Универзитет „Кирил и Методиј“ – Скопје

Факултет за информатички науки и компјутерско инженерство

Анализа на големи податоци во банкарскиот сектор

Докторска дисертација

Кандидат:

м-р Фисник Доко

Ментор:

проф. Д-р Игор Мишковски

Скопје, 2021

Комисија

Проф. д-р Димитар Трајанов, претседател
Факултет за информатички науки и компјутерско инженерство
Универзитет Св. “Кирил и Методиј”, Скопје, Северна Македонија

Проф. д-р Игор Мишковски, ментор
Факултет за информатички науки и компјутерско инженерство
Универзитет Св. “Кирил и Методиј”, Скопје, Северна Македонија

Проф. д-р Иван Чорбев, член
Факултет за информатички науки и компјутерско инженерство
Универзитет Св. “Кирил и Методиј”, Скопје, Северна Македонија

Проф. д-р Слободан Калајџиски, член
Факултет за информатички науки и компјутерско инженерство
Универзитет Св. “Кирил и Методиј”, Скопје, Северна Македонија

Проф. д-р Беким Фетаји, надворешен член
Факултет за Информатика
Универзитет “Мајка Тереза”, Скопје, Северна Македонија

Анализа на големи податоци во банкарскиот сектор

Фисник доко

Апстракт

Живееме во свет каде експоненцијалниот раст на податоците го трансформираат начинот на работење обезбедувајќи нови знаења и шаблони кои се клучни за носење на информирани одлуки. Големите податоци претставуваат потенцијално богатство за компаниите кои успешно ќе се справат и ќе ги искористат истите. Финансискиот сектор се води од податоците за да се справи со различните ризици при секојдневното работење. Примената на големите податоци и науката за податоците покрај многуте придобивки на банките им овозможува да се справат подобро со различните ризици и да се стекнат со конкурентска предност. Како најчест предизвик во финансиските институции е одредувањето на кредитниот ризик, чие точно предвидување е од клучно значење за нив. Постојат многу истражувања за кредитниот ризик меѓутоа сите предвидувања се основаат на користење на податоците од комерцијалните банки, и според мое знаење не постои истражување кое користи податоци од база Кредитен регистар која ја има само во централните банки. Финансиските вести се исто така важен инструмент кои влијаат и предвидуваат различни финансиски инструменти. Истите може да се искористат и за анализа на компании, личности од финансискиот свет, финансиски инструменти, продукти и други финансиски ентитети. Докторската дисертација предлага платформа за предвидување на кредитниот ризик која ќе им помогне на финансиските институции со помош на новите знаења кои се споени во платформата. Во платформата е интегрирано предвидување на кредитниот ризик за клиент или компанија со користење на податоците на Кредитниот регистар. За финансиските вести е применето препознавање на ентитети и анализа на сентиментот. Овие две посебни модели се направени со користење финансиски вести на македонски и албански јазик. Уникатноста на дисертацијата е дека досега нема некое истражување кое го предвидува кредитниот ризик со помош на базата Кредитен регистар и со извлекувањето на знаење од финансиските вести на македонски и албански јазик.

Платформата овозможува подобро справување со кредитниот ризик со помош на дополнително мислење од централната банка. Анализата за кредитниот ризик се помага со финансиските вести преку кои се анализира компанијата во која е вработен клиентот или компанијата која е барател за кредит, соодветно. Платформата е наменета за економисти и брокери, овозможувајќи им на лесен начин уникатни придобивки за справување со кредитниот ризик.

Клучни зборови: кредитен ризик, кредитен регистар, процесирање на природни јазици, препознавање ентитети, анализа на сентимент

Analysis of Big Data in banking sector

Fisnik Doko

Abstract

We live in a world where exponential growth of data is transforming the way that companies work by providing new knowledge and insights that are valuable for making informed decisions. Big Data represents potential treasure for companies which will know to successfully manage and utilize it. The financial sector is guided by data in its decisions, also managing the various risks in day-to-day operations. The usage of Big Data and Data Science in banks provides many benefits, enabling them to better manage various risks and gain a competitive advantage. The most common challenge in financial institutions is the credit risk assessment, and its correct prediction is crucial for them. There is vast amount of research on credit risk but all predictions are based on the use of data sets from commercial banks, and to my knowledge there is no research that uses data from the Credit Registry database which are available only in central banks. Financial news represents an important information that influence and predict various financial instruments. They also can be used for analysis of companies by analyzing public textual content for them. The uniqueness of the dissertation is that so far there is no research that predicts credit risk using the Credit Registry database and using the additional knowledge from financial news in Macedonian and Albanian language.

The doctoral dissertation proposes a platform for predicting credit risk. The platform integrates credit risk prediction for a client or company using data from the Credit Registry. In the platform there is also integrated analysis of financial news using Named Entity Recognition and Sentiment Analysis. These two separate models are created using financial news in Macedonian and Albanian language.

The platform enables better credit risk management with the help of additional opinion from the central bank. Credit risk analysis is helped additionally by financial news that analyzes the company in which the client is employed or the company that is applying for a loan. The platform is designed for economists and brokers, providing them the unique benefits for dealing with credit risk.

Key words: credit risk, credit registry, natural language processing, named entity recognition, sentiment analysis

Благодарност

Најпрво огромна благодарност за несебичната поддршка на мојот ментор Игор Мишковски, без чија помош моите докторски студии ќе немаа патоказ, мотивација и успех. Неговата посветеност и помош за сите двоумења несомнено ми помогна и ме насочи да ги завршам успешно сите активности во овие студии. Ме охрабруваше во секој предизвик и неговите совети бескрајно ми помогнаа во текот на моите докторски студии.

Изразувам голема благодарност до моето семејство за нивната поддршка, мотивација и трпение при целото време посветено на докторските студии.

Благодарност до моите родители кои ме насочија во изборот на додипломските студии а потоа и на нивната поддршка за магистерските и докторските студии. Оваа одлука беше клучна во мојот живот и ме насочи кон кариера во ИТ индустријата.

Бескрајно благодарност до мојата сопруга Алма, за нејзината љубов, поддршка, жртвување и трпение без кои немаше да биде возможна изработката на оваа докторска дисертација. На крај, мора да го спомнам и синот Лис кој последната година ми беше извор на енергија и радост.

Содржина

Листа на слики

Листа на табели

1. Вовед

Живееме во свет и време кога апликациите и податоците се во секој наш чекор, почнувајќи од мобилните телефони, социјалните мрежи и сè до банкарските трансакции. Создавањето на нови податоци расте експоненцијално последните години, и овој голем раст на податоците создава потреба за нивна обработка и анализа со нови технологии. Овие големи податоци имаат огромен потенцијал за откривање нови знаења и шаблони кои ќе помогнат за носење на информирани и вистински одлуки. Науката за податоци и машинското учење се главните двигатели за овие големи податоци, и со нивна помош придобивката од анализата на овие податоци е огромна.

Експоненцијалниот раст на технологијата и зголемувањето на генерирањето на податоци го трансформираат начинот на кој работат повеќето компании и организации. Финансискиот сектор се смета еден од главните сектори што се води од податоците и затоа секогаш е во потрага по нови можности за обработка, анализа и искористување на податоците на корисен начин. Големите податоци сè повеќе се потенцијално богатство за финансиските институции и потребни им се вистинските алатки за да ги монетизираат нив. Примената на големите податоци и на науката за податоците обезбедуваат банките подобро да ги анализираат различните ризиците, да се стекнат со конкурентска предност, да ги подобрат нивните работни процеси, да предвидат различни ризици и информации поврзани со корисниците, и многу други придобивки.

Токму поради овие придобивки, главната истражувачка цел во оваа дисертација е да се помогнат финансиските институции со помош на нови знаења откриени од податоците добиени од Кредитниот регистар и од финансиски вести. Секако, покрај истражувачката цел, оваа докторска дисертација има и практична цел, со која финансиските податоци од Кредитниот регистар и од финансиските вести дополнети со новите знаења од истражувањето ќе се интегрираат во единствена платформа која ќе го помага справувањето со кредитниот ризик. Банките ќе можат да ја користат ова платформа за поинформирани одлуки при издавање кредит, односно предвидување за профилот на клиентот преку моделот на Кредитниот регистар и анализа за компанијата во која работи клиентот преку процесирање на финансиските вести. Оваа платформа ќе може да се користи и од брокери кои ќе ги користат придобивките од анализата на сентиментот и препознавањето на ентитети во финансиските вести и ќе може да инвестираат на по информиран начин.

1.1. Мотивација

Големите податоци претставуваат големо множество на податоци кое е големо и сложено за работа и обработката со помош традиционални алатки и технологии е скоро невозможно. Банките генерираат голем обем на податоци во пета бајти.

Револуцијата на податоците во изминатите години веќе има дополнителен ефект врз економските истражувања. Квалитетот и квантитетот на податоците за економијата и финансиските институции брзо се шират. Доколку банките сакаат да останат релевантни и конкурентни, тие треба да го преиспитаат своето работење, да прифатат пристапи водени од податоци и некако да преземаат ризик да ги пробаат новите

технологии. Важно е да се знае дека големите податоци во банкарскиот сектор можат да помогнат за подобрување на придобивката и напредување на бизнисот. Големите податоци се сметаат како синоним за анализа на клиенти, аналитика во реално време или аналитика за предвидување.

Банкарскиот сектор има огромни количини на податоци за клиентите кои растат експоненцијално (на пр. депозити, уплати, исплати, трансакции преку Интернет, податоци за клиенти итн.). Оваа огромна количина на податоци претставува одлична можност за банкарскиот сектор. Во исто време, овој сектор се соочува со предизвици за управување и анализа на овие податоци. Банките треба да ги имплементираат придобивките од обемот на собраните податоци и да ги следат трендовите на дигиталната револуција за да обезбедат подобри услуги за своите клиенти.

Користењето на големи податоци за деловните субјекти може да им помогне да добијат целосен преглед, од моделот на однесување на клиентот до внатрешните процеси, па дури и пошироките трендови на пазарот.

Колку се поголеми податоците, толку е поголем ризикот за нивна злоупотреба. Решенијата за големи податоци во финансискиот сектор им овозможуваат на деловните субјекти да имаат подобра видливост во секојдневното работење и зголемена можност за решавање на било какви проблеми.

Технологијата ги натера банките да ги користат податоците за интелигентни одлуки. Способноста за имплементирање на технологии за големи податоци за поддршка на одлуки во реално време ќе го зголеми јазот помеѓу успешните компании и тие што заостануваат со технологиите.

Често поради финансиски кризи или настани кои влијаат индиректно, се случува клиентите да изгубат доверба во банките, а тоа, доведува до масовно повлекување на средствата, што пак влијае на стабилноста на банкарскиот систем и кредитниот систем на банките. Поради овие причини банките се посветени на пронаоѓање решение за управување со ризикот и предвремено предвидување на истиот.

За предвидување на кредитниот ризик веќе долго време се користи машинското учење за негово моделирање, чија цел е да го предвиди ризикот користејќи финансиски податоци помогнати со дополнителни множества на податоци. Кога приватно или физичко лице аплицира за кредит, мора истиот да се процени дали е веројатноста голема дека ќе може да го врати кредитот (главницата и каматата) во планираниот рок.

Иако постојат различни агенции за проценка на кредитниот ризик кои нудат резултати и извештаи за одредени износи, сепак истражувачите продолжуваат да истражуваат различни техники на машинското учење за да ја подобрат проценката на кредитниот ризик.

Постојат многу истражувања за кредитниот ризик, но никое од нив не користи предвидување на кредитен ризик со базата на податоци Кредитен регистар која е централизирана база за сите издадени кредити во Република Северна Македонија. Целта на оваа дисертација е да го предвиди кредитниот ризик користејќи ги податоците од реална база на податоци - Кредитен регистар на Народна банка на Република Северна

Македонија (НБРСМ). Моделот кој е развиен во ова дисертација ќе биде дополнителен извор за предвидување на кредитниот ризик за клиенти кој претставува вредна информација за комерцијалните банки. Уникатноста е дека Кредитниот регистар ги складира податоците од сите банки за неколку години, односно историјата на кредити за сите жители на државата и од сите банки. Во ова дисертација се споредени пет алгоритми на машинско учење со цел да се предвиди кредитниот ризик: логистичка регресија, дрво на одлуки, случајна шума, векторски поддржани машини (SVM) и ациклична невронска мрежа. Според резултатите на овие експерименти е предложен модел основан на податоците за Кредитен регистар од централната банка со детална методологија што може да го предвиди кредитниот ризик врз основа на кредитната историја на населението во државата.

Во дисертацијата е направена и напредна анализа на големи податоци користејќи ја алатката Microsoft Power BI за анализата на Кредитниот регистар. Анализата ги визуелизира податоците и нивните трансформации на многу брз начин со цел да овозможи полесно и побрзо анализирање на големо податочно множество.

Примената на големите податоци за оценување на кредитниот ризик е проширено и со користење на процесирање на природните јазици (Natural Language Processing – NLP). Според досегашните истражувања е увидено е дека негативниот сентимент на вестите има повеќе влијание за кредитниот ризик, отколку позитивниот сентимент [1] [2]. Во дисертацијата се анализирани и повеќе финансиски вести од он-лајн порталите и се анализирани вестите на македонски и албански јазик со помош на библиотеките за процесирање на природни јазици. Анализата на вестите се состои во одредување на сентиментот на веста како и за кои финансиски ентитети, настани, инструменти, итн. станува збор во вестите.

Спојувањето на наведените применети истражувања во дисертацијата, се постигнува во платформата која на едно централно место ги вклучува сите наведени анализи. Платформата како алатка овозможува повеќе функционалности. Прво може да се користи за предвидување на ризик за идни клиенти и компании, преку користење на знаењето што го има учено од базата Кредитен ризик, која претставува агрегирање од кредитната историја од сите банки и како таков овој модел лесно може да се користи како готов и од комерцијалните банки. Анализата на сентиментот и препознавањето на ентитети овозможува да се земе во предвид актуелната состојба од вестите и со тоа ќе се увиди кои компании моментално се со негативни вести, и преку тоа тие компании и лицата кои работат во истите ќе се поризични при издавање кредит. Платформата може исто така да се користи и од брокери со цел да се користат придобивките од анализата на вестите (сентиментот и препознавањето на ентитети) со цел да тргуваат побезбедно и да остварат поголем профит и помал ризик, помагајќи се од информациите од вестите кои особено делуваат за тргувањата на берзата.

Според целите на дисертацијата, има повеќе придонеси каде главниот придонес е дефинирање на модел за предвидување на кредитниот ризик од уникатната база на податоци Кредитен регистар која ја има само една во секоја централна банка. Во рамки на дисертацијата предложена е и детална постапка за напредна анализа на големите

податоци преку алатка која може да ја користат луѓе од финансиските сектор. Овој начин ќе овозможи побрзо и полесно анализирање на големи податоци во банкарскиот сектор, и ќе овозможи самите вработени да вршат анализи на нивните големи податочни множества. Анализата на сентиментот на економските вести на македонски и албански јазик е исто така значаен придонес на оваа дисертација бидејќи моментално не постои таква услуга и примена во нашиот банкарски сектор. Истото се дополнува и со препознавање и анализа на ентитетите во вестите што ќе придонесе за препознавање на трендовите на финансиските информации и нивната поврзаност со компаниите, поединците како и со други финансиските ентитети.

На крајот на дисертацијата горе наведените експерименти се споени во една платформа која може лесно да се користи од луѓе од финансискиот сектор без некое знаење за позадинските ИТ процеси. Платформата ја унапредува анализата и користењето на големите податоци во банките и влијае за подобро предвидување на кредитниот ризик. Заедничката платформа како централно место за помагање за носење услуги овозможува и автоматско следење на економските вести, односно препознавање на ентитетите и детектирање на сентиментот од економските вести.

Во оваа докторска дисертација се користат методите анализа, синтеза и експериментирање. За да се дефинира моделот за предвидување користејќи го Кредитниот Регистар е анализирана проблематиката и идентификувани се сите чекори и атрибути кои треба да се користат при градење на моделот. Притоа се применуваат и евалуираат повеќе модели за предвидување на Кредитниот Регистар. Откако ќе се одредат најдобрите модели за предвидување и за анализа на вестите, со методот на синтеза составени се заклучоците со што е дефиниран моделот за предвидување кредитен ризик.

Истражувањето во докторската дисертација ги има следните цели:

1. Дефинирање на модел за предвидување на кредитниот ризик користејќи го Кредитниот регистар
2. Анализа на податоците со визуелна алатка наменета за луѓе од финансискиот сектор
3. Анализа на сентиментот на финансиските вести
4. Препознавање на ентитетите во вестите и нивното поврзување со финансиските инструменти
5. Интегрирање на горе наведените точки во една заедничка платформа

1.2. Хипотези на истражувањето

Главната хипотеза на докторската дисертација гласи:

Кредитниот регистар има голем обем на историски податоци за секој издаден кредит, и според ова има доста знаење во него кое може да помага за предвидување на кредитниот ризик.

Дополнително, ги дефинираме и следниве поспецифични хипотези:

- Предвидувањето од Кредитниот регистар ќе биде со точност од околу 90% поради големиот обем на податоци.
- Анализата на сентиментот и ентитетите на вестите ќе овозможат поефикасно предвидување на кредитниот ризик. Дополнително, развиените модели ќе допринесат и за збогатување на информациите во банкарските, односно финансиските институции.

1.3. Преглед на докторската дисертација

Во Глава 2 е даден вовед за големите податоци и примената на големи податоци во банкарскиот сектор. Прво се опишува примената на големи податоци општо во банките, потоа анализата се фокусира на примената на големи податоци во централните банки, дополнето со придобивките и предизвиците за нивната примена. Во ова Глава е даден и краток осврт на науката за податоците, машинското учење и алгоритмите кои се користат во оваа дисертација. Потоа се воведува проблематиката за кредитен ризик кој дел продолжува со главен фокус на базата Кредитен регистар која е и главниот извор на податоци во дисертацијата.

Следно, во Глава 3 е направена детална напредна анализа на податоците на Кредитниот регистар со алатката Power BI. Во анализата се вклучени и други извори на податоци за да се увиди корелацијата со други случувања во финансискиот свет. Податоците се визуелизирани во различни извештаи и прикази и со овој дел е постигнато првично разбирање и претставување на податоците од базата Кредитен регистар.

Понатаму, во Глава 4 првично е даден осврт на досегашните истражувања и анализи поврзани со кредитниот ризик. Во анализата се споредуваат научни трудови од последните години и како анализата еволуирала низ годините. Во оваа глава е предложена методологија за предвидување на кредитниот ризик со користење на податоците на Кредитниот регистар. Уникатноста на овој дел е дека досега нема некое истражување кое го предвидува кредитниот ризик со помош на оваа база. По деталната методологија и чекорите за пред обработка на податоците, направена е најдобрата селекција на атрибутите по што се применети пет различни алгоритми за машинско учење и на крај е претставено детална анализа на резултатите.

Во Глава 5 е опишана обработката на текстуалните податоци со цел да може да се користат за машинско учење и да се препознаваат ентитети во македонските вести. По описот на различните техники за анализа на текст и процесирање на природните јазици е дадено и опис за примена на модели со: меморија во кратки интервали на долги

периоди - LSTM, двонасочна меморија во кратки интервали на долги периоди - Bi-LSTM и како и со условно случајно поле - CRF. Опишана е потребата за векторска репрезентација на зборовите со фокус на контекстуалната векторска репрезентација и на векторска репрезентација со користење на BERT моделот. И на крај на оваа глава е опишана методологијата, и резултатите од новиот модел за препознавање на ентитети на македонски јазик, кој се применува за анализа на финансиските вести. За препознавање ентитети со 13 лабели, за вестите на македонски јазик е постигнат F1 резултат од 0.75.

Во Глава 6 се користи анализа на сентиментот и како тие влијаат на банкарскиот сектор. Направено е краток осврт на развојот на анализа на сентиментот и различните пристапи за анализирање на истиот. Главната цел е анализа на чувства со користење на модели за длабоко учење. Потоа се опишуваат рекурентните и конволуциските невронски мрежи и нивните разлики. Посебен акцент во оваа глава имаат и трансфомерите и нивната архитектура. Во практичниот дел е применет Bert multilingual [3] и XLM-RoBERTa [4] за вестите на македонски и албански јазик.

Глава 7 е крајниот производ од оваа дисертација, односно платформата во која се интегрирани сите претходно наведени теоретски истражувања и нивната практична примена. Платформата како краен производ е интеграција на сите наведени придобивки од големите податоци со цел банкарскиот сектор да може да ги искористи придобивките во реалност.

На крај, во Глава 8 се резимираат придонесот и резултатите од истражувањата опишани во оваа докторска дисертација. Освен заклучоците се наведени и идни насоки за истражувања и се посочени предизвиците кои треба да се надминат.

2. Примена на големи податоци во банкарскиот сектор

Концептот на големи податоци е релативно нов во денешно време. Луѓето почнаа да сфаќаат колку многу податоци се генерираат со користењето на онлајн услугите. Развојот на социјални мрежи и мобилните телефони е од суштинско значење за растот на големите податоци. Компаниите веќе се свесни за присутноста на големите податоци и се трудат да ги користат придобивките во нивната секојдневна работа. Поимот големи податоци е поради нивниот обем, брзината со која се генерираат и разновидноста на податоците. Тие користат големи податоци и наука за податоци за да откријат нови знаења и да ги искористат податоците за носење на информирани одлуки. Големите податоци во банкарскиот сектор овозможуваат на банките да ги подобрат своите процеси, да ја подобрат добивката и да останат пред конкуренцијата. Големите податоци имаат потенцијал да ги трансформираат работните процесите на компаниите, особено во банкарскиот сектор, бидејќи тие имаат огромна количина на податоци за нивните клиенти. Големите податоци претставуваат збир на податочни множества кои се големи и сложени за обработка со помош на традиционалните алатки за управување со базите на податоци. Поради тоа се појави потребата за нови алатки и техники за да може истите во реално време да се обработат и навремено да се користат новите здобиени знаења и заклучоци. Банките користат големи податоци за да ги променат нивните деловни процеси, нивната организација и целата индустрија. Исто така, тие користат големи податоци за да го предвидат движењето на акциите и хартиите од вредност преку развој на алгоритми за предвидување. Во финансискиот сектор во последните години се повеќе се користи и процесирањето на природните јазици и поврзување на соодветните заклучоци со финансиските анализи и проценки.

Целта на оваа глава е да даде преглед на различните пристапи и предизвици што постојат во големите податоци во банкарскиот сектор. Првично се опишани големите податоци во банкарскиот сектор (комерцијални и централни банки), потоа соодветните придобивки и предизвици. Бидејќи следува дел за анализа на кредитниот ризик каде се користи машинско учење, направен е краток осврт на користените алгоритми за машинско учење. На крај на оваа глава е опишан кредитниот ризик, досегашните истражувања за истиот и опис на Кредитниот регистар.

2.1. Големи податоци во банки

Не постои единствена официјална дефиниција за концептот големи податоци. Под концептот големи податоци се подразбира собирање и анализа на големи количини на структурирани и неструктурирани податоци, потенцијално во реално време за да се создаде вредност за компаниите. Концептот на големи податоци во финансиски контекст е различен од другите индустрии [5]. Дефиницијата за столбовите на големи податоци, еволуирала низ годините. На почеток постоеле три столбови според дефиницијата на Gartner, потоа најпозната дефиниција за големите податоци е со 5 столбови (Слика 1) (5Vs) [6]:

1. **Волумен**: со зголемената употреба на мобилните телефони и социјалните мрежи, експоненцијално се зголеми креирањето и обемот на податоците. Овие големи множества на податоци не е можно да се чуваат и анализираат од

традиционалните релациони системи. Банките[1] во текот на годините се справуваат со голем обем на податоци и секогаш биле во групата со големи податочни множества. Комерцијалните банки собираат податоци за трансакциите од клиенти, додека централните банки собираат податоци од комерцијални банки и финансиски институции.

2. **Разновидност**: постои разновидност на податоците од структурирани (релациони податоци) до неструктурирани (нерелациони податоци, слики, видеа, аудио). Големите податоци им овозможуваат на корисниците не само да ги анализираат структурираните податоци што постојат во банкарскиот сектор, туку и големиот обем на комплексни неструктурирани податоци кои стануваат се порелевантни со цел да откријат нови сознанија и наоди.
3. **Веродостојност**: се однесува на доверливоста на податоците во големите податоци, особено ако се добиени од јавни извори на трети лица. Веродостојноста се зголемува експоненцијално со обемот на податоците.
4. **Брзина**: е концепт кој ја опишува брзината на генерирање на податоци од различни извори. Податоците често се ажурираат и може брзо да се анализираат со можност за анализи во реално време.
5. **Вредност**: со откривање и предвидување на нови знаења основани од анализата на постоечки и историски податоци, банките можат да создаваат вредност за клиентите нудејќи им нови услуги.

Слика 1. 5 столбови (5Vs) на големи податоци

Големите податоци се модерно поле каде технологијата и науката за податоци обезбедуваат нови начини за извлекување вредност од океанот на нови информации.

[1] https://www.ibm.com/downloads/cas/E4BWZ1PY - Analytics: The real-world use of big data in financial services

Како клучна предност пред конкуренцијата е можноста за ефикасно управување со увидите (insights) и извлекување знаења[2] [5]. Банките ја користат науката за податоците за поддршка и предвидување на различните финансиски ризици и откривање измами [6].

Банките треба да ги искористат придобивките од обемот на податоци што ги собираат и трендовите на дигиталната револуција за да обезбедат подобро прилагодени услуги за своите клиенти во сѐ по конкурентниот дигитален свет[3]. Комерцијалните банки вклучуваат широк спектар на банкарски услуги кои ги нема во централните банките, како што се трансакциски сметки, штедни сметки, кредити, кредитни картички, услуги за е-банкарство, телефонски банкарски услуги итн. Дополнително, Интернетот и мобилното банкарството го сменија банкарскиот сектор, правејќи го поразличен од пред една деценија. Користењето на социјалните мрежи и мобилните апликации доведе до намалување на интеракциите лице во лице помеѓу клиентите и банките, а во меѓувреме доведе до зголемување на виртуелните интеракции со што се зголеми обемот на податоците за клиентите. Податоците што банките ги имаат за своите клиенти се големи по обем и се разновидни. Обработувајќи ги големите податоци може ефикасно да се користат податоците за клиентите, помагајќи да се прошират персонализираните производи и услуги.

2.2. Големи податоци во централните банки

Големите податоци во централните банки се со голем обем, бидејќи податоците се пријавуваат од повеќе банки и финансиски институции, се собираат и анализираат. Пример се податоците за кредит, хартии од вредност, итн. Недостаток е што централните банки имаат поопшти агрегирани податоци од комерцијалните банки и податочните множества најчесто се само нумерички.

Централните банки веќе имаат големи податочни множества за статистика, структурирани податоци и информации, кои редовно се користат во нивниот процес на носење одлуки. Кредитниот регистар најчесто е најголемото податочно множество во повеќето централни банки.

Искуството на централните банки со големи податоци е анкетирано и од меѓународната анкета BIS-IFC Big Data [7]. Во резултатите од 2015 година, нема јасно разбирање за дефиницијата за големи податоци. Примарниот фокус на централните банки е пристапот и обработката на податоците.

Ветувачко истражување е истражувањето спроведено од Централното банкарство во соработка со BearingPoint во текот на средината на 2017 година. Истражувањето докажува дека големите податоци се активна област за нови проекти во централните банки и за клучните проекти за големи податоци во централните банки се забележува користењето на Кредитниот регистар, потоа проектите со консолидација на внатрешни системи и извори [8]. Во годишниот извештај на IFC 2018 [9], повратните информации од централните банки се однесуваат на сложените импликации врз

[2] https://www.cognizant.com/InsightsWhitepapers/Banking-on-Data-Science.pdf S. Dubey and S. Nainwani, Cognizant, 2019

[3] https://www.fintechbusiness.com/blogs/759-big-data-and-customer-engagement

приватноста при работењето со големи податоци. Последното истражување спроведено во 2018 година од страна на Централното банкарство во соработка со BearingPoint [10] извештаи за пристапот на централните банки кон големите податоци. Централните банки започнаа да бараат надворешни извори за да добијат повеќе придобивки. Комплексните импликации врз приватноста при справувањето со големите податоци и компликациите за користење на другите извори се зголемуваат со текот на времето. Проектите за големи податоци сеуште се сметаат за истражувачки процеси и не е евидентирано некоја значителна примена во продукција.

2.3. Придобивки од големите податоци за банкарскиот сектор

Големите податоци и науката за податоци како современа технологија, ги поддржува секоја банка со цел поефикасно водење на своите деловни активности. Тие се однесуваат на полето на примена на софтверски технологии во комбинација со напредни алгоритми и методи со цел да се добие поголем увид, да се донесат подобри одлуки или да се предвидат различни ризици.

Во следната листа се наведени некои од најпознатите предности на употребата и анализата на големите податоци во банкарскиот сектор и случаите на примена на науката за податоците. Наведените примени ги претставуваат и придобивките на банките од овие технологии во периодот од 2018 до 2020 година[4, 5]:

- **Справување со ризици**. Големите податоци можат да бидат насочени кон потребите на организацијата и да се применуваат за подобрување на различните области на ризик: кредитен ризик [11], [12], ликвидност, оперативен ризик и ризик на финансискиот пазар [13].
- **Увид на задоволството и барањата на клиентите**[6]. Банките го искористуваат големиот обем на податоци за клиентите преку повеќе канали за да ги откријат моделите на однесување на клиентите и да го зголемат знаењето за потрошувачите.
- **Подобрено откривање на измама**. Технологиите за големи податоци овозможуваат анализа во реално време на поголеми множества на утврдување на измамата [14]. Тие овозможуваат корелација на податоците од различни извори или инциденти за откривање на измама. Со оглед на огромната количина на податоци што треба да се анализираат, архитектурата за големи податоци овозможува побрзо и поточно откривање на измами.
- **Алгоритамско тргување и предвидување на берзата**. Комбинирањето на различни групи на податоци од повеќе пазари и берзи обезбедува подобрен поглед на пазарот што може да генерира подобри сигнали и профит [15]. Алгоритамското тргување користи компјутерски програми за автоматско тргување без човечка интервенција. Ова се постигнува со користење на огромни историски податоци за да го пресмета односот на успехот на

[4] https://www.techiexpert.com/top-data-science-use-cases-in-finance

[5] https://intetics.com/blog/top-5-machine-learning-use-cases-for-the-financial-industry

[6] https://www.digitalistmag.com/customer-experience/2019/01/29/data-driven-analytics-practical-use-cases-for-financial-services-06195123/

алгоритмите и да оцени илјадници хартии од вредност со комплексни математички алатки. Тие исто така комбинираат и анализираат податоци од релевантни финансиски извори за направат поточно предвидување и одлуки.

- **Предвидување**. Предвидување идни настани кои може да се случат, преку анализа и разбирање на социјалните медиуми, трендовите на вестите и другите извори на податоци. Ги предвидува движењата на пазарот и финансиските средства на клиентите.

Во продолжение се наведени најважните анализи на големите податоци во областа на маркетингот на банките:

- **Анализа на сентиментот**. Следење на социјалните мрежи за да се зголеми успехот во маркетинг и правилно прилагодување на кампањите за маркетинг, идентификување на клиенти со големо влијание во социјалните мрежи бидејќи тие се клучни за исполнување на целите [16].
- **Профил на клиенти**. Идентификување на профилот на клиентот користејќи повеќе атрибути за да се испитаат навиките и да се изгради целосен холистички профил на клиентот. Ова анализа овозможува разбирање на навиките на клиентот со цел да му се испрати наменета маркетинг порака за нови кампањи за кредити. Пример банки кој го имплементирале: OCBC Bank, HDFC Bank, Austria Bank.
- **Сегментација на клиенти**[7]. Големите податоци овозможуваат побрза и поостра класификација на клиентите во различни сегменти кои споделуваат слични карактеристики или однесувања основани според однесувањето на потрошувачите и различните атрибути.
- **Следната најдобра понуда**[8]. Овозможува зголемување на можностите предвидувајќи што следно сака клиентот, користејќи систем на препораки за да предвиди желби на клиентите (според историски трансакции).

2.4. Предизвици

Револуцијата на големи податоци е многу значајна за подигнување на свеста. Анализата на релевантни големи податочни множества ни дава поголемо разбирање за светот, и овозможува да се направат предвидувања и решавање на проблеми. Од друга страна, ваквата анализа сепак со себе носи етички импликации во однос на незаконско користење на цели овие информации за нарушување на приватноста и деталното следење на корисниците.

Примената, процесирањето, анализата како и извлекувањето на знаење од големите податоци е многу сложен процес што подразбира многу промени во ИТ системите на банките. Постојат бројни предизвици за примена на големите податоци во банкарскиот сектор, како што се инфраструктурата, приватноста на информациите и

[7] https://medium.com/activewizards-machine-learning-company/top-9-data-science-use-cases-in-banking-6bb071f9470c

[8] https://activewizards.com/blog/top-9-data-science-use-cases-in-banking/

трошоците за складирање. Во продолжение се опишани некои генерални предизвици што треба да се надминат за да се има успешен проект за големи податоци.

Приватност и безбедност. Меѓу клучните етички прашања во прибирањето на податоците е правото на приватност, кое им овозможува на луѓето до го ограничат кој има пристап до нивните лични податоци. Анонимизацијата е минималниот потребен услов за да се заштити приватноста на податоците и корисниците. Со зголемувањето на обемот на личните информации кои се чуваат, се зголемува и потребата за повеќе правила за регулирање на оваа социјална трансформација. Корисниците треба да се запознаени со начинот на прибирање на податоците, како се користат тие, како се чуваат и споделуваат истите[9]. Справувањето со големи податоци е пофлексибилно во облакот, но прописите за приватност и безбедност честопати ја ограничуваат оваа одлука за миграција во облакот. Големите податоци се соочуваат со критики за надминување на границите на приватноста. Анализите на големите податоци се ограничени со бројни регулативи за заштита на податоците и приватноста што влијаат на анализите, бидејќи поединците можат да одбијат да се користат нивните лични податоци од обработка во одредени околности. Етичките проблеми треба да бидат разгледани во вакви сценарија, бидејќи анализите за предвидување ќе ги идентификуваат луѓето со ниски социјални услови и поради немање интерес да работат со нив новите услуги ќе ги насочат само за повисоки социјални класи [17].

Складирање и обработка. Додека банкарските структурирани податоци постојано растат, неструктурираните податоци растат побрзо и стануваат сè повлијателни. Ова ја зголемува потребата да се има неструктурирани бази на податоци од неколку тера бајти. Конвенционалните техники за управување со податоци веќе не се доволни за да се справат со масовната големина, големата брзина и хетерогеноста на податочните множества [18]. Кога се користат има видливи ефекти во перформансите на податоците поставени во облакот за заедничка анализи во реално време со податоци што се наоѓаат во компанијата. Сепак најголемите проблеми при користење на облакот се приватноста и регулаторните импликации.

Технички предизвици. Големината и обемот на податоците брзо се зголемуваат, затоа е многу важно да се користат соодветни техники и технологии што може да се справат со огромно количество на големи и разновидни комплексни множества на податоци. Банкарскиот сектор заостанува во примената на новите инфраструктурни компоненти како Hadoop, NoSQL, Map Reduce [19].

