Катедри за Рачунарство и информатику
Математичког факултета
Универзитета у Београду

ИЗВЕШТАЈ

о прегледу мастер рада
„Имплементација и евалуација техника класификације текста заснованих на Бајесовој теореми“
кандидата Андрије Ђуришића

Одлуком Наставно-научног већа Математичког факултета донетом на 315. редовној седници одржаној 20.06.2014. год. именовани смо за чланове комисије за преглед и одбрану мастер рада под насловом „Имплементација и евалуација техника класификације текста заснованих на Бајесовој теореми“ кандидата Андрије Ђуришића, студента мастер студија на студијском програму Информатика на Математичком факултету.

I Област рукописа
Рукопис „Имплементација и евалуација техника класификације текста заснованих на Бајесовој теореми“ кандидата Андрије Ђуришића, бави се применом техника машинског учења и пробабилистичког резоновања заснованог на Бајесовој теореми о условним вероватноћама на проблем класификације текста тј. проблем категорисања новинских чланака у фиксиран скуп класа на основу области коју чланак покрива. У раду се користи знање машинског учења, вероватноће, аутоматске анализе природног језика и програмирања.

II Структура рукописа и кратак приказ

Рукопис се састоји од 44 стране, организоване у 7 поглавља и додатак.

У уводном поглављу кандидат даје мотивацију за свој рад, уводи проблем аутоматске класификације тј. категоризације текста и даје кратак историјски осврт на покушаје решавања тог проблема.

У поглављу "Класификација текста" кандидат даје математичку дефиницију тог проблема и описује најзначајније мере квалитета и технике евалуације класификације.

У поглављу "Најзначајнији алгоритми надгледане класификације" кандидат даје преглед техника машинског учења које се користе у решавању проблема класификације текста. Обрађују се стабла одлучивања, вештачке неуронске мреже, као и наивни Бајесов метод који представља основу имплементације и експерименталног дела овог рада. За сваку од ових метода укратко је приказана математичка формулација и приказана је примена методе на малом, вештачком примеру, класификације текста.

У поглављу "Имплементација" кандидат описује имплементацију система Newsy који врши аутоматску класификацију текстова на српском језику и представља централни допринос овог рада. Кандидат описује арихтектуру тог система, као и софтверске алатке који је користио током израде система.

У поглављу "Експерименти" кандидат приказује резултате експерименталне евалуације система Newsy на два новинска корпуса - корпусу Ebart и корпусу B92.

У поглављу "Корекција грешака и побољшања" кандидат врши анализу експерименталних резултата, идентификује проблеме који смањују одзив и прецизност класификације и предлаже и евалуира побољшања заснована на елиминацији тзв. стоп речи и на примени TF-IDF методе.

У поглављу "Закључак" кандидат врши завршно разматрање и износи закључке.

Додатак "Литература" садржи списак од 16 библиографских јединица које је кандидат користио приликом писања рада.

III Анализа рукописа

У свом првом делу рукопис приказује теоријске основе проблема класификације текста, укључујући и детаљне описе алгоритама који се обично користе за решавање овог проблема (стабла одлучивања и ID3 алгоритам, неуронске мреже и перцептрон као и наивни Бајесов класификатор). Централни допринос аутора чини други део рада у коме је извршена имплементација једног класификатора (програм Newsy) заснованог на Бајесовој теореми и његова детаљна експериментална евалуација. Експериментални резултати указују на задовољавајућ квалитет класификације (прецизност класификације за поједине категорије се креће од око 67% у најлошијем до око 99% у најбољем случају, а одзив од око 50% до око 98%). Додатно, показано је како се квалитет може донекле побољшати једноставним оптимизацијама попут елиминације стоп речи и TF-IDF методе. Као чланови Комисије пратили смо писање овог рукописа и дали аутору низ примедби, захтева и сугестија, које је он усвојио и обрадио у финалној верзији текста на задовољавајући начин.

IV Закључак и предлог

Приказом теоријских основа примењених техника, реализацијом експерименталног истраживања и обрадом, анализом и приказом добијених резултата кандидат је показао да је у стању да самостално усвоји, имплементира и тестира технике машинског учења у домену обраде природног језика, чиме је приказао потребан степен научно-стручног знања. На основу свега наведеног Комисија предлаже да се рукопис под насловом:

„Имплементација и евалуација техника класификације текста заснованих на Бајесовој теореми“

кандидата Андрије Ђуришића прихвати као мастер рад и да се закаже његова јавна усмена одбрана.

Комисија:
др Филип Марић, ментор

проф. др Гордана Павловић-Лажетић

др Јелена Граовац

Београд, 17.11.2014.