Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/16BE42Ogrizovic

AUTOMATSKA SUMARIZACIJA VESTI O KRIPTOVALUTAMA

AUTOMATIC SUMMARIZATION OF CRYPTOCURRENCY NEWS

Uroš Ogrizović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Kriptovalute su dosegle veliku popularnost 2021. godine. U ovom radu su evaluirani aktuelni transformer modeli na zadatku automatske sumarizacije vesti o kriptovalutama. Ulaz modela je sekvenca tokena koja predstavlja tekst vesti. Izlaz modela je sekvenca tokena koja predstavlja sumarizaciju vesti sa ulaza. Kao labela je korišćen naslov vesti. Predloženi su koraci za poboljšanje performansi modela.*

Ključne reči: *mašinsko učenje, transformer, automatska sumarizacija teksta*

Abstract – *This paper utilizes several popular transformer models to perform automatic text summarization of cryptocurrency news. The model's input is a sequence of tokens representing the content of the news, while the model's output is a sequence of tokens representing a summary of the input. The title of each piece of news was used as the label. The paper discusses the steps for further performance improvements.*

Keywords: *machine learning, transformer, automatic text summarization*

1. UVOD

Kriptovaluta je oblik digitalne imovine koja se koristi kao sredstvo razmene. Početkom 2009. godine, programer ili grupa programera pod pseudonim Satoši Nakamoto je predstavio Bitkoin, prvu decentralizovanu kriptovalutu. Od tada je stvoren veliki broj novih kriptovaluta.

Kriptovalute su doživele naglu ekspanziju 2017. godine, kada je njihova ukupna tržišna kapitalizacija porasla sa 11 na 177 milijardi dolara. Na početku 2017. godine, cena Bitkoina je iznosila 450 dolara, a na kraju iste godine 19500 dolara.

Novi rast popularnosti kriptovalute su doživele 2021. godine, kada su i konačno u većoj meri prihvaćene od renomiranih institucija poput kompanije *Visa*[1].

Sa druge strane, kriptovalute su i dalje veoma volatilne, na šta ukazuje činjenica da je vrednost Bitkoina na početku ove godine iznosila 29 hiljada dolara, nakon čega je porasla i dostigla vrhunac od preko 64 hiljade dolara u aprilu, da bi se zatim vratila na 30 hiljada dolara u julu, nakon čega je ponovo rasla do 50 hiljada u septembru.

Jedan od faktora koji je uticao na promenu vrednosti Bitkoina bili su tvitovi Ilona Maska: u prvom[2], iz marta, naveo je kako njegova kompanija Tesla prihvata Bitkoin kao sredstvo plaćanja (i pritom je Tesla kupila određenu količinu Bitkoina), dok je u drugom[3], iz maja, naveo kako Tesla više neće prihvatati Bitkoin kao sredstvo plaćanja jer ta kriptovaluta nije dobra po životnu sredinu. Mnogi su postupke Maska protumačili kao pokušaj manipulacije tržištem.

Oblast sumarizacije teksta je veoma izazovna za mašine zbog toga što one ne poseduju sposobnost razumevanja jezika na nivou na kom je poseduju ljudi. Sa druge strane, s obzirom na to da mašine mogu da čitaju tekst mnogo brže nego ljudi, razvijanje alata za sumarizaciju vesti o kriptovalutama omogućava brže sticanje saznanja o kriptovalutama.

2. PRETHODNA REŠENJA

Postoje dve tehnike sumarizacije teksta:

1. ekstraktivna - iz teksta se biraju najkvalitetnije rečenice po nekoj funkciji za vrednovanje značaja rečenice. Na primer, moguće je odrediti značaj rečenice kao sumu težina reči koje se u njoj nalaze, pri čemu težina reči može da odgovara frekventnosti te reči u tekstu. Izvučene rečenice se spajaju u celinu koja se prosleđuje kao ulaz modelu
2. apstraktivna - ne vrši kopiranje rečenica, već na vrši apstrahovanje suštine teksta. Ovaj pristup više podseća na ono što bi čovek uradio kada bi se od njega zatražilo da izvrši sumarizaciju teksta

U ovom poglavlju su predstavljena prethodna rešenja koja su se bavila apstraktivnom automatskom sumarizacijom teksta.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, vanr. prof.