Аналитички предизвици. За да имаме придобивки од големите податоци, мора да се користат аналитички способности и алатки. На банките им се потребни квалификувани научници за податоци (data scientists) за да може да ги користат можностите на големите податоци. Поради строгите правила на управување во банкарството, на банките им недостасуваат посебни работни позиции за научниците за податоци. Исто така, поради нивните стандарди и регулативи, банките имаат застој во

[9] http://www.rss.org.uk/Images/PDF/influencing-change/2016/rss-report-opps-and-ethics-of-big-data-feb-2016.pdf

примена на технологии за аналитичките способности на анализата на текстот, чувството, гласот и видеото.

Сопственост. Сопственоста е комплексен концепт, кој се однесува на правото на редистрибуција и промена на податоците, како и придобивките од интелектуалната сопственост и иновациите направени преку анализи на овие податоци. Редистрибуцијата и промената на податоците може да се ограничи од сопственикот, меѓутоа останува дискутабилен пристапот за анализа и развој со тие податоци. Сето ова вовлекува две форми на сопственост, права за контрола на податоците и права за придобивка од извлечените информации.

Транспарентноста. Транспарентноста како еден од главните столбови на човековото општество, може да помогне да се спречи злоупотребата на институционалната моќ, а воедно да ги поттикнува корисниците да се чувствуваат побезбедно во размена на порелевантни информации кои би помогнале за да се подобрат предвидувањата извлечени од големите податоци.

Користејќи ги алатките и технологиите на науката за податоците, банките можат поефикасно да се информираат при стратешкото донесување одлуки, намалувајќи ја неизвесноста и елиминирајќи ја неизвесноста. И покрај сите предности на големите податоци, тие имаат ограничувања кога станува збор за имплементација во банкарството. Комерцијалните и централните банки користат големи податоци на различни начини, но тие мора прво да ги надминат сите споменати предизвици за да бидат во чекор со технологијата и да добијат максимална корист од големите податоци. Банките треба постојано да ги ревидираат политиките и регулаторните стандарди за да можат да усвојат нови технологии. Горенаведената анализа покажува дека сè уште има предизвици за истражување на сите нивоа и вклучуваат широк дијапазон на различни технологии. И покрај технолошките аспекти, постојат организациски, културни и правни фактори кои диктираат како банките ќе го продолжат справувањето со големите податоци во деловните активности и процеси.

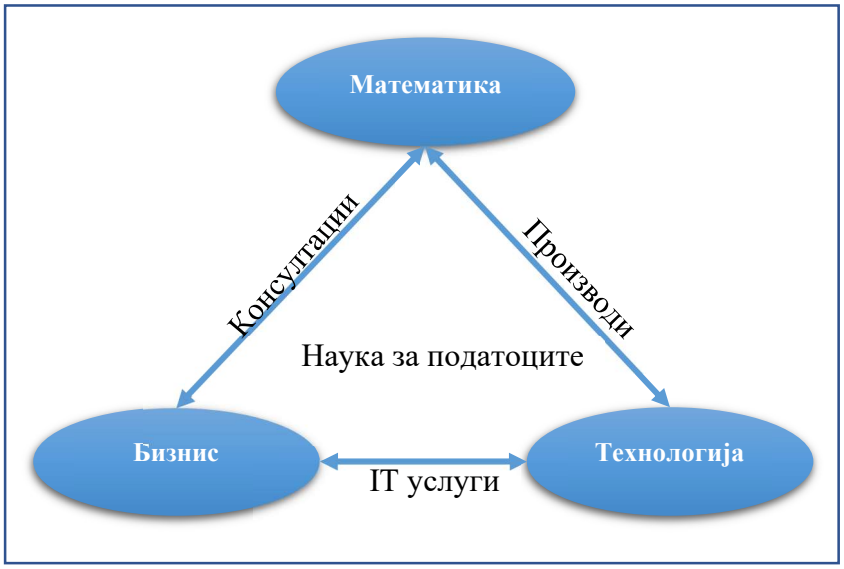
2.5. Наука за податоците

Компаниите во денешно време се соочуваат со огромни податочни множества, затоа секоја индустрија е насочена кон искористување на податоците за да постигнат предност пред конкуренцијата. Покрај растот на големината и разновидноста на податоците, со тек на времето и компјутерското процесирање стана помоќно. Сите овие промени доведоа до сè пошироко распространета примена на науката за податоци. Науката за податоците (Data Science) [20] е мултидисциплинарна област насочена кон откривање знаење од големи податочни множества. Истата вклучува обработка на големи структурирани и неструктурирани податоци, вклучувајќи нивна подготовка, селекција и анализа.

Науката за податоци може да се дефинира како наука за носење одлуки управувани од податоците. Областа е интердисциплинарна комбинација на множества на вештини и збир на принцип кои овозможуваат исцрпување на информации и знаење од податоците (Слика 2).

Науката за податоци е во основа комбинација на:

- Математика - Статистика, линеарна алгебра, веројатност итн.
- Бизнис - Познавање на областа на проблематиката
- Технологија - Вештини за програмирање



Слика 2. Наука за податоците

Со присутноста на големите податоци стигнаа и нови предизвици за одлуките управувани од податоците, и со тоа се појави и потребата за трансформација на големите податоци во информации и знаење. Затоа науката за податоците е неразделен дел од големите податоци. Како главни столбови на науката за податоците се податоците, технологијата и луѓето [21]. Податоците се насекаде, технологиите се развиваат брзо, меѓутоа луѓето се компонентата која заостанува повеќе. Оваа нова професија има потреба од луѓе кои ќе ја играат својата улога со примена на науката за податоците за решавање на проблеми со големи податоци, спојувајќи ја празнината помеѓу податоците и технологијата. Големите податоци и науката за податоци се користат скоро насекаде и во комерцијални и во некомерцијални цели. Големите податоци имаат важно влијание во многу сектори и светски економии, како што се здравството, производството и трговијата, државниот сектор, финансиските услуги итн. Технологијата не само што ќе им помогне на финансиските институции да ја зголемат вредноста на податоците кои ги поседуваат, но им помага да стекнат конкурентски предности, да ги минимизираат трошоците, да ги претворат предизвиците во можности и да ги минимизираат ризиците во реално време.

2.6. Кредитен ризик

Кредитен ризик е веројатноста за загуба како резултат на неуспехот на клиентот да изврши исплата на долгот во планираниот временски рок. За предвидување на кредитниот ризик веќе долго време се користи машинското учење, чија цел е да го предвиди ризикот користејќи финансиски и други податоци. Кога приватно или физичко

лице аплицира за кредит, мора за истиот да се процени веројатноста дека ќе може да го врати кредитот (главницата и каматата) во планираниот рок.

Иако постојат различни агенции за проценка на кредитниот ризик кои нудат резултати и извештаи за одредени износи, сепак истражувачите продолжуваат да истражуваат различни техники на машинското учење за да го подобрат оценувањето на кредитниот ризик.

Моделите можат да се подобрат кога има и други извори на податоци со можност за комбинирање и спојување на податоците од повеќе извори. Добро трениран модел може потоа да извршува автоматско оценување за кредитниот ризик и да им помогне на вработените да работат многу побрзо и попрецизно.

Трудот [22] покажува дека моделот на кредитно оценување базиран на невронска мрежа е поефикасен во скрининг на стандардните апликации. Во [23] се воведува двостепен систем за предвидување на кредибилитет на заем, кој користи алгоритам за индукција на дрвото на одлуки за предвидување. Во [24] се претставува нов пристап на комбинација основан на консензус на шест класификатори што создава групно рангирање. Трудот [25] покажува дека употребата на податоците за плаќање на клиентот се многу важен фактор за подобрување на предвидувањето на кредитниот резултат. Во [26] се демонстрира дека најзначајни атрибути при утврдување на резултатот од кредитната апликација се приходот, годините на стаж, кредитниот резултат и дали клиентот успешно ги исплатил претходни кредити. Податоците од социјалните мрежи се голем фактор за одредување на кредитниот резултат [27]. Сеопфатниот труд [12] кој анализира голем број на трудови, презентира поголем увид во употребата на големите податоци за моделите за кредитно оценување. Ова обезбедува докажана основа дека банките треба да развиваат нови модели и да воведат нови извори на податоци што значително ќе го подобрат моделот за предвидување на кредитниот ризик. Авторите на овој сеопфатен труд предлагаат модели со големи податоци кои ќе вклучуваат различни податоци за клиентите, вклучувајќи каде купуваат, нивните набавки, нивните профили на социјалните мрежи и други фактори кои не се директно поврзани со кредитната способност.

2.7. Кредитен регистар

Кредитниот регистар претставува база на податоци и информации за кредитната изложеност на банките и штедилниците во Република Северна Македонија кон нивните клиенти, чијашто основна намена е да придонесе за подобрување на квалитетот на кредитите и за одржување на стабилноста на банкарскиот систем. Кредитниот регистар е збирка на лични податоци, чијшто контролор е НБРСМ.

Целта на Кредитниот регистар е да овозможи:

- централизирање на податоците и информациите за кредитната изложеност кон клиентите, доставени од страна на банките и штедилниците;
- користење на податоците и информациите за кредитната изложеност кон клиентите, од страна на банките и штедилниците, за потребите на управувањето со кредитниот ризик;

- користење на податоците и информациите за кредитната изложеност кон клиентите, на одделните банки и штедилници и банкарскиот систем во целина, од страна на Народната банка, за потребите на вршењето на супервизорската функција.

Податоците се праќаат за секој месец и за секое физичко и правно лице вклучуваат информации за кредити, кредитни картички и дозволени минус плати.

Кредитниот регистар се користи како веб сервис на институциите, преку кој ќе добиваат податоци за кредитоспособноста на клиентите во Република Северна Македонија. Исто така, во рамките на НБРСМ, регистарот се користи како моќна алакта за добивање на предефинирани или диманички генерирани извештаи.

Сите функции на превземање и доставување на податоци во системот се вршат со соодветент веб сервис, преку негови предефинирани методи.

Размената на податоците со институциите се врши преку XML датотеки кои имаат предефинирани XSD шеми и кои треба да бидат дигитално потпишани со сертификат со XML стандард. Бидејќи овој систем се базира на податоци за клиенти, потребно е најпрвин да се внесат во системот сите клиенти за кои институцијата ќе доставува податоци. Внесувањето на податоците за клиенти претходи на сите останати доставувања на податоци. Овие податоци треба целосно да се точни пред внесување на останатите податоци. При отварање на нов клиент, институцијата мора да ги внесе податоци за клиентот пред да ги внесува останатите податоци. Ако институцијата не ги внесе податоците за клиентот, останатите податоци нема да бидат прифатени во системот. По внес на податоците за клиенти, институцијата треба да ги внесува податоците дефинирани по кредитна партија на клиент, вонбилансно прекнижените побарувања и отпишаните кредити. Секоја XML датотека е порака која има предефинирано значење. Во структурата се содржи код-назив, преку кој се идентификува постапката која треба да се направи со податоците во датотеката.

3. Напредна анализа на податоците на Кредитниот регистар, со користење на Power BI

Големите податоци се растечки тренд кои носат и потреба за поефикасна анализа и визуелизација на податоците. Анализата и визуелизацијата на голем обем на податоци во банкарскиот сектор честопати страда од перформанси во традиционалните системи со користење на традиционалните алатки. За визуелно разбирање на податоците на базата Кредитен регистар, во оваа докторска дисертација се користеше Power BI. Со експоненцијалното зголемување на количината на податоците, се покренува потребата да се разберат трендовите во бизнисот и да се добие важен увид од постојните податоци. Различни бизниси треба да ги разберат аналитичките концепти користејќи статистички методи, предвидување на податоци и машинско учење. Овие анализи во минатото ги правеа само програмери, но сега овие современи алатки ги обезбедуваат овие способности и за луѓето од бизнисот. Microsoft Power BI е алатка за напредна анализа, што им овозможува на обичните корисници извлекување на корисно знаење од податоците, визуелизации на податоци и интеграција со R. Power BI овозможува анализи за предвидување на излезни променливи со бројни вредности, користејќи машинско учење без никакво програмирање.

По разгледување на повеќе алатки, избрано за користење е Power BI поради големата ефикасност и брзина на манипулација и анализа на податоците. Во наредниот пример, дел од базата на податоци од 13 GB во SQL при импорт со SQL изрази е намалена на 330mb во .pbix форматот на Power BI, формат кој е прилагоден за работа со големи податоци [28].

Power BI е алатка за бизнис аналитика која дава увид за да се овозможат брзи и информирани одлуки. Power BI е множество на софтверски услуги, апликации и конектори кои работат заедно за да ги претворат неповрзаните извори на податоци во кохерентни, визуелни и интерактивни прикази. Податоци може да бидат табела во Excel на облак или локална база на податоци. Алатката овозможува лесно поврзување со изворите на податоци, визуелизирање, откривање на важни информации, и споделување на истите преку веб и мобилна апликација.

Според Gartner[10], Power BI е апликација лидер за бизнис интелигенција која за брзо време стана најпозната и вредна за големите корпорации надминувајќи ги Click и Tableau.

Power BI се состои од:

- **Power BI Desktop** - Апликација за Windows десктоп.
- **Power BI Service** - Интернет SaaS (Софтвер како услуга) услуга .
- **Power BI Mobile** - Мобилна апликации за уреди со Windows, iOS и Android.

Обработката со Power BI почнува со користење на Power BI Desktop односно импортирање на податоците, поставување релации и визуелизација. Откако анализите

[10] https://info.microsoft.com/ww-landing-2020-gartner-magic-quadrant-for-analytics-and-business-intelligence.html?LCID=EN-US?LCID=EN-US

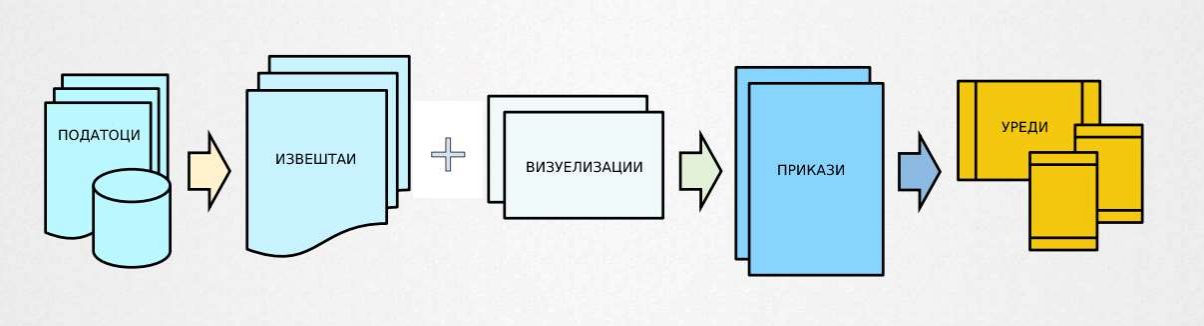
ќе завршат на локалната околина, истите може да се публикуваат на SaaS верзијата на Power BI на облак (Слика 3) каде може да се додели пристап на корисници преку кориснички сметки на Azure Active Directory. Во ист момент кога се публикуваат на SaaS, податоците, извештаите и приказите се достапни и преку официјалната мобилна апликација. За уредно прикажување на мобилен, е потребно уредување на изгледот за мобилен преку десктоп апликацијата. На ваков начин брзо и ефикасно се постигнува преглед и увид на анализите преку десктоп, веб и мобилна апликација.

Слика 3. Power BI апликации и нивно поврзување

3.1. Процес на изведување на анализите

За изведување на анализите се направени неколку фази на обработка на податоците за да се постигнат крајните резултати како што е прикажано на Слика 4.

Првично податоците се импортирани од повеќе извори, уредени и дополнети во Power BI Desktop. Креиран е модел, како и соодветни извештаи и прикази кои на крај се публикувани на верзијата на Power BI на облак и потоа истите се достапни преку Веб и мобилна апликација.



Слика 4. Процес на изведување на анализите

Во продолжение е детализниот процес почнувајќи од изворот на податоци до крајно прикажување на анализите.

3.1.1. Анализа со SQL Server

- **Анализа –** изворот на податоци во SQL, првично имаше 52 атрибути. Анализирани се минималните и максималните вредности. Важен е фактот дека поради законски измени низ годините, не сите полиња имаат податоци.
- **Празни вредности –** Празните вредности првично се лоцирани преку SQL код, и потоа за бројните колони е внесено нула.
- **Одбирање на атрибути (Feature selection) –** Лоцирањето на најважните атрибути е направено со помош на колеги економисти кои се специјализирани во делот за менаџирање на Кредитниот регистар. Со нивна помош се лоцирани околу 15 најважни атрибути.
- **Инженерство на атрибути (Feature engineering) –** Во интерес на ефикасноста на моделот се креирани и нови изведени атрибути.
    - **GoleminaNaBanka** – изведена колона според шифрарник на големина на банки
    - **BrojKrediti** – за секое лице со помош на SQL код е најден бројот на кредити во тековниот период за кој се известува
    - **DatumStartKredit** – Увидено е дека колоната за датум на прв прилив на парите има доста нелогични датум поради претходно отсуство на контрола за таа колона, затоа е изведена нова колона која го поставува датумот на почнување на кредитот во моментот кога прв пат се појавува таа кредитна партија во база за соодветниот клиент.
    - **GodiniTraenjeKredit** – Изведена колона со помош на колоната за датум на завршување на кредитот.
    - **Vozrast** – Возраста е изведена за физичките лица преку нивниот матичен број. Овој процес е внимателно изведен од администраторите на бази на податоци кои од продукција на развој префрлија табела со колона анонимизиран матичен број кој се користи за спојување со постоечката табела на развој, и возраста која се добива со издвојување на шестиот и седмиот карактер од оригиналниот матичен број од продукција.

3.1.2. Дополнителни извори на податоци

За подобрување на моделот и за поставување и анализа со надворешни фактори, преземени се од Интернет неколку множества на податоци од јавен карактер, кои се важен економски фактор.

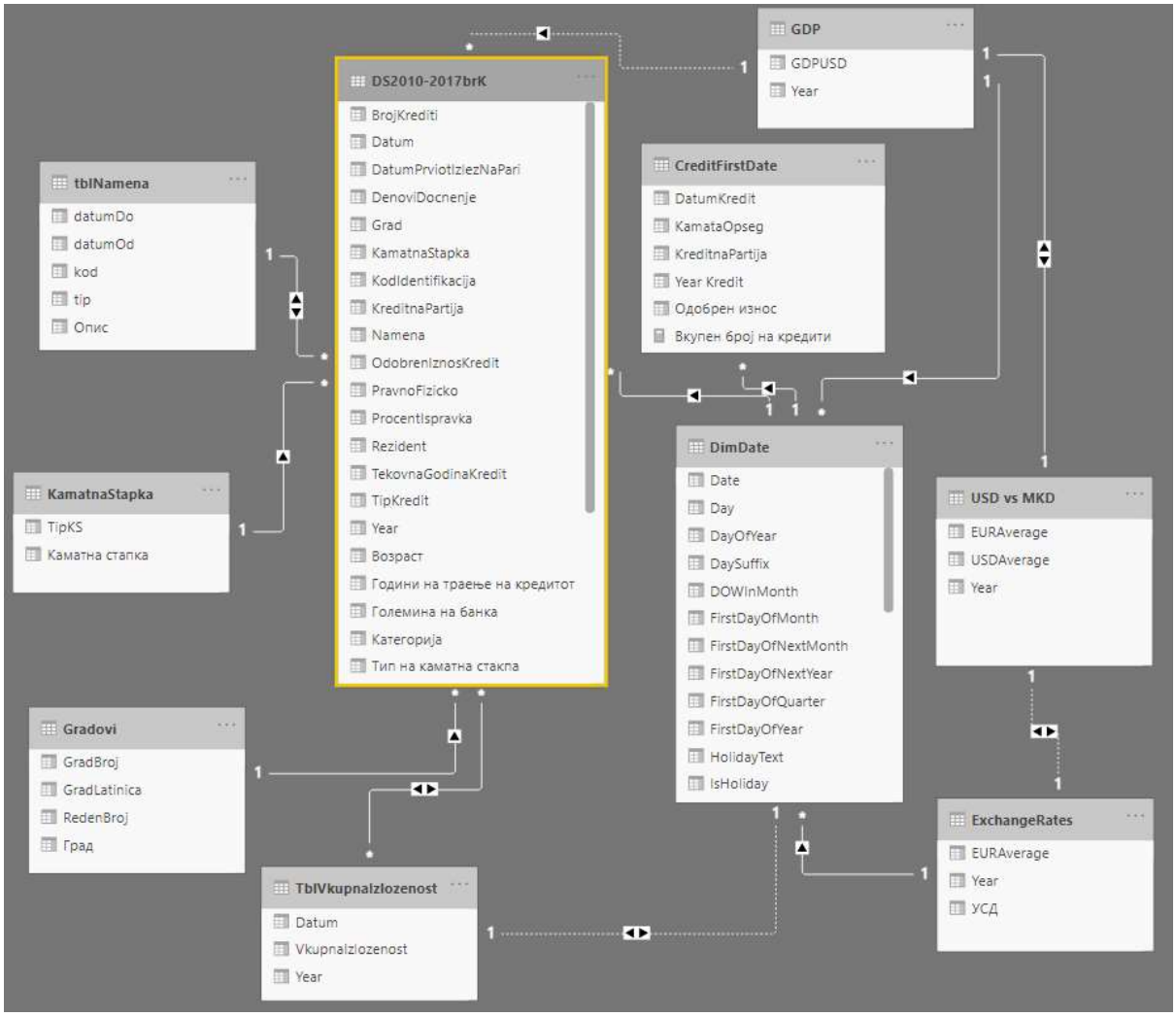
- Курсна листа на USD, EUR за периодот од 2010 година, за да се види како тоа влијае на бројот на кредити и нивната изложеност.
- Податоци за ГДП изразени во долари. Целта е да се увиди како е поврзаноста на ГДП со издадените кредити.
- Табела на општините на Република Северна Македонија која соодветно има и број на општина потребен за спојување и прикажување на името на општината.

3.2. Анализа и обработка со Power BI Desktop

Целосната анализа е направена со главната апликација од множеството односно со Power BI Desktop. Важен дел е пред процесирањето и креирањето на моделот. Во оваа алатка се направени следните активности:

- **Импортирање на податоците од SQL & Excel** – Податоците првично се импортирани во Power BI, и потоа податоците се независни од SQL и Excel изворите. При промена на податоци во изворите, во Power BI се прави само освежување на податоците и ги вчитува од почеток.
- **Креирање калкулации** - Калкулациите се изведени колони кои се дополнително пресметан износ на ниво на секој ред
- **Креирање мерки** - Мерките се износи на ниво на цела табела, не на ниво на секој ред.
- **Креирање хиерархии** - За овозможување на функционалноста drill down, се креирани хиерархии на атрибути односно нивно групирање.
- **Дизајнирање на модел и релации** - За да може сите колони да се однесуваат како да се во иста табела, затоа е креиран моделот со спојување на сите изворни табели преку релации.
- **Креирање извештаи со визуелизации** - Извештаите се креирани со помош на готовите визуелизации, потоа се конфигурирани соодветните колони кои се прикажуваат при анализата.
- **Креирање предвидување** - Power BI има вградена анализа со помош на линеарна регресија.
- **Креирање прегледи за мобилен** - За уредно прикажување на мобилни уреди, се дизајнирани и прикази за преглед од мобилни телефони.
- **Публикување на Power BI Service** - По завршување на сите анализи и изработки локално, дата сетот заедно со сите визуелизации се публикувани на онлајн верзијата на Power BI односно Power BI Service.

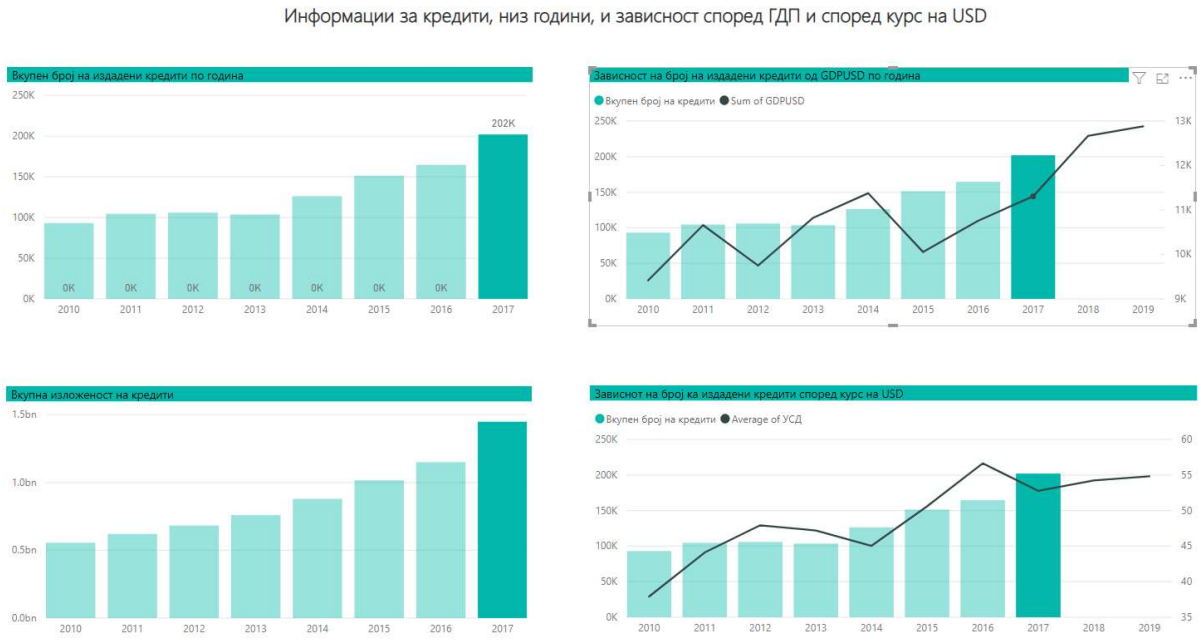
Во продолжение е креираниот модел кој е вид на ѕвезда шема и сите изворни табели се споени со главната табела за кредити (Слика 5).



Слика 5. Дизајн на моделот

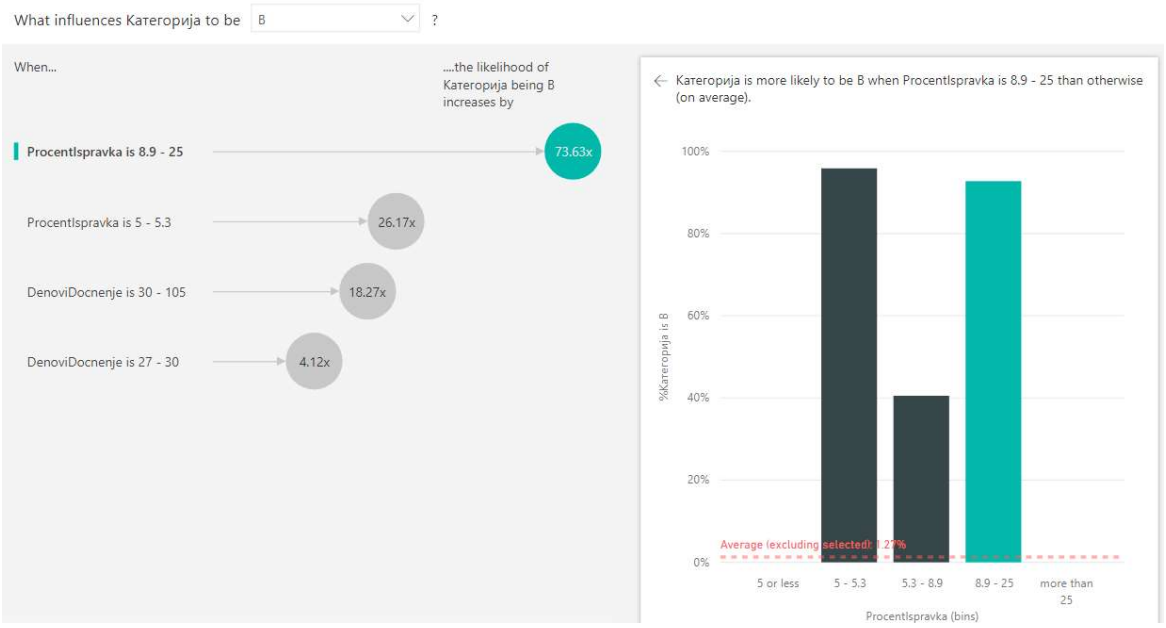
По дизајнирање на моделот и креирање на новите калкулации и мерки, следен чекор е во секоја страна (извештај) да се постават визуелизации и истите да се поднесат посебно. Значаен е фактот дека по конфигурирање на сите визуелизации посебно, истите на ниво на извештај се интерактивни и при одбирање на некоја информација истата се ажурира во сите визуелизации.

На Слика 6 се прикажани прикази за вкупниот број на кредити и вкупната изложеност, односно како тие од 2010 до 2017. Од приказот може да се увиди дека и бројот и изложеноста растат полека. Исто така на Слика 6 е анализирано дали има зависност бројот на издадени кредити со курсот на доларот и со бруто домашниот производ изразен во долари, и е увидено дека немаат никаква корелација и зависност.



Слика 6. Извештај за број на кредити, кредитна изложеност, и зависност на број на кредити според курс на долар и според ГДП

Алатка е моќна и за наоѓање на зависности помеѓу различни атрибути [29]. На Слика 7 е претставена анализа дека категоријата на ризик зависи од деновите на доцнење на исплата на рата, и според процентот на исправка кое го внесуваат самите банки. Алатката преку визуелизацијата Influencer открива со голема точност опсегот на денови во кои клиентот треба да биде за секоја категорија на ризик, а исто така со поголема точност ја открива зависноста од процентот на исправка. Слика 7 прикажува пример дека категоријата е поверојатно да биде Б кога процентот на исправка е 8,9-25 со точност од 73,63%, а исто така и кога одложеното плаќање на рата е помеѓу 30-105 дена.



Слика 7. Визуелизација на зависности на категорија преку Inlfuencer визуелизацијата

3.3. Аналитика со Power BI Service

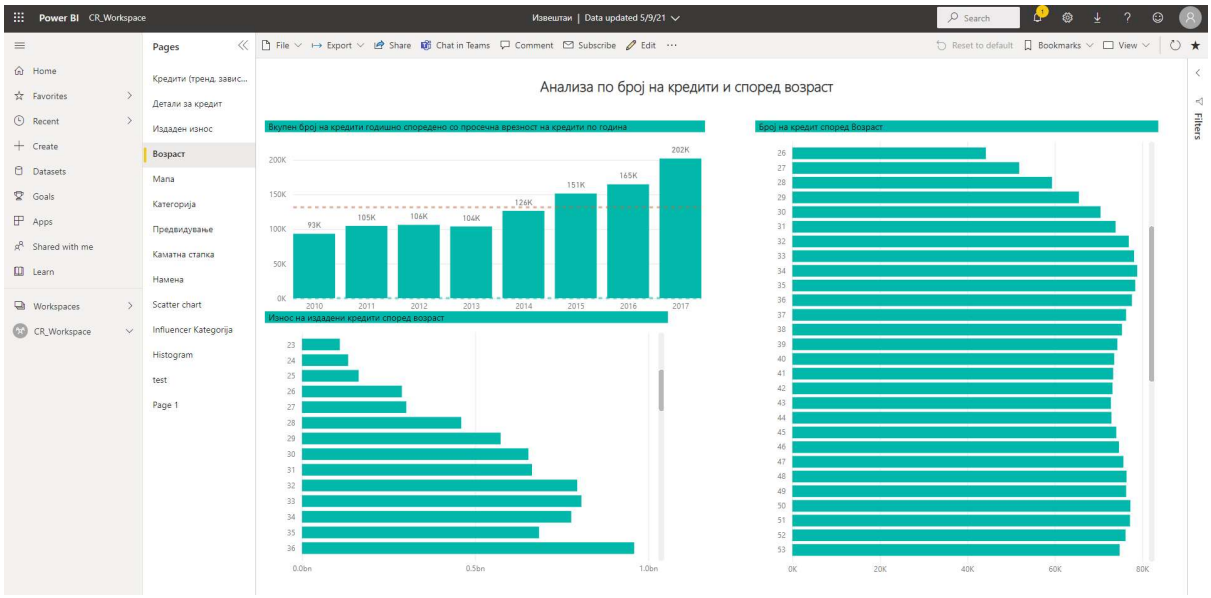
Power BI Service овозможува целосно поставување на податоците и на визуелизациите на облакот на Power BI и нивна достапност секаде и секогаш.

Во Power BI Service односно облак верзијата на оваа алатка, се извршени следните анализи:

- **Креирање приказ (Dashboard) –** Нов приказ се креира со избирање на важните визуелизации и нивно прикажување на нова страна.
- **Get insights -** Power BI Service има можност сам да пребарува да дознае нови знаења и визуелизации од податоците. Овие знаење не секогаш се од помош.
- **Ask question -** Опција која овозможува пишување на англиски прашања текстуално и се труди да прикаже резултати од податоците.
- **Публикување на веб -** Можност да се постават графиконите во надворешни страни, меѓутоа ваквиот пристап сепак ќе бара автентикација на корисниците.

Извештаите и приказите се тестирани и на мобилната апликација Power BI Mobile, во која по најавување со корисничката сметка за Power BI, автоматски се прикажуваат сите визуелизации и прикази.

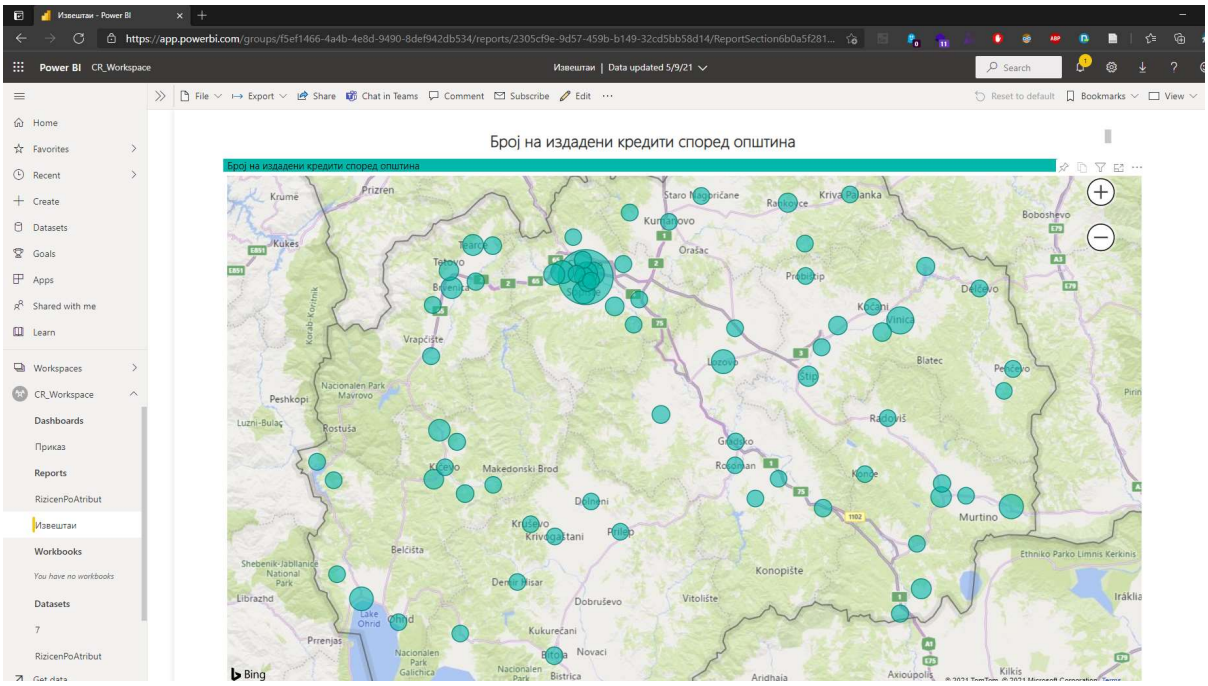
Во прилог се неколку слики од решението и нивен соодветен опис. Слика 8 прикажува анализа според возраста, бројот на кредити според возраста и распределбата на износот по возраст. Анализата на бројот на кредити според возраст покажува дека најмногу кредити имаат возрасните лица околу 34 години и околу 50 години. Износот на кредити според возраста нема некоја очигледна корелација со возраста, освен дека лицата до 30 годишна вредност имаат кредити со помал износ.



