Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/04BE13Milutinovic

PREDIKCIJA VREDNOSTI KRIPTOVALUTA ANALIZOM ISTORIJSKIH CENA, BLOKČEJN INFORMACIJA I SENTIMENTA TVITOVA

PREDICTING THE VALUES OF CRYPTOCURRENCIES USING HISTORICAL VALUES, BLOCKCHAIN INFORMATION AND TWITTER SENTIMENT

Milica Milutinović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Ovaj rad bavi se problemom predikcije cene Bitkoina na osnovu istorijskih cena i blokčejn informacija. Cenu kriptovaluta velikim delom diktiraju špekulacije pa je ispitano u kojoj meri sentiment analiza tvitova doprinosi padu ili rastu cena. Pored podataka o ceni i sentimentu, korišćeni su blokčejn podaci i istorijski podaci za još tri popularne kriptovalute: Lajtkoin, Etereum i Ripl. Za sentiment analizu ispitane su tri tehnike: konvolucione neuronske mreže, ansambli leksikona i metoda zasnovana na jezičkim pravilima. Rezultati sentiment analize upotrebljeni su u daljem procesu predviđanja kriptovaluta. Izabrani model za predikciju je rekurentna neuronska mreža sa GRU ćelijama. U fazi evaluacije posmatrala se tačnost predviđanja kretanja cene, odnosno da li će ona da raste ili opada. Takođe, upotrebljena je i relativna tačnost u kojoj se posmatra odnos stvarne cene i prediktovane. Najbolji rezultat dostigao je 57,3% tačnosti u predikciji kretanja cene, odnosno 99,39% relativne tačnosti.*

Ključne reči: *sentiment analiza; predikcija; Bitkoin;*

Abstract – *This paper studies the problem of Bitcoin price prediction based on historical prices and blockchain information. As the price of a cryptocurrency is dictated largely by speculation we also examine the influence of twitter sentiment on the performance of our predictive models. We have also experimented with historical data for three other popular cryptocurrencies: Litecoin, Ethereum and Ripple. We experimented with three techniques for sentiment mining: convolutional neural networks, ensemble of lexicons and linguistic rules. The results of sentiment analysis were integrated with other data to train a recurrent neural network with GRU cell as our price prediction model. We evaluated our model both as a binary classifier (predicting whether the price will go up or down) and a regression predictor (predicting the actual price). The accuracy of the classification model was 57.3%, while the relative accuracy of the regression model was 99.39%.*

Keywords: *sentiment analysis; prediction; Bitcoin;*

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, van. prof.

1. UVOD

Kriptovalute predstavljaju oblik digitalne imovine koje su zasnovane na kriptografskom protokolu koje omogućavaju korisnicima da ih čuvaju i razmenjuju putem mreže [1]. Satoshi Nakamoto je u januaru 2009. implementirao Bitkoin (en. *Bitcoin*), što predstavlja prvu kriptovalutu. Od tada njihova popularnost stalno raste, najviše zbog osobina kao što su decentralizovanost, transparentnost i sigurnost. U ovom radu će biti opisani pristupi i rešenja za predviđanje cene Bitkoina. U poređenju sa „klasičnim“ akcijama na berzi, za kriptovalute se mogu vezati informacije o samoj blokčejn tehnologiji koje su javno dostupne što uvodi dodatnu semantiku [2].
U narednoj sekciji je izložen pregled postojeće literature vezane za problem predviđanja cena kriptovaluta i sentiment analize podataka. Treća sekcija sadrži informacije o prikupljanju podataka. Četvrta sekcija opisuje metodologije i alate korišćene u radu. U petoj sekciji objašnjena je eksperimentalna evaluacija rezultata, dok poslednja sekcija zaključuje rad predlaže pravce mogućeg daljeg istraživanja.

