# Predikcija cene finansijske aktive analizom sentimenta finansijskih tvitova

Dušan Stević
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
stevic.r233.2020@uns.ac.rs

Dušan Stević
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
stevic.r233.2020@uns.ac.rs

***Sažetak*—Sentiment analiza finansijskih tvitova javila se kao potreba u sve dinamičnijem tržišnom ambijentu. Tradicionalni strukturirani izvori podataka više nisu dovoljni u finansijskoj analizi. U nastojanju da se ostvari što veći profit investitori se okreću novim tehnikama rudarenja podataka kako bi iz njih izvlačili korisne informacije. Različite tehnike analize sentimenta koriste se kako bi se sirovi podaci transformisali u lukrativne informacije. Tehnike rudarenja podataka i sentiment analize koriste se kao ulaz u prediktivne modele koji za zadatak imaju predikciju kretanja cene finansijske aktive. Osnovi cilj ovog projekta je razvoj jedinstvenog sistema koji bi predstavljao fuziju sistema za analizu sentimenta i sistema za predikciju cene finansijske aktive. Za potrebe analize sentimenta u radu su obrađene tehnike koje se zasnivaju na jezičkim pravilima, rečnicima, tradicionalnim modelima mašinskog učenja i modelima zasnovanim na neuronskim mrežama. Dominantnost modela zasnovih na neuronskim mrežama u pogledu performansi potvrdili su BERT i DistilBERT. Izlazi iz sentiment analize koriste se kao ulaz u regresione modele koji za cilj imaju ispitivanje hipoteze da li postoji korelacija između sentimenta finansijskih tvitova i cene finansijske aktive. Sentiment finansijskih tvitova treba da posluži investitorima kao signal kojim treba da se rukovode prilikom donošenja investicionih odluka. Sprovedeno istaživanje potvrdilo je hipotezu da postoji pozitivna korelacija između sentimenta finansijskih tvitova i kretanja cene finansijske aktive. Pozitivan sentiment uslovljava rast cena finansijske aktive, dok negativan sentiment uslovljava pad cena finansijeke aktive. Korišćenjem ovog sistema investitorima se pruža mogućnost da donose racionalne investicione odluke.**

***Ključne reči— rudarenje podatakat; sentiment analiza; finansije; DistilBERT; regresija***

## I. UVOD

Motivaciju za predikciju cene finansijske aktive možemo naći na nivou svetske ekonomije, na nivou država, kao i na nivou pojedinaca.

Na nivou svetske ekonomije uzimajući u obzir prethodne tržišne krahove, The Dot – Com Bubble 90-tih i The US Housing Bubble 2000-tih, postavlja se pitanje da li će bitkojn (engl. *bitcoin*[1]), finansijski derivati (engl. *financial derivatives*[2]) i drugi proizvodi fintech (engl. *fintech*[3]) industrije biti uzrok naredne Svetske ekonomske krize [1]. Novokovanica fintech predstavlja zajednički imenitelj za sve finansijske inovacije koje su se pojavile u poslednjih nekoliko godina. Etimološko poreklo reči fintech derivirano je od engleskih reči finansije (engl. *finance*) i tehnologija (engl. *technology*). Nagli prodor informacionih tehnologija u domen specifične oblasti uslovio je nastanak ove složenice. Automatizacija i optimizacija procesa u finansijama zahtevali su uvođenje i implementaciju informacionih tehnologija na polju finansija, a kao rezultat ovog procesa bilo je generisanje pojma fintech. Digitalne valute (engl. *cryptocurrency*), automatsko uparivanje ponude i tražnje, finansijski blokčejn (engl. *financial blockchain*), itd tipični su predstavnici fintech-a [8]. Pojava velikog broja kriptovaluti, ogroman broj participanta tržišne utakmice i veliko dnevno oscilovanje kriptovaluti predstavljaju indikatore za nastanak novog finansijskog sloma. Kao potreba javlja se nastojanje da tržišni ambijent učinimo sigurnijim predviđanjem kretanja cene finansijske aktive [2].

Porast državnog intervencionizma i protekcionizma u nastojanju da se reguliše finansijsko tržište zahteva od države da izvrši diverzifikaciju vlastitih sredstava. Rezerve kojima država interveniše na otvorenom tržištu (engl. *open market operations*) prestruktuiraju se sa tradicionalnih komoditeta, kao što su zlato i devizne rezerve, na kriptovalute i finansijske derivate[3]. Nagla apresijacija[4] ili depresijacija[5] kriptovaluti dovela bi do stanja da centralne banke nisu u stanju da vrše operacije na otvorenom tržištu, čime ne bi bile u stanju da obavljaju jednu od svojih osnovnih funkcija.