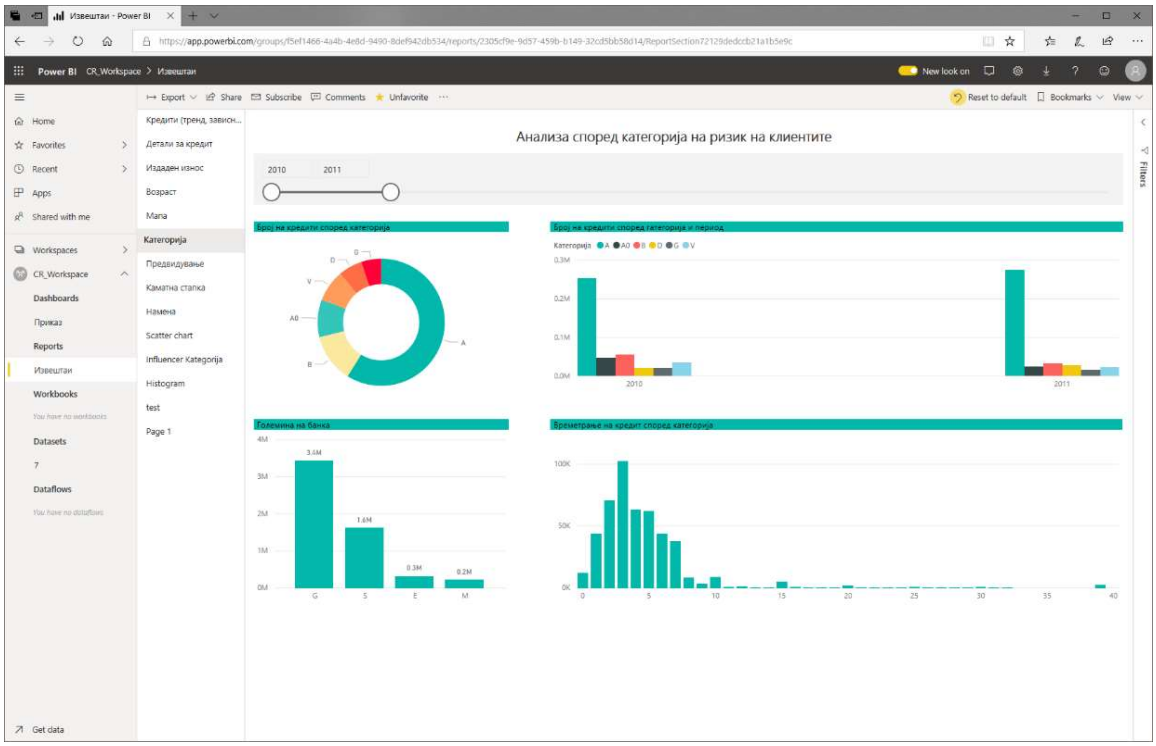
Слика 8. Визуелизација на категорија на ризик, број на кредити по категорија, по големина на банка, времетраење на кредит споерд категорија

Слика 9 ја прикажува распределбата на бројот на кредити по општини. Мапата е Bing мапа и со користење на Power BI визуелизација се прикажува бројот на кредити по

општина, односно зелените кругови со соодветна големина го претставуваат вкупниот број на кредити.



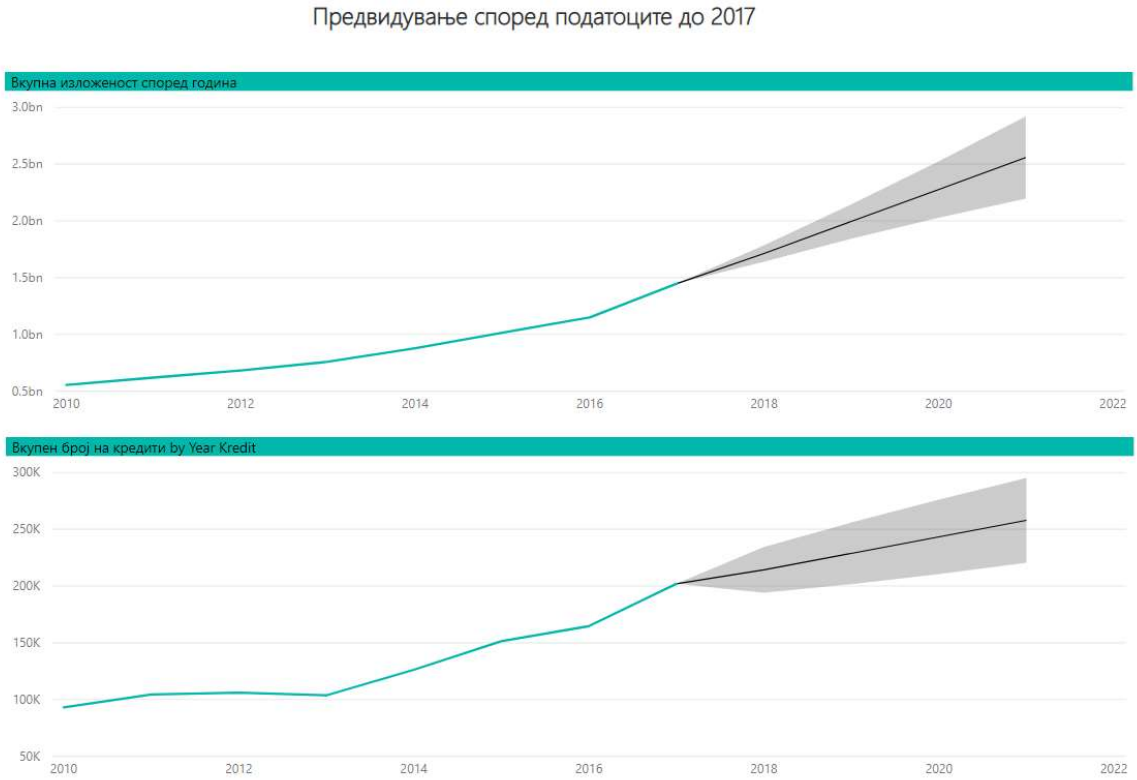
Слика 9. Мапа на број на кредити според општина



Слика 10. Визуелизација на категорија на ризик, број на кредити по категорија, по големина на банка, времетраење на кредит споерд категорија

На Слика 10 е претставен приказ во кој може да се филтрира за одреден период и за соодветниот период ја јавува распределбата по категории. Од приказот за 2010 и 2011 година над пола од клиентите се означени со најдобрата категорија А (не-ризични).

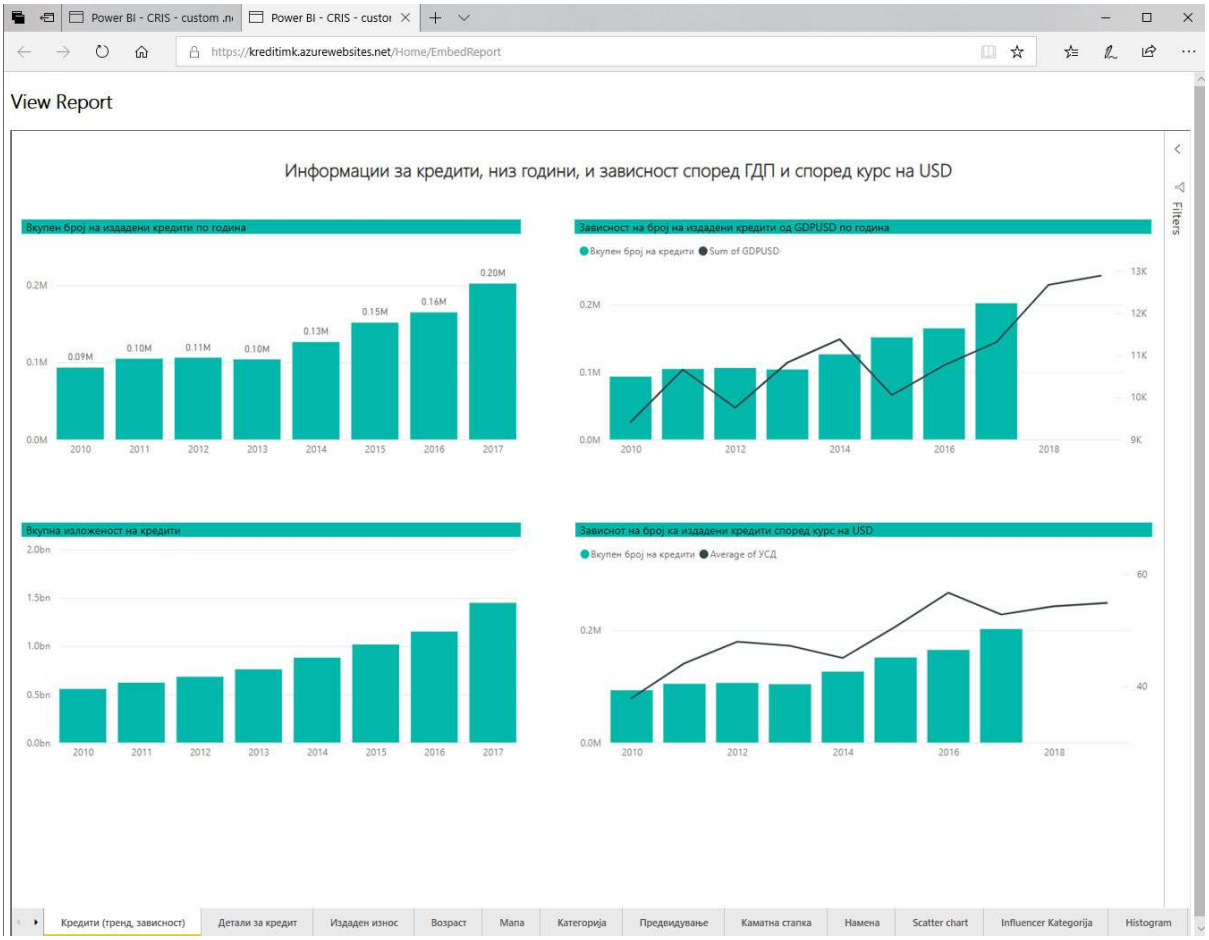
Power BI е во можност и да предвиди со користење на линеарна регресија, и тоа е направено на Слика 11. На сликата се претставени предвидување за вкупниот број на кредити и за вкупната изложеност за периодот 2018-2022 година. Според анализата се увидува дека расте вкупниот број на кредити и вкупната изложеност. Линеарната регресија е обучена со податоците од периодот 2010-2017 и потоа со мали подесувања за периодот на предвидување и за прецизноста на предвидување се добиени прикажаните резултати.



Слика 11. Визуелизација на предвидување на вкупна изложеност и на број на кредити по година

3.4. Интеграција на анализите во надворешна веб апликација

Извештаите развиени во Power BI може да се интегрираат и во надворешна апликација преку програмски код Слика 12. За да може да се вгради во надворешна апликација е потребно да се поседува Power BI Pro сметка која за автентикација користи Azure Active Directory. Апликацијата треба да се регистрира во Azure Active Directory за да може да пристапи до REST API-то на Power BI. Регистрирањето на апликацијата овозможува да се постави идентитет на апликацијата, и истата да се препознае при читање на податоците.



Слика 12. Power BI извештаи вградени во MVC 5 апликација, поставена на Azure како App Service

Со помош на Power BI е постигнато детален увид и анализа на зависности на самите атрибути и корелација со надворешни фактори. Анализата на големи податоци со помош на ова алатка е ефикасен начин за откривање на знаење и визуелизирање на многу брз начин, кое за разлика од традиционалните алатки е неспоредливо брзо и моќно, помогнато со модерни можности за приказ, зависност и предвидување. Со користење на алатката се најдени и отстранети доста нелогичности во некои од колоните. Исто така увидено се и промени на законите кои влијаеле и во износите за известување, пример за износ на кредит до 2010 година известувале со износ денари, а потоа со илјада денари.

Направениот проект овозможи целосно запознавање со големите податоци на Кредитниот регистар. Увидено е дека бројот на ново издадени кредити расте во 2010-2012 и 2014-2017, растот е отприлика линеарен. Не е најдено никаква зависност на бројот на издадени кредити со курсот на USD и со ГДП. Лицата на возраст околу 34 години и околу 50 години имаат повеќе кредити. Износот на кредитот нема корелација со возраста. Највеќе кредити се издадени во Скопје, што е и очекувано. Поголемиот број од клиентите се во најдобрата категорија на ризичност. Имплементацијата на направените извештаи во надворешни апликација е едноставна и моќна.

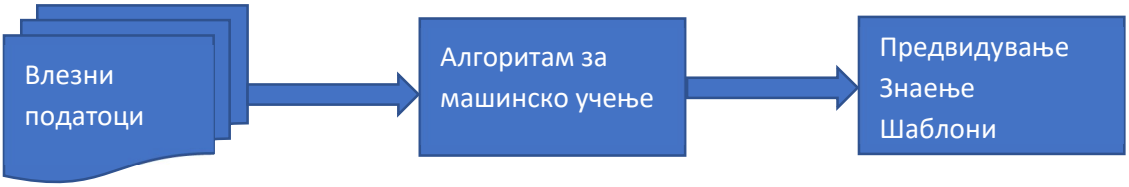
4. Модел за предвидување кредитен ризик, од Кредитниот Регистар

Науката за податоци и техниките за машинско учење им помагаат на банките да го оптимизираат работните процеси, да ги подобрат анализите на ризиците и да стекнат конкурентска предност. Постојат многу истражувања во врска со кредитниот ризик, но според мое знаење никој од нив не го користи Кредитниот регистар како извор на податоци за да моделира и предвиди кредитниот ризик. Целта на оваа глава е нов модел за кредитен ризик користејќи ги податоците од базата на податоци за реалниот Кредитен регистар на НБРСМ. Моделот кој е развиен во ова глава ќе биде дополнителен извор за оценување на кредитниот ризик.

4.1. Машинско учење

Алгоритмите за машинско учење автоматски градат математички модел користејќи податоци за обучување за да донесуваат автономни одлуки без да бидат специфично програмирани за донесување на тие одлуки. Со континуираното учење, моделите за машинско учење се подобруваат и стануваат попрецизни. Последните години, примери за револуционарни примени на машинско учење од најголемите компании се следните: Facebook користи машинско учење за автоматско тагирање на сликите, Netflix за препорака на нови филмови, Google за подобрено пребарување и филтрирање на спам мејлови, Siri и Alexa како за лична асистенција, Paypal за заштита од перење пари, итн.

Моментално оваа технологија е една од највлијателните во ИТ светот, која се подобрува континуирано и се применува во голем број на многу важни области во реалниот свет. Машинското учење се применува преку алгоритми кои користат статистика за преставување на шаблоните во податоците. Големината на множество на податоци овозможува подобро обучување на алгоритмите и тоа допринесува за поефикасно предвидување. Многу важна улога во процесот на машинско учење има квалитетот на податоците и процесот на пред обработка, селекција и трансформација на податоците. Најчесто податоците се делат во множество за тренирање (обучување) и тестирање. Перформансите на моделот се испитуваат со податоците за тестирање, и според нив се гледа колку добро е трениран моделот со податоците за тренирање. Постојат неколку вида на машинско учење: надгледувано, ненадгледувано, полунагледувано и засилено учење. Секој вид има различен пристап за учење и се користи во различна проблематика. Во дисертацијата се користи само надгледувано учење [30].



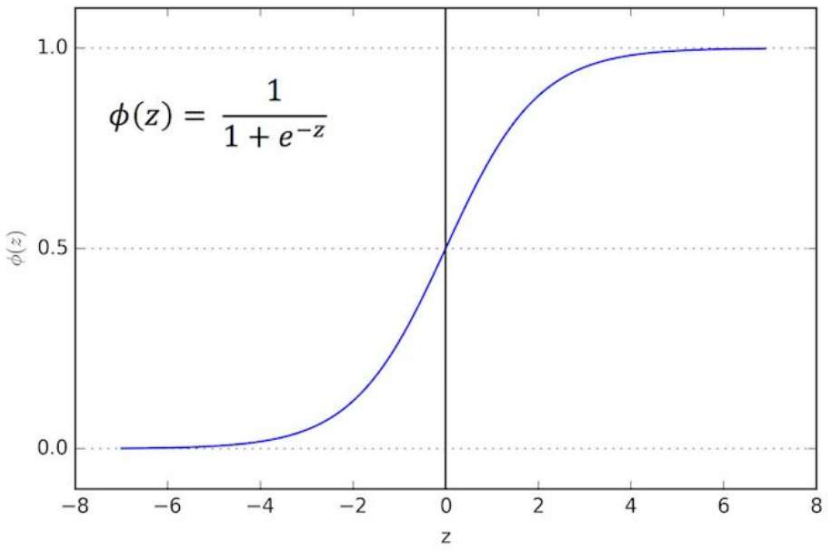
Во наредните под заглавија е даден краток опис на алгоритмите за надгледувано учење кои се користени во оваа глава.

4.1.1. Дрво на одлуки

Дрвото на одлуки (Decision Trees) [31] помага во решавањето на проблемите со класификација и регресија користејќи методи основани на дрвја. Дрвото на одлуки може да се опише како структура на проток во кое внатрешните јазли претставуваат тест атрибути, каде секоја гранка претставува исход од тестот додека секој лист претставува одлука донесена по евалуација на сите претходни атрибути. Визуелизацијата на дрвото на одлуки овозможува полесно разбирање за начинот на класифицирање, и по интуитивен приказ за самите корисници на проблематиката. Дрвото на одлуки е техника која може да се користи за да се помогне во донесувањето одлуки. Алгоритамот дрво на одлуки дефинира модел кој се креира со повторувања или рекурзија врз основа на дадените податоци. Целта на овој алгоритам е да ја предвиди вредноста на атрибутот според множеството на влезни податоци. Секоја класификација се претвора во множество на ако-тогаш правила. Податочното множество го дели на помали подмножества и паралелно се креира дрвото на одлуки кое е составено од одлучувачки јазли и јазли-листови. Одлучувачките јазли имаат две или повеќе гранки, додека јазлите-листови ја претставуваат одлуката односно класификацијата. Најпрвин, се идентификува карактеристиката којашто најточно ги разделува класите, а потоа рекурзивно се продолжува со одбирање на следна најдобра карактеристика, сè додека не се добие јасно разграничување помеѓу класите.

4.1.2. Логистичка регресија

Логистичка регресија (Logistic Regression) [32] е алгоритам кој што се заснова на концептот на веројатност за класификациските проблеми. Користи логистичка крива, S – крива (eng. Sigmoid Function) за да ги мапира предвидените вредности во веројатности (Слика 13) од 0 до 1. Сигмоидната функција е едноставна математичка функција што има карактеристична крива во облик на буквата S. Притоа одбира праг на веројатноста за класифицирање на примероците во една или друга класа. Излезната вредност на логистичката регресија е број кој ја претставува веројатноста примерокот да припаѓа на една од класите.



Слика 13. Логистичка крива

4.1.3. Случајна шума

Честа оптимизација на дрвото на одлуки е употребата на случајна шума (Random Forest) [33]. Ова претставува множества на различни дрва на одлуки.

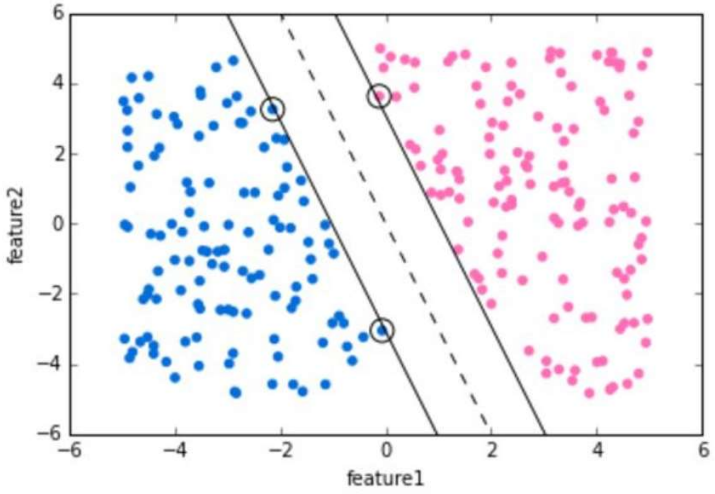
Случајната шума е составена од повеќе дрва на одлуки, притоа за секое дрво на одлуки користи различни подмножества од податочното множество за обука со што спречува и претренирање (overfitting) на моделот. Секое дрво на одлуки од шумата има свое предвидување за класата, а излезот од класификаторот случајна шума е класата со најмногу гласови од дрвата на одлуки

За примена на Случајна шума за предвидување на кредитниот ризик, во дисертацијата се користи пакетот randomForest[11] во R.

4.1.4. Векторски поддржани машини (SVM)

Векторски поддржани машини (Support Vector Machine) [34] е исто така линеарен модел кој се користи за решавање на проблеми за класификација и регресија. За проблемите со класификација, повлекува линија или рамнина во зависност од димензијата на податоците која ги раздвојува двете класи (Слика 14).

Рамнината што го прави раздвојувањето во просторот е наречено хипер рамнина (hyperplane). Бидејќи може да постојат повеќе хипер рамнини коишто ги раздвојуваат класите, SVM ја наоѓа најоптималната. Се наоѓаат примероците кои што се најблиски до линиите од двете класи, овие линии се наречени помошни вектори (support vectors), растојанието помеѓу нив е наречено маргина, која треба да биде оптимизирана (Слика 14). Доколку податоците не може да се разграничат линеарно, со линија или со рамнина доколку се тродимензионални, SVM функционира на тој начин што ја зголемува димензијата на податоците во простор во кој што ќе може да повлече хипер рамнина и да се направи раздвојување на класите.



Слика 14. Векторски поддржани машини

[11] https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

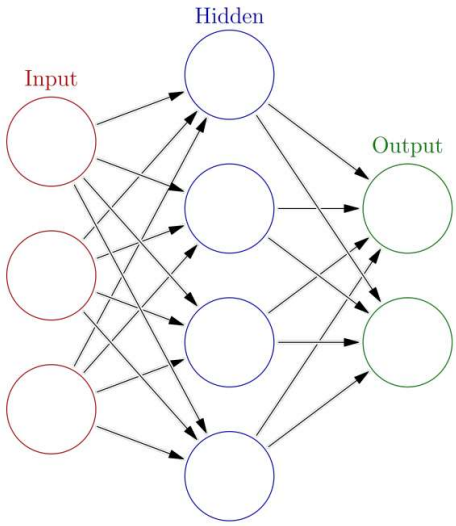
4.1.5. Невронски мрежи

Невронските мрежи (Neural network) [35] се моделирани според биолошките невронски мрежи и целта е да овозможат компјутерите да учат на сличен начин како човекот. Невронските мрежи се организирани во слоеви. Слоевите се составени од многу меѓусебно поврзани јазли кои содржат активациска функција. Невронските мрежи може да состои од следните 3 слоеви (Слика 15):

- Влезен слој
- Скриен слој
- Излезен слој

Обрасците се претставуваат на невронската мрежа преку влезниот слој, кој комуницира си еден или повеќе скриени слоеви, каде што вистинската обработка се прави преку систем на длабоки врски. Скриените слоеви се поврзани со излезниот слој каде се добива одговорот од моделот.

Постојат неколку видови на невронски мрежи, меѓутоа во оваа дисертација е користено ациклична невронска мрежа за предвидување на кредитниот ризик, рекурентна (опишано во Глава 6.1.1) и конволуциска (опишано во Глава 6.1.2) невронска мрежа за процесирање на природни јазици. Ацикличната невронска мрежа се користи за класификација и влезните податоци патуваат само во една насока (напред).



Слика 15. Невронска мрежа

4.2. Поврзана работа

Постојат многу истражувачки трудови поврзани со кредитниот ризик користејќи различни алгоритми за машинско учење. Во следниот текст, се анализирани некои од најважните трудови подредени според годината на публикација.

Во [36] авторите го анализирале кредитниот ризик во комерцијалните банки користејќи векторски поддржани машини (SVM), и експериментите покажале дека

бинарниот модел има висока точност на класификација, додека во [37] авторите препорачуваат хибриден SVM базиран модел на кредитно оценување, кој ги пребарува оптималните параметри на моделот и подмножеството на атрибути за да ја подобри точноста на кредитното оценување.

Во [38] авторите споредиле седум методи за избор на атрибути за кредитно оценување, применети на австралискиот и германскиот јавни множества на податоци и нагласува дека дрвото на одлуки за класификација и регресија (CART) (classification and regression tree) е метод за одбирање на атрибути со поголема вкупна точност. CART може да го исече дрвото и да го намали времето на извршување, притоа задржувајќи го оптималното предвидување. Слична работа е направена и во [39], каде што авторите го анализирале кредитниот ризик на небалансирани податоци и откриле дека дрвјата за логистичка регресија, класификација и регресија (CART) и случајните шуми даваат добри резултати над небалансираните податоците за кредитен ризик.

Авторите во [40] за проценка на кредитниот ризик споредиле повеќе алгоритми и увиделе дека SVM, дрвото на одлуки и логистичка регресија се најдобри модели за предвидување за класифицирање на апликантите за кредит.

Детално истражување е направено во [41], каде што авторите примениле петнаесет алгоритми за машинско учење за бинарна класификација и откриле дека сите алгоритми генерираат резултати со точност од 76 до над 80%. Тие исто така откриле дека дури и со три атрибути од вкупно 23, нема значителна разлика во нивната точност за предвидување и во другите мерки.

Наивен баесовиот класификатор, невронската мрежа и дрвото на одлуки се користени во [42] за предвидување на кредитниот ризик. Резултатите од оваа истражување покажаа дека дрвото на одлуки е најдобриот алгоритам основано според точноста.

Во [43] по споредување на неколку алгоритми за предвидување како што се дрво на одлуки, SVM, логистичка регресија, случајна шума и невронска мрежа, откриле дека најдобриот алгоритам за класификација на кредитен ризик е алгоритамот за случајна шума. Тие исто така покажаа дека најголемо влијание имаат следните атрибути: возраста, времетраењето и износот на кредитот. Од друга страна, во [44], авторите испитале дваесет и пет алгоритми за бинарна класификација за предвидување на кредитниот ризик и откриле дека невронските мрежи попрецизно ја извршуваат класификацијата на клиентите.

Во [45], авторите предложија модел со високи перформанси за оценување на кредитот наречен NCSM, основан на одбирање на атрибути и мрежа на пребарување (grid search) за да се оптимизира алгоритамот на случајна шума. Овој модел во споредба со другите линеарни модели покажа подобри перформанси во однос на точноста на предвидувањата како резултат на намалувањето на влијанието на ирелевантните атрибути. Во [46], авторите го предвидоа кредитниот ризик користејќи линеарна регресија, SVM и невронски мрежи.

Нивната истражувачка работа ги споредува показателите за успешноста на методите за предвидување пред и по балансирање на податоците. Нивните резултати покажуваат дека имплементацијата на стратегиите за земање примероци (како што е техниката на синтетичка техника која користи прекумерни семплирање на класата која е помалку застапена (SMOTE) ги подобрува перформансите на моделите за предвидување во споредба со небалансираните податоци. Во мојата дисертација, ја земам во предвид стратегијата за семплирање SMOTE при проценка на моделите.

Сеопфатна анализа за кредитно оценување е направена во истражувачки труд [12] преку анализа на 258 трудови за кредитно оценување. Во трудот се сумира дека повеќето студии спроведуваат само еден статистички метод во периодот по 2010 година, а потоа следат студии кои спроведуваат повеќе статистички методи за истата база на податоци. Откриено е дека најкористената техника е логистичката регресија.

Авторите во [47] предлагаат нов модел на класификација основан на невронските мрежи и техниките за оптимизација на класификаторот за небалансирано оценување на кредитниот ризик. Нивниот предложен модел постигнува поголема вкупна точност во споредба со седум широко користени модели. Експериментите се направени врз германската и австралиската база на податоци. Слично на тоа, во [48]–[50] се фокусираат на хибридни модели со комбинирање на постоечки избор на одлуки и ансамбл класификатори за да се подобри предвидувањето на кредитното оценување. Експериментите се потврдени врз множества на податоци што се користат во комерцијалните банки.

Еден неодамнешен истражувачки труд [51] го моделира кредитниот ризик со употреба на логистичка регресија, дрвја на одлуки и случајна шума. Откри дека логистичката регресија и случајната шумата имаат подобри резултати и ги имаат истите вредности за точност, чувствителност и специфичност. Слични резултати се презентирани во [52] каде што е спроведена компаративна проценка на успешноста на моделите за кредитно оценување со користење наивен баесов класификатор, анализа на логистичка регресија, случајна шума, дрво на одлуки и класификаторот К-најблизок сосед. Резултатите покажуваат дека случајната шума работи подобро од другите во однос на прецизноста, повторното повикување, AUC и точноста.

Скоро сите споменати трудови и истражувања експериментираат со јавни мали бази на податоци или со бази на податоци на комерцијални банки, но никој од нив не се обидел да го моделира кредитниот ризик користејќи ја базата на податоци на Кредитниот регистар на која било централна банка. Затоа во оваа истражување е користено базата на податоци на Кредитниот регистар и се презентирани сите потребни чекори од собирање на податоци до предвидување и проценка. На оваа база на податоци, обучуваме модели со употреба и споредување на најчесто користените алгоритми за машинско учење. Дополнително е разгледано синтетичка техника која користи прекумерни семплирање на класата која е помалку застапена, како што е пристапот во [46]. Експериментите и анализите со оваа база на податоци и претставениот модел обезбедуваат дополнителен придонес во полето на проценка на кредитниот ризик. Главниот недостаток е тоа што добиените резултати не се споредуваат со резултатите со

друга слична база на податоци, бидејќи е невозможно да се добијат такви податоци од соседните или друга централна банка.

4.3. Методологија

Проценката на кредитниот ризик е многу важна и клучна мерка за разликување на ризични и не-ризични клиенти, односно бинарна класификација. Кредитен ризик е атрибут на класификација, кој ги класифицира клиентите со цел пред време да се предвидуваат ризичните клиенти. Во ова истражување се предвидува дали клиентот е ризичен односно дали ќе го исплати кредитот во планираниот рок. За ваков тип на сериозни клиенти ризичноста е блиску до 0, додека атрибутот ризичен клиент е еднаков на 1 кога клиентот е ризичен односно може да ја одложи исплатата или во некои ситуации и нема целосно да го исплати кредитот.

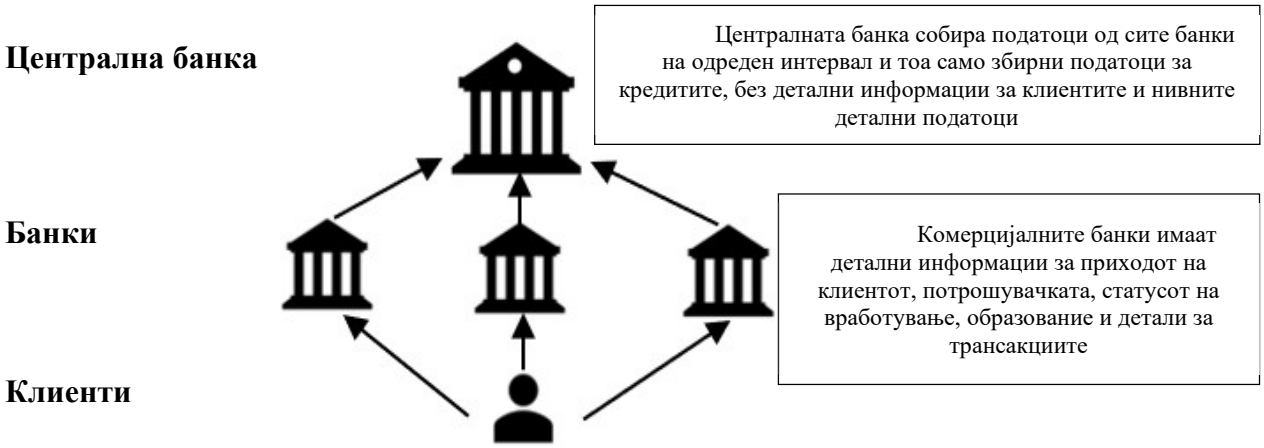
На Слика 16 се претставени чекорите на методологијата што се користи во ова истражување, каде што секој чекор е објаснет во следните потточки.

Слика 16. Чекори на применетата методологија

4.4. Прибирање на податоци

Во оваа дисертација се користи базата на податоци на Кредитниот регистар на Република Северна Македонија, кој се состои од неколку милијарди записи, што ја прави најголем база според големината и според бројот на записи во централната банка.

Базата на податоци се состои од 52 финансиски и нефинансиски атрибути [53], а сите приватни полиња се анонимизирани поради приватноста на податоците и законот за заштита на личните податоци. Базата на податоци е централна точка за сите кредити во државата и собира податоци од сите други комерцијални банки и штедилници (Слика 17). Во оваа база на податоци, секој запис го претставува месечниот статус за секоја кредитна партија.



Слика 17. Процес и архитектура на прибирање на информациите за кредити. Оваа слика покажува дека комерцијалните банки имаат повеќе детали за клиентите. Сликата исто така покажува дека централната банка собира агрегирани податоци од сите комерцијални банки во ред

Оваа база на податоци, како најголемата база на податоци во централната банка (НБРСМ) ги исполнува повеќето карактеристики на големите податоци, како што се обемот и брзината. Постојат многу контроли на валидација при внесување на податоците, а квалитетот на податоците е добро контролиран пред и при нивниот внес.

