УНИВЕРЗИТЕТ У БЕОГРАДУ

ЕКОНОМСКИ ФАКУЛТЕТ

Александар С. Дамјановић

Рударење текста – Утицај јавно доступних текстова на приносе крипто валута

Докторска дисертација

Београд, 2024.

UNIVERSITY OF BELGRADE

FACULTY OF ECONOMICS

Aleksandar S. Damjanović

Text mining – Impact of publicly available texts on crypto currency returns

Doctoral Dissertation

Belgrade, 2024.

ПОДАЦИ О МЕНТОРУ И ЧЛАНОВИМА КОМИСИЈЕ ЗА ОДБРАНУ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Ментор: др Ирена Јанковић, ванредни професор, Универзитет у Београду – Економски факултет

Чланови комисије за одбрану докторске дисертације:

Датум одбране:_____

Milieni

Рударење текста – Утицај јавно доступних текстова на приносе крипто валута

С а ж е т а к

Дисертација демонстрира употребну моћ алтернативних извора података у анализи приноса на крипто-валуте. Применивши алате за рударења текста на вести о крипто-валутама преузетих са онлајн портала конструисана су три предиктора приноса. Скуп разматраних предиктора чинили су сентимент вести о самој крипто-валути, њихова читљивост и сентимент вести о тржишном лидеру, Биткоину. За процењивање сентимента вести коришћен је иновирани приступ оцењивања пондера сентимента који заступа ова дисертација. Истраживање је показало да су у првом кварталу 2022. године изабрани предиктори добро објашњавали кретање приноса за осам одабраних крипто-валута. Сентимент Биткоина се показао као најзначајнији међу предикторима, док се читљивост вести показала као најслабији предиктор. Идентификоване везе искоришћене су за изградњу модела машинског учења из породице ансамбла за сваку крипто-валуту посебно. Изграђени ансамбли искоришћени су за испитивање предвидљивости будућег кретања приноса и квалитета иновираних оцена сентимента. И тачкасте оцене и статистички тестови потврдили су да ансамбли имају већу предиктивну моћ када се користи иновирана методологија мерења сентимента. Поред стандардних статистичких тестова, за поређење квалитета прогнозе коришћен је и Омнибус тест који је развио аутор. Будући да је постојао известан степен предвидљивости приноса истраживањем су документовани сигнали потенцијалне нарушености слабе форме хипотезе о тржишној ефикасности код анализираних крипто-валута. Спроведени тестови показали су да цене крипто-валута *AVAX, BTC, DOGE* и *ETH*, нису случајан ход, као и да се приноси крипто-валута *ADA, DOT* и *LUNA* могу описати *ARMA* процесом. Слаба форма тржишне ефикасности потврђена је само код крито-валуте *SOL*.

Кључне речи: сентимент, рударење текста, приноси, крипто-валуте, машинско учење, алгоритам, предвиђање, скреповање, финансије, хипотеза о ефикасности тржишта

Научна област: економија

Ужа научна област: статистика

JEL: C650, C880, G400

Text mining – Impact of publicly available texts on crypto currency returns

A b s t r a c t

The thesis demonstrates the utility of alternative data sources in analyzing crypto-currency returns. By applying text mining tools on crypto-currency news downloaded from online portals, three predictors of returns were constructed. The set of considered predictors consisted of crypto-currency news sentiment, news' readability and news sentiment about the market leader, Bitcoin. To measure news sentiment, an innovative approach of estimating sentiment weights, proposed by this thesis, was used. Bitcoin sentiment emerged as the most significant predictor, while news readability has turned out to be the weakest predictor. The identified relations were used to construct machine learning model from the ensemble family for each crypto-currency separately. Constructed ensemble models were used to examine the predictability of future returns movements and the quality of the innovative sentiment estimates. Both point estimates and statistical tests confirmed that ensembles have greater predictive power when using the innovative sentiment estimation methodology. In addition to the standard statistical tests, the Omnibus test developed by the author was used to compare the quality of the forecasts. Since there was a certain degree of returns' predictability, the research documented signals of a potential violations of the weak form of efficient market hypothesis in the case of analyzed crypto-currencies. The conducted tests showed that the prices of crypto-currencies AVAX, BTC, DOGE and ETH doesn't follow random walk, and that the returns of crypto-currencies ADA, DOT and LUNA can be described by the ARMA processes. A weak form of market efficiency was confirmed only in case of the crypto-currency SOL.

Key words: sentiment, text mining, returns, crypto-currencies, machine learning, algorithm, prediction, scraping, finance, efficient market hypothesis

Scientific field: economics

Scientific subfield: statistics

JEL: C650, C880, G400

С А Д Р Ж А Ј

С П И С А К С Л И К А

С П И С А К Т А Б Е Л А

1. Увод

Како традиција налаже, прво поглавље сублимира суштину научног истраживања и пружа читаоцу увид у контекст неопходан за његово разумевање. Из тог разлога читаоцу ће, најпре, бити предочени предмет истраживања, истраживачке идеје и мотивације. Затим ће се кроз преглед литературе читаоцу представити дубина до које сежу постојећа економска сазнања о теми истраживања. Ослањајући се на претходно, читалац ће моћи да сагледа циљеве истраживања, као и његове научне доприносе. Коначно, на самом крају читалац ће имати прилику да сагледа изворе података неопходних за спровођење истраживања, али и структуру даљег излагања.

1.1 Предмет истраживања

У контексту финансија, сентимент инвеститора представља његов став или уверење о будућим кретањима на финансијском тржишту, будућим новчаним токовима и ризицима улагања, који не мора да буде оправдан чињеницама (Бејкер (*Baker*) и Вурглер (*Wurgler*) 2007). Вероватно једну од најинтересантнијих дефиниција сентимента у финансијском контексту дали су Каплански (*Kaplanski*) и Леви (*Levy*) (2010). Према поменутим ауторима, сентимент је свака субјективна перцепција инвеститора која може довести до погрешног одређивања фундаменталне вредности неке активе. Из приложеног става се јасно види да сентимент као такав комбинован са ефектом крда може бити покретач растућег или опадајућег тржишта (енгл. *bull and bear market*), што га издваја као веома важну променљиву у свету финансија.

Сентимент инвеститора се не може непосредно мерити. Из тог разлога истраживачи и привредни актери користе одређене мерљиве показатеље блиско повезане са сентиментом инвеститора како би га проценили. Анкетирање инвеститора, индекси тржишне волатилности, композитни индикатори и други показатељи засновани на историјском кретању економских променљивих само су неки од показатеља коришћених за процену сентимента. Иако се дуго и традиционално користе, пракса је показала да је употребна моћ издвојених показатеља ограничена (Крајевелд (*Kraaijeveld*) и Де Смет (*De Smedt*) 2020, Ша (*Xia*) 2022 и др.). То је био повод да се крајем прве деценије двадесет и првог века пажња јавности усмери на нове приступе у мерењу сентимента. Становиште од којег се том приликом пошло је да на сентимент инвеститора битно утичу вести којима су инвеститори изложени (Питерсон (*Peterson*) 2016). У том случају довољно би било пратити и мерити само сентимент новопридошлих вести, јер ће се он на тржишту преточити у сентимент инвеститора. Овај приступ је постао веома популаран и често је укључен у опсежније техничке анализе цена финансијских инструмената.

Како се вести претежно саопштавају писаним путем, фокус аналитичара премешта се са нумеричких на текстуалне податке. За такав радикалан заокрет било је потребно да се стекну одређени услови. У почетку је главни проблем представљало прикупљање свих доступних

вести у току дана (географске потешкоће, недостатак кључне инфраструктуре, асиметричне информације и др.). Иако је истакнути проблем превазиђен са појавом рачунара и распрострањивањем процеса умрежавања, на његово место су дошли други проблеми који су последица масовности доступних вести. Наиме, у савременом дигитализованом свету дневно се широм планете објави на милијарде текстова (објава на друштвеним мрежама, новински чланци, блогерски текстови и други текстуални садржаји). Као последица тога селекција релевантних извора текстуалних вести, прикупљање, читање, анализирање и обрада велике количине текстова зарад добијања релевантних информација главни су изазови са којима се истраживач суочава. Тако велики посао, чак иако се он делегира на велики број људи, није могуће обавити довољно брзо да би добијени резултати могли практично да се примене. Последично, било је неопходно развити алгоритме преко којих ће машине обавити поменуте задатке уместо људи. Тек од тог тренутка текстуални записи постају употребљив извор података за истраживаче. Ово може звучати невероватно, с обзиром на то да су текстуални записи у животу људи присутни преко 5000 година. Упркос њиховој старости, текстуалне записе сврставамо у породицу тзв. алтернативних извора података. Алтернативни извори података (енгл. *alternative data sources*) представљају све изворе података који се могу користити за спровођење истраживања који су нам постали доступни са развојем савремене технологије. Поред текстуалних записа у алтернативне изворе података спадају и *GPS* сигнал, интернет претраге, активности на мрежи или интернет сајту, па чак и апарати који региструју одређено понашање (на пример: паркинг апарати и рампе).

Дакле, сентимент вести је потребно проценити на основу текстова. То изискује дефинисање појма „сентимента текста“. Сентимент текста је квантитативни одраз позитивности, односно негативности, изнетих чињеница, али и става, тона, призвука и нивоа оптимизма, то јест песимизма, са којим је текст написан. Једно од најцитиранијих истраживања које је у великој мери заслужно за популарност анализе сентимента текстова у финансијама, спровео је Тетлок (*Tetlock*) (2007). Поменути аутор је заступао субјективан приступ мерењу сентимента текстова заснованом на изради лексикона. Како би избегли ефекат субјективности, каснији истраживачи су се преоријентисали на машинско учење у мерењу сентимента. Међу њима се посебно истиче модел Џагадиша (*Jegadeesh*) и Вуа (*Wu*) (2019) који дозвољава да сентимент речи узима произвољне вредности са реалне осе. Њихов рад је представљао огроман искорак у анализи сентимента, будући да су друга истраживања, углавном, дозвољавала да сентимент речи може узимати вредности из неког коначног скупа. Ипак, метод оцењивања који су развили Џагадиш и Ву (2019) није беспрекоран. Оцењени пондери сентимента речи у изворном облику садржаће грешку мерења. Ова дисертација понудиће модификован приступ (у даљем тексту „алтернативни приступ“) њиховом оцењивању који ће обезбедити прецизније предвиђање економских променљивих. Алтернативни приступ инспирисан је Јохансеновом (*Johansen*) (1996) процедуром коришћеном за оцену коинтеграционих параметара у класичној анализи временских серија. Као резултат добијене су нове оцене из којих је одстрањен део грешке мерења. Остатак грешке мерења може се апроксимирати или потпуно отклонити нормирањем уз увођење одређених претпоставки. Резултати добијени у овом истраживању сугеришу да сентимент измерен на бази алтернативних оцена помаже у прецизнијем предвиђању кретања економских величина. Другим речима, предложене алтернативне оцене прецизније процењују сентимент инвеститора.

Крајњи циљ конструкције прецизније мере сентимента је подизање његове употребне моћи у финансијским истраживањима. То се огледа кроз два аспекта. Први од њих је подизање прецизности анализе утицаја. Конкретно, прецизнијим мерењем сентимента омогућиће да прецизније сагледамо утицај који сентимент има на кретање економских величина. Други аспект је прецизније предвиђање. Прецизнијом проценом сентимента текстуалних вести прецизније ће се проценити сентимент инвеститора. Самим тим ће се прецизније предвидети притисак на тржишне силе и цену (тј. приносе) и друге економске величине. За поређење квалитета прогнозе коришћени су МекКракенов (*McCracken*) (2000) тест и Омнибус тест који предлаже аутор заснован на модификацији теста Диеболд-Маријана (*Diebold-Mariano*) (1991).

Ослањајући се на предложено побољшање у мерењу сентимента, дисертација ће дати свој допринос анализи тржишта крипто-валута. Спроведено истраживање провериће какав је утицај информација добијених из онлајн вести на приносе код осам високо популарних[1] крипто-валута у време започињања истраживања. У те сврхе из преузетих вести изрударена су три показатеља за које се верује да могу утицати на приносе крипто-валута. Први од три показатеља је сентимент новопридошлих вести о самој крипто-валути. Када говоримо о сентименту не смемо изгубити из вида да крипто-валуте немају интринзичну вредност, колатерал ни државну гаранцију. У таквим околностима очекивано је да утицај сентимента на њихову вредност (а самим тим и приносе) постоји, као и то да је овај утицај израженији него код осталих класа финансијске активе. Избор наредног показатеља инспирисале су две важне идеје. Из микроекономске теорије је познато да се на тражњу за неким добром могу одразити ставови појединаца о другим сродним добрима. Из тог разлога дисертација жели да испита утицај тзв. унакрсног (енгл. *cross-section*) сентимента. Под унакрсним сентиментом подразумеваћемо утицај на приносе неке крипто-валуте који има сентимент друге крипто-валуте. Друга крипто-валута ће у свим анализираним случајевима бити *BTC*-а, осим у сопственом случају када ће га одменити *ETH*. *BTC* је природни избор за представника крипто-валута будући да је реч о најстаријој, најпопуларнијој и највећој (по тржишној капитализацији) крипто-валути. Осим тога, многа истраживања указују на то да крипто-валуте прате кретање *BTC*-а (Кумар (*Kumar*) и Аџаз (*Ajaz*) 2019, Адедоку (*Adedoku*) 2019, Озајдн (*Ozaydin*) 2021 и др.), који се успоставио као тржишни лидер. У таквим околностима очекивано је да његов сентимент има утицај на тражњу, а самим тим и приносе, других крипто-валута. Једно од пионирских истраживања унакрсног-сентимента документовано је радом Дамјановића и Дреновака (2023). Поменути рад је објављен као предистраживање ове дисертације. Последњи разматрани показатељ је читљивост преузетих вести. Читљивост може бити важан предиктор приноса будући да делимично или потпуно неразумевање текста може имати непредвидив утицај на понашање инвеститора. Истраживање је показало да су приноси на крипто-валуте могли у доброј мери да се објасне и предвиде информацијама добијеним рударењем текстова публикованих вести.

Истраживања предвидљивости приноса је важно за тржишта и/или тржишне сегменте у чију се ефикасност сумња, будући да оно може бити сигнал потенцијалне тржишне неефикасности. Уколико је тржиште ефикасно, вредност финансијске активе мора да одражава све доступне информације. Последично, нико не може да искористи доступне информације да предвиди кретање приноса. На неефикасном тржишту могуће је предвидети кретање цена и

[1] Према броју објављених чланака на порталу Крипто Њуз (више о овој теми у одељку **1.5**).

профитирати од тржишних грешака. Због тога инвеститори непрестано трагају за неефикасностима које могу искористити за стицање профита. Тренутна сазнања подстичу сумње да је тржиште крипто-валута неефикасно. Конкретно, реч је о финансијском тржишту које је веома нерегулисано, још увек је младо (цене се на овом тржишту прате само око једну деценију), није стационирано, и даље се развија, и чија актива нема фундаменталну вредност, те је подложна стварању ценовних балона (енгл. *bubble behavior*). И ово истраживање наишло је на сигнале потенцијалне неефикасности тржишта крипто-валута. То је мотивисало аутора да спроведе формалне тестове слабе форме тржишне ефикасности. Тестови су показали да су анализиране крипто-валуте већином биле неефикасне.

Оно што даје посебну релевантност добијеним резултатима јесте период у којем се предвидљивост приноса испитује. Узорак за тестирање обухвата први квартал 2022. године, односно почетак глобалне кризе која се одразила и на тржиште крипто-валута. Наиме, рат на истоку Европе покренуо је негативну фазу привредних циклуса на глобалном нивоу и жестоко је уздрмао тржиште крипто-валута. Ово је први пут у историји да имамо глобалну кризу и кризу на тржишту крипто-валута у исто време. Као што је документовано у Саркодие (*Sarkodie*) и сарадници (2022), пандемија КОВИД19 није нашкодила тржишту крипто-валута за разлику од остатка светске привреде. Напротив, тржиште крипто-валута је у то време доживело један од својих највећих процвата. Стога се са правом може рећи да је анализа утицаја информација добијених из вести на приносе крипто-валута у условима свеопште кризе још увек недовољно истражено поље које ова дисертација планира да испита.

1.2 Преглед литературе

Пре даљег излагања, осврнимо се на преглед достигнућа у постојећој литератури. Због разноликости тема покривених у овој дисертацији, али и због њених методолошких и емпиријских доприноса, преглед литературе изложен у наставку одељка биће подељен у три подсекције. До краја овог одељка читалац ће бити у прилици да се упозна са сазнањима до којих сеже постојећа литература у домену мерења сентимента, рударењу текста, крипто-валута, тржишној ефикасности и слично.

1.2.1 Различити приступи у мерењу сентимента

Велики број радова анализирао је постојеће мере сентимента постављајући питање њихове класификације (видети Браун (*Brown*) и Клиф (*Cliff*) 2004., Бандопађаја (*Bandopadhyaya*) и Џонс (*Jones*) 2006., Брхарт (*Burghardt*) 2011., Амдуни (*Amdouni*) 2021., Прасад (*Prasad*) и сарадници 2023., и други). Иако су идентификованим групама у литератури давани различити називи, генерално се може рећи да савремена литература препознаје четири категорије мера сентимента. Прву групу мера чине оне засноване на анкетирању инвеститора. Ове мере се сматрају најнепосреднијим, јер испитујући директно инвеститоре покушавају да разоткрију

њихове ставове и очекивања, односно, њихов сентимент. Међу овим показатељима посебно се издвајају они добијени из анкета за мерење сентимента које спроводе следеће организације: Америчко удружење индивидуалних инвеститора (енгл. *American Association of Individual Investors*), Инвеститорске информације (енгл. *Investors' Intelligence*) и *UBS*. Ипак индикатори сентимента базирани на анкетирању са собом носе два кључна недостатка. Први недостатак су класични анкетни проблеми (попут искрености испитаника, квалитета постављених питања, намерног изостанка одговора на осетљива питања, пристрасности и слично). Други недостатак односи се на фреквентност доступности ових индикатора (анкете се не могу обављати свакодневно).

Другу групу индикатора чине они које обрачунавамо из економских и финансијских података (нпр. цена, обима трговања, бројности одређених инструмената, и слично). Највећи број поменутих показатеља израђен је за потребе техничке анализе. *TRIN* статистика, индекс непотпуних лотова, пут-кол рацио (енгл. *put-call ratio*), интересовање за продају на кратко и друго, само су неки од показатеља сентимента које је понудила техничка анализа (видети Боди (*Bodie*) и сарадници 2007). Највећа слабост ових показатеља је њихова тзв. аутодеструктивност. Према Бодију и сарадницима (2007) свако техничко правило које постане јавно доступно тежи да се поништи. Добијене сигнале покушаће да експлоатише велики број трговаца и чартиста чиме се врши притисак на тренутну корекцију цене, те сигнал остаје неупотребљив. Како се истакнути показатељи већ дуго и традиционално користе и добро су познати у финансијској јавности, предиктивна моћ сентимента процењеног на бази њих обично није велика.

Међу ове показатеље литература сврстава и индексе тржишне волатилности, будући да се и они обрачунавају на бази економских података. Добар пример је *VIX* индекс (али и његови еквиваленти[2]) који представља будући ниво тржишне волатилности антиципиран од стране тржишних актера (Деметерфи (*Demeterfi*) и сарадници 1999). Значајан део његове популарности дугује се чињеници да је лако доступан финансијским актерима. Ипак, овај индекс као индикатор сентимента инвеститора доводи се у везу са бројним недостацима. Хестла-Барнхарт (*Hestla-Barnhart*) (2015) истиче да је *VIX* мера сентимента тржишта као целине. Као такав неће бити прецизан индикатор сентимента инвеститора у појединачну финансијску активу или поједине тржишне сегменте, чији сентимент не мора бити усаглашен са општим сентиментом тржишта. Примера ради, док је током пандемије КОВИД19 читава америчка привреда била у паници, тржиште крипто-валута је просперирало (тржишни сегмент), док је компанија Зум (енгл. *Zoom*) остварила енормни раст (појединачна актива). Грифин (*Griffin*) и Шамс (*Shams*) (2018) критиковали су поузданост *VIX*-а. Показавши да је *VIX* подложан манипулативним утицајима, аутори скрећу пажњу да његово кретање не одражава увек прави ниво сентимента. Ипак, можда најинтересантнију критику налазимо у раду Бекарта (*Bekaert*) и сарадника (2013). *VIX* није чиста мера сентимента, већ, као што је истакнуто, мера антиципиране волатилности. Кретање овог показатеља може се декомпоновати на две компоненте. Прва компонента одражава тржишну неизвесност и представља стварну очекивану тржишну волатилност. Друга компонента је резидуална те скупља све нелинеарне ефекте настале услед страха, аверзије према ризику и тржишног сентимента. Постојање ове компоненте за последицу има да *VIX* често прецењује стварни ниво будуће волатилности

[2] Примера ради *CVI* индекс који прати волатилност тржишта крипто-валута.

(видети Кришнамрфи (*Krishnamurthy*) 2021. и Едвардс (*Edwards*) и Престон (*Preston*) 2017.). Како се само један део кретања *VIX*-а дугује сентименту, *VIX* не може бити идеална мера сентимента. Коначно, предиктивна моћ *VIX*-а је ограничена. Пре свега, како је у питању један од најпопуларнијих индикатора тржишног сентимента, изложен је поменутој аутодеструкцији. Чен (*Chen*) и сарадници (2012) скренули су пажњу да је *VIX* лош у „дан за дан" предикцијама као и на високо-фреквентним подацима, док су Бандопађаја и Џонс (2008) показали да је пут-кол рацио прецизнија мера сентимента од *VIX*-а.

У наредну категорију мера сентимента спадају тзв. композитни индикатори. Реч је о показатељима који се добијају комбиновањем неколико појединачних показатеља сентимента. Главни циљеви израде композитних индикатора су ублажавање недостатака употребе појединачних показатеља и узимање у обзир више доступних информација приликом процене сентимента. Међу најпопуларнијима су: индекс страха и похлепе *CNN*-а (енгл. *CNN's fear and greed index*), композитни индекс Бејкера и Вуглера (2006) и композитни индекс Брауна и Клифа (2004). Први од поменутих се ослања на седам појединачних индикатора, други користи чак осам појединачних индикатора, док се трећи ослања на шест (два анкетна и четири заснована на тржишним подацима) појединачна индикатора. Истакнути показатељи су се показали прецизнијим од појединачних мера сентимента. Ипак, и они имају своје недостатке које су идентификовали сами њихови аутори. Постојећи индикатори обухватају само неколико појединачних мера сентимента, те се један део сигнала игнорише. Обрачун композитних индикатора прате одређене потешкоће (неки индикатори касне за другима, индикатори могу имати различите фреквенције и јединице мере и сл.) што повећава могућност грешке. Такође, композитни индикатори некада могу наследити мане показатеља од којих су изграђени и могу бити непрактични за интерпретацију.

Коначно, четврту групу чине показатељи базирани на алтернативним изворима података. Њихова идеја је да анализирајући садржај којем је изложена одређена група инвеститора можемо проценити какав ће сентимент инвеститори формирати. У те сврхе најчешће се користи анализа различитих текстуалних садржаја са интернета, међу којима су и онлајн вести (ову тему детаљно разматрамо у посебном пододељку). Поред текстова користе се и различите врсте интернет претраге (видети Кристуфека (*Kristoufek*) 2013), слике које се појављују у медијима доступним инвеститорима (видети Обаид (*Obaid*) и Пуктуантонг (*Pukthuanthong*) 2022), али и други алтернативни извори. Емпиријски резултати предочени у постојећој литератури сугеришу да су алтернативни извори података тренутно најуспешнији у процени сентимента инвеститора. Хајних (*Heinig*) и Нанда (*Nanda*) (2018) су анализирали употребу композитних мера сентимента и мера сентимента базираних на алтернативним изворима података на тржишту некретнина. Оба индикатора су у значајној мери могла да објасне кретање приноса на некретнине, али су се индикатори базирани на алтернативним изворима података показали значајно бољим. Слично томе, Мао (*Mao*) и сарадници (2011) нису могли да објасне кретање и неколико важних економских променљивих (Дау Џонс индекса, цена злата, укупног тржишног обима трговања и *VIX*-а) помоћу традиционалних мера сентимента. Са друге стране, индикатори базирани на алтернативним изворима података успели су да статистички значајно објасне кретање издвојених променљивих. Анализа је показала и да су традиционалне мере сентимента углавном касниле са давањем сигнала, док то није био случај са индикаторима сентимента базираним на алтернативним изворима података. Ова дисертација покушаће да да свој допринос у развоју управо ове категорије мера сентимента.

1.2.2 Приступи у мерењу сентимента

Велика заслуга за ширење популарности примене текстуалних записа као алтернативних извора података у сфери финансија дугује се Тетлоку. Тетлок (2007) је анализирао сентимент колумне „У току са тржиштем“ (енгл. *Abreast of the Market*) која се дневно публикује у Волстрит Журналу. Сентимент је мерен на бази априори дефинисаног лексикона. Лексикон разврстава често коришћене речи у енглеском језику на позитивне, негативне и неутралне према мишљењима стручњака из области лингвистике и психологије[3]. Израчунати сентимент Тетлок (2007) је упарио са укупним обимом трговања на тржишту (са *NYSE*) и приносима Дау Џонс индекса (енгл. *Dow Jones*) кроз *VAR*(5) модел. Добијени резултати сугерисали су да само негативне вести успевају да објасне кретање тржишта и обима трговања, као и да предвиде њихово будуће кретање. Прва замерка упућена на рад Тетлока (2007) дошла је од Лохрана (*Loughran*) и Мекдоналда (*McDonald*) (2011). Они сматрају да је коришћење општег лексикона неадекватно, јер речи могу имати различити сентимент у различитим областима. У складу са тим аутори истичу да је потребно израдити посебан лексикон сентимента за сваку научну област који ће уважити њене језичке специфичности. За своје истраживање аутори су израдили лексикон сентимента специјализован за сферу финансија. Истраживање које су спровели потврдило је њихове сумње будући да су текстови из 10-К извештаја имали већу предикативну моћ када се њихов сентимент мерио финансијским лексиконом двојице аутора уместо до тада коришћеног општег лексикона. Слично томе, Хенри (*Henry*) (2008) је у свом истраживању базираном на саопштењима за јавност водећих Америчких корпорација истакао да је слабост општих лексикона и то што се не баве питањем синонимије и полисемије. Имајући у виду да синоними и хомоними варирају од контекста до контекста, Хенри такође препоручује израду специјализованих лексикона. Ипак, главни недостатак анализе сентимента на бази лексикона је субјективност истраживача приликом израде лексикона (речи немају исту тежину за све људе). Упркос томе, лексикони су и даље присутни у многим истраживањима. Примера ради, Банеа (*Banea*) и сарадници (2018) представили су приступ који смањује субјективност приликом израде лексикона заснован на бутстраповању. Са друге стране, неки истраживачи користе посебне врсте лексикона да потпомогну своја истраживања. Добар пример је рад Шапкоте (*Sapkota*) (2022) у којем је аутор израдио лексикон емоција.

Објективност је разлог због којег је један број истраживача прибегао оцени сентимента речи машинским учењем[4]. Ова миграција представља другу етапу у развоју анализе сентимента. Посматрајући речи присутне у тексту и понашање одређене економске варијабле на дан објаве текста (нпр. у финансијама понашање приноса, волатилности, обима трговања и др.), рачунар је учио о томе које речи имају позитиван, односно негативан или неутралан утицај. За класификацију речи користе се најразноврснији методи међу којима су: логистичке регресије, наивни Бајес, технике кластерисања, случајне шуме, неуронске мреже и др. Занимљиво је да је један од пионирских радова о рударењу текста у финансијама управо био базиран на овом приступу. Антвеилер (*Antweiler*) и Френк (*Frank*) (2004) су користећи Наивни Бајес и метод подражавајућих вектора (енгл. *Support Vector Machine*, у даљем тексту *SVM*) утврдили сентимент публикација на порталима: *Yahoo!Finance, Raging Bull* и Волстрит Журнал. На бази

[3] Лексикон је израђен на бази Харвардовог IV психолошког речника.

[4] Будући да се применом исте технике машинског учења на исте податке добијају исте процене сентимента речи, овај приступ сматра се објективнијим мерењем сентимента у поређењу са лексиконским приступом.

добијених оцена сентимента аутори су затим направили успешне предикције приноса, обима трговања и волатилности. Још један интересантан пример оваквог истраживања је рад Шумахера (*Schumaker*) и сарадника (2012). Аутори су користили *SVM* за класификацију сентимента, док су за предвиђање кретања приноса и других повезаних варијабли користили систем који су сами израдили (назван *Arizona Financial Text System*). Ипак, њихов рад се издваја по добијању негативног предзнака везе између сентимента и приноса, што је интерпретирано као специфичност (аномалија) периода истраживања. Овај приступ заступали су и Касем (*Qasem*) и сарадници (2015), Дај (*Day*) и Ли (*Lee*) (2016), Рено (*Renault*) (2020), Шашмаз (*Şaşmaz*) и Тек (*Tek*) (2021), али и многи други.

Оба приступа (израда лексикона и машинско учење) подразумевају класификацију речи у три групе према њиховом сентименту. Међутим, оваква класификација речи имплицира да све речи из исте групе са собом носе исти ниво (тј. магнитуду) сентимента. Другим речима, подразумевало би се да су све позитивне речи подједнако позитивне, као и да су све негативне речи подједнако негативне. Ипак, вероватније је да то није случај. Примера ради, читаоцу се оставља да сам (субјективно) упореди тежину следећих речи: „волети“ и „свиђати се“, „банкрот“ и „дуговати“, „убити“ и „ранити“ и сл. Евидентно је да би класификација речи према сентименту у 3 групе била исувише груба. С тим у вези, литература се окреће методама оцене сентимента текста који препознају разлику у тежини међу речима из исте групе. Неки аутори су проблем решили повећањем броја категорија које сентимент може узети. Примера ради, Болен (*Bollen*) и сарадници (2011) су речи према сентименту груписали у 6 категорија, а Бонато (*Bonato*) и сарадници (2020) у 9. Други аутори који су уочили овај проблем покушали су промене приступ у моделирању сентимента. Тако, на пример, аутори попут Лохрана и Мекдоналда (2011) или Јанга (*Yang*) и сарадника (2015) покушали су да развију своје формуле за пондерисање сентимента базиране на фреквенцијама речи и њиховим трансформацијама. Једно занимљиво решење засновано на факторској анализи понудио је и пионирски рад о примени рударења текста у финансијама који су публиковали Фрејжер (*Frazier*) и сарадници (1984). Ипак, међу радовима из ове групе посебно се истиче рад Џагадиша и Вуа (2019) који су оценили магнитуду сентимента уз помоћ класичног линеарног регресионог модела. Њихов приступ биће разматран детаљније у секцији **4.1**.

1.2.3 Примена сентимента и других показатеља добијених из текстова

Сентимент није једини производ рударења текста. Постоје бројни други технички индикатори који се могу добити анализом текста, а који могу послужити у финансијским истраживањима. Бројни су радови који документују употребну моћ ових индикатора. Једно од најзапаженијих истраживања овог типа спровели су Коен (*Cohen*) и сарадници (2020). Посматрајући корпоративне извештаје, аутори су показали да што су они сличнији из године у годину то ће вредност корпорације више расти. Према интерпретацији аутора, добијени резултат сугерише да сличност садржаја корпоративних извештаја из године у годину указује на стабилност компаније, а та стабилност се на тржишту претаче у раст њене вредности. Додатно, аутори су показали да је могуће формирати стратегију трговања базирану на сличности корпоративних извештаја која би на дуги рок обезбедила принос већи од тржишног. Слично томе, Ли (*Li*) (2006) је указао да добра читљивост годишњих корпоративних извештаја може бити

индикатор позитивних приноса у наступајућем периоду. Хенри (2008) је анализирајући саопштења за јавност показао да њихова дужина и нумерички интензитет (тј. заступљеност квантитативних вредности у тексту) имају скромну предикативну моћ, док њихова читљивост нема никакву предикативну моћ. Постоје и технички показатељи повезани са текстуалном анализом који нису индикатор везан за појединачне текстове. Добар пример је медијска покривеност, односно, број објављених текстова у неком периоду. Међу бројним радовима који се баве овим индикатором издваја се истраживање Енгелберга (*Engelberg*) и Парсонс (*Parsons*) (2011) који су кроз панел модел са фиксним ефектима показали да медијска покривеност битно објашњава обим трговања.

Ипак, неспорно је да међу техничким показатељима добијеним рударењем текста анализа сентимента има примат у постојећој литератури. Сентимент инвеститора анализиран је преко сентимента текстова за разне врсте финансијских актива. Највећи број радова испитује утицај сентимента код обичних акција и берзанских индекса, што се јасно види из до сада представљене литературе. У њиховом фокусу није само америчко тржиште. Примери оваквих радова су: Хо (*Ho*) и сарадници (2020) и Одрино (*Audrino*) и Тетерева (*Tetereva*) (2019). Први су пратили утицај сентимент текстова доступних на порталу *RavenPack Dow Jones News Analytics* на приносе и волатилност водећих банака са берзи у Народној Републици Кини. Други ауторски пар је анализирао утицај вести са портала *Thomson Reuters* на приносе водећих корпорација у ЕУ. Оба рада су користила Греинџеров тест узрочности да покажу да сентимент вести заиста узрокује кретање анализираних економских варијабли. Када је реч о утицају сентимента на трговачку робу (енгл. *commodities*) издвојимо рад Бонатоа и сарадника (2020). Аутори су испитивали како нафта реагује на сентимент (праћен преко индекса среће, енгл. *the happiness index*) објава на друштвеној мрежи Твитер. Истраживање је показало да постоји само краткорочна веза између сентимента и волатилности приноса на нафту, што објашњавају спекулативним понашањем инвеститора. Ни хартије од вредности са фиксним приносом нису занемарене. Капорале (*Caporale*) и сарадници (2018) испитивали су утицај макроекономских вести на спредове између Немачких десетогодишњих државних обвезница и државних обвезница осам одабраних земаља Еврозоне. Резултати су показали да поменути утицај постоји, али и да је он израженији код негативних вести.

Крипто-валуте су идеалан примерак финансијске активе за коју треба спровести анализе сентимента због одсуства њихове фундаменталне вредности и склоности ка креирању ценовних балона (енгл. *bubble behavior*). У извештају које је израдио Голдмен Сакс (*Top of Mind: Crypto – A New Asset Class*), графички је анализиран утицај објављених вести на волатилност Биткоина. Из презентованих резултата недвосмислено се могло закључити да су крипто-валуте изузетно осетљиве на нове вести тј. нове информације. Упркос томе, када су 2017. Каралевицијус (Karalevicius) и сарадници публиковали свој рад о употреби анализе сентимента за предвиђање приноса на Биткоин и дефинисање инвестиционих стратегија, истакли су да је мало радова који користе анализу сентимента иако она има велики потенцијал у финансијама. Од тада па до данас ситуација се променила и све је више радова који испитују крипто-валуте не само текстуалном анализом (пре свега анализом сентимента), већ и другим алтернативним изворима података.