Banke, osiguravajuća drušva, hedž (engl. *hedge*[6]) fondovi, itd. vrše ulaganje sopstvenih i tuđih sredstava zarad ostvarivanja što većeg profita. U želji da ostvare što veći profit, finansijske institucije odlučuju da svoj portfolio prošire

[1] Vrsta kriptovalute

[2] Vrsta visokorizične finansijske aktive

[3] Predstavlja upotrebu savremenih informacionih tehnologija u domenu finansija [7]

[4] Porast cene finansijske aktive [6].

[5] Pad cene finansijske aktive [6].

[6] Vrsta institucionalnih investitora[6]

finetech inovacijama, često zanemarujući rizik koje svako ulaganje nosi sa sobom. Finansijske institucije (ne)svesno kontaminiraju svoju finansijsku aktivu dovodeći u pitanje stabilnost celog finansijskog sistema [4].

Imajući u vidu racionalnog donosioca odluka, ispred svakog investitora postavlja se nagodba prinosa i rizika. Često, ne poznajući tehnologiju koja stoji iza finansijskih inovacija, zaslepljeni lukrativnim mogućnostima, investitori donose neracionalne odluke i na taj način ugrožavaju vlastiti portfolio. Kao imperativ ispred svakog individualnog i kolektivnog investitora postavlja se zahtev za upoznavanjem sa finansijsko- tehnološkim inovacijama pre bilo kakvog ulaganja, kako bi se sprečio domino efekat koji bi propagirao na nivo svetske ekonomije [5].

Domen finansija pokazao se kao izuzetno privlačan i izazovan za primenu tehnika veštačke inteligencije. Osnovna motivacija za izradu ovog rada je produbljivanje znanja iz oblasti veštačke inteligencije u domenu finansija ali i aplikativnost softverskog rešenja u procesu donošenja investicionih odluka. Multidisciplinarni pristup u izradi rada omogućio je primenu *state of the art* algoritama veštačke inteligencije na polje finansija. Rad se sastoji od tri tematske celine. Prva tematska celina bavi se problemima rudarenja i obrade podataka (engl. *data mining*). Kako podaci predstavljaju sirovine za modele veštačke inteligencije posebna pažnja se posvećuje raznorodnosti podataka. Na performanse ovih modela direktno utiče kvalitet obrađenih podataka. Raznorodnost podataka koja se ogleda u različitim izvorima i tipovima podataka obezbeđuje da se u radu pored tradicionalnih strukturiranih tipova podataka koriste i nestrukturirani tipovi podataka. U poslednjih nekoliko godina beleži se dominantnost nestrukturiranih tipova podataka nad strukturiranim. Eksplozija nestrukturiranih podataka doprinela je da se u radu prilikom predikcije cene finansijske aktive pored tradicionalnih tabelarnih finansijskih izveštaja koriste i finansijski tvitovi (engl. *tweets*). Na ovaj način u radu je postignuta raznorodnost podataka koja se manifestuje u simbiozi strukturiranih (tabelarni berzanski izveštaji) i nestrukturiranih (finansijki tvitovi i vesti) podataka. Druga tematska celina bavi se klasifikacionim problemima konkretno klasifikacijom sentimenta finansijskih tvitova i vesti. Sentiment analiza (engl. *sentiment analysis*) ili ekstrakcija mišljenja i stavova (engl. *opinion mining*) za cilj ima da utvrdi da li neki tekst ima negativan, neutralan ili pozitivan prizvuk. Na donošenje investicionih odluka pored kretanja berzanskih indeksa utiče i investiciono raspoloženje javnog mnjenja. Iz tog razloga neophodno je prilikom predikcije cene finansijske aktive uključiti i analizu sentimenta finansijskih tvitova i vesti. Treća tematska celina bavi se regresionim problemima, konkretno pronalaženjem korelacije između kretanja cene finansijske aktive i sentimenta finansijskih tvitova. Osnovni zadatak ove celine je da ispita uticaj koji finansijski tvitovi i vesti, kao izraz javnog mnjenja, imaju na berzanska kretanja. U nastavku rada će detaljnije biti objašnjeni različiti aspekti rešavanih problema kao i način realizacije samog rešenja. U poglavlju 2 je napravljen kratak pregled pronađenih radova koji se bave istom ili sličnom problematikom. Za svaki rad je detaljnije obrazložena metodologija kojom je problem rešavan. Poglavlje 3 opisuje proces formiranja ciljnih skupova podataka korišćenih u daljim analizama. U poglavlju su detaljan način opisani skupovi podataka o kretanju cene finansijske aktive i finansijskim tvitovima. U nastavku poglavlja predstavljen je postupak obrade i temporalne fuzije[7] sakupljenih podataka u nastojanju da se dobiju ciljni skupovi podataka. Poglavlje 4 daje osvrt na metodološke pristupe korišćene u radu. Poglavlje 5 prezentuje dobijene rezultate sentiment analize i predikcije cene finansijske aktive. Poglavlje 6 je zaključno poglavlje u kojem se sumarizuje celokupan rad i daju predlozi za buduće pravca rada.