Базата на податоци Кредитниот регистар ги има следните информации кои се доставуваат од банките:

- Тип на клиент (правно лице, физичко лице, домаќинства и сл.)
- Идентификација на клиентот (матичен број, даночен број и активност доколку е правно лице, седиште и сл.)
- Изложеност од страна на кредитна страна (износ, структура, податоци за одобрување, одложени денови, редовна камата, ниво на каматна стапка, вид на каматна стапка, намена, итн.)
- исплата на обврските
- Други податоци и информации во врска со видот на обезбедувањето, видот на исправката на вредноста, намената и карактеристиките на кредитната изложеност и / или клиентот
- Отпишани побарувања

Карактеристично за оваа база на податоци е фактот дека нема информации за приходите, трошењето, навиките за купување, деталите на социјалните мрежи. Во базата има само општи податоци за клиентите и статусот на нивните кредити во сите банки во државата. Атрибутот категорија се користи за означување на кредитен ризик за клиентот и го класифицира секој клиент во една од петте претходно дефинирани категории, кои се означени како А, Б, Ц, Д, В. Категоријата А е најдобрата, и тоа значи клиент со најмал ризик, а секоја следна категорија е полоша категорија, при што категоријата В е најлошата [54]. Одлуката за категоријата на клиентот ја носат службените лица на комерцијалните банка, а во централната банка податоците се собираат и потоа им се на располагање на сите банки кои учествуваа во Кредитниот регистар. Во Табела 1, претставена е статистиката на дистрибуција податоците во нашето податочно множество односно на клиенти по категорија.

Табела 1. Дистрибуција на клиентите според категорија

Категорија	Записи
А	898279
Б	35599
Ц	36581
Д	11511
В	18030

Во истражувањето е користено под множество на оваа база на податоци што содржи 1.000.000 редови, односно претставува статус на 1.000.000 различни кредити само за физички (приватни) клиенти и нивниот статус (веројатност за неисполнување на обврските) во планираниот датум на целосната исплата на кредитот. Податочното множество нема атрибут дали клиентот е во можност да го отплати кредитот (ризичен) во планираниот временски рок или не. Овој атрибут го изведовме од атрибутот за категорија на клиентот во целата база на податоци и ја најдовме нивната најлоша категорија во која беа класифицирани повеќе од 20% од времето. Откако е добиено најлошата категорија за клиентот, ги поделивме во две категории во согласност со дефиницијата на Базелскиот договор [55] и одлуката на Народната Банка на Република на Северна Македонија [54] според кои А и Б се не-ризични категории додека останатите категории претставуваат ризична категорија односно клиент.

Не-ризичните клиенти се означени со вредност 0 за кредитен ризик што значи дека тие успешно ќе ги извршат сите плаќања, а другите ризични клиенти се означени со вредност 1 за кредитен ризик и претставува неисполнување на обврските за да не ги извршат потребните плаќања на кредитот.

Пред да се направи оваа поделба, на почеток истите експерименти кои се во продолжение се направени за повеќе класно предвидување односно 5 класи (категории). Од резултатите е увидено дека се добива многу лошо предвидување кои во реалност не се воопшто корисни. Максималниот F1 резултат беше 0.485 постигнато со примена на дрво на одлуки. Сето ова се случува бидејќи времетраењето на кредитите изразено во месеци е најчесто двоцифрен број и низ овој период често комерцијалните банки им ги менуваат категориите според нивните внатрешни правила, што во реалност го буни

целосно алгоритмите за машинско учење и не постигнуваме никаква придобивка. Исто така сите наведени трудови во Глава 4.2 користат само бинарна класификација.

4.5. Подготовка и пред-обработка на податоците

За подобро разбирање на базата на податоци и за визуелизација на податоците е користено алатката Power Business Intelligence (Power BI) како што е опишано во Глава 3. Користејќи ги можностите за напредна аналитика, визуелизирани се атрибутите, нивни зависности, трендовите и зависност со дополнителни извори на податоци. Анализата е извршена во повеќе фази, вклучувајќи и дополнителни пресметани колони, мерки и модел на ѕвездена шема. По дизајнот на моделот и односите, се направени повеќе извештаи на многу ефикасен начин, што помогна за прегледување на базата на податоци и подобар увид во податоците.

Подготовката и пред обработката на податоците е неизбежен чекор за добивање на квалитетен резултат од примената на алгоритмите за откривање на знаење. Повеќето од следните чекори се направени во повеќе циклуси за да се постигнат посакуваните резултати. Во продолжение се опишани чекорите во процесот на пред-процесирање.

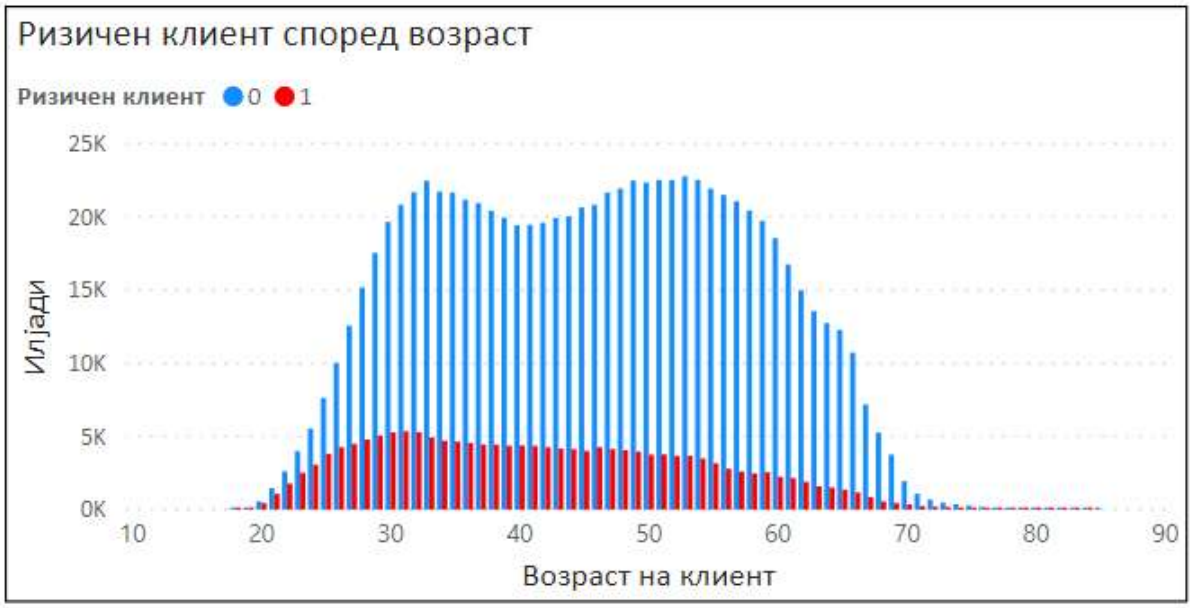
4.5.1. Подготовка на податоците со SQL скрипти

Бидејќи податоците кои се користени се зачувани во SQL Server, затоа првично се користени Transact-SQL (TSQL) скрипти за создавање под-множества на оригиналната база на податоци, каде е намален бројот на атрибути со отстранување на непотребни приватни информации, особено на текстуални податоци (телефон, адреса, итн.) и некои агрегирани нумерички вредности. За секоја колона, и покрај спроведените контроли проверени се минимални/максимални вредности за нумеричките податоци додека за не-нумерички колони проверено е должината на содржината. Изведени се и нови колони од постоечките податоци, како што се возраста од матичниот број и вкупниот број на успешно исплатени кредити. По операциите во SQL, базата на податоци е подготвена за следната фаза односно визуелна анализа преку R Studio [2].

4.5.2. Опис на податоците и визуелно истражување

Првичното запознавање и разбирање на базата на податоци е направено со Power BI. За подмножеството за истражување односно со 1.000.000 записи, направено е детално истражување со постојните пакети на R (статистички програмски јазик), што помогна за анализира на дистрибуциите на променливите, постојната корелација помеѓу променливите и исфрлените вредности. Користејќи прикази (plots), хистограми и прикази за корелации (box plots) помеѓу променливите, се увиде дека нема исфрлени екстремни вредности. Според Слика 18, Слика 19, Слика 20 првичните визуелизации обезбедуваат подобар увид за дистрибуциите и ризичните клиенти, што помогна во разбирањето на податоците за понатамошна анализа.

Слика 18 покажува ризик според возраста (претставена на X-оската како години) и според времетраењето на кредитот. Сликата покажува дека по ризичните клиенти (со црвена боја) се на возраст околу 32 години и дека бројот на кредити е поголем кај околу 32-та година и околу 55-та година (со сина боја).



Слика 18. Ризик според возраст

Слика 19 го покажува ризикот според времетраењето на кредитот. Слика покажува споредба на вкупниот број на клиенти (со ознака 0 – сина боја) и ризичните клиенти (со ознака 1 – црвена боја) во зависност од времетраењето на кредитот во години.

Слика 19. Ризик според времетраењето на кредитот

Слика 20 ја покажува распределбата на ризичните клиенти (со ознака 1 – црвена боја) според намената на кредитот. Како најзастапена намена е 1802 односно потрошувачки кредити и 1801 кредити за купување и реновирање станбен простор. Опис за останатите кодови на намена е претставен во [53].

Слика 20. Ризични клиенти според намена на кредит.

Според Табела 2, прилагодливата каматна стапка (Р) е претставена повеќе во базата на податоци, потоа се појавува фиксната каматна стапка (Ф) која е непроменлива и последната е променливата каматна стапка (V) која зависи и се менува од движењата на одредена референтна каматна стапка. Што се однесува до типот на кредит, ануитетски (А) кредити се повеќе застапени во базата на податоци отколку (Е) еднократните кредити.

Табела 2. Дистрибуција по тип на каматна стапка и тип на кредит

Ризичен клиент по		**0 – Не ризичен**	**1 - Ризичен**
Тип на каматна стапка	Р - Прилагодлива	763293	161142
	F – Фиксна	55994	3972
	V – Променлива	12698	2901
Тип на кредит	А – Ануитетски	820834	162581
	Е - Еднократен	11151	5434

4.5.3. Трансформација на податоци и креирање фактори

За категоричните колони со конечно множество на вредности создадени се фактори за претставување на категорични податоци. Како фактори, се променливите: големина на банка (мала, средна, голема), тип на кредит (ануитет и поединечни вратени заеми), каматна стапка (серија од опсези: 1, 2, 3, 4), тип на каматна стапка (прилагодлива, фиксна и променлива каматна стапка), намена, возраст и ризичен клиент (зависната излезна променлива со вредност 0 или 1). Нумеричките колони се: број на кредити, времетраење во години, тековна година на кредитот, денови на доцнење на плаќање и успешно исплатени кредити.

4.5.4. Вредности кои недостасуваат и екстремни вредности (outliers)

Вредностите кои недостасуваа во нумеричките колони ги заменивме со нула вредности. Поради промените во законските регулативи на Кредитниот регистар низ

времето, има некои колони што се воведуваат подоцна, а оние вредности што недостасуваат се поставени на нула. Поради проверките и логичките контроли од двете страни на централните и комерцијалните банки, немаше дупликати. По идентификување на исфрлените вредности на нумеричките атрибути, истите се избришани. Податочното множество на податоци по оваа фаза нема податоци што недостасуваат.

4.6. Одбирање на атрибути

Со цел побрза обработка, полесна имплементација и посигурен модел, применето е и одбирање на најважните атрибути [56]. За оваа цел е користено анализа на информационата вредност и корелациона анализа за избирање на најважните атрибути. Податочно множество првично имаше 52 атрибути [53], потоа со техниката за информациона вредност на атрибути сведено е на 11 колони, како што е опишано во Глава 4.6.3, а потоа на шест колони како што е опишано во Глава 4.6.4.

4.6.1. Инженерство на атрибути (Feature Engineering)

За да ја подобриме ефикасноста на моделот, со инженерство на атрибути, го дополнивме нашето множество со следниве колони:

- **Големина на банка** - изведена категоричка колона според кодот на големината на банката.
- **Број на кредити** - за секој клиент најдовме број на активни кредити во тековниот период.
- **Број на успешно исплатени кредити** - претставува број на успешно платени кредити во историјата на клиентот.
- **Времетраење на кредитот** - изведена колона за должината на времетраењето во години.
- **Возраст** - возраста е добиена само за физички лица преку нивниот матичен број.

На Табела 3, претставуваме преглед статистика за изведените колони користејќи инженерство на атрибути.

Табела 3. Преглед статистика

Големина на банка	Број на кредити	Времетраење на кредитот	Број на успешно исплатени кредити
E: 49636	Min. :1.000	Min. :1.0	Min. : 0.00000
G: 669363	1st Qu.: 1.000	1st Qu.: 4.0	1st Qu.: 0.00000
M: 42070	Median: 1.000	Median: 5.0	Median: 0.00000
S: 238931	Mean: 1.238	Mean: 5.7	Mean: 0.08596
	3rd Qu.: 1.000	3rd Qu.: 7.0	3rd Qu.: 0.00000
	Max. : 11.000	Max. : 31.000	Max. : 5.00000

4.6.2. Дискретизација и скалирање на атрибути

Поради небалансираната природа на базата на податоци и за да ја зголемиме точноста на предвидувањето, спроведовме дискретизација како што е прикажано и на Слика 21. Како што е направено во [57] за кредитно оценување, исто и тука е применето квантилен (quantile) метод за дискретизација на возраста во 20 бинови (серија од опсези)

и четири бинови за каматната стапка. Исто така пробан е и оптималниот метод за бинирање на возраста во четири бинови, меѓутоа квантилната дискретизација обезбеди подобри резултати.

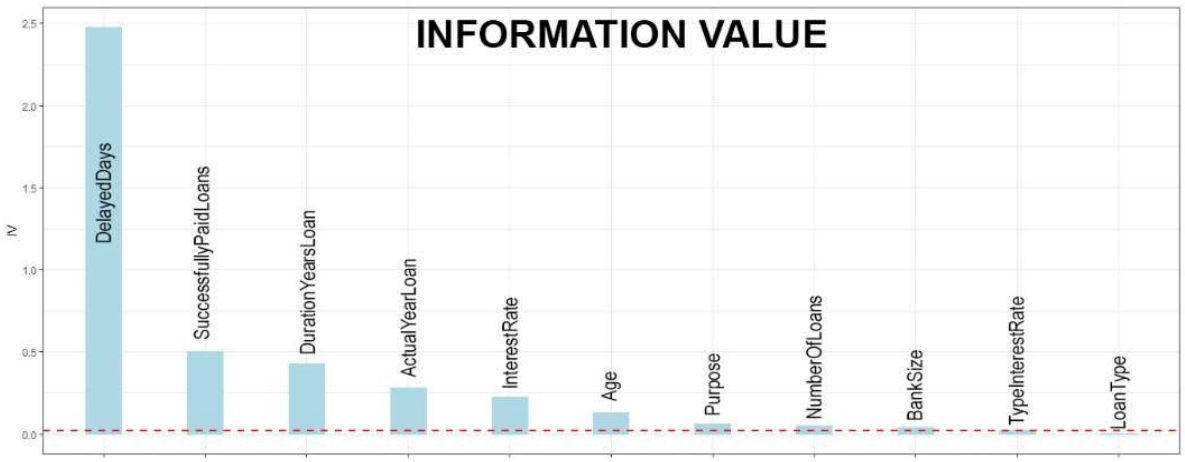
Слика 21. Бинирање (серија од опсези) на атрибутот Возраст

Атрибутите со континуирани вредности, имаа вредности во различен опсег. Со цел подобра визуелизација и влијание за точноста на моделот, колоните со континуирани вредности се скалирани во заедничка скала.

4.6.3. Информациона вредност (Information Value)

За да се увиди моќта на предвидување на секој од атрибутите во однос на зависната променлива, користено е информациона вредност, која е доста применета во проблемите со кредитно оценување. После неколку повторувања (циклуси), се увидени и отстранети атрибутите кои имаат предвидлива моќ помала од 0,02 [56].

Овие резултати исто така се потврдени и со користење на случајна шума за важните променливи. Резултатите на Слика 22 ја покажуваат анализата на моќта на предвидување за релевантните атрибути, што покажува идентично подредување како анализата за информациона вредност. Најважниот атрибут е бројот на денови на доцнење, проследен со бројот на успешно платени кредити и времетраењето на кредитот во години.

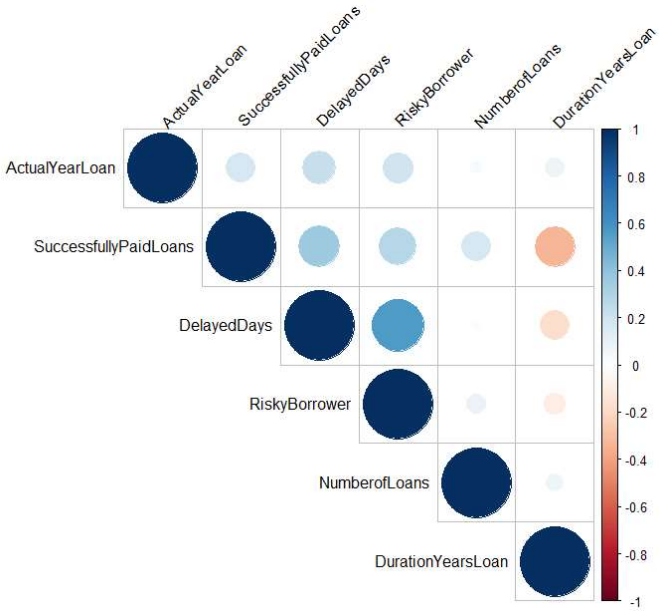


Слика 22. Информациона вредност

4.6.4. Корелациона анализа на атрибутите

За да се избегнат непотребни или нерелевантни атрибути, кои би влијаеле негативно на перформансите на моделот, користено е матрица за корелација користејќи кругови со соодветни бои и големина кои ја претставуваат корелацијата меѓу соодветните атрибути. Корелациите ги увидов со користење на тестот за корелација на Pearson. Оваа корелација има вредности помеѓу −1 и 1, каде што r = 1 или r = −1 претставува совршена линеарна корелација, додека r = 0 не претставува корелација помеѓу атрибутите.

Според Слика 23, од 11-те колони во поглавје 3.3.3, поголема корелација се наоѓа со деновите на задоцнета исплата отколку со бројот на успешно платени кредити и со тековната година на постоечкиот кредит. Позитивните корелации помогнаа за подобро предвидување на ризични клиенти, а корелациите со резултат помал од 0,2 не се користени во моделот на податоци.



Слика 23. Корелациона анализа на атрибути

4.7. Резултати и дискусија

За да се применат алгоритмите на моделите за машинско учење и да се анализираат резултатите, податочното множество е поделено на тренинг множество и тест множество. Поделбата на податочно множество во овој случај е направена во сооднос 4:1 односно 80% од податоците како множество за учење и 20% од податоците како множество за тестирање.

По првичната поделба и визуелизирање на дистрибуцијата на податоци, откриено е дека и двете множества се небалансирани.

За фазата на учење од податоците се користени пет најчесто користени алгоритми за кредитен ризик односно: логистичка регресија која во реалност е параметарски статистички модел, дрво на одлуки, случајна шума, SVM и невронска мрежа. За проверка на ефективноста на алгоритмите е применето 10-пати вкрстена валидација (10-fold cross validation) за да се провери стабилноста на моделот со различни множества на податоци.

Небалансираноста на множество на податоци влијае на алгоритмите за машинско учење, поради игнорирање на малцинската класа. Ова е многу битно во овој случај бидејќи точното предвидување на малцинската класа е од голема важност, односно одредување на ризичните клиенти. Најчест пристап за справување со проблемот на небалансираност кај податочните множества е ребалансирање на податочното множество што го тренираме. Односно ова подразбира креирање на нова трансформирана верзија на тренинг множеството во кое ќе се балансира присуството на двете класи. За да се надминеме проблемот со небалансираната база на податоци кој може да доведе до негативни ефекти со перформансите, применето е синтетичка техника. Оваа техника користи прекумерни семплирање на класата која е помалку застапена – SMOTE [47], [58] за да се провери дали ќе влијае за подобри резултати.

SMOTE вештачки го зголемува бројот на класата која е помалку застапена. Ова ќе помогне да се надмине ситуацијата кога мнозинската класа би ги влијаела резултатите. По многу повторувања и калибрации на функцијата SMOTE, достигнавме прилично балансирано множество за тестирање како што не е прикажано во Табела 4. Функцијата SMOTE вештачки генерира примероци од малцинската класа и ја балансира застапеноста на двете класи.

Табела 4. Дистрибуција пред балансирање и по балансирање

Пред балансирање		**После балансирање**	
Ризични клиенти -1	168015	Ризични клиенти -1	478050
Не-ризични клиенти - 0	831985	Не-ризични клиенти - 0	466050

Бидејќи во базата на податоци има и нумерички и категорични атрибути, затоа користени се различни комбинации со и без скалирање/SMOTE на влезните податоци за да се увиди дали ќе тоа влијае на резултатите за предвидување.

Со наведените комбинации ги имаме следните четири множества на податоци на кои се применети различни алгоритми на машинско учење: небалансирани податоци без

скалирање, небалансирани податоци со скалирање, балансирани податоци со SMOTE без скалирање, балансирани податоци со SMOTE со скалирање. Небалансираните податоци покажуваат подобри резултати од балансираните, бидејќи SMOTE генерира вештачки редови, што во оваа база на податоци, не помогнаа за добивање подобри резултати.

Со цел да се испита најдобриот модел на базата на податоци, применети се пет алгоритми за машинско, користејќи го програмскиот јазик R на четирите наведени множества на податоци. За невронски мрежи користено е пакетот R nnet[12] со 20 скриени слоеви, пропаѓање на тежините (weight decay) (регулирање за да се избегне прекумерно вклопување) 0,001 и 20 повторувања. Како алатка е користено RStudio Desktop[13], Open Source Edition.

Слика 24 ја претставува кривата ROC [4] за сите модели обучени на не/балансирано податочно множество со и без скалирање. Во Табела 5 се прикажани резултатите за точноста (accuracy) (1), прецизност (2), отповик (recall) (3) и F1 резултат (4) (функција на прецизност и отповик) за пет модели за истото податочно множество.

$$\boldsymbol{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$\boldsymbol{Precision} = \frac{TP}{TP + FP} \quad (2)$$

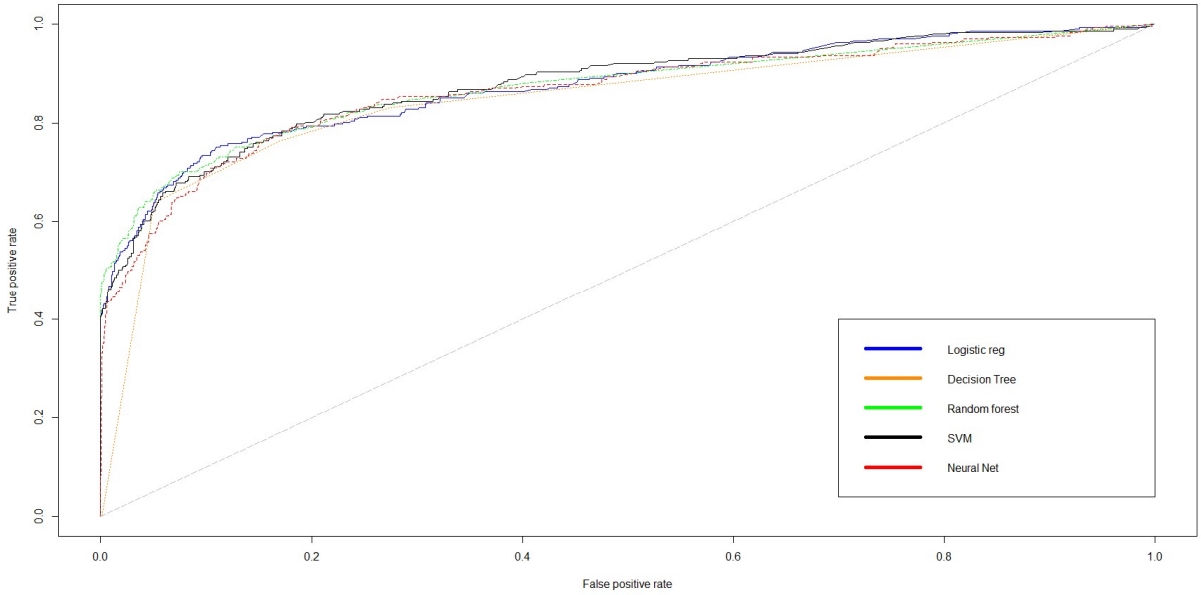
$$\boldsymbol{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\boldsymbol{F1\_Score} = \frac{2TP}{2TP + FP + FN} \quad (4)$$

Прикажаните резултати во Табела 5 покажуваат најдобри предвидувања за небалансираните податочни множества, каде има повеќе примероци од класа 0 (не-ризични клиенти) и помалку од класа 1 (ризични клиенти). Кредитниот регистар има висока небалансираност на податоците бидејќи нормално ризичните клиенти ги има помалку.

Класификацијата на кредитниот ризик е осетлива и ризична затоа е важно да не се пропушти ризичен клиент да остане неоткриен (recall), но исто така е важно да се знае дека предвидувањето е точна (прецизност). Евалуацијата за прецизност-отповик е поинформативна отколку ROC приказот, при проценка на бинарни класификатори на небалансирани множества на податоци, како што е нашата база на податоци за Кредитениот регистар [59].

[12] https://cran.r-project.org/web/packages/nnet/nnet.pdf
[13] https://www.rstudio.com/products/rstudio/



Слика 24. ROC крива – Споредба на резултатите на моделот за балансирано податочно множество, со SMOTE без скалирање

Табела 5. Резултати на перформансите на моделите

		Точност	Прецизност	Отповик	F1 Резултат
Небалансирани податоци без скалирање	Логистичка регресија	0.9008	0.9263	0.9841	0.9543
	Дрво на одлуки	0.9205	0.9221	0.9899	0.9548
	Случајна шума	0.9185	0.9229	0.9864	0.9536
	SVM	0.9145	0.9098	0.9982	0.9520
	Невронски мрежи	0.9090	0.9258	0.9705	0.9477
Небалансирани податоци со скалирање	Логистичка регресија	0.92	0.9277	0.9823	0.9543
	Дрво на одлуки	0.9205	0.9226	0.9962	0.9894
	Случајна шума	0.9215	0.9255	0.9871	0.9553
	SVM	0.915	0.9104	0.9982	0.9523
	Невронски мрежи	0.9065	0.93	0.9623	0.9459
Балансирано множество за тренирање, без скалирање	Логистичка регресија	0.8985	0.9410	0.9394	0.9402
	Дрво на одлуки	0.9025	0.9367	0.9494	0.9430
	Случајна шума	0.9095	0.9367	0.9582	0.9473
	SVM	0.9060	0.9285	0.9635	0.9457
	Невронски мрежи	0.8870	0.9355	0.9311	0.9333

Резултатите покажуваат дека сите модели предвидуваат со висока точност и прецизност користејќи ги небалансираните податоци. Во Табела 5, не се прикажани резултатите од балансираното податочно множество со SMOTE и со скалирање, бидејќи резултатите беа многу слаби. Според резултатите најдобар модел избран според F1

резултатот се дрвото на одлуки, случајната шума и логистичката регресија. Од сите експерименти најлошата комбинација е кога податоците се балансирани и скалирани. Добиените резултати од споредувањето на пет модели на машинско учење покажуваат подобри резултати во споредба со постоечките трудови кои како извор користат бази на податоци за кредити на комерцијалните банки (опишани во Глава 4.2). Од четирите комбинации со балансирање и скалирање, резултатите покажуваат резултати со највисока точност се добиваат со небалансираното и скалираното податочно множество. Најдобар алгоритам во овој случај е дрвото на одлуки, проследено од случајна шума, логистичка регресија, SVM и невронските мрежи. Со висока точност исто така се и експериментите со небалансираното податочно множество и без скалирање, потоа и балансираното податочно множество со SMOTE и без скалирање. Примената на скалирање на атрибутите во користеното податочно множество покажува дека има сосема мало влијание врз резултатите. Ова се случува бидејќи атрибутите веќе беа во скоро истиот опсег и немаат голема разлика. Со користеното податочно множество интересен е фактот дека балансирањето со SMOTE не ги обезбеди очекуваните подобрувања на резултатите како што е опишано погоре за примената на SMOTE. Според резултатите, можеме да заклучиме дека тоа е поради односот на податоците во база на податоци помеѓу главната и помалата класа е 4:1, што во реалност не е голема разлика. Во Табела 6 е прикажана матрицата на конфузност на најдобриот модел, односно при користење на Дрво на одлуки. Од матрицата може да се увидат бројките на точно предвидените ризични и не-ризични клиенти, со бројот на неточно предвидени ризични и не-ризични клиенти

Табела 6. Матрица на конфузност за најдобриот случај

	Предвидени не-ризични клиенти	**Предвидени ризични клиенти**
Актуелни не-ризични клиенти	817563	14422
Актуелни ризични клиенти	3081	164934

Најлоши резултати се добиени со небалансираните податоци и скалираните атрибути. Ова може да биде поради додавањето шум на веќе постоечкиот шум (балансирање потоа скалирање). Во делот за поврзана работа веќе е спомнато дека дрвото за одлуки е често пати најдобар алгоритам за податочно множество за кредити, што се потврди и со извршените експерименти во дисертацијата. Дрвото на одлуки исто така може да биде вредно и за визуелизација и одобрување на кредити.

Целосно истата методологија, со исклучок на атрибутот Возраст е применета и тестирана и за податочно множество кое се состои само од компании. Возраста за компаниите го нема и не може да се изведе од самата база Кредитен регистар. Увидено е дека резултатите се послаби за околу 0.03. Односно најдобар резултат се добива со

користење на Случајна шума и тоа F1 резултат 0,9605 (Табела 7). Овој дополнителен експеримент докажува дека предложениот модел може успешно да се користи за предвидување на ризичните физички и правни лица (компании).

Табела 7. Резултати на перформансите за предвидување ризични компании

		Точност	Прецизност	Отповик	F1 Резултат
Небалансирани податоци со скалирање	Логистичка регресија	0.927	0.9307	0.9876	0.9583
	Дрво на одлуки	0.9265	0.9204	0.9992	0.9585
	Случајна шума	0.931	0.9329	0.9899	0.9605
	SVM	0.9265	0.9205	0.9992	0.9585
	Невронски мрежи	0.8495	0.8465	0.1	0.9186

Презентираната методологија и резултатите од ова истражување може да ги зајакнат автоматизираните или полу автоматизираните одлуки за одобрување на кредит и да го намалат кредитниот финансиски ризик на пазарот. Специфичното податочно множество ќе помогне да се појават и други примени на науката за податоци на истото, со цел да се извлечат други знаења и предвидување, минимизирајќи ги ризиците за кредитите и банките. Освен ефикасното предвидување на кредитниот ризик од базата Кредитен регистар, во горниот дел детално е опишано и деталната методологијата која се применува на оваа специфична база на податоци. Покрај предложената методологија и споредувањето на моделите што ги оценуваме, база на податоци е различна од другите најчесто користени бази на податоци од банките затоа што е вистинска голема база на податоци за Кредитен регистар која ја има само во централни банка. Податоците во оваа база се разликуваат од податоците во комерцијалните банки бидејќи се збирни за кредити и нема податоци за приходите, трошењето, статусот на вработување и деталите за трансакциите. Базата на податоци што се користи во дисертацијата ја има само во централните банки, има историски податоци за сите клиенти во земјата и има поголем потенцијал да го предвиди кредитниот ризик, поради огромната количина на информации од сите комерцијални банки. Од друга страна, недостаток на базата на податоци е недостапноста на личните информации на клиентот, како што е платата на клиентот и трансакциите за трошење.

Ова предложено истражување може да биде дополнителен извор на вредни информации, што ќе им помогне на банките да донесат соодветни одлуки за одобрување кредит. По имплементацијата на овој модел во централна банка, комерцијалните банки треба само да го испратат матичниот број на клиентот, а моделот ќе го врати своето предвидување за ризикот користејќи историски податоци и однесувањето на клиентот во сите банки во земјата, со што се обезбедува информирана одлука добиена од централната банка. Со користење на овој пристап, банките ќе имаат поголеми придобивки, односно наместо да добиваат историски податоци за клиентот од централната банка, тие исто така можат да добијат предвидување за кредитниот ризик за даден клиент. Врз основа на овој модел, за да се одобри кредит, клиентот не смее да

доцни со исплата на претходните кредити и треба успешно да ги исплатил претходните кредити, и возраста да не биде во биновите со висок ризик.

Како недостаток на горе наведеното истражување е дека е користено само едно база на податоци за Кредитен регистар, меѓутоа за да се добие пристап до ваква база на податоци од друга земја е невозможно поради приватноста, сензитивноста и тајноста на секоја централна банка и земја. Врз основа на усогласувањата на Република Северна Македонија со Европската централна банка, опишаната методологија може лесно да се примени и во други земји. Друг недостаток е тоа што нема никакво истражување што користи податоци од кредитен ризик и не може за се изврши споредба на резултатите.

За имплементирање на моделот во продукција, истиот е вграден во SQL Server бидејќи и оригиналното податочното множество е складирано во SQL Server. Ова е направено поради подобри перформанси за да се избегне копирање и пренос на податоците. За оваа цел е користено функционалноста на SQL Server 2016 наречена R Services (SQL Machine Learning Services) која овозможува код напишан во R или Python да се вгради во сторирана процедура и извршувањето се случи директно во SQL Server. Моделот е зачуван во база и потоа во апликација се користи како обична сторирана процедура. Моделот пожелно е на неколку месеци да се тренира пак со нови податоци и така ќе биде поефикасен и по ажурен. Моделот може да се сподели и со комерцијалните банки и истиот може да се надгради со дополнителни информации со кои располагаат истите.

Слика 25. Користење на R Services во SQL Server

5. Примена на NLP за препознавање на ентитети во финансиските информации

Во денешно време имаме пристап до голем број на текстуални документи кои претежно се достапни во неструктурирана форма. Овие документи можат да се користат за автоматизирање на деловните процеси што може да ја олесни целокупната работа на една компанија или институција. Обработката на текстуални податоци во оригинална форма е предизвикувачка задача бидејќи тие се неструктурирани, специфични за нивниот автор и во некои случаи двосмислени поради постоењето на зборови со различно значење во различни контекстуални употреби. Во последно време поради големиот раст на текстуалните податоците во секој сектор, се зголеми и потребата за сумаризација на текстот. Многу компании и институции имаат голема количина на податоци, кои можат да се искористат за да се извлечат информации поврзани со ново знаење. Анализата на текстот односно рударството на текст помага за анализа на клиентите, можностите за маркетинг, за спречување измами, за подобрување на оперативните активности и за развој на нови деловни модели. Кога станува збор за банкарскиот сектор, може да се користат два примарни извори на податоци за анализа на текст: надворешни и внатрешни извори на податоци. Внатрешни податоци се податоците за трансакции, податоци за локација и останати податоци од апликациите. Надворешните податоци се податоци од социјалните мрежи и веб сајтовите. Кога зборуваме за процесирање на природни јазици и извлекување информации од текст, поголемиот дел од времето се мисли на препознавање ентитети и дел од говор (part-of-speech) кои вклучуваат лабелирање на текст. Текстуалната содржина поврзана директно или индиректно за ентитетот може да се користи за автоматизирање на деловните процеси што ја олеснува целокупната работа на една компанија или институција. Извлекување на информации од необработените текстуалните податоци е многу предизвикувачки и затоа е потребна обработка на текстот за да биде по разбирлив за компјутерот и алгоритмите за машинско учење

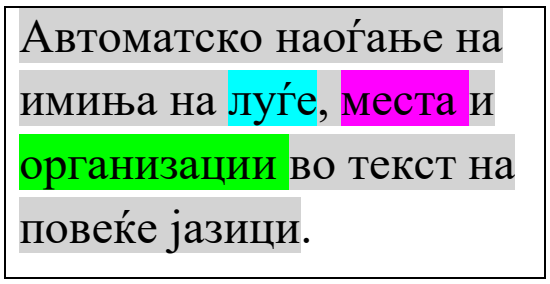
Банките веќе имаат одредени податоци за нивните клиенти. Анализата на текстот може да се интегрира во деловните процеси и тоа овозможува дополнителен придонес за успешноста и сигурноста на овие процеси. Анализата на текст односно рударството на текст вклучува примена на јазични, статистички и техники за машинско учење преку кои ќе се обезбедат информации кои потоа можат да се интегрираат со други извори на податоци како што се структурираните бази на податоци. Едната од првите задачи при анализа на неструктуриран текст е организирањето и трансформирањето во соодветна форма за понатамошна квалитативна и квантитативна анализа. Анализата на текст извлекува релевантни зборови и врски помеѓу нив, со цел нивна категоризација и примена за донесување одлуки. Најчеста примена на анализата на текст во финансискиот сектор вклучува препознавање на ентитети и анализа на чувствата (сентиментот). Рударството на текст стана популарно при анализа на големи податоци во финансискиот сектор. Преку анализа на текстот и процесирање на природните јазици, банките можат да препознаваат вредни информации од клиентите: мислења, документи, објавувања на социјалните мрежи, е-пошта, регистар на повици, измами, ризици итн.