О атрактивности теме доста говори и дисперзија извора текстова на бази којих се сентимент инвеститора у крипто-валуте процењује. Најинтересантнији пример је рад Шапкоте и Грбиса (*Grobys*) (2023) у којем аутори анализирају сентимент садржаја такозваних „белих књига“ (енгл. *whitepaper*)[5], што је јединствен случај у постојећој литератури. Њихово истраживање је показало да је успешност садржаја беле књиге да пробуди оптимизам код инвеститора, отклони страх и приближи функционисање дате крипто-валуте, важан фактор за успех у прикупљању фондова приликом иницијалне јавне продаје. Насупрот томе, постојећа литература је преплављена примерима у којима се сентимент оцењује из објава на друштвеним мрежама[6], првенствено Твитеру. Овај феномен се једноставно може објаснити тиме да се на друштвеним мрежама лако може доћи до текстова за анализу. За овај приступ су се определили Маи (*Mai*) и сарадници (2015) у случају Биткоина, Крајевелда и Де Смета (2020) у случају Итиријума, Шашмаз и Тек (2021) у случају Неа, али и многи други. Неупоредиво мање радова сентимент инвеститора у крипто-валуте оцењује из вести. Међу њима разликују се два приступа. Према првом приступу анализира се утицај макроекономских вести. Ове вести су општије и самим тим доступније истраживачима од вести уско повезаних са појединачним крипто-валутама. Међутим, њихова предиктивна моћ је доста слабија. Корбет (*Corbet*) и сарадници (2020) су показали да само вести о незапослености и производњи трајних потрошних добара имају негативан утицај на кретање Биткоина, будући да стимулишу инвеститоре да улажу у друге облике финансијске активе. У случају вести о осталим макроекономским величинама статистички значајна веза није пронађена. Сличан резултат добили су и Ентроп (*Entrop*) и сарадници (2020) показавши да макроекономске вести немају никакав утицај на кретање фјучерса на Биткоин.

Други приступ подразумева анализу вести о самим крипто-валутама које су предмет истраживања. Ове вести су специфичније и већински их преносе портали који су доминантно оријентисани на крипто-валуте. Да би се до текстова дошло често је неопходно да истраживач добро познаје релевантне медије и информатичке технике за преузимање онлајн садржаја. Осим тога, овај приступ нуди мањи број текстова на дневном нивоу од приступа заснованог на друштвеним мрежама. Коначно, истраживања базирана на друштвеним мрежама су атрактивнија због популарности коју данас уживају саме друштвене мреже. Све су то разлози из који су истраживања заснована на текстовима са онлајн портала мање бројна. Без обзира на то, употребна моћ текстова са онлајн портала је изразито висока, о чему сведочи дисперзија њихових примена. Лмон (*Lamon*) и сарадници (2017), Во (*Vo*) и сарадници (2019) и Анамика (*Anamika*) (2022) су путем сентимента вести анализирали приносе крипто-валута. Бернарди (*Bernardi*) и сарадници (2017) су показали да се вредност под ризиком (енгл. *Value at Risk*) може прецизније проценити уколико се узме у обзир сентимент вести. До закључка да се ризик може прецизније проценити када се у обзир узме сентимент вести дошли су и Ђанкаја (*Cankaya*) и сарадници (2019) и Шапкота (2022) анализирајући волатилност Биткоина. Једно интересантно истраживање са специфичном идејом спровели су Роњоне (*Rognone*) и сарадници (2020). Аутори су покушали да дају одговор да ли се крипто-валуте понашају више као средство плаћања (тј. фиат валуте) или као финансијска актива (инструменти са

[5] Беле књиге представљају документа у којима се публикују ставови, мишљења и предлози владе, удружења, институције, предузећа или неке друге организације о одређеном питању. Назив су добиле по белим корицама уз које се штампају у САД-у како би се разликовале од других докумената које доноси влада. Када говоримо о крипто-валутама, у белим књигама се публикују визија и циљеви децентрализоване мреже, организација дистрибутивних евиденција (најчешће блокчеина), коришћена технологија, тип крипто-валуте и друге техничке карактеристике пре започињања самог пројекта.

[6] Поред Твитера, често коришћене друштвене мреже су и Редит, БиткоинТок, СикингАлфа, Фејсбук и др.

финансијског тржишта). Аутори су посматрали волатилност и приносе крипто-валута и њихове реакције на текстуалне вести поредећи их са онима код девизних курсева и обичних акција. Резултати су показали да се крипто-валуте тренутно понашају више као финансијска актива него као средство плаћања.

1.2.4 Ефикасност савремених тржишта

Као што се може видети из досадашњег прегледа, постојећа литература сугерише да се понашање економских променљивих може предвидети анализом сентимента. Према хипотези о ефикасности тржишта, коју су независно један од другог развијали Семјулесон (*Samuelson*) (1965) и Фама (*Fama*) (1963, 1965), тржиште је ефикасно уколико цене одражавају све доступне информације. Последично, кретање цена мора бити непредвидиво (тј. може се описати као случајан ход). Другим речима, никакви технички и фундаментални показатељи не би могли да предвиде будуће промене цена. У светлу нових доказа које је литература пружила, морамо се запитати да ли су финансијска тржишта заиста ефикасна. У последњих тридесетак година појављују се радови који тврде да цене не следе случајан ход, те да су у одређеној мери и у одређеном периоду предвидиве (Ло (*Lo*) и Мекинли (*MacKinlay*) 1988, Батлер (*Butler*) и Малаика (*Malaikah*) 1992, Кавусанос (*Kavussanos*) и Докери (Dockery) 2001, Галахар (Gallagher) и Тејлор (Taylor) 2002, Киан (Qian) и Рашид (*Rasheed*) 2007 и др.). Поред наведеног и успех појединих тржишних актера у остваривању абнормалног приноса, такође, сугерише да тржишта нису тако ефикасна као што је сматрано у конвенционалној литератури. То је навело поједине ауторе попут Де Лонга (*De Long*) и сарадника (1990) или Барбериса (*Barberis*) и сарадника (1998) да развију нове моделе финансијске економије. Аутори релаксирају претпоставку о рационалним инвеститорима и дозвољавају да један број актера у привреди буде вођен тренутним сентиментом. Овај тип инвеститора одлуке не доноси рационално већ тргује у складу са својим тренутним ставовима, расположењима и емоцијама. При томе формира се ефекат крда којим се врши притисак на цену тако да она може одступати од фундаменталне вредности на кратак рок. Због перзистентности притисака у кратком року, али и због других ограничавајућих фактора, до корекције цене неће доћи одмах, као што то сугерише хипотеза о ефикасности тржишта. Де Лонг и сарадници (1990) изводе закључак да на цену утичу две групе фактора: фундаментални фактори и тржишни ставови тј. сентимент инвеститора. Последично, тржиште се понаша неефикасно на кратак рок у којем је могуће предвиђати промене цене. Додатну аргументацију дали су Ле Барон (*LeBaron*) и сарадници (1999). Аутори су дизајнирали алгоритме којима се симулирају тржиште и тржишни актери према моделу Де Лонга и сарадника (1990). На тржишту су се повремено појављивале нове информације, затим би инвеститори доносили најбољу могућу одлуку у складу са њима. Симулација је показала да је заиста потребно неко време како би се тржиште довело у равнотежу и да је на кратак рок могуће профитирати и из техничке и из фундаменталне анализе. Једно објашњење овог феномена понудио је Ђидофалви (*Gidofalvi*) (2003). Аутор заступа теорију двадесето-минутног прозора прилика (енгл. *window of opportunity*) у којем је трговање на бази сентимента могуће и профитабилно. Прозор прилика постоји пре и после објављивања нове информације тј. нове вести. Први прозор је инсајдерски. Ђидовалви сматра да је двадесетак минута потребно да нова информација стигне у редакцију (или до особе која жели да је објави), припреми се за објављивање и, коначно, објави се. За то време један мањи број људи има информацију пре осталих, и може да је искористи за трговање. Други прозор је

вођен сентиментом. Овај прозор настаје зато што објављене информације неће одмах стићи до свих заинтересованих људи. Људи користе различита средства информисања, те је потребно мало времена док је сви медији пренесу. Истовремено, потребно је да протекне одређено време док људи не постану свесни да се нова информација појавила, као и да предузму одређене акције када информацију коначно добију. У том периоду, који према Ђидофалвију (2003), такође, траје око двадесетак минута, појединци који су први дошли до информација пожуриће да их искористе и покренуће ефекат крда. Из тог разлога тржиште је неефикасно само на веома кратак рок, те је превиђање цена и профитабилно трговање могуће само унутар дневног трговања (енгл. *intraday trading*).

Овакви резултати навели су један број истраживача који се баве анализом сентимента да се у оквиру својих истраживања осврну и на сигнале о тржишној ефикасности које добијају из њих. Њихови закључци темељили су се на покушајима предвиђања кретања приноса на бази сентимента, на проверавању значајности везе између ове две величине или на тестовима узрочности. Аутори међу којима су Болен и сарадници (2011), Шумахер и сарадници (2012), Касем и сарадници (2015) добили су сигнале који сугеришу да анализирана финансијска тржишта нису краткорочно ефикасна. До истих сигнала дошли су и Во и сарадници (2019) и Крајевелд и Де Смет (2020) анализирајући сентимент крипто-валута. Насупрот њима, Рено (2020) није успео да на бази сентимента направи задовољавајуће предикције приноса, те истиче да његови резултати не дају разлога за сумњу у тржишну ефикасност.

Поред радова који због предвидљивости приноса сумњају на тржишну неефикасност код крипто-валута, постоје и други разлози који априори буде исте сумње. Пре свега, до пре пар година могло се говорити о некомплетности тржишта крипто-валута. Овај проблем се дуговао мањку финансијских инструмената којима се могло трговати на тим тржиштима. О томе довољно говори чињеница да су први деривати на крипто-валуте уведени тек крајем 2017. године на Чикашкој берзи. Крајевелд и Де Смет (2020) сугеришу да је тржиште крипто-валута још увек младо и да је потребно да протекне одређено време док оно не сазри и инвеститори не постану потпуно упознати са њим. Џајан (*Ciaian*) и сарадници (2016) и Собјетов (*Sovbetov*) (2018) приметили су да су цене крипто-валута различите на различитим берзама, што их чини рањивим на арбитражу и продубљује њихову осетљивост на прилив сваке нове информације. Још један аргумент у прилог неефикасности тржишта крипто-валута је њихово понашање које ствара ценовне балоне. Бројни су радови који тематизују ово питање. Међу њима је и пионирска примена алтернативних извора података на крипто-валуте. Реч је о раду Кристуфека (2013) у којем се испитује веза између фреквенције интернет претрага на Гуглу и Википедији и вредности Биткоина. Резултати су демонстрирали да је реч о симултаној, односно, двосмерној вези. Што је интересовање људи веће (отелотворено кроз већи број интернет претрага) то ће цена Биткоина бити већа. Истовремено, што је цена већа, то је и већи број интернет претрага (тј. интересовање расте). Према речима аутора, оваква веза основ је за формирање само-испуњујућег спекулативног балона на тржишту крипто-валута. До истог закључка дошли су и аутори радова новијих датума попут: Фреја (*Fry*) и Чеја (*Cheah*) (2016), Филипса (*Phillips*) и Горса (*Gorse*) (2017), Корбета и сарадника (2018) и Алборга (*Aalborg*) и сарадника (2019).

Ако говоримо о формалним тестовима тржишне ефикасности највећи број радова се позабавио питањем слабе форме тржишне ефикасности. Њима ће се придружити и ова дисертација. Једно од првих и најцитиранијих истраживања спроведених на ову тему публиковао је Урхарт (*Urquhart*) (2016). Поменути аутор је посматрао приносе *BTC*-а и показао да су исти аутокорелисани и неслучајни, као и да се не понашају као прирасти случајног хода. У те сврхе употребљени су тестови случајности, тестови аутокорелисаности, аутоматски тест количника варијанси и Хрустов рацио. Сви тестови су једногласно потврдили нарушеност слабе форме тржишне ефикасности на тржишту BTC-а у периоду од 2010-2016. Међутим, уколико би се узорак скратио и посматрао се само период ближи 2016. години тестови више не би били једногласни, што према речима аутора сугерише да тржишна ефикасност варира кроз време. Чеј и сарадници (2018) показују да цене *BTC*-а са 5 одабраних тржишта (цене *BTC*-а на 5 тржишта (серије: *BTC/USD, BTC/CAD, BTC/AUD, BTC/GBP* и *BTC/EUR*)) одликује дуга меморија, те да се не понашају као случајан ход. Аутори закључују да су тржишта *BTC*-а умерено до веома неефикасна и анализа меморије у ценама може помоћи инвеститорима да остваре спекулативни профит. Слично томе, Банди (*Bundi*) и Вилди (*Wildi*) (2019) су оспорили тржишну ефикасност показавши да се приноси *BTC*-а у периоду 2014-2019. године могу описати као MA(6) процес. На бази претходних резултата аутори су формирали неколико стратегија трговања и демонстрирали да оне могу да донесу статистички значајан профит. Једно иновативно истраживање спровео је Кристуфек (2018) који је искомбиновао неколико постојећих мера ефикасности у јединствен показатељ назван индекс ефикасности (енгл. *the efficiency index*). Истраживање је дало чврсте доказе да је неефикасност *BTC*-а била перзистентна кроз време у периоду 2014-2017. године. Са растом популарности крипто-валута и са повећањем њихове бројности, аутори су полако почели да се окрећу и другим припадницима ове класе финансијске активе. Истраживање које су спровели Паламалаи (*Palamalai*) и сарадници (2021) тематски је најближе ономе што ће читалац моћи да види у овој дисертацији. Поменути аутори су применили више тестова слабе форме тржишне ефикасности (међу којима су тестови случајности, тестови јединичног корена, тестови количника варијансе и моделирање случајних процеса за описивање кретања приноса) на 10 одабраних крипто-валута. Резултати које су добили указивали су на нарушеност слабе форме тржишне ефикасности. Поред претходног, и рад Кима (*Kim*) и сарадника (2023) добар је пример да закључци о одсуству ефикасности тржишта крипто-валута опстају до данашњег дана. Рад је успео да оспори слабу форму тржишне ефикасности код чак 15 крипто-валута тестовима јединичног корена, при чему је посебна пажња поклоњена нелинеарном квантилном тесту јединичног корена (енгл. *quantile unit root*).

Радови који проналазе доказе о важењу слабе форме тржишне ефикасности код крипто-валута далеко су мање бројнији. Баривиера (*Bariviera*) (2017) је дошао до закључка да се тржиште *BTC*-а понашало ефикасно током 2014. године ослањајући се на динамички приступ испитивању слабе форме тржишне ефикаснсоти заснован на Хурстовом експоненту и анализи флуктуације без тренда (енгл. *detrended fluctuation analysis – DFA*). Сенсој (*Sensoy*) (2019) је први који је испитивао тржишну ефикасност крипто-валута на унутар дневним подацима. Користећи мере ентропије и спектралне зависности, истакнути аутор је показао да унутар дневне цене *BTC*-а почињу да се понашају у складу са слабом формом тржишне ефикасности током 2016. године, али да су се до 2018. године смењивали периоди тржишне ефикасности и тржишне неефикасности. Аутор сугерише да су његови резултати потврда да постоји цикличност у кретању тржишне ефикасности BTC-а. Ји (*Yi*) и сарадници (2023) су користили квантни хармонијски осцилатор (енгл. *quantum harmonic oscillator – QHO*) и тест количника

варијанси да испитају промене у слабој форми тржишне ефикасности *BTC*-а кроз време. Њихови општи закључак је да тржиште BTC-а није ефикасно, али да је близу тога да постане ефикасно, као и да се кроз време умањио простор за профитабилне спекулативне стратегије. Кумар (*Kumar*) и сарадници (2020) испитивали су слабу форму тржишне ефикасности на бази приноса на *CRIX* индекс. Како *CRIX* индекс представља апроксимацију за кретање целокупног тржишта крипто-валута, циљ њиховог истраживања био је да се дође до генералних закључака о тржишној ефикасности за све крипто-валуте. Ослањајући се на резултате неколико тестова слабе форме ефикасности дошли су до закључка да се тржиште током 2020. године понашало ефикасно. Наведени докази из постојеће литературе сугеришу да тржишта крипто-валута још увек нису ефикасна, као и да њихова ефикасност варира кроз време. У складу са тим, код крипто-валута уместо класичне хипотезе о ефикасности тржишта примереније је говорити о такозваној алтернативној хипотези о тржишној ефикасности коју је установио Ло (*Lo*) (2004). Према овој хипотези у савременим условима тржишта нису непрестано ефикасна, али су адаптивна и конкурентска. Ниво њихове ефикасности зависиће од промена у окружењу, популацији инвеститора и њиховим ставовима. Из тог разлога у савременим условима постоје периоди када се тржиште понаша ефикасно, као и они у којима се понаша неефикасно. Према томе, адаптивна хипотеза о тржишној ефикасности дозвољава коегзистирање рационалности и бихевиоризма што је можда добар оквир за описивање тржишта крипто-валута.

1.3 Циљеви и научни доприноси

Након осврта на актуелну литературу, стечени су услови да се читалац упозна са циљевима дисертације, али и доприносима које иста намерава да пружи. Кроз њих ће читалац јасно сагледати тежње ка којима дисертација стреми, али и истраживачки јаз који ће њоме бити премошћен.

Дисертација има два важна усмерења. Прво је методолошко, и односи се на одређена статистичка и информатичка унапређења. Поред тога, дисертација ће понудити алтернативни приступ оцењивању пондера сентимента из модела Џагадиша и Вуа (2019) у циљу подизања њихове употребне моћи. Побољшани модел пружиће могућност новим истраживачима да ефикасније оцењују сентимент из текстова и самим тим успешније предвиђају кретање економских величина или анализирају утицај сентимента на њих. Додатно, дисертација ће представити и свој тест за вишеузорачко испитивање квалитета прогнозе два модела. Као што је раније истакнуто, реч је о модификацији теста Диеболд-Маријана (1991). Предложени тест доноси суд о квалитету прогнозе два модела поредећи њихове предикције на N независних узорака, а не на једном као што је то до сада био случај. Дисертацији ће приказати и свој алгоритам рударења текста и општи лексикон променљивих речи и синонима. Дизајн алгоритма за рударење текста је важан јер он припрема текстове за квантитативне анализе. Предност алгоритма презентованог у одељку **3.3.1** је то што је он специјално прилагођен потребама рударења текстова о крипто-валутама (због жаргона, терминологије и других лингвистичких појединости). Израђени лексикон дизајниран је да буде подршка алгоритму за рударење текста. Овде није реч о лексикону сентимента, у којем се поред сваке речи назначава њен сентимент, већ о техничком лексикону, који обухвата често коришћене променљиве речи у енглеском језику и њихове најближе синониме. На тај начин додатно се смањује могућност

грешке приликом лематизација, уводе се и неологизми повезани са крипто-валутама и води се рачуна о синонимији. О овим питањима детаљније ће се говорити у одељку **3.3.1**.

Друго усмерење дисертације је емпиријско и уско је повезано са претходним. Израђени алати биће упослени у анализи сентимента, али и у рударењу других показатеља из текста са циљем њиховог коришћења за предвиђање приноса и испитивање постављених хипотеза. Реч је о питањима која привлаче пуно пажње у постојећој литератури, будући да су крипто-валуте млада и још увек недовољно истражена класа финансијске активе. Ова дисертација покушаће да да своје виђење датих проблема. Наиме, спроведено истраживање издваја се од већине постојећих из више разлога. За почетак, истраживање се ослања на иновативну методологију. Такође, према најбољим сазнањима аутора, у тренутку писања дисертације по први пут у постојећој литератури примењен је модел Џагадиша и Вуа (2019) за оцену сентимента на текстове о крипто-валутама. Истраживање доприноси постојећој литератури и по томе што своје анализе заснива на текстуалним вестима са интернет портала. Заступљеност оваквих истраживања је мања у поређењу са онима базираним на друштвеним мрежама. Већина истраживача опредељује се за текстове са друштвених мрежа због лакоће њиховог прикупљања, њихове масовности на дневном нивоу, и популарности самих друштвених мрежа. При томе, истраживање се базира на другачијем извору вести. Реч је о порталу *cryptonews.net*. Поменути портал се показао високо поузданим и последњих година стекао је високу популарност међу привредним актерима. Портал прикупља и објављује искључиво вести повезане са крипто-валутама. По томе се издваја од других извора вести коришћених у постојећој литератури које одликује општији опус интересовања (на пример: *Yahoo!Finance, RavenPack Dow Jones News Analytics, NewsNow, LexisNexis* и сл.). С обзиром на то да је приступ сајту бесплатан, текстови су прикупљени путем скреповања, а не комерцијалним путем. Још један битан аспект који издваја дисертацију од постојеће литературе је ширина истраживања. За разлику од већине истраживача, који анализирају једну или пар крипто-валута, ова дисертација ће истовремено пратити осам најпопуларнијих[7] међу њима. Код сваке од њих, поред сентимента саме валуте, биће анализиран и утицај унакрсног сентимента и читљивост објављених текстова. Испитивање утицаја унакрсног сентимента, према најбољим сазнањима аутора, издваја ову дисертацију од постојеће литературе. Коначно, истраживањем је обухваћен специфичан период из више разлога. У питању је први квартал 2022. године, који обухвата почетак кризе на истоку Европе и представља прекретницу у глобалном привредном циклусу. Почетак светске рецесије узео је свој данак и на тржишту крипто-валута. То је био специфичан тренутку у њиховом развоју. Изузев Биткоина, током светске финансијске кризе друге крипто-валуте нису постојале. Прва наредна крипто-валута, настала је тек у октобру 2011. године, док је на почетку 2013. постојало тек 4 крипто-валуте. Уједно први подаци о цени Биткоина доступни су нам тек од средине 2010. године, док се за остале крипто-валуте бележе тек од 2014. године[8]. Са друге стране, период пандемије изазване вирусном инфекцијом КОВИД19 погодовао је крипто-валутама. Док је поменути период у осталим секторима обележила криза, популарност крипто-валута доживела је невероватан успон. У 2022. години коначно нам је пружена прилика да сагледамо понашање крипто-валута у условима трансмисије глобалне кризе на њихова тржишта. С тим у вези, ово истраживање ће проверити каква је могућност примене анализе сентимента за предвиђање приноса крипто-валута у кризном и преткризном периоду. Додатно, дисертација ће испитати валидност слабе форме хипотезе о ефикасности

[7] Према броју објављених текстова.
[8] Према: *Yahoo!Frinance* и *CoinMarketCap*.

тржишта крипто-валута у првом кварталу 2022. године. Питање тржишне ефикасности у једном овако специфичном периоду је посебно интересантно узевши у обзир доказе из постојеће литературе да тржишна ефикасност код крипто-валута варира кроз време.

Да резимирамо, дисертација даје свој допринос у три научне дисциплине. Када је реч о емпиријском доприносу, дисертација ће испитати утицај неколико текстуалних показатеља на крипто-валуте кроз специфично дизајнирано истраживање (шири обухват валута, иновативна методологија и иновативни извори вести). Међу њима посебно је важно издвојити испитивање утицаја унакрсног сентимента. Паралелно са претходним, провериће се могућност предвиђања кретања приноса на крипто-валуте и ефикасност тржишта на зачетку светске рецесије. Са становишта методолошког доприноса, дисертација ће понудити алтернативни приступ оцењивању пондера сентимента по узору на модел Џагадиша и Вуа (2019). Поред тога, дисертација се може похвалити и тиме што ће представити прилагођену верзију Диеболд-Маријановог (1991) теста за вишеузорачко тестирање квалитета прогнозе (која ће бити изложена у оквиру теореме 2). Додатно, представиће се и алгоритам погодан за предвиђање приноса заснован на показатељима добијеним рударењем текста и ансамблима. Коначно, дисертација развија свој алгоритам за рударење текста са лексиконом променљивих речи у енглеском језику са блиским синонимима и обезбеђује програмерску имплементацију свих представљених решења. Поред истакнутих доприноса, оно чиме ова дисертација може посебно да се похвали је чињеница да ће ово бити једно од првих истраживање овог типа спроведено у Републици Србији.

1.4 Полазне хипотезе

Дисертација ће проверити валидност три тесно повезане хипотезе означене ћириличним словима А, Б и В. У наставку ове секције постављамо и анализирамо сваку од њих појединачно. Начин на који ће постављене хипотезе бити испитане дисертација ће представити у методолошком поглављу.

Н$_А$: „Постоји веза између сентимента (ставова) вести о крипто-валутама објављених на популарним онлајн порталима и приноса на крипто-валуте.“

Првопостављена хипотеза претпоставља да информације које долазе из вести са интернет портала могу утицати на инвестиционе одлуке крипто-трговаца (купи, продај или држи) и пословне одлуке крипто-рудара (рудари или не рудари). На тај начин, информације из вести директно утичу на понуду и тражњу за крипто-валутама, а самим тим и на њихову цену (тј. приносе). Изнети став се додатно аргументује чињеницом да популацију крипто-рудара и крипто-трговаца чине млади људи. Млађа популација типично не поклања пуно пажње класичним медијима, већ за своје информисање користи практичније изворе. Порталима се може приступити бесплатно, брзо и једноставно путем интернета, а већина их је доступна и као апликација за паметни телефон. Поред тога, често су уско специјализовани за одређени садржај и концептуално су ближи млађим генерацијама. Старосно доба заинтересоване

јавности игра још једну важну улогу. Млађи људи су много флексибилнији и отворенији ка новитетима и спољашњим утицајима. О томе сведоче бројне полемике у друштву, медијима и научним радовима о утицају интернета, друштвених мрежа и нове технологије уопште на младе људе. С тога је природно очекивати да ће информације из онлајн вести имати значајан утицај на њихове ставове и економске одлуке. Овај утицај је можда чак и значајнији од утицаја који вести имају код других врста финансијских актива, због одсуства фундаменталне вредности крипто-валута. Међутим, постојање ове везе има своје економске последице. Према класичној хипотези о полу-јакој форми ефикасности финансијских тржишта, тржиште се сматра ефикасним уколико тржишна цена одражава све јавно доступне информације. Другим речима, нико не може да искористи јавно-доступне информације да предвиди будуће кретање цена, односно, приноса. С тим у вези, валидност овако постављене хипотезе била би сигнал потенцијалне нарушености полу-јаке хипотезе о тржишној ефикасности код крипто-валута, што, наравно, треба проверити посебним истраживањем.

Нб: „Постоји веза између приноса крипто-валута и читљивост вести написаних о њима, као и веза између приноса крипто-валута и сентимента вести о Биткоину.“

Друга хипотеза у разматрање уводи две додатне променљиве добијене рударењем текста. Претпоставка о утицају читљивости изводи се из првопостављене хипотезе о утицају текстова на ставове заинтересоване јавности. Наиме, ако је текст слабо читљив, његова порука читаоцу неће бити јасна, а самим тим неће ни имати утицај на њега. Анализа читљивости нашла је своје место у овом истраживању јер се у постојећој литератури показало плодоносним укључивање неког техничког својства текста у анализу утицаја вести на економске променљиве. Питањем читљивости детаљније ћемо се позабавити у одељку **3.5**.

Друга променљива коју уводи хипотеза Б је сентимент вести о Биткоину. Као што је већ раније истакнуто, утицај сентимента инвеститора о једном финансијском инструменту на приносе другог финансијског инструмента називамо унакрсним сентиментом. Питање унакрсног сентимента покренуто је у предистраживању ове дисертације публикованом у раду Дамјановића и Дреновака (2023). Оно је посебно интересантно код крипто-валута јер су оне међусобно веома сличне[9] и прате кретање Биткоина[10]. Додатно, ово питање је важно и са микроекономског становишта. Добро је позната чињеница да се на тражњу за неким добром могу одразити ставови појединаца о другим сродним добрима, што крипто-валуте међусобно јесу. Будући да се Биткоин успоставио као суверени лидер који руководи тржиштем крипто-валута, биће занимљиво сагледати какав утицај имају ставови инвеститора о њему на друге крипто-валуте. Осим тога, биће занимљиво сагледати и утицај унакрсног сентимента на сам Биткоин. У те сврхе послужиће нам сентимент инвеститора о другој највећој (по тржишној капитализацији) и другој најпопуларнијој крипто-валути, Итиријуму (*ETH*).

Прихватање ове хипотезе сугерисало би постојање везе између кретања приноса и показатеља добијених рударењем текста (које сврставамо у домен техничке анализе). То је, такође, у

[9] На висок степен њихове сличности указали су и Џајан и сарадници (2016).
[10] Видети Кумар и Ацаз (2019), Аседоку (2019) и др.

супротности са хипотезом о полу-јакој форми тржишне ефикасности, што ће бити још један сигнал потенцијалне тржишне неефикасности крипто-валута.

Нв: „Грешке у прогнози приноса на бази сентимента текстова онлајн вести ће бити мање када се користи иновирана методологија оцењивања сентимента уместо оригиналне методологије Џагадиша и Вуа (2019)“

Џагадиш и Ву (2019) дефинисали су модел за мерење нивоа, односно, магнитуде сентимента појединачних речи директно из података, односно, из резултата добијених рударењем текста. Ниво сентимента речи аутори су назвали пондерима сентимента. Оцене пондера које су аутори предложили, у ознаци $\hat{B}_j$, садржале би грешку мерења, те процена сентимента текста на бази њих не би била адекватна. Полазећи од те чињенице, јасно је да би се отклањањем или умањивањем грешке мерења добила супериорнија, односно, прецизнија процена магнитуде сентимента. Добијање прецизнијих оцена је од велике важности за даље анализе, будући да се сентимент користи за предвиђање кретања економских променљивих. Ова дисертација покушаће да да свој допринос у решавању овог проблема налажењем примерније оцене пондера сентимента, у ознаци $\widehat{w}_j$. Сентимент процењен на бази предложених оцена биће искоришћен за предвиђање кретања приноса на крипто-валуте. Квалитет добијених прогноза биће упоређен са онима добијеним на бази сентимента измереног уз помоћ оцена Џагадиш и Ву (2019). Добијање квалитетнијих прогноза говорило би у пролог супериорности оцене које ова дисертација заступа и њихову примену у будућим анализама сентимента. Квалитетније прогнозе сугерисале би и да је кретање приноса код крипто-валута предвидљивије него што се верује. То би био још један сигнал потенцијалних тржишних неефикасности.

Сагледавши све три хипотезе, јасно је да ће се током целог тока истраживања посредно прикупљати сигнали о тржишној ефикасности за тржиште крипто-валута. Провера истинитости добијених сигнала остављена је за крај истраживања. На тај начин формално ће се испитати слаба форма тржишне ефикасности анализираних крипто-валута, док ће испитивање полу-јаке форме тржишне ефикасности бити одложено за посебно истраживање. На тај начин дисертација пружа својеврстан допринос дебати у корист и против тржишне ефикасност крипто-валута и допуниће литературу сазнањима о ефикасности ових тржишта у условима зачетка глобалне рецесије. Овај аспект истраживања имаће посебну важност због економских последица које неефикасност имплицира.

1.5 Извори података

Зарад спровођења истраживања неопходне су две врсте података. Пре свега неопходно је обезбедити податке о ценама како би се из њих обрачунали приноси чије кретање желимо да објаснимо и предвидимо. Други скуп података чине онлајн вести (текстови) на којима ће се анализа заснивати. До података о ценама крипто-валута много је лакше доћи. У те сврхе, рад се послужио чувеном америчком онлајн платформом за финансијске податке – *YahooFinance!*.

Како би узорак приноса обухватио период 01.01.2021. – 22.03.2022., цене крипто валута су преузете за период 31.12.2020. – 22.03.2022. (погледати секцију о обрачуну приноса **3.6**). Преузимање је обављено програмерским путем. Серије које су анализиране у овом истраживању односе се на крипто-валуте: *ADA*, *AVAX*, *BTC*, *DOGE*, *DOT*, *ETH*, *LUNA* и *SOL*. Са сваком од ових крипто-валута читалац ће имати прилику да се детаљније упозна у одељку **2.6**. У међувремену, дисертација овде читаоца упознаје са техничким особинама преузетих серија. Са тим у вези у наставку су дате основне дескриптивне статистике у Табели **1** и графички прикази серија на Слици **1**.

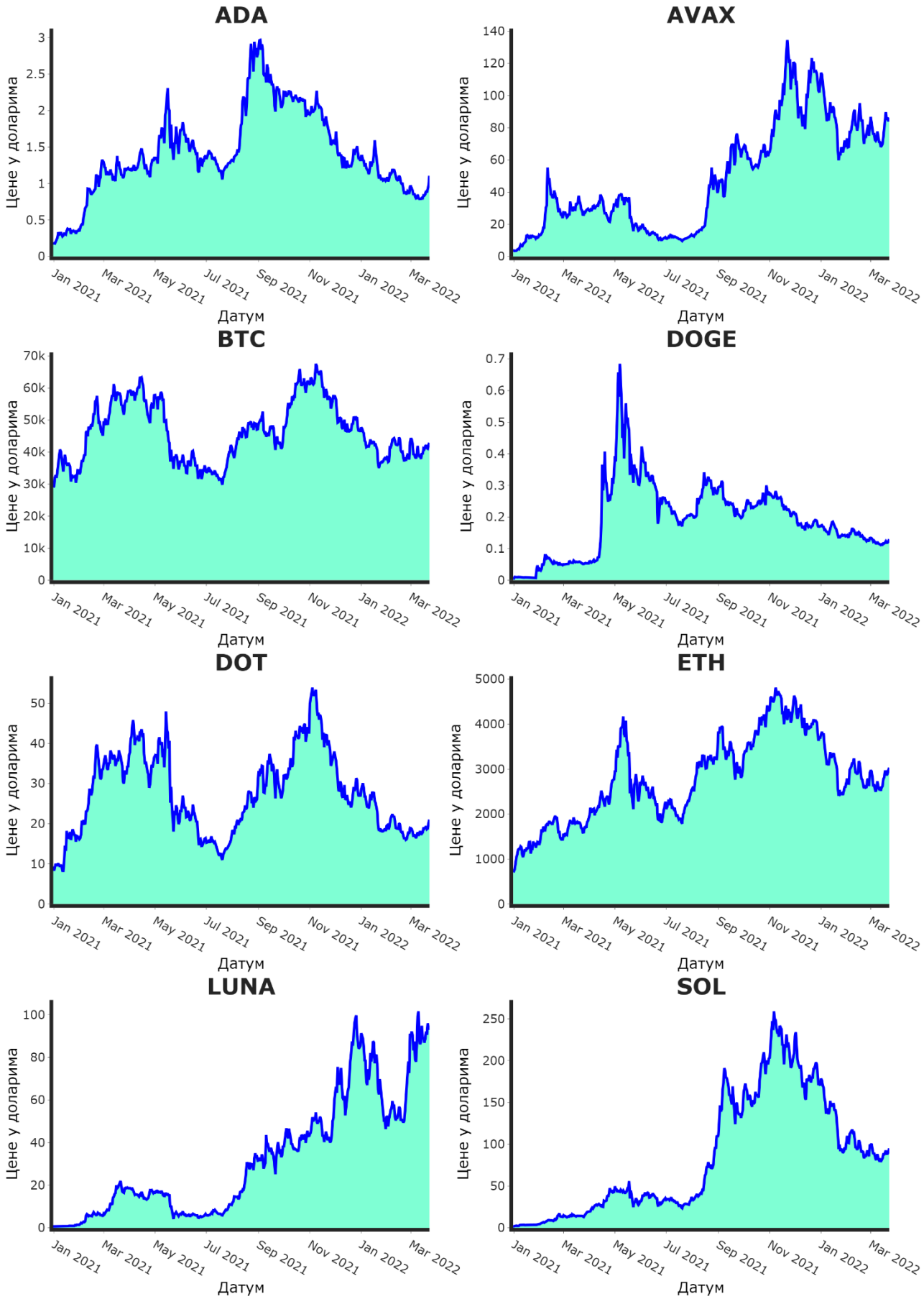
Табела 1: Дескриптивне статистике цена осам одабраних крипто-валута

	ADA	**AVAX**	**BTC**	**DOGE**	**DOT**	**ETH**	**LUNA**	**SOL**
Медијана	1.32	38.25	46033.88	0.19	26.64	2741.43	19.66	47.28
Аритмт. сред.	1.42	49.03	46190.92	0.19	27.45	2796.77	32.51	84.97
Минимум	0.18	3.14	29001.72	0.01	8.26	730.37	0.63	1.51
Максимум	2.97	134.53	67566.83	0.68	53.88	4812.09	101.59	258.93
Ст. дев.	0.59	33.15	9303.68	0.12	10.18	941.20	28.04	71.39
Асиметрија	0.34	0.51	0.26	0.76	0.31	0.11	0.76	0.63
Спљоштеност	0.02	-0.89	-0.99	1.42	-0.73	-0.83	-0.59	-0.93

Са слике се види да је кретање анализираних крипто-валута прилично усаглашено. Примера ради, готово све серије одликује умерен пад вредности средином 2021. године, а затим оштар и нагли раст. Приметимо и да се ниво цена драстично разликује од валуте до валуте. Код неких говоримо о центима, док код неких говоримо о хиљадама долара. Такође, приметна је и велика волатилност. Нивои цена пуно осцилирају и у кратким периодима, присутни су пикови (екстремне вредности) и велике разлике између минималне и максималне цене у узорку. То се може сагледати и из дескриптивних статистика. Стандардне девијације су поприлично високе, док су минимум и максимум поприлично удаљени. Додатно, у већини случајева медијана је била нешто нижа од аритметичке средине што такође потврђује присуство позитивних екстремних вредности.