## II. PRETHODNA REŠENJA

Za potrebe realizacije ovog rada analizirano je više postojećih rešenja među kojima se izdvajaju sledeća tri rada. Cilj *"Web Services for Stream Mining: A Stream-Based Active Learning Use Case"* [9] rada je bio da se izvrši analiza sentimenta finansijskih tvitova upotrebom tehnika mašinskog učenja konkretno SVM (engl. *support-vector machines*) klasifikatora. Usled nedovoljne količine labeliranih podataka autori su se odlučili za hibridan pristup koji objedinjuje nadgledano (klasifikaciju) i nenadgledano učenje (klasterovanje). AL-SVM predstavlja sintezu SVM i K – means algoritma. Ideja je da se pomoću klasterovanja dođe do nedostajućih labela. Autori su na raspolaganju imali 379390 javno dostupnih tvitova koji imaju tag "$AAPL". Konvencija je da $ govori da je reč o akcijama neke korporacije dok karakteri iza $ označavaju *ticker* odnosno ime korporacije. Doprinos koji ovaj rad ima ispoljava se u činjenici da se kao izvor podataka koriste tvitovi sa tagom "$AAPL". Takođe za rešvanje problema klasifikacije sentimenta finansijskih tvitova upotrebljene se standardne tehnike mašinskog učenja (SVM, NB (engl. *Naive Bayes*), slučajne šume (engl. *Random Forest*)). Za evaluaciju klasifikacije koristi se ista metrika tj. tačnost (engl. *accuracy*). Drugi rad od značaja za ovaj projekat je *"Fine-Grained Analysis of Financial Tweets"* [10]. Cilj ovog rada je bio da se izvrši analiza sentimenta finansijskih tvitova upotrebom neuronskih mreža (CNN (engl. *convolutional neural network*), RNN (engl. *recurrent neural network*) i LSTM (engl. *long short-term memory*)). Pored analize sentimenta finansijskih tvitova ideja je da se izvrši i predikcija kretanja cena Apple korporacije u budućnosti. Rađena je komparativna analiza više modela zasnovanih na neuronskim mrežama. Metodološki se problem klasifikacije tvitova uliva u problem regresije. Kada se dobije sentiment tvitova on predstavlja dodatno obeležje (engl. *feature*) za regresioni problem. Autori su na raspolaganju imali 334K javno dostupnih tvitova koji imaju tag "$AAPL". Evaluacija rešenja vršna je podelom na trening i test skup. Za potrebe evaluacije klasifikacije koristi se tačnost i F1 mera. Dok se za potrebu regresije koristi $R^2$ (engl. *coefficient of determination*) i MSE (engl. *mean squared error*). Doprinos ovog rada se manifestuje u komparativnom pristupu izučavanju modela. Jedinstvenom pristupu disekcije realnog problema na probleme klasifikacije i regresije. Primena sveobuhvatnog metričkog sistema za evaluaciju koji će omogućiti komparaciju različitih modela. Treći rad od značaja za ovaj projekat je *"Sentiment Analysis Based on Financial Tweets*

[7] Spajanje heterogenih (potiču iz različtih izvora) podataka po vremenu nastanka

*and Market Information"* [11]. Cilj ovog rada je bio da se izvrši leksički zasnovana analiza sentimenta finansijskih tvitova. Takođe u radu je izvršeno ispitivanje korelacije između sentimenta finansijskih tvitova i kretanja cene akcija. Leksički pristup podrazumeva postojanje unapred pripremljenog rečnika u kojima je dostupan polaritet reči (reči su anotirane kao pozitivne, negativne i neutralne na osnovu toga koliko se često pojavljuju u odgovarajućim kontekstima). Autori su na raspolaganju imali 15K javno dostupnih tvitova koji imaju tag "$AAPL". Za evaluaciju rešenja koristi se koeficijent korelacije. U radu koeficijent korelacije iznosi 0.611639 što nam govori da između prosečne vrednosti sentimenta i cene akcije postoji jaka pozitivna korelacija tj. objave na tviteru utiču na tržišna kretanja. Doprinos koji ovaj rad ima ispoljava se uvođenjem novog metodološkog pristupa zasnovanog na leksičkoj analizi sentimenta kao i uvođenjem nove metrike zasnovane na koeficijentu korelacije.