[1] Vest o prihvatanju kriptovaluti kao sredstava plaćanja od strane *Visa* kompanije, pristupljeno 15.10. 2021. https://www.forbes.com/sites/ninabambysheva/2021/03/29/visa-to-start-settling-transactions-with-bitcoin-partners-in-usdc/?sh=489b3d5f5228

[2] Tvit Ilona Maska o prihvatanju Bitkoina kao sredstva plaćanja, pristupljeno 15.10.2021. https://twitter.com/elonmusk/status/1374617643446063105

[3] Tvit Ilona Maska o prestanku prihvatanja Bitkoina kao sredstva plaćanja, pristupljeno 15.10.2021. https://twitter.com/elonmusk/status/1392602041025843203

U radu [1] su autori predstavili model BART (***B**idirectional **A**uto-**R**egressive **T**ransformer**s**). Cilj ovog rada jeste razvijanje autoenkodera za pretreniranje jezičkog modela (eng. *Language Model*), koji zatim može da se koristi za mnoštvo različitih zadataka, uključujući i sumarizaciju teksta. Glavna ideja se sastoji iz dva koraka: ”kvarenje“ teksta ubacivanjem šuma u isti pomoću proizvoljne funkcije i obučavanja modela da rekonstruiše originalni tekst. Korišćena je arhitektura neuronske mreže zasnovana na transformerima. Autori su dobili najbolje rezultate kada su promešali redosled rečenica u tekstu i iskoristili tačno jedan maskirni token da zamene nijednu, jednu ili više reči u tekstu. Na ovaj način, postignuta je generalizacija *masked language modeling* (MLM) i *next sentence prediction* (NSP) zadataka iz BERT [2] modela, jer model u ovom slučaju ne zna tačnu dužinu rečenice.

U radu [3] predstavljen je *sequence-to-sequence* (Seq2Seq) model ProphetNet zasnovan na transformerima koji uvodi samonadgledani zadatak koji se zove *future n-gram prediction*, kao i *n-stream self-attention* mehanizam. Pored klasičnog Seq2Seq modela koji optimizuje predviđanje za jedan korak unapred, ProphetNet takođe vrši predikciju n koraka unapred koristeći *future n-gram prediction* mehanizam, koji forsira model da ne vrši preprilagođavanje (*overfitting*) nad lokalnim korelacijama u tekstu. Odstupanje od klasičnog transformera prisutno je na strani dekodera, gde se, umesto predviđanja samo sledećeg tokena u svakom vremenskom koraku, vrši predviđanje narednih *n* tokena istovremeno.

Cilj rada [4] jeste uvođenje TLDR generisanja, nove vrste ekstremne sumarizacije naučnih radova, čiji je ulaz naučni rad, a izlaz jedna rečenica koja ga sumarizuje, i ta rečenica se naziva TLDR. Generisanje TLDR-a je alternativa apstraktu, koji se uvek nalazi na početku naučnog rada i predstavlja sumarizaciju tog rada. Za pisanje TLDR-a je potrebno ekspertsko znanje i poznavanje jezika specifičnog za domen (*domain-specific language*, DSL). TLDR-ovi se fokusiraju na ključne delove rada, izbacujući suvišne detalje. Imajući u vidu činjenicu da se broj godišnje objavljenih naučnih radova duplira svakih devet godina [5], evidentna je korist koju TLDR-ovi donose.

U radu [6] uveden je radni okvir koji pretvara sve probleme jezika zasnovane na tekstu u tekst-na-tekst (*text-to-text*) format, to jest, tekst je ulaz, i tekst je izlaz. Autori su na raznim zadacima iz oblasti obrade pri-rodnih jezika (*Natural Language Processing*, NLP) testirali različite arhitekture, skupove podataka, tehnike pretreniranja, pristupe transfer učenju i još drugih aspekata obrade prirodnog jezika. Postignuti su *state-of-the-art* rezultati na mnogim zadacima, poput sumarizacije teksta, odgovaranja na pitanja i klasifikacije teksta. Takođe, ovaj rad je uveo C4 skup podataka (*Colossal Clean Crawled Corpus*). Pristup koji podrazumeva tretiranje svakog problema kao tekst-na-tekst problem omogućava direktno korišćenje jednog modela sa istom funkcijom gubitka I istim vrednostima hiperparametara na više različitih zadataka. Modelu, koji je sličan originalnom transformer modelu iz rada [7], kao prefiks je prosleđivan naziv zadatka koji treba da izvrši. Na primer, za prevođenje rečenice: ”That is good.“ sa engleskog na nemački jezik, ulaz je: ”translate English to German: That is good.“, a izlaz je: ”Das ist gut.“, dok je prefiks za sumarizaciju: ”summarize:“.

Cilja rada [8] jeste vršenje apstraktivne sumarizacije rečenica pomoću pristupa sumarizaciji zasnovanog na pažnji koji autori nazivaju *Attention-Based Summarization* (ABS). Autori su kao model upotrebili kombinaciju arhitekture modela jezika (*language model*, LM) sa kontekstualnim enkoderom ulaza.