Када говоримо о вестима (текстовима), за потребе овог истраживања преузето је око 40 000 текстова написаних на енглеском језику о изабраним крипто-валутама. Реч је о текстовима који су били објављивани у периоду: 01.01.2021. – 22.03.2022. год. на порталу *cryptonews.net* (у даљем тексту: *Крипто Њуз*). Текстови су, такође, преузети програмским путем уз помоћ технике веб скрепинга (алгоритам је представљен у секцији **3.2**). У наставку ове секције дискутујемо релевантност портала Крипто Њуз као извора вести о крипто-валутама.

Слика 1: Графички приказ кретања цена осам одабраних крипто-валута



Извор: *Yahoo!Finance* (приступљено: 23.03.2022.)

1.5.1 Зашто Крипто Њуз?

У данашње време, интернет корисницима је доступан огроман број онлајн портала. У прилог томе најбоље говори чињеница да су заинтересованој јавности услед ефекта глобализације данас доступни чак и портали из њима далеких земаља. То пред интернет кориснике и истраживаче ставља тежак задатак избора добрих и релевантних извора информација из тог „океана“ потенцијалних извора које интернет нуди. Ова одлука је од изузетног значаја јер избором неадекватног извора информација истраживачи и корисници добијају погрешне или застареле информације, што ће се, последично, одразити на доношење погрешних економских одлука и добијање нетачних резултата из истраживања. Са друге стране, никада није добро ослонити се само на један извор информација. Ово питање је посебно важно уколико не изгубимо из вида могућност да аутори субјективно приступе неком питању, фаворизују одређени став под притиском нечијег лобирања или да им одређена важна вест промакне.

Портал Крипто Њуз је агрегатор вести. Као такав, овај портал објављује најважније вести (за сваку крипто-валуту посебно) преузете у потпуности из других релевантних извора. Свакодневно међународни и мултикултурални тим стручњака ручно прати преко 300 извора вести из области крипто-валута, блокчејна и финтека. Међу изворима које Крипто Њуз користи су и: *NewsBTC, CryptoGlobe, CoinGecko, CoinTelegraph, CoinDesk, Forbes, AMBCrypto, TheBlock, U.Today, CryptoDaily* и *CoinGape* (али и многи други). Новински чланци су одабрани на начин да покривају широк спектар тема. Приступ „за сваког по нешто” омогућио је да се сви чланови заинтересоване јавности, без обзира на циљеве, информишу на истом месту. Теме као што су: иновација у технологији која прати процес рударења, берзански догађаји, техничке анализе кретања цена, употреба крипто-валута у пракси широм света, трачеви и дискусије везане за дебату „за и против“ крипто-валута итд. , само су део онога што овај портал нуди. Из наведених разлога за платформу Крипто Њуз, можемо рећи да је добар диверсификован, непристрасан и релевантан извор информација.

Још једно важно питање је популарност портала. Популарност је доказ да ће објављене вести допрети до шире јавности. Крипто Њуз је себи за циљ поставио да буде најбоље и најсвеобухватније дигитално средство информисања јавности заинтересоване за крипто-валуте широм света. Овако високо постављени циљеви, доступност портала у виду апликације за паметне телефоне и превођење објављених текстова на четири светска језика допринели су да портал придобије велику популацију читалаца широм планете. Осим тога, посебан допринос популарности портала даје ширина тема које портал обрађује, као и самих извора које портал користи.

Додатна предност коришћења агрегатора вести је што је анализа ефикаснија. Истраживачи добијају вести из различитих извора дефинисањем само једног алгоритма за скреповање.

1.6 Оквиран преглед садржаја

У наставку ове дисертације излагање ће бити организовано на следећи начин. Друго поглавље бави се крипто-валутама. Оно ће читаоцу омогућити сагледавање свих аспекта повезаних са крипто-валутама, који су релевантни за потпуније разумевање финалних резултата. У оквиру њега биће објашњен њихов настанак, затим њихово функционисање, њихове предности, њихови недостаци, њихова примена и њихове врсте. На самом крају друге секције укратко ће бити представљене и саме крипто-валуте које учествују у овом истраживању. Ово поглавље је важно и због тога што је дисертација за један од својих циљева поставила ширење знања о крипто-валутама. Наредна секција биће методолошка. У њој ће читалац најпре моћи да се упозна са током истраживања и његовим дизајном. Затим ће бити представљени израђеним информатичким алатима за прикупљање и рударење текстова. Након тога уследиће преглед метода коришћених за обрачун приноса и читљивости текстова. Потом следи представљање начина на који ће хипотезе бити тестиране, опис алгоритма за предвиђање приноса, опис коришћених тестова прогнозе и слабе форме тржишне ефикасности. Поглавље ће такође представити вишеузорачки тест квалитета прогнозе који заступа ова дисертација. Четврто по реду поглавље увешће читаоца у проблематику мерења сентимента. У њему ће бити представљена иновирана методологија оцењивања пондера сентимента коју предлаже ова дисертација. У оквиру истог поглавља биће представљене и финалне оцене сентимента које ће бити коришћене у емпиријском истраживању. Претпоследње поглавље представиће добијене резултате. Најпре ће бити изложене оцене сентимента. Затим ће бити анализиране везе између показатеља добијених рударењем текста и приноса одабраних крипто-валута. Потом ће се размотрити и успешност конструисаног ансамбл модела у предвиђању приноса. Овде ће се, такође, анализирати алтернативни облици финалног ансамбл модела и закључци добијени на бази њих. На крају овог поглавља биће дати резултати тестова слабе форме тржишне ефикасности код анализираних крипто-валута. Последња секција даће завршну реч. У оквиру ње сублимираће се добијени резултати и још једном ће се подвући њихов научни допринос. Дисертација се завршава прегледом литературе, биографијом аутора и пратећим изјавама.

2. Крипто-валуте и информатичке иновације у финансијској технологији које су их створиле

Нове појавне облике финансија настале развојем информационих технологија (хардвера и софтвера), али и распрострањивањем процеса умрежавања, заједнички називамо дигиталне финансије (енгл. *digital finance*). Неки од примера дигиталних (савремених) финансија су: онлајн и мобилно банкарство (енгл. *web- and m-banking*), мобилно плаћање (енгл. *m-payments*), алгоритамско трговање (енгл. *algo-trading*) и роботизовано саветовање (енгл. *robo-advising*). Заједнички назив за све технолошке иновације настале у сфери финансија од појаве интернета до данашњег дана је енглеска кованица финтек (енгл. *fintech*). Ова реч је добијена спајањем речи: „финансије“ и „технологија“, а њена употреба широко је распрострањена у пракси.

Ова дисертација анализира само један аспект дигиталних финансија – крипто-валуте. С тим у вези, у наставку овог одељка ће читаоцу бити приближен концепт крипто-валута и финтека који стоји иза њих. Циљ овог поглавља је да осигура да читалац разуме начин на који крипто-валуте функционишу и све њихове специфичности. Остваривање овог циља је важно за разумевање финалних резултата представљених у последњем одељку дисертације. Зарад систематичности, у наставку поглавља биће представљен финтекови повезани са крипто-валутама (интернет децентрализација, криптографија и блокчејн), а затим ће се прећи на анализу самих крипто-валута њихових особина и подврста.

2.1 Развој интернета – стварање потребе за децентрализацијом

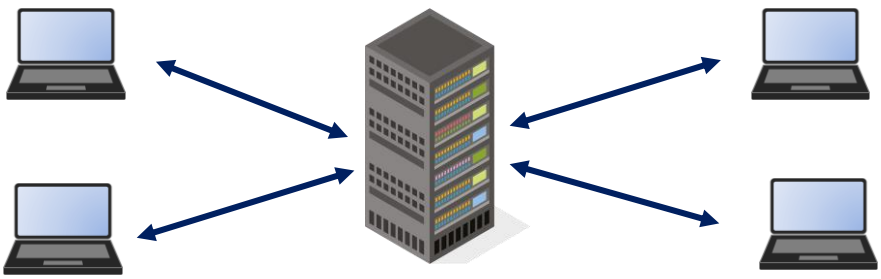
Интернет је први пут ушао у домове људи 90-их година прошлог века путем телефонског кабла (енгл. *dial-up system*). Ова технологија са собом је донела могућност да корисник не напуштајући удобност свог дома добије све информације од интереса једноставном онлајн претрагом. Ипак, морамо истаћи да су суштински могућности крајњег корисника биле скромне, будући да је корисник могао само да „чита“ садржај постављен на интернету, али не и да сам нешто поставља на њега. Истовремено многа предузећа препознала су интернет као прилику да приближе своје производе и услуге потрошачима путем сопствених веб страница, те се један део привредне активности сели у дигитални простор. Ови бенефити условили су да се интернет веома брзо преселио из Сједињених Држава у друге земље, па и на друге континенте. Ову еру у развоју интернета називамо Вебом (енгл. *Web*) 1.0.

Даљи развој интернета донео је са собом бројне промене почетком 2000-их. У том периоду појавили су се бесплатни претраживачи, онлајн трговина је доживела велику експанзију, телефонски кабл је заменио мрежни кабл, и сл. Ипак, ни једна од промена није имала тако далекосежно дејство као појава кориснички оријентисаних платформи. Реч је о платформама које су омогућавале кориснику да самостално дели неки садржај на интернету (пише поруке, шаље слике, објављује видео клипове и слично). У односу на дотадашње могућности интернета, кориснички оријентисане платформе биле су револуционаран корак унапред. Иза ових платформи стајали су сервери великог капацитета у власништву приватних корпорација. Сервери су прикупљали садржај који су корисници желели да поделе и дистрибуирали их ка другим корисницима. Кориснички оријентисане платформе биле су основ за настанак

друштвених мрежа (Фејсбука, Твитера, Инстаграма…), блогова (Вордпреса, Блогера, Тамблер…), веб-апликација (Воцапа, Вибера, Слека…), портала за дељење садржаја (Јутјуба, Фликера, Пинтереста…) и других посредничких окружења. Последично, и други садржаји директно или индиректно повезани са интернетом постали су отворени за повезивање корисника и дељење садржаја између њих. У складу са тим, другу фазу у развоју интернета (тј. фазу у којој су корисници добили могућност да се кроз дељење садржаја на интернету повезују) називамо друштвени веб (енгл. *Social Web*) или Веб 2.0.

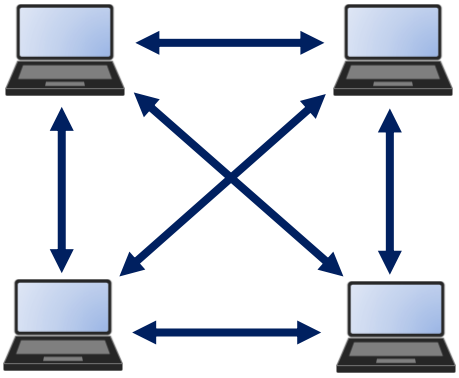
Средином друге деценије 21. века отпочела је трећа фаза у развоју интернета. Ову фазу одликују два паралелна процеса. Први процес представља преоријентацију интернет стандарда у правцу оспособљавања машина да разумеју интернет садржај, као и развоја програмерских алата којима ће се машина оспособити за наведени задатак. Овај процес за циљ има повећање употребне моћи интернета за корисника кроз побољшање квалитета резултата интернет претраге, идентификацију садржаја који би потенцијално био интересантан за корисника, олакшано дељење знања међу корисницима и слично. Међутим, имплементирање овог процеса у постојећем веб окружењу повезано је са централизованим прикупљањем података. Последично, постоји бојазност корисника за приватност њихових података. Томе додатно доприноси и чињеница да кориснички оријентисане платформе функционишу по принципу посредника. То је илустровано сликом 2. Наиме, корисник са свог рачунара даје налог централном серверу да садржај који је припремио (поруку, слику, или неку другу објаву) подели са неким другим корисником (или корисницима), након чега сервер извршава задату наредбу. Све објаве, и подаци повезани са њима, остаће сачувани на серверу. Иако се ове платформе чине бесплатним, оне за своје услуге од корисника траже право на располагање подацима прикупљеним од њих. Међутим, корпорације често не поступају савесно са подацима својих интернет корисника. Сведоци смо бројних скандала повезаним са злоупотребом приватних корисничких података од стране популарних корисничких платформи. Ово је условило потребу за паралелним развојем још једног процеса, а то је процес децентрализације. Процес децентрализације захтева да се, уместо путем сервера, рачунари повезују директно један са другим. На тај начин избегава се посредничка улога сервера и сви подаци остају на умреженим рачунарима. Децентрализовано организована мрежа, тј. мрежа који омогућава директно повезивање рачунара без сервера као посредника, назива се „друг другу“ мрежа (енгл. *peer to peer* или скраћено *P2P*). Пример „друг другу“ мреже приказан је на слици **3**. Узевши све наведено у обзир, трећу етапу у развоју интернета називамо семантичким децентрализованим вебом (енгл. *Semantic decentralized web*) или краће Вебом 3.0.

Слика 2: Приказ централизоване мреже



Извор: приказ аутора

Слика 3: Приказ децентрализоване мреже



Извор: приказ аутора

2.2 Појава криптографије

Избацивање посредника из структуре мреже може деловати као довољан услов за обезбеђивање приватност података, јер ће они остати на умреженим рачунарима. Међутим, приватност ће и даље бити угрожена, јер ће дељени садржај бити доступан свим рачунарима на мрежи. Овај проблем се у литератури назива проблем поверљивости (енгл. *confidentiality*). Осим тога, другим корисницима са мреже биће дозвољено да постављени садржај модификују што нас доводи до проблема интегритета (енгл. *integrity*). Коначно, у децентрализованим мрежама појавиће се и проблем идентификације (енгл. *authentication*). Наиме, корисник у сваком тренутку мора са сигурношћу да зна коме доставља садржај који жели да подели, као и од кога прима садржај који су други желели да поделе са њим. У централизованим мрежама, сервер као посредник са лакоћом решава све наведене проблеме. Дистрибутивна улога сервера гарантује да ће дељени садржај бити достављен само оним корисницима мреже којима је намењен. Поред тога, сервер је гарант да ће садржај постављен преко њега бити заштићен од манипулација других корисника. Коначно, сервер мотри на све кориснике мреже и у сваком тренутку може да их идентификује. Тиме се гарантује да су подаци о примаоцима и пошиљаоцима исправни.

Поменути проблеми претили су да угрозе опстанак идеје о децентрализованим мрежама. Ипак, решење за све њих нађено је кроз концепт криптографије (енгл. *cryptography*). Реч је о посебном поступку шифрирања и дешифровања дељеног садржаја уз помоћ јединствених кључа. У шифрирању, кључеви представљају одређени улазни податак на бази којег је могуће шифрирати или дешифровати одређени садржај. У зависности од броја кључева који су нам потребни за дешифровање или шифровање садржаја, разликујемо два приступа у криптографији – симетричну и асиметричну криптографију. У наставку разматрамо сваку од њих.

Према концепту симетричне криптографије, када год корисник отпочне комуникацију са новим уређајем на мрежи, оба уређаја добијају по један нови симетричан кључ. Ови кључеви су јединствени за посматрани пар корисника, и познати су само њима. Према томе, када један од корисника пожели да подели одређени садржај са неким другим корисником, искористиће

њихов заједнички кључ за шифрирање садржаја пре него га постави на мрежу. Иако је шифровани садржај доступан свим уређајима на мрежи, корисници којима он није намењен неће моћи да га дешифрују јер немају одговарајући кључ за то. Према томе, дељени садржај остаће непознат за кориснике којима није намењен. Са друге стране, једини ко ће моћи да дешифрује дељени садржај јесте корисник којем је тај садржај намењен, јер само он поседује одговарајући кључ. Мана овог приступа огледа се у чињеници да комуникација са сваким новим уређајем захтева генерисање и чување једног новог кључа. Последично, уколико је на мрежи умрежено N рачунара, постојаће $N \cdot (N-1)$ различитих парова симетричних кључева додељених сваком пару уређаја посебно. Овакав начин шифровања није ефикасан јер захтева памћење великог броја кључева. Ипак, овај приступ шифрирању је оправдан само уколико је реч о малој мрежи, те број симетричних парова кључева није велики. Приметимо да су оваквим поступком шифрирања решена сва три проблема децентрализованих мрежа: садржај могу отворити само корисници којима је намењен (поседоваће кључ којим могу да га дешифрују), садржај се не може модификовати од нежељених корисника (јер немају одговарајући кључ) и пошиљалац и прималац се увек могу идентификовати (само њихови симетрични кључеви могу да дешифрују дати садржај).

Како би се избегло непотребно памћење великог броја симетричних кључева, прелази се на концепт асиметричне криптографије који обезбеђује употребу минималног броја кључева. Према овом концепту, сваки корисник на мрежи уместо N симетричних има два асиметрична кључа – приватни и јавни. Јавни кључ корисника познат је свим члановима мреже, док је приватни кључ познат само кориснику. Приватни и јавни кључ су асиметрични. То значи да ако је за шифровање садржаја искоришћен један од њих, други се мора употребити за дешифровање. С тим у вези, поверљивост се код асиметричне криптографије постиже на следећи начин. Корисник који жели да одређени садржај подели (у даљем тексту пошиљалац) са другим корисником (у даљем тексту прималац) садржај мора да шифрира два пута. Први пут за шифрирање користи свој приватни кључ. Затим тако шифрирани садржај поново шифрира јавним кључем примаоца. Да би се тако шифровани садржај отворио потребно је најпре употребити приватни кључ примаоца, а затим јавни кључ пошиљаоца. Информацију о приватном кључу примаоца има само прималац, те су проблеми очувања приватности и интегритета садржаја решени. Са друге стране, како је за дешифровање садржаја такође потребан јавни кључ пошиљаоца, идентитет истог се једнозначно може одредити те је и проблем идентификације решен. Због чињенице да асиметрични кључеви једнозначно одређују корисника на мрежи називамо их и дигиталним потписом или дигиталним отиском корисника (енгл. *digital signature or digital fingerprint*). Овакав приступ шифрирању далеко је ефикаснији, јер захтева да мрежа са N уређаја има $2N$ кључева (што је далеко мање од $N \cdot (N-1)$ кључева, колико је потребно код симетричне криптографије). Поред тога, сваки корисник мора да зна само свој приватни кључ, док су му јавни кључеви свих осталих корисника јавно доступни. Из тог разлога, овај приступ криптографији искоришћен је за изградњу система плаћања по којем крипто-валуте функционишу.

2.3 Систем децентрализованих финансија

Крипто-валуте настају из потребе да се створи средство плаћања које неће бити ни под чијом контролом. Да би се елиминисали посредници (држава или трећа лица) потребно је формирати децентрализовану мрежу путем које ће се плаћања одвијати. Оваква одлука условљава стварање другачијег финансијског система и повезана је са многим проблемима, али и

бенефитима, које разматрамо у наставку ове подсекције. Такође, у наставку уводимо и неке основне појмове важне за разумевање система по којем крипто-валуте функционишу.

2.3.1 Децентрализоване мреже и њихови корисници

Формирање децентрализоване мреже намеће питање складиштења базе података. Ово питање се логично намеће из чињенице да такву мрежу не опслужује један централни сервер. Поред меморијског простора за складиштење, сервер је обезбеђивао сигурност и интегритет базе, али и идентификацију свих корисника који учествују у комуникацији на мрежи. Постављањем базе на једном уређају са децентрализоване мреже поново бисмо добили један облик централизованог система (јер ће сви остали уређаји морати да комуницирају са уређајем на којем је база ускладиштена). Уз то, ризик губитка базе је велики, јер не постоје копије (резерве) исте. Коначно, власник уређаја домаћина базе био би у искушењу да се понаша хазардно, а чак ни претња избацивањем са мреже, контрола од стране других корисника мреже, па ни периодична промена домаћина базе не би били гаранти да до тога неће доћи. Једноставно решење за овај проблем пронађено је имплементирањем дистрибутивних евиденција (енгл. *distributed ledger*). Реч је о сасвим новој парадигми складиштења и организовања базе података специфично дизајнираној за децентрализоване мреже. Један вид дистрибутивних евиденција је блокчеин систем. Зарад разумевања даљег излагања уведимо појам чворишта. Сваки уређај повезан на децентрализовану мрежу назива се чвориштем (енгл. *node*). Према блокчеин систему по једна копија базе података биће сачувана на свим чвориштима мреже заинтересованим за њено одржавање. Ова чворишта називамо потпуним или тешким чвориштима (енгл. *full or heavy nodes*), а њихове кориснике рударима крипто-валута или само крипто-рударима (енгл. *crypto-miners*). Свака промена у бази података мора бити прослеђена свим потпуним чвориштима на мрежи. Након тога, потпуна чворишта верификују дату промену и у складу са њом ажурирају своје копије базе. Како ни један корисник мреже не може да направи промену у бази без верификације крипто-рудара, и како велики број међусобно независних крипто-рудара има своје копије базе, овај систем се сматра високо сигурним. Целокупан процес функционисања блокчеина и рударења крипто-валута биће објашњен у наредној подсекцији.

Крипто-рудари нису једини актери у децентрализованој мрежи плаћања коју формирају крипто-валуте. Заправо, већину актера чине финални корисници. Реч је о корисницима који су чланови мреже зато што желе да користе крипто-валуте као средство плаћања и/или за трговање. Једним именом називамо их коинерима тј. кованичарима[11] (енгл. *coiner* – особа која кује или сакупља метални новац). У литератури се може пронаћи и термин ходлер (енгл. *hodl*) као синоним за реч коинер, иако ова два појма нису сасвим еквивалентна. Термин ходлер потиче од погрешно написане енглеске речи „држати“ (енгл. *hold*)[12]. Под овим појмом подразумевамо кориснике које држе крипто-валуте у свом поседу и не планирају нити да их потроше нити да са њима тргују. Овде треба нагласити да је реч о свесној и стратешкој одлуци да се прибављене крипто-валуте не употребљавају, јер се дугорочно очекује раст њихове вредности. Са друге стране, како постоје корисници који тренутно држе крипто валуте, али са њима слободно тргују и/или обављају трансакције, није до краја исправно користити реч

[11] Слободан превод аутора.

[12] Конкретно, реч потиче из твита постављеног од стране корисника „GameKyuubi“, који је веома брзо постао виралан у крипто-свету. Међутим, један број познаваоца прилика воли да користи овај појам као акроним за реченицу „држи се (крипто-валута) ако ти је живот мио“ (енгл. ***H**old **O**n for **D**ear **L**ife*).

ходлер као синоним за реч коинер. Коначно, уведимо и појам крипто-трговаца или крипто-трејдера (енгл. *crypto-traders*). Они представљају кориснике који користе крипто-валуте као финансијску активу, односно, за трговање на берзи.

Коинере као финалне кориснике не интересује одржавање базе података. Према томе, коинери на својим уређајима не морају да имају копију целокупне базе података. Уместо тога, довољно је да имају адекватан узорак из ње који је неопходан за обављање редовних активности. Из тог разлога се уређаји које коинери користе за приступање мрежи називају делимична или лагана чворишта (енгл. *partial or light nodes*). Међутим, заједничко за оба типа чворишта (потпуна и делимична) је то да се са њих мора приступити децентрализованој мрежи како би корисници обављали своје активности. Приступ децентрализованим мрежама омогућен је путем такозваних децентрализованих апликација (енгл. *dapps*)[13]. Ове апликације имају два различита појавна облика у зависности од популације корисника којима су намењени. Конкретно, крипто-рудари користе софтвере за рударење (енгл. *mining software*), док коинери користе електронске новчанике (енгл. *e-wallet*). У наставку ћемо се укратко осврнути на оба типа децентрализованих апликација. Софтвери за рударење су апликације које обезбеђују радно окружење за крипто-рударе. Ове апликације користе рачунарску снагу (енгл. *computing power*) како би рудариле валуте, дају могућност кориснику да прати структуру блокова и целог ланца (тј. базе података), обезбеђују извештавање о зарађеном новцу и слично. Са друге стране, електронски новчаници су апликације путем којих се могу електронским путем обављати финансијске активности. Коинери путем њих отварају своје рачуне, обављају трансакције и конвертују дигиталне крипто-валуте у фиат валуте[14]. Важно је истаћи да у већини случајева не постоје ексклузивне децентрализоване апликације за приступ мрежи плаћања неке крипто-валуте. Уместо тога, корисник сам бира које ће апликације користити према својим преференцијама. Такође, децентрализоване апликације користе крипторграфију засновану на асиметричним кључевима приликом сваке комуникације (трансакције) обављене на мрежи. На тај начин сви приватни подаци корисника остају заштићени, интегритет трансакције и стања на рачуну је неповређен, а учесници у трансакцијама се у сваком тренутку могу једнозначно одредити.

2.3.2 Децентрализоване финансије

Финансијски систем који је изграђен на бази децентрализоване мреже потпомогнуте дистрибутивном евиденцијом и заштићене криптографијом назива се децентрализованим финансијама или ДиФај-ем (енгл. *DeFi*). У оваквом систему финансијске активности обављају се директно између учесника, без присуства посредника. Последично, настаће и сва ограничења које посредовање са собом доноси. За почетак, посредничке провизије више неће постојати, те ће се трансакциони трошкови значајно смањити. Надаље, корисник ни са ким не мора да дели своје податке. Безбедности система доприноси и чињеница да је теже хаковати блокчеин базу података (о чему дискутујемо у наредној подсекцији), него сервер појединачних банака. Додатно, време потребно за извршење трансакција ће се скратити, што овакав

[13] Неким читаоцима ће вероватно бити познате децентрализоване апликације које се користе у свакодневном животу, попут: *BitMessage, BitTorrent, Tor, Popcorn* и слично.

[14] Поред електронских новчаника, постоје и специјализовани терминали путем којих се обавља конверзија дигиталне имовине (крипто-валутама) у фиат новац (физички или преко банковног рачуна) и обрнуто. Ове терминале називао крипто-мењачницама (енгл. *crypto-exchanges*). Неке од најпознатијих су: *Binancecoin1, Coinbaseexchange* и *KuCoin.*

финансијски систем чини ефикаснијим од традиционалног. Такође, нестаће и ограничења у погледу максималног дозвољеног износа по трансакцији, те велике трансакције више неће морати да се најављују. Уједно, проблем не извршених трансакција због превазилажења дневног трансакционог лимита више неће постојати. Са друге стране, децентрализовани финансијски систем обезбеђује одређене погодности и на глобалном нивоу. Пословање са земљама или деловима земаља у којима финансијски сектор није присутан или није развијен, биће олакшано, јер се за обављање трансакција захтева само приступ интернету. Поред тога, нема додатних трошкова за међународне трансакције, јер се сви учесници третирају подједнако. Најзад, ствара се простор за употребу крипто-валута као светских валута. На овај начин валута ни једне земље неће бити у повлашћеном положају и отвара се пут ка стварању праведнијег света.

Ипак, нису сви аспекти децентрализације финансијског система позитивни. Пре свега, монетарна политика државе, као средство за утицање на ниво привредне активности у земљи, ће ослабити или чак нестати. Неки познаваоци прилика ово виде као позитивну страну децентрализованих финансија, док је неки доживљавају као негативну. Присталице децентрализације истичу да се на овај начин степен државних мешања у привредне токове смањује, и повећава се ослонац на тржиште. Са друге стране, скептици сматрају да се привреда без државне интервенције излаже већем ризику, што представља опасност по стабилност целокупног привредног система. Друга замерка коју скептици упућују односи се на изостанак државне регулативе. Без адекватне регулативе, ревизије и контроле простор за манипулације, злоупотребе и преваре се повећава. Неки истакнути стручњаци, попут Ворена Бафета (*Warren Buffett*), упозоравају да ће овакав систем убрзо постати параван за прање новца и обављање финансијских трансакција повезаних са илегалним активности. Они истичу да без државне контроле овакав развој догађаја нико неће моћи да спречи. Следећи сет замерки односи се на крипто-валуте које децентрализовани финансијски систем пропагира као средство плаћања. Већина крипто-валуте нема фундаменталну или интризичну вредност (о томе дискутујемо у одељку **2.5**), а ни државну гаранцију (попут фиат валуте), те је инвестирање у њих изузетно ризично. Последично, њихова вредност је предоминантно изложена очекивањима појединаца на тржишту. Уједно, вредности крипто-валута изразито су нестабилне, што се огледа и кроз изразито високу волатилност њихових приноса. Тачније, њихова волатилност је већа од волатилности других финансијских актива о чему сведоче бројни радови као што је истакнуто у прегледу литературе. Коначно, процес рударења и одржавања децентрализоване мреже повезани су са великом потрошњом електричне енергије и нерационалним трошењем рачунарске опреме. Ова питања су изразито актуална у данашње време с обзиром на светску енергетску кризу и феномен глобалног загревања, а на њих се још увек тражи одговор.

2.4 Блокчејн технологија и рударење крипто-валута

Према Шерману (*Sherman*) и сарадницима (2019) блокчеин систем за вођење базе података се први пут предлаже као идеја 1982 у дискусији „Рачунарски системи успостављени, одржавани и подржавани од стране међусобно неповерљивих група“ коју је водио истакнути амерички информатичар Дејвид Чом. Ипак, прва практична имплементација овог система уследила је тек 26 година касније. Иноватор познат под псеудонимом Сатоши Накамото (*Satoshi Nakamoto*) је 2008. године, у жељи да створи прву децентрализовану валуту, унапредио и успоставио први блокчеин систем. Од тог дана па до данас систем се непрестано развија, а његова популарност не јењава. Како је блокчеин систем база података виталан за

функционисање крипто-валута и одвијање процеса рударења, у наставку ове подсекције детаљније ћемо се осврнути на важне аспекте његовог функционисања. На овај начин, читалац тезе стећи ће потпунију слику о крипто-валутама. У наставку најпре разматрамо основне градивне елементе блокчеин система, а затим се осврћемо на његову архитектуру и функционисање.

2.4.1 Хеш функције

Блокчеин је облик децентрализоване базе података. Као што је истакнуто у претходној подсекцији, таква база података је лоцирана на свим пуним чвориштима мреже. Како би лични подаци коинера остали приватни, потребно је осигурати да крипто-рудари не могу да их прочитају из базе података. Из тог разлога у блокчеин систем уводе се хеш функције (енгл. *hash*). Реч је о функцији која произвољан улазни податак претвара у хексадецималну вредност фиксне дужине. На тај начин сви записи у бази података постају кодирани. Сви крипто-рудари имаће увид само у кодирани запис, док ће само корисник моћи да зна шта стоји иза њих. Хеш функција има још једну важну улогу у блокчеин систему. Она се користи у процесу верификације блокова, односно, у процесу рударења (о чему детаљно говоримо у одељку **2.4.4**).

Да би хеш функција могла да се користи у блокчеин систему она мора да поседује 6 важних особина. У наставку укратко разматрамо сваку од њих:

1. **Одређеност** (енгл. *Deterministic*): Хеш функција мора да се понаша као математичка функција. Другим речима, сваки задати улазни податак мора имати свој хексадецимални запис. Поред тога, тај хексадецимални запис мора бити детерминистичан. То значи да сваки пут када унесемо исти улазни податак, морамо добити исти хексадецимални запис.

2. **Стандардизованост**: Сваки улаз, без обзира на његову меморијску величину или облик, мора да буде пресликан у хексадецималну вредност фиксне дужине (рецимо, 64 карактера). Тиме се постиже да сви хексадецимални кодови буду стандардизовани, чиме се олакшава даљи рад са њима. Последица стандардизованости је фиксан меморијски простор који ће заузимати сваки хексадецимални код (рецимо, 252 бита). То нас доводи до друге предности стандардизације – уштеда меморијског простора (сви улазни подаци произвољне меморијске величине претварају у кодове фиксне, углавном мање, меморијске величине).

3. **Брзина**: Да би коинер брзо обавио своје трансакције (обављене трансакције се морају регистровати и стање у бази након трансакције се мора ажурирати), и да би крипто-рудари брзо обављали верификацију блокова, потребно је да и сама хеш функција ради брзо. Такође, веома је важно да функција улазни податак брзо прекодира у хексадецимални код без обзира на његову (меморијску) величину или облик. На тај начин повећава се ефикасност целог блокчеин система.

4. **Отпорност на инверз** (енгл. *inverse resistance*): Хеш функција мора бити дефинисана на такав начин да је немогуће конструисати њену инверзну функцију. Функција инверзна хеш функцији би хексадецимални код могла да претвори назад у оригинални улазни податак, чиме би безбедност података корисника била угрожена. Ово је згодно место да истакнемо разлику између хешовања и шифрирања. Шифрирање подразумева да се шифровани садржај на одређени начин може дешифровати. Шифрирање користимо када желимо да још неко поред нас разуме шифровани садржај, али не желимо да ту могућност имају сви они који дођу у контакт са шифрованим садржајем. Са друге стране, хешовање подразумева да се кодирани садржај ни на који начин не може декодирати. Једини начин да се одгонетне улазни податак који је произвео одређени хеш код је да се он случајно погоди методом покушаја и грешака. Последично, хешовање користимо када не желимо да ико други, сем нас, разуме кодирани садржај. На тај начин лични подаци корисника сачувани у бази података у виду хексадецималног кода биће разумљиви само њему, али не и осталим учесницима са мреже.