## III. FORMIRANJE SKUPOVA PODATAKA

Sve veća prisutnost nestrukturiranih tipova podataka stvorila je potrebu da se i ovakav tip podataka uključi u finansijsku analizu. Tradicionalni stukturirani tipovi podataka pokazali su se kao nedovoljni da se njima opiše dinamičan tržišni ambijent. Stoga kao imperativ nameće se težnja da se finansijski skupovi podataka ekspanduju novim nestrukturiranim tipovima podataka kao što su finansijski tvitovi i vesti. Upliv informacionih tehnologija u domen finansija uslovio je hiperprodukciju nestrukturiranih podataka. U uslovima turbulentnih tržišnih dešavanja gde se kao najvažniji komoditet pojavljuju podaci, izostavljanjem nestrukturiranih podataka iz finansijske analize smanjio bi njen sveobuhvat. Homogeni podaci tj. podaci koji potiču iz jednog izvora nisu više dovoljni. Dosadašnje iskustvo je pokazalo da se za donošnje ozbiljnih investicionih odluka moraju koristiti heterogeni podaci. Podaci moraju da potiču iz različitih izvora i da po svojoj prirodi budu strukturirani i nestrukturirani. Iz ovih razloga u ovom radu se pojavljuju kako strukturirani tipovi podataka tako i nestrukturirani tipovi podataka. Strukturirani tipovi podataka obuhvataju tabelarne berzanske izveštaje. Nestrukturirani tipovi podataka obuhvataju dve grupe finansijskih podatka. Prvu grupu predstavljaju labelirani/anotirani finansijski tvitovi i vesti, dok drugu grupu predstavljaju nelabelirani/neanotirani finansijski tvitovi i vesti. Labelirani finansijski tvitovi i vesti koriste se za potrebe obučavanja algoritama nadgledanog učnja. Nelabelirani tvitovi i vesti koriste se za potrebe predikcije cena finansijske aktive.

### *A. Akvizicija podataka*

Akvizicija podataka odnosi se na prikupljanje strukturiranih i nestrukturiranih podataka. U narednom delu biće detaljno opisane tehnike rudarenja podataka koje su se koristile kako bi se došlo do željenih sirovih podataka.

Akvizicija tabelarnih berzanskih izveštaja vršena je pomoću Yahoo Finance API-a [12]. Parametri pretrage za ovaj skup podataka obuhvataju: početak vremenskog intervala za koji se podaci sakupljaju, kraj vremenskog intervala za koji se podaci sakupljaju i berzanski simbol finansijske aktive za koju se podaci sakupljaju. Prikupljeni podaci su tabelarnog oblika i predstavljaju strukturirani tip podataka. Istorijski podaci sa berze koriste se pri predikciji kretanja cene finansijske aktive.

Akvizicija labeliranih/anotiranih finansijskih tvitova i vesti vršena je manuelno sa sajtova specijalizovanih za deponovanje skupova podataka kaggle [14], metatext [15], data.world [16] i lionbridge.ai [17]. Prikupljeni podaci su tekstualnog oblika i predstavljaju nestrukturirani tip podataka. Labelirani finansijski tvitovi i vesti koriste se za obučavanje modela nadgledanog učenja.

Akvizicija nelabeliranih/neanotiranih tvitova vršena je pomoć Twint API-a [13]. Parametri pretrage za ovaj skup podataka obuhvataju: početak vremenskog intervala za koji se podaci sakupljaju, kraj vremenskog intervala za koji se podaci sakupljaju, lista korisnika čiji tvitovi se sakupljaju i lista termina po kojima se tvitovi filtriraju. Listu korinika čiji tvitovi se sakupljaju čine najeminentiji ljudi, televizijske kuće i korporacije iz sveta finansija. Prilikom sakupljanja tvitova filtrirani su se samo oni tvitovi koji u sebi sadrže termine iz liste termina. Prikupljani podaci su tekstualnog oblika i predstavljaju nestrukturirani tip podataka. Nelabelirani finansijski tvitovi i vesti koriste se za pronalaženjem korelacije između kretanja cene finansijske aktive i sentimenta finansijskih tvitova.