Model jezika je prilagođen po uzoru na rad [9]. Enkoder ulaza je modelovan po uzoru na enkoder zasnovan na pažnji (*attention-based encoder*) iz rada [10]. Ključna razlika između modela koršćenog u ovom radu i klasičnih modela jezika je što su enkoder i model jezika trenirani istovremeno.

Više različitih vrsta enkodera je testirano: *bag-of-words* (BOW), konvolutivni enkoder, kao i enkoder zasnovan na pažnji. Za treniranje je korišćen anotirani *Gigaword* skup podataka [11].

2.6 OBJAŠNJENJA METRIKA ZA EVALUACIJU SUMARIZACIJE TEKSTA

Po radu [12], tehnike evaluacije sumarizacija dele se na:

1. ekstrinzične – ove tehnike evaluiraju koliko dobro se sistem koji vrši sumarizaciju ponaša na nekom zadatku za koji je potrebna sposobnost vršenja sumarizacije, poput odgovaranja na pitanja sa više ponuđenih odgovora ili pretrage dokumenata po nekim kriterijumima.
2. intrinzične – ove tehnike vrše poređenje automatskih sumarizacija sa ”zlatnim standardom“, to jest, sa ljudskim sumarizacijama. Medu intrinzične metrike ubrajaju se: *Perplexity*, ROUGE, BLEU, METEOR. Prednost intrinzičnih metrika nad ekstrinzičnim je što programerima daju eksplicitnu povratnu informaciju o kvalitetu sumarizacija. Ekstrinzične metrike tu vrstu informacija pružaju implicitno, kroz rezultate postignute pri obavljanju nekog zadatka. Sa druge strane, nedostatak intrinzičnih metrika je što zahtevaju skup referentnih sumarizacija, to jest, anotiran skup podataka.

ROUGE-N je intrinzična metrika koja evaluira preklapanje N-grama između mašinske i ljudske (referentne) sumarizacije. Tačnije, računa se koliki procenat N-grama iz ljudske sumarizacije se nalazi u mašinskoj sumarizaciji. ROUGE-1 posmatra unigrame, to jest, reči, dok ROUGE-2 posmatra bigrame, to jest, parove reči.
BLEU je intrinzična metrika koja koristi modifikovanu verziju preciznosti da izračuna sličnost između dve rečenice. Razlika između preciznosti i modifikovane preciznosti je što modifikovana preciznost ne razmatra samo da li se neka reč pojavljuje u referentnim rečenicama, već i koliko puta se u njima pojavljuje. U praksi, korišćenje unigrama nije optimalno, pa se koriste N-grami. Empirijski je utvrđeno da korišćenje tetragrama (N=4) daje rezultate najsličnije ljudskom razmišljanju, pa se iz tog razloga BLEU-4 često koristi u praksi. Za

primere eksploatacije problema BLEU metrike pogledati poglavlje 3 rada [13].

3. SPECIFIKACIJA I IMPLEMENTACIJA

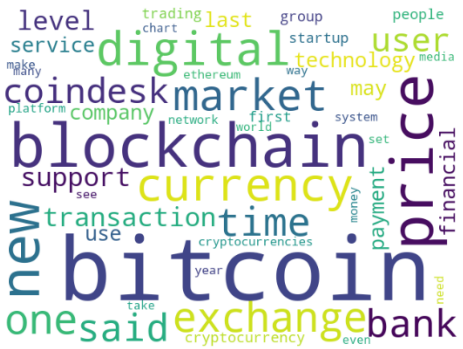
Cilj ovog rešenja je korišćenje T5 i BART modela za vršenje apstraktivne sumarizacije vesti o kriptovalutama. U potpoglavlju 3.1 su izloženi alati koji su korišćeni za implementaciju rešenja. T5 i BART modeli su dotrenirani nad skupom podataka koji je opisan u potpoglavlju 3.2, i njegove performanse su upoređene sa verzijom istog modela koja nije dotrenirana nad tim skupom podataka.

3.1 Korišćeni alati

Korišćeni su modeli iz okvira otvorenog koda *Hugging Face* u programskom jeziku *Python* verzije 3.8: $T5_{BASE}$, $BART_{BASE}$. Računar na kom su razvijani i trenirani modeli ima sledeće specifikacije: procesor *Intel Core i7-8750-H*, grafička kartica *Nvidia GeForce GTX 1050 Ti 3GB RAM*, 32GB radne memorije.