2. PRETHODNA REŠENJA

Dosadašnja literatura pokriva veliki broj problema koji teže da se objasne u radu. Proučavanja sentiment analize teksta su dostigla visok nivo, a pristupi koji su rađeni variraju od metoda koje ne koriste nikakav vid učenja odnosno koja se baziraju na pravilima i/ili leksikonima, do metoda koje koriste mašinsko učenje za te svrhe. Metod zasnovan na pravilima [3], VADER alat, teži da odredi sentiment na osnovu jezičkih konstrukcija, odnosno pravila koje vladaju u njima. Zbog izuzetnih rezultata koji su dobijeni (0,96 F1 mera tačnosti za sentiment tvitova), implementacija ovog metoda je korišćena kao jedan od načina za određivanje sentimenta. Pristup koji određuje sentiment teksta baziran na ansamblima leksikonima [4] traži reči u formiranim leksikonima i u zavisnosti od broja ponavljanja u njima, dodeljuju svakoj reči jedinstven sentiment. Ovaj metod unapredio je proste metode bazirane na leksikonima koji su imali mnogih nedostataka [5], a dostigao je približnu tačnost kao i metode mašinskog učenja u određivanju sentimenta teksta. Iz tih razloga, implementiran je, a detalji će biti objašnjeni kasnije u radu. Analiziranje sentimenta u tekstu često se rešava primenom algoritama dubokog učenja. Rad zasnovan na konvolucionim neuronskim mrežama [6] koristi konvolucionu arhitekturu neuronske mreže i tzv. Bag-of-Words model [7] kao

ulazne podatke u mrežu. Iako arhitektura u pomenutom radu nije velike dubine, podešavanjem parametara mreže i korišćenjem regularizacije u procesu učenja, ovaj model pokazao je bolje rezultate od drugih koji su upoređivani u istom radu.
U radu koji se bavi predikcijom volatilnosti akcija na kineskom tržištu [8], korišćenjem rekurentne neuronske mreže sa sentimentalnim podacima, dobijena je tačnost od 65%. Međutim, u tom radu predikcija je vršena samo za jedan vremenski korak unapred.

3. SAKUPLJANJE PODATAKA

Skup podataka koji je analiziran dobijen je prikupljanjem tvitova influensera u domenu kriptovaluta, istorijskih cena i informacija o blokčejnu u periodu od Januara 2016. godine do kraja Juna 2018. godine.
Tvitovi su prikupljani zahvaljujući projektu [9] koji formira url na osnovu koga skrejpuje (eng. *scrapping*) stranicu preko koje sakuplja tvitove. Zahvaljujući ovom pristupu preuzet je veliki broj tvitova u kratkom vremenskom intervalu. Nakon procesa sentiment analize, dobijeni rezultati prosleđeni su u fazu predikcije. Iskorišćeno je 7 numeričkih atributa u daljem toku:

- vader_sent i vader_pol – intezitet i polaritet sentimenta dobijen korišćenjem VADER alata
- lex_sent - sentiment dobijen korišćenjem modela ansambla leksikona
- cnn_sent i cnn_pol - intezitet i polaritet sentimenta dobijen korišćenjem CNN arhitekture
- favorites - predstavlja broj "lajkovanja" tvita
- retweets – predstavlja broj deljenja tvita

Istorijski podaci o cenama su preuzeti sa Poloniex API-ja za Bitkoin, Lajtkoin, Etereum i Ripl. Podaci se nalaze u vremenskom koraku od 5 minuta za period od Januara 2016. godine do Aprila 2018. godine. Eksplorativnom analizom atributa je utvrđeno da podaci o ceni imaju veliku međusobnu korelaciju (približno 1) te je za dalju obradu korišćeno: close (cena na kraju intervala), date (vreme u *unix timestamp*-ovima), volume (trgovani iznos u poslednja 24h u dolarima) i quoteVolume (trgovani iznos u poslednja 24h u datoj kriptovaluti). Takođe je zaključeno da postoji korelacija između različitih valuta, tako da su za konačnu predikciju cene Bitkoina validirani i podaci od drugih valuta.

Preostalih 10 atributa su podaci o Bitcoin blokčejnu koji su preuzeti sa blockchain.info API-ja i unose dodatnu semantiku o transakcijama ove kriptovalute.

S obzirom da prikupljeni tvitovi nemaju pravilan vremenski interval, a podaci o blokčejnu se nalaze na uniformnom intervalu od 1 dan, nije ih bilo moguće spojiti sa podacima o cenama koje imaju unforman interval od 5 minuta. Stoga su oni interpolirani na isti vremenski raspon. U radu je upotrebljena interpolacija na 1 sat i na 10 minuta.