### *B. Data wrangling*

Data wrangling je postupak transformacije sirovih podataka iz jednog oblika u drugi. Pri čemu podaci i dalje ostaju u sirovom obliku ali u formatu koji je pogodniji za dalju analizu U ovom radu izvršene su neke od sledećih data wrangling tehnika:

- Rukovanje velikim skupovima podataka koji prevazilaze kapacitete RAM-a (engl. *random-access memory*). Skupovi podataka su toliko veliki da ne mogu kompletni da stanu u RAM stoga je potrebno primeniti posebne tehnike učitavanja podataka u delovima (engl. *chunks*).
- Uklanjanje nepotrebnih kolona iz skupova podataka kako bi se smanjilo memorijsko zauzeće.
- Izbacivanje NA/null vrednosti iz skupova podataka
- Uklanjanje kompletnih duplikata iz skupova podataka. Kompletni duplikati su oni tvitovi koji imaju isti tekst i isti sentiment.
- Uklanjanje parcijalnih duplikata iz skupova podataka. Parcijalni duplikati su oni tvitovi koji imaju isti tekst, a različit sentiment.
- Postavljanje jedinstvenog zaglavlja za sve skupove tvitova. Svaki skup tvitova (labelirani i nelabelirani) ima jedinstveno zaglavlje koje čine: date kolona, tekst kolona i sentiment kolona.
- Postavljanje jedinstvenog skupa vrednosti sentimenta za sve skupove tvitova. Svaki skup tvitova (labelirani i nelabelirani) ima jedinstven skup vrednosti sentimenta koji čine sledeće vrednosti: -1 za negativan sentiment, 0 za neutralan sentiment i 1 za pozitivan sentiment.

- *Inplace* transformacije skupova podataka radi uštede memorijskog zauzeća.

### C. *Pretprocesiranja podataka*

Pretprocesiranje podataka podrazumeva takve operacije nad podacima koje prevode podatke iz sirovog u rafinirani oblik pogodan za modele mašinskog učenja. U ovom radu izvršene su neke od sledećih tehnika pretprocesiranja podataka:

- Konverzijom teksta u mala slova prevenira se različita reprezentacija reči koje imaju isto značenje.
- Eliminacija linkova. Sadržaj linkova predstavalja šum u podacima jer ne unosi dodatnu semantiku.
- Eliminacija pominjanja korisnika (engl. *mentions*). Sve desno od @ znaka se uklanja jer ne unosi dodatnu semantiku.
- Eliminacija stop reči. Eliminacija frekventnih gramatičkih konstrukata koji sa sobom ne nose semantiku (predlozi, pomoćni glagoli, rečce).
- Eliminacija znakova interpunkcije.
- Svođnje tvitova na alfanumeričke karaktere.
- Morfološka normalizacija reči upotrebom steminga (engl. *stemming*). Na primer "walking" postaje "walk" nakon steminga.
- Morfološka normalizacija reči upotrebom lematizacije (engl. *lemmatization*). Na primer "better" postaje "good" nakon lematizacije.

### D. *Exploratory data analysis (EDA)*

Nakon što su podaci rafinirani može se pristupiti eksplorativnoj analizi. Eksplorativna analiza ima za cilj detaljniji uvid u podatke. U ovom radu eksplorativna analiza je izvršena na dva nivo.

Prvi nivo eksplorativne analize obuhvata eksplorativnu analizu istorijskih podataka sa berze. Eksplorativna analiza berzanski podataka za cilj ima da utvrdi kakva je volatilnost finansijske vremenske serije. Ova analiza treba da nam pruži informacije da li je došlo do šokova na finansijskom tržišu u posmatranom period. Fig. 1 predstavlja finansijsku vremensku seriju akcije Apple kompanije u vremenskom periodu od 2010. godine do 2020. godine. EDA-om je utvrđeno da je volatilnost akcije Apple bila veća na početku posmatranog perioda. Već volatilnost direktna je posledica promena u top menadžmentu Apple kompanije početkom 2010. godine.

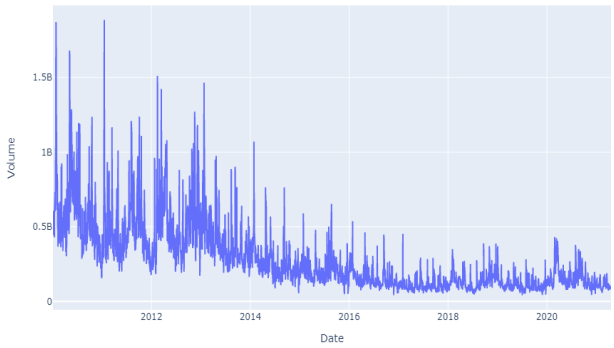


Fig. 1. EDA finansijske vremenske serije

Drugi nivo eksplorativne analize obuhvata eksplorativnu analizu finansijskih tvitova i vesti. Fig. 2 predstavlja distribuciju sentimenta finansijskih tvitova i vesti po labelama. EDA-om je utvrđeno da je skup podataka finansijskih tvitova i vesti blago nebalansiran što za posledicu ima lošije performanse modela veštačke inteligencije.