За анализа на големите податоци во банкарскиот сектор најкористени техники за анализата на текст и процесирањето на природните јазици се следните:

- Извлекување на клучните зборови
- Препознавање ентитети
- Анализа на чувства (сентимент)
- Екстракција на теми и
- Анализа на социјалните мрежи

Излекувањето на клучни зборови (keyword extraction) [60], [61] игра клучна улога во финансиските апликации за рударство на текст. Едноставна имплементација на ова техника е извлекување на коментари и статии за дадени клучни зборови, додека понапредна примена е автоматско извлекување на клучни зборови. Поради големиот обем на податоци, истите не може да се читаат сите, затоа цел на оваа техника е да извлече низа на зборови. Во [62] се претставени четири пристапи за извлекувањето на клучните зборови: статистички (фреквенција на поим, обратна фреквенција на документ), лексички (WordNet, n-Gram, Part-of-Speech), машинско учење и хибриден пристап (комбинација од претходните пристапи).

Препознавање на ентитети (Named entity recognition) [63] е една од клучните техники при анализата на текст Слика 26. Целта на препознавање ентитети е да ги идентификува зборовите во текстот кои претставуваат однапред дефинирани категории како што се лице, локација, организација, финансиски инструмент инт. Повеќе сектори користат препознавање на ентитети на големи податочни множества. Скоро сите техники за препознавање ентитети користат машинско учење, кое има потреба за големи податочни множества со цел да се обучи подобро алгоритамот за класификација. Препознавање на ентитети е процес кој ги открива имињата во текстот, ги класифицира според типот на ентитетот и открива врски помеѓу ентитетите.



Слика 26. Препознавање на ентитети

Анализата на чувството (Sentiment analysis) или анализата на мислењето се користи во банкарскиот сектор за да го идентификува мислењето и задоволството на клиентите. Ова техника на процесирање на природни јазици помага за утврдување на ставот за одредена тема или настан. Анализа на чувствата се користи и во банкарскиот сектор за откривање и проценување ризици [64].

Екстракција не теми (Topic extraction) [65] се основа според бројот и дистрибуцијата на поимите низ документите со броење на веројатноста да припаѓаат на одредена тема. Една таква примена е направена во трудот [66] за анализирање на трудовите поврзани со бизнис интелигенција и банкарство. По групирање на трудовите

во повеќе теми увидено е дека најзастапени се темите за ризик, откривање измама, одобрување кредит и банкрот. Со овој пристап, може да се процени веројатноста секој документ да припаѓа на одредена тема. На овој начин може да се идентификуваат темите кои привлекуваат поголемо внимание.

Анализата на социјалните мрежи е процес што се базира на теоријата на графови и се користи за подобро разбирање на социјалните структури. Во случај на Твитер, секој јазол би претставил еден корисник на Твитер, и секој раб е врската помеѓу двајца корисници. Анализата на социјалните мрежа е различен вид на анализа во споредба со анализата на текст, но може да се користи за резултатите од анализата на текстот да се интегрираат во постоечки апликативни решенија [67]. Во трудот [68] користена анализата на социјалните мрежи за да се увиди начинот на кој клиентите на банките влијаат едни на други со цел да пронајдат највлијателните клиенти.

Во последните неколку години се појавија повеќе библиотеки на Python за процесирање на природни јазици, како SpaCy, NLTK и FLAIR. Бидејќи полето на препознавање ентитети се развива како моќна техника, имплементациите во реални апликации стануваат сè побројни. Во ова дисертација ќе се користи систем базиран на машинско учење со помош на библиотеката FLAIR и Python, која веќе покажала оптимални резултати за препознавање на ентитети на неколку светски јазици (англиски, германски, руски, француски итн.). Библиотеката FLAIR е отворена за проширување на функционалностите за повеќе јазици со додавање на новите а податочни множества и обука на повеќејазични векторски репрезентации на зборови. Во дисертацијата се користи FLAIR за препознавање финансиски ентитети во податочното множество од македонски вести. Постигнатиот F1-резултат за македонски јазик е околу 0,75. Детално цел овој процес е опишан во наредните поглавја.

Моментално, најсовремените модели за процесирање на природни јазици користат Bi-LSTM (Двонасочна меморија во кратки интервали на долги периоди). Исто така многу актуелни и распространети се и други методи како што се трансформери со слоеви на внимание (attention layers) [69]. Главната цел на препознавање на ентитети е да се додели однапред дефинирана лабела на зборовите, во зависност од концептот што го опишуваат и припаѓаат. Постојат неколку пристапи за да се изгради таков модел, односно лексички пристап, системи основани на правила, системи основани на машинско учење (длабоко учење) и хибридни системи. Во ова дисертација целта е да се пробаат моделите основани на длабоко учење со користење на библиотеката FLAIR од ZalandoResearch. Целта е собирање и структурирање на податоци од македонските веб агрегати на вести, градење корпус од нив и нивно користење за препознавање ентитети за финансискиот сектор.

5.1. Архитектура на модел на јазик (FLAIR)

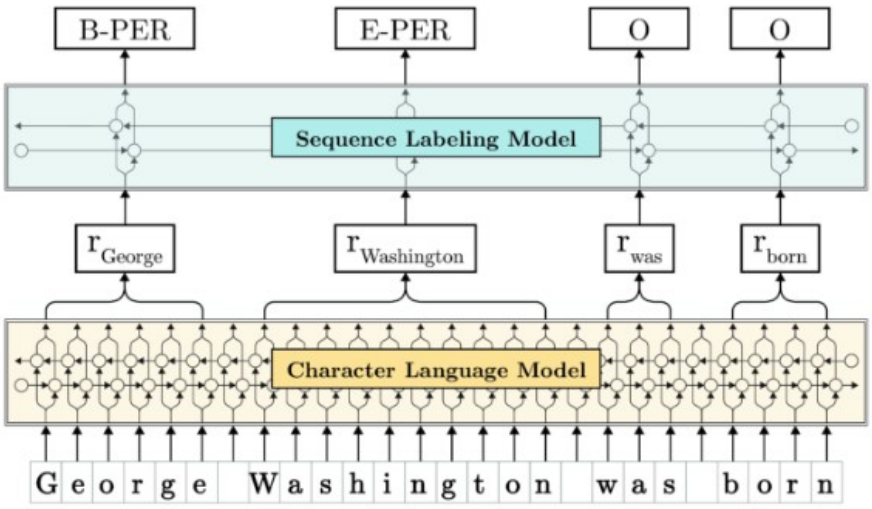
Користењето на машинското учење за разбирање на природните јазици станува се по важна алатка во автоматизацијата на деловните процеси. Клучна цел на овие алатки е обучување на невронските мрежи со минимален обем на податоци во најкраток временски период. Алатката што ја постигнува оваа цел е FLAIR на Zalando Research, која моментално нуди најсовремени решенија за повеќе класични задачи во полето на

процесирање природни јазици, користејќи претходно вградени јазични модели за повеќе од 20 јазици.

FLAIR е библиотека основана на PyTorch и тоа го прави прилично интуитивен да се користи во Python за обука на сопствени модели користејќи ги вградените векторски репрезентации. Библиотеката користи пристап за моделирање на природните јазици со рекурентни невронски мрежи со кои преку дадениот корпус учи моќни и контекстуални информации за јазикот. Овој тип на репрезентација содржи многу семантички и синтаксички информации, кои се клучни за проблемите на препознавање ентитети.

Основната структура на податоците се состои од корпусот и реченициите, кои како влезни податоци се претставени како низа од знаци во претходно обучен јазичен модел. Од овој модел, за секој збор добиваме векторска репрезентација (контекстуална или статичка) што подоцна се користи како влез за Bi – LSTM - CRF (Двонасочна меморија во кратки интервали на долги периоди - условно случајно поле) модел.

Архитектура на FLAIR за лабелирање на секвенца (sequence labeling) е прикажана на Слика 27. Реченицата претставена како низа од знаци е влезен податок во претходно обучен двонасочен јазичен модел од знаци. Потоа векторската репрезентација за секој збор се пренесува во BiLSTM-CRF постигнувајќи најоптимални резултати.



Слика 27. Модел на архитектура со секвенцијално лабелирање

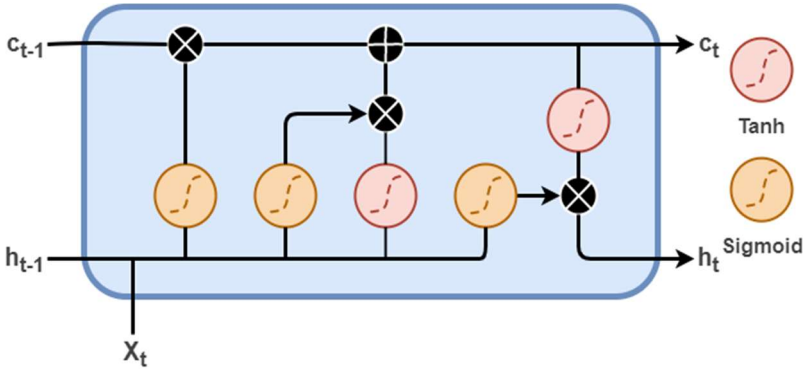
5.1.1. Меморија во кратки интервали на долги периоди - LSTM

Во минатото за препознавање на ентитетите повеќе се користело класичното машинско учење кое бара квалитетен процес на одбирање на атрибутите бидејќи од нив зависи и резултатот на истите. Ова мануелна работа се надмина со појавата на длабоко учење кое на автоматски начин учи и ги открива најдобрите атрибути, и со тоа постигнува и подобри резултати. Постојат повеќе пристапи за примена на длабоко учење при процесирање на природните јазици меѓутоа во последните години, приматот за процесирање на природни јазици го имаат рекурентните невронски мрежи.

Рекурентните невронски мрежи имаат меморија во која може да ја запаметат состојбата на претходните пресметки. Овие мрежи се во можност да моделираат влезни вектори со произволна должина и со зависности на долг опсег [70]. Потоа рекурентните невронски мрежи при обработката на податоците во низата ги земаат во предвид и контекстуалните информаци. Иако традиционалните рекурентни невронски мрежи покажале добри резултати при моделирање на зависностите на долг опсег, тие сепак се ограничени на одреден степен на опсег помеѓу зависностите. Нивното ограничување е поврзано со векторот на градиентниот кој се зголемува или намалува пропорционално со зависностите во долг опсег. Овој проблем предизвикан од ова зголемување и намалување се нарекува градиент на експлозија и градиент кој исчезнува [71]. Овој недостаток се надминува со примена на меморија во кратки интервали на долги периоди (Long Short Term Memory - LSTM).

Архитектурата на LSTM се состои од ќелија со три мултипликативни порти (регулатори), како што е прикажано на Слика 28. LSTM ја додава портата за заборавање која овозможува мемориската ќелија да ги чува информациите долго време. Ќелијата е одговорна да ја чува скриената состојба за одреден временски период и со тоа да ги следи зависностите на хронолошки подредените податоци. Трите регулатори - влезниот, излезниот и портата за заборавање, се одговорни за протокот на информации низ ќелијата, кои информации да ги забораваат и кои да ги пренесат напред.

На Слика 28 е прикажана архитектурата на LSTM единица. Трите σ ги означуваат портите (Влезна, Заборавање, Излезна), x_t е влезот (векторската репрезентација на зборови во нашиот случај), h_t е скриена состојба на ќелијата и c_t е вектор на состојба на ќелијата.



Слика 28. Архитектура на LSTM Единицата

Состојбата на LSTM единицата во определено време t математички е претставена со:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (5)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (6)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (7)$$

$$\tilde{c}_t = tahn(W_c x_t + U_c h_{t-1} + b_c) \quad (8)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (9)$$

$$h_t = o_t \odot tahn(c_t) \quad (10)$$

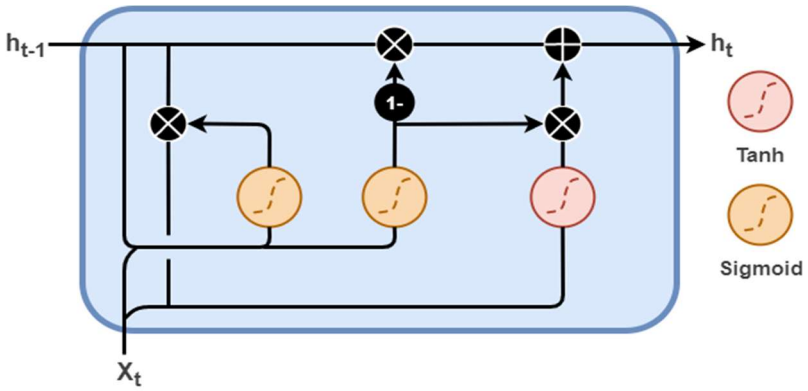
Големите букви означуваат матрици, додека малите букви се вектори. $x_t \in \mathbb{R}^d$ е влезниот вектор за LSTM единицата, $f_t, i_t, o_t \in \mathbb{R}^h$ се векторите за активација на заборавање, влезни и излезни порти соодветно, $\tilde{c}_t \in \mathbb{R}^h$ е ќелијата за активирање на влезниот вектор и $c_t \in \mathbb{R}^h$ е векторот за состојба на ќелијата, $W \in \mathbb{R}^{hxd}$ и $U \in \mathbb{R}^{hxh}$ се матриците на тежина и $b \in \mathbb{R}^h$ е поместување (bias). Матриците на тежина и на поместување на почеток се поставени на нивните почетни вредности и подоцна се учат (оптимизираат) во фазата на обука. d и h се димензиите на влезниот вектор и бројот на скриени состојби.

5.1.2. Двонасочна меморија во кратки интервали на долги периоди - Bi-LSTM

При лабелирање на збор од низата, истиот не зависи само од зборовите што се појавиле претходно, туку и од оние што ќе се појават напред. Архитектурата за скриена состојба LSTM ги зема во предвид само појавите во минатото, кое не е доволно за формирање на целиот контекст. Bi-LSTM [72] освен обуката напред, обучува посебен модел и со обука назад со истата архитектура и потоа ги спојува двете скриени состојби во една, која го кодира минатото и иднината во една низа. Оваа структура овозможува да има информации за назад и напред на секој чекор. Користејќи двонасочна меморија во кратки интервали на долги периоди влезните информации ги обработува двонасочно, зголемувајќи ја количината на информации на мрежата и подобрувајќи го контекстот за алгоритамот. При користење на Bi-LTSM, алгоритамот за учење се храни со оригиналните податоци еднаш од почеток на крај и еднаш од крај до почеток, односно во реалност тоа е користење на две независни рекуренти невронски мрежи.

5.1.3. Рекурентна единица со механизам на порта

Рекурентна единица со механизам на порта (Gated Recurrent Unit) [73] или GRU е едноставна варијанта на мрежата LSTM бидејќи работи на истиот принцип, но и недостасува регулаторот за излез (Слика 29), и со тоа има помалку параметри што го прави посоодветен за помали и небалансирани податочни множества. Рекурентната единица со механизам на порта има два слоеви за разлика од LSTM кој има три слоеви. Првата порта е наречена порта за ресетирање и управува со комбинирање на нов влез и претходните пресметки, односно одлучува колку информации ќе заборави. Втората порта е наречена порта за ажурирање и одредува кои информации од претходните пресметки ќе се чуваат за иднина. Сигмоидните функции ги означуваат портите (Ресетирање и Ажурирање), x_t е влез (векторска репрезентација во нашиот случај) и h_t е скриена состојба на ќелијата.



Слика 29. Архитектура на рекурентна единица со механизам на порта

Една целосна рекурентна единица со механизам на порта ги има следните параметри:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (11)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (12)$$

$$\hat{h}_t = tahn(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \quad (13)$$

$$h_t = (z - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \quad (14)$$

z_t е вектор за ажурирање на порта, h_t е излезниот вектор, $\hat{h}_t$ е кандидат вектор за активирање и r_t е вектор за ресетирање на порта.

5.1.4. Условно случајно поле - CRF

Анализата опишана погоре се однесува на двонасочна насока и како да ги кодираме соседите на зборот со цел да го извлечеме неговиот целосен контекст. Дополнително може да го подобриме извлекувањето на контекстот ако ги земаме во предвид не само соседните зборови туку и нивните лабели. Кога зборуваме за процесирање на природни јазици, може да увидиме дека поголемиот дел од времето гледаме одредена група на лабели заедно, а од друга страна сме прилично сигурни дека ознаката t_1 никогаш не може да биде проследена од ознаката t_2 [74]. Ако го научиме нашиот модел со вакви правила, тогаш веројатноста за подобри резултати е голема и за тоа може да ни помогне само декодерот.

Условно случајно поле (Conditional Random Field - CRF) е дискриминирачка статистичка техника за моделирање на еднонасочни графови кои се користат за пресметување на условната веројатност на вредностите (лабелите) на назначените излезни јазли што ја зема во предвид распределбата на вредностите на соседните јазли при доделување лабели [75]. Да претпоставиме дека имаме случајна променлива X која се состои од зборови што треба да бидат обележани и случајна променлива Y со соодветна ознака за секој $x_i \in X$ и Y е подмножество на множеството $\mathcal{Y}$ кој е просторот на сите можни ознаки во нашиот свет и $P(Y|X)$ е условна веројатност. Потоа според [76] ќе го дефинираме CRF како граф $G = (V, E)$, s.t. $Y = (Y_v)_{v \in \mathcal{V}}$, или со други зборови, нашето подмножество на лабели Y се означува со темиња на G и ние велиме дека (X, Y) е условно случајно поле кога X, случајната променлива Y_v го задоволува својството на

Markov (распределбата на условната веројатност на моменталната состојба, зависи само од претходната состојба) во однос на графот:

$$P(Y_v|X, Y_u, u \neq v) = P(Y_v|X, Y_u, u \sim v) \quad (15)$$

Симболот $\sim$ означува соседство помеѓу две темиња, или со други зборови состојбата Y_i, со оглед на нејзините соседи е условно независна од сите други состојби на графот G.

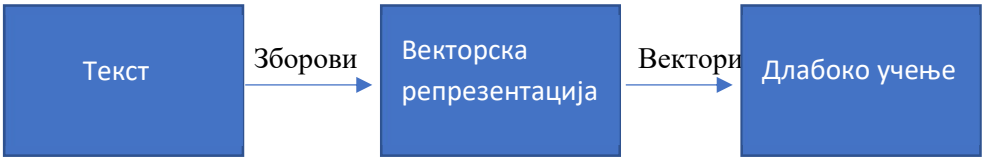
Експериментите во дисертацијата ја потврдија важноста на условното случајно поле, резултатите се прикажани во Табела 8. Од резултатите можеме да забележи дека само со помош на RNN е добиен F1 резултат скоро 0,1 понизок во споредба со случајот кога беше додаден CRF за LSTM и GRU.

Табела 8. Импактот на типови на векторска репрезентација на зборови за проценка на NER_MK моделот

NER – Македонски јазик		
RNN	**F1-резултат (макро)**	**F1-резултат (микро)**
GRU + CRF	0.77	0.75
LSTM + CRF	0.73	0.70
LSTM	0.68	0.65

5.2. Векторска репрезентација на зборови

Невронските мрежи не можат да ги разбираат реченициte како што ги разбираме ние луѓето затоа мора зборовите да ги претвораме во векторска репрезентација (Слика 30). Процесот на претворање на збор, реченица или параграф во вектор на броеви е од суштинско значење со цел да се постигне значаен излез при користење на невронските мрежи. Ова конверзија се нарекува векторска репрезентација на зборови. Невронската мрежа што ја прави конверзијата се нарекува кодер. Векторските репрезентации на зборови се претстави во d-димензионален простор кодирани како густи нумерички вектори [77]. Векторската репрезентација на зборови е тип на репрезентација на зборови која има за цел да го претстави значењето на зборовите во форма на вектори (Слика 30), каде што зборови со слично значење и контекст се претставени со слични вектори. Векторската репрезентација на зборови се смета за важен столб во анализата на чувствата како и во другите анализи при процесирање на природни јазици. Ова репрезентација служи и како прв слој за обработка на податоците при користење на длабоко учење [78].



Слика 30. Пример за примена на длабоко учење за процесирање на текст

Векторска репрезентација на зборови е основен концепт при процесирање на природни јазици. Секој збор што го имаме во корпусот мора да се претвори во вектор со

помош на овие претходно обучени модели со цел да биде разбирлив за компјутерот. Постојат и се јавно достапни различни претходно обучени векторски репрезентации на зборови. Овие се резултат од комбинацијата на различни модели за векторска репрезентација на зборови. Популарен модел за векторска репрезентација на зборови е GloVe, развиен од група истражувачи на Универзитетот Стенфорд [79] во 2014 година. GloVe ги надмина недостатоците на Word2Vec во фазата на обука, подобрувајќи ги генерираните векторски репрезентации. GloVe како кодер на зборови е користен во многу истражувања за анализа на сентименти при процесирање на природни јазици [80]. Facebook Research во 2016 го публикуваше проектот за векторска репрезентација на зборови fastText [81], кој претставува брз и ефикасен метод за учење на претставите на зборовите. Главната карактеристика на fastText е дека ја зема во предвид внатрешната структура на зборовите која овозможува да се научат различните морфолошки форми на зборовите. fastText работи со лизгачки прозорец над влезниот текст и учење на централниот збор од преостанатиот контекст. Последно, голем успех постигна векторската репрезентацијата со користење на BERT моделот. BERT користи трансфомер кој е основан на механизам за самостојно внимание што го истражува контекстуалниот однос помеѓу зборовите или под-зборовите во текстот.

Во FLAIR секоја класа Embeddings наследува од интерфејсот TokenEmbedding или DocumentEmbedding. Двете интерфејси наметнуваат имплементација на методот .embed() кој се користи за конвертирање на Sentence објект или листа од Sentence објекти во тензор. Овие векторски репрезентации на зборови можеме да ги поделиме на две категории: статични и контекстуални.

5.2.1. Статичка векторска репрезентација

Статичката векторска репрезентација (Static Embeddings) е најстарата векторска репрезентација на зборови. Ова репрезентација има голем корпус на зборови, и во реалност е како речник кој за секој збор враќа соодветен вектор како вредност. Статичката векторска репрезентација има иста претстава за секој збор што има иста форма, без оглед на неговото семантичко значење, односно не ја зема во предвид распределбата на зборовите пред и по целниот збор за кој ја моделираме векторската репрезентација.

Статичките векторски репрезентации кои се вклучени во FLAIR се:

- WordEmbedding
- BytePairEmbeddings
- CharacterEmbeddings
- FastTextEmbeddings
- OneHotEmbeddings

5.2.2. Контекстуална векторска репрезентација во FLAIR

Контекстот на зборовите е многу важен кога се работи со задачи за процесирање на природни јазици. Контекстуалната векторска репрезентација на зборови (Contextual String Embeddings) ги користи внатрешните состојби на обучен модел на јазик за да создаде нова векторска репрезентација за зборовите. Ова репрезентација користи

одредени внатрешни принципи на обучен модел за знаци, за да генерира соодветен вектор бидејќи истиот збор во различни реченици има различно значење.

Контекстуалната векторска репрезентација (Contextual string embeddings) се смета како дел на современата обработка на природните јазици, особено поради следните важни аспекти на ова поле:

1. Можност да се претходно обучени на голем нелабелиран корпус
2. Го препознаваат семантичкото значење на зборот и овозможуваат различни претстави на полисемни зборови во зависност од контекстот во кој се користени
3. Ги моделираат зборовите како низа од карактери (знаци), што ги надминува проблемите кога некој збор го нема во дадениот речник [82]

Најчесто векторската репрезентација на зборовите во форма на вектор (или тензор) се врши со помош на верзијата LSTM на рекурентната невронска мрежа. Како пософистицирана векторска репрезентација е користењето на BERT кој користи трансформатори на јазични модели со посебни слоеви за внимание.

Тие работат на ниво на карактер и нивната главна цел е да најдат доволно добро предвидување на дистрибуцијата $P(c_{0:T})$ за низа на карактери $(c_0, c_1, \ldots, c_T) = x_{0:T}$. Со обука на моделот, се добива $P(c_t|c_0, \ldots, c_{t-1})$ или предвидување за дистрибуција на следниот знак. Заедничката дистрибуција на целата реченица може да се претстави како производ на предвидените дистрибуции на сите карактери, условени од дистрибуциите на нивните претходници:

$$P(c_{0:T}) = \prod_{t=0}^{T} P\left(c_t|c_0, \ldots, c_{t-1}\right) \tag{16}$$

При архитектурата на LSTM, условната веројатност $P(c_t|c_0, \ldots, c_{t-1})$ е приближно, функција на излезот на мрежата o_t.

$$P(c_t|c_0, \ldots, c_{t-1}) \approx \prod_{t=0}^{T} P\left(c_t|o_t; \theta\right) \tag{17}$$

каде θ е вектор на сите параметри на моделот.

Тајната за успешноста на векторските репрезентации на FLAIR е ефикасното користење на скриените слоеви на LSTM мрежата. Покрај моделот напред, моделот наназад исто така се обучува на целосен идентичен начин, но во целосна спротивна насока.

$$P^r(c_t|c_{t+1}, c_{t+2}, \ldots, c_T) = \prod_{t=0}^{T} P^r\left(c_t|c_{t+1}, \ldots, c_T\right) \tag{18}$$

$$P^r(c_t|c_{t+1:T}) \approx \prod_{t=0}^{T} P^r\left(c_t|o_t^r, \theta\right) \tag{19}$$

r претставува обратен редослед и o_t^r е функција на излезот на LSTM мрежата, која на некој начин ги кодира сите податоци што моделот ги видел за дадениот карактер.

Може да види дека моделот за напред го доловува семантичкото значење на реченицата до точката на набљудуваниот карактер, додека моделот за назад ја доловува истото значење но почнувајќи од крајот на реченицата до тој знак. Клучната поента е да се спојат двете модели за да се добијат семантичките информации и контекстот на целиот збор и неговите соседи со напластување на излезите на моделите за напред и за назад:

$$we_i^{charLM} = \begin{bmatrix} o^f_{t_{i+1}-1} \\ o^r_{t_i-1} \end{bmatrix} \tag{20}$$

5.2.3. Векторска репрезентација со користење на BERT моделот

Во 2018 [83] користејќи ја архитектурата на трансформери се појави револуционерниот модел за векторска репрезентација на јазици наречен BERT (Bidirectional Encoder Representations from Transformers). BERT е развиен од Google и е претходно обучен модел користејќи ги податочните множества на Wikipedia и BooksCorpus (16 GB податоци, односно 3.3 милијарди зборови). Моментално BERT обезбедува претходно обучени модели за англиски и за 103 други јазици, кои потоа може да ги прилагодиме според нашите потреби. Овој револуционерен модел е почеток на нова ера во процесирањето на природни јазици, постигнувајќи врвни резултати. BERT го надминува ограничувањето на претходните модели за јазици кои вклучуваат само еднонасочни претстави на зборовите во реченициte, воведувајќи двонасочен маскиран модел на јазик, кој случајно ги предвидува маскираните зборови во реченицата, збогатувајќи ги контекстуалните информации на зборовите. Во дисертацијата се користи BERT за доусовршување за класификација на финансиските вести. Во оваа дисертација со користење на претходно обучениот модел BERT и по дополнително обучување и комбинирање со векторски репрезентации на зборови се постигнаа успешни перформанси.

BERT во основа е обучен трансформер за кодирање со 12 кодери во основната верзија и 24 во големата верзија. BERT користи голема ациклична невронска мрежа (768 јазли во основната верзија и 1024 во големата верзија) и повеќе глави на внимание односно 12 и 16 за соодветната верзија. BERT е обучен со Wikipedia и со Book Corpus. BERT има повеќе варијанти кои се обучени на соодветни корпуси.

BERT како влезни податоци прима низа на зборови (реченица) и потоа како излез за овие низи генерира соодветен вектор за секој збор. Векторот за соодветниот збор зависи од соседните зборови. Пример за зборот падна во реченициte „Падна цената на дизелот“ и „Топката падна во река“ ќе се генерираат различни вектори. Во првата реченица векторот за зборот падна е поблиску до зборот цена додека во втората реченица зборот падна е поблиску до векторот на зборот река.

BERT креира векторска репрезентација за збор на следниот начин:

- BERT има вокабулар со фиксен број на зборови и под зборови. Зборовите кои ги има во вокабуларот ги мапира директно со соодветниот вектор, додека тие што ги нема ги дели во под зборови и за нив се труди да најде соодветен збор во вокабуларот.
- При обучување на моделот, BERT ја учи векторската репрезентација користејќи ги главите за внимание и другите матрици за трансформација кои исто така се учат за време на обучувањето. Потоа обучениот модел заедно со научените вектори се користи за тестирање, кој потоа генерира вектор за секој збор.
- За генерирање на вектор за секој збор, BERT исто така ја зема во предвид и позицијата на зборот во реченицата. За BERT клучен придонес имаат и соседните зборови.
- Секој слој на BERT моделот има повеќе глави на внимание (12 во основниот и 16 во големиот) и нелинеарна ациклична невронска мрежа која ги зема излезите од главите на внимание и овозможува да комуницираат едни со други пред да се проследат во наредниот слој.

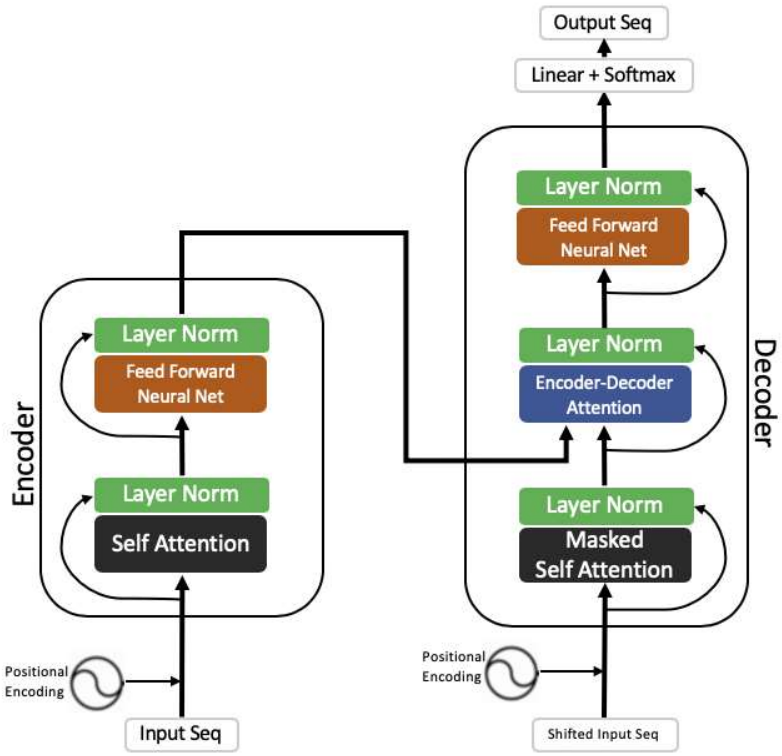
Како претходно тренирани модели на BERT се: BioBERT (за биомедицински текст), SciBERT (за научни публикации), VideoBERT, TransBERT, ClinicalBERT (за клинички белешки), M-BERT, G-BERT (за дијагностика во медицина), итн. Овие специфични модели постигнуваат подобри резултати во споредба со примената на генералниот BERT модел.

DistilBERT е дестилирана верзија на BERT со 40% помала големина на моделот, 60% побрз, задржувајќи ги 97% перформансите, и користејќи само пола од бројот на параметри [84]. Техниката за дестилација на големата невронска мрежа ја намалува невронската мрежа по обучување и дистрибуциите ги претставува со помала мрежа. DistilBERT се обучува за четири пати побрзо од BERT и има послаби перформанси до 3%.

FinBERT [85] е верзија на BERT наменета за финансиски сектор, кој е претходно обучен на финансиски корпус кој се состои од 1,8 милиони вести од Ројтерс објавени помеѓу 2008 и 2010 година. Овој модел на BERT постигнува до 15% подобрување на точноста класификација на финансиски текстови.

Целта на BERT, исто како и претходните модели е да изгради однапред обучен модел од неозначени податоци користејќи длабоки двонасочни претстави од двете страни на зборот.

Структурата на BERT модел (Слика 31) се состои од делот за кодер кој е напластена рекурентна единица (како LSTM или GRU) за процесирање на еден елемент од влезната низа и проследување на добиената информација, и од декодерот кои има рекурентни единици кои генерираат предвидување за излез y_i, по еден со временски чекор i. Структурата на архитектурата на BERT е визуелно прикажана на Слика 31.



Слика 31. Трансформери. Аркитектура со слоеви за самостојно внимание

Кодер е множество на N идентични рекурентни слоеви на единици, обично 6, 12 или 24 и секој од слоевите има посебен вектор на тежина. Потоа секој од слоеви се состои од две под-единици, под слојот за само-внимание и ациклична невронска мрежа. Влезот на кодерот е првичната векторска репрезентација на збор (репрезентација на токен, сегмент и позиција). Влезот прво поминува низ слојот за само-внимание со цел моделот да бара други зборови во реченицата, потоа излезот на овој под слој се пренесува како влез во ациклична невронска мрежа.

Под слој за самостојно внимание (Self-Attention sublayer) е концепт за наоѓање на врските помеѓу одредени зборови во реченицата. Пример ако ја имаме следната реченица „Брзиот автомобил падна на реката Вардар“, целта е алгоритамот да научи дека зборот „брзиот“ се однесува за „автомобилот“, и зборот „реката“ се однесува за „Вардар“.

Од гледна точка на алгоритам, само-вниманието е процес кој се состои од 6 чекори [69]:

- Конструирање на вектори за пребарување, клуч и вредност за секој од влезните кодери: За време на фазата на обука, се оптимизирани 3 матрици θ^Q, θ^K и θ^V, секој од влезовите се множи со една од овие матрици за да се добие соодветниот вектор со фиксна должина.
- Пресметување на резултат: Ако пресметуваме векторска репрезентација на збор во реченица, тогаш го пресметуваме производот со точки QueryVector.KeyVector за овој збор наспроти преостанатите зборови во реченицата. Резултатот за набљудуваниот збор со самиот себе е секогаш

најголем затоа што правиме производ со точка помеѓу истиот вектор и ја добиваме неговата квадратна норма, колку е поголем производот, толку е посилна врската.