3.2 Skup podataka

Korišćen je skup vesti o kriptovalutama koji je dostupan na sajtu *Kaggle*[4]. Taj skup sadrži 40 hiljada vesti o kriptovalutama u periodu od 2013. do 2018. godine koje su preuzete sa popularnih sajtova, među kojima su CCN, CoinDesk i NewsBTC. Svaka vest sadrži sledećih sedam obeležja: URL, naslov, tekst tela vesti, HTML tela vesti, godina objavljivanja, ime autora, izvor. Autor skupa podataka je naveo da ovaj skup podataka može da se koristi kao merilo kvaliteta algoritama sumarizacije teksta. Sve do 2017. godine, Bitkoin je bio najpopularnija kriptovaluta. Stoga, ne čudi što je među tekstovima iz skupa podataka najučestalija reč ”Bitcoin“ sa preko 12000 pojavljivanja. Ona se pojavljuje u 80% tekstova i u 52% naslova iz skupa podataka. Na slici 1 je dat oblak reči sa 50 najfrekventnijih reči u tekstvoima iz skupa podataka.



Slika 1: *50 najfrekventnijih reči u tekstovima vesti iz skupa podataka*

U radu [4] naslov je korišćen kao pomoćni signal pri obučavanju, što ukazuje na to da se kvalitetan naslov makar u nekoj meri može smatrati adekvatnom sumarizacijom. Još jedan razlog zašto se naslov može smatrati adekvatnom sumarizacijom je taj što se 93% najčešćih reči iz tekstova tela vesti o kriptovalutama nalazi i u naslovima vesti. Ipak, jasno je da su naslovi naučnih radova većeg kvaliteta od naslova vesti o kriptovalutama, kako zbog toga što ih pišu stručnjaci iz oblasti, tako i zbog toga što nemaju potrebu za senzacionalizmom. Stoga, može se zaključiti da naslovi vesti iz skupa podataka ponekad jesu, a ponekad nisu adekvatna sumarizacija. U okviru potpoglavlja 4.1 je diskutovan primer gde naslov nije adekvatna sumarizacija vesti.

U proseku, tekst tela vesti sadrži 535,78 reči, dok je prosečna dužina naslova 8,85 reči.

4. EVALUACIJA REŠENJA I REZULTATI

Dotrenirana su dva modela - jedan za T5 arhitekturu, a drugi za BART arhitekturu. U pitanju su BASE verzije modela, koje su manje od LARGE verzija. Svaki od modela je treniran nad 10000 vesti, i evaluiran nad 2000 vesti pomoću tri metrike: ROUGE-1, ROUGE-2, BLEU. Takođe, verzije T5 i BART modela koje nisu dotrenirane nad vestima iz oblasti kriptovaluta su evaluirane nad istih 2000 vesti, i rezultati su upoređeni.

Ciljna sumarizacija, koja se sastoji iz jedne ili više rečenica, generisana je na osnovu teksta tela vesti. Kao labela je korišćen naslov. Dužina ulazne sekvence je pri dotreniranju ograničena na 50 tokena. Za *batch size* je korišćena vrednost 64. Treniranje je vršeno u tri epohe, na procesoru, zbog toga što nije bilo dostupno dovoljno radne memorije na grafičkoj kartici.

Pri pretprocesiranju teksta uklonjeni su karakteri šuma i veći broj razmaka je zamenjen jednim razmakom. Prebacivanje teksta u mala slova nije poboljšalo rezultate, kao ni lematizacija. U tabeli 1 dat je prikaz rezultata evaluacije modela za ulaznu sekvencu ograničenu na dužinu od 50 tokena.

metrika / model	ROUGE-1	ROUGE-2	BLEU-4
T5*	31,40	**12,99**	**0,099**
BART*	**31,71**	12,89	0,097
T5	23,41	7,16	0,029
BART	20,97	6,84	0,025

Tabela 1: Vrednost metrike po modelu. * označava dotrenirane verzije modela.

Iz rezultata je uočljivo da je dotreniranje modela doprinelo poboljšanju performansi. Takođe, rad [13] navodi da ne postoje statistički značajne razlike u performansama između T5 i BART modela, što je u skladu sa rezultatima iz tabele 1.

Empirijski su ispitane različite vrednosti za dužinu ulazne i dužinu izlazne sekvence. S obzirom na to da su dobijeni isti rezultati za ograničenje dužine ulazne sekvence na 50, 200 i 500 tokena, radi uštede resursa je izabrana najmanja od ispitanih vrednosti za ograničenje dužine sekvence.