5. **Случајност**: Хеш функција мора да обезбеди да слични улазни подаци дају потпуно различите хексидецималне кодове. Ова особина је од изузетног значаја за подизање безбедности базе. Захваљујући овој особини поступак проналажења улазног податка који стоји иза одређеног хексидецималног кода методом покушаја и грешака уз сирову снагу (енгл. *brute force*) изузетно је отежан. Отуда и назив „хеш“ за ову функцију (преведено на српски глагол *to hash* значи смућкати, направити папазјанију, помешати на чудан начин). Ово се постиже укључивањем псеудо-случајних алгоритама у дефиницију хеш функције.

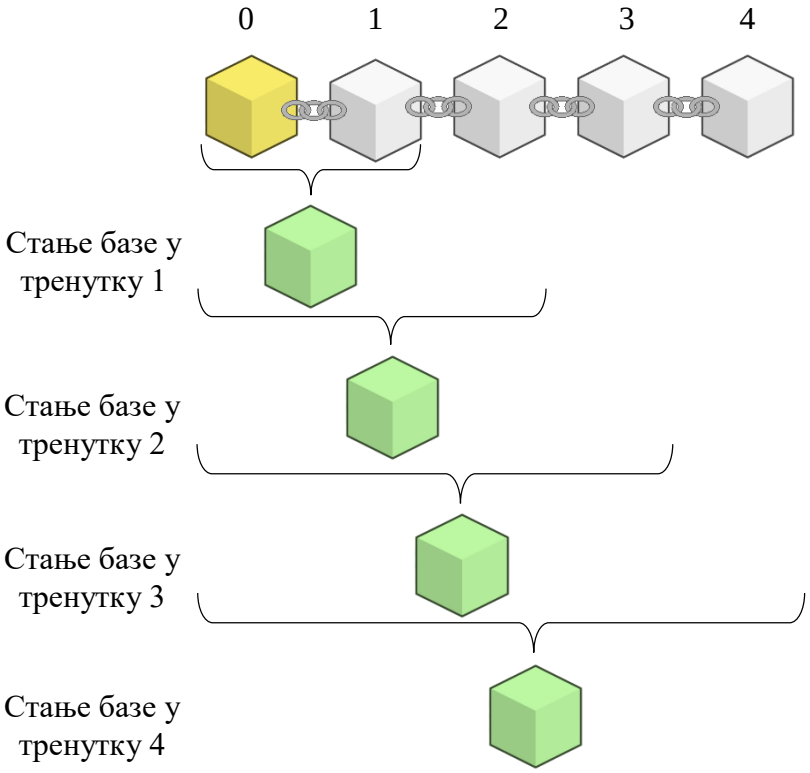
6. **Одсуство колизије** (**гранична инјективност**): Ситуацију у којој два или више улазна података имају исти хексадецимални код назива се колизија. Хеш функција мора бити дефинисана тако да сваки улазни податак има јединствени хексидецимални код. Ова особина у математици се назива инјективност. Међутим, због елемената псеудо-случајности ова особина може бити нарушена. Ипак, ако се одабере довољно дугачак стандард за хексидецималан код вероватноћа појаве колизије тежиће ка нули. Зато у случају хеш функција можемо говорити о граничној инјективности. Овде наилазимо на треидоф. Избор довољно великог стандарда смањиће вероватноћу колизије и повећаће сигурност података (експоненцијално ће се повећати број покушаја који хакер треба пробати како би дошао до личних података корисника). Са друге стране, већи стандард захтеваће већи меморијски простор за складиштење података у бази и успорава процес кодирања што ће умањити ефикасност система. У време писања ове тезе, опште прихваћени стандард величине хексидецималног кода био је или 252 или 512 бита.

Неке од најпознатијих и најкоришћенијих хеш функција у пракси су: функција за сажимање поруке (енгл. *message-digest function, MD*), безбедни алгоритам за хешовање (енгл. *secure hashing algorithm, SHA*), и кечак (енгл. *Keccak*, што је назив Индијско-Индонежанског ритуалног плеса пореклом са Балија). Уз назив хешинг функције обично се придодаје меморијска величина кода који она производи (нпр.: *SHA252* или *Keccak512*).

2.4.2 Од блока до ланца

Блокчеин је децентрализована база података организована у виду меморијских блокова повезаних у ланац. Отуда долази назив блокчеин. Преведеног са енглеског блокчеин значи блоковски ланац[15] тј. ланац блокова (енгл. *blockchain – chain of blocks*). Оваква организација базе података је неопходна зарад подизања нивоа безбедности података на децентрализованој мрежи. Наиме, како велики број појединаца има приступ комплетној бази, постоји опасност да ће неко од њих злоупотребити своју уређивачку моћ да направи промене у бази које му одговарају (тј. да се понаша хазардно). Међутим, организовањем базе у виду повезаних меморијских блокова могућности за хазардно понашање се значајно могу смањити, о чему дискутујемо у одељку **2.4.3**. У наставку овог одељка дискутоваћемо архитектуру блокчеин система, и структуру блокова.

Слика 4: Илустрација блокчеин система



Извор: приказ аутора

Први блок у ланцу представља почетно стање базе и назива се генерички блок (енгл. *generic block*). У блокчеин систему промене се не праве директно у почетној бази података, као што је то случај код традиционално организованих база података. Уместо тога, све промене се евидентирају као засебан меморијски блок. Последично, тек заједничким увидом у генерички блок и сва његова ажурирања (тј. у меморијске блокове који долазе после њега) моћи ћемо да сагледамо тренутно стање базе. Примера ради, други блок по реду представља прво

[15] Слободан превод аутора.

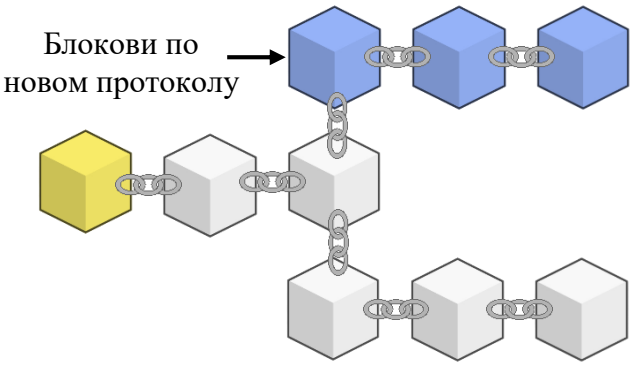
ажурирање полазне базе података, а заједно са генеричким блоком даће нам увид у стање базе након првог ажурирања. Сваки наредни блок ажурира претходно стање базе. Блокови се на тај начин повезују у хронолошки ланац према времену свог настанка. Према томе, да бисмо имали увид у тренутно стање базе података потребно је да сагледамо све блокове заједно као целину. Једна илустрација блокчеин система дата је сликом **4**.

Сваки блок у себи садржи податке о обављеним трансакцијама са крипто-валутама на бази којих се ажурира претходно стање базе. Поменути подаци обухватају податке о трансакторима и исходу трансакције. Како би се заштитила приватност корисника, сви подаци су кодирани хеш функцијом. Пре него што се нови блок формира, подаци о свим обављеним трансакцијама се привремено чувају у меморијском простору који се назива базен трансакција (енгл. *transaction pool*). Када количина прикупљених података у базену трансакција достигне ниво од око 1MB подаци се групишу у блок. Након тога се свим потпуним чвориштима на мрежи упућује захтев за ажурирање њихове копије базе додавањем новог блока. До ажурирања ће доћи уколико блок успешно прође верификацију у процесу рударења. О овој процедури детаљније говоримо у одељку **2.4.4.**

Поред података о трансакцијама, сваки блок у себи садржи и мета податке који се једним именом називају заглављем блока (енгл. *block's header*). У заглављу блока налазе се информације које ће једнозначно одредити сваки блок:

- **Временска ознака** (енгл. *timestamp*): За сваки блок морамо знати прецизно време генерисања. Како би се избегле намерне или случајне грешке у навођењу времена креирања, ова информација ће се аутоматски забележити приликом генерисања блока. Последично, приликом сваке промене блока остаје временски траг. На овај начин мрежа је додатно заштићена од могућег хазардног понашања њених корисника.

Слика 5: Утицај промене протокола или верзије постојећег протокола на ланац блокова



Извор: приказ аутора

- **Верзија пратећих протокола** која је коришћена приликом генерисања блока: Протоколи су програми који раде у позадини система и омогућавају његово несметано функционисање. Неки примери познатих протокола су: хајперлеџер (енгл. Hyperledger), кворум (енгл. Quorum),

корда (енгл. Corda), итд. Као и сви други програми, и протоколи се временом побољшавају и ажурирају. Како верзија протокола може имати утицај на процес креирања блокова, битно је да тачно знамо која верзија је коришћена приликом његове израде. Ситуација до које долази услед промене верзије постојећег протокола називамо меком виљушком (енгл. *soft fork*). Са друге стране, ситуацију у којој се стари протокол замењује новим називамо тврдом виљушком (енгл. *hard fork*). Назив „виљушка“ последица је чињенице да се након оваквих промена обично воде два паралелна ланца (ланац по старој верзији или старом протоколу и ланац по новој верзији или новом протоколу). У том случају, добијени ланац подсећаће на виљушку (што је илустровано сликом **5**).

- **Једнократни део** или **нонс** (енгл. *nounce* – потрошно, нешто што је за једнократну употребу) и **циљана вредност** (енгл. *target*): Реч је о концептима који заједно чине доказ о раду (енгл. *proof of work*). Њима ћемо се детаљно бавити у секцији **2.4.4**. За сада довољно је истаћи да он представљају доказ да је блок прошао кроз валидацију у процесу рударења.

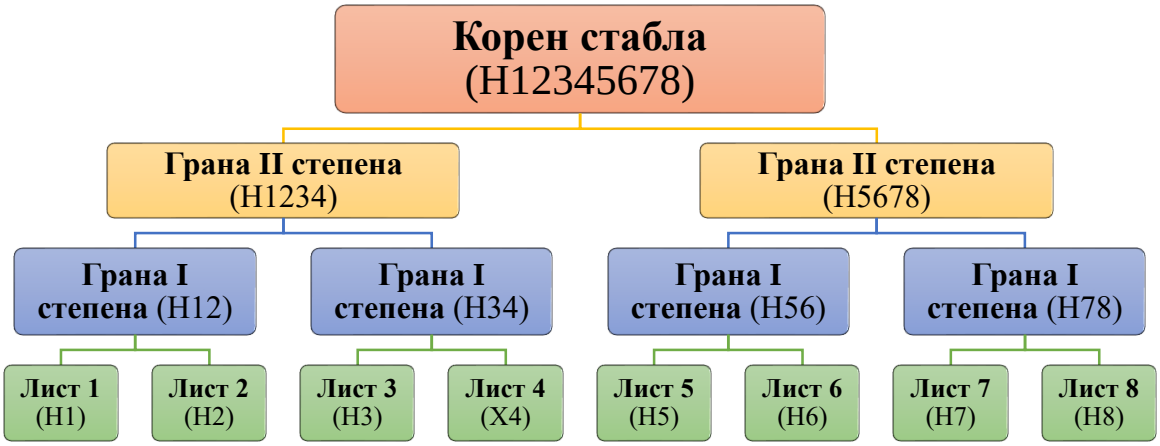
- **Агрегатни хексидецимални код** свих трансакција обухваћених блоком (тзв. **корен Меркеловог стабла трансакција**): Да бисмо такав код генерисали, потребан нам је алгоритам назван хеш стабло или Меркелово стабло[16] (енгл. *Merkle tree*). Меркелово стабло је структура у облику бинарног стабла која служи за агрегирање појединачних хексидецималних кодова у један агрегатни хексидецимални код. Листови (крајеви) овог стабла су хексидецимални кодови појединачних трансакција. Последично, листова има онолико колико и трансакција. Уколико је број листова паран број чији су сви фактори парни бројеви стабло је симетрично. У том случају процес агрегирања је изразито једноставан. Наиме, свака два[17] суседна листа агрегираће се у гране првог степена, након чега ће се сваке две суседне гране првог степена агрегирати у гране другог степена, и тако редом. Другим речима, процес агрегирања се наставља и на вишим нивоима стабла све док не дођемо до једног заједничког хексидецималног кода за цело стабло. Тај код називамо Меркеловим кореном или кореном хеш стабла (енгл. *Merkle root*). Под агрегирањем подразумевамо примену хеш функције на спојени хексидецимални код две суседне трансакције. На овај начин од два хексидецимална кода добијамо један. Процес агрегирања за случај симетричног стабла изграђеног за блок који има 8 трансакција (фактори броја 8 су парни бројеви: 2, 2 и 2) илустрован је на слици **6**. Са друге стране, уколико је број листова непаран број или паран број чији је бар један фактор непаран број поступак агрегације се мало модификује, јер је стабло асиметрично. У том случају кроз процес агрегације, пре или касније, наићи ће се на непаран број грана (или листова) на неком нивоу стабла. Да би се процес бинарног агрегирања наставио, потребно је да се стабло на спорном нивоу учини симетричним. То се постиже дуплирањем гране која нема свог пара у стаблу. Након тога агрегирање се несметано наставља на уобичајен начин. Уколико се непаран број грана појави на још неком нивоу стабла, проблем ће се решити на исти начин (дуплирањем гране која нема свог пара). Слика **7** илуструје случај асиметричног стабла на примеру блока са 10 трансакција (фактори броја 10 су 2 и 5, те како је један од њих непаран број, појавиће се проблем асиметричности). На крају, треба истаћи да увођење агрегатног

[16] Ралф Меркел (*Ralph Merkle*) је чувени информатичар и иноватор који је патентирао алгоритам хеш стабла. Из тог разлога се ово стабло често у литератури назива по њему Меркелово стабло. Осим хеш стабла, Меркел је створио концепт криптографије са асиметричним кључевима дискутован у одељку **2.2.**

[17] Како се сваке две трансакције групишемо у грану, Меркелово стабло је бинарно. Да је број трансакција потребних за агрегирања три, стабло би било триномно.

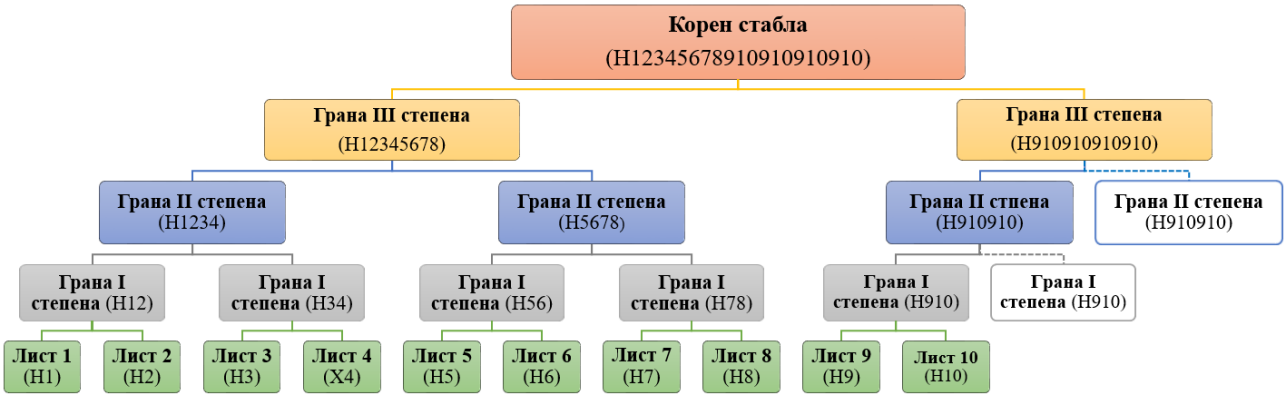
хексидецималног кода свих трансакција у заглавље блока игра кључну улогу за одржавање безбедности целокупног блокчеин система (о чему говоримо у одељку **2.4.3**).

Слика 6: Пример Меркеловог стабла симетричног облика (на свим нивоима стабла у заградама се налазе ознаке за хексидецималне кодове)



Извор: приказ аутора

Слика 7: Пример Меркеловог стабла асиметричног облика (на свим нивоима стабла у заградама се налазе ознаке за хексидецималне кодове, док су дуплиране гране означене провидним пољима и спојене су са стаблом испрекиданим линијама)



Извор: приказ аутора

- **Хеш код посматраног и претходног блока**: Као што свака трансакција има свој хексадецимални код, тако ће и сваки блок имати свој хексадецимални код. Хеш код блока представља агрегатни хексидецимални код свих претходно наведених података садржаних у заглављу блока (тј. временске ознаке, верзије протокола, нонса, циљане тежине, агрегатног хеш кода свих трансакција и хеш кода претходног блока). Агрегација хешованих вредности података из заглавља се такође обавља путем Меркеловог стабла на претходно описан начин. Хексидецимални код датог блока је последњи податак који се уноси у његово заглавље. Међутим, као што је истакнуто, заглавље блока мора садржати и хексидецимални код

претходног блока[18]. На тај начин сваки блок ће бити једнозначно повезан са блоком који му претходи. Повезивањем свака два суседна блока на овај начин формира се ланац блокова. Поред тога, уношење хеш кода претходног блока у заглавље посматраног блока кључно је за одржавање безбедности блокчеин система јер ће хеш код посматраног блока зависити од хеш кода претходног блока (о чему дискутујемо у наредном одељку).

2.4.3 Безбедност у блокчеин систему

Блокчеин систем је вишеструко заштићен, што је илустровано сликом **8**. Сви назначени аспекти безбедности функционишу заједно као целина, и систем чине изразито безбедним за све кориснике. У наставку укратко сублимирамо сваки аспекат заштите:

1. **Повезаност блокова**: Повезивање суседних блокова се постиже тако што се у заглавље сваког блока укључује хексидецимални код блока који му претходи. Ово чињеница има енорман утицај на безбедност базе података и управо је она разлог зашто се на децентрализованој мрежи користи блокчеин систем, а не класичан систем база података. Наиме, претпоставимо да је дошло до хакерског напада на базу тако да је у једном од блокова промењен или додат неки податак[19]. То ће резултовати другачијим скупом листова (тј. другачијим улазним хексидецималним кодовима) Меркеловог стабла приликом одређивања агрегатног хексидецималног кода свих трансакција. Самим тим резултат агрегирања (корен стабла) биће другачији. Последично, промениће се хексидецимални код блока, јер је дошло до промене једног од података из његовог заглавља. То ће за резултат имати прекид ланца, јер блок више неће бити повезан са наредним блоком. Наиме, наредни блок ће садржати стари хексидецимални код посматраног блока, док ће посматрани блок сада имати нови код. Како се ова два кода више не поклапају, ланац је компромитован и база неће функционисати. Да би база поново могла да функционише, хакер мора да промени податке из заглавља наредног блока. Последично, то ће променти хексидецимални код наредног блока, те он неће бити повезан са блоком који га следи. Према томе, да би база функционисала, хакер мора да прекодира не само нападнути (тј. хаковани) блок, него и све наредне блокове, што је исцрпан посао.

2. **Доказ о раду** (енгл. *proof of work*): Иако је прекодирање свих блокова почевши од нападнутог блока исцрпан посао, супер компјутер би могао да га изврши за коначно дуго време. Да би се тако нешто спречило, потребно је пролонгирати време рада. То се постиже постављањем захтева да сваки блок мора да прође кроз процес рударења (енгл. *mining*) и добије доказ о раду како би потврдио своју валидност (овај процес дискутујемо у наредном одељку). Целокупан процес траје око 10 минута по блоку, те брзо прекодирање свих блокова не би могао да изврши чак ни супер компјутер. Примера ради, уколико хакер нападне

[18] Изузетак од овог правила је генерички блок. Наиме, како генерички блок нема блок који му претходи, у заглављу генеричког блока у поље предвиђено за хексидецимални код претходног блока стоји вредност 0.

[19] Примера ради, хакер може бити постојећи корисник који жели да модификује неку стару трансакцију коју је обавио, тако да му се исплати већи новчани износ. Слично, хакер може бити нови корисник који додаје нову измишљену трансакцију у блок у којој велики број корисника мреже плаћа одређени износ на рачун хакера.

генерички блок Биткоина (те мора да прекодира апсолутно све блокове у његовом ланцу) целокупан период прекодирања би трајао нешто више од осам година[20]. Како напад не може да се изврши довољно брзо, неко од корисника ће већ приметити да база више не функционише и пријавити проблем надлежнима. Последично, напад ће бити разоткривен.

3. **Консензус** (енгл. *consensus*): Да би се блок сматрао валидним, поред доказа о раду, он мора имати и консензус рудара на мрежи. Крипто-рудар који је први дошао до доказа о раду, блок са доказом шаље свим осталим потпуним чвориштима на мрежи. Потпуна чворишта још једном прегледају послати блок и гласају о његовом прикључивању или не прикључивању у ланац. Консензус о валидности се постиже уколико је удео потпуних чворишта на мрежи која су гласала за прикључивање бар 51%. Уколико блок не добије консензус о валидности сматраће се неисправним, и неће моћи да се прикључи мрежи. Овде уводимо појам напад 51 (енгл. *51 attack*). До оваквог напада на мрежу долази уколико се 51% рудара на њој организује у криминалну групу са циљем да хакују базу да би остварили против правну имовинску корист. Како они чине 51% мреже, има их довољно да постигну задовољавајући консензус и наметну другима своје промене. Ипак, како је број потпуних чворишта на мрежи јако велики, и како су она дистрибуирана широм света и њих чине људи најразличитијих профила, мало је вероватно да ће доћи до криминалног удруживања неке групе њих. Са друге стране број рудара стално расте, те како би задржали већину, рудари који су се удружили у криминалну групу такође морају да проширују своје редове. Додатно, за такав напад потребно је издвојити пуно новца и компјутерских ресурса, али и адекватна организација и одржавање лојалности у групи, што није лако обезбедити. Коначно, и доказ о раду одвраћа потпуна чворишта од покушаја да се удруже у криминалну групу. Наиме, због доказа о раду напад не може да се изврши брзо, па је велика вероватноћа да ће неко од корисника приметити проблеме са базом и разоткрити напад. Према томе, вероватноћа да ће до оваквог напада доћи је минорна и опада са растом броја потпуних чворишта на мрежи.

4. **Копије базе**: Чињеница да база није сачувана само на једном месту, већ постоје њене комплетне копије на свим потпуним чвориштима представља важан аспект безбедности децентрализоване мреже. Иако је хакер хаковао једну копију базе, он мора да убеди кориснике на свим осталим потпуним чвориштима да у својим копијама базе прихвате промене које је он направио. За то ће му бити потребни консензус и доказ о раду. Постојање више од једне копије базе даје основ за функционисање претходно наведена два стуба сигурности базе на децентрализованој мрежи. Поред тога, постојање копија базе има још једну важну улогу. Уколико дође до губитка дела или целе базе услед више силе на једном чвориште, увек постоји копија на другом чвориште. Такав луксуз не постоји код централизованих база података.

5. **Временска ознака**: Сваки блок у заглављу садржи временску ознаку тренутка генерисања. Временска ознака се аутоматски мења приликом сваке модификације блока. Уколико се у неком од старих блокова из ланца нешто буде модификовало, време генерисања из заглавља тог блока биће млађе од времена блокова који га следе. То ће бити јасан индикатор свим крипто-рударима на мрежи да је у посматраном старом блоку неко извршио накнадне

[20] Процењено на бази тренутне величине блокчеина Биткоина. У време писања овог рада она је износила приближно 440GB.

модификације и да је ланац компромитован. Према томе, чак и да дође до удруживања 51% крипто-рудара у хакерски напад путем којег ће модификовати податке блокова у своју корист, иза њих ће остати временски траг промене. На бази тога остали рудари моћи ће да препознају да је до манипулација дошло. Према томе, временска ознака може се искористити као доказ приликом пријаве напада надлежним органима.

6. **Трошкови**: Процес рударења повезан је са високим трошковима. То није случајно, јер су управо тако високи трошкови један од основа за доказивање валидности рударења блока. Наиме, високи трошкови одвратиће хакере од покушаја да нападну ланац. Уколико се напад разоткрије, блок који је хакер израдио биће одбијен. Енергија и компјутерски ресурси ће бити узалуд утрошена, а хакер остаје без профита. Насупрот хакерима, легитимни рудари ће бити спремни да се изложе тако високим трошковима, јер знају да блок који буду изрударили неће бити одбијен те могу да профитирају од провизије. Последично, рударима који су спремни да се изложе овако високим трошковима можемо веровати. Што се више рудара прикључи потрази за доказом о раду, то смо сигурнији да је реч о валидном блоку. Из тог разлога уводимо показатељ сигурности мреже назван хеш стопа (енгл. *hash rate*). Хеш стопа је број покушаја да се проналажење доказ о раду у секунди на целој децентрализованој мрежи. Што је она већа, то је мрежа сигурнија.

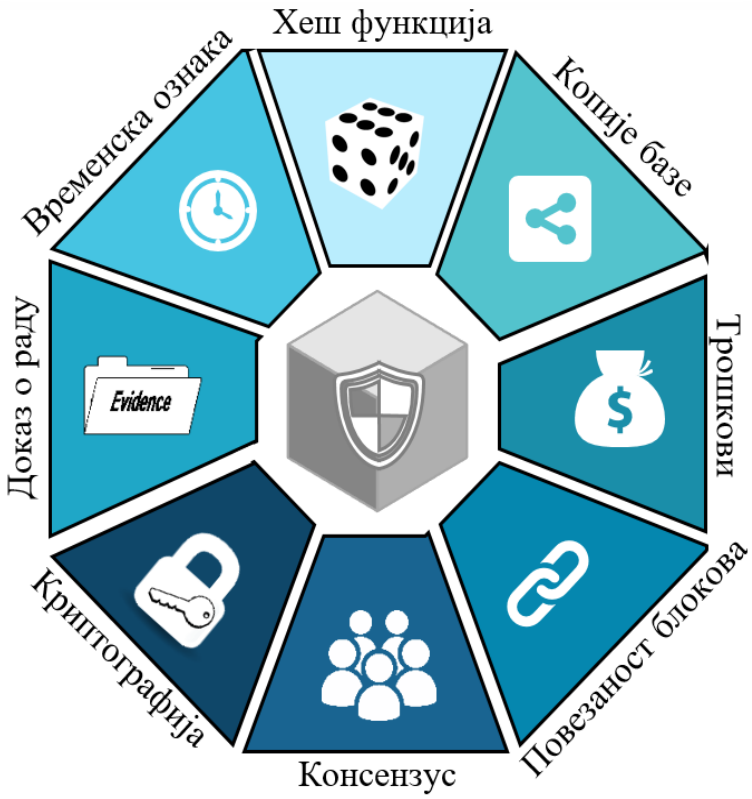
7. **Криптографија**: Као што је већ истакнуто криптографија осигурава да се комуникација између корисника на децентрализованој мрежи одвија директно, без могућности уплитања и увида других корисника. На тај начин корисници имају сигурност да нико неће имати могућност модификовања или увида у њихових трансакција и личних података. Такође, уз помоћ криптографије сваки учесник се једнозначно може идентификовати. Према томе, уколико неки од учесника предузме одређене недозвољене радње (покуша да хакује базу), увек се једнозначно може одредити о коме је реч путем његових кључева.

8. **Хешовање**: Да би подаци корисника, који су сачувани у бази, остали тајни кодирани су хеш функцијом. На тај начин, нико од рудара не може да има увид у личне податке осталих корисника мреже. Поред тога хеш функција штити интегритет блокова и саме базе. Уколико дође до било какве модификације, захваљујући хеш функцији доћи ће до драстичне промене кода блока и самим тим до компромитовања ланца. Чак и да је реч о минималној модификацији (примера ради, да се износ неке трансакције повећа за неку малу вредност) због особина хеш функције хексидецимални код блока ће се драстично променити, те ће неповезаност бити очигледна. Ова особина важна је за функционисање верификације на бази доказа о раду о чему дискутујемо у одељку **2.4.4**.

Да би сви наведени сегменти безбедности могли ефикасно да обављају своје улоге, неопходно је да на мрежи постоји довољно велики број потпуних и делимичних чворишта. Уколико је њихов број мали, лако може доћи до удруживања једног броја рудара у неки облик криминалне организације чија ће превара проћи испод радара осталих рудара и малобројних коинера. Са друге стране, уколико је број корисника велики, рудари ће се теже организовати зарад напада на базу, док ће у случају напада неко од корисника сигурно приметити нефункционалност базе. Према томе, потребно је радити на поверењу и обезбеђивању адекватних награда како би се

велики број корисника и привукао. Ипак, некада ни то није довољно, с обзиром на то да се у данашње време готово све може лажирати. Добар пример за то је Ванкоин (енгл. *OneCoin*) пирамидална шема из нама суседне Бугарске која је од корисника широм света украла милионе долара. Компанија је успела да обмане своје кориснике да се трансакције обављају на децентрализованој мрежи по блокчеин систему на којем ради велики број рудара, иако ништа од тога није заиста постојало. Новац прикупљен од нових корисника коришћен је за исплате старим корисницима, све док власници шеме нису прикупили довољно новца и нестали са њим. Лондон Тајмс је описао овај догађај као једну од највећих превара у историји света.[21]

***Слика 8**:* Сегменти безбедности блокчеин система



Извор: приказ аутора

2.4.4 Рударење и доказ о раду

Да би се доказала валидност блока, он мора да прође кроз поступак добијања доказа о раду. Овај поступак се још назива и рударењем валута (енгл. *crypto-currency mining*), а његов циљ је да се пролонгира време израде сваког блока. Једном када количина трансакција у базену трансакција достигне меморијску тежину од око 1MB, потребно их је сместити у један блок и додати их бази. То је задатак крипто-рудара. Да би рудари били плаћени за обављени посао, они датом скупу трансакција додају још једну нову трансакцију која представља његову провизију. Додата трансакција обухвата две операције. Прво операција је стварање нове

[21] Бартлет Џејми (*Bartlett Jamie*) (2019)

количине дате крипто-валуте у оптицају (тј. „рударење новца“[22]). Друга операција представља исплату изрударено новца рудару који је први обезбедио доказ о раду. Последично, новостворена количина управо је једнака провизији која треба да му буде исплаћена. Ова провизија се утврђује према сету правила специфичном за сваку крипто-валуту. Приликом постављању тих правила треба водити рачуна о томе да провизија мора бити подстицајна за рударе и не сме да произведе инфлацију (јер се провизија исплаћује стварањем нових монетарних јединица дате крипто-валуте, чиме се повећава количина дате валуте у оптицају).

Након што дода своју провизију у блок, рудар одређује агрегатни хексидецимални код свих трансакција путем Меркеловог стабла. Добијени агрегатни хексидецимални код се затим користи за добијање доказа о раду. Да би добио доказ о раду, рудар мора да пронађе једнократни део (тзв. нонс) чијим се додавањем на крај агрегатног хексидецималног кода свих трансакција добија запис чија вредност хеш функције мања од циљане вредности. У наставку ћемо објаснити да је реч о веома једноставној идеји. За почетак треба истаћи да је сваки хексидецимални код (тј. излазни резултат хеш функције) ништа друго до одређени број записан у хексидецималном, уместо у децималном, запису. Примера ради, број 1.000 се у хексидецималном запису пише као 3E8[23]. Да би се добио доказ о раду, потребно је наћи хексидецимални код који представља број мањи од неког задатог броја. Задати број се назива циљаном вредношћу или само циљем (енгл. *target*). Циљана вредност одређује се тако да се обезбеди да просечно време генерисања блока буде око 10 минута и ажурира се на сваких 2016 блокова. То се постиже уз помоћ следеће формуле[24]:

$$target_t = \frac{target_{t-1} \cdot AT2016}{10} \qquad (1)$$

где је:

- $target_t$: нова циљана вредност која ће се користити за добијање доказа о раду за наредних 2016 блокова;

- $target_{t-1}$: стара циљана вредност која је коришћена за добијање доказа о раду претходних 2016 блокова;

- $AT2016$: просечно време генерисања за претходних 2016 блокова.

Како број рудара непрестано расте, а просечно време генерисања блока мора да остане око 10 минута, постављени циљ се мора стално отежавати. У супротном, растом броја рудара расте и број покушај да се дође до решења, па ће се решење брже наћи. За мерење тежине користимо показатељ тежина циља (енгл. *target difficulty*), који пореди циљану вредност коришћену за генерисање генеричког блока ($target_0$) и циљану вредност коришћену за генерисање посматраног блока. Историјско кретање тежине циља у случају Биткоина приказано је сликом **9**, док је формула за њено израчунавање дата изразом испод (Накамото, 2008.):

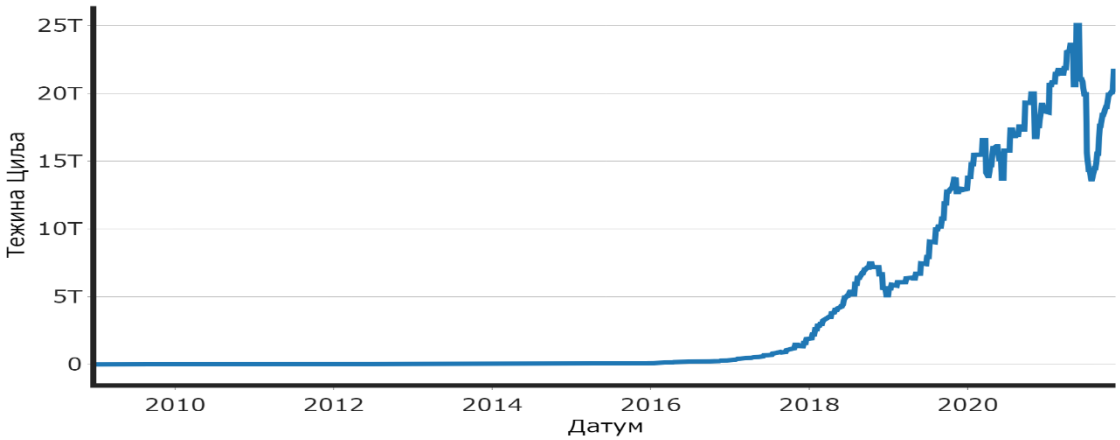
[22] У време када се новац ковао од племенитих метала, да би се нове монетарне јединице издале било је потребно изрударити додатну количину племенитог метала. По аналогији са овим историјских чињеницама, процес добијања доказа о раду у којем се стварају нове крипто-валуте добио је назив рударење.

[23] Уколико бисмо желели да запис овог броја има *N* карактера, само бисмо додали онолико нула на почетак овог броја, колико је потребно да укупан број карактера буде *N*. Примера ради, да смо желели да број 1.000 буде записан у облику хексидецималног кода са 4 карактера, написали бисмо: 03E8.

[24] У формули делимо бројилац са 10, како бисмо обезбедили да просечно време генерисања буде око 10 минута.

$$difficulty_t = \frac{target_0}{target_t} \quad (2)$$

***Слика 9**:* Историјско кретање тежине циља за Биткоин у периоду 2009-2022



Извор: *blockchain.com* (приступљено: 15.11.2022.)

Вратимо се полазном проблему налажења хексидецималног кода који представља вредност мању од циљане. Никада се неће десити да агрегатни хексидецимални код свих трансакција буде вредност мања од циљане вредности. Последично, добијени код је потребно модификовати како би се поновном применом хеш функције на модификовани код добио нови код који ће представљати вредност мању од циљане. Модификација се одвија тако што се на крај агрегатног хексидецималног кода свих трансакција дода неки случајан стринг[25]. Додати стринг назива се једнократни део или нонс, јер се користи само привремено зарад добијања доказа о раду и нема никакво посебно значење. Како је хеш функција резистентна на инверз, није могуће априори одредити како нонс треба да изгледа да би резултат био мањи од циљане вредности. Последично, одговарајући нонс се мора наћи методом покушаја и грешака што може да потраје. Рудар који први успе да нађе одговарајући нонс, произвешће доказ о раду и његов блок биће додат ланцу. Последично, само ће он бити исплаћен за извршено рударење.