- Делење на резултатот: Третиот чекор е да се подели резултатот со квадратниот корен на димензијата на клучните вектори за да се добие постабилен градиент
- Нормализирање: излезот од претходниот чекор сега се пренесува преку функцијата softmax со цел сите димензии да се направат позитивни и да се нормализира резултатот на векторот
- Множење на резултатот со вредниот вектор: Во овој чекор се множи вредниот вектор со резултатот на softmax со цел да се ослободиме од нерелевантни врски со мали резултати на softmax
- Збир: последниот чекор е да се сумираат сите вектори со пондерирана вредност и да се добие излез за само-внимание за набљудуваниот збор

$$Attention\left(\vec{Q},\vec{K},\vec{V}\right) = softmax\left(\frac{\vec{Q}\vec{K}^{T}}{\sqrt{d_k}}\right)\vec{V} \quad (21)$$

Декодерот се состои од ист N број на единици поделени во под-единици исто како кај кодерот. Влез за секоја од единиците за декодер се векторите за само-внимание K и V што се користат за одредување на соодветниот фокус на декодерот во низата на влез. Излезот на декодерот кој е вектор се пренесува на крајната линеарна единица која го трансформира во многу поширок вектор во зависност од големината на речникот. Оваа единица го преведува векторот на бројот во краен резултат односно збор. Излезот од овој слој е вектор на веројатности за зборовите во вокабуларот и според веројатноста ќе се избере ќелијата за зборот со најголема веројатност.

Напластени вектори (Stacked Embeddings) - За подобри резултати, најчесто имаме потреба за комбинирање на повеќе векторски претстави на зборови. Во FLAIR ова е овозможено со користење на класата StackedEmbeddings што е иницијализирана со наведените векторски репрезентации, напластени една со друга,

$$we_i = \begin{bmatrix} we_i^{charLM} \\ we_i^{static} \end{bmatrix} \quad (22)$$

и потоа тие функционираат како редовна класа Embedding, односно тие го наследуваат интерфејсот TokenEmbedding и го имплементираат методот embed(). FLAIR освен напред-назад и BERT векторските репрезентации, исто така вклучува PooledFlairEmbeddings, ELMoEmbeddings и TransformerWordEmbeddings кои вклучуваат неколку класични варијанти на претходно обучени трансформатори, како што се RoBERTa, XLM итн.

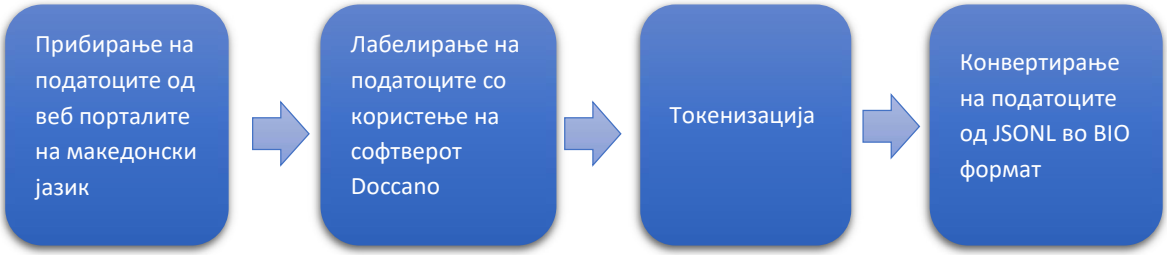
Иако постои изјава од ZalandoResearch дека најдобри резултати за NER се добиваат со комбинирање на контекстуалните и статичките векторски репрезентации, моделот во дисертацијата работи подобро кога ќе се користат само контекстуалните векторски репрезентации (Табела 9).

Табела 9. Импактот на типови на векторска репрезентација за резултатот на NER_MK моделот

NER – Македонски јазик		
Векторска репрезентација	**F1 - Резултат (макро)**	**F1 – Резултат (микро)**
FlairEmbeddings + BERT	0.77	0.75
FlairEmbeddings + BytePair + BERT	0.67	0.64
WordEmbeddings + BytePair + BERT	0.63	0.59

5.3. NER модел – Македонски јазик

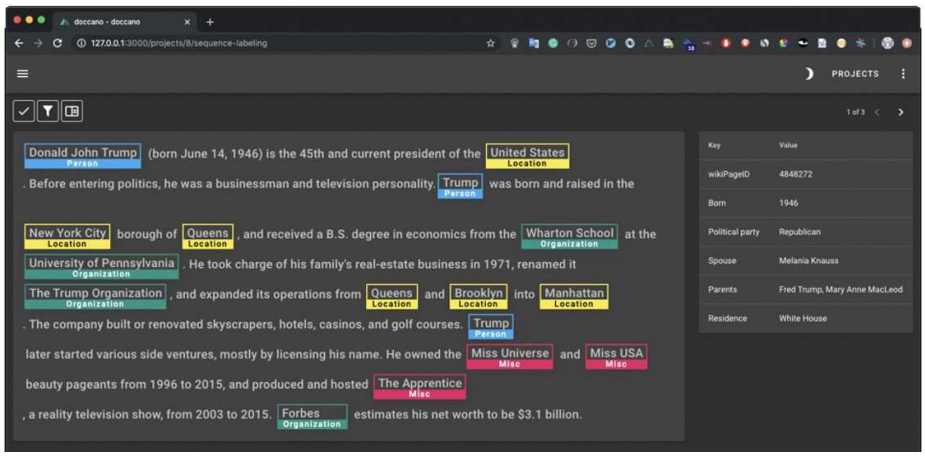
Експериментот за моделот на препознавање ентитети на македонски јазик, е поделен и изведен во неколку фази (Слика 32). Во првата фаза се прибираат податоците од веб порталите на македонски јазик, потоа е направено лабелирање на насловите од вестите со користење на софтверот Doccano, а потоа направена е токенизација и конвертирање на податоците од JSNONL во BIO формат. На крај тестирани се различни модели и податочни множества за обука, со цел да се стигне до најдобриот резултат.



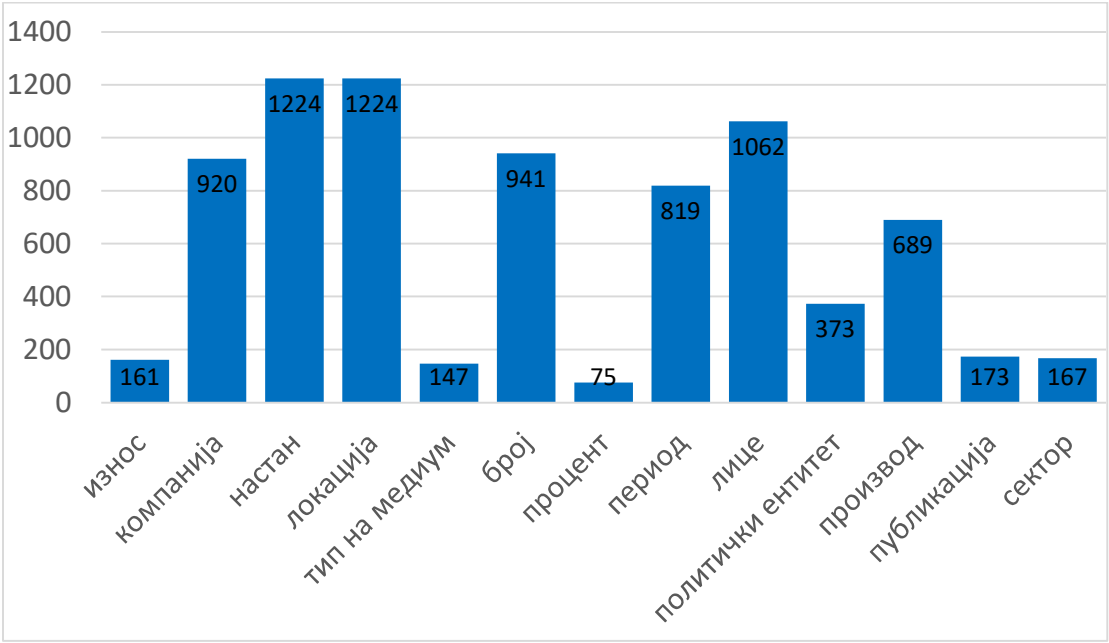
Слика 32. Прибирање и пред обработка на податоците

5.3.1. Податоци

Податоците што се користат во овој експеримент се преземени од македонските веб страници за on-line вести, односно time.mk и greed.mk. Од овие веб страни се вчитани насловите на вестите од Македонија, Балканот и светот, за следните категории: финансии, економија, политика, живот, хроника, култура, технологија и сцена. Вкупно податоци се 2000 вести на македонски јазик, и 500 на албански јазик. Дел од податоците се добиени со користење на fetchRSS во формат csv, а другиот дел е добиени со помош на библиотеката BeautifulSoup во Python. Лабелирањето на податоците е направено со користење на софтверот Doccano (Слика 33), со користење на 13 релевантни класи со доволно голем број на примероци за обука како што е прикажано на Слика 34.



Слика 33. Апликација Doccano



Слика 34. Лабели

FLAIR ги користи сите CoNLL формати за претставување на податоци во текстуални датотеки. Секој токен се доделува во нов ред со соодветна лабела и ознака, односно ознака B, ознака I и ознака O (ознаката BIO означува Почеток, Внатре, надвор [86]). Кога нема лабела за одреден токен, тој се обележува со ознаката O. Кога токенот е дел од некој ентитет, тој се обележува или со ознака I или ознака B (како што е прикажано на Слика 35). Ознаката I се користи само кога даден токен го има истиот ентитет како и претходниот без O-ознаки помеѓу нив (ентитети со повеќе токени). Постојат и екстензии на овој формат, како што е форматот BIOES.

Велика	*B-Location*
Британија	*I-Location*
и	*O*
САД	*B-Location*
почнаа	*B-Event*
преговори	*I-Event*

Слика 35. IOB-тагирање (пример)

5.3.2. Токенизација на податоци

При процесирање на природни јазици, многу важен чекор е токенизирање на реченициte. Точноста на распределбата на лабелите во текстот е директно зависна од видот на токенизаторот што се користи за разделување на реченициte на одделни зборови, т.е. токени. Токенизаторите можат да бидат едноставни, односно за разделување на реченициte на празните простори, но исто така постојат понапредни и современи модели како SpaCy, NLTK и Segtok токенизатори.

Изборот на токенизатор главно зависи од природата на проблемот со кој работиме и грануларноста на токените што очекуваме да ги постигнеме. Некои проблеми за процесирање на природни јазици кои се многу зависни од граматичкиот состав на зборовите, т.е. сумирање на текстот, каде што излезот е низа зборови што очекуваме да бидат граматички точни, може да не функционираат добро ако го разделуваме текстот според празните простори. За вакви проблеми потребни се претходно обучени јазични модели како токенизатори кои можат да ја поделат реченицата во зависност од нејзината граматичка формулација.

За експериментот во продолжение за токенизација на текст се споредени двата различни пристапи, односно едноставните функции за разделување на празните простори и токенизаторот Segtok кој е стандардна техника за токенирање во FLAIR. Од експериментите се увиде дека токенизаторот Segtok работи најдобро кога ги изоставивме знаците за интерпункција. Бидејќи за проблемот на процесирање на ентитети важно е да се земаат во предвид и интерпункциските знаци, затоа во продолжение е користен едноставниот начин за токенизација.

5.3.3. Обука

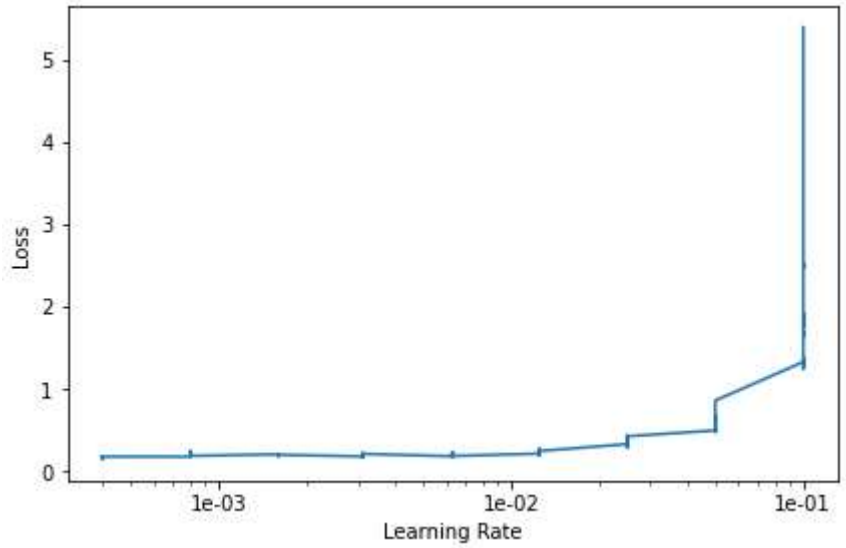
Обуката на моделот е направена во многу циклуси, експериментирајќи со векторските репрезентации на зборови, типот на рекурентна невронска мрежа и хипер параметрите на моделот. Земајќи ја предвид сличноста на македонскиот јазик со другите словенски јазици (српски, руски, итн.), идејата беше да се редат сите овие WordEmbeddings заедно со BytePairEmbeddings за македонскиот јазик за да се добијат задоволителни резултати. Меѓутоа не само што не се добија добри резултати, туку и процесот на обука траеше многу долго поради многуте димензии на влезните векторски репрезентации

Како следен чекор се користени под множества на претходните јазици за да се комбинираат со контекстуални векторски репрезентации како FlairEmbeddings и BERTEmbeddings. Овој пристап ги подобри малку резултатите, но како што е наведено погоре, најдобрите резултати се добиени кога се користени контекстуални повеќе јазични векторски репрезентации, во овој случај тоа се FlairEmbeddings ('multi-forward'), FlairEmbeddings ('multi-backward') и BERT (како што е прикажано во Табела 11).

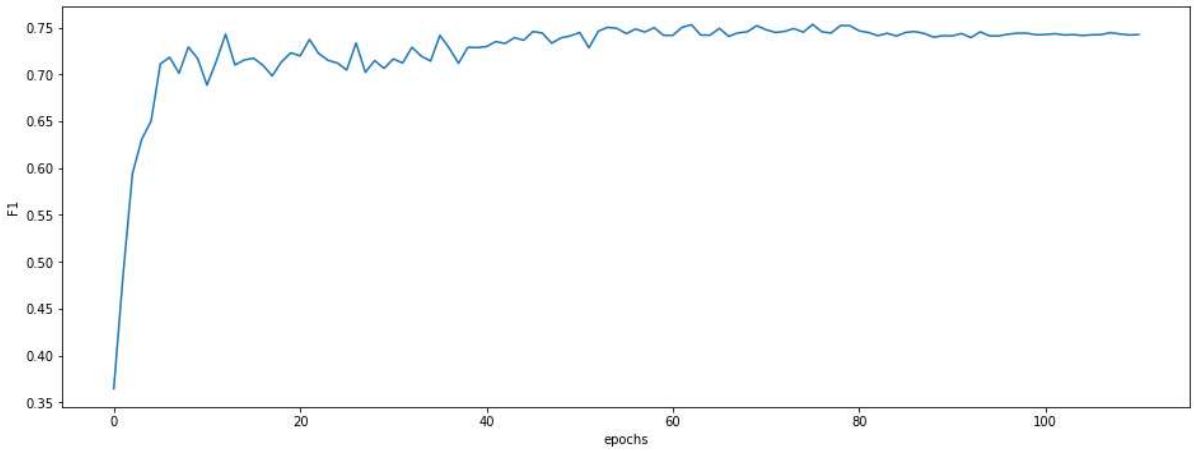
Кога станува збор за кој тип на рекурентна невронска мрежа да се користи при препознава ентитети, клучното прашање е колку е големо податочното множеството на податоците. Разликите помеѓу LSTM и GRU веќе се дискутирани погоре, и може да се заклучи дека кога има огромно и комплексно податочно множество тогаш најдобри резултати се очекуваат да се постигнат со користење на LSTM. Кога податочното множество е помало, тогаш варијантата GRU се очекува да има подобри резултати во повеќето случаи. Наведената граница за големината на податоците не е јасно дефинирана затоа како најбезбедна варијанта е да се проверат и LSTM и GRU.

Најверодостоен хипер-параметар во кој било модел на машинско учење е најчесто стапката на учење, кој чекор се користи во алгоритмите за спуштање на градиентот за оптимизирање на функцијата на загуба. Најчестата и препорачана вредност за овој параметар во анализите на NER е 0,1, што се покажа точно и во овој експеримент. Зголемувањето на стапката на учење ќе го зголеми бројот на потребни епохи за да се спојат функциите за загуба, додека намалување на истото би ја направила функцијата на загуба да се спои прерано и да го остави моделот да има помалку информации за податоците. Друг корисен хипер-параметар што може да се користи е параметарот reproject_embeddings, кој додава слој за обучување на врвот на целосно поврзаниот слој за повторно проектирање на векторските репрезентации. Честопати има проблем со небалансирано податочно множество и одредена лабела ја покажува како поважна од другите класи. Постојат и неколку други хипер-параметри кои оставени со стандардните вредности на FLAIR.

Моделот во дисертацијата е обучен со 200 епохи преку мрежа со скриена големина од 256 слоеви, почнувајќи со стапка на учење од 0,1. Од Слика 36 можеме да заклучиме дека моделот ги научил повеќето информации со стапка на учење од 0,1, 0,05 и 0,025. Како што е прикажано на Слика 37, F1 од множеството за обука започнува да се спојува во епоха 80.



Слика 36. Загуба споредено со стапка на учење



Слика 37. f1 споредено со број на епоки

5.4. Резултати и дискусија

Главната задача на овој модел беше обележување зборови во нови невидени реченици со една од 13-те претходно споменати класи. Од резултатите кои се прикажани во Табела 10, класите со повеќе статички соседи како што се „ПРОЦЕНТ“ и „ИЗНОС“ воопшто не претставуваа проблем за претставениот модел. Причината за добрите резултати е што со голема веројатност лабелираните сегменти од класата „ПРОЦЕНТ“ го содржат знакот % или зборот „проценти“. Истото важи и за лабелираните сегменти од класата „ИЗНОС“ кои секако содржат валута како збор или симбол. Попредизвикувачката класа која сè уште работи добро е класата „БРОЈ“. Причина за збунетост за оваа лабела е дека се мешаат со класата "PERIOD" која содржи нумерички временски опсег. Лабелирањето на ентитетите со класата „НАСТАН“ дава најлоши резултати бидејќи настаните во текстовите се многу стохастички по природа и тие обично немаат фиксна дистрибуција на околните зборови како класата „ПРОЦЕНТ“ и „ИЗНОС“. Како можен начин за подобрување на резултатите за одредени

класи со незадоволителни резултати е да се зголеми нивната тежина, но во овој случај ресетирањето на тежините влијаеше негативно на резултатите на моделот. Поединечно имаше подобрување за класата „НАСТАН“, меѓутоа не вреди да се намали вкупниот резултат на моделот поради поединечни класи.

Табела 10. Дистрибуција на TP, FP и FN

NER – Македонски јазик						
Лабела	**TP**	**FP**	**FN**	**Прецизност**	**Отповик**	**F1 Резултат**
Износ	9	0	4	1.00	0.69	0.82
Компанија	78	22	33	0.78	0.70	0.74
Настан	71	58	78	0.55	0.48	0.51
Локација	110	21	26	0.84	0.81	0.82
Тип на медиум	16	3	9	0.84	0.64	0.73
Број	117	40	23	0.75	0.84	0.79
Процент	3	0	0	1.00	1.00	1.00
Период	94	18	20	0.84	0.82	0.83
Лице	116	20	14	0.85	0.89	0.87
Политички ентитет	32	8	7	0.80	0.82	0.81
Производ	59	32	26	0.65	0.69	0.67
Публикација	19	7	7	0.73	0.73	0.73
Сектор	10	4	5	0.71	0.67	0.69

За споредба истиот корпус со податоци на македонски јазик се проба и со во моделот BERT Моделот со слој на внимание, споделувајќи ги истите хиперпараметри. Меѓутоа моделот FLAIR даде подобри резултати споредено со BERT (како што е прикажано во Табела 11). Ова може да се должи на фактот што со моделот на BERT воопшто не ги нагодувавме параметрите и не направивме детална анализа, бидејќи фокусот за постигнување подобри резултати е со двонасочниот јазичен модел на FLAIR.

Табела 11. FLAIR споредено со BERT моделот

NER – Македонски јазик		
	F1 – Резултат (макро)	**F1 – Резултат (микро)**
FLAIR	0.77	0.75
BERT	0.55	0.52

Бидејќи македонскиот јазик е дел од јужната група на словенски говорни јазици, како предизвик беше да се увиди какви информации може да извлече нашиот модел од друг јазик од истата група. Поради сличноста првично е тестиран со корпус на српски јазик, односно со податоци од српските веб портали за вести. Повеќето од вообичаените класи како "ЛИЦЕ", "ЛОКАЦИЈА" и "ПОЛИТИЧКИ ЕНТИТЕТ" беа откриени со висока точност, со f1 резултат од 0,83, 0,8 и 0,66. И во овој случај класата „НАСТАН“ не појави добри резултати. Микро f1-резултатот за целиот корпус е 0,61 и макро-f1-резултатот е

0,48. Причина за овие послаби резултати е малото податочно множество за обука односно отсуството на голем број на настани во множеството.

Анализите се продолжени и со испитување на повеќе јазичниот модел како ќе се однесува со податоци од други различни јазици. За оваа цел е тестирано со вести на албански јазик. Како што и се очекуваше, повеќе јазичниот модел на FLAIR заврши одлична работа во комбинирањето на овие два јазици, така што немаше пречки од еден на друг јазик во фазата на обука.

Ова се покажа како одлична техника кога се работи со корпуси со мало количество податоци, бидејќи целиот процес делува на начин на зголемување на податоците. Македонскиот корпус беше поддржан од албанскиот и обратно, што на крајот помогна функцијата на загуба да се спои побрзо, бидејќи податоци за обука се зголемија и тоа ги подобри резултатите на тестирање за двете јазици.

Поради големиот обем на податоци и информации, се зголеми и потребата за промена на начинот на кој ги извршуваме секојдневните задачи од рачно во автоматско начин. Алгоритмите за процесирање на природен јазик се дизајнирани да анализираат и процесираат голема количина на текстови и да извлекуваат информации од нив, што во реалност истото не може да се направи мануелно од човекот. Препознавањето на ентитети е само една гранка од широката палета на можности за процесирање на природни јазици. При создавање на ваков модел треба да се има во предвид за која област е наменет моделот и кој ќе го користи истиот. FLAIR е едно модерно решение за овој проблем, кое ја минимизира рачната работа со податоците и ги оптимизира резултатите, но тоа што е најважно го поддржува и олеснува развојот на модели за помали јазици (како што е македонскиот) со малку или без никакви ресурси. Потоа, во експериментите користен е корпус од јазик од иста јазична група со македонски и добивме задоволителни резултати за истите класи како и за македонскиот јазик, докажувајќи колку е моќен концептот на повеќе јазични модели на јазик. Предложениот модел за препознавање ентитети на македонски и албански јазик претставува важен чекор за процесирање и препознавање информации на македонски и албански јазик.

6. Примена на NLP за одредување на сентиментот на финансиските информации

Во минатото пред ерата на социјалните мрежи, за да се дознае мислењето на другите за одредено нешто, се прашувале пријателите и познаниците. Кога на компаниите или организациите им требало мислењето од јавноста, применувале анкети за да го дознаваат нивното мислење. Знаењето за мислењето на јавноста или потрошувачите претставува вистинска вредност за носење на многу одлуки кое влијае директно на нивната успешност. Анализата на чувствата (сентиментот), исто така именувана како рударство на мислења ги анализира мислењата на луѓето, чувствата, проценитe, ставовите и емоциите кон субјектите како што се производи, услуги, организации, поединци и настани. Мислењата играат важна улога во скоро сите луѓе активности бидејќи тие се индикатори за однесување и за успешност. Често при носење одлуки, компаниите и поединците се основаат и според мислењата на другите. Во денешно време се почесто се применува анализата на мислењето на корисниците. Исто така самите потрошувачи сакаат да ги знаат мислењата на постојните корисници за одредениот производ, услуга или настан. Со големиот раст и експанзија на социјалните мрежи во форма на коментари, слики, и објавувања на социјалните мрежи, индивидуалците и компаниите сè повеќе ја користат оваа достапна содржина за носење на информирани одлуки. Во денешно време кога некој сака да купи производ, првично се основа од мислењата на другите корисници на социјалните мрежи, наместо да ги праша своите пријатели како што било во минато. За компаниите веќе е многу полесно и ефикасно прибирање на потребните мислења преку филтрирање на релевантните информации на социјалните мрежи, при секојдневното работење. За да се постигне овој ефикасен начин на извлекување и идентификување на чувствата потребно е да се примени автоматска анализа која ќе им обезбеди готови информации.

Поимот анализа на сентимент односно анализа на чувствата се појавува за да увиди позитивна и негативна класа во податоците [87], [88]. Апликациите за големи податоци во финансискиот сектор вклучуваат анализи на чувствата за управување со кредитниот ризик [89]. Еден од можните патишта за да се извлечат информации од огромното количество големи податоци е рударството на текст или анализата на текст. Анализата на чувствата во финансискиот сектор најчесто се основа од финансиските вести, дали текстот е објективен или субјективен и се обидува да утврди дали текстот содржи позитивни или негативни чувства [90]. Ова примена во реалност претставува класификација во позитивни и негативни чувства. Важен чекор за ова имплементацијата претставува прибирањето во реално време на финансиските вести од веб изворите. За прибирање на финансиските вести најлесно е истите да се прибираат од сајтови кои се агрегати на вести од повеќе извори и потоа нивно копирање и процесирање во своја база на податоци.

Банкарскиот сектор постојано ги следи финансиските и економските вести, бидејќи според поуките во банкарскиот сектор вестите секогаш преодат на промените во финансиските пазари [91]. Со тоа новите вести се важно знаење за предвидување на идните промени и за носење на информирани одлуки, особено во банкарскиот сектор.

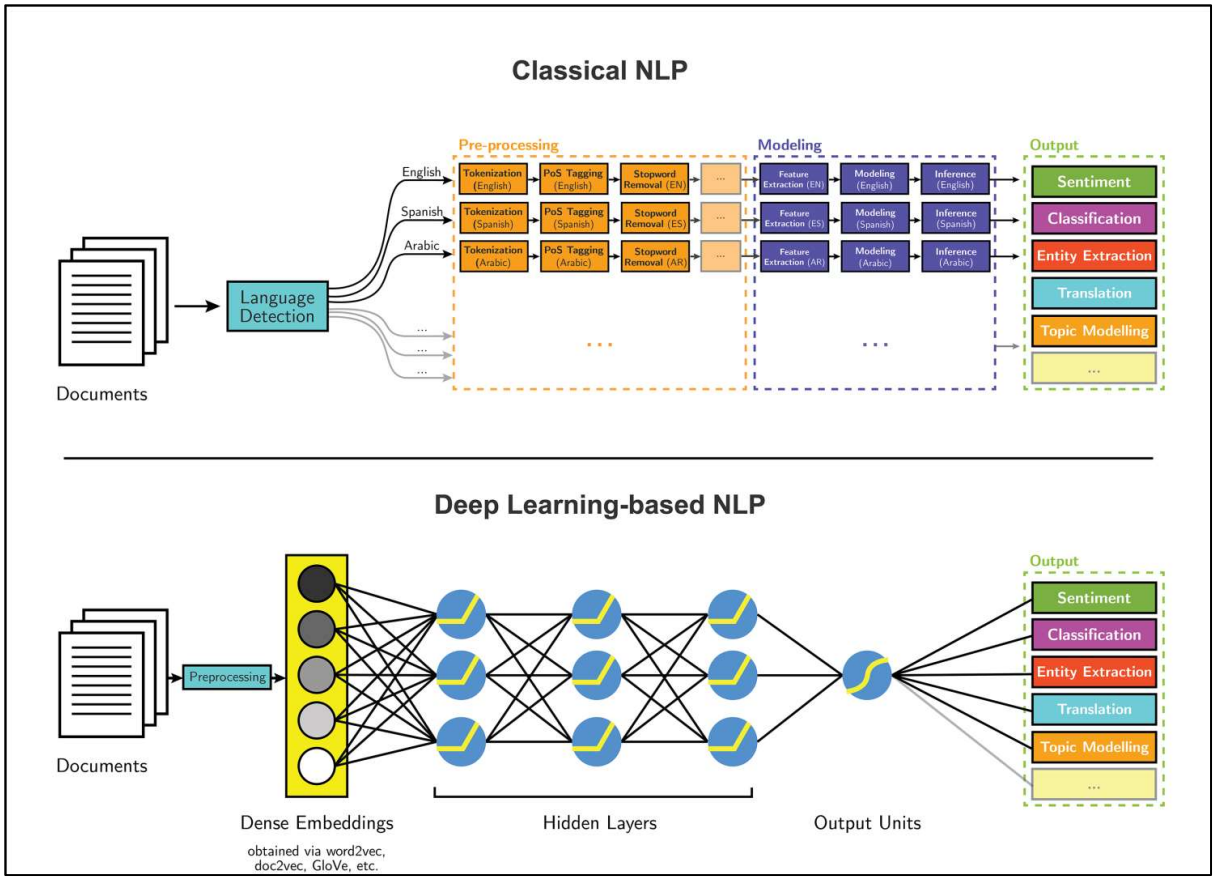
Моделите за анализа на чувствата може да обезбедат вредни информации и индикации за успешно справување со ризиците. Успешноста на анализата на овие вести зависи од јазикот на вестите, односно дали има означени податочни множества за одредениот јазик и област. За да го надминеме овој предизвик, во продолжение ќе се користат вести прибирани од агрегатори на вести од нашата земја и за истите ќе се применат комбинации на методи и класификатори за машинско учење.

Секој ден се објавуваат многу вести за компаниите кои директно влијаат на инвеститорите и клиентите. Мануелно читање на овие вест и лабелирање како позитивни или негативни е неизводливо поради големиот обем на генерирани вести што само се зголемува. Затоа е потребно автоматската анализа на чувството која ќе овозможи текстуалните информации да се трансформираат во нумеричка форма за да се има вредност во нивната примена во одлуките [92]. Технологијата за анализа на текстот односно чувствата овозможува информативна предност преку обработка на содржините на автоматски начин, пред корисникот да може да ги чита или сублимира сите нив [93]. Растот на финансиските текстови во експанзијата на големите податоци ги предизвика повеќето организации и донесе зголемени барања за алатки за анализа. Анализата на текст е потешка за анализа и извлекување знаење, во споредба со нумеричките податочни множества [94].

Во сè повеќе глобалниот интернет пазар, чувството на вестите наменети за една компанија не само што ќе ја влијаат истата туку може да ги влијаат и останатите компании од истиот сектор. Направена е ваква анализа користејќи ги финансиските вести од Reuters за период од 7 години за анализа на чувствата за 87 компании. Истражувањето покажа дека во одредени сектори чувствата на вестите може да ги влијае останатите сродни компании во финансиската мрежа [95]. Анализата на чувствата е користено за предвидување акциите на берза користејќи финансиски вести [96], курсна листа [97], како и за предвидување на добивка за компанијата [98].

За да се изгради точен модел за анализа на чувства постојат неколку пристапи. Некои пристапи се осврнуваат на овој проблем од поглед на процесирање на природен јазик, други од поглед на машинско учење, а во последно време како проблем со длабоко учење. Првиот пристап основан на процесирање на природен јазик, е да се создаде речник на познати негативни и позитивни зборови. За овој пристап потребни се само екстремни поларитети и зборови што можат правилно да се поврзат со поларитет. Врз основа на развиениот речник, чувството во реченицата се пресметува со едноставно броење на зборовите што се наоѓаат во речниците, потоа се класифицира поларитетот во зависност од тоа кој поларитет победил. Во истражувачката литература има голем број на истражувања и трудови со примена на надгледувани методи на класификација, како што се векторски поддржани машини [99], [100], наивен баесов [101] и дрво на одлуки [102] кои вршат анализа на чувства во повеќе истражувачки проекти. Овие техники за машинско учење користат модел со торба со зборови (bag of words) [103]. Во моделот со торба на зборови текст се претставува како множество на зборови, не земајќи го во предвид редоследот на зборовите во речениците. Меѓутоа токму редоследот на зборовите во реченицата може да го смени значењето и чувството на зборот.

Следниот пристап кој е базиран на машинско учење, создава големо податочно множество на податоци, кое содржи примери кои класифицирани рачно од страна на човек. Врз основа на класификацијата, може да се развие модел со машинско учење кој ќе овозможи автоматска класификација. Класификацијата може да биде со две класи (позитивни или негативни) или повеќе класи (пример од 1-5 за одредување на интензитет на чувството). Во архитектурата за големи податоци, моделот за машинско учење може да се користи класификација на содржина во реално време. Точноста може да биде поголема од 80% и со користење на едноставни алгоритми со правилно одбирање на атрибути и отстранување на шумот од податоците [104]. Последниот пристап заснован на длабоко учење, го анализа чувството со користење на векторска репрезентација на зборови (word embeddings), како што се word2vec [105], GloVe [106]. Моделот word2vec користи методи на длабоко учење за да произведе високо-димензионална векторска претстава за секој збор. Овој модел ја користи локацијата на релевантните зборови во реченицата за да ја најде семантичката врска помеѓу нив. За разлика од моделот bag of words, word2vec може да ја увиди сличноста на сентиметот меѓу зборовите. Векторска репрезентација на зборови се користи за да ги претстават зборовите како вектори. Со оваа техника сличните зборови се мапираат во соседни точки во континуиран векторски простор. Длабокото учење претставува подобрување во споредба со другите пристапи, затоа во дисертацијата се анализира и применува користењето на длабоко учење за анализа на сентиментот на насловите на финансиските и економските вести. Примените на длабоката невронска мрежа за процесирање на природните јазици се разликуваат во векторската репрезентација на зборови, бројот на скриени слоеви помеѓу влезот и излезот и бројот на излезни единици (Слика 38). Во традиционалните пристапи за машинско учење, одбирањето и инженерството на атрибути се дефинираат и извлекуваат рачно или со употреба на методи за селекција на атрибути. Ова не е случај кај моделите за длабоко учење кои автоматски ги извлекуваат атрибутите и со тоа постигнуваат подобра точност и перформанси.



Слика 38.Споредба на класичното машинско учење и процесите на длабоко учење за NLP [107]

Сентиментот може да се анализира на три нивоа и тоа на ниво на збор, реченица и текст. Моментално постојат три пристапи за решавање на проблемот со анализата на чувството [30]:

- **Техники основани на лексикон** - Овие техники се користеле првично за анализа на чувства. Се делат на две пристапи: базирани на речник и базирани на корпус. За овие техники не е потребно податочно множество за обука, меѓутоа во денешно време е голем предизвик да се одржува ваков речник и корпус.
- **Техники основани на машинско учење** – Овие техники може да се поделат на две групи: традиционални модели и модели за длабоко учење. Традиционалните модели се однесуваат на класичните техники за машинско учење, како што е наивен баесов и векторски поддржани машини. Точноста на овие модели зависи од избраните атрибути. Моделите основани на длабоко учење постигнуваат подобри резултати од традиционалните модели. Како најпознати примени се длабоките невронски мрежи, конволуциските невронски мрежи и рекурентните невронски мрежи.
- **Хибридни пристапи** – се комбинација на техниките базирани на лексикон и на техниките базирани на машинско учење.

