Sistem za podršku investicionom odlučivanju (Investment Decision Support System)

Motivacija

Rad bi predstavljao nastavak primene veštačke inteligencije u domenu finansija obogaćen novim saznanjima i tehnikama koje su tema predmeta SIAP. 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.

Obrađivane teme

1. Klasterovanje finansijske aktive (ORI). Primena klasterovanja u postupku određivanja tipa finansijske aktive u nastojanju da se izvrši diverzifikacija.
2. Regresiona analiza fluktuacije cene bitcoin-a (PIGKUT). Komparacija različitih regresionih modela na primeru fluktuacije cene bitcoin-a.
3. Analiza i predikcija vremenskih serija (SOFT). Komparacija tradicionalnih statističkih modela za predikciju (AR, MA, ARIMA) i state of the art modela zasnovanih na neuronskim mrežama (LSTM).
4. Sistem za podršku investicionom odlučivanju (SIAP). Finansijski LDA (FinLDA), Analiza i prikupljanje finansijskih vesti (Web Scraper) i analiza i prikupljanje tvitova (Sentiment Analysis of Financial Tweets) akcenat u ovom radu u odnosu na prethodne je na raznorodnosti podataka.

Pegled vladajućih stavova i shvatanja u literature

1. Martin Saveski and Miha Grčar, (2011) "Web Services for Stream Mining: A Stream-Based Active Learning Use Case", Jožef Stefan Institute, Ljubljana, Slovenia
   https://web.media.mit.edu/~msaveski/assets/publications/2011_active_learning/paper.pdf
   Cilj rada:
   Cilj ovog rada je bio da se izvrši analiza sentimenata finansijskih tvitova upotrebom tehnika mašinskog učenja konkretno SVM klasifikatora.
   Metodologija:
   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.
   Skup podataka:
   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.
   Evaluacija rešenja:
   U radu je korišćena podela na trening i test skup. Koristeći AL-SVM postignuta je tačnost od 85%.
   Prednosti rešenja:
   Ideja da se problem klasifikacije sentimenta finansijskih tvitova može rešavati upotrebom standardnih tehnika mašinskog učenja.
   Doprinos za naš rad:
   Kao izvor podataka koristiće se tvitovi sa tagom “$AAPL”. Upotreba standardnih tehnika mašinskog učenja za rešvanje problema klasifikacije sentimenata finansijskih tvitova (SVM, NB, Random Forest). Za evaluaciju klasifikacije koristiće se iste metrika tj. tačnost.
   Nedostaci rešenja:

Autori su se ograničili samo na upotrebu jedne tehnike za klasifikaciju, a nisu uzeli u obzir neke naprednije tehnike kao sto su leksički zasnovana analiza sentimenta ili analiza sentimenta na osnovu neuronskih mreža (RNN i BERT) i NLP tehnika.

2. Chen, C.-C., Huang, H.-H., & Chen, H.-H. (2018). Fine-Grained Analysis of Financial Tweets. Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18. doi:10.1145/3184558.3191824
https://sci-hub.do/10.1145/3184558.3191824
Cilj rada:
Cilj ovog rada je bio da se izvrši analiza sentimenata finansijskih tvitova upotrebom neuronskih mreža (CNN, RNN i LSTM). Pored analize sentimenata finansijskih tvitova ideja je da se izvrši i predikcija kretanja cena Apple korporacije u budućnosti.
Metodologija:
Rađena je komparativna analiza više modela zasnovanih na neuronskim mrežama. U ovom radu se pored problema klasifikacije finansijskih tvitova rešava i problem predikcije buduće cene akcije. Metodološki se problem klasifikacije tvitova uliva u problem regresije. Kada se dobije sentiment tvitova on predstavlja dodatni feature za regresioni problem.
Skup podataka:
Autori su na raspolaganju imali 334K 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.
Evaluacija rešenja:
U radu je korišćena podela na trening i test skup. Za potrebe evaluacije klasifikacije koristi se tačnost i F1 mera. Dok se za potrebu regresije koristi R^2 i MSE.
Prednosti rešenja:
Koristi se komparativni pristup istraživanju modela. Jedinstven konglomerat više različitih tipova problema. Sveobuhvatan metrički sistem za evaluaciju modela.
Doprinos za naš rad:
Komparativni pristup izučavanju modela. Jedinstveni pristup disekcije realnog problema na probleme nadgledanog učenja i regresije. Primena sveobuhvatnog metričkog sistema za evaluaciju koji će omogućiti komparaciju različitih modela.
3. Ao, S. (2018). Sentiment Analysis Based on Financial Tweets and Market Information. 2018 International Conference on Audio, Language and Image Processing (ICALIP). doi:10.1109/icalip.2018.8455771
https://sci-hub.do/10.1109/icalip.2018.8455771
Cilj rada:
Cilj ovog rada je bio da se izvrši leksički zasnovana analiza sentimenata finansijskih tvitova. Takođe u radu je izvršeno ispitivanje korelacije između sentimenata finansijskih tvitova i kretanja cene akcija.
Metodologija:
Leksički pristup podrazumeva postojanje unapred pripremljenog rečnika u kojima je dostupan polaritet reči (reči su anotirane kao pozitivne, negativnei neutralne na osnovu toga koliko se često pojavljuju u odgovarajućim kontekstima). Napomena: Zbog mogućnosti posmatranja problema klasifikacije kao problema regresije gde se iz binarnog domena (pozitivan/negativan sentiment) prelazi u kontinualni domen [-1,1] imamo mogućnost da računamo koeficijent korelacije između prosečne vrednosti sentimenata za posmatrani trgovački dan i cene posmatrane akcije.
Skup podataka:
Autori su na raspolaganju imali 15K 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.