Za proces formiranja rečnika u metodologiji koja koristi ansamble leksikona upotrebljeno je 5 skupova podataka koji sadrže anotirane tvitove: *Stanford Twitter Sentiment Test Set* [10], *Sentiment Strength Twitter Dataset* [11] , skup od 9637 tvitova skinutih sa *Kaggle*-a, skup od 50800 tvitova skinutih sa *Kaggle*-a i *Senticnet* [12].

4. METODOLOGIJA

U cilju istraživanja sentimenta tvitova, isprobana su tri popularna rešenja. Rezultati metoda su upotrebljeni da se odredi u kojoj meri doprinose poboljšanju predikcije.
U sva tri pristupa, tvitovi su pretprocesirani primenom nekoliko koraka: izbacivanjem emotikona, akronima i skraćenica, pretvaranjam velikih u mala slova, uklanjanjem *URL*-ova, praznih mesta, stop reči, znakova interpunkcije i specijalnih karaktera. Potom je primenjivana morfološka normalizacija reči. Skupovi podataka korišćeni u procesu formiranja leksikona pretprocesirani su na sličan način, s tim da su im i ciljne labele svedene na isti interval, odnosno 1 za pozitivan, -1 za negativan, a 0 za neutralan tvit.
Prvi isprobani pristup koristi jezička pravila za računanje sentimenta [3]. U pomenutom pristupu posmatra se grupa reči koja predstavlja neku celinu, pa se njoj određuje sentiment, a ne samo pojedinačna reč. Najbolji rezultati se dobijaju ukoliko se radi analiza tzv. mikro-blogova, odnosno kratkih statusa. Implementacija ovog pristupa je javno dostupna i preuzeta je.
Drugi način bazira se na formiranju leksikona, koji se potom koriste da se pronađu reči koje imaju pozitivan ili negativan sentiment. Ovom metodologijom najpre su generisani leksikoni koristeći pomenute skupove. Leksikoni su kreirani tako što su se pronalazile reči koje su najzastupljenije u pozitivnim, odnosno negativnim recenzijama. Na ovaj način, napravljen je skup od 10 rečnika za navedenih 5 skupova, odnosno rečnici sa po 5 i 25 najpozitivnijih i najnegativnijih reči za svaki skup posebno. Formiran je i jedan rečnik koji je agregira sve skupove podataka zajedno i među njima traži skup od 25 najpozitivnijih i najnegativnijih reči. Preostala 4 rečnika formirana su na osnovu reči koje jednoznačno doprinose promeni inteziteta sentimenta i preuzeti su iz rada [13].
Sledeći korak predstavlja klasifikaciju sentimenta na pozitivan ili negativan na osnovu ansambla rečnika. Ovaj model uzima pojednačne reči iz rečenice i predviđa njihov sentiment tako što traži da li se ta reč pojavljuje više kao pozitivna ili u negativna. Ukoliko preovladava pojavljivanje reči u leksikonima sa negativnim rečima, vrednost sentimenta reči će biti -1, odnosno 1 u suprotnom. Ukoliko se reč ne pojavljuje ni u jednom leksikonu, vrednost njenog sentimenta će biti 0. Nakon što se svakoj reči dodeli sentiment, računa se ukupna vrednost sentimenta na nivou teksta, odnosno tvita u našem slučaju. Ukupan sentiment za leksikon l i tvit t računa se po formuli:

$$s_l = \begin{cases} 1, \wedge\, if \sum (l,t) > 0 \\ -1, \wedge\, if \sum (l,t) < 0 \\ 0, \wedge\, otherwise \end{cases}$$

Konačan rezultat sentiment analize za jedan tvit dobija se usrednjavanjem rezultata svih leksikona. Ukoliko je izračunata vrednost veća od 0, uzima se da je tvit pozitivan, ukoliko je manja od 0, smatra se da je tvit negativan, a u suprotnom je tvit neutralan.
Treći pristup korišćen u ovom radu zasniva se na primeni konvolucionih neuronskih mreža, koje koriste konvolucioni operator da proizvode nove osobine. Ulaz u mrežu dobijen je formiranjem tzv. Bag-of-Words modela i predstavlja matricu fiksne dužine $N \times K$, gde N predstavlja

broj tvitova u datom intervalu, a *K* fiksnu dužinu rečenica na kojima je izvršeno upotpunjavanje (*padding)* ili odsecanje (*truncate)* u odnosu na fiksnu dužinu. Izlaz iz mreže daje dve vrednosti, tj. odnos rasta ili pada cene i oznaku da li cena raste ili opada.