Tri metode kojima bi ove performanse mogle da se poboljšaju su:

[4] Skup vesti o kriptovalutama, pristupljeno 15.10.2021. https://www.kaggle.com/kashnitsky/news-about-major-cryptocurrencies-20132018-40k/version/2

1. korišćenje LARGE verzija modela umesto BASE verzija
2. korišćenje kvalitetnijih referentnih sumarizacija, koje se mogu generisati pomoću neke SOTA arhitekture kao što je PEGASUS, koja je uvedena u radu [14].
3. Vršenje rekurizvne sumarizacije, po uzoru na rad [15]

4.1 Analiza grešaka modela

Performansama modela škodi činjenica da je naslov korišćen kao referentna sumarizacija. Na primer, za pojedine vesti model je proizveo smislenu sumarizaciju, ali naslov nije smislen, i zbog toga je ostvarena mala vrednost za metriku evaluacije. Jedan takav primer dat je na slici 2.

text: summarize:the emblem of the peoples' republic of china. editor's note: btc china has been dropped by tenpay, the 2nd largest third party payment processor in china after alipay, and is currently using yeepay (an obscure smaller third party payment processor) for deposits and withdrawals. btc china is also located in a special zone in shanghai and it is my opinion that btc china will continue legal operations in the best interest of their users; however, the waters in the meantime will be tumultuous. the timeline is this: by chinese new years, chinese bitcoin exchanges will have to have found new ways for users to deposit and withdraw. virtual commodity payment processors anyone? the bitcoin chart remains somewhat precarious following the latest round of news and rumours out of china. all markets hate uncertainty, so,

title: translation of sina.com article by cryptocoinsnews editor, caleb chen,

summary: chinese Bitcoin Exchanges Will Have to Find New Ways to Deposit and Refund,

rouge1: 0.0

Slika 2: *Greška T5 modela usled neadekvatnog naslova*

Prisutni su i primeri gde je model napravio grešku u sumarizaciji zbog toga što nije video dovoljan deo teksta vesti.

Uzrok najvećeg broja loše ocenjenih sumarizacija kada je u pitanju BLEU metrika jeste činjenica da su sumarizacije mnogo duže od naslova, te ne daju visoku vrednost metrike, iako su veoma smislene.

5. ZAKLJUČAK

U ovom radu su evaluirani aktuelni transformer modeli na zadatku automatske sumarizacije vesti o kriptovalutama. Korišćen je javno dostupan skup podataka sa sajta *Kaggle*. Ulaz modela je sekvenca tokena koja predstavlja tekst vesti. Izlaz modela je sekvenca tokena koja predstavlja sumarizaciju vesti sa ulaza. Kao labela je korišćen naslov vesti. Za evaluaciju su korišćene ROUGE i BLEU metrike. Rezultati pokazuju da je dotreniranje modela doprinelo poboljšanju njihovih performansi. Izložene su i analizirane uočene greške u performansama. Predloženi su koraci za poboljšanje performansi modela.

6. LITERATURA

[1] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

[2] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[3] Qi, Weizhen, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. "Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training." arXiv preprint arXiv:2001.04063 (2020).

[4] Cachola, Isabel, Kyle Lo, Arman Cohan, and Daniel S. Weld. "TLDR: Extreme summarization of scientific documents." arXiv preprint arXiv:2004.15011 (2020).

[5] Van Noorden, Richard. "Global scientific output doubles every nine years." Nature news blog (2014).

[6] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683 (2019).

[7] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008. 2017.

[8] Rush, Alexander M., Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." arXiv preprint arXiv:1509.00685 (2015).

[9] Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. "A neural probabilistic language model." The journal of machine learning research 3 (2003): 1137-1155.

[10] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[11] Fan, James, Raphael Hoffmann, Aditya Kalyanpur, Sebastian Riedel, Fabian Suchanek, and Partha Talukdar. "Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)." In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX). 2012.

[12] Mani, Inderjeet. "Summarization evaluation: An overview." (2001).

[13] Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. "Re-evaluating the role of BLEU in machine translation research." In 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006.

[14] Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." In International Conference on Machine Learning, pp. 11328-11339. PMLR, 2020.

[15] Wu, Jeff, Long Ouyang, Daniel M. Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. "Recursively Summarizing Books with Human Feedback." arXiv preprint arXiv:2109.10862 (2021).

Kratka biografija:

Uroš Ogrizović rođen je 1997. godine. Osnovne akademske studije završio je 2020. godine na Fakultetu tehničkih nauka, na kom brani i master rad 2021. godine iz oblasti Elektrotehnike i računarstva – Softversko inženjerstvo i informacione tehnologije.
Kontakt: uros.ogrizovic@uns.ac.rs