Рачунари који имају јаке процесорске (*CPU*) и графичке (*GPU*) јединице моћи ће да изврше већи број покушаја у јединици времена. Због тога може деловати да ће само рудари са најбољом опремом зарађивати у овом такмичењу, што рударење чини неправичним. Међутим, то није случај. Ако мало боље погледамо овај проблем, видећемо да задатак није идентичан за све рударе. Наиме, сваки рудар у блок додаје своју верзију финалне трансакције (у којој се ново-креиране монетарне јединице исплаћују баш њему). Последично, агрегатни хексидецимални код свих трансакција ће бити другачији код свих рудара. Према томе, сваком рудару ће требати различити број покушаја да нађе нонс који одговара његовом агрегатном хексидецималном коду. Према томе, чак и рудар са лошом опремом може први да нађе одговарајући нонс и добије доказ о раду. Употреба високо квалитетне опреме може само да

[25] Текстуални запис у програмирању се назива стринг.

повећа вероватноћу победе (тј. конкурентност), али не и да је гарантује. Из тог разлога се процес рударења сматра фер такмичењем за све рударе.

Рударење није јефтино. Зарад подизање конкурентности, многи рудари ангажују велику количину рачунарске опреме која је сама по себи скупа. Поред трошкова њеног прибављања, ту су и повећани трошкови замене и трошкови отклањања кварова (опрема која се интензивно користи брже се хаба), али и високи трошкови електричне енергије (рад велике количине рачунарске опреме и опреме за расхлађивање троши пуно струје). Из тог разлога многи предлажу редефинисање поступка валидације блокова. Неки од њих су: доказ о улагању (енгл. *proof of stake*), доказ о депоновању (енгл. *proof of deposit*), доказ о протеклом времену (енгл. *proof of elapsed time*), доказ о капацитету (енгл. *proof of capacity*) и сл. Међутим, њих овде нећемо разматрати. Са друге стране, високи трошкови пружају индиректну гаранцију валидности блокова и служе за одвраћање хакера, као што је дискутовано у претходном одељку.

Занимљиво је истаћи да доказ о раду није концепт развијен специјално за блокчеин систем. Заправо, овај концепт је настао још 1997. године као предлог за решавање проблема спам мејлова, односно проблема нежељене електронске поште. Идеја је да се трошак електричне енергије слање мејла подигне. Обичан корисник не би приметио повећање трошка, док то не би био случај са корисником који шаље на хиљаде или више мејлова. Пошиљалац спамова не би био спреман на тако високе трошкове, те не би био у стању да обезбеди доказ да је у генерисању свог мејла утрошио довољну количину електричне енергије (тј. довољну количину рачунарског рада). Базирано на присуству или одсуству доказа о раду, прималац мејла би лако могао да диференцира између нежељене и жељене поште.

2.5 Крипто-Валуте

Крипто-валуте су облик дигиталног новца заштићеног криптографијом који је под контролом свих корисника децентрализоване мрежи на којој се трансакције са њим одвијају. Већ из саме дефиниције крипто-валута можемо идентификовати кључне разлике између њих и фиат новца. Пре свега, крипто-валуте су присутне само као запис у рачунару, тј. у дигиталном, али не и у физичком облику. За разлику од фиат новца, трансакције са крипто-валутама не обављају се уз посредовање треће стране (финансијских институција, у првом реду банака), већ директно између корисника мреже. За директно повезивање корисника заслужене су децентрализоване мреже и криптографска заштита (о чему је дискутовано у подсекцијама **2.2** и **2.3**). Овакав систем чува приватност корисника, док паралелно смањује трансакционе трошкове и његову зависност од трећих лица, што су слабости централизованих система. Децентрализованост значи и одсуство регулаторне контроле и државне гаранције, што није случај са фиат валутама. Коначно, корисници сами стварају нове монетарне јединице крипто-валута кроз процесе рударења или ковања, док би у централизованом систему о томе одлучивала држава.

Из досадашњег излагања јасно је да је архитектура система по којем крипто-валуте функционишу комплексна и пажљиво организована, али да то не утиче на једноставност са којом крајњи корисници обављају своје трансакције. Слика **10** илуструје процедуру обављања једне трансакције са крипто-валутама. Трансактор који жели да исплати одређени износ другом кориснику шаље захтев за обављање трансакције. Захтев се шифрира и хешује, а затим се доставља у базен трансакција. Трансакције из базена се рударе у блокове који затим пролазе кроз процес верификације. Уколико се постигне консензус о валидности блока, трансакција се сматра валидном и након дешифровања друга страна у трансакцији добија исплаћени износ. Приметимо да све кораке обављају корисници мреже, без уплитања треће привилеговане стране. Корисници који одржавају мрежу немају увид у детаље трансакција који су познати само укљученим странама. На тај начин обезбеђно је да све релевантне информације остају у кругу трансактора, иако је сама трансакција обављена на транспарентан начин и прошла је валидацију осталих корисника мреже.

Слика 10: Процедура обављања трансакције на децентрализованој мрежи

Како крипто-валуте немају државну гаранцију, природно је запитати се шта одређује њихову вредност. Према томе да ли крипто-валуте имају неку интринзичну вредност или им вредност даје поверење (односно очекивања) корисника, разликоваћемо три групе крипто-валута: кованице, новчиће или коини (енгл. *coins*), жетоне или токене (енгл. *tokens*) и стабилне кованице или стабилне новчиће (енгл. *stablecoins*). У наставку разматрамо сваку од наведених група. Поред њихових кључних особина, размотрићемо и процес одређивања њихове почетне вредности.

2.5.1 Кованице/Новчићи/Коини

Коини су крипто-валуте у ужем смислу речи. Они представљају дигитални новац у циркулацији на децентрализованој мрежи чија је примарна функција да служе као средство плаћања. Из тог разлога познаваоци прилика коине посматрају као крипто-валутама прве генерације. Коини своју вредност изводе из поверења корисника, будући да иза себе немају никакво покриће или гаранцију којима би своју вредност поткрепили. Рачуноводствено говорећи, њихова вредност је чист гудвил (енгл. *goodwill*). Последично, они немају праву интринзичну вредност. Из тог разлога велика пажња се поклања управо изградњи поверења код корисника валуте. Да би поверење постојало корисници морају да се осећају сигурним и да имају контролу над својим трансакцијама. Зато се природно намеће обавеза да коини морају да имају сопствену децентрализовану мрежу и сопствену дистрибутивну евиденцију (тј. блокчеин). Међутим, развој једне овакве мреже траје дуго и захтева заједнички рад бројних стручњака. По креирању треба непрестано радити на одржавању постојеће инфраструктуре целог система, али и на ширењу свести о датом коину и његовом маркетингу. Из тог разлога стварање потпуно новог коина је тежак и озбиљан задатак. Ипак, једном када се поверење

изгради и коин уђе у животе људи он постаје изузетно ликвидна финансијска актива као и сам фиат новац.

Једно од важних питања приликом покретања новог коина је његова почетна вредност. По узору на акционарска предузећа која пролазе кроз процес иницијалне јавне понуде акција (енгл. *initial public offering – IPO*) у којем одређују своју почетну вредност, и коини организују иницијалну јавну понуду (енгл. *initial coin offering – ICO*). У оквиру овог процеса потребно је направити бизнис план датог коина, одредити иницијалну понуду (количину коина која ће бити у оптицају), представити јавности изграђену децентрализовану мрежу, систем изабран за дистрибутивну евиденцију и протоколе по којима они раде, затим анализирати трошкове изградње мреже, дати доказе о сигурности, указати на подршку из привреде (уколико има оних за које се унапред зна да ће прихватити дати коин као средство плаћања) и слично. Иако не постоји никаква законска регулатива која захтева обелодањивање ових података пре емитовања коина (као што је то случај код акционарских друштава), покретачи коина се сами опредељују на тај корак како би изградили поверење у јавном мњењу. Након обелодањивања процењује се расположење заинтересоване јавности. Добијене процене се комбинују са досадашњим и будућим трошковима (уколико је бизнис планом предвиђено да се накнадно обаве још неке развојне операције читавог система) и одређује се цена по којој ће коин бити понуђен на првој продаји. Након прве продаје, цену ће одређивати тржишне силе које ће примарно бити руковођење ставовима појединаца.

2.5.2 Токени/Жетони

Токени су крипто-валуте у ширем смислу речи. За њих се може рећи да представљају другу генерацију крипто-валута. Примарна улога токена је да својим корисницима пруже неку функционалност, односно да обезбеде неку употребну вредност. За разлику од коина, секундарна сврха токена је да служе као средство плаћања. Функционалност се обезбеђује тако што се кориснику или пружа нека финансијска актива или нека услуга, што се постиже кроз паметне уговоре (енгл. *smart contract*). Паметни уговори су кондиционални програми уграђени у дистрибутивну евиденцију који уговорним странама дозвољавају приступ уговорним правима тек након што изврше своје уговорне обавезе. Главна предност оваквих уговора је то што се елиминише потреба за посредницима. Примера ради, замислимо уговор којим се продавац обавезује да испоручити одређену робу купцу након што роба буде плаћена. Паметни уговор ће са рачуна купца резервисати износ неопходан за плаћање продавцу. Међутим, тај износ неће бити пребачен на рачун продавца све док купац не добије поручену робу (што морају потврдити и купац и курирска служба). Према томе, купац добија робу након што су средства резервисана, а продавац добија резервисана средства након што испоручи обећану робу. Како програм преузима улогу посредника, елиминише се потреба за увођењем треће стране (попут клириншких кућа у овом примеру) као посредника у уговору. Из изложеног је јасно да токени имају интринзичну вредност, јер иза њих стоји неки паметни уговор. Последично, токени не настају у процесу рударења, већ у процесу ковања или минтовања

(енгл. *minting*)[26]. Под овим процесом подразумевамо стварањем нових паметних уговора и издавање токена повезаних са њима. Из тог разлога, издаваоци токена не морају да развију сопствену децентрализовану мрежу и дистрибутивну евиденцију. Уместо тога, довољно је да закупе простор на некој од постојећих мрежа и користе њихову дистрибутивну евиденцију за своје потребе. Многе децентрализоване мреже (попут Итиријумове или Бајнансове мреже) имају спремне шаблоне по којима друге фирме или предузетници могу да дизајнирају и емитују своје токене. Осим тога, многе мреже имају и своје огранке за тестирање које се називају тест мреже (енгл. *test network*). Тест мреже су делови децентрализоване мреже на којима корисници могу да испробају како раде паметни уговори по основу који желе да емитују своје токене. Захваљујући овим погодностима убрзава се процес емитовања новог токена на тржишту и смањују се повезани трошкови. Из истакнутих чињеница јасно је да је токен неупоредиво лакше емитовати у односу на коин. Ипак, токени су обично мање ликвидни од коина. Њихова ликвидност у првом реду зависи од њихове намене. Према класификацији швајцарске Управе за надзор финансијских тржишта (енгл. *Swiss Financial Market Supervisory Authority – FINMA*)[27] токени се према својој намени могу поделити у три категорија:.

- **Токени вредностних папира** (енгл. *Security tokens* или *Asset tokens*) су токени емитовани на бази паметних уговора који се понашају као хартије од вредности. Другим речима, корисник њиховом куповином де факто стиче право на неку хартију од вредности. Како је на овај начин могуће емитовати све три групе хартија од вредности (финансијске деривате, дужничке и власничке хартије од вредности) њихово емитовање подлеже регулативи комисије за хартије од вредности (енгл. *Securities and Exchange Commission* – SEC).

- **Услужни токени**[28] (енгл. *Utility tokens*) су токени емитовани на бази паметних уговора који нуде неку погодност, услугу или производ. Њихови власници могу добити право да користе одређени садржај на децентрализованим апликацијама, приступе одређеним ресурсима, учествују у лутријама, остварују попусте, и слично. Због специфичности своје намене обично су мање ликвидни.

- **Хибридни токени** (енгл. *Hybrid tokens*) су врста токена која по својој конструкцији представља комбинацију претходне две врсте токена или комбинацију неке од њих са коинима. Другим речима, токени који се понашају као хартије од вредности, а свом власнику нуде неку специјалну погодност представљали би хибридне токене. Слично томе, хибридним токенима ће се сматрати и они који су примарно емитовани да служе као средство плаћања, а својим власницима дају право на неку специјалну погодност или су конструисани као нека хартија од вредности. Најбољи пример су обмотани или враповани коини (енгл. *wrapped coins*). Реч је о токенима чији се паметни уговори понашају као размењиви фондови (енгл.

[26] Када рударите племените метале, никада унапред не можете да знате колику ћете количину изрудартити. Са друге стране, када кујете новац, увек можете да испланирате колико ћете новца исковати. Слично, како емитенд самостално унапред одређује колико ће паметних уговора створити, процес стварања токена назива се ковање.
[27] *FINMA* на коине гледа као на посебну врсту токена која постоји без паметних уговора. Како истичу, реч је о токенима који не дају никакво право свом власнику, већ се користе као средство плаћања.
[28] По овој класи токени (жетони) су добили своје име. Наиме, када би одређена трговинска радња хтела да пружи одређене погодности једној групи својих купаца емитовали би купоне или жетоне путем којих би приступ погодностима био омогућен.

exchange traded funds – ETF)[29] који прате кретање одређеног коина. На овај начин се корисницима неке децентрализоване мреже пружа могућност де факто трговања коином којем та мрежа није домицилна. Примера ради, обмотани Биткоин је токен са Итиријумове мреже који прати кретање Биткоина. Захваљујући обмотаном Биткоину корисник који жели да тргује и са Итиријумом и са Биткоином не мора да буде члан обе од њихових децентрализованих мрежа. Уместо тога довољно је да буде члан само Итиријумове мреже, а затим да тргује са Итиријумом и обмотаним Биткоином. Додатна погодност обмотаних коина је смањење трошкова трговања јер средства не морају да се преносе са мреже на мрежу. Осим тога, како провизије крипто-мењачница варирају од мреже до мреже, корисник може да изабере да тргује обмотаним коинима на оној мрежи чије су провизије мењачница најмање.

Иницијална јавна продаја токена вредносних папира (енгл. *security token offering*) је изразито захтеван процес јер је регулатори виде као иницијалну јавну продају хартија од вредности. Из тог разлога на њу се примењују сва правила прописана за иницијалну јавну понуду хартија од вредности. Исто важи и за емитовање хибридних токена, јер и њихови паметни уговори функционишу по принципу хартија од вредности. Са друге стране, за иницијалну јавну понуду услужних токена не постоји никаква законска регулатива, те се она може обавити по сопственом избору.

2.5.3 Стабилни новчићи/кованице

Стабилни новчићи су коини емитовани уз одређено покриће. Последично, стабилни новчићи имају интринзичну вредност и много су мање волатилни у односу на коине (отуда и њихов назив стабилни новчићи). Процес њиховог настајања се такође назива ковање[30], а не рударење. Основна намена стабилних новчића је да служе као средство плаћања, те и они, као и коини, морају имати сопствену децентрализовану мрежу и дистрибутивну евиденцију. Стабилни новчићи су обично ликвиднији од коина, јер се лако могу разменити за активу која служи као покриће. Заменом стабилног новчића за активу која служи као покриће, размењена количина стабилних новчића се повлачи из оптицаја док се утрошена актива поново не обезбеди. У зависности од активе која служи као покриће, стабилне новчиће можемо поделити на:

1. **Фиат покривене**: реч је о најчешћој врсти стабилних новчића. Њихово покриће је обезбеђено одређеном количином неке фиат валуте (нпр. амерички долар). Емитенд приликом издавања стабилног новчића овог типа поставља одређени курс по којем се стабилни новчић може разменити за фиат валуту којом је покривен (на пример, 1 стабилни новчић : 1 долар). Међутим, већина емитената фиат новац који служе као покриће држе у портфолијима хартија од вредности због временске вредности новца. Како вредност портфолија варира кроз време, стабилни новчић није увек савршено покривен обећаном количином новца. Из тог разлога,

[29] У конвенционалним финансијама, размењиви фондови се користе за праћење приноса неког берзанског индекса или стране активе којима се не може трговати на домаћем финансијском тржишту.

[30] За увођење нове монетарне јединице стабилног новчића захтева се априори обезбеђивање додатне количине активе која служи као покриће. Како се количина новца у оптицају увек може планирати, као када се новац кује, овај процес је такође назван ковањем.

главни задатак је адекватно управљати портфолиом који служи као покриће. Примери стабилних новчића из ове групе су: *XSGD* (покривен сингапурским доларима), *USDT* (покривен америчким доларима), *EURS* (покривен еврима) и сл.

2. **Робно покривене**: реч је о стабилним новчићима покривени материјалном активом којом се тргује на берзи. За сада се у слободном промету као покрића појављују нафта, нафтни деривати, непокретности и племенити метали. Велика предност ове врсте стабилних новчића је то што тровање њима замењује трговање активом која служи као њихов колатерал (покриће). Провизије за трговање стабилним новчићима су мање, а количине које су предмет трансакција нису ограничене (чиме се отвара пут за мале инвеститоре). Примери ове врсте стабилних новчића су: *DGX* (повезан са 1 грамом злата) и *SRC* (повезан са некретнинама).

3. **Крипто покривене**: реч је о стабилним новчићима који користе друге крипто-валуте или портфолије крипто-валута као своје покриће. Избор јако волатилне финансијске активе за покриће може деловати упитно. Ипак, ефекат диверсификације и обезбеђивање портфолија много веће вредности од вредности емитованих стабилних новчића ипак улива поверење у стабилност ове активе. Уколико би вредност покрића пала изнад неког унапред обећаног нивоа, емитенд би морао да депонује додатна средства, промени врсту покрића или укине дату валуту (њеним откупом за преостали износ покрића). Добар пример оваквог стабилног новчића је *DAI* који је покривен портфолиом крипто-валута који између осталог чине: *USDC*, *USDP*, *WBTC*, *ETH* и многе друге.

4. **Непокривене**: реч је о стабилним новчићима који формално немају никакво покриће. Ова категорије може деловати контрадикторно узевши у обзир шта су стабилни новчићи. Стабилност ових новчића остварује се кроз алгоритам (вештачку интелигенцију) за контролу њихове понуде. Алгоритам прати кретања тражње на тржишту и предузима адекватне кораке како би цена датог новчића остала стабилна. Постоје различити алгоритми, а овде кратко истичемо два могућа приступа. Према првом приступу алгоритам директно утиче на понуду стабилног новчића повећавајући је (емитовањем и продавањем нових монетарних јединица) или смањујући је (тј. откупом и повлачењем из оптицаја постојећих монетарних јединица). Други приступ подразумева свесно стварање арбитражних прилика које ће корисници мреже моћи да искористе. Мотивисани арбитражним профитом корисници ће својим појединачним акцијама повећавати или смањивати укупну понуду стабилних новчића и тиме одржати њихову стабилност. Адекватни примери за ову групу стабилних новчића су *AMPL* (код којег алгоритам на дневном нивоу врши корекције понуде) и *UST* (који је арбитражно оријентисан).

2.6 Одабране крипто-валуте

Сада када је читалац упознат са свим значајним аспектима крипто-валута, можемо се позабавити питањем њихове селекције. Кроз ово истраживање испитиваће се само утицај онлајн доступних текстова на коине (уз један специфичан изузетак). Таква одлука је донета као последица неколико чињеница. У одсуству фундаменталне вредности коина, очекивања

заинтересоване јавности преузимају примат у формирању цене. Последично, коини су најподложнији утицају ставова изнетих у јавно доступним онлајн чланцима. Одсуство фундаменталних фактора доприноси и томе да се утицај онлајн текстова може најчистије сагледати управо код коина у односу на било коју другу финансијску активу. Са друге стране, коини су познатији широј јавности од осталих типова крипто-валута. Пример ради, читалац је сасвим сигурно и пре читања овог рада имао прилике да чује о Биткоину и Итиријуму, што се вероватно не може рећи за већину токена. Последично, за њима влада веће интересовање и о њима се објављује много више чланака. Коначно, коини су основни тип крипто-валута, тј. крипт-валуте у ужем смислу. Анализом утицаја онлајн текстова на њих, сагледаћемо основне аспекте њиховог утицаја на крипто-валуте као целину. Имајући ово на уму, у наставку представљамо седам одабраних коина и један хибридни токен. Изузетак је направљен због изразите популарности коју је уживао изабрани хибридни токен у тренутку започињања истраживања. Међутим, без обзира на то, оваква одлука не угрожава концепт целог истраживања. Изабрани хибридни токен се понаша као крипто-валута, и осмишљен је на специфичан начин (о чему дискутујемо у наставку). Што се критеријума појединачног одабира тиче, избор сваке од крипто-валута је био руковођен искључиво њеном популарношћу у тренутку спровођења истраживања (у погледу броја објављених чланака). У наставку укратко представљамо сваку од њих.

2.6.1 ADA

ADA је крипто-валута (коин) који је развила децентрализована мрежа Кардано крајем 2017. године. Иако је на њиховом развоју био укључен огроман број стручњака, *ADA* и њен блокчеин развијани су преко две године. Захваљујући томе, Кардано мрежа се данас може похвалити чињеницом да представља једну од највећих и најсавременијих децентрализованих мрежа међу крипто-валутама. Назив коина, *ADA*, инспирисан је Августом Адом Кинг, првом женом програмером у историји човечанства. Од емитовања *ADA* коина, Кардано мрежа је поставила високу лествицу циљева које треба остварити. Амбициозни Кардано кроз *ADA* коин жели да створи трећу генерацију крипто-валута која ће покушати да реши све постојеће проблеме. У првом реду, мрежа је најавила иновације у погледу међусобног повезивања бројних децентрализованих мрежа, унапређењу процеса рударења, приближавању крипто-валута банкама и другим финансијским институцијама и у финансирању даљег развоја мреже. Дакле, јасно је да је реч о веома младој крипто-валути која пуно пажње поклања свом даљем развоју и иновацијама.

2.6.2 AVAX

AVAX је крипто-валута (коин) развијена од стране децентрализоване мреже Аваланч (срп. Лавина) 2020. године. Аваланч је мрежа која велику пажњу поклања ефикасности, те зато и не чуди што се сматра једном од најефикаснијих децентрализованих мрежа у крипто свету. Захваљујући својој конструкцији мрежа је способна да обрађује неограничен број трансакција у секунди, а време чекања корисника на финализовање трансакција од тренутка њиховог

подношења је испод две секунде. Аваланч је познат и по томе што води више од једног блокчеина и даје могућност својим корисницима да на њему покрећу нове блокчеинове. Мрежа је себи за циљ поставила да постане главна база за све који желе да издају своје паметне уговоре. Све операције[31] и трансакције на њој обављају се уз помоћ коина *AVAX*. Због свог специфичног лога, *AVAX* је међу својим корисницима познат и под надимком „црвени коин".

2.6.3 BTC

Реч је о крипто-валути коју не треба посебно представљати. Биткоин је родоначелник крипто-валута као класе финансијске активе. Према Ћупи (*Ciupa*) (2019), настао је 2008. године у јеку светске економске кризе – тачно 46 дана након краха познате банке „Лиман браћа" (енгл. *Lehman Brothers*). Због пионирских револуционарних промена које је Биткоин са собом донео (у погледу безбедности, децентрализованости, организације базе података и слично) брзо је стекао вишемилионску популарност широм планете. Захваљујући томе, вредност ове крипто-валуте забележила је експлозивни раст са кумулативним приносима који су оставили *S&P500* у прашини у периоду растућег тржишта које је уследило после кризе. Ипак, Биткоин је највише пажње скренуо на себе у периоду кризе изазване КОВИД-ом 19. У том периоду велики број људи се окреће инвестирању у дигиталну активу, а вредност Биткоина вртоглаво расте. Кулминација његове популарности резултирала је његовим увођењем за званично средство плаћања у Сан Салвладору 2021. године. Како се његова вредност мери десетинама хиљада долара, зарад задржавања малих инвеститора, Биткоин је подељен на ниже монетарне јединице које носе назив по псеудониму његовог креатора Сатошија. Биткоин данас представља најстарију, највећу (у погледу тржишне капитализације), најпопуларнију и највреднију крипто-валуту на глобалној сцени.

2.6.4 DOGE

Прича о Догкоину (скраћено *DOGE*) почела је као шала. Један од оснивача његове мреже, Џексон Палмер, 2013. године поставио је твит у виду рекламе (видео записа) у којем се препоручује инвестирање у фиктивну крипто-валуту Догкоин. У видеу је за маскоту ове валуте изабрао младунче јапанске расе паса Шиба Ину, по којем је валута и добила име. Твит је требао да буде шала на рачун крипто-валута као нестабилних инвестиција и растуће „инфлације" кућних љубимаца, алудирајући на то да ће власници кућних љубимаца желети да инвестирају у тако нешто. Шала је убрзо постала вирална, а интересовање за Догкоином се брзо пренело и у реални свет у којем се велики број људи распитивао како се може инвестирати у ову крипто-валуту. Запањен оваквом реакцијом, Палмер је решио да покрене Догкоин у жељи да покаже да је то крипто-валута у коју нико озбиљан не би инвестирао. Уз помоћ твитера Палмер је пронашао још присталица своје идеје и са њима је обавио цео развојни процес без превише труда. Међутим, ни ова шала није прошла по плану будући је Догкоин уживао огромну популарност и одржао се до дан-данас. Врхунац своје популарности Догкоин доживљава у

[31] Операције повезане са паметним уговорима и депоновањем улога зарад добијања доказа о улагању.

првој половини 2021. када се чувени предузетник и корпоративни магнат, Илон Маск, придружује његовој породици. Том пликом је за јако кратко време вредност овог коина скочила за преко 15 000% (видети слику **1**).

2.6.5 DOT

Полкадот (срп. *Полка тачкица*) или само *DOT* (срп. *тачка*) је крипто-валута (коин) коју је покренула Швајцарска фондација *Web3*. Свој специфичан назив дугује графичком приказу децентрализованих мрежа који подсећа на насумичне разбацане тачкице сличне онима које се појављују на Пољској народној ношњи за Полку. Иако је његова децентрализована мрежа у потпуности завршена тек крајем 2021. године након 7 година развоја, *DOT* је почео са радом још 2020. године. Мрежа ове крипто валуте препознатљива је према својим усмерењима да обезбеди мостове за лако повезивање са другим мрежама и дистрибутивним евиденцијама и највиши могући степен приватности за своје кориснике. Из тих разлога *DOT* и његова мрежа првенствено су интересантни девелоперима.

2.6.6 ETH

Још једна светски добро позната крипто-валута којој посебно представљање није потребно. Итиријум, етар или само *ETH* је друга највећа и најпопуларнија крипто-валута и највећи ривал Биткоина. Назив је добила по етру, петом елементу (вода, ватра, ваздух, земља и етар) из античко-средњовековних теорија, за који се веровало да испуњује свемир. Примера ради, у античкој Грчкој се веровало да је етар ваздух који дишу Богови. Из тог разлога једна од ознака за ову крипто-валуту је велико грчко слово Ξ. Са развојем његове мреже почело се још 2013. године, а коин је први пут емитован средином 2015. године. Иако је у почетку радио на бази доказа о раду, Итиријум се преоријентисао на доказ о улагању 2020. године (комплетна миграција је завршена 2022. године). Ова одлука допринела је енормном расту популарности овог приступа рударењу. Поред тога, Итиријум је заслужан за прве практичне имплементације паметних уговора и стварање крипто-валута друге генерације. Позицију мреже број један за емитовање нових паметних уговора задржао је до данас.

2.6.7 LUNA

LUNA је хибридни токен који је емитовала Тераформ лабораторија. Ова крипто-валута је настала као нус производ приликом емитовања Тераформ непокривеног стабилног новчића (*UST*). Замисао девелопера је била да *LUNA* буде токен који служи као средство плаћања, а чија вредност слободно флуктуира на тржишту. При томе, свака јединица *LUNA* токена издаје се уз паметни уговор по којем је могуће разменити је у било ком тренутку за *UST*. Захваљујући

томе, у случају значајнијих промена у тражњи за *UST*-ом, алгоритам постављен на мрежи ствара арбитражне могућности којима се одржава стабилност *UST*-а. Међутим, алгоритам брине искључиво о стабилности *UST*-а, али не и *LUNE* која је у потпуности препуштена тржишним силама. Иако су изградили сопствену децентрализовану мрежу, Тераформ лабораторија никада није покренула сопствени коин сматрајући да би на тај начин произвела непотребну конкуренцију за *LUNA* токен. *LUNA* је брзо стекла популарност и пронашла своје место на тржишту. Ипак, криза на истоку Европе скупо је коштала ову крипто-валуту будући да је њена вредност за свега месец дана стоструко смањена. Покушаји да се врати поверење у њу нису уродили плодом, па је *LUNA* убрзо била угашена[32].

2.6.8 SOL

Солана или *SOL* је крипто-валута (коин) која је почела са радом 2020. године. Са развојем њене мреже почело се још 2017. године, за шта је заслужан развојни тим из Украјине познат по томе што је први развио концепт доказа о историји (енгл. *proof-of-history*). Иницијално је планирано да овај коин носи назив *LOOM*. Међутим, како исти назив носи и мрежа коју су заједно лансирали Бајнанс, Итиријум и Трон, одлучено је да се назив промени у Солана (што преведено са Шпанског значи „*сунчев сјај*“). Мрежа на којој он функционише успела је да се издиференцира као изузетно јефтино и брзо место за обављање разних активности повезаних са блокчеином[33], што јој је донело велику популарност. Упркос томе, ова мрежа је повезана са бројним контроверзама. Још приликом иницијалне продаје, 48% коина завршило је у рукама инсајдера. Осим тога, мрежа је два пута на кратко престала да ради због техничких проблема. Без обзира на то, мрежа и даље ужива велику популарност међу својим корисницима.

[32] Уместо ње у септембру 2022. године емитована је нова крипто-валута, *LUNC*, али овога пута као коин.
[33] Између осталог, мрежа изнајмљује свој блокчеин за потребе онлајн игрица.

3. Методологија

Методолошко поглавље ће представити читаоцу све коришћене методолошке концепте, али и дати увид у дизајн истраживања. На тај начин ће се паралелно и олакшати праћење даљег излагања и обезбедити транспарентност и репродуктабилност истраживања. Сви концепти изложени у наставку ове секције биће дати по хронолошком редоследу (тј. оним редом којим се користе у истраживању). Из тог разлога, читалац ће најпре бити упознат са хронолошким током истраживања, а затим и са сваким концептом појединачно.

3.1 Хронолошки ток истраживања

Руковођен постављеним циљевима, аутор дисертације је истраживање организовао као троетапну процедуру. Етапе су структуриране тако да обезбеђују ефикасност спровођења истраживања, очувају интегритет његовог тока, задрже објективност у анализи резултата и прилагоде се потребама за тестирање предложених методолошких побољшања. У наставку ове секције биће изложене основне активности у свакој од етапа, са циљем да се читаоцу олакша праћење тока истраживања. Детаљан опис процедуре представљен је у раду Дамјановића и Дреновака (2023). Свакој од етапа појединачно поклоњено је много више пажње у емпиријском делу рада у којем су представљени резултати добијени у свакој од њих.

Прва етапа подразумева припрему истраживања и технички је најсложенија. У оквиру ње потребно је прикупити податке и припремити их за даље истраживање. Ова етапа се првенствено ослања на информатичке алате повезане са анализом текста. Поменуте алате испрограмирао је сам аутор специјално за потребе овог истраживања. Комплетан код је урађен у програмском језику Пајтон, познатом по својој флексибилности и применама у рударењу текста и машинском учењу. Први од алата задужен је за скреповање. Путем њега преузети су релевантни текстови и историјски подаци о кретању приноса одабраних крипто-валута. Како би се преузети текстови припремили за даље анализе они подлежу процесу рударења, зашта је задужен други испрограмирани алат. На овај начин се од текстова добијају вектори речи, који се потом трансформишу у математичке (нумеричке) векторе. Захваљујући томе сви текстови постају квантификоване величине. У овој етапи потребно је обезбедити још само један састојак неопходан за спровођење даљег истраживања. Реч је о обрачуну нивоа сентимента сваке речи, о чему дискутујемо у наредном поглављу ове тезе. Након овог корака бићемо у стању да измеримо сентимент сваког текста.

У другу етапу истраживања улазимо са обрачунатим показатељима добијеним из текстуалних вести. Циљ ове етапе је да се међу посматраним показатељима препознају они који се могу сматрати добрим предикторима приноса. То се постиже оцењивањем помоћног модела на бази којег се испитује значајност утицаја издвојених предиктора на приносе. На крају ове етапе из даље анализе биће одстрањени предиктори који не могу да објасне кретање приноса, док ће остали учествовати у изградњи финалног модела. У оквиру ове фазе анализирају се и економске импликације уочених веза и испитују прве две постављене хипотезе.

У трећој и финалној етапи анализираће се предиктивна моћ текстуалних вести. Предиктори идентификовани као значајни у претходној етапи формираће модел машинског учења из породице ансамбла. У те сврхе испрограмиран је алгоритам који за сваки дан посебно оцењује нови ансамбл регресија, а затим на бази текстова доступних у току тог дана, предвиђа приносе на сутрашњи дан. У овој етапи биће упоређене предикције добијене по оригиналној методологији и предикције добијене уз помоћ методолошких побољшања које заступа ова дисертација. На овај начин ће се проверити квалитет и употребна моћ предложених побољшања. Предложена побољшања су главни мотив да се истраживање бави појединачним крипто-валутама, а не њиховим портфолиом. На овај начин ћемо побољшања испитати на осам уместо на једном узорку, што закључке о квалитету побољшања чини независним од избора узорка. Поред тога, за овакву одлуку постоји још један мотив. Техничка анализа се традиционално бави појединачним финансијским инструментима или берзанским индексима. Последично, истраживање је комплетно спроведено у духу техничке анализе, а његови резултати демонстрираће корисност савремених аспеката техничке анализе за финансијска предвиђања.

Поред наведеног, дисертација ће спровести још две додатне анализе. У очекивању да резултати претходних процедура дају сигнале који указују на потенцијалну тржишну неефикасност биће спроведени и формални тестови слабе форме тржишне ефикасности. Осим тога, провериће се и да ли се ансамбл модел може побољшати увођењем још једног предиктора базираног на кластеризацији методом *K*-средњих вредности.

Сада кад је читалац начелно упознат са хронолошким током истраживања, детаљније ћемо се позабавити појединачним корацима у оквиру сваке од идентификованих етапа. До краја овог поглавља анализираће се методолошки аспекти који стоје иза њих.

3.2 Скреповање

Подаци су неизоставни део сваког истраживања. Последично, истраживања се типично започињу прикупљањем различитих података. Скреповање (енгл. *scraping*) је термин који означава различите алгоритме, методе и процедуре које се користе за прикупљање информација са целог интернета. С обзиром на то да је реч о изузетно младом концепту који је настао и развијао се пре свега неформалним путевима, терминологија повезана са њим није егзактна. Из тог разлога се у литератури могу појавити и следећи енглески називи за овај концепт: *web scraping, web data extraction, screen scraping* и *web harvesting*. Најчешће коришћени назив, скреповање (срп. стругање), инспирисан је идејом о насилном скидању мурала са зида, зашта је потребно стругање.