6.1. Анализа на чувства со користење на модели за длабоко учење

Длабокото учење се појави во 2006 како дел од длабоката невронска мрежа [108]. Длабока невронска мрежа е невронска мрежа со повеќе од два слоеви, од кои некои се скриени слоеви. Овие невронски мрежи користат софистицирани математички моделирања со цел обработување на податоците на многу различни начини. Невронската мрежа е прилагодлив модел на излези како функции на влезови, кој се состои од неколку слоеви односно: влезен, скриени и излезен слој. Длабоките невронски мрежи може да се применат за надгледувано и ненадгледувано учење [109]. Како најпознати претставиле на длабокото учење се конволуциски невронски мрежи и рекурентните невронски мрежи. Невронските мрежи се многу корисни за: генерирање текст, векторска претстава, класификација на реченици, моделирање на реченици итн [110].

Моделите за длабоко учење [77] постигнуваат многу добри резултати кога се применуваат на компјутерско препознавање на говор, како и на процесирање на природни јазици. Употребата на вектори на зборови за да се претстават зборовите се нарекува векторска репрезентација. Идејата за векторска репрезентација на зборови не е нова меѓутоа стана популарна откако Google го објави претходно обучениот Word2Vec [111] алгоритам во 2013 година. Потоа се појавија и други векторски репрезентации на зборови како GloVe и fastText [112].

За дизајнирање модел за анализа на чувствата користејќи длабоко учење е потребно податочно множество кое ќе се обучи и тестира. Постојат неколку јавно достапни податочни множества за анализа на чувствата, меѓутоа ти се наменети за производи и за филмови. Моделите за соодветните индустрии постигнуваат добри перформанси [83], меѓутоа примената на овие модели во други индустрии и сектори е предизвик затоа што секој домен има уникатно множество на зборови за изразување. Банкарскиот домен се карактеризира со единствен вокабулар, кој има специфични поими карактеристични за анализа на чувствата за доменот.

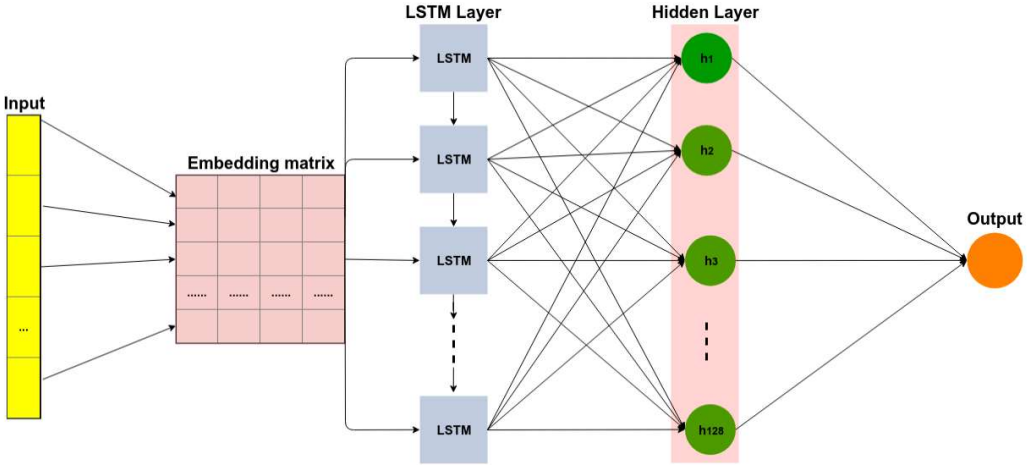
Методите за длабоко учење [113] користат каскада од повеќе слоеви на нелинеарни единици за обработка за извлекување на атрибути, потоа секој последователен слој го користи излезот од претходниот слој како влез. Ефикасноста на моделите за длабоко учење се докажува во многу студии и трудови користејќи рекурентна невронска мрежа (recurrence neural network) [114] и конволуциски невронски мрежи [81], [115] за откривање на чувството во финансиски текстуални содржини. Овој успех на примената на длабоко учење за процесирање на природни јазици се случува поради воведувањето и подобрувањето на методите за претставување на текст, односно кодери за зборови [79] и кодери за реченици [116]. Овие кодери ги конвертираат зборовите и речениците во векторска претстава. Моментално најдобри перформанси се постигнати во трудот [117] користејќи ги трансформерите BERT и RoBERTa за анализа на чувството.

6.1.1. Рекурентна невронска мрежа

Рекурентна невронска мрежа (Recurrent Neural Network) е класа на невронски мрежи чии врски помеѓу невроните формираат насочен циклус, што создава повратни

јамки во рамките на оваа мрежа. Рекурентна невронска мрежа е поопшта форма на ациклична невронска мрежа која има внатрешна меморија. Се вика рекурентна бидејќи ја извршува истата функција за секој влез на податоци, додека излезот на тековниот влез зависи од претходната пресметка. За разлика од ацикличните невронски мрежа, рекурентните невронски мрежи може да ја користат својата внатрешна меморија за да процесираат низа од влезни информации, што ги прави популарни за обработка на секвенцијални информации. Значењето на меморијата е дека рекурентната невронска мрежа ја извршува истата задача за секој елемент од низата, при што секој излез зависи од сите претходни пресметки, што претставува памтење на досега обработените информации. Рекурентната невронска мрежа претпоставува дека влезните податоци не се независни, односно знаењето на податоците од претходните повторувања ќе ја подобри точноста за предвидување на следниот збор во низата на зборови. Како недостатоци на овие мрежи се: градиентот на експлозија, градиентот кој исчезнува [71] и времетраењето на обука со овој алгоритам. Овие недостатоци се надминуваат со примена на меморија во кратки интервали на долги периоди (Long Short Term Memory - LSTM).

Главната разлика помеѓу рекурентната невронска мрежа и LSTM [118] е ќелијата со механизам на порта (gated cell). Ќелиите со механизам на порта му помагаат на LSTM да зачува повеќе информации во споредба со рекурентната невронска мрежа. Информациите можат да се чуваат, да се запишуваат или да се читаат од ќелија. Ќелиите одлучуваат дали да ги избришат или зачуваат информациите со отворање и затворање на портите. Ќелијата е составена од четири главни елементи: влезна порта, неврон со саморекурентна врска, порта за заборавање и излезна порта. Портата за заборавање е елемент што на клетката и овозможува да ја запамети или заборави својата претходна состојба. На Слика 39 е претставена архитектурата на LSTM. Влезните податоци се претходно обработени за вградување во матрицата (слично како во конволуциските невронски мрежи). Следниот слој е LSTM кој вклучува 200 ќелии. Последниот слој е целосно поврзан слој, кој вклучува 128 ќелии за класификација на текст. Последниот слој ја користи сигмоидната функција на активација за да го намали векторот со висина 128 до еден излезен вектор, со оглед на тоа што треба да се предвидат две класи односно позитивна и негативна.



Слика 39. Меморија во кратки интервали на долги периоди [119]

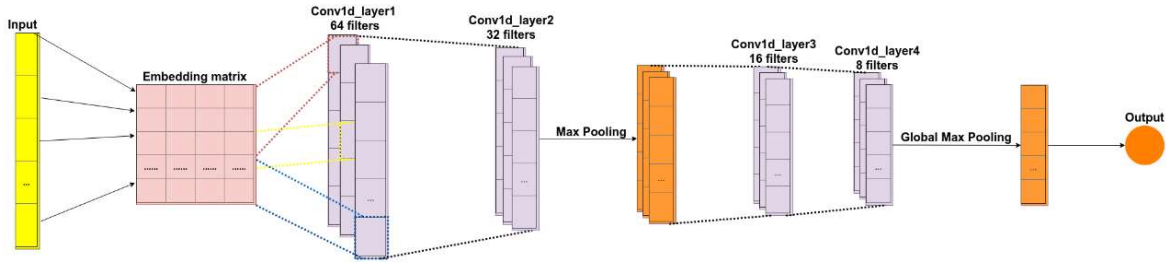
6.1.2. Конволуциски невронски мрежи

Еден од најчесто користените модели на длабоко учење е целосно поврзаната невронска мрежа. Иако целосно поврзаните невронски мрежи се сметаат за добро решение за класификација, сепак огромниот број на врски во овие мрежи може да доведат до проблеми. Овие проблеми се појавуваат уште повеќе при обработката на текстот, поради големиот број потребни неврони. Целосно поврзаните невронски мрежи ги третираат сите влезни зборови како да се блиску едни до други во реченица. Ова претставува уште поголем проблем кога се работи со големи податоци и често доведува до непосакувани резултати.

Конволуциските невронски мрежи (Convolutional neural network) [120] може да ги надминат горе споменатите проблеми и може да се применат за анализа на големи податоци. Секој неврон во првиот скриен слој наместо да се поврзе со сите влезни неврони се поврзува само со мал број од нив. Ова намалување на комплексноста на врските ги намалува и потенцијалните проблеми за процесирање на истата. Додека користењето на истите тежини за секој од скриените неврони овозможува да се открие истиот атрибут на различни локации во влезниот текст. Конволуциската невронска мрежа е посебен тип на ациклична невронска мрежа, која успешно се применува во областа на компјутерски вид, системи за препораки и процесирање на природни јазици.

Конволуциската невронска мрежа е тип на ациклична невронска мрежа и претставува архитектура на длабока невронска мрежа која се состои од слоеви за конволуција и слоеви за здружување. Конволуциските слоеви ги филтрираат нивните влезови за да извлечат атрибути и потоа излезите на повеќе филтри можат да се комбинираат. Влезниот слој ги зема необработените податоци и генерира векторска репрезентација за зборовите. Потоа, слоевите за извлекување на атрибутите кои вклучуваат слоеви на конволуција и здружување, ги учат релевантните атрибути. Конволуцискиот слој применува филтри познати како детектори на атрибути и произведува мапа на атрибути. Слојот на здружување, кој исто така е познат како метод на намалување на димензиите, се користи за извлекување на релевантните атрибути. На крај избраните атрибути се пренесуваат на слојот за класификација кој се состои од целосно поврзана мрежа со класификатор. Слоевите на здружување ја намалуваат резолуцијата на одликите, што може да ја зголемува отпорноста од шум и деформација.

На Слика 40 е претставена конволуциска невронска мрежа со влезна вградена матрица обработена од четири слоеви на конволуција и два слоеви за здружување. Првите два слоја на конволуција имаат 64 и 32 филтри, кои се користат за обука на различни атрибути, проследени од слој на здружување кој се користи за да се намали комплексноста на излезот и да се спречи претренирање на податоците. Третиот и четвртиот слој на конволуција имаат по 16 и 8 филтри, кои исто така се проследени од слој на здружување. Последниот слој е целосно поврзан слој што ќе го намали векторот со висина 8 до еден излезен вектор односно позитивна или негативна класа.



Слика 40. Конволуциска невронска мрежа [119]

Освен за препознавање на ентитетите и за одредување на сентиментот е користено библиотеката FLAIR, која е дизајнирана да олесни обука и дистрибуција на најсовремено лабелирање, класификација на текст и модели на јазици. Целта на FLAIR е на едноставен начин да претстави едноставен интерфејс за различни типови на векторска репрезентација за зборови и документи, криејќи ја на инженерите комплексноста на векторска репрезентација. За користење на FLAIR потребна е последна инсталација на Python. За експериментите во дисертацијата е користено Google Colab за извршување на експериментите на облакот на Google. Google Colab овозможува бесплатно користење на GPU поддршка за подобри перформанси, и има добра интеграција со Google Drive и со Github. GLoVe, Bert и другите техники за векторска репрезентација на зборови обезбедуваат импресивни резултати, сепак секогаш има простор за подобрување и тука FLAIR може да помогне. FLAIR вклучува тренирани модели за анализа на чувство, векторска репрезентација на текст и препознавање на ентитети.

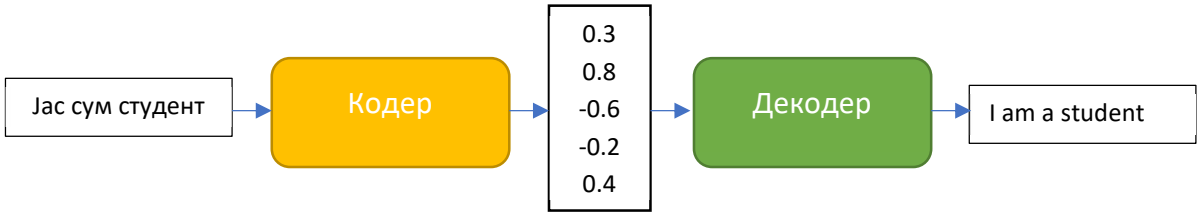
Класификација на текст со користење на векторска репрезентација на FLAIR се состои од следните чекори

1. Собирање на економските и финансиските вести од веб агрегатите на вести
2. Импортирање на податоците во Colab околината
3. Инсталирање на FLAIR
4. Подготовка на текстот за работа со FLAIR
5. Векторска репрезентација на зборови со FLAIR
6. Поделба на податоците во множество за обука и тестирање
7. Предвидување
8. Резултати

6.2. Трансформери

Претходно обучените векторски репрезентации на зборови и реченици како Word2Vec, GloVe и FastText покажуваат добри перформанси при процесирање на природни јазици. Сепак овие репрезентации за секој збор имаат фиксен вектор кој го генерираат секогаш за истиот збор, што во многу случаи не е соодветно и истиот збор може да има друго значење. Затоа најновите истражувања предложиле методи кои произведуваат различен вектор за истиот збор, во зависност од контекстот на зборот во реченицата. Моделите кои работат со секвенци (низи) [121] овозможија голем напредок

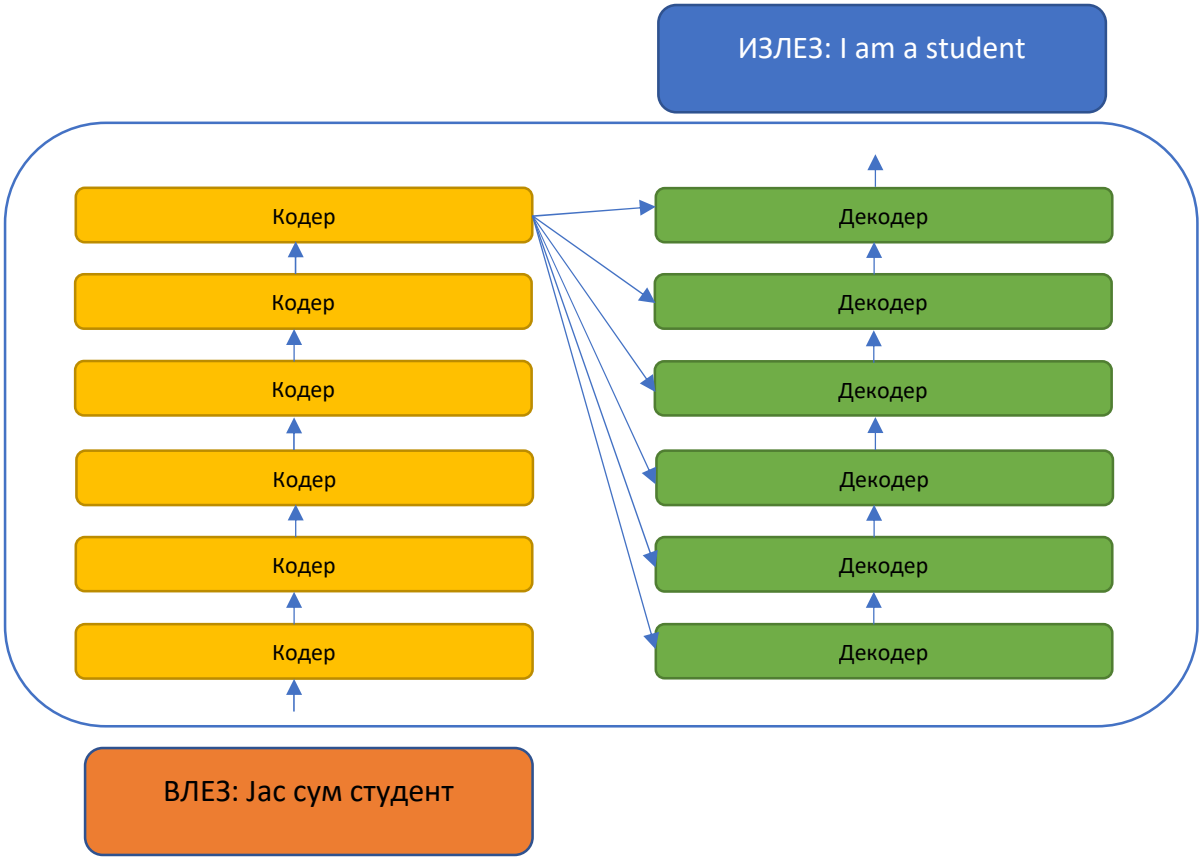
на времето, овозможувајќи кодирање на текст во одреден вектор и потоа декодирање на истиот во различни јазици (Слика 41). Овие модели најчесто се подобруваат поврзувајќи го кодерот и декодерот со механизам за внимание.



Слика 41. Кодирање и декодирање на секвенца

Трансформерите имаат различна архитектура за трансформација на секвенцата, користејќи кодер и декодер. Трансформер е модел за машинско учење кој е предложен во [69], кок поради паралелизацијата го надмина и моделот Google Neural Machine Translation за специфични задачи. Трансформерите претставуваат ново семејство на архитектура на нервонски мрежи, кои се создадени да ги надминат проблемите на конволуциските и рекурентните невронски мрежи и најчесто се користат за задачи со секвецна на секвенца односно машинско преведување и за креирање векторски репрезентации на зборови, како што е BERT кој креира векторска репрезентација за зборовите кои може да се користат за други задачи. Придобивка од трансформерите е тоа што може да се обучат брзо и дека механизмите за внимание го игнорираат редоследот и со тоа поедноставно ги откриваат релациите помеѓу оддалечени ставки во секвенцата. Сепак за да се користат трансформери треба голема процесирачка моќ и меморија за обучување, кое потоа генерира големи модели.

Оригиналниот труд ја препорача архитектурата на Слика 41 односно стек со шест идентични кодери и стек со шест идентични декодери. Секој кодер има слој за самостојно внимание и ацикличен слој. Во кодерот податоците прво поминуваат низ слојот за самостојно внимание кој помага да се анализираат другите зборови во реченицата и да се генерира соодветен вектор. Потоа излезот од слојот за самостојно внимание е како влез во ацикличната невронска мрежа. И декодерот ја ги има овие две слоеви, меѓутоа помеѓу нив има и слој за внимание што му помага на декодерот да се фокусира на релевантните делови од реченицата. Во првиот кодер се случува конвертирањето на зборовите во вектори со големина 512, потоа низата на вектори го проследува паралелно секој збор на наредниот кодер како влезен параметар.



Слика 42. Архитектура на Трансформери

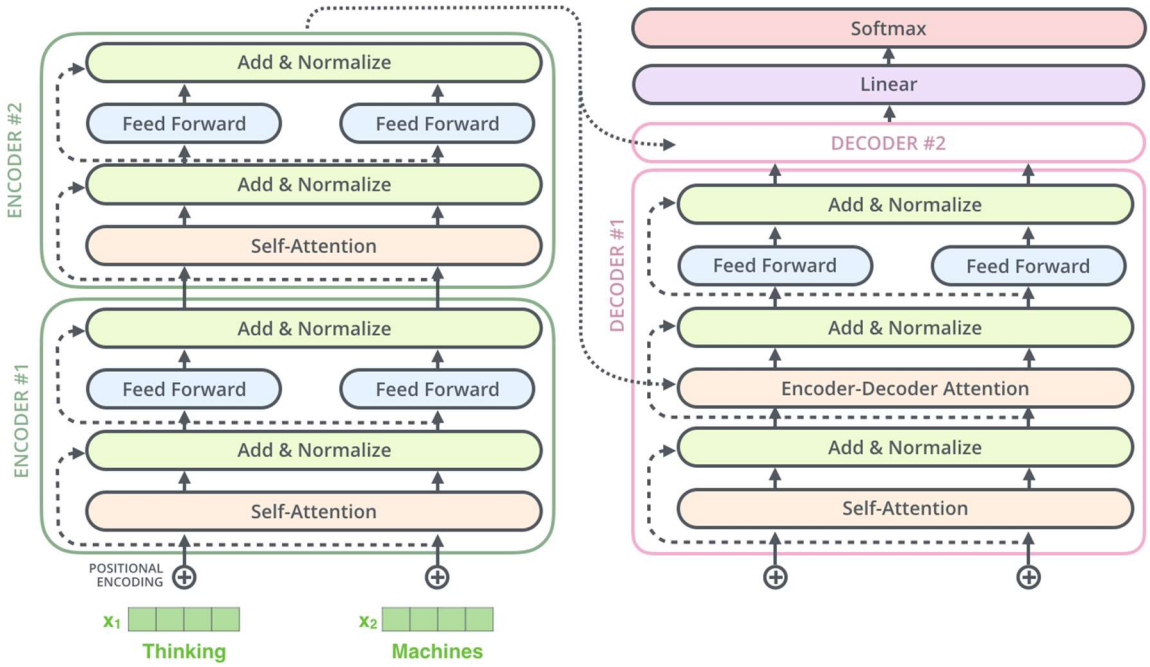
Самостојно вниманието овозможува да се разгледа соодветниот збор со другите позиции во влезната секвенца и да се најде најдоброто кодирање за зборот. Самостојно вниманието е механизам што трансформерот го користи за да го извлече најдоброто разбирање за зборот кој се процесира во тој момент. Трансформерот го пресметува своето внимание користејќи вектори со 64 димензии. За секој влезен збор создава вектор за пребарување q, вектор за клуч v и вектор за вредност v. Овие вектори заедно ги создаваат матриците Q, K и V. Овие матрици се создаваат со множење на векторските репрезентации на секој збор X со трите матрици W^q, W^k и W^v кои се иницијализираат и обучуваат при процесот на обука. За пресметување на самостојно вниманието се множи векторот за пребарување на зборот од интерес со векторот за клуч на другиот збор во реченицата, и истото се продолжува за секој останат збор во низата. Потоа резултатите се делат со 8 (коренот на 64 – димензиите на векторот за клуч) и со помош на softmax се нормализираат во опсегот 0 до 1. Нареден чекор е да се множи секој вектор со резултатот на softmax, со што ќе останат зборовите од интерес и ќе се елиминираат со мали вредности небитните зборови. Како последен чекор за соодветниот збор е да се сумираат векторите на тежина и со тоа се добива излезот од слојот за самостојно внимание кој се проследува на ачикличната невронска мрежа. Во реалност опишаната пресметка се прави со форма на матрица поради подобра обработка.

Перформансите на механизмот за самостојно внимание се подобруваат со користење на повеќе глави за самостојно внимание. Овој начин ја проширува способноста на моделот за да се фокусира на различни позиции и создава осум множества на наведените матрици. Секое од овие множества се иницијализира по

случаен избор и генерира различна векторска репрезентација за зборот. Бидејќи слојот за ациклична невронска мрежа очекува една матрица (вектор за зборот), затоа осумте матрици се спојуваат во една матрица.

За да се зема во предвид редоследот на зборовите во влезната секвенца, трансформерот додава вектор на секоја влезна векторска репрезентација. Овие вектори помагаат да се одреди позицијата на секој збор или растојанието помеѓу различните зборови во секвенцата.

Во секој кодер, излезот од слојот за самостојно внимание поминува низ слој за нормализација кој го намалува и времето за обучување. Потоа излезот од овој слој се внесува во слојот за ациклична невронска мрежа. Резултатите од слојот за ациклична невронска мрежа се собираат преку слој за нормализација и истото продолжува и во наредните кодери.



Слика 43. Детална архитектура на трансформери[14]

Излезот на последниот кодер се трансформира во множество на вектори за самостојно внимание K и V. Секој чекор во фазата на декодирање генерира по еден елемент од излезната секвенца. Овој процес се повторува се додека не се стигне до специјален симбол, и со тоа се препознава дека се стигнало до крај на секвенцата и тука декодерот го завршува својот излез. Слојот за самостојно внимание во декодерот е одобрен да присуствува само во претходните позиции во излезната секвенца. Ова се постигнува со маскирање на идните позиции пред чекорот за softmax. Слојот внимание за кодер-декодер (Encoder-Decoder Attention) работи на ист начин како кај самостојното внимание со повеќе глави. Стекот со декодери на крај како излез има вектор на реални броеви, кој мора да се конвертира во збор. Конвертирањето се прави со помош на Linear

[14] The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)

слојот, кој потоа се проследува со слојот softmax. Linear слојот претставува целосно поврзана невронска мрежа која ги проектира генерираните вектори во многу поголем вектор наречен logits вектор. Ако при обучување моделот препознал 1000 уникатни зборови, тогаш logits векторот ќе се состои од 1000 ќелии, и во секоја ќелија ќе се постави вредноста на секој збор. Потоа слојот softmax ги претвора тие резултати во веројатност, се избира ќелијата со најголема веројатност и како излез од овој чекор е соодветниот збор од ќелијата.

6.2.1. Bert multilingual

BERT е специфичен голем трансформер кој се состои од повеќе кодери кои генерираат векторска репрезентација на зборови [83]. Оригиналниот модел BERT може да се прилагоди користејќи дополнителни податоци, и со тоа се ажурираат тежините на оригиналниот модел и се постигнуваат подобри резултати за соодветна примена и област. Со тек на времето се создале и изведени архитектури со цел да се направи модел кој е помал и поефикасен за извршување (пример RoBERTa, DistilBERT, AlBERT). BERT може да се користи за конвертирање на текстот во вектори и како претходно обучен модел. Сепак BERT е на англиски јазик, и за потребата да се користи BERT за друг јазик ни треба претходно тренирана верзија на BERT за соодветниот јазик. За повеќето јазици иако постои соодветна верзија, сепак не е толку точна колку што е BERT за англискиот јазик и не ги добиваме очекуваните перформанси. Друг начин за да се искористи оригиналниот BERT е друг јазик да се преведе со помош на машинско учење на англиски јазик и потоа да се примени BERT.

Повеќе јазичните модели се модели за машинско учење кои можат да разберат повеќе јазици [3]. Пример за повеќејазичен модел е mBERT (Multilingual BERT) создаден од Google [3]. mBERT наместо само со англиски е обучен со сите јазици, без да му се каже од кој јазик се податоците. Првата верзија на mBERT се појави после еден месец, откако се појави оригиналниот BERT на англиски јазик. Овој модел поддржува и разбира 104 јазици[15], додека генералните (еднојазични) модели можат да разберат само еден јазик. Повеќе јазичните модели веќе постигнале добри резултати за одредени задачи. Покрај придобивките, овие модели се поголеми, им требаат повеќе податоци и време за обучување. Голема придобивка од повеќе јазичните модели е можноста да се тренира моделот со друг јазик, и потоа да се тестира и примени на друг јазик. Според резултатите во [3] најдобри перформанси при тестирање се постигнуваат кога mBERT ќе се обучува дополнително и со податоци од соодветниот јазик.

6.2.2. XLM-RoBERTa

RoBERTa е воведен од Facebook и претставува оптимизирана верзија на BERT која е обучена од ново со подобрена методологија, со 100% повеќе податоци и пресметувачка моќ. За подобрување на методологијата на обучување, RoBERTa ја отстранува задачата за предвидување на следната реченица (Next Sentence Prediction) и воведува динамично маскирање. RoBERTa користи 160 GB податоци за претходно обучување дополнувајќи ги податоците на BERT со база на податоци на CommonCrawl

[15] bert/multilingual.md at master · google-research/bert · GitHub

News (63 милиони статии, 76 GB), корпус на веб-текст (38 GB) и Stories од Common Crawl (31 GB). Овој модел постигнува подобри перформанси од BERT за 2-20%.

Моделот XLM-RoBERTa (CrossX-Lingual-Model RoBERTa) [4] е повеќејазичен модел обучен на сто различни јазици со користење 2,5 тера бајти филтрирани податоци и се базира на моделот RoBERTa на Facebook. XLM-R постигнува солидни перформанси при меѓу јазични задачи за пренос, вклучително и класификација на текст. Тоа што го прави привлечен овој модел е дека XLM-RoBERTa нуди можност за повеќејазично моделирање без намалување на перформансите. Постигнува подобри резултати од mBERT за околу 14,6%. Повеќе јазичните модели се многу моќни. XLM-RoBERTa поддржува 100 јазици, и сепак со перформансите е конкурентен на моделите за еден јазик.

6.3. Анализа на резултатите

Класификација на текст претставува метод за надгледувано машинсло учење кој се користи за класификација на реченициите во една од дадените класи. Кога станува збор за сентиментот тоа најчесто се претставува со две класи односно позитивна и негативна класа. Во дисертацијата е користено Flair и HuggingFace[16] поради докажаната ефикасност и предност над другите библиотеки и решенија. За да се обучи сопствен класификатор, треба да имаме наше податочно множество во кое ќе бидат означени сите реченици во соодветната класа. Во нашиот случај имавме 540 наслови на финансиски и економски вести на македонски јазик, 540 на албански јазик, и 540 на англиски јазик. Сите податочни множества содржат еднаков број на позитивни и негативни вести. За Flair форматот на податочното множество треба да биде според форматот FastText, односно во секоја линија ќе има ознака за сентиментот (__label__pos или __label__neg) и потоа со таб е разделена соодветната реченица. Во користеното податочно множества нема дупликати и истите се претходно избришани. За обука на моделот со mBERT се користени различни комбинации на напластени вектори користејќи mBERT и векторските репрезентации на Flair. Во зависност од податоците за обучување и од бројот на епохи, обучувањето на моделот траеше од 15 минути до 3 часа на Google Colaboratory користејќи GPU.

За обучување на моделите се користени различни комбинации на податочни множества, односно:

- Множество за обучување: македонски (440 вести); множество за тестирање: македонски (50 вести); множество за валидација: македонски (50 вести);
- Множество за обучување: албански (440 вести); множество за тестирање: албански (50 вести); множество за валидација: албански (50 вести);
- Множество за обучување: македонски+албански (880 вести); множество за тестирање: македонски+албански (100 вести); множество за валидација: македонски+албански (100 вести);

[16] https://huggingface.co/

- Множество за обучување: македонски+албански+англиски (1320 вести); множество за тестирање: македонски+албански (100 вести); множество за валидација: македонски+албански (100 вести);

Со mBERT најдобри резултати се добиени користејќи ги следните векторски репрезентации:

- mBERT ('bert-base-multilingual-cased')
- BytePairEmbeddings ('multi')
- FlairEmbeddings ('multi-forward')
- FlairEmbeddings ('multi-backward').

BytePairEmbeddings е векторска репрезентација за под зборови, додека FlairEmbeddings претставуваат контекстуални векторски репрезентации. Тестирани се и другите комбинации, меѓутоа сите резултати беа послави споредено со наведената комбинација. Во Табела 12 се претставени резултатите од користење на различни податочни множества за обучување, тестирање и валидација. Според прикажаните резултати, најдобри резултати за обучување на моделот се постигнати кога се користи податочно множество за обучување од 1320 вести, по 440 вести на македонски, албански и англиски јазик. Потоа за тестирање и валидација се користат само македонските и албанските вести. Бројот на епохи е поставен до 200, меѓутоа резултатите најчесто се појавуваа до околу епоха 70. На Слика 44 е прикажано испитување на сентиментот за неколку вести со зачуваниот модел добиен со користење на mBERT. На сликата освен предвидување на сентиментот (позитивен/негативе) се прикажува и соодветната точност на предвидувањето.

```
from flair.models import TextClassifier
from flair.data import Sentence

classifier = TextClassifier.load('./resources/taggers/sentiment_ml/final-model.pt')


sentence = Sentence('Алкалоид - резултати од работењето за периодот јануари - септември 2019 година')

classifier.predict(sentence)

print(sentence.labels)
```

```
2021-05-30 20:27:35,391 loading file ./resources/taggers/sentiment_ml/final-model.pt
[pos (0.9442)]
```

Слика 44. Тестирање со зачуваниот модел

Табела 12. Импактот на различните податочни множества за F1 резултатот на одредување на сентиментот со Multilingual BERT

Обучување / тест / валидација	F1 - Резултат
440 МК вести / 50 МК вести / 50 МК вести	0.63
440 АЛБ вести / 50 АЛБ вести / 50 АЛБ вести	0.62
880 МК+АЛБ вести / 100 МК+АЛБ вести / 100МК+АЛБ вести	0.66
1320 МК+АЛБ+АНГ / 100 МК+АЛБ / 100 МК+АЛБ	0.78

Истите податочни множества се користени и за обучување на моделот со користење на XLM-RoBERTa. За овој модел подобри резултати се постигнати користејќи го трансформерот на HuggingFace за XLM-RoBERTa. Како што е опишано во Глава 6.2.2, XLM-RoBERTa постигнува подобри резултати од BERT и во нашите експерименти. Дообученият модел со вестите на македонски и албански јазик, постигна подобри резултати отколку mBERT (Табела 13). И за овој модел, големината на корпусот за обучување значително влијае на резултатите на предвидување. За сите наведени експерименти во Табела 13 се користени по 10 епохи. Во двата модели, пероформансите на моделот за специфичниот јазик, може се подобруваат ако се внесат податоци и на англиски јазик.

Табела 13. Импактот на различните податочни множества за F1 резултатот на одредување на сентиментот со XLM Roberta

Обучување / валидација	F1 - Резултат
495 МК вести / 45 МК вести	0.76
495 АЛБ вести / 45 АЛБ вести	0.74
990 МК+АЛБ вести / 90 МК+АЛБ вести	0.80
1455 МК+АЛБ+АНГ / 90 МК+АЛБ	0.90

На Слика 45 е претставено загубата според стапка на обучување и валидација на најдобриот модел од Табела 13, при 10 епохи. Од сликата се гледа дека за обучување и добивање задоволителни резултати се доволни само 4-5 епохи.

