком имена и презимена сваког члана, звања, назива уже научне области за коју је изабран у звање, датума избора у звање и назив факултета, установе у којој је члан комисије запослен:**
   1. др Душан Јаковетић, ванредни професор ПМФ-а, председник, ужа научна област: математичко моделирање, датум избора у звање: 15.11.2020.
   2. др Милош Савић, ванредни професор ПМФ-а, ментор, ужа научна област: рачунарске науке, датум избора у звање: 17.11.2020.
   3. др Владимир Курбалија, редовни професор ПМФ-а, члан, ужа научна област: рачунарске науке, датум избора у звање: 1.7.2021.

II ПОДАЦИ О КАНДИДАТУ

1. **Име, име једног родитеља, презиме:**
   Стефан, Срђан, Димитријевић
2. **Датум рођења, општина, република:**
   19.11.1994., Лесковац, Република Србија
3. **Година уписа на дипломске академске студије, смер/усмерење:**
   2017., Примењена математика – наука о подацима, модул: рачунарство високих перформанси

III НАСЛОВ МАСТЕР РАДА

Анализа алгоритама машинског учења и дубоког учења за детекцију сентимента текста

ПРЕГЛЕД МАСТЕР РАДА

Мастер рад „Анализа алгоритама машинског учења и дубоког учења за детекцију сентимента текста“ је написан на 99 страна у 10 поглавља и садржи 44 референце, 5 табела и 45 фигура. Рад почиње предговором, садржајем и списком табела, фигура и скраћеница.

Садржај рада чине следећа поглавља: 1. Увод, 2. Релевантна истраживања, 3. Скупови података, 4. Атрибути (енг. *Features*), 5. Концепти машинског учења, 6. Модели машинског учења, 7. Модели дубоког учења, 8. Оптимизација, 9. Експерименти и резултати и 10. Закључак.

На крају рада дат је списак референцираних радова.

У уводу рада дефинисан је проблем детекције сентимента у тексту, истакнут је његов значај и области могуће примене. Детекција сентимента је проблем класификације у коме се за неки текст изводи класа сентимента (нпр. позитивна или негативна рецензија неког производа). Стога је овом проблему могуће приступити техникама машинског и дубоког учења тренирајући одговарајуће класификационе моделе.

У другом поглављу рада је дат приказ претходних истраживања који су се бавили експерименталним анализама техника машинског и дубоког учења у домену детекције

сентимента текста.

У трећем поглављу рада се описују експериментални скупови података који су коришћени у овом раду.

Четврто поглавље рада описује различите начине векторизације текста које су разматране у раду.

Пето поглавље даје преглед основних концепата машинског учења који су релевантни за спроведено истраживање.

Поглавља 6 и 7 описују алгоритме машинског и дубоког учења који су експериментално испитивани у овом раду. Поглавље 8 се надовезује на претходно поглавље дајући преглед техника оптимизације који се користе приликом тренирања модела дубоког учења базираних на неуронским мрежама.

У поглављу 9 даје се опис спроведених експеримената и приказ добијених резултата са пратећом дискусијом.

Последње поглавље сумира добијене резултате и напрам њих даје смернице за даљи рад.

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА МАСТЕР РАДА

Циљ рада је био да се спроведе опсежна анализа постојећих алгоритама машинског и дубоког учења за проблем детекције сентимента текста. Да би се могли обучити класификатори који детектују сентимент текста неопходно је прво спровести векторизацију текстуалних докумената из тренинг скупа података (конвертовање тектуалног корпуса у табелу која садржи нумеричке атрибуте где је сваки текстуални документ представљен нумеричким вектором односно ембедингом). У раду се на 3 реална скупа података (*IMDB Reviews*, *YELP reviews* и *Corona tweets*) испитује ефективност 5 различитих алгоритма машинског и дубоког учења у спрези са 5 различитих техника векторизације текста.