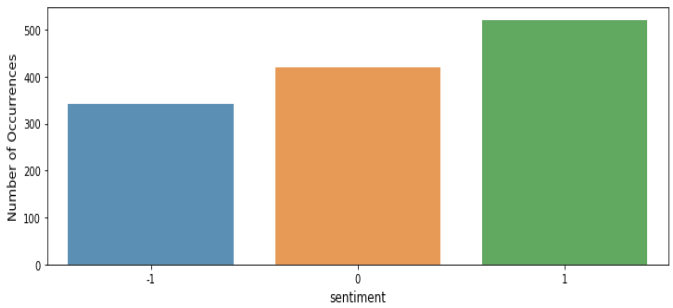


Fig. 2. Distribucija sentimenta tvitova po labelama

Pored ispitivanja distribucije sentimenta tvitova po labelama EDA-a se koristi i za formiranja oblaka reči (engl. *word cloud*). Oblak reči se u finansijama koristi kao signal za aktivnost na tržištu. Fig. 3 predstavalja oblak reči negativnih tvitova. Negativne reči koje se pojavljuju u ovom oblaku trebaju da signaliziraju investitorima da je tržište na silaznoj putanji. Negativne signalne reči u ovom oblaku su: *short*, *low*, *volatility*, *drop*, *decline*, *lower*, *sell*, *cut*, *bear*, *bearish*, *crash* i *loss*.

Fig. 3. Oblak reči negativnih tvitova

Fig. 4 predstavalja oblak reči neutralnih tvitova. Neutralne reči koje se pojavljuju u ovom oblaku trebaju da signaliziraju investitorima da je tržište u fazi stagnacije. Neutralne signalne reči u ovom oblaku su: *market*, *option*, *chart*, *stocks*, *trade*, *gold*, *investor*, *equity*, *commodity*, *investment*, *price* i *week*.

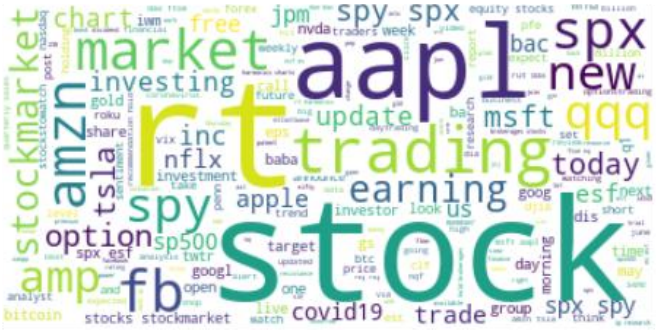


Fig. 4. Oblak reči neutralnih tvitova

Fig. 5 predstavalja oblak reči pozitivnih tvitova. Pozitivne reči koje se pojavljuju u ovom oblaku trebaju da signaliziraju investitorima da je tržište na uzlaznoj putanji. Pozitivne signalne reči u ovom oblaku su: *great*, *good*, *profit*, *long*, *gain*, *best*, *bull*, *bullish*, *buy*, *call*, *strong* i *high*.


Fig. 5. Oblak reči pozitivnih tvitova

## IV. METODOLOGIJA

U prethodnom poglavlju na detaljan način je opisana prva tematska celina u radu. Prva tematska celina u radu bavila se problemom prikupljanja i obrade podataka. U ovom poglavlju obrađuju se preostale dve tematske celina. Nakon što su podaci prikupljeni i obrađeni neophodno je iste proslediti modelima veštačke inteligencije. Rafinirani podaci prvo se prosleđuju klasifikacionim modelima. Primena klasifikacionih modela na rafinirane podatke je problematika kojom se druga tematska celina bavi. Nako što klasifikacioni modeli vrate rezultate sa tim podacima se ulazi u regresione modele. Primena regresionih modela na rezultate klasifikacionih modela je problematika kojom se treća tematska celina bavi.

### A. *Klasifikacioni Problemi*

Klasifikacioni problemi u ovom radu svode se na analizu sentimenta finansijskih tvitova i vesti. Analiza sentimenta ima za zadataka da utvrdi polaritet nekog tvita ili vesti. Sentiment analiza treba za određeni tvit ili vest da vrati da li je taj tvit ili vest negativnog, neutralnog ili pozitivnog prizvuka. Kako svaki tvit ili vest može da ima samo jednu sentiment labelu sentiment analiza se svodi na problem više-klasne klasifikacije (engl. *multi-class classification*) tj. na problem jedne labele (engl. *single-label problem*) [18]. U radu je isprobano više različitih modela sentiment analize. Svi primenjeni modeli sentiment analize mogu da se podele na dve velike grupe u zavisnosti od tipa učenja koji se nalazi u pozadini modela. U nenadgledane modele spadaju modeli zasnovani na pravilima i modeli zasnovani na rečnicima. U nadgledane modele spadaju tradicionalni modeli mašinskog učenja i modeli zasnovani na neuronskim mrežama.