Evaluacija rešenja:
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.
Prednosti rešenja:
Nema striktne podele na pozitivan negativan sentiment vec uvodi fuzzy logiku koja nam omogućava primenu koeficijenta korelacije kao metrike.
Doprinos za naš rad:
Uvodi novi metodološki pristup zasnovan na leksičkoj analizi sentimenta kao i novu metriku zasnovanu na koeficijentu korelacije.
Nedostaci rešenja:
Mali skup podataka. Nedostatak metrika za evaluaciju klasifikacje.

Podaci

Mana prethodnih modela i projekata je ta što se koriste homogeni izvori podataka. Za prethodne modele koristili su se podaci iz jednog izvora podataka koji su strukturirani po svojoj prirodi (tabelarni izveštaji sa berze). 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.

Yahoo Finance REST API

Prikupljanje tabeliranih berzanskih izveštaja (strukturirani podaci).

Web Scraper

Prikupljanje podataka o fundamentalnim pokazateljima poslovanja korporacija koji se tiču profitabilnosti i likvidnosti (nestrukturirani podaci). Podaci bi se preuzimali sa sajtova korporacija ili sa specijalizovanih sajtova za praćenje ekonomskih tema cnbc, bloomberg i yahoo. Podaci koji su od interesa za analizu: da li je došlo do nekih merdžera i akvizicija, da li je došlo do isplate dividendi, da li se najavljuje neki novi prototip proizvoda, da li su se pojavile neke tužbe protiv korporacija, podaci o konkurenciji, podaci o zasićenosti tržista proizvodima, podaci o strajkovima radnika, geopolitički podaci, etc.

Twitter REST API

Primećeno je da pored podataka o samoj korporaciji (fluktuacija cena i pokazatelji uspešnosti poslovanja) na odluku da li će se investirati u neku korporaciju utiče i opšta slika javnog mnjenja o dotičnoj korporaciji. Na formiranje opšte slike javnog mnjenja o nekoj korporaciji utiču objave u vestima ali i objave na društvenim mrežama konkretno Twitter. Potrebno je pokazati da postoji korelacija između raspoloženja twitter korisnika i berzanskog indeksa tj. da postoji korelacija između objavljenih novinskih članaka i vrednosti akcija. U nastavku će biti navedeni samo neki od primera koji su se odrazili na vrednost korporacije nakon objave u medijima:

1. Najava snižavanje kamatne stope od strane Alan Greenspan
2. Najava smene top management –a Microsoft-a
3. Problemi sa baterijom Apple telefona
4. Pobeda na predsedničkim izborima (Demokrate vs Republikanci)

Sve tipove podataka potrebno je zasebno izanalizirati i agregirati kako bi se investitoru predočila što jasnija slika o potencijalnim ulaganjima. Usled tržišne nesavršenosti (nemaju svi akteri na tržištu jednak volume i kvalitet informacija) potrebno je više izvora informacija angažovati kako bi se povećala profitabilnost investitora i smanjio rizik koji sa sobom nosi svako ulaganje.

Tvitovi koji bi se prikupljali bili bi na engleskom jeziku sa tagom “$AAPL”. Skupljali bi se tvitovi o korporaciji Apple.

Metodologija

Prilikom donošenja investicionih odluka mora se pristupiti tehničkoj (Technical Analysis) i fundamentalnoj analizi finansijskih tržišta (Fundamental Analysis). Tehnička analiza ima zadatak da prati kretanje cena akcija neke korporacije i da na osnovu fluktuacija u ceni zaključi da li je lukrativno ulagati u tu korporaciju. Sa druge strane pored tehničke analize imamo i fundamentalnu analizu koja se bavi fundamentalnim pokazateljima poslovanja jedne korporacije (Profitabilnost (PE Ratio), Zarada po akciji (EPS), Dividende, know-how). Tehnička analiza bi se bavila podacima prikupljenim preko Yahoo Finance REST API-a. Dok bi se fundamentalna analiza bavila fundamentalnim pokazateljima poslovanja koji su prikupljeni preko Web Scraper-a. Usled velike kolicine podataka koja se generiše svakodnevno u medijima, potrebno je obratiti posebnu pažnju i na te podatke ako se želi opsežnija analiza intrinsičke vrednosti neke korporacije. Na unutrašnju vrednost neke korporacije pored fundamentalnih pokazatelja poslovanja i tržišnog kretanja akcija utiče i opšta slika javnog mnjenja o korporaciji. Preko Twitter REST API bi se prikupili podaci relevantni za neku korporaciju. Tako prikupljeni podaci predstavljali bi ulaz u modele ML. Ideja je da se odradi sentiment analiza finansijskih tvitova i da se odradi finansijski LDA tj. detekcija finansijskih tema.