Слика 45. Загуба споредено со стапка на обучување и валидација

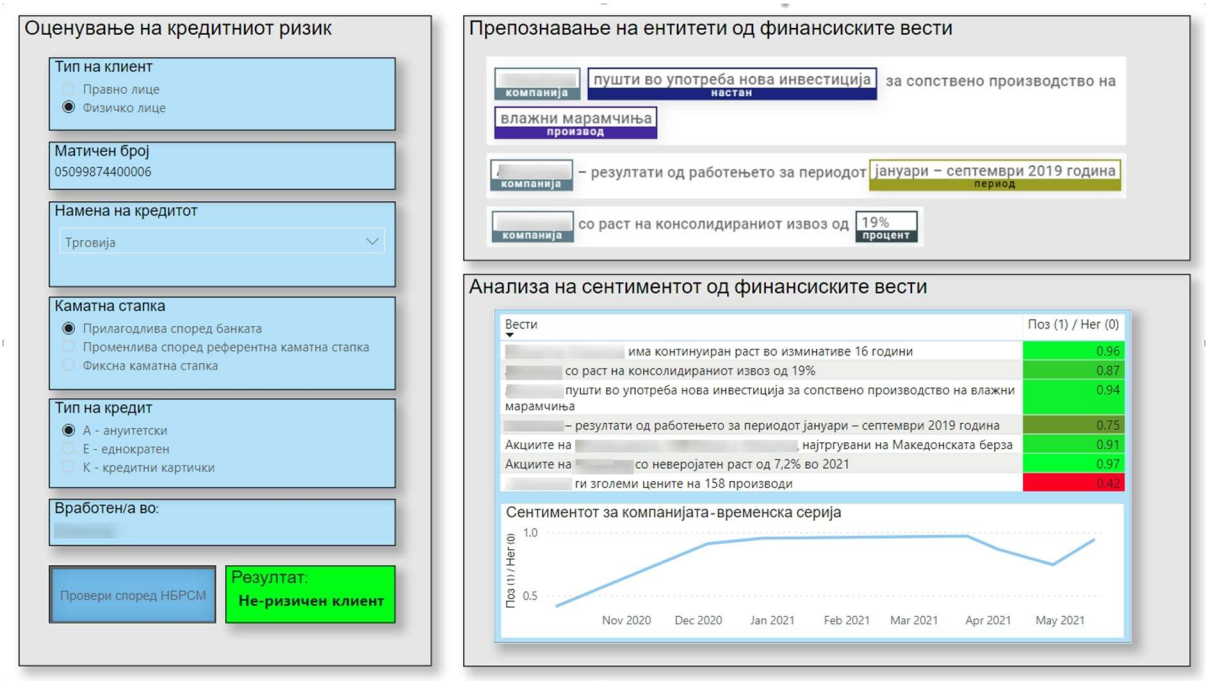
Опишаната методологија може да се користи многу лесно и за друг јазик, и да се постигнат подобри резултати за соодветниот јазик со користење на наменети модели наместо генералните и англиските јавни модели кои се претходно обучени.

7. Интеграција во Power BI платформа

Реалната примена на досега опишаните експерименти, модели и знаења е применето во платформата која претставува апстракција над големите податоци и машинското учење применето во претходните глави. Платформата претставува софтверско решение кое е спремно за користење во продукциска околина.

Бизнис секторот особено економистите се подготвени да користат готови алатки односно алатки кои ќе им овозможат манипулација и прикажување на податоците во повеќе форми. Овие алатки ќе им овозможат да бидат независни од програмерите и сами да можат детално да ги анализираат и визуелизираат податоците со модерни и богати прикази. Како што е опишано во Глава 3, алатката Power BI е моќна алатка за визуелизација и за анализирање на податочни множества. Поради голема сличност со алатките на Microsoft Office, сите економисти лесно може да ја користат истата постигнувајќи напредна анализа

Со цел сите истражувања и експерименти од дисертацијата да се направат во лесен формат за користење особено за економистите, е избрано алатката Power BI. Во ова алатка освен различните извештаи од податочните множества е изградена и платформа која ги спојува: предвидувањето на кредитниот ризик со моделот добиен од Кредитниот регистар, препознавањето на ентитети во финансиските вести и анализата на сентиментот од истите. Алатката може да се користи при секој дневно работење за истражување и анализа на било какви извори на податоци. Поради таа цел во ова алатка е направена и платформата која ќе им служи при секојдневна работа како дополнителна помош со повеќе вредни информации за кредитниот ризик и за анализа на соодветните вести.



Слика 46. Power BI платформа

Како што е прикажано на Слика 46, економистот (корисникот) потребно е да избере тип на клиент, матичен број за (лице или компанија), намена на кредит, тип на каматна стапка, тип на кредит и да внесе компанија во која е вработен клиентот. По избирање на копчето провери според НБРСМ податоците се праќаат на моделот добиен во Глава 4 и се добива соодветното предвидување за ризичноста на клиентот. На десната страна автоматски се појавуваат сите вести за наведената компанија собрани од повеќе агрегати на вести, обележувајќи ентитетите во прикажаните вести, и анализа на сентиментот за истите вести. За секоја вест се прикажува сентиментот обележан со соодветна боја односно зелен за не-ризичен клиент и црвена боја за ризичен клиент. За истата компанија е претставено и визуелизација за сентиментот на прикажаните вести според временска серија. Овој приказ овозможува полесно следење како се движело сентиментот на вестите за компанијата и дали има пад или подобрување на истиот. Сите овие визуелизации овозможуваат за било кој корисник на апликацијата во период од неколку секунди да се препознае ризичноста односно неризичноста на соодветниот клиент. Визуелниот приказ обезбедува лесно препознавање на ризичноста на клиентот и обезбедува информирани одлуки за корисниците на платформата користејќи знаење од голем обем на податоци. Приказите од платформата може да се интегрираат и во надворешни апликации, што ја олеснува интеграцијата и користењето во постоечките апликации.

Платформата овозможува апстракција на сите наведени истражувања во дисертацијата, меѓутоа најбитно е дека придобивките на овие истражувања на лесен начин може да се користат од било кој и со тоа да се управува подобро со кредитниот ризик, помогнато од знаењето на Кредитниот регистар и од финансиските вести. Дополнително, платформата може де се користи со податочни множества и соодветни вести од друга област, и да се искористат придобивките на истата само со промена на изворите и препорачливо да се обучат нови модели за соодветната област.

Платформата може да се користи за оценување на физички и правни лица. За правни лица полето вработен/а во не се појавува. Оваа платформа ќе им овозможи на комерцијалните банки, централно место кое ги спојува придобивките од големите податоци користејќи модерни истражувања во едно место. Со новите знаења и информации од платформата ќе се овозможат поинформирани одлуки, помал кредитен ризик и поефикасно услужување на клиентите. Истата може да се користи и од брокери за испитување на акциите на компаниите преку следење на соодветните вести за истите, што ќе им помогне за повеќе придобивки и поуспешно тргување со акциите.

8. Заклучоци и идна работа

Растот на податоците односно големите податоци претставуваат нова ера во финансискиот сектор. Обемот на податоците вклучува и нови информации односно знаења кои ги помагаат одлуките и претставуваат дополнително знаење за деловните процеси. Успешната примена на науката на податоците врз големите податоци ги издвојува успешните компании на пазарот. Банкарскиот сектор има огромни количини на податоци за клиентите кои растат експоненцијално (на пр. депозити, уплати, исплати, трансакции преку Интернет, податоци за клиенти итн.). Банките ги имплементираат придобивките од обемот на собраните податоци и ги следат трендовите на дигиталната револуција за да обезбедат подобри услуги за своите клиенти.

Во Глава 2 е опишано примената на големите податоци во банкарскиот сектор. Увидено е дека комерцијалните банки се чекор напред со користењето на придобивките на големите податоци, подобрувајќи ги деловните процеси и намалувајќи ги различните ризици (кредитен, ликвидносен, монетарен). При примената на големи податоци во централните банки посебен акцент е ставен на предизвиците и примената за кредитниот ризик. Во продолжение дел на ова глава е воведена проблематиката за кредитен ризик, односно предвидување на истиот и споредување на референти трудови. Во последниот дел се воведува базата Кредитен регистар која е и главниот извор на податоци во дисертацијата.

Во Глава 3 е направена напредна анализа на големо податочно множество, со користење на алатка за кое не е потребно познавање на ИТ вештини. Алатката не ефикасен начин овозможува анализирање, и детално визуелизирање на податоците, запознавање со нив, откривање трендови, екстремни вредности и интегрирање на извештаите во надворешни апликации. Во анализата се вклучени и други дополнителни извори на податоци за да се увиди корелацијата со други случувања во финансискиот свет. Дизајниран е и модел во вид на ѕвезда шема и сите изворни табели се споени со главната табела за кредити. Податоците се визуелизирани во различни извештаи и прикази и со овој дел е постигнато првично разбирање и визуелно претставување на податоците од базата Кредитен регистар.

Во Глава 4 се анализарани и споредени научни трудови од последните години и како анализата на кредитен ризик еволуирала низ годините. Постоечките трудови се однесуваат на кредитниот ризик користејќи податоци само од комерцијаланите банки. Во оваа глава е предложена детална методологија за предвидување на кредитниот ризик со користење на податоците на Кредитниот регистар. Уникатноста на овој дел е дека досега нема некое истражување кое го предвидува кредитниот ризик со помош на оваа податочно множество. По деталната методологија и чекорите за пред-обработка на податоците, направено е одбирањето на атрибутите по што се применети пет различни алгоритми за машинско учење. Според резултатите најдобар модел избран според F1 резултатот се дрвото на одлуки, случајната шума и логистичката регресија. Резултатите покажуваат резултати со највисока точност се добиваат со небалансираното и скалираното податочно множество.

Во Глава 5 е опишана обработката на текстуалните податоци со цел да може да се користат за машинско учење и да се препознаваат ентитети во македонските и албанските вести. Споредени се модели со: меморија во кратки интервали на долги периоди – LSTM, двонасочна меморија во кратки интервали на долги периоди - Bi-LSTM и со условно случајно поле - CRF. Од големо значење е векторска репрезентација на зборовите, односно контекстуалната векторска репрезентација и векторска репрезентација со користење на BERT моделот. По примена на детална методологија, за препознавање ентитети со 13 лабели, за вестите на македонски јазик е постигнат F1 резултат 0.75. Резултатот не е многу висок, меѓутоа е задоволителен за мало податочно множество. Увидено е дека обемот на податоци, дури и дополнување на податоци од англиски јазик, помага за постигнување подобар резултат.

Во Глава 6 посебен акцент имаат трансфомерите и нивната архитектура. Во практичниот дел е применет Bert multilingual и XLM-RoBERTa за вестите на македонски и албански јазик. За овој модел подобри резултати се постигнати користејќи го трансформерот на HuggingFace за XLM-RoBERTa односно 0.90 F1 резултат.

Глава 7 го претставува крајниот производ од оваа дисертација, односно платформата во која се интегрирани сите претходно наведени теоретски и практични делови. Платформата како краен производ е интеграција на сите наведени придобивки од големите податоци со цел банкарскиот сектор да може да ги искористи придобивките во реалност. Платформата како централно место ги спојува предвидувањето на кредитниот ризик со моделот добиен од Кредитниот регистар, препознавањето на ентитети во финансиските вести и анализата на сентиментот од истите. Платформата овозможува апстракција на сите наведени истражувања во дисертацијата, меѓутоа најбитно е дека придобивките на овие истражувања на лесен начин може да се користат од било кој и со тоа да се управува подобро со кредитниот ризик, помогнато од знаењето на Кредитниот регистар и од финансиските вести. Опишано е како платформата може да се користи за оценување на физички и правни лица.

Во дисертацијата е направено напредна анализа на големата база Кредитен регистар и истата е визуелизирана преку модерни извештаи и прикази направени во Power BI. Извештаите се интерактивни и секој економист кој има знаење за алатките на Microsoft Office може лесно да манипулира со извештаите. Потоа извештаите се интегрирани и во надворешни веб апликации што укажува дека економистите може сами да креираат извештаи според нивните потреби и потоа без помош на програмери. Напредни извештаи вклучуваат и мапа на градовите со соодветниот број на кредити, предвидување со помош на линеарна регресија и истражување на корелации и влијание за предвидување на соодветни атрибути. Дисертацијата овозможи и помоќна и полесна анализа на големи податоци во банките, со тоа процесот на напредно анализирање е поефикасен и ќе овозможи одлуки кои ќе се во реално време според големо множество на податоци.

Примарна цел на дисертацијата беше да се помогне оценувањето на кредитниот ризик со нов пристап, бидејќи сите истражувања кои моментално постојат се ограничени само на податоци од некои јавни податочни множества или податочни множества кои

постојат само во комерцијалните банки. Најголема причина за ова реалност е невозможноста за пристап до други бази, односно базата Кредитен регистар која ја има само во централните банки. Претставената методологија и резултати се дополнителен извор и во истражувачката и во реалната примена за кредитниот ризик. Со применетите алгоритми е увидено дека со помош на истите кои се користеле во многу истражувања, се постигнуваат подобри резултати со податоците на Кредитен регистар. Примената на методологијата е направена во голема релациона база и скриптите на R јазикот се вградени во сторирани процедури што овозможи да се искористат постоечките ресурси и алатки, што допринесува за полесна имплементација на цело оваа истражување.

Во истражувачката литература е докажано дека особено вестите влијаат во финансискиот сектор. Поради ова причина избрани се финансиските вести поврзувајќи ги со соодветните компании и други ентитети од финансискиот свет. Досега постојат истражувања особено на англискиот јазик и на некои најпознати јазици, меѓутоа во дисертацијата придобивките од процесирањето на природните јазици се овозможени за помалку користени јазици односно македонскиот и албанскиот јазик. Целта беше дисертацијата да има придонес за нашиот регион и да помогне оценување на кредитниот ризик. Овој дел има директен придонес за компаниите бидејќи за нив има повеќе вести. Во платформата е направен приказ основан на временска серија кој го визуелизира сентиментот во соодветниот ден на објава на вестите и со тоа може да се увиди во каква насока се движи компанијата според јавното мислење односно вестите. Овие примени ќе поттикнат иновативност и користење и во други области во нашиот регион поради иновативноста и придобивките кои ги нудат.

Како резултат од дисертацијата е имплементирана и платформа која на банките ќе им овозможи помал ризик при нивното работење со кредитниот ризик поради дополнителното мислење кое ќе го добијат од моделот на централната банка и информациите од финансиските вести. Моделот ќе биде од особено значење за комерцијалните банки за да не претрпат загуби, а за централната банка ќе значи повеќе стабилност во банкарскиот систем. Во платформата голем придонес имаат и моделите за процесирање на природните јазици односно детектирање на сентиментот на финансиските вести и препознавање на ентитетите во истите вести. Приказот овозможува сублимат на повеќе придобивки и визуелизирани информации кои лесно може се разберат и применат за банкари, економисти и брокери. Делот за процесирање на природни јазици може да се нуди и јавно бидејќи сите податоци се од јавно публикувани вести, и поставувајќи филтер за ентитетите е вредна алатка за анализа и истражување на соодветните ентитети и нивниот сентимент.

Сите горе наведени примени ќе може лесно да се уредат и да се применат и во други големи сектори кои ќе ги искористат придобивките на големите податоци при нивното работење. Платформата исто така претставува и место каде се соединуваат придобивките од машинското учење и процесирањето на природни јазици односно на македонскиот и албанскиот јазик. Примената на платформата е наменета за ризиците на кредитниот ризик, меѓутоа истата платформа може лесно да се прилагоди и за различни

ризици и цели. Со тоа претставеното решение претставува потенцијал кој ќе овозможи придобивки за повеќе области.

Како идна работа останува да се обезбеди пристап на податочно множество кое ќе се спои со Кредитниот ризик и да се провери дали ќе се добијат подобри знаења за предвидување на кредитниот ризик. Готовиот модел од дисертацијата може да се користи и од комерцијалните банки, меѓутоа како дел за истражување останува да се до обучува моделот со податоци од комерцијални банки кои би се однесувале за клиентите и да се увиди дали може да се подобри истиот модел. И покрај напорите за да се најде корелација на движењата на кредитниот ризик со соодветни јавни финансиски настани и инструменти, сепак не е најдена таква корелација. Останува да се истражува овој дел и да се проба и со други финансиски настани од светски и домашен опсег. Моделите за анализа на вестите на македонски јазик, може да се подготват со податоци за различни сектори и области со цел да се искористат и да се обезбедат придобивки и за другите области користејќи модели специјализирани за одреден јазик. Од клучна вредност е зголемувањето на податочното множество кое помага за подобрување на прецизноста на моделите за препознавање ентитети и одредување сентимент. За овој дел останува да се увиди ефикасноста на моделот за препознавање ентитети за други области. Предизвик е и изборот на соодветни лабели за одредените области, и ова е тема која треба да се истражува.

Дисертацијата претстави нови и решенија кои не се увидени во истражувачкиот свет, и со тоа придонесе не само за економистите туку и за истражувачкото поле вклучувајќи машинско учење, големи податоци, наука за податоците и препознавање на ентитети.

9. Референци

[1] J. Roeder, M. Palmer, and J. Muntermann, "Utilizing News Topics for Credit Risk Management: The Explanation of Bank CDS Spreads," *J. Decis. Syst.*, pp. 1–13, 2020.

[2] F.-T. Tsai, H.-M. Lu, and M.-W. Hung, "The impact of news articles and corporate disclosure on credit risk valuation," *J. Bank. Finance*, vol. 68, pp. 100–116, 2016.

[3] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?," *ArXiv Prepr. ArXiv190601502*, 2019.

[4] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," *ArXiv Prepr. ArXiv191102116*, 2019.

[5] B. Fang and P. Zhang, "Big data in finance," in *Big data concepts, theories, and applications*, Springer, 2016, pp. 391–412.

[6] J. Anuradha, "A brief introduction on Big Data 5Vs characteristics and Hadoop technology," *Procedia Comput. Sci.*, vol. 48, pp. 319–324, 2015.

[7] Bank for International Settlements and Irving Fisher Committee on Central Bank Statistics, *IFC report: Central banks' use of and interest in "big data" : 2015 Survey conducted by the Irving Fisher Committee on Central Bank Statistics (IFC)*. Basel: Bank for International Settlements, 2015.

[8] "Big data in central banks: 2017 survey," *Central Banking*, Nov. 06, 2017. http://prod.centralbanking.bb8.incinsight.net/node/3315546 (accessed Mar. 28, 2021).

[9] "2018 IFC Annual Report," p. 14, 2018.

[10] "Big data in central banks: 2018 survey results," *Central Banking*, Aug. 02, 2018. https://www.centralbanking.com/node/3661931 (accessed Mar. 28, 2021).

[11] S. J. Shiv, S. Murthy, and K. Challuru, "Credit Risk Analysis Using Machine Learning Techniques," in *2018 Fourteenth International Conference on Information Processing (ICINPRO)*, 2018, pp. 1–5.

[12] C. Onay and E. Öztürk, "A review of credit scoring research in the age of Big Data," *J. Financ. Regul. Compliance*, 2018.

[13] O. M. Araz, T.-M. Choi, D. L. Olson, and F. S. Salman, "Role of analytics for operational risk management in the era of big data," *Decis. Sci.*, vol. 51, no. 6, pp. 1320–1346, 2020.

[14] S. Lakshmi and S. D. Kavilla, "Machine learning for credit card fraud detection system," *Int. J. Appl. Eng. Res.*, vol. 13, no. 24 Pt. 1, pp. 16819–16824, 2018.

[15] X. Qin, "Making use of the big data: next generation of algorithm trading," in *International Conference on Artificial Intelligence and Computational Intelligence*, 2012, pp. 34–41.

[16] M. Pejić Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," *Sustainability*, vol. 11, no. 5, p. 1277, 2019.

[17] A. Chandani, M. Mehta, B. Neeraja, and O. Prakash, "Banking on Big Data: A case study," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 5, pp. 2066–2069, 2015.

[18] D. Bholat, "Big data and central banks," *Big Data Soc.*, vol. 2, no. 1, p. 2053951715579469, 2015.

[19] A. Jaiswal and P. Bagale, "A Survey on Big Data in Financial Sector," in *2017 International Conference on Networking and Network Applications (NaNA)*, 2017, pp. 337–340.

[20] R. Elshawi, S. Sakr, D. Talia, and P. Trunfio, "Big data systems meet machine learning challenges: towards big data science as a service," *Big Data Res.*, vol. 14, pp. 1–11, 2018.

[21] I.-Y. Song and Y. Zhu, “Big data and data science: what should we teach?,” *Expert Syst.*, vol. 33, no. 4, pp. 364–373, 2016.
[22] A. Byanjankar, M. Heikkilä, and J. Mezei, “Predicting credit risk in peer-to-peer lending: A neural network approach,” in *2015 IEEE Symposium Series on Computational Intelligence*, 2015, pp. 719–725.
[23] M. Sudhakar and C. V. K. Reddy, “Two step credit risk assessment model for retail bank loan applications using Decision Tree data mining technique,” *Int. J. Adv. Res. Comput. Eng. Technol. IJARCET*, vol. 5, no. 3, pp. 705–718, 2016.
[24] M. Ala’raj and M. F. Abbod, “Classifiers consensus system approach for credit scoring,” *Knowl.-Based Syst.*, vol. 104, pp. 89–105, 2016.
[25] E. Tobback and D. Martens, “Retail credit scoring using fine-grained payment data,” *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 182, no. 4, pp. 1227–1246, 2019.
[26] “(PDF) Credit Approval Analysis using R.” https://www.researchgate.net/publication/321002603_Credit_Approval_Analysis_using_R (accessed Mar. 28, 2021).
[27] J. Lohokare, R. Dani, and S. Sontakke, “Automated data collection for credit score calculation based on financial transactions and social media,” in *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, 2017, pp. 134–138.
[28] S. M. Ali, N. Gupta, G. K. Nayak, and R. K. Lenka, “Big data visualization: Tools and challenges,” in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 2016, pp. 656–660.
[29] C. Cantú, S. Claessens, and L. Gambacorta, “How do bank-specific characteristics affect lending? New evidence based on credit registry data from Latin America,” *J. Bank. Finance*, p. 105818, 2020.
[30] A. Dzelihodzic and D. Donko, “Data Mining Techniques for Credit Risk Assessment Task,” *Recent Adv. Comput. Sci. Appl.*, vol. 6, 2013.
[31] L. Rokach and O. Maimon, “Decision trees,” in *Data mining and knowledge discovery handbook*, Springer, 2005, pp. 165–192.
[32] R. E. Wright, “Logistic regression.,” 1995.
[33] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[34] W. S. Noble, “What is a support vector machine?,” *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006.
[35] M. Anthony and P. L. Bartlett, *Neural network learning: Theoretical foundations*. cambridge university press, 2009.
[36] W. Sun, C. Yang, and J. Qi, “Credit Risk Assessment in Commercial Banks Based on Support Vector Machines,” in *2006 International Conference on Machine Learning and Cybernetics*, 2006, pp. 2430–2433.
[37] C.-L. Huang, M.-C. Chen, and C.-J. Wang, “Credit scoring with a data mining approach based on support vector machines,” *Expert Syst. Appl.*, vol. 33, no. 4, pp. 847–856, 2007.
[38] P. Yao, “Feature selection based on SVM for credit scoring,” in *2009 International Conference on Computational Intelligence and Natural Computing*, 2009, vol. 2, pp. 44–47.
[39] S. Birla, K. Kohli, and A. Dutta, “Machine learning on imbalanced data in credit risk,” in *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2016, pp. 1–6.

[40] S. Purohit and A. Kulkarni, “Credit evaluation model of loan proposals for Indian Banks,” in *2011 World Congress on Information and Communication Technologies*, 2011, pp. 868–873.
[41] R. E. Turkson, E. Y. Baagyere, and G. E. Wenya, “A machine learning approach for predicting bank credit worthiness,” in *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, 2016, pp. 1–7.
[42] A. J. Hamid and T. M. Ahmed, “Developing prediction model of loan risk in banks using data mining,” *Mach. Learn. Appl. Int. J. MLAIJ Vol*, vol. 3, no. 1, 2016.
[43] A. Gahlaut and P. K. Singh, “Prediction analysis of risky credit using Data mining classification models,” in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2017, pp. 1–7.
[44] P. Singh, “Comparative study of individual and ensemble methods of classification for credit scoring,” in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, 2017, pp. 968–972.
[45] X. Zhang, Y. Yang, and Z. Zhou, “A novel credit scoring model based on optimized random forest,” in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, 2018, pp. 60–65. doi: 10.1109/CCWC.2018.8301707.
[46] S. Khemakhem, F. B. Said, and Y. Boujelbene, “Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines,” *J. Model. Manag.*, 2018.
[47] F. Shen, X. Zhao, Z. Li, K. Li, and Z. Meng, “A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation,” *Phys. Stat. Mech. Its Appl.*, vol. 526, p. 121073, Jul. 2019, doi: 10.1016/j.physa.2019.121073.
[48] D. Tripathi, D. R. Edla, and R. Cheruku, “Hybrid credit scoring model using neighborhood rough set and multi-layer ensemble classification,” *J. Intell. Fuzzy Syst.*, vol. 34, no. 3, pp. 1543–1549, 2018.
[49] V. Kuppili, D. Tripathi, and D. Reddy Edla, “Credit score classification using spiking extreme learning machine,” *Comput. Intell.*, vol. 36, no. 2, pp. 402–426, 2020.
[50] D. Tripathi, D. R. Edla, V. Kuppili, and A. Bablani, “Evolutionary Extreme Learning Machine with novel activation function for credit scoring,” *Eng. Appl. Artif. Intell.*, vol. 96, p. 103980, 2020.
[51] R. S. Kovvuri and R. Cheripelli, “Credit Risk Valuation Using an Efficient Machine Learning Algorithm,” in *Advances in Decision Sciences, Image Processing, Security and Computer Vision*, Springer, 2020, pp. 648–657.
[52] Y. Wang, Y. Zhang, Y. Lu, and X. Yu, “A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data,” *Procedia Comput. Sci.*, vol. 174, pp. 141–149, Jan. 2020, doi: 10.1016/j.procs.2020.06.069.
[53] “Regulativa_Upatstvo_krediten_registar.pdf.” Accessed: Apr. 11, 2021. [Online]. Available: https://nbrm.mk/WBStorage/Files/Regulativa_Upatstvo_krediten_registar.pdf
[54] “Одлука_кредитен_ризик_КОНЕЧНА.pdf.” Accessed: Apr. 11, 2021. [Online]. Available: https://nbrm.mk/content/%D0%9E%D0%B4%D0%BB%D1%83%D0%BA%D0%B0_%D0%BA%D1%80%D0%B5%D0%B4%D0%B8%D1%82%D0%B5%D0%BD_%D1%80%D0%B8%D0%B7%D0%B8%D0%BA_%D0%9A%D0%9E%D0%9D%D0%95%D0%A7%D0%9D%D0%90.pdf

[55] S. Oreski, D. Oreski, and G. Oreski, “Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment,” *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12605–12617, 2012.

[56] E. Zdravevski, P. Lameski, A. Kulakov, and D. Gjorgjevikj, “Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets,” in *2014 Federated Conference on Computer Science and Information Systems*, 2014, pp. 387–394.

[57] L. Zhang, H. Ray, J. Priestley, and S. Tan, “A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data,” *J. Appl. Stat.*, vol. 47, no. 3, pp. 568–581, 2020.

[58] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[59] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PloS One*, vol. 10, no. 3, p. e0118432, 2015.

[60] K. S. Hasan and V. Ng, “Automatic keyphrase extraction: A survey of the state of the art,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1262–1273.

[61] D. M. Eler, D. Grosa, I. Pola, R. Garcia, R. Correia, and J. Teixeira, “Analysis of document pre-processing effects in text and opinion mining,” *Information*, vol. 9, no. 4, p. 100, 2018.

[62] S. K. Bharti and K. S. Babu, “Automatic keyword extraction for text summarization: A survey,” *ArXiv Prepr. ArXiv170403242*, 2017.

[63] C. J. Saju and A. S. Shaja, “A survey on efficient extraction of named entities from new domains using big data analytics,” in *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, 2017, pp. 170–175.

[64] C. Nopp and A. Hanbury, “Detecting Risks in the Banking System by Sentiment Analysis,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Sep. 2015, pp. 591–600. doi: 10.18653/v1/D15-1071.

[65] W. S. Lee and S. Y. Sohn, “Identifying Emerging Trends of Financial Business Method Patents,” *Sustainability*, vol. 9, no. 9, Art. no. 9, Sep. 2017, doi: 10.3390/su9091670.

[66] S. Moro, P. Cortez, and P. Rita, “Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation,” *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1314–1324, 2015.

[67] S. A. Rios, F. Aguilera, J. D. Nuñez-Gonzalez, and M. Graña, “Semantically enhanced network analysis for influencer identification in online social networks,” *Neurocomputing*, vol. 326, pp. 71–81, 2019.

[68] H. Mao, L. Zhu, and X. Jin, “Methods of Measuring Influence of Bank Customer Using Social Network Model,” *Am. J. Ind. Bus. Manag.*, vol. 05, pp. 155–160, Jan. 2015, doi: 10.4236/ajibm.2015.54017.

[69] A. Vaswani *et al.*, “Attention is all you need,” *ArXiv Prepr. ArXiv170603762*, 2017.

[70] W.-J. Ko, B.-H. Tseng, and H.-Y. Lee, “Recurrent neural network based language modeling with controllable external memory,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5705–5709.

[71] S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and problem solutions,” *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 6, no. 02, pp. 107–116, 1998.

[72] T. Chen, R. Xu, Y. He, and X. Wang, “Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN,” *Expert Syst. Appl.*, vol. 72, pp. 221–230, Apr. 2017, doi: 10.1016/j.eswa.2016.10.065.

[73] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, “Coupled multi-layer attentions for co-extraction of aspect and opinion terms,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31, no. 1.

[74] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” *ArXiv Prepr. ArXiv160301360*, 2016.

[75] P. M. Shishtla, K. Gali, P. Pingali, and V. Varma, “Experiments in telugu ner: A conditional random field approach,” 2008.

[76] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” 2001.

[77] L. M. Rojas-Barahona, “Deep learning for sentiment analysis,” *Lang. Linguist. Compass*, vol. 10, no. 12, pp. 701–719, 2016.

[78] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, “Sentiment analysis leveraging emotions and word embeddings,” *Expert Syst. Appl.*, vol. 69, pp. 214–224, 2017.

[79] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[80] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, “Sentiment analysis based on improved pre-trained word embeddings,” *Expert Syst. Appl.*, vol. 117, pp. 139–147, 2019.

[81] I. Santos, N. Nedjah, and L. de Macedo Mourelle, “Sentiment analysis using convolutional neural network with fastText embeddings,” in *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 2017, pp. 1–5.

[82] A. Akbik, D. Blythe, and R. Vollgraf, “Contextual string embeddings for sequence labeling,” in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1638–1649.

[83] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *ArXiv Prepr. ArXiv181004805*, 2018.

[84] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *ArXiv Prepr. ArXiv191001108*, 2019.

[85] D. Araci, “Finbert: Financial sentiment analysis with pre-trained language models,” *ArXiv Prepr. ArXiv190810063*, 2019.

[86] L. A. Ramshaw and M. P. Marcus, “Text chunking using transformation-based learning,” in *Natural language processing using very large corpora*, Springer, 1999, pp. 157–176.

[87] T. Nasukawa and J. Yi, “Sentiment analysis: Capturing favorability using natural language processing,” in *Proceedings of the 2nd international conference on Knowledge capture*, 2003, pp. 70–77.

[88] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, “Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques,” in *Third IEEE international conference on data mining*, 2003, pp. 427–434.

[89] J. C. Salinas Alvarado, K. Verspoor, and T. Baldwin, “Domain Adaption of Named Entity Recognition to Support Credit Risk Assessment,” in *Proceedings of the Australasian*

Language Technology Association Workshop 2015, Parramatta, Australia, Dec. 2015, pp. 84–90. Accessed: Apr. 25, 2021. [Online]. Available: https://www.aclweb.org/anthology/U15-1010

[90] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning subjective language," *Comput. Linguist.*, vol. 30, no. 3, pp. 277–308, 2004.

[91] R. P. Schumaker, Y. Zhang, and C. Huang, *Sentiment Analysis of Financial News Articles*.

[92] A. Yadav, C. K. Jha, A. Sharan, and V. Vaish, "Sentiment analysis of financial news using unsupervised approach," *Procedia Comput. Sci.*, vol. 167, pp. 589–598, 2020.

[93] T. Loughran and B. McDonald, "Textual analysis in accounting and finance: A survey," *J. Account. Res.*, vol. 54, no. 4, pp. 1187–1230, 2016.

[94] S. W. K. Chan and M. W. C. Chong, "Sentiment analysis in financial texts," *Decis. Support Syst.*, vol. 94, pp. 53–64, Feb. 2017, doi: 10.1016/j.dss.2016.10.006.

[95] X. Wan, J. Yang, S. Marinov, J.-P. Calliess, S. Zohren, and X. Dong, "Sentiment correlation in financial news networks and associated market movements," *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41598-021-82338-6.

[96] L. Dodevska *et al.*, *Predicting companies stock price direction by using sentiment analysis of news articles*. 2019.

[97] S. F. Crone and C. Koeppel, "Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons," in *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 2014, pp. 114–121.

[98] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. Chitkushev, W. Souma, and D. Trajanov, "Forecasting corporate revenue by using deep-learning methodologies," in *2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*, 2019, pp. 115–120.

[99] A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. R. Venugopal, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," *World Wide Web*, vol. 20, no. 2, pp. 135–154, 2017.

[100] S. Liu, F. Li, F. Li, X. Cheng, and H. Shen, "Adaptive co-training SVM for sentiment classification on tweets," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 2079–2088.

[101] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in *International semantic web conference*, 2012, pp. 508–524.

[102] V. Sehgal and C. Song, "Sops: stock prediction using web sentiment," in *Seventh IEEE international conference on data mining workshops (ICDMW 2007)*, 2007, pp. 21–26.

[103] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Intell. Res.*, vol. 37, pp. 141–188, 2010.

[104] V. Narayanan, I. Arora, and A. Bhatia, "Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model," in *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, Berlin, Heidelberg, 2013, pp. 194–201. doi: 10.1007/978-3-642-41278-3_24.

[105] X. Rong, "word2vec parameter learning explained," *ArXiv Prepr. ArXiv14112738*, 2014.

[106] R. P. Schumaker, Y. Zhang, C.-N. Huang, and H. Chen, "Evaluating sentiment in financial news articles," *Decis. Support Syst.*, vol. 53, no. 3, pp. 458–464, 2012.

[107] J. Thanaki, *Python natural language processing*. Packt Publishing Ltd, 2017.

[108] M.-Y. Day and C.-C. Lee, “Deep learning for financial sentiment analysis on finance news providers,” in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 1127–1134.
[109] P. Vateekul and T. Koomsubha, “A study of sentiment analysis using deep learning techniques on Thai Twitter data,” in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2016, pp. 1–6.
[110] Y. Zhang, J. E. Meng, R. Venkatesan, N. Wang, and M. Pratama, “Sentiment classification using comprehensive attention recurrent models,” in *2016 International joint conference on neural networks (IJCNN)*, 2016, pp. 1562–1569.
[111] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *ArXiv Prepr. ArXiv13013781*, 2013.
[112] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017.
[113] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, p. e1253, 2018.
[114] D. Tang, B. Qin, and T. Liu, “Document modeling with gated recurrent neural network for sentiment classification,” in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1422–1432.
[115] N. Nedjah, I. Santos, and L. de Macedo Mourelle, “Sentiment analysis using convolutional neural network via word embeddings,” *Evol. Intell.*, pp. 1–25, 2019.
[116] Z. Lin *et al.*, “A structured self-attentive sentence embedding,” *ArXiv Prepr. ArXiv170303130*, 2017.
[117] L. Zhao, L. Li, and X. Zheng, “A BERT based sentiment analysis and key entity detection approach for online financial texts,” *ArXiv Prepr. ArXiv200105326*, 2020.
[118] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[119] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, “Sentiment analysis based on deep learning: A comparative study,” *Electronics*, vol. 9, no. 3, p. 483, 2020.
[120] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” *ArXiv Prepr. ArXiv14042188*, 2014.
[121] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *ArXiv Prepr. ArXiv14093215*, 2014.