Kako se o tvitovima čuvaju informacije o danu kada su kreirani, izlazni podaci u procesu treniranja se dobijaju tako što se računa odnos cena tog dana i sutrašnjeg dana i oznaku da li cena raste ili opada. Na primer, ukoliko je tvit postavljen dana 10.12.2017. posmatra se cena tog dana i cena 11.12.2017, računa se količnik cena ta dva dana, a postavlja se i indikator da li cena raste ili opada. Zbog nemogućnosti da se odredi čiji je tvit važniji, svi tvitovi objavljeni istog dana imaju isti odnos cena i isti indikator koji govori o kretanju cene kriptovalute.

Za proces računanja predikcije korišćene su rekurentne neuronske mreže jer imaju sposobnost da adresiraju vremenske podatke koristeći unutrašnju memoriju.

S obzirom da su svi atributi numerički, odrađena je normalizacija u intervalu [0,1] na svim atributima. Kako bi se kreirala veća količina podataka upotrebljen je preklapajući obučavajući i trening skup. Ova redudantnost omogućena je korišćenjem „klizećeg" vremenskog okvira za svaki skup ulaznih i izlaznih vrednosti.

Odabir arhitekture RNN mreže je dobijen eksperimentalno nakon što je testirano više različitih arhitektura. GRU koristi manje parametara od LSTM (eng. *Long short-term memory*), a performanse se neznatno razlikuju. Korišćene su GRU ćelije koje sadrže 96 neurona (ćelija) u jednom sloju. Na izlazu je Fully-Connected sloj sa 24 izlaza koji predstavljaju vremenske korake predikcije. Kako bi se sprečio overfitting, dodata je Dropout regularizacija. Testirani su modeli sa jednim i dva unutrašnja skrivena sloja. Za funkciju greške odabrana je MSE (*mean squared error*), gde je takođe korišćena Adamova optimizacija.

5. EKSPERIMENTI I REZULTATI

Za osnovnu meru tačnosti je uzet broj uspešno predviđenih kretanja valuta – pad ili rast što može da se svede na binarnu klasifikaciju. Formula po kojoj se ona izračunava:

$$Acc = \frac{successfullpredictions}{totalpredictions}$$

Relativna tačnost, dobija se računanjem odnosa stvarne cene i prediktovane. Osim pomenutih metrika, korišćene su i srednja kvadratna i srednja apsolutna greška, radi upoređivanja tačnosti rezultata.

U postupku evaluacije kreirano je više skupova kako bi se validirali podaci dobijeni u postupku sentiment analize. Urađeni su i testovi radi ispitivanja parametara koji dovode do poboljšanja rezultata. Tim putem ispitano od kolikog je značaja vremenski raspon u kojem su interpolirani podaci, broj ulaznih vremenskih koraka u proces obučavanja, kao i informacije o drugim vrednostima kriptovaluta.

Interesantno je primetiti da informacije o sentimentu influensera dovode do boljih rezultata u određenoj meri. Uočeno je da povećanje broja ulaznih jedinica, kao i informacije o vrednostima drugih kriptovaluta ne doprinose boljim rezultatima. Ipak, najveće poboljšanje postignuto je u slučaju smanjenja intervala interpolacije sa 1. sata na 10 minuta. U Tabeli 1. prikazani su rezultati za najbolji validirani test, odnosno za za slučaj interpolacije na 10 minuta, gde se prediktuju vrednosti za naredna 2 sata, na osnovu vrednosti u prethodna 24 sata. Oznake skupova sastoje se od atributa koji ih sačinjavaju: **b** (blokčejn podaci), **h** (istorijske podaci), **s** (sentiment podaci), **v** (sentiment podaci dobijeni vader metodom), **c** (sentiment podaci dobijeni cnn metodom) i **l** (sentiment podaci dobijeni ansamblima leksikona).