Model zasnovan na pravilima VADER (engl. Valence *Aware Dictionary and sEntiment Reasoner*) [19] nastoji da odredi sentiment finansijskih tvitova i vesti na osnovu jezičkih pravila. Ne zahteva prethodno obučavanje i kao takav spada u modele nenadgledanog učenja. Zbog specifičnosti jezičkih pravila ovaj model ne zahteva prethodno pretprocesiranje podataka.

Loughran McDonald model [20] zasnovan je na specijalizovanim finansijskim rečnicima. Specijalizovani finansijski rečnici sadrže finansijske izraze i žargon. Za potrebe analize sentimenta finansijskih tvitova i vesti rečnik opšte namen je proširen finansijskom terminologijom. Svakoj reči u rečniku pridružen je sentiment. Na taj način formiraju se liste sa pozitivnim i negativnim finansijskim rečima. Ovaj model spada u kategoriju nenadgledanih modela.

Za potrebe klasifikacije finansijskih tvitova i vesti u radu su se koristili tradicionalni modeli mašinskog učenja (*Support-vector machine*, *Random forest*, *Logistic regression*, *Naive Bayes*). Ovi modeli spadaju u modele nadgledanog učenja.

DistilBERT [21] je model zasnovana na neurosnkim mrežama koji se u radu koristi za sentiment analizu. Spada u grupu modela nadgledanog učenja. Razvio se kao manja, brža i jeftinija verzija BERT-a [21]. Za potrebe analize sentimenta finansijskih tvitova koristio se postupak *transfer learning*. Opšti pretrenirani DistilBERT se postupkom *fine-tuning* kalibrisao na finansijsku terminologiju.

### B. *Regresioni problemi*

Regresioni problemi predstavljaju treću tematsku celinu u ovom radu. Nakon što se odredi sentiment finansijskih tvitova i vesti potrebno je utvrditi korelacije između kretanja cene finansijske aktive i sentimenta finansijskih tvitova. Osnovni zadatak ove celine je da ispita uticaj koji finansijski tvitovi i vesti, kao izraz javnog mnjenja, imaju na kretanje cene finansijske aktive. Regresioni modeli su formirani sa ciljem testiranja sledeće hipoteze: Da li će dodavanje rezultata sentiment analize u regresioni model doprineti smanjenju greške regresionog modela i poboljšati prediktivnost regresionog modela.

Kao regresioni model u ovom radu se koristila LSTM mreža. Dok se za potrebe utvrđivanja korelacije između kretanja cene finansijske aktive i sentimenta finansijskih tvitova koristi Pirsonov koeficijent korelacije.

## V. Rezultati i diskusija

U ovom poglavlju predstavljeni su rezultati sentiment analize finansijskih tvitova i rezultati predikcije cene finansijske aktive.

### A. *Rezultati sentiment analize*

Evaluacija modela sentiment analize se vrši nad test podacima koji su izdvojeni kao 20% od ukupnog skupa podataka, dok se preostalih 80% opet deli u odnosu 80:20 na trening skup podataka i validacioni skup podataka.

Za evaluaciju modela klasifikacije finansijskih tvitova koristila se tačnost. TABLE I. predstavlja komparativnu analizu performansi modela sentiment analiza. Ono što se uočava je da model DistilBERT postiže bolje rezultate od svih ostalih modela. Razlog zašto DistilBERT postiže bolje rezultate u odnosu na sve ostale modele leži u činjenici da je DistilBERT treniran na izuzetno velikom skup podataka i kao takav ima bolju moć generalizacija i prilagodljivosti. Loše performanse prikazali su modeli nenadgledanog učenja. Razlog zašto ovi modeli imaju loše performanse je taj što su ovi modeli zasnovani na isključivo finansijskoj terminologiji dok su podaci koji su bili na raspolaganju više opšteg karaktera.

TABLE I. Performanse modela sentiment analize

| ***Model*** | ***Preciznost*** |
|---|---|
| VADER | 0.55 |
| Loughran McDonald | 0.57 |
| Support-vector machine | 0.70 |
| Random forest | 0.74 |
| Logistic regression | 0.73 |
| Naive Bayes | 0.68 |
| DistilBERT | 0.88 |

### B. *Rezultati predikcije cene finansijske aktive*

Evaluacija regresionog modela vrši se nad test podacima koji su izdvojeni kao 20% od ukupnog skupa podataka, dok se preostalih 80% opet deli u odnosu 80:20 na trening skup podataka i validacioni skup podataka.

Za evaluaciju regresije koristila se $R^2$ i MSE metrika, dok se za evaluaciju korelisanosti koristio koeficijent korelacije.