Sentiment analiza finansijskih tvitova

Potrebno je da se odradi sentiment analiza prikupljenih tvitova i da se oni klasifikuju u pozitivne ili negativne. Pozitivan tvit bi predstavljao tvit koji bi u sebi oslikavao prosperitet kompanije u koju bi se ulagalo. Dok bi negativan setntiment sadržao informacije o dekadenciji kompanije. Reakcija na pozitivan odnosno negativan sentiment ogledala bi se u tržišnom kretanju. Pozitivan sentiment uzrokovao bi rast tržišne krive dok bi negativan sentiment uzrokovao pad tržišne krive. Sentiment analiza finansijskih tvitova bi omogućila da se utvrdi temporalna veza izmedju pojave finansijskih tvitova i kretanja berze. Na osnovu timestamp –a bi se povezivali podaci (vremenska serija i vreme objave tvita). Ovako sintetisan podatak bi poslužio da se utvrdi koji su to događaji uticali na tržišno kretanje i da li bi se takvi događaji mogli predvideti u slučaju pojave sličnih tvitova.

Ideja je da se uradi komparacija više modela za analizu sentimenata finansijskih tvitova:

1. Pristup zasnovan na leksičkoj analizi sentimenata i upotrebi standardnih ML algoritama (SVM, NB, Random Forest)
2. Pristup zasnovan na NLP-u i NN-u (BERT i RNN)

Na ovaj način napravila bi se paralela između tradicionalnih ML algoritama i algoritama zasnovanih na NN.

Finansijski LDA

Nakon sto su prikupljeni podaci iz vodećih ekonomskih časopisa i sajtova potrebno je izvrši detekciju topika. Ideja je da se na osnovu dominantnih oblasti koje su prisutne u medijima uspostavi korelacija željenog ulaganja sa nekom od oblasti . Npr. ako se često spominju tehnološki giganti , stimulišuće mere monetarne i fiskalne politike možda je pravi momenta da se investitor opredeli za ulaganje u neku od spomenutih kompanija zato što one predstavljaju “Hot stuff” i vredno je u njih ulagati. Finansijski LDA bi mogao da nam ponudi koje su to oblasti vredne ulaganja, a koje bi oblasti trebali da izbegnemo ukoliko želimo sigurno i profitabilno ulaganje.

Predikcija budućeg kretanja cene

Rad bi predstavljao sintezu dva problema: problema klasifikacije finansijskih tvitova i problema predikcije budućeg kretanja cene akcija. Rešenje problema klasifikacije se tretiraju kao novi feature-i u problemu regresije. Ovako se pravi jedan pipe modela kroz koji podaci protiču. Za potrebe predikcije koristiće se MLP.

Softver

Aplikacija će biti kompletno izrađena u programskom jeziku Python. Za potrebe skladištenja podataka koristiće se SQL/NoSQL baza podataka.

Evaluacija

Isptitivala bi se vremenska koreliranost (kauzalnost) između vremena objave tvita i reakcije berze, kao i uticaj detektovanih tema na investicionu odluku. Evaluacija će se vršiti nad test podacima koji će biti izdvojeni kao 20% od ukupnog skupa podataka, dok će preostalih 80% opet biti podeljeno u odnosu 80:20 na trening skup podataka i validacioni skup podataka. Nakon evaluacije modela biće odrađena analiza grešaka. Analiza grešaka je izuzetno važan deo rada. Ručno će se izvojiti određeni podskup primera na kojima modeli greše i analizirati zbog čega dolazi do grešaka.

Metrički sistem

Za evaluaciju klasifikacije finansijskih tvitova koristiće se tačnost ili F1 mera u zavisnosti od da li jedna grupa (pozitivni/negativni tvitovi) dominira u ukupnom broju tvitova. Ako jedna grupa dominira (neuravnotežen odnos klasa)koristiće se F1 mera za evaluaciju.

Za evaluaciju regresije koristice se R^2 i MSE.

Za evaluaciju korelisanosti koristiće se koeficijent korelacije.

Plan rada

Plan rada na projektu obuhvata sledeće tačke:

1. Eksplorativna analiza podataka (EDA)
2. Obrada podataka
3. Obučavanje modela
4. Evaluacija modela

Članovi tima

Dušan Stević R2-33/2020