Tabela 1. *Rezultati*

	acc (%)	rel_acc (%)	mse	mae
bps	53.31	98,83	9180.78	79.47
bp	52.74	97.8	27721.35	151.15
bpv	47.36	96.1	79079.49	272.38
bpl	54.32	99.2	5056.30	54.41
bpc	52.76	97.22	40463.36	190.12
s	71.22	38.57	56364.57	100.56
ps	57.3	99.39	3731.67	40.63

Slika 1. *Grafik predikcije za skup ps*.

Na slici 1. dat je prikaz najboljeg skupa atributa ps, za date parametre. Može se uočiti da su prediktovani rezultati prilično korelirani sa stvarnim.

6. ZAKLJUČAK

U ovom radu predložen je novi pristup predviđanja cena kombinacijom različitih skupova podataka u cilju precizniji ih rezultata. Podaci o sentimentalnoj analizi uticajnih ljudi na Tviteru doprineli su poboljšanju predikcije, međutim, potrebno je dodatno istraživanje kako bi se ispitao stepen uspešnosti. Ovim rezultatima je potvrđena težina i kompleksnost problema predviđanja cene na berzi. Moguća obrazloženja lošijih rezultata je relativno mali period skupa podataka (2 godine i 6 meseci) i činjenica da je u proteklih godinu dana zabeležen rast nekih kriptovaluta čak i do 10,000% te je stvoren tržišni balon (*market bubble*) koji je polako počeo da puca u januaru 2018. godine.

Budući pravci istraživanja bi mogli da razmotre korišćenje nekih eksperimetalnih modela neuronskih mreža u rešavanju problematike. Takođe, bilo bi pogodno istražiti koliko pojedini tvitovi utiču na promenu cena, pa u skladu sa tim da im se dodele odgovarajuće težine. Pored ovoga, bilo bi zanimljivo istražiti i kako finansijske vesti i neki drugi indikatori utiču i na druge cene na berzi i pored kriptovaluta.

7. LITERATURA

[1] Wikipedia contributors. (2018, April 11). Cryptocurrency. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:46, April 13, 2018, from https://en.wikipedia.org/w/index.php?title=Cryptocurrency&oldid=835965914

[2] Blockchain. (2018, April 13). Retrieved April 13, 2018, from https://en.wikipedia.org/wiki/Blockchain

[3] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014

[4] Augustyniak, L., Kajdanowicz, T., Szymanski, P., Tuliglowicz, W., Kazienko, P., Alhajj, R., & Szymanski, B.K. (2014). Simpler is better? Lexicon-based ensemble sentiment classification beats supervised methods. 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), 924-929.

[5] S.-H. Na, Y. Lee, S.-H. Nam, and J.-H. Lee, "Improving opinion retrieval based on query-specific sentiment lexicon," in Advances in Information Retrieval, ser. Lecture Notes in Computer Science, vol. 5478. Springer Berlin / Heidelberg, 2009, pp. 734–738.

[6] K., & Y. (2014, September 03). Convolutional Neural Networks for Sentence Classification. Retrieved April 13, 2018, from https://arxiv.org/abs/1408.5882

[7] Wikipedia contributors. (2018, March 26). Bag-of-words model. In *Wikipedia, The Free Encyclopedia*. Retrieved 20:07, April 13, 2018, from https://en.wikipedia.org/w/index.php?title=Bag-of-words_model&oldid=832576977

[8] Liu, Y., Qin, Z., Li, P., & Wan, T. (2017). Stock Volatility Prediction Using Recurrent Neural Networks with Sentiment Analysis. *Advances in Artificial Intelligence: From Theory to Practice Lecture Notes in Computer Science,*192-201. Doi:10.1007/978-3-319-60042-0_22

[9] J. (2018, January 29). Jefferson-Henrique/GetOldTweets-python. Retrieved April 13, 2018, from https://github.com/Jefferson-Henrique/GetOldTweets-python

[10] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009)

[11] Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology 63(1), 163–173 (2012)

[12] SenticNet. (n.d.). Retrieved from https://sentic.net/

[13] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011, pp. 142–150.

Kratka biografija:

Milica Milutinović rođena je 6.1.1995. godine u Rumi. Osnovne studije je upisala 2013. godine na Fakultet tehničkih nauka, odsek Računarstvo i automatika. Osnovne studije je završila 2017. godine, nakon čega upisuje master akademske studije na Fakultetu tehničkih nauka, smer Inteligentni sistemi.