Osnovi zadatak regresionog modela bio je da ispita hipotezu: Da li će dodavanje rezultata sentiment analize u regresioni model doprineti smanjenju greške regresionog modela i poboljšati prediktivnost regresionog modela. Fig. 6 nam je potvrdila da postoji pozitivna korelacija između kretanja cene finansijske aktive i sentimenta finansijskih tvitova. Dodavanjem rezultata sentiment analize u regresioni model potvrdila se hipoteza tj. metrika $R^2$ je porasla dok je MSE metrika opala. Fig. 7 pokazuje visoku prediktivnu moć regresionog modela koja je posledica proširivanja regresionog modela rezultatima sentiment analize.

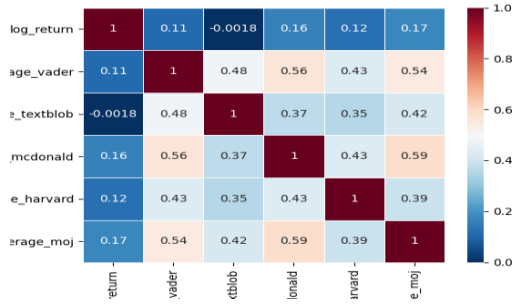


Fig. 6. Pirsonov koeficijent korelacije

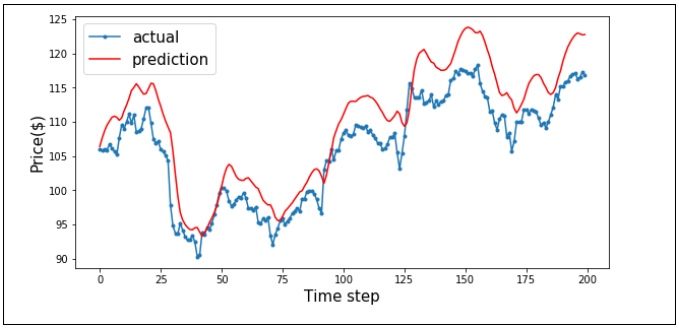


Fig. 7. Rezultati predikcije nakon dodavanja rezultata sentiment analize u regresioni model

## VI. Zaključak

### A. *Sumarizujte rad*

Rad predstavlja primenu veštačke inteligencije u domenu finansija. Domen finansija pokazao se kao izuzetno privlačan i izazovan za primenu tehnika AI. Osnovna motivacija je produbljivanje znanja iz oblasti veštačke inteligencije u domenu finansija ali i aplikativnost softverskog rešenja u procesu donošenja investicionih odluka. Na donošenje investicionih odluka pored kretanja berzanskih indeksa utiče i investiciono raspoloženje javnog mnjenja.

### B. *Predložite pravce budućeg rada*

Pravci budućeg rada ogledaju se u činjenici da su za bolje performanse modela neophodni kvalitetniji podaci stoga bi se sledeće istraživanje usmerilo na kvalitetnije labeliranje podataka.

## VII. Bibliografija

[1] T. Picketty, “Capital in the Twenty-First Century,” 2013.

[2] P. Vigna and M. Casey, “The Age of Cryptocurrency,” 2015.

[3] Y. Kitao, “Learning Practical FinTech from Successful Companies,” 2018.

[4] B. Graham, “The Intelligent Investor,” 2006.

[5] M. Lewis, “Moneyball: The Art of Winning an Unfair Game,” 2003.

[6] H. Haylitt, “Economics in One Lesson: The Shortest & Surest Way to Understand Basic Economics,” 1946.

[7] P. Goldfinch, “A Global Guide to FinTech and Future Payment Trends,” 2018.

[8] G. Heleman and M. Rauchs, “Global cryptocurrency benchmarking study, Cambridge Centre for Alternative Finance,” 2017.

[9] M. Saveski and M. Grčar, “Web Services for Stream Mining: A Stream-Based Active Learning Use Case,” 2011.

[10] C. C. Chen, H. H. Huang and H. H. Chen, “Fine-Grained Analysis of Financial Tweets,” 2018.

[11] S. Ao, “Sentiment Analysis Based on Financial Tweets and Market Information,” 2018.

[12] [Online]. Available: https://python-yahoofinance.readthedocs.io/en/latest/api.html

[13] [Online]. Available: https://github.com/twintproject/twint

[14] [Online]. Available: https://www.kaggle.com

[15] [Online]. Available: https://metatext.io/datasets

[16] [Online]. Available: https://data.world

[17] [Online]. Available: https://lionbridge.ai

[18] [Online]. Available: https://atheros.ai/blog/text-classification-with-transformers-in-tensorflow-2

[19] C. H. E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” 2014.

[20] [Online]. Available: https://sraf.nd.edu/textual-analysis/resources/

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