У другом поглављу рада дат је преглед научно-истраживачких радова публикованих у међународним часописима који се баве проблематиком детекције сентимента текста користећи технике машинског и дубоког учења. На основу датог прегледа може се закључити да су претходне компаративне анализе фокусиране или само на класично машинско учење или само на дубоко учење, те да недостаје компаративна анализа алгоритама из обе групе.

У трећем поглављу рада даје се преглед експерименталних скупова података који су коришћени приликом тренирања и евалуације модела машинског и дубоког учења (број инстанци, број различитих класа који одређује сентимент текста и расподела инстанци по класама). Поред тога објашњене су технике препроцесирања текста пре његове векторизације (нормализација и токенизација текста, уклањање хеш тегова у твитовима).

У четрвртом поглављу се детаљно објашњавају анализиране технике за векторизацију текста: TF-IDF, Word2Vec, Glove, FastText и BERT. Подвучене су концептуалне разлике између ових приступа: TF-IDF је *bag-of-words* векторизација код које се игнорише редослед и контекст речи (те је сходно томе брза и једноставна за имплементацију), Word2Vec, Glove и FastText су претренирани текст ембединзи, док је BERT контенстуални текст ембединг базиран на трансформерима.

Пето поглавље даје преглед основних метрика и методологија за евалуацију модела машинског и дубоког учења (дефиниције метрика перформанси класификационих модела и методологија за њихову оцену). Модели анализирани у овом раду су упоређивани на

бази добијених F_1 скорова који су оцењени крос валидацијом у 10 фолдова.

У шестом, седмом и осмом поглављу се детаљно описују алгоритми машинског и дубинског учења који су коришћени за тренирање модела који детектују сентимент текста: *naive Bayes*, логистичка регресија, стабла одлучивања, *random forest*, *support vector machine*, LSTM неуронске мреже и BERT трансформери.

Девето поглавље, поред приказа и дискусије добијених резултата, даје опис хардвера који је коришћен приликом експерименталних анализа, списак библиотека програмског кода за реализацију експеримената и спецификације хипер-параметара за моделе базиране на неуронским мрежама.

VI ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Добијени експериментални резултати показују да модели тренирани *support vector machine* алгоритмом на TF-IDF векторизацијама постижу највећу тачност класификације на сва три експериментална скупа података. F_1 скорови добијени за овај приступ износе 0.89 (*IMDB Reviews*), 0.55 (*YELP reviews*) и 0.64 (*Corona tweets*) што указује да модели поседују средње (*YELP*, *Corona*) до високо (*IMDB*) задовољавајући степен тачности. Разлике у F_1 скоровима између најбољег и најлошијег предиктивног модела варирају између 12% и 14%. Код модела базираних на техникама дубинског учења је примењиван тјунинг хипер-параметара грид методом, али на малом опсегу могућих вредности због изразито великог времена које оваква операције захтева на расположивом хардверу. Стога је закључак рада да модели базирани на дубоком учењу захтевају опсежнији, финији тјунинг хипер-параметара како би се њихове перформансе поспешиле.

VII КОНАЧНА ОЦЕНА МАСТЕР РАДА

Мастер рад је у потпуности урађен у складу са одобреном темом. Садржај рада показује да кандидат поседује разумевање широког спектра алгоритама машинског и дубоког учења и оперативно знање да их имплементира у области обраде природног језика. Рад је прегледан и добро написан, а добијени резултати су коректно анализирани и приказани.

VIII ПРЕДЛОГ

На основу коначне оцене, Комисија предлаже да се мастер рад „Анализа алгоритама машинског учења и дубоког учења за детекцију сентимента текста“ прихвати, а кандидату Стефану Димитријевићу одобри одбрана.

Нови Сад, 13.1.2023

ПОТПИСИ ЧЛАНОВА КОМИСИЈЕ

др Душан Јаковетић, председник

_____

др Милош Савић, ментор

_____

др Владимир Курбалија, члан

_____