За скреповање је потребно израдити програм који симулира људско претраживање интернета у циљу прикупљања одређених информација са различитих веб локација. Предност овог

приступа је то што машина брже претражује и прикупља информације од човека, као и то што је могуће насилно преузети садржај који није намењен за преузимање. Међутим, та могућност преузимања садржаја без дозволе подигла је пуно контроверзи у вези са практичном применом скреповања. Иако су временом интернет садржаји постали заштићенији (динамички сајтови, проверавање да ли човек врши претрагу и сл.), и ресурси за скреповање су се паралелно развијали и постали много способнији. Још једна предност скреповања је и то што оно представља јефтин начин да се прикупи велика количина корисних информација. Упркос томе, бенефити скреповања доступни су само истраживачима са програмерским знањима, која још увек нису довољно распрострањена. Каоману овог приступа истичемо и то што га није могуће генерализовати. Наиме, није могуће израдити један програм који ће бити универзалан за преузимање произвољног садржаја са произвољних интернет страница. За сваку страницу и за сваки тип садржаја (текст, слике, линкове, мејл адресе, видео записе…) потребно је израдити посебан алгоритам, што може бити исцрпан посао. Без обзира на то, скреповање обећава да ће постати популаран и моћан начин прикупљања информација.

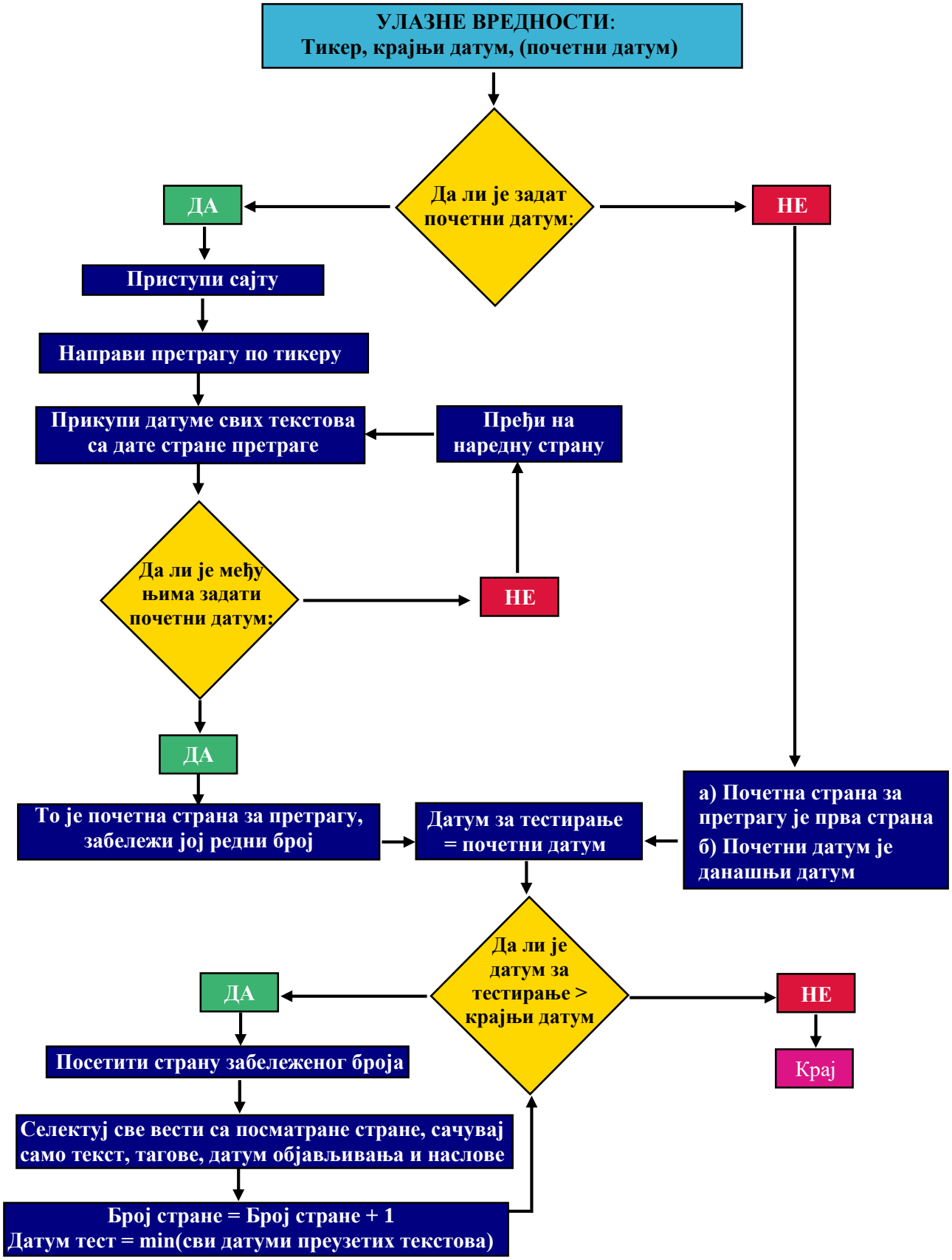
За потребе овог истраживања неопходне су две групе података – цене и онлајн чланци. За скреповање цена (нумерички податак) са портала *YahooFinance!* већ постоје развијена програмска решења[34]. Међутим, то није случај са прикупљањем текстуалних чланака, те је било неопходно израдити алгоритам за скреповање портала Крипто Њуз. У те сврхе аутор је израдио програм који сам приступа енглеској верзији сајта Крипто Њуз и преузима текстове чланака[35]. У наставку алгоритам ће укратко бити описан.

Зарад подизања ефикасности, кориснику алгоритма даје се могућност да спецификује списак крипто-валута чије онлајн чланке жели да преузме. Поред тога, корисник треба да спецификује временски период за који треба преузети текстове. Алгоритам започиње претрагом базе чланака портала Крипто Њуз у потрази за чланцима о крипто-валути коју корисник спецификује. Из резултата претраге алгоритам преузима само оне веб странице на којима се налазе текстови објављени у задатом истраживачком периоду. Када алгоритам пронађе све релевантне странице, са сваке од издвојених страница алгоритам извлачи само текст вести публикованих на њима (без слика, реклама, линкова ка другим страницама и сл.). Поред текстова, за сваки преузети чланак алгоритам бележи и датум и време његовог објављивања, придружене му кључне речи (енгл. *tags*) и његов наслов. Када су преузети сви текстови објављени у задатом временском интервалу, прва итерација је готова и прелази на наредну крипто-валуту. Алгоритам престаје са радом тек након што преузме текстове вести за све крипто-валуте које је корисник спецификовао. Преузети садржај корисник може да сачува у привременој меморији рачунара или да их извезе у екстерне фајлове. Приказ једне итерације, односно операција које алгоритам обавља за сваки задати тикер (тј. крипто-валуту) посебно дат је сликом **11**.

[34] Библиотека *YFinance*.

[35] Са протоком времена, портали мењају дизајн, архитектуру, организацију, мере заштите и сл. Из тог разлога и изграђени алгоритам за скреповање је потребно повремено ажурирати како би наставио са успешним функционисањем и у новим околностима. Алгоритам је завршен и употребљен 23.03.2022.

Слика 11: Приказ једне итерације алгоритма за скреповање вести (псеудо код)

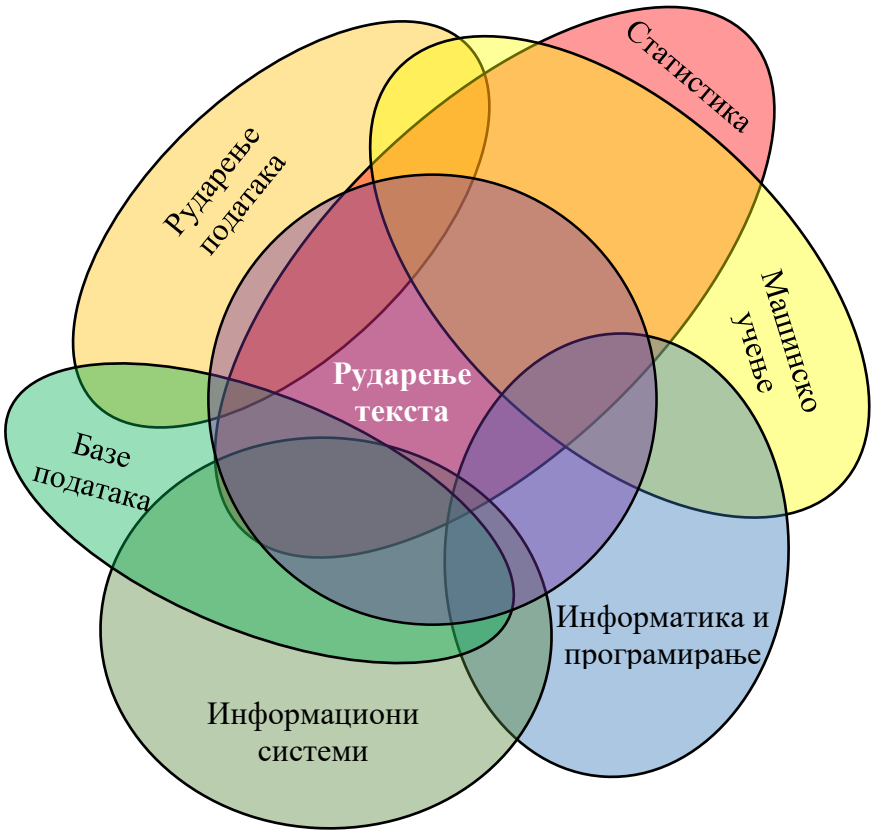


Извор: приказ аутора

3.3 Рударење текста

Захваљујући савременој технологији, данашње машине су способне да прикупљају и анализирају информације из разноврсних извора података који нам раније нису били доступни. Иако су текстуални записи увек били присутни у животу људи, због ограничених ресурса (пре свега људског рада и времена) могућности за њихову употребу у пракси су биле скромне. Оспособљавањем машине да процес разумевања и анализе текстова преузме на себе отклоњене су препреке за укључивање текстуалних података у процес моделирања. Тако се из науке о подацима (енгл. *data science*) развила нова научна дисциплина названа рударењем текста (енгл. *text mining*). Рударење текста је аутоматизовани процес структурирања текстова написаних на природном језику у сврхе њихове анализе како би се дошло до важних сазнања. Ова процедура представља комбинацију програмирања, статистике, лингвистике и машинског учења. Циљ ове секције је да читаоцу додатно приближи и објасни овај процес. Њено место, обухват, али и веза са другим научним дисциплинама илустровани су сликом **12**.

Слика 12: Место и обухват рударења текста као научне дисциплине



Извор: Miner и сарадници (2012)

Иако су текстуални записи јако интуитиван концепт људском бићу, машине у раду са њима наилазе на бројне проблеме који су последица њихове неструктурираности. Текстуални записи

могу бити форматирани на различите начине (форме писања, фонтови, специјални карактери, дужина текста…), у њима се могу појавити техничке и правописно-граматичке грешке, могу садржати скраћенице, стране речи и изразе, стручне термине, фразе, жаргон, некњижевни говор и сл. Проблеми ове врсте не постоје код конвенционалних података будући да су они обично високо структурирани (углавном су дати у облику табела које садрже нумеричке вредности). Да би машина могла да анализира садржај текстуалних података, она мора да се оспособи за разумевање природних језика. Из тог разлога, прва етапа у рударењу текста је процесирање природног језика (енгл. *natural language processing – NLP*). Циљ ове етапе је добијање структурираних података који се могу користити за даље анализе. Ово је најтежи и најважнији део посла, и он представља рударење текста у ужем смислу речи.

Као резултат прве етапе сви текстови биће претворени у векторе речи. Међутим, такви вектори и даље нису употребљиви за машинско учење и статистичко моделирање. Зато је потребно векторе речи мапирати у математичке (нумеричке) векторе. Ова процедура се назива „од текста до вектора“ (енгл. *text-to-vec* или скраћено *t2v*). Мапирање се обавља тако што се за сваки елемент вектора речи рачуна одређена врста фреквенције или неке статистике базиране на фреквенцијама. За потребе овог истраживања *t2v* је обављен коришћењем *TF-IDF*-а (ову статистику дискутујемо у наредном одељку).

Коначно, добијањем математичких вектора стечени су услови да се отпочне са другом етапом у рударењу текста. То је разлог зашто се процедура *t2v* сматра природним мостом између прве и друге етапе. У оквиру друге етапе на добијене математичке векторе потребно је применити одговарајући модел машинског учења или статистике како би се дошло до циљних сазнања. Због своје аналитичке природе, ова етапа се још назива и аналитиком текста (енгл. *Text Analytics*). Неки од циљева које је могуће остварити у овоје етапи су: класификација текстова, идентификовање кључних речи, мерење сентимента, проналажење информација и сл.

Рударење текста је релативно млада процедура за коју још увек нема пуно развијених стандарда. Последично, још увек нема пуно развијених комерцијалних софтвера[36] и готових програмских решења за ову процедуру и већини њих није могуће бесплатно приступити. Осим тога, како се текстови из различитих области могу драстично разликовати (по стилу писања и изражавања, по коришћеном језику и сл.), дизајн алгоритма за рударење текста је пожељно прилагодити њима. С тим у вези, за потребе ове докторске дисертације аутор је израдио свој програм за рударење текста. Алгоритам аутора ће бити укратко представљен у наставку ове секције. Међутим, модел машинског учења коришћен у другој етапи рударења текста биће представљен у одељку **3.8.1**. Чињеница да овај алгоритам хронолошки долази на ред у последњој етапи истраживања условила је да се њиме позабавимо касније.

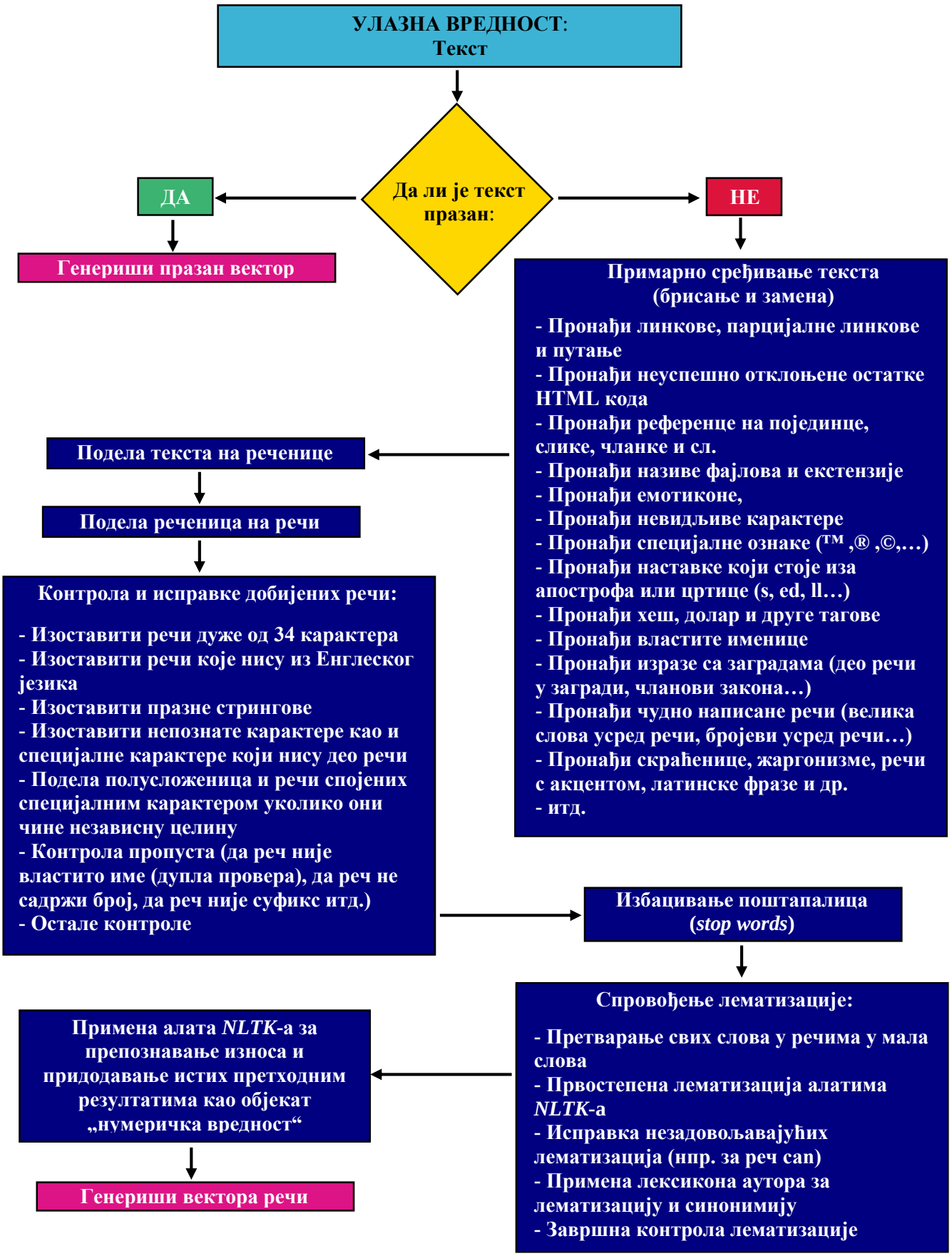
[36] У време писања овог рада неке од популарних комерцијалних софтвера израдиле су корпорације Амазон, Гугл, или *IBM*.

3.3.1 Алгоритам за рударење текста

Да би се текстуални запис претворио у вектор речи потребно је обавити низ различитих операција у циљу његове стандардизације. Извршавање ових операција је основни задатак алгоритма за рударење текста израђеног од стране аутора. Са жељом да изграђени алгоритам не представља црну кутију за читаоца ове дисертације, у наставку наводимо неке од операција које он обавља. Сажет приказ операција које алгоритам примењује на сваки текст посебно дат је и сликом **13** како би читалац лакше пратио даље излагање.

Најдужа фаза у раду алгоритма је сређивање текста. Пре свега алгоритам препознаје и отклања све елементе текста који немају никакву употребну вредност у анализи сентимента (што ће бити један од задатака друге етапе рударења текста у овом истраживању). Овде се пре свега мисли на следеће елементе текста: линкове, адресе, тагове, властита имена, датуме, неуспешно отклоњене остатке *HTML* кода приликом преузимања текста, емотиконе и сл. Алгоритам, такође, замењује скраћенице и акрониме са речима које стоје иза њих, како их машина не би третирала као различите појмове. Примера ради, скраћенице „*ака*“ и „*а.к.а*.“ треба третирати подједнако као израз „*also known as*“. Затим, алгоритам препознаје и отклања делове текста који нису написани на енглеском језику. Као што је већ истакнуто, јавност заинтересовану за крипто-валуте чине људи из свих крајева света. Из тог разлога њихове објаве, анализе, цитати, коментари и други садржаји који преносе чланци са онлајн портала могу бити написани на матерњем језику аутора или садржати делове написане на матерњем језику аутора. Међу најзаступљенијим страним језицима су: шпански, кинески, руски и португалски. Надаље, алгоритам је оспособљен за рад са текстуалним нумеричким вредностима. Алгоритам препознаје нумеричке вредности (тј. бројеве и цифре) записане као речи у тексту, разликује да ли нумеричка вредност представља новчани износ или количину неке друге величине, препознаје називе валута и јединица мере и слично. Ипак, најважнија операција коју алгоритам обавља је лематизација. Лематизација (енгл. *lemmatization*) представља свођење променљивих врста речи на њихов основни облик. Без лематизације машина би појавне облике исте речи третирала као различите појмове. Примера ради, у одсуству лематизације појавни облици глагола „*to be*“ (а то су „*be*“, „*am*“, „*are*“, „*is*“, „*was*“, „*were*“ и „*been*“) третирали би се као различите речи, а не као иста реч. Да би се лематизација спровела успешно, поред прецизног лексикона потребно је да улазне речи буду технички исправне. Из тог разлога алгоритам води рачуна и о третману великог слова, чисти идентификоване речи од сувишних симбола и предузима друге техничке кораке не би ли се лематизација спровела исправно. Додатно, алгоритам ниско фреквентне речи (оне које се у целокупном опусу преузетих текстова појављују мали број пута) спаја са најближим синонимима по значењу. Присуство оваквих речи би могло да произведе шум у финалној анализи, што се овим путем спречава. Алгоритам води рачуна и о латинским фразама и изразима који се често користе у говорном језику, полусложеницама које се не могу третирати као две одвојене реч, вулгаризмима, заградама спојеним са речима (примера ради „(де)централизоване финансије“, „(не)ефикасно тржиште“), невидљивим и непознатим карактерима, али и бројним другим аспектима неопходним за стандаризовање текста.

Слика 13: Скраћени приказ рада алгоритма за рударење текста (псеудо код)



Извор: приказ аутора

Изграђени алгоритам се ослања на три групе ресурса[37]. Први су регуларни изрази (енгл. *regular expressions* или скраћено *RegEx*). Регуларни изрази представљају одређене маске или шаблоне путем којих је могуће идентификовати жељене делове текста. Изузетно су корисни за препознавање елемената текста неупотребљивих у моделирању, за отклањање техничких и/или правописно-граматичких грешака, али и за рад са карактеристичним конструкцијама у тексту (попут скраћеница, синтагми, специфично написаних речи и сл.). Како би се читаоцу приближио овај концепт, овде наводимо пример употребе једног регуларног израза. Претпоставимо да је циљ пронаћи и одстранити властита имена људи[38] из текста. Једна могућа маска би проверавала да ли нека од реченица у себи садржи две суседне речи које почињу великим словом, након којих следи зарез, па реч која означава занимање или вршиоца радње, при чему између њих може стајати једна или више произвољних речи. Уколико се оваква конструкција пронађе, суседне речи написане великим словом представљају властита имена која треба одстранити. Оваква маска би у реченицама попут: „*This was proposed by Elon Musk, controversial entrepreneur and investor.*“ и „*According to Vitalik Buterin, co-creator of Ethereum, an update will be available before the end of the month*.“ идентификовала речи „Илон“, „Маск“, „Виталик“ и „Бутерин“ као властита имена. Наравно, ово није једина маска која се мора направити да би се детектовала властита имена у тексту. Због сложености природних језика различите конструкције се могу појавити у тексту на различите начине што отежава израду адекватног алгоритма. Другим речима, имена људи у тексту не препознајемо само по занимањима, већ и по другим појмовима и реченичним конструкцијама. Осим тога, имена људи могу бити дата на различите начине (иницијали, пуно име са средњим словом или средњим именом, само име, само презиме, надимак и др.). Такође, властита имена могу имати различит облик (попут властитог имена са апострофом (*O'Connor*), властититог имена које има велико слово усред речи (*McDonalds*), властитог имена на страним језицима (*Витáлий*) итд.). Све то изискује прављење великог броја маски и усложњавање алгоритма. При томе не сме се изгубити из вида да је ово само једна операција коју алгоритам обавља, а има их још пуно.

Други коришћени ресурс је библиотека *NLTK* (скраћено од енгл. *Natural Language Toolkit*). Реч је о библиотеци која нуди одређене функционалности из области процесирања природног језика. Поједини готови алати из ове библиотеке коришћени су за изградњу алгоритма за рударење текста. Неки од поменутих алата омогућили су убрзање активности попут: поделе текста на реченице и реченица на речи, препознавање врсте речи, детекцију поштапалица (енгл. *stop words*), пружање подршке приликом лематизације и сл. Ипак, оваквим модификацијама додатно би се закомпликовао изложени алгоритам, док би се његов рад знатно успорио. Насупрот томе, постојећом структуром се постиже задовољавајући однос ефикасности и прецизности.

Коначно, трећи важан ресурс је лексикон енглеских речи који је израдио сам аутор. Највећи значај израђеног лексикона је то што гарантује исправну лематизацију. Наиме, комбинујући

[37] Поред поменутих ресурса, у изградњи су коришћене и друге библиотеке које представљају основни градивни елемент кодова у Пајтону (попут библиотека *NumPy* и *PanDas*). Ипак, најзначајнију улогу одиграле су поменуте три групе ресурса.

[38] Властита имена представљају неупотребљив елемент текста у анализи сентимента због немоућности да се њихов сентимент објективно одреди.

већ постојеће алате из библиотеке *NLTK*[39] са лексиконом аутора, могућност алгоритма да направи било какву грешку приликом лематизације је значајно смањена. Осим спровођења лематизације, овај лексикон представљао и основ за спајање ниско фреквентних речи са најближим синонимима, отклањање једног дела честих правописно-граматичких и техничких грешака из текста, препознавање вулгаризама, јединица мере и слично. Лексикон, између осталог, садржи појавне облике, заједничке корене и синониме за око 5000 речи. Израђени лексикон ослања се на онлајн доступне речнике енглеског језика *Google Dictionary*[40] (лиценциран од стране *Oxford University Press*) и *Merriam Webster*[41] (највећи и најтиражнији амерички издавач лингвистичких дела са преко 170 година искуства у изради речника).

Ипак, израђени алгоритам има и својих мана. Прецизност алгоритма се може подићи продубљивањем семантичке анализе текста формирањем к-торки (енгл. *k-grams*) речи уз помоћ покретних прозора. Такође, прецизност резултата би се повећала и уколико би се уместо фреквенција за претварање текстова у векторе користиле напредније процедуре. Примера ради, процедура ембедовања (енгл. *embedding* – уграђивање/усађивање) речи у векторе која не формира математичке векторе за сваки текст, већ за сваку реч посебно. Поред поменутог, рад алгоритма би могао да се убрза осмишљавањем ефикаснијих маски, али и побољша додавањем неких нових маски за претрагу.

Као резултат рударења сваки текст ће бити претворен у вектор речи. То су вектори чији су елементи речи и синтагме које се појављују у тексту. Речи и синтагме једним именом називамо концептима. Димензија сваког вектора речи зависиће од укупног броја концепта који је алгоритам изрударио. Према томе различити текстови биће представљени векторима речи различитих димензија. Да би анализа могла да се настави, потребно је мапирати векторе речи у математичке (нумеричке) векторе.

3.4 TF-IDF

TF-IDF је статистика развијена у теорији информација. Она показује важност неке речи за дати текст узевши у обзир неки корпус текстова. Другим речима, *TF-IDF* нам омогућава да меримо информативну моћ сваке речи у анализираном тексту. Из тог разлога овај показатељ је пронашао своје место у рударењу текста, претраживачима, кориснички оријентисаним системима препоручивања (колаборативно филтрирање) и слично. О популарности ове статистике сведочи истраживање Бреитингера (*Breitinger*) и сарадника (2015), који су показали да 83% система за претраживање и препоручивање своје резултате базира на *TF-IDF*-у. Назив

[39] Верзија пакета *NLTK* која је била доступна у време писања овог рада није могла сама по себи да обезбеди исправну лематизацију. Пакет је био нарочито склон прављењу грешке уколико би реч била написана великим почетним словом или верзалом, уколико би се реч нашла уз специјални карактер и др.

[40] Чијем садржају је приступљено преко екстензије: *https://chrome.google.com/webstore/detail/google-dictionary-by-goog/mgijmajocgfcbeboacabfgobmjgjcoja* и гуглове веб странице за превођење: https://translate.google.com/. (последњи пут посећени дана 28.08.2022)

[41] Чијем садржају је приступљено преко сајта: *https://www.merriam-webster.com/* (последњи пут посећени дана 28.08.2022)

ове статистике је скраћеница за израз „однос фреквенције речи/појма и инверзне фреквенције документа/текста“ (енгл. *term frequency – invers document frequency*). Статистика је назив добила због тога што се за одређивање информативности, односно важности, неке речи у тексту користе и фреквенција дате речи и инверзна фреквенција текстова. Иза овакве конструкције стоје две премисе. Прва премиса тврди да што се реч више појављује у тексту, то је већа вероватноћа да она представља неки битан појам или информацију. Друга премиса допуњује претходну и тврди да уколико се иста реч пречесто појављује и у другим текстовима, онда она представља неки општи појам, а не важну информацију. Из тог разлога, само истовременим сагледањем обе фреквенције можемо да разазнамо информативност датог појма. На важност инверзне фреквенције документа прва је указала Спарк-Џоунс *(Spärck-Jones*) (1972), али је теоријско оправдање за увођење овог појма у теорији информација дао тек Аизава (*Aizawa*) (2003) тридесет година касније. Постоји више начина на који се *TF-IDF* може обрачунати. Сваки приступ има своје предности и мане, те избор обрачуна ове статистике зависи од њене примене. У овом раду биће коришћена оригинална дефиниција *TF-IDF*-а, дата формулом испод (Вајс (*Weiss*) и сарадници 2015):

$$TFIFD_{i,j} = tf_{i,\mathrm{j}} \cdot idf_{i,j} = \frac{f_{i,j}}{\sum_{i=1}^{n_j} f_{i,j}} \cdot \ln\left(\frac{N}{N_{i,j}}\right) \quad (3)$$

где је: $tf_{i,j}$ фреквенција *i*-те речи у *j*-том тексту, $idf_{i,j}$ инверзна фреквенција текстова који садрже *i*-ту реч из *j*-тог текста, *N* укупан број текстова, $N_{i,j}$ број текстова који у себи садрже *i*-ту реч из *j*-тог текста, $f_{i,j}$ апсолутна фреквенције *i*-те речи у *j*-том тексту и n_j број речи у *j*-том тексту. Приликом израчунавања инверзне фреквенције документа (у ознаци $idf_{i,j}$) искоришћен је природни логаритам да би се ублажио њен утицај на коначну вредност *TF-IDF*-а. Без ове корекције количник $\frac{N}{N_{i,j}}$ би могао да буде значајно велики број, будући да је $N_{i,j} \leq N$.

Ова статистика има и својих слабости које ћемо укратко анализирати. Пре свега *TF-IDF* је могуће користити само када се рударење текста врши на бази појединачних речи, те није компатибилна са напреднијим процедурама. Такође, статистика занемарује чињеницу да важност речи може зависити од њене позиције у реченици, али и у самом тексту. Статистика занемарује и чињеницу да речи могу имати синоним. То би значило да иако се реч директно не појављује у тексту, можда је нека друга реч која носи њено значење и семантички је замењује присута у тексту, а то остаје игнорисано. Коначно, будући да за сваку реч мора да провери колико се пута појавила у датом тексту и да ли се појављује у осталим текстовима, њен обрачун може да буде захтеван за рачунар и да потраје дуже од обрачуна других статистика. Ипак, наведене слабости нису престављале препреку за имплементацију *TF-IDF*-а у ово истраживање.

TF-IDF ће у овом раду бити употребљен у оквиру *t2v* процедуре за мапирање вектора речи у математичке (нумеричке) векторе. Сваки добијени нумерички вектор представљаће један текст. Димензија нумеричких вектора биће једнака укупном броју јединствених речи (у ознаци *I*) идентификованих на бази целог корпуса анализираних текстова. Дакле, овде није реч о укупном броју јединствених речи у тексту, већ о укупном броју јединствених речи за цео корпус текстова. Такође, приметимо да су математички вектори конструисани да буду истих димензија за разлику од вектора речи чије димензије варирају од текста до текста. Да би се формирао нумерички вектор неког текста, потребно је обрачунати *TF-IDF* за сваку од *I*

идентификованих јединствених речи. Уколико се посматрана реч не појављује у датом тексту, њен *TF-IDF* ће бити 0 (јер је њена $tf_{i,j} = 0$). У супротном, *TF-IDF* посматране речи ће бити нека позитивна вредност. Сваки елемент конструисаног нумеричког вектора, у ознаци $TFIDF_{ij}$, представља вредност *TF-IDF*-а *i*-те јединствене речи израчунатог за *j*-ти текст. Спајањем свих овако конструисаних вектора у једну матрицу добићемо матрицу квантификованих текстова која се користи као улазна вредност за анализе које ће уследити. Општи запис ове матрице дат је изразом (4):

$$\begin{array}{c} \\ \text{реч}_1 \\ \text{реч}_2 \\ \vdots \\ \text{реч}_I \end{array} \begin{array}{c} \begin{array}{cccc} \text{текст}_1 & \text{текст}_2 & \cdots & \text{текст}_N \end{array} \\ \begin{pmatrix} TFIDF_{11} & TFIDF_{12} & \cdots & TFIDF_{1N} \\ TFIDF_{21} & TFIDF_{22} & \cdots & TFIDF_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ TFIDF_{I1} & TFIDF_{I2} & \cdots & TFIDF_{IN} \end{pmatrix} \end{array} \tag{4}$$

Из матрице квантификованих текстова потребно је уклонити две врсте вектора: векторе за неуспешно скинуте текстове[42] (нула векторе) и векторе за текстове дупликате. У супротном добијена матрица ће бити сингуларна. Овде кратко дискутујемо узроке оба проблема. До неуспешног скидања текстова долази када је на селектованој страници портала Крипто Њуз постављен само линк ка неком другом порталу. На овај начин Крипто Њуз даје могућност својим читаоцима да прочитају одређени текст за који нису добили дозволу аутора за постављање на својој страници. У овом случају, алгоритам за скреповање неће скинути ништа, будући да је испрограмиран да преузима само текстуални садржај и игнорише све остало. Са друге стране, до појаве дупликата може доћи из два разлога. Први разлог је последица подударности. Наиме, постоје текстови који истовремено пишу о неколико крипто-валута. Ово може бити проблем будући да истраживање обухвата и текстове о Биткоину због анализе унакрсног-сентимента. Уколико постоји текст који пише и о посматраној крипто-валути и о Биткоину (на пример пореди њихове перформансе, анализира како ће одређене економске околности утицати на сваки од њих појединачно и сл.) биће скинут два пута (једном из угла посматране крипто-валуте и једном из угла Биткоина). Други разлог је последица људског фактора. Конкретно, платформа грешком може да објави два идентична текста на исти дан, у приближно исто време. Појава ових проблема не утиче на исход овог истраживања из два разлога. Прво, Крипто Њуз дневно објави велики број текстова о свакој крипто-валути. Изостављањем неког од њих не губимо на општости. Са друге стране, поменути проблеми се јављају у малом броју случајева (јављају се у свега 3,02% случајева). Након што се изврше обе трансформације матрице квантификованих текстова, матрица ће бити спремна за изградњу модела машинског учења у другој етапи рударења текста

3.5 Индекс замагљености

Поруке и ставове које текст пропагира, као и тон са којим је исти написан, не могу утицати на понашање читалаца уколико им нису јасни. Другим речима, уколико је текст написан тако да није разумљив просечном читаоцу, његов ефекат ће бити мали, непостојан или чак обрнут од

[42] Упитању су вектори настали од празних вектора речи (видети слику **13**).

очекиваног. Треба напоменути да овде није реч о физичкој читљивости (која се, на пример, појављује код нечитког рукописа), већ о лексичкој читљивости теста. Под лексичком читљивошћу подразумевамо јасноћу са којом је текст написан. На то утичу стил писања, вештина приповедања и објашњавања, избор речи, сложеност реченица, ниво стручности и други аспекти. Због утицаја који читљивост има на разумевање текста (а самим тим и на одлуке појединаца) осмишљене су технике њеног мерења. Овај рад ће представити једну такву меру – индекс замагљености (енгл. *the fog index*). Овај показатељ је развио Ганинг (*Gunning*) (1952) у лингвистичке сврхе. Основна премиса од које се полази приликом његове конструкције јесте да текстови који садрже дугачке реченице са пуно многосложних речи (речи које имају више од два слога) нису једноставни за читање обичном човеку. С тим у вези, Ганингов (1952) индекс замагљености дефинише на следећи начин:

$$FI = 0.4(ASL + PoCW) \quad (5)$$

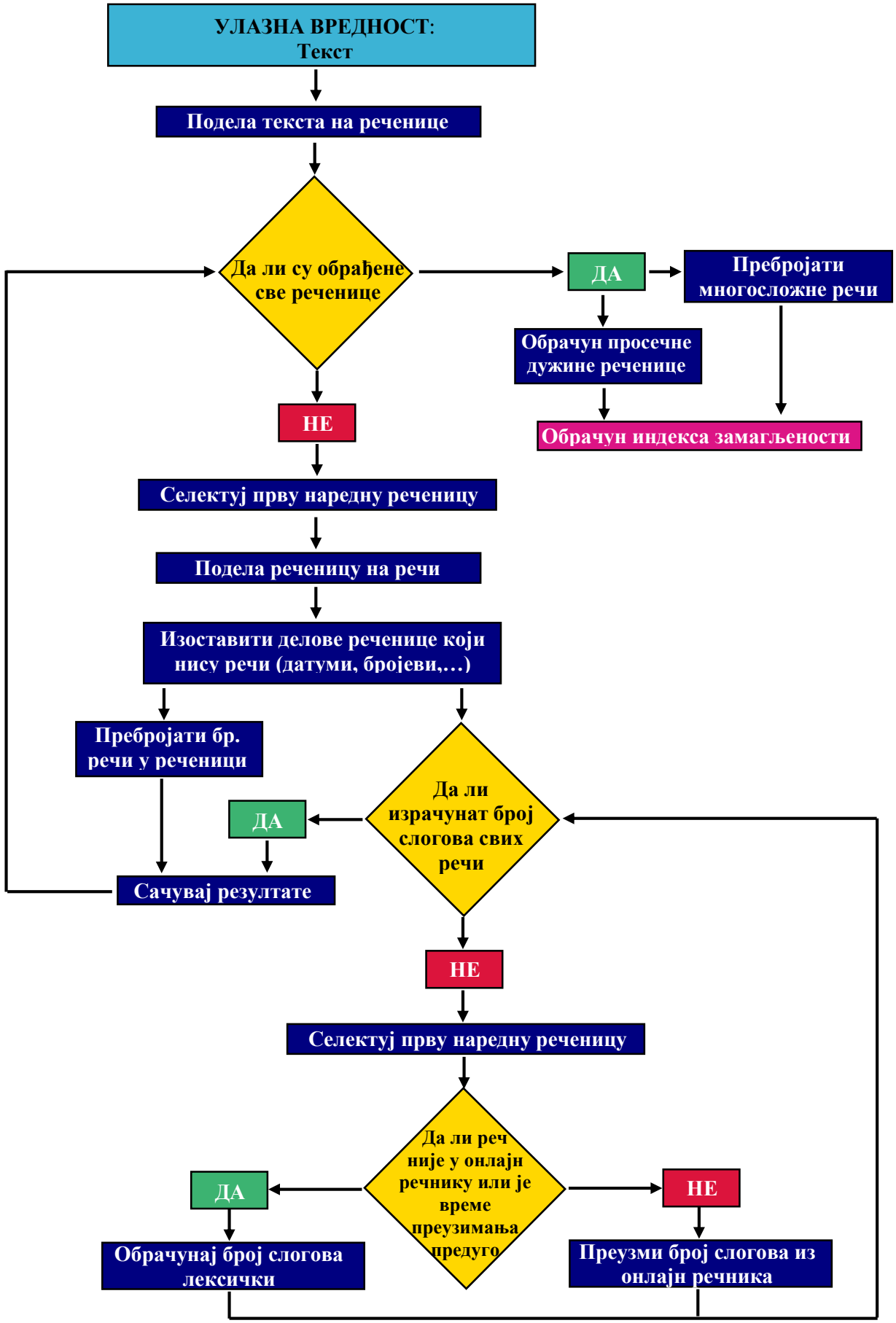
где је: *ASL* (енгл. *Average sentence length*) просечна дужина реченица (израчуната као просечан број речи по реченици у тексту), а *PoCW* (енгл. *Proportion of complex words*) удео многосложних речи у укупном броју речи у тексту.

Што је вредност индекса већа, то је текст "замагљенији" читаоцу, те поруке који исти са собом носи неће допрети до читаоца у пуној мери. Сматра се да је читљивост добра уколико је вредност овог индекса око 7 или мање. Са друге стране, уколико је вредност овог индекса преко 12 сматра се да је текст тежак за разумевање просечном човеку, већ да је за његово разумевање потребно високо образовање и/или стручност у датој области.

Одлука да се у истраживање укључи и техничка карактеристика текста као што је читљивост није случајна. Коен и сарадници (2020) показали су да сличност текстова објављених у корпоративним извештајима може имати утицај на њихову вредност. Претходни резултат популаризовао је употребу техничких својстава текста у финансијским истраживањима и мотивисао је укључивање неког од њих у ову дисертацију. Како је питање утицаја читљивости текста на крипто-валуте још увек недовољно истражено, постављено је следеће истраживачко питање: може ли јасноћа текста бити индикатор кретања приноса крипто-валута? Прецизније речено, ова дисертација ће покушати да провери да ли постоји веза између конфузно или јасно написаних текстова о крипто-валутама из претходног периода и будућег кретања његових приноса.

За потребе овог истраживања аутор је израдио програм који рачуна индекс замагљености према формули (5) за сваки текст из датог корпуса текстова. Поменути програм се такође ослања на процесирање природног језика, будући да је потребно идентификовати и пребројати речи и реченице, али и пребројати слогове сваке идентификоване речи. За његову израду коришћени су исти ресурси као и приликом израде алгоритама за рударење и скреповање текста. Зарад брзине утврђивања броја слогова, програм најпре покушава да скрепује ту информацију са истакнутих онлајн речника. Уколико то није могуће, програм рачуна број слогова по комплексном сету граматичких правила енглеског језика. Како би се читаоцу олакшало праћење рада програма, направљена је шема његовог алгоритма. Приказ је дат сликом **14**.

Слика 14: Приказ алгоритма за обрачун читљивости појединачних текстова (псеудо код)



Извор: приказ аутора

3.6 Обрачун приноса

Временске серије цена економских добара, па и финансијских инструмената, обично одликује присуство јединичног корена. Занемаривањем његовог присуства у процесу моделирања се могу добити неадекватни резултати. Из тог разлога се у финансијским истраживањима, без губитка општости, традиционално прелази на моделирање временске серије приноса уместо цена. Приноси представљају релативну промену вредности инвестиције. Уједно приноси су релативна мера периодичних профита остварених по основу инвестирања. За потребе овог истраживања приноси ће бити обрачунати на дневном нивоу уз претпоставку континуалног времена, тј. по логаритамској формули датој испод:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \tag{6}$$

где су P_t цене.

Овај приступ обрачуна одабран је због пожељних својстава које имају логаритамски приноси (видети Цаи (*Tsay*) 2006 или Кристоферсен (*Christoffersen*) 2012). Између осталог, овако обрачунати приноси су прва диференца логаритма цене што ће се показати корисним у тестовима слабе форме тржишне ефикасности (видети секцију **3.11**).

3.7 Анализа утицаја и испитивање постављених хипотеза

Како би се утврдило који од анализираних показатеља може да послужи као предиктор приноса, дисертација ће размотрити један помоћни модел. Модел је назван помоћним због чињенице да је ултимативни циљ истраживања предвиђање приноса. Из тог разлога главни модел биће финални ансамбл алгоритам представљен у наредној секцији. Помоћни модел ће послужити као подршка за изградњу поменутог финалног ансамбл алгоритма, јер ће на бази њега бити извршена селекција релевантних предиктора. Потенцијални предиктори разматрани у овом истраживању су: ранији ниво приноса, сентимент текстова о посматраној крипто-валути, замагљеност датих текстова и сентимент текстова о Биткоину. Из досадашњег излагања јасно је да сваки од истакнутих потенцијалних предиктора може да има утицај на кретање приноса крипто-валута, а циљ помоћног модела је да испита да ли су ти утицаји заиста постојали. Оваквим избором предиктора истраживање у дисертацији жели да провери да ли је могуће да се само на бази информација из текста и прећашњег кретања приноса добро предвиди кретање будућих приноса. На тај начин се симулира спровођење савремене техничке анализе и испитује њена употребна моћ на тржиштима крипто-валута. Додатан разлог за овакав избор предиктора је то да нам онда помоћни модел може послужити за испитивање валидности прве две постављене хипотезе. О томе дискутујемо у наставку ове секције.

3.7.1 Поставка помоћног модела за анализу утицаја

Приликом постављања помоћног и финалног модела пошло се од претпоставке да се предвиђање врши један дан унапред. То би значило да се на бази дешавања у току данашњег дана могу наслутити и предвидети сутрашња дешавања. Ослањајући се на изнету премису, можемо поставити следећу регресиону једначину у којој фигуришу сви потенцијални предиктори:

$$r_{k,t} = c + \beta_1 r_{k,t-1} + \beta_2 S_{k,i,t-1} + \beta_3 S_{BTC,j,t-1} + \beta_4 FI_{k,i,t-1} + \varepsilon \quad (7)$$

где су: $S_{k,i,t}$ сентимент *i*-тог текста о крипто-валути *k* објављеног на *t*-ти дан, $S_{BTC,j,t}$ сентимент *j*-тог текста о *BTC*-у објављеног на *t*-ти дан, $FI_{k,i,t}$ индекс замагљености *i*-тог текста о крипто-валути *k* објављеног на *t*-ти дан.

Као што је већ речено избор предиктора мотивисан је жељом да се истраживање спроведе у духу техничке анализе и истовремено испитају две од три постављене хипотезе. Захваљујући помоћном моделу, постављање математичке форме ових хипотеза биће тривијално. Да подсетимо, прва хипотеза претпоставља постојање везе између приноса и сентимента вести. Њена валидност испитаће се провером значајности параметра нагиба уз сентимент вести о датој крипто-валути у помоћном моделу. Нумеричка форма ове хипотезе дата је у наставку:

$$\text{H}_{\text{A},0}\text{: } \beta_2 = 0$$

$$\text{H}_{\text{A},1}\text{: } \beta_2 \neq 0$$

Слично томе, друга хипотеза претпостављала је значајну везу између приноса и сентимента вести о Биткоину, као и између приноса и читљивости написаних вести. За проверу валидности ове сложене хипотезе у помоћној регресији анализираће се значајност параметара нагиба уз читљивост текстова и сентимент вести о Биткоину. У квантитативном запису ову сложену хипотезу можемо представити као две просте – по једну за сваку променљиву (означене са Б1 и Б2):

$$\text{H}_{\text{Б1},0}\text{: } \beta_3 = 0 \qquad \text{H}_{\text{Б2},0}\text{: } \beta_4 = 0$$

$$\text{H}_{\text{Б1},1}\text{: } \beta_3 \neq 0 \qquad \text{H}_{\text{Б2},1}\text{: } \beta_4 \neq 0$$

Осврнимо се још укратко на улогу ауторегресивне компоненте у помоћној регресији. Подсетимо читаоца да техничка анализа претпоставља да постоје шаблони у кретању цена (а самим тим и приноса) који се могу искористити за предвиђање будућег кретања. Увртавањем ауторегресивне компоненте у модел истраживање проверава истинитост те тврдње на тржиштима крипто-валута. Додатан мотив за ову одлуку је и то што су Џагадиш и Ву (2019) у свом оригиналном раду такође укључили ауторегресивну компоненту у свој финални модел за предвиђање кретања приноса финансијских инструмената. Осим тога, ова компонента игра још једну важну улогу. Она ће нам послужити и као сигнал за потенцијално присуство слабе

форме тржишне ефикасности. Постојање лагова у кретању приноса је у супортоности са слабом формом хипотезе о тржишној ефикасности (видети секцију **3.11**), те би статистичка значајност ове компоненте дала мотив више да се слаба форма тржишне ефикасности формално и испита.

3.7.2 Трансформација података и финални запис помоћног модела

Из иницијалне поставке помоћног модела одмах је видљив проблем неусаглашености фреквенција изабраних предиктора. Приметимо да бесплатно јавно доступне податке о ценама (приносима) имамо једном дневно, док се у току једног дана обично објави неколицина чланака о посматраној крипто-валути. Додатно, број објављених чланака се мења из дана у дан. Другим речима, фреквенције објављивања нису фиксне. Анализу додатно компликује чињеница да се фреквенције објављивања разликују и између валута. Примера ради, број објављених чланака о Итиријуму на одређени дан не мора нужно бити једнак броју објављених чланака о Биткоину на исти тај дан. Напротив, у општем случају ове две вредности ће се готово извесно разликовати сваког дана. Постављени проблем се може класификовати као специфичан облик *MIDAS* (енгл. *Mixed Data Sampling*) проблема[43].

Дисертација ће понудити једно једноставно решење датог проблема погодно са аспекта машинског учења тестирано у раду Дамјановића и Дреновака (2023). Идеја је да све фреквенције сведемо на фреквенцију најважније променљиве у моделу, тј. на фреквенцију објављивања чланака о посматраној крипто-валути. То значи да је укупан број опсервација искоришћених за моделирање једнак укупном броју објављених чланака (N). Ова вредност израчунавамо на следећи начин:

$$N = \sum_{t=1}^{T-1} n_t \tag{8}$$

где је n_t укупан број чланака о датој крипто-валути објављених на дан t.

Постоји још један битан разлог из којег је одлучено да се фреквенције свих променљивих у моделу сведу на фреквенције објављивања чланака о посматраној крипто-валути. Захваљујући оваквој одлуци, модели ће имати неупоредиво више опсервација за тренинг, будући да се у току једног дана објави велики број чланака. Да су све фреквенције сведене на фреквенције зависне променљиве (приноса) то не би био случај. У тој ситуацији мањак опсервација би било могуће надоместити значајним проширењем узорка. Ипак, таква одлука наишла би на две препреке. Прво, враћањем превише далеко у прошлост губи се актуелност резултата јер ће на оцене веза између предиктора и приноса утицати вредности из периода значајно удаљеног од периода од интереса. Друго, крипто-валуте су изразито млада финансијска актива. Многе крипто-валуте постоје само пар година, те велике узорке у овом тренутку свакако није могуће формирати. Осим тога, популарност крипто-валута у ранијим периодима није била на оном

[43] Назив представља скраћеницу за израз „мешовито узорковање података“ (енгл. ***Mi**xed **Da**ta **S**ampling*). Реч је о добро познатој породици проблема и у статистичком моделирању и у машинском учењу.

нивоу на којем је данас. Заинтересована јавност за ову финансијску активу чинила је много мањи део популације, те је било и мање портала и текстова који прате актуелна дешавања повезана са крипто-валутама.

Вектор зависне променљиве (тј. приноса) је сада потребно прилагодити вектору сентимената објављених чланака. Да би то постигли, сваком чланку придружујемо принос који је остварен дан након што је текст написан. То значи да је вектор од T приноса потребно редефинисати у вектор од N приноса, тако да принос стечен на дан t одговара сваком тексту објављеном на претходни дан. Нови вектор приноса представљен је следећим изразом:

$$r = \begin{pmatrix} r_T \\ \vdots \\ r_T \\ r_T \\ \vdots \\ r_3 \\ \vdots \\ r_3 \\ r_3 \\ r_2 \\ \vdots \\ r_2 \\ r_2 \end{pmatrix} \begin{matrix} n_{T-1} \\ \vdots \\ 2 \\ 1 \\ \vdots \\ n_2 \\ \vdots \\ 2 \\ 1 \\ n_1 \\ \vdots \\ 2 \\ 1 \end{matrix} \Big\} N \qquad (9)$$

Нешто сложенији проблем се појављује приликом редефинисања вектора сентимента вести о Биткоину. Не само да се фреквенције објављивања чланака о посматраној крипто-валути не поклапају са фреквенцијама објављивања чланака о Биткоину, већ се оне не могу ни једнозначно повезати. Другим речима, не може се једнозначно одредити с којим од укупно $n_{BTC,t}$ чланака објављених о Биткоину на дан t треба упарити i-ти текст о посматраној крипто-валути објављен на исти тај дан. Како би се тај проблем превазишао, рачуна се просечан ниво сентимента на одређени дан, t, за чланке о Биткоину.

$$\bar{S}_{BTC,t} = \frac{1}{n_{BTC,t}} \sum_{j=1}^{n_{BTC,t}} S_{BTC,j,t} \qquad (10)$$

где је $\bar{S}_{BTC,t}$ просечан ниво сентимента за чланке о Биткоину објављених на дан t, а $n_{BTC,t}$ укупан број текстова о Биткоину објављених на дан t.

На тај начин од вектора са N_{BTC} елемената добијамо вектор од T елемената. Након ове трансформације могуће је свакој од N опсервацији једнозначно придружити по један елемент вектора просечних сентимената чланака о Биткоину. Поступак упаривања ће бити готово исти као у случају приноса. Једина разлика биће то што се уместо просечног сентимента наредног дана, узима просечан сентимент истог дана. Новодобијени вектор дат је изразом испод:

$$\mathrm{X}_{BTC} = \begin{pmatrix} \bar{S}_{BTC,\mathrm{T}-1} & n_{T-1} \\ \vdots & \vdots \\ \bar{S}_{BTC,\mathrm{T}-1} & 2 \\ \bar{S}_{BTC,\mathrm{T}-1} & 1 \\ \vdots & \vdots \\ \bar{S}_{BTC,2} & n_2 \\ \vdots & \vdots \\ \bar{S}_{BTC,2} & 2 \\ \bar{S}_{BTC,2} & 1 \\ \bar{S}_{BTC,1} & n_1 \\ \vdots & \vdots \\ \bar{S}_{BTC,1} & 2 \\ \bar{S}_{BTC,1} & 1 \end{pmatrix} \Bigg\} N \tag{11}$$

Коначно, како је за сваки чланак о посматраној крипто-валути рачунат ниво индекса замагљености, ових вредности ће бити исто колико и чланака (тј. N). Из тог разлога није потребно правити никакве модификације њиховог вектора зарад његовог укључивања у модел.

Узевши све наведено у обзир, можемо поставити коначни облик помоћног модела:

$$r_{k,t} = c + \beta_1 r_{k,t-1} + \beta_2 S_{k,i,t-1} + \beta_3 \bar{S}_{BTC,t-1} + \beta_4 FI_{k,i,t-1} + \varepsilon_{k,i,t} \tag{12}$$

Приметимо да израз (12) подсећа на панел једначину у којој су приноси и просечан сентимент чланака о Биткоину временске променљиве. Додатно, приметимо и то да ће овај запис помоћног модела бити искоришћен за испитивање постављених хипотеза и селекцију предиктора, будући да је из њега отклоњен проблем неусаглашености фреквенција.

3.8 Ансамбли

Ансамбли су једна од најмлађих класа модела машинског учења чији настанак се повезује са такмичењем „Нетфликсова Награда“. Такмичење је организовала компанија Нетфликс, у то време највећи *DVD* видеоклуб у Сједињеним Државама. Циљ организовања такмичења био је да компанија дође до нових идеја за унапређење свог система за препоручивање филмова корисницима. Нетфликс је такмичарима дао приступ великом делу своје базе податак о филмовима и ставовима њених корисника, а њихов задатак је био да осмисле што бољи модел машинског учења за препоручивање филмова на бази датих података. Како су године пролазиле, појавио се интересантан тренд. Све више тимова одлучивало се на спајање са другим тимовима, како би се обједињавањем њихових модела добиле прецизније процене. Уочивши овај тренд, убрзо је и сам компанијски стручни жири спознао да се обједињавањем различитих процена добија значајно прецизнија процена непознате вредности. Практична имплементација уочене идеје резултирала је значајним подизањем квалитета препорука и невероватним растом популарности компаније. Приступ је убрзо добио назив ансамбл, јер у њима више процена учествује у стварању прецизније процене, баш као што више музичара из

музичког ансамбла заједно ствара лепшу музику од њиховог појединачног свирања. Остало је историја.

Иако ансамбли постоје краће од две деценије, идеја обједињавања или комбиновања која стоји иза њих много је старија. О томе сведочи њена широка распрострањеност у бројним сферама људског интересовања попут музике, фармације, кулинарства и др. Пример примене ове идеје можемо наћи чак и у финансијама. Комбиновањем више финансијских инструмената у портфолио добијамо мање ризичну финансијску активу од сваког појединачног инструмента. Обједињавање није новина ни у статистичком моделирању. Истакнути викторијански статистичар Френсис Галтон (*Francis Galton*) је показао употребну моћ упросечавања, једног од приступа обједињавању процена. Галтон је посматрао сеоске вашаре и сточне пијаце на којима су се организовала такмичења погађања тежине изложене стоке. Када би израчунао аритметичку средину процена свих људи који су учествовали у такмичењу, добио би процену која је одговара тачној вредности тежине стоке или јој је изузетно блиска. Управо је овако описано упросечавање, најпопуларнији приступ у обједињавању процена, који користе највећи број ансамбла. Поред аритметичке средине[44], за обједињавање се користе и медијана, пондерисана аритметичка средина, обрезана аритметичка средина (енгл. *trimmed mean*), геометријска средина и др.

Ансамбл се може формирати на неколико начина. Први приступ је да се дизајнира неколико модела чијом ће се применом на исти узорак добити више предикција које треба објединити. Други приступ подразумева да се исти модел примени на више узорака на бази којих се предвиђа иста непозната вредност. Описани приступ је погодан када је циљ избећи претерано прилагођавање модела једном узорку (енгл. *over-fitting*). Коначно, последњи приступ подразумева да се више модела примени на више различитих узорака тако да сваки узорак одговара тачно једном моделу. Све тако добијене предикције се обједињују на један од већ дискутованих начина. Највећа предност свих варијанти ансамбла је подизање прецизности предикција. Ипак, та прецизност има своју цену. Главни недостатак ансамбла је то што су изразито рачунски захтевни. Последично, њихов рад захтева ангажовање значајних рачунарских ресурса и изискује дуже време израчунавања. То ансамбле чини неадекватним приступом предвиђања уколико је брзина прављења прогнозе важна.

3.8.1 Израђени ансамбл алгоритам за предикцију

Овај одељак представља израђени алгоритам из породице ансамбла који предвиђа приносе на дан t на бази релевантних предиктора идентификованих из модела (12). Идеја алгоритма је да оцени адекватан облик модела (12) на бази двомесечног узорка текстова који непосредно претходе тренутку $(t-1)$. Добијени модел се примењује на све текстове објављене у тренутку

[44] Побројане технике обједињавања се користе када је циљна променљива нумеричка. У случају да је реч о дискретној или атрибутивној променљивој користе се други приступи обједињавања. Један од најпопуларнијих приступа коришћен у описаном случају је систем „већине гласова“.

$(t-1)$ како би се добило n_{t-1} предикција приноса у тренутку *t*. Добијене предикције се потом обједињују. У наставку одељка изложене су све етапе алгоритма:

1. Уколико је циљ предвидети принос у тренутку *t*, у првом кораку потребно је селектовати све текстове који су објављени у претходна два месеца у односу на тренутак $(t-1)$. Међу ове текстове не треба сврстати оне објављене у тренутку $(t-1)$, тј. објављене претходног дана у односу на дан за који правимо предикцију.

2. На бази селектованих текстова потребно је оценити једначину (12) у којој фигуришу само предиктори идентификовани као значајни у помоћном моделу описаном у секцији **3.7**.

3. Након тога потребно је селектовати све текстове објављене на изостављени дан, односно, све текстове објављене у тренутку $(t-1)$.

4. Претпоставимо да је на изостављени дан било n_{t-1} текстова о посматраној крипто-валути. Убацивањем показатеља добијених њиховим рударењем у оцењени модел из другог корака добићемо n_{t-1} предикција могућих вредности сутрашњих приноса.

5. Коначна предикција сутрашњих приноса биће аритметичка средина свих n_{t-1} предвиђених вредности.

$$\hat{\bar{r}}_{t-1}(1) = \frac{1}{n_{t-1}} \sum_{j=1}^{n_{t-1}} \hat{r}_{j,t-1}(1) \tag{13}$$

где је: $\hat{r}_{j,t-1}(1)$ предикција приноса из тренутка *t* на бази *j*-тог текста објављеног дан раније, а $\hat{\bar{r}}_{t-1}(1)$ обједињена предикција приноса коју ансамбл предлаже.

6. Након успешно завршене предикције, прелази се на предвиђање наредног приноса. За то је потребно померити алгоритам за 1 дан унапред, по принципу покретних прозора и поновити целокупну процедуру. Алгоритам стаје са радом када су предикције за све дане из трећег подузорка креиране.

Очекивано је да овакав приступ да прецизније резултате предвиђања из неколико разлога. Пре свега за сваку предикцију оцењује нови модел из двомесечног узорка ранијих текстова (тј. учи се из непосредне прошлости). Додатно, свака предикција се добија обједињавањем n_{t-1} прво оцењених потенцијалних предикција. Осим тога, алгоритам обезбеђује да се приликом прављења предикција за наредни дан у обзир узимају све доступне релевантне информације у току претходног дана. Предност алгоритма је и то што се захваљујући упросечавању смањује могућност пристрасности предикција. Наиме, ниједан текст не може битно да опредељује процену приноса јер ансамбл упросечује процене сваког од њих.

Ипак, конструкција алгоритма не може бити једини гарант квалитета његовог рада. Потребно је адресирати и *GIGO* проблем (енгл. скраћеница за: *garbage in – garbage out*). Другим речима, да би се обезбедио квалитет прогноза, важно је да и сами улазни подаци буду квалитетни. То, пре свега, подразумева адекватно мерење показатеља добијених рударењем текста и да избор периода за испитивање њиховог утицаја не буде исувише удаљен од периода за који се прогнозе врше. Правилан избор периода за анализу значајности предиктора је нарочито важан код крипто-валута, због њихове осетљивости на нове информације. Изабрани период мора да одражава релевантне факторе који опредељују кретање приноса око периода за који се прогноза врши. Последично, избором превише удаљеног периода не би био сагледан само утицај тренутно релевантних предиктора, већ и предиктора који су били релевантни пре промена околности. Из тог разлога пожељно је да период у којем се испитује значајност предиктора не буде превише дугачак, нити превише удаљен од периода предвиђања. Уколико се прогнозе врше за дужи временски период, препоручује се и да се избор предиктора ревидира протоком времена. Избор адекватних периода из аспекта овог истраживања дискутујемо у одељцима **5.2** и **5.3**.

3.9 Квалитет прогнозе

За испитивање треће постављене хипотезе потребно је проверити да ли финални ансамбл модел боље предвиђа приносе када су променљиве изрударене из текста добијене по методологији Џагадиша и Вуа (2019) или по методологији коју предлаже ове дисертација. У овом одељку биће прецизиране мере квалитета прогнозе и статистички тестови на бази којих ће бити спроведена бенчмарк анализа добијених резултата. Да би се осигурала робусност добијених резултата, тестирањем ће се квалитет прогнозе истовремено проверити на чак осам узорака (тј. код осам крипто-валута). Овом приликом подсећамо читаоца да је то био један од разлога да се анализа спроведе над појединачним крипто-валутама, а не над портфолиом валута или панелом. Ово поглавље ће представити и нумерички запис треће истраживачке хипотезе, али и предложени иновирани приступ за вишеузорачко тестирање квалитета прогнозе.

3.9.1 Корен из средње квадратне грешке прогнозе

Савремена статистика нуди више показатеља путем којих се прати квалитет прогнозе. Ипак, међу њима се посебно истиче корен из средње квадратне грешке прогнозе (енгл. *Root Mean Square Error – RMSE*), као једна од најпопуларнијих мера у пракси (а нарочито у сфери машинског учења). Овај показатељ се дефинише на следећи начин:

$$RMSE = \sqrt{\frac{1}{g}\sum_{t=1}^{g}\left(r_{T+t} - \hat{r}_T(t)\right)^2} = \sqrt{\frac{1}{g}\sum_{t=1}^{g}\hat{e}_{T+t}^2} \tag{14}$$

где је: g укупан број периода за које правимо прогнозу, T последњи период укључен у моделирање, r_{T+t} права вредност анализиране варијабле у тренутку $T+t$ и $\hat{r}_T(t)$ вредност коју модел предвиђа у тренутку $T+t$, а $\hat{e}_{T+t}$ грешка прогнозе у тренутку $T+t$.

Модел који има мањи корен из средње квадратне грешке прогнозе сматраће се прецизнијим јер су његове грешке биле мање. Ипак, наведени показатељ представља само тачкасту оцену квалитета прогнозе, те је за генерализацију закључака потребно спровести статистичко тестирање. Из тог разлога трећа хипотеза ће истовремено тестирати да ли су грешке прогнозе прављене када се у моделу појављује сентимент измерен по методологији Џагадиша и Вуа (2019) веће од грешака прогнозе на бази алтернативне методологије код свих узорака. Квантитативно записано, трећа хипотеза гласи:

$$H_{B,0}: RMSE_{A,i} = RMSE_{JW,i}\ \forall i$$

$$H_{B,1}: \exists i\ RMSE_{A,i} < RMSE_{JW,i}$$

где је $RMSE_{A,i}$ грешка прогнозе направљена применом алтернативне методологије мерења сентимента на i-том узорку, а $RMSE_{JW,i}$ грешка прогнозе направљена применом оригиналне методологије Џагадиша и Вуа (2019) на i-том узорку. Прихватање алтернативне хипотезе значило би да су побољшања допринела смањењу грешака прогнозе код свих узорака, док би прихватање нулте хипотезе сугерисало да постоје узорци код којих то није случај.

3.9.2 Тестирање квалитета прогнозе

Некада је потребно да се статистичким тестом провери да ли један модел даје прецизнију прогнозу од другог. Класични економетријски тестови прогнозе дизајнирани су тако да утврде који од два модела даје боље резултате на конкретном узорку. То углавном радимо када је потребно извршити селекцију модела, те бирамо онај који ће у конкретно датом случају пружити боље резултате. Ипак, некада је потребно генерализовати закључке, односно, проверити да ли један модел у општем случају тежи да пружи боље резултате прогнозе од другог. Тада моделе није довољно поредити на једном случају (тј. на једном узорку), већ на већем броју њих. Од једног таквог проблема полази и ово истраживање. Наиме, потребно је проверити да ли модел израђен на бази побољшаних оцена (у даљем тексту алтернативни модел) ради боље од модела изграђеног на бази оригиналних оцена (у даљем тексту оригинални модел). Имајући у виду да је ултимативна сврха конструкције ових модела свакако предвиђање приноса, природно је трагати за оним чија је употребна вредност за прогнозирање већа. Модели ће бити упоређени код, чак, осам различитих крипто-валута (тј. код осам различитих узорака), како би се извукли генералнији закључци о њиховом квалитету. Другим речима, провериће се да ли један од њих у општем случају даје квалитетније прогнозе.

Један могући приступ овом питању је вишеструко појединачно тестирање. У том случају применили бисмо неки од класичних економетријских тестова прогнозе (попут Диеболд-Маријановог 1991 теста) на сваки од осам различитих узорака посебно. Уколико би већина тестова показала да је алтернативни модел успешнији, могли бисмо да генерализујемо закључак да је алтернативни модел генерално прецизнији у предвиђању. Ипак, овај приступ има одређене слабости. Прва од њих је значајно редуковање нивоа значајности. Статистички тестови се углавном проверавају на нивоу значајности од 5%. Он представља вероватноћу доношења погрешног закључка прве врсте – односно одбацивање тачне нулте хипотезе. Поставља се питање колика ће бити вероватноћа прављења грешке прве врсте на нивоу целокупне процедуре. Грешку прве врсте у целокупној процедури нећемо направити уколико је не направимо ни на једном од осам тестова. Како је реч о независним (елементарним) догађајима, вероватноћу њихове уније (сложеног догађаја) можемо израчунати као производ вероватноћа елементарних (појединачних) догађаја. Ослањајући се на претходно, вероватноћу прављења грешке прве врсте на нивоу целог истраживања можемо израчунати преко формуле за супротни догађај на следећи начин: $1 - 0{,}95^8 = 0{,}337$. Дакле, ризик прављења грешке прве врсте је много већи него што испрва може деловати. Из тог разлога статистичка пракса често предлаже прелазак или на омнибус (здружене) статистичке тестове или на примерну Бонферонијеве корекције. Омнибус статистички тестови дизајнирани су тако да једним тестом провере валидност неке сложене статистичке хипотезе при чему задржавају ниво значајности на циљаних 5%. Ова дисертација се ослања на два таква теста како би се осигурала поузданост добијених резултата. Први од њих биће омнибус варијанта Диеболд-Маријановог (скраћено ОДМ) теста коју дефинише сам аутор. Други коришћени омнибус тест је МекКракенов (2000) тест квалитета прогнозе. Он ће послужити као додатна провера добијених резултата будући да је и робуснији и флексибилнији тест од оригиналног Диеболд-Маријановог (1991) теста. Конкретно, МекКракенов (2000) тест може да функционише и као појединачни и као омнибус тест при чему се може базирати на различитим мерама квалитета прогнозе (попут: средње апсолутне грешке прогнозе, средње процентуалне апсолутне грешке прогнозе, средње квадратна грешка прогнозе и сл.). Са друге стране, до ове дисертације омнибус варијанта Диеболд-Маријановог (1991) теста није постојала, те се тек са њом флексибилност теста поправља. Истовремено, МекКракенов (2000) тест даје значајно прецизније резултате за мале узорке. Диеболд-Маријано (1991) тест, као најстарији статистички тест за проверу квалитета прогнозе, је јако осетљив на величину узорка[45]. То примену теста у економским истраживањима чини рањивом, будући да су економски узорци за предвиђање обично кратки[46]. Насупрот томе, МекКракенов (2000) тест је значајно робуснији у малим узорцима што га чини природно привлачним и за ово истраживање. Оба теста представићемо у наставку ове секције. Пре тога, зарад лакшег разумевања извођења расподеле ОДМ статистике теста, дисертација ће читаоца упознати са оригиналном верзијом Диеболд-Маријановог (1991) теста, али и са резултатима из постојеће литературе који ће се појавити као кораци током њеног извођења.

[45] Како тест даје изузетно непрецизне резултате на малим узорцима, мора се конструисати његова корекција за мале узорке како би се резултати макар у извесној мери поправили.

[46] Протоком времена економске околности се мењају. Последично, параметри модела се морају ревидирати тако да одражавају актуелну ситуацију. Ревидирањем модела се задржава његова прецизност. Према томе, не можемо исти модел користити за предвиђање кретање економских променљивих дуго у будућност.

3.9.3 Диеболд-Маријанов тест

За почетак, укратко представимо идеју оригиналне (тј. једнодимензионе) варијанте Диеболд-Маријановог (1991) теста. Разумевање идеје оригиналног теста ће читаоцу олакшати праћење излагања о омнибус варијанти теста коју предлаже аутор дисертације. Поступак тестирања започиње дефинисањем нове временске серије дате изразом испод:

$$rg_t = \left(r_{T+t} - \hat{r}_{I,\ T}(t)\right)^2 - \left(r_{T+t} - \hat{r}_{II,\ T}(t)\right)^2 = \hat{e}^2_{I,\ T+t} - \hat{e}^2_{II,\ T+t} \qquad (15)$$

где: rg_t означава новодобијену временску серију која представља разлику у квадратима грешке прогнозе два модела у тренутку $T + t$, а ознаке I и II означавају редом први и други модел који се тестом пореде.

Тестом се испитује валидност следећих хипотеза:

$$H_0\text{: } E(rg) = 0$$

$$H_1\text{: } E(rg) \neq 0$$

У циљу провере валидности постављених хипотеза, оцењује се регресија у којој се разлике квадрата грешака прогнозе (тј. rg_t) регресирају само на константу. Уколико је оцењена константа статистички значајно различита од нуле, то значи да постоји константна разлика у погледу квалитета прогнозе два модела у корист једног од њих. Алтернативно, уколико оцењена константа није статистички значајна, не постоји разлика у квалитету прогнозе два модела. Уколико је оцењена константа негативна, из израза (15) закључујемо да други модел прави веће грешке прогнозе, те је први бољи. У случају да је оцењена константа била позитивна закључак би био супротан. За тестирање се користи класичан т-тест.

Приметимо да је помоћна регресија на бази које се тест спроводи јако оскудна (временска серија се регресира само на константу). Последично, постоји велика вероватноћа да ће оцењену једначину одликовати присуство хетероскедастичности и/или аутокорелације, што за последицу има пристрасност стандардних грешака оцена. Проблем се разрешава употребом Њуи-Вестове (*Newey-West*) корекције (видети Њуи и Вест 1987a). Може се показати да t-статистика задржава t_{n-k} расподелу[47] и након примене Њуи-Вестове корекције стандардне грешке, где је n обим узорка, а k број оцењених параметра (видети Младеновић и Нојковић 2018 или Деметреску (*Demetrescu*) и сарадници 2015)[48]. Овај резултат биће важан чинилац извођења расподеле ОДМ статистике теста које следи у наредној секцији.

[47] У случају Диеболд-Маријановог (1991) теста у питању је t_{g-1} расподела будући да је узорак за тестирање пропорционалан хоризонту прогнозе, g, а да је оцењен само један параметар – константа.

[48] Приказани резултат не важи само за Њуи-Вестову корекцију већ за све корекције стандардних грешака зарад отклањања последица аутокорелације и хетероскедастичности (енгл. *HAC estimates*). Једно додатно својство које важи само за Њуи-Вестову корекцију доказали су сами аутори, а то је да се расподела t-статистике може апроксимарати стандардном нормалном расподелом ако однос проточности (енгл. *bandwidth*) узорка и обима узорка тежи нули, када обе ове вредности теже бесконачно (видети Њуи и Вест 1987b).

3.9.4 Омнибус Диеболд-Маријанов (ОДМ) тест

Омнибус варијанта теста поредиће квалитет прогнозе два модела на L узорака, уместо на једном. У складу са тиме, претпоставимо да је за сваки од L узорака израчуната средња квадратна грешка прогнозе, MSE, као агрегатна мера квалитета прогнозе на нивоу узорка. Ова мера представља квадрат корена из средње квадратне грешке прогнозе, *RMSE*, а њен обрачун дат је формулом испод:

$$MSE_l = \frac{1}{g}\sum_{t=1}^{g} \hat{e}^2_{T+t,\ l} = RMSE_l^2 \tag{16}$$

где је g хоризонт прогнозе.

Слично једнодимензионој варијанти теста, направимо нову серију података org_l[49]. Овога пута новодобијену серију дефинишемо као разлику средње квадратних грешака прогнозе два модела за различите узорке. Новодобијена серија дата је изразом (17). Приметимо да је овако добијена серија структурна, а не временска.

$$org_l = MSE_{I,\ l} - MSE_{II,\ l} \tag{17}$$

Алгоритам тестирања је исти као код оригиналног Диеболд-Маријановог (1991) теста. Новодобијена временска серија регресирала би се на константу. Уколико је константа у таквој регресији статистички значајна то би значило да један од два типа модела у већини узорака доминира над другим, будући да даје мање средње квадратне грешке прогнозе. У случају да константа није статистички значајна, закључили бисмо да су оба модела подједнако добра, јер се средње квадратне грешке прогнозе мало разликују. Значајност константе би се испитала класичним t-тестом уз Њуи и Вест (1987a) корекцију, док би се формално проверила валидност следећих хипотеза:

H_0: E(org) = 0

H_1: E(org) ≠ 0

Да би предложени омнибус Диеболд-Маријанов тест могао да се користи потребно је доказати да ће статистика теста имати *t* расподелу. То ће бити случај ако и само ако новодобијена серија org_l има нормалну расподелу. Да бисмо то доказали, осврнимо се на неке важне резултате из теорије статистичког закључивања. Најважнији међу њима је централна гранична теорема. У питању је заједнички назив за породицу резултата који показују да под одређеним условима расподела збира великог броја случајних променљивих конвергира ка нормалној расподели. Конкретно, овде наводимо један од њих:

[49] Назив долази од скраћеног омнибус *rg*.

Теорема 1 (Централна гранична теорема Хинчина): Нека је $\{X_1, X_2, \dots, X_n\}$ низ независних и идентично расподељених случајних променљивих са коначно варијансом чији је збир $S_n = \sum_{i=1}^{n} X_i$, тада расподела случајне променљиве Z, дефинисане као:

$$Z = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \tag{18}$$

конвергира у расподели ка стандардизованој нормалној расподели, $\mathcal{N}(0,1)$[50].

Последица ове теореме је да случајна променљива S_n има асимптотски $\mathcal{N}\big(E(S_n), Var(S_n)\big)$ расподелу. Другим речима, збир великог броја ($n \to \infty$) независних и идентично расподељених случајних променљивих имаће нормалну расподелу. Јасно је да резултат не зависи од полазне расподеле које имају све појединачне случајне променљиве у збиру.

Други важан резултат је лема о линеарној комбинацији нормално расподељених случајних променљивих. Овај резултат разматрамо у наставку:

Лема 1 (лема о линеарној комбинацији нормално расподељених случајних променљивих): Нека су редом $X_1, X_2, \dots, X_n$ нормално расподељене случајне променљиве, тада је случајна променљива $Y = \sum_{i=1}^{n} \alpha_i X_i$ $(\forall i, \alpha_i = const)$ нормално расподељена случајна променљива. Конкретно, $Y \sim \mathcal{N}(\sum_{i=1}^{n} \alpha_i \mu_i, \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2)$, где су μ_i и σ_i^2 редом очекивање и варијанса *i*-те случајне променљиве из линеарне комбинације[51].

Сада су стечени услови да изведемо расподелу ОДМ статистике теста. Извођење је дато као доказ теореме 2, која је дата у наставку.

Теорема 2 (ОДМ статистика теста): Нека је $\hat{c}$ оцена константе добијена методом обичних најмањих квадрата (енгл. *ordinary least squares*) у регресији структурне серије разлика средње квадратних грешака прогнозе два модела код L узорака само на константу. Тада количник оцене $\hat{c}$ и њене стандардне грешке има t_{L-1} расподелу.

Доказ

Уколико су класичне економетријске претпоставке испуњене, стандардне грешке прогнозе представљаће низ независних и идентично расподељених случајних променљивих (енгл. *i.i.d.*) са нормалном расподелом $\mathcal{N}(0, \sigma_{\text{гп}}^2)$, где је $\sigma_{\text{гп}}^2 < \infty$ варијанса грешака прогнозе (Младеновић и Петровић 2018). Последично, њихови квадрати ће такође представљати низ независних и идентично расподељених случајних променљивих. Према централној граничној теореми, збир великог броја квадрата грешака прогнозе имаће асиптотски нормалну расподелу. Из тога следи да ће и средње квадратне грешке, *MSE*, асиптотски имати нормалну расподелу.

[50] Видети Петровић (2015).
[51] Видети Петровић (2015).

Последица претходног закључка је да ће променљива $org_l = MSE_{I,l} - MSE_{II,l}$ представљати линеарну комбинацију две нормално расподељене случајне променљиве. Према леми о линеарној комбинацији нормално расподељених случајних променљивих, и сама променљива org_l имаће нормалну расподелу. Захваљујући томе моћи ћемо да применимо стандардно статистичко закључивање будући да је тиме осигурано да ће оцене параметара дате регресије имати нормалну расподелу. Покажимо сада чему ће бити једнака оцена константе по методу обичних најмањих квадрата. Пођимо од модела:

$$org_l = c + \varepsilon_l \tag{19}$$

Минимизирањем суме квадрата резидуала добијамо:

$$\min f\,(\varepsilon_l) = \sum_{l=1}^{L} \varepsilon_l^2 = \sum_{l=1}^{L} (org_l - c)^2$$

Из услова првог реда имамо:

$$\frac{\partial f}{\partial c} = -2 \sum_{l=1}^{L} (org_l - c) = 0$$

Израз изнад ће бити нула само ако је сума из њега једнака нули, односно:

$$\sum_{l=1}^{L} (org_l - c) = 0$$

$$\sum_{l=1}^{L} org_l - L \cdot c = 0$$

Одакле коначно следи да је оцена константе једнака просечној разлици средње квадратних грешака прогнозе:

$$\hat{c} = \frac{1}{L} \sum_{l=1}^{L} org_l = \overline{org} \tag{20}$$

Ослањајући се на лему о линеарној комбинацији нормално расподељених случајних променљивих, и на чињеницу да су разлике средње квадратних грешака прогнозе нормално расподељене, и сама оцена $\hat{c}$ ће бити нормално расподељена. Последично, када обрачунамо стандардизовано одступање, трансформисана статистика ће имати стандардну нормалну расподелу:

$$Z = \frac{\overline{org} - E(org)}{\sqrt{\frac{1}{L} V(org)}} \tag{21}$$

где су $E(org)$ и $V(org)$ редом очекивање и варијанса променљиве org_l.

За оцену варијансе $V(org)$ узмимо узорачку варијансу $S^2(org)$ дату изразом испод:

$$S^2(org) = \frac{1}{L-1}\sum_{l=1}^{L}(org_l - \overline{org})^2 \tag{22}$$

Без доказа узимамо добро познату чињеницу да ће количник узорачке варијансе помножен са бројем степени слободе и подељен правом вредношћу непознате варијансе имати χ^2_{L-1} расподелу (видети Петровић 2015 или Младеновић и Петровић 2018). Одатле следи да ће статистика дата изразом (23) имати t_{L-1} расподелу.

$$t = \frac{Z}{\sqrt{\frac{(L-1)S^2(org)}{V(org)}}} = \frac{\frac{\overline{org} - E(org)}{\sqrt{\frac{1}{L}V(org)}}}{\sqrt{\frac{S^2(org)}{V(org)}}} = \frac{\overline{org} - E(org)}{\sqrt{\frac{1}{L}S^2(org)}} \tag{23}$$

Како се нултом хипотезом претпоставља да је $E(org) = 0$, под претпоставком њене тачности статистике теста се своди на:

$$t = \frac{\overline{org}}{\sqrt{\frac{1}{L}S^2(org)}} \tag{24}$$

Искористимо раније поменут резултат по којем t статистика обрачуната на бази узорачке стандардне грешке коригованом за ефекте хетероскедастичности и аутокорелације задржава оригиналну расподелу. Тако долазимо до финалног облика ОДМ статистике теста који има t_{L-1} расподелу:

$$t = \frac{\overline{org}}{\sqrt{\frac{1}{L}S^2_{HAC}(org)}} \tag{25}$$

где је $S^2_{HAC}(org)$ узорачка стандардна девијација коригована за ефекте аутокорелације и хетероскедастичности (у овом истраживању она ће бити обрачуната по методологији Њуиа и Веста 1987а).

Будући да се тест ослања на алгоритам Диеболд-Маријана (1991), наследио је и његове слабости. Као што је већ истакнуто, слабост оригиналног теста оличена је кроз захтев за великим узорком за предвиђање. Другим речима, да би закључци били поуздани, потребно је да се обезбеди довољан број опсервација за конструкцију помоћне регресионе једначине. Слично томе предложени омнибус тест захтева да број узорака, L, буде велики како би поузданост била адекватна. То је ограничавајући фактор у практичним истраживањима, будући да није увек могуће обезбедити значајно велики број узорака. Тај недостатак је свакако надомешћен МекКракеновим (2000) тестом.

3.9.5 МекКракенов тест

Тест започиње избором мере квалитета прогнозе на којој ће се алгоритам тестирања базирати. У овом истраживању ослонићемо се на корен из средње квадратне грешке прогнозе. Одабир мере квалитета предодредиће изглед нове временске серије коју треба конструисати. Нова временска серија се конструише на бази трансформације грешака прогнозе која се користи за обрачун изабране мере квалитета. Примера ради, уколико се тест базира на средњој апсолутној грешци биће нам потребне апсолутне вредности грешака прогнозе. Са друге стране, у случају корена из средње квадратне грешке прогнозе биће нам неопходни квадрати грешака прогнозе. Нова временска серија, $\not{f}_t$, представљаће разлике између одабраних трансформација грешака прогнозе два модела (исто као rg_t код Диеболд-Маријано 1991 теста):

$$\not{f}_t = \left(r_{T+t} - \hat{r}_{T\ I}(t)\right)^2 - \left(r_{T+t} - \hat{r}_{T\ II}(t)\right)^2$$

$$\not{f}_t = \hat{e}^2_{I,\ T+t} - \hat{e}^2_{II,\ T+t} \tag{26}$$

Будући да желимо да добијемо генерални закључак и спроведемо омнибус тест потребно је да обрачунамо нову временску серију $\not{f}_t$ за сваки узорак (тј. за сваку крипто-валуту). У том случају добићемо векторску временску серију f_t.

$$f_t = (\not{f}_{t1} \quad \not{f}_{t2} \quad \cdots \quad \not{f}_{tL}) \tag{27}$$

У развијеном облику векторска временска серија f_t представља матрицу чији су елементи разлике трансформација (тј. у овом случају квадрата) грешака прогнозе два модела. За произвољан хоризонт прогнозе t $(1 \leq t \leq g)$, матрица ће бити димензија $t \times L$,:

$$f_t = \begin{pmatrix} \not{f}_{11} & \not{f}_{12} & \cdots & \not{f}_{1L} \\ \not{f}_{21} & \not{f}_{22} & \cdots & \not{f}_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \not{f}_{t1} & \not{f}_{t2} & \cdots & \not{f}_{tL} \end{pmatrix} \tag{28}$$

Циљ теста је да проверимо да ли у свих L случајева модели раде подједнако добро или један од њих доминира над другим. Технички говорећи, проверавамо да ли је корен из средње квадратне грешке једног од модела мањи од истог показатеља код другог модела у свих L случајева. Из тог разлога постављамо следећи пар хипотеза:

$$\mathrm{H}_0\text{: } \mathrm{E}(f_t) = \mathbf{0}$$

$$\mathrm{H}_1\text{: } \mathrm{E}(f_t) \neq \mathbf{0}$$

Нулта хипотеза тврди да је очекивана вредност векторске временске серије f_t нула вектор, односно, да је очекивана вредност сваке временске серије $\not{f}_{tl}$ нула. Уколико је нулта хипотеза тачна, то би значило да су квадрати (односно, у општем случају трансформације) грешака прогнозе два модела једнаки. То нас наводи на закључак да оба модела дају приближно једнаке грешке прогнозе, односно, да су подједнако квалитетна. Са друге стране, алтернативна

хипотеза би нас наводила на супротан закључак, будући да у том случају очекивана вредност векторске временске серије f_t није нула вектор. Да одредимо који модел даје боље резултате у том случају посматрамо знак елемената у вектору математичког очекивања. Уколико су његови елементи негативни, први модел је супериорнији јер он даје мање грешке. Закључак проистиче из чињенице да су временске серије $\not{f}_t$ дефинисане као разлике трансформација грешака прогнозе првог и другог модела. Уколико су њихове вредности негативне, то значи да су умањеници (грешака првог модела) мањи од умањилаца (грешака другог модела). Уколико су елементи вектора математичког очекивања позитивне вредности, иста логика би нас навела на закључак да је други модел супериорнији премда су његове грешке мање. Оцена очекиване вредности векторске временске серије f_t (у ознаци $\bar{f}$) се обрачунава на следећи начин:

$$\bar{f} = \frac{1}{g}\sum_{t=1}^{g} f_t \tag{29}$$

Еквивалентан приступ за утврђивање супериорнијег модела након тестирања био би да се погледају разлике између тачкастих оцена корена из средњих квадратних грешака прогнозе. Разуме се, модел са мањим вредностима корена из средњих квадратних грешка прогнозе је супериорнији. Закључци ће бити исти као у претходном случају.

Статистика теста, χ, коју МекКракен (2000) предлаже дата је изразом (30):

$$\chi = \frac{1}{\sqrt{g}}\Omega^{-0,5}\sum_{t=1}^{g} f_t \tag{30}$$

где је Ω коваријациона матрица која се одређује на посебан начин.

Статистика теста χ има L-димензиону стандардну нормалну расподелу, $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$[52]. Како би се спровело тестирање, статистика теста се трансформише у секуларни облик који има χ_L^2 расподелу[53]:

$$\chi_L^2 = \chi'\chi \tag{31}$$

Да би се осигурало да статистика теста има назначену расподелу неопходно је: да су испуњене претпоставке класичног линеарног регресионог модела, да је временска серија чије се предвиђање врши стационарна и да је узорак за оцену модела значајно већи или пропорционалан хоризонту прогнозе (тј. важи једна од следећих једнакости $\lim_{T,g\to\infty}\frac{g}{T} = 0$ или $\lim_{T,g\to\infty}\frac{g}{T} = const.$).

[52] Где је **I** јединична матрица димензије $L \times L$.

[53] Будући да се добија као збир квадрата L независних случајних променљивих са стандардном нормалном расподелом.

Поступак закључивања је веома једноставан. Уколико је реализована вредност статистике теста већа од критичне, прихватићемо алтернативну хипотезу и закључити да један од модела ради боље.

Најтежи део поступка тестирања је одредити адекватну коваријациону матрицу Ω. Њен облик зависи од избора мере квалитета прогнозе на којој ће тестирање бити базирано. То ставља додатну тежину на ову одлуку, будући да поједине мере квалитета условљавају да матрица Ω има изузетно компликован облик. У случају корена из средње квадратне грешке прогнозе матрица Ω се добија као:

$$\Omega = \sum_{j=-(g-1)}^{g-1} \text{К(j)}\Gamma_{ff}(j) \tag{32}$$

где су: $\Gamma_{ff}(j)$ аутоковаријационе матрице векторске временске серије f_t на доцњи j, а К(j) прозор за пондерисање (енгл. *kernel*) на доцњи j.

МекКракенов (2000) дозвољава истраживачу употребу произвољног прозора. Такође, допуштен је и изостанак прозора уколико желимо да се свим аутоковаријационим матрицама да исти значај. За потребе овог истраживања коришћен је најпопуларнији статистички прозор, Бартлетов прозор:

$$\text{К}_\text{B}(j) = 1 - \frac{j}{g+1} \tag{33}$$

Подсетимо читаоца и на дефиницију аутоковаријациону матрицу векторске временске серије:

$$\Gamma_{ff}(j) = \frac{1}{g}\sum_{t=j}^{g}(f_t - \bar{f})'(f_{t-j} - \bar{f}) \tag{34}$$

Како су све матрице $\Gamma_{ff}(j)$ димензија $L \times L$ и матрица Ω ће имаће исте димензије. Захваљујући начину на који је конструисана у изразу (32), матрица Ω ће бити симетирчна матрица иако $\Gamma_{ff}(j)$, за $j > 0$ није симетрична матрица (видети МекКакен 2000).

3.10 Кластеризација методом *K*-средњих вредности

Кластеризација методом *K*-средњих вредности (енгл. *K-means cluster analysis*) је техника машинског учења и статистике која разврстава опсервације у унапред одређени број група. Овде је реч о специфичној врсти група које називамо кластерима. Кластери представљају групе које су унутар себе хомогене, а између себе хетерогене. Другим речима, све опсервације сврстане у исти кластер биће међусобно сличне, а значајно ће се разликовати од опсервација из других кластера. Опсервације се разврставају у групе према њиховој сличности. Одлике група нису унапред познате, већ само њихов број. Жељени број група типично се означава са

K. Анализа започиње идентификацијом *K* најразличитијих опсервација. Њихова различитост манифестоваће се кроз њихову удаљеност у *N*-димензионом простору. Аналитичар сам одређује коју меру удаљености жели да користи. У пракси се за то најчешће користи Махаланобисово одстојање. *K* идентификованих опсервација чиниће почетних *K* кластера. У сваком наредном кораку, рачуна се удаљеност свих опсервација од центра постојећих кластера. Центре кластера називамо центроидима. Свака опсервација се придружује оном кластеру чијем центроиду је најближа. Након тога, поново се израчунавају центри кластера и удаљеност свих опсервација од њих. Збир удаљености сваке опсервације од центра кластера којем припада назива се укупном удаљеношћу. Уколико је могуће смањити укупну удаљеност премештањем опсервација између кластера алгоритам наставља са радом, тј. претходни поступак се понавља. Алгоритам ће стати са радом тек онда када не постоји начин да се укупна удаљеност смањи даљим размештањем (због чега се задаје минимално прихватљива разлика као критеријум конвергенције). Једном када се кластери формирају, особине сваког од њих разазнајемо анализом центроида. Предности оваквог приступа у кластерисању је брзина израчунавања и мања рачунарска захтевност, што га чини идеалним за велике базе података. Поред тога, захваљујући итеративној структури алгоритма опсервације које су биле погрешно класификоване на почетку, могу касније бити премештене у исправан кластер, чиме се смањује могућност грешке и повећава стабилност решења. Главна мана кластеризације методом *K*-средњих вредности је априори одређивање броја кластера, за разлику од хијерархијског кластерисања.

За добијање квалитетних резултата важан је правилан одабир броја кластера, односно, одређивање оптималне вредности за *K*. Постоји неколико могућих приступа решавању овог проблема. Најједноставнији приступ се базира на експертизи истраживача, доменском знању и потребама истраживања. Када истраживач у току експлоративне анализе сагледа податке, на бази свог искуства може претпоставити колики би број кластера био одговарајући. Слично, уколико на бази доменског знања зна да се подаци из одређене области типично групишу у одређени број група, може ту вредност поставити за циљни број кластера. Алтернативно, број кластера може наметнути само истраживање уколико је из практичних разлога пожељно да постоји одређени број кластера. Други приступ би подразумевао експериментисање са различитим бројем кластера или спровођење крос-валидације. Коначно, трећи приступ подразумева анализу дијаграм одрона (енгл. *scree plot*) тј. дијаграм лакта (енгл. *elbow chart*). Реч је о графичком приказу који ставља у однос просечну удаљеност опсервација од центара кластера којима припадају и број задатих кластера. Што је број задатих кластера већи то су конструисани кластери хомогеније скупине. Ипак, нема додавање сваког новог кластера исти значај. У почетку, када је број кластера мали, додавање сваког новог кластера значајно доприноси подизању ниво хомогености кластера. Касније допринос сваког новог кластера постаје све мањи и мањи. Дијаграм одрона/лакта нам омогућава да графичким путем анализирамо однос хомогености и броја кластера. У тачки после које допринос нових кластера постаје исувише мали налази се оптималан број кластера. Принцип доношења одлуке се може објаснити и путем визуелне аналогије са људском руком. Дијаграм одрона/лакта има облик руке савијене у лакту при чему се формира туп угао. У тачки која представља лакат, тј. теме тупог угла, налази се оптималан број кластера, *K*.

Кластеризација методом *K*-средњих вредности биће искоришћена за груписање прикупљених текстова у кластере текстуалних категорија према њиховој сличности. Категорије текста биће

моделиране као вештачке променљиве додате у модел (12). На тај начин испитаће се да ли писање одређених врста вести (текстова) може бити индикатор будућег кретања приноса. То је релевантно питање будући да се на тему крипто-валута објављују текстови најразноврснијих природа, као што је дискутовано у одељку **1.5**. Добијени резултати дискутоваће се у одељку **5.4.2**.

3.11 Хипотеза о слабој форми ефикасности тржишта и њена провера

Предвидљивост тржишних промена била би сигнал потенцијалне тржишне неефикасности. Из тог разлога се ово истраживање завршава управо формалном провером ефикасности тржишта осам одабраних крипто-валута. Ефикасност тржишта гарантовала би објективност вредновања активе којом се на њему тргује, али и њену правилну алокацију. Постојање ефикасности говорило би у прилог тржишног здравља и представљало би доказ да се тржишном механизму може веровати. Импликације овог закључка су и то да је фундаментална анализа беспотребна (јер су све цене фер), да је техничка анализа неупотребљива, а да је једина логична инвестициона стратегија пасивно инвестирање (праћење тржишта). Према томе, испитивање тржишне ефикасности је од великог значаја и са финансијског и са микроекономског становишта.

Хипотеза о нивоу ефикасности финансијских тржишта има три појавна облика. Први облик назива се хипотезом о слабој форми ефикасности тржишта. Према овом облику, тржиште се сматра ефикасним уколико све цене одражавају информације о својим ранијим кретањима. Хипотеза о полу-јакој форми ефикасности тржишта тврди да поред информација о ранијим кретањима цена, цене одражавају и све остале јавно доступне информације. Коначно, хипотеза о јакој форми ефикасности тржишта тврди да цене одражавају све могуће информације, па чак и оне које нису јавно доступне, односно инсајдерске информације.

Ово истраживање формално ће се позабавити питањем слабе форме тржишне ефикасности. Уколико текуће цене у себи садрже информације о свим претходним ценама онда се њихово кретање може описати као случајан ход (енгл. *random walk*). У питању је случајан процес који трајно памти све шокове из прошлости (информације о претходним променама су већ укључене у цену) и чије се промене дешавају на случајан (непредвидив) начин. Математички се случајан ход може записати као прост ауторегресивни модел првог реда чији је аутокорелациони коефицијент једнак јединици:

$$x_t = x_{t-1} + e_t \tag{35}$$

где је x_t процес случајан ход, а e_t процес бели шум.

Како бисмо сагледали претходно истакнута два својства случајног хода запишимо овај процес на алтернативни начин. Рекурзивном заменом израза (35) случајан ход се може представити као збир свих претходних шокова реализованих до тренутка t:

$$x_t = x_0 + \sum_{s=1}^{t} e_s \tag{36}$$

где је x_0 почетна вредност процеса.

Ослањајући се на израз (36) једноставно се може показати да су прве диференце (прирасти или инкременти) случајног хода пропорционалне белом шуму, што промене овог процеса чини случајним:

$$dx_t = x_t - x_{t-1} = e_t \tag{37}$$

где је dx_t ознака прве диференце.

У финансијским истраживањима типично се уместо цена посматрају њихови логаритми (видети: Ло и Мекинли 1988 и Младеновић и Нојковић 2018). Та одлука ће олакшати даље истраживање, будући да ће у том случају прва диференца анализиране временске серије (логаритмованих цена) бити управо једнака логаритамским приносима, описаним у секцији **3.6**. Последично, уколико је тржиште ефикасно цене се морају понашати као случајан ход, а приноси као бели шум. То је основна премиса од које полазе сви тестови слабе форме тржишне ефикасности. Статистички тестови који су у те сврхе коришћени током овог истраживања представљени су у наставку овог одељка.

Ипак, не треба сметнути с ума да резултати добијени у пређашње описаним етапама истраживања могу представљати сигнал потенцијалне полу-јаке форме тржишне ефикасности коју, такође, треба проверити. Уколико полу-јака форма тржишне ефикасности постоји не би било могуће предвиђати кретање цена на бази јавно доступних информација, будући да су оне већ укључене у цену. Последично, веза између цена и индикатора добијених из вести не би смела да постоји. У овом истраживању, јавно доступне информације о крипто-валутама прикупљене су као вести са интернет портала. Реч је о најсвеобухватнијем извору информација, јер интернет портали преносе и вести иницијално објављене у другим изворима. Захваљујући томе, посредно је обухваћен и утицај осталих медија, што интернет портале издваја као добар прокси за јавно доступне информације. Анализирајући везу интернет вести и приноса, као и кроз испитивање могућности предвиђања будућих кретања приноса на бази дате везе, долазимо до сигнала потенцијалних тржишних неефикасности. Да бисмо формално могли да тврдимо да овај вид ефикасности постоји потребно је проверити да ли се може формирати стратегија трговања која ће донети вишак приноса у односу на преузети ризик. Тим питањем ће се позабавити посебно истраживање.

3.11.1 Проширени Дики-Фулеров тест

Једноставан начин да се провери да ли је нека временска серија случајан ход или не јесте да се оцени као ауторегресивни модел првог реда:

$$x_t = \phi_1 x_{t-1} + e_t \tag{38}$$

где је ϕ_1 ауторегресивни коефицијент првог реда.

Уколико је посматрана временска серија заиста случајан ход онда ће оцењени ауторегресивни коефицијент бити приближно једнак јединици. Према томе, тестирајући да ли је $\phi_1 = 1$ истовремено проверавамо да ли је посматрана временска серија случајан ход или не[54]:

H_0: $\phi_1 = 1$ (серија је случајан ход)

H_1: : $\phi_1 \neq 1$ (серија није случајан ход)

У економетријској пракси преферира се да се тестира да ли је вредност оцењеног параметра нула или не. Из тог разлога се модел (38) типично замењује моделом (39). Модел (39) се добија од модела (38) тако што се обема странама одузме прва доцња посматране временске серије:

$$dx_t = \varphi_1 x_{t-1} + e_t \tag{39}$$

где је $\varphi_1 = \phi_1 - 1$.

Статистика теста се формира на уобичајен начин као *t*-однос параметра φ_1 и његове стандардне грешке. Међутим, статистика теста неће имати стандардну *t* расподелу, будући да се у случају испуњености нулте хипотезе губе пожељна статистичка својства временске серије (проблем нестационарности). Уместо тога *t*-однос ће имати модификацију *t* расподеле асиметричну улево, названу по ауторима који су је први пронашли Дики-Фулерова расподела (енгл. *Dickey-Fuller's distribution*)[55]. Будући да је реч о нестандардној расподели њен облик није једнозначан, већ зависи од величине узорка и детерминистичких компоненти садржаних у моделу. То ће се последично одразити на критичне вредности теста. Из тог разлога, пре започињања тестирања потребно је проверити које детерминистичке компоненте треба укључити у модел (39). За то се типично користи Сток-Вотсонов (*Stock-Watson*) (1989) тест. Сток-Вотсонов тест регресира прву диференцу посматране временске серије само на константу[56]. Уколико је константа статистички значајна то значи да модел (39) мора да

[54] Из тог разлога тестови који проверавају да ли је временска серија случајан ход обично носе назив тестови јединичног корена.
[55] Дики-Фулер (1979).
[56] Слично тесту Диеболда-Маријана (1991) и овде је реч о сиромашној регресији коју типично карактерише присуство аутокорелације и/или хетероскедастичности, те се *t*-однос формира уз Њуи-Вестову корекцију.

садржи тренд[57]. Уколико то није случај, модел (39) може да садржи само константу. Одлука о томе да ли ће је садржати или не доноси се графичком анализом временске серије[58].

Поред детерминистичких компоненти, тест је осетљив и на присуство аутокорелације. Из тог разлога у модел (39) додаје се онолико доцњи прве диференце посматране временске серије колико је неопходно да корелограм[59] резидуала из модела (39) изгледа као корелограм белог шума[60]. Финални запис модела у том случају изгледа овако:

$$dx_t = \varphi_1 x_{t-1} + \sum_{i=1}^{p} \delta_i dx_{t-i} + e_t \qquad (40)$$

где је p број доцњи неопходан за отклањање аутокорелације, а δ_i регресиони параметри уз доцње прве диференце.

Будући да се полазни модел проширује доцњама прве диференце зарад отклањања аутокорелације, финална верзија Дики-Фулеровог теста названа је проширени (енгл. *augmented*) Дики-Фулеров тест. Сходно томе, *t*-однос параметра φ_1 и његове стандардне грешке се типично означава као *ADF*(p) статистика, где је p број доцњи неопходан за отклањање аутокорелације из модела (40)[61].

Уколико се проширеним Дики-Фулеровим тестом покаже да временска серија логаритмованих цена прати случајан ход тржиште се може сматрати слабо ефикасним. У супротном, можемо закључити да тржиште није слабо ефикасно.

3.11.2 КПСС тест

Добро је позната емпиријска чињеница да стандардни тестови јединичног корена не успевају да одбаце нулту хипотезу о јединичном корену за многе економске временске серије (видети Квиатовски (*Kwiatkowski*) и сарадници 1992). Да би се тај проблем ублажио Квиатовски и сарадници (1992) предлажу две модификације класичних тестова. Прва модификација се

[57] Константа у првој дифренеци може бити присутна ако и само ако серија у нивоу има линеарни детерминиситчки тренд.
[58] Провером да ли се серија креће око нулте или ненулте средње вредности. Ненулта средња вредност сугерисала би да је исправна спецификација модела за тестирање она која садржи константу.
[59] Корелограм је важан алат у анализи временских серија. У питању је графички приказ коефицијената обичне и парцијалне аутокорелационе функције. На основу њега добијамо наговештаје за адекватну спецификацију модела, проверавамо стабилност модела (одсуство ауткорелације резидуала), али и контролишемо тестове јединичног корена. Када је реч о последњој намени, добро је позната чињеница да корелограм случајног хода мора изгледати тако да је значајан само први парцијални ауткорелациони коефицијент (док је његова вредност висока и блиска јединици) при чему су сви обични аутокорелациони коефицијенти такође значајни, а њихов ниво опада изразито споро (за више детаља видети Младеновић и Нојковић 2018).
[60] Будући да је бели шум неаутокорелисани процес све доцње обичне и парцијалне аутокорелацион функције морају бити статистички незначајне.
[61] За више детаља о поступку тестирања погледати Младеновић и Нојковић (2018).

односи на замену редоследа хипотеза. Начин на који се спроводи класично статистичко тестирање хипотеза осигурава да се нулта хипотеза прихвати осим када нема јаких доказа против ње. Из тог разлога су класични тестови јединичног корена (који нестационарност тврде по нултој хипотези) склони да прогласе серију за случајан ход, чак и када она то није. Друга модификација се односи на промену самог приступа тестирању. Конвенционални тестови јединичног корена су углавном модификације проширеног Дики-Фулеровог теста, те као и он стационарност испитују преко ауторегресивног коефицијента. Уместо тога, Квиатовски и сарадници (1992) предлажу другачију перспективу истог проблема. Како бисмо објаснили поступак тестирања који предлажу аутори претпоставимо најпре да је посматрана временска серија случајан ход. Посматрајмо регресију исте само на детерминистичке компоненте (константу и тренд):

$$x_t = c + b \cdot t + e_t^* \tag{41}$$

где је x_t анализирана временска серија, c константа, b регресиони параметар уз тренд, док су e_t^* резидуали модела.

Под таквим околностима целокупно стохастичко понашање пренеће се на резидуале модела, те ће они бити случајан ход:

$$e_t^* = e_{t-1}^* + v_t \tag{42}$$

где је v_t случајна компонента резидуала која прати бели шум и чија је варијанса σ_v^2.

Претпостављајући да је почетна вредност резидуала $e_0^* = 0$ и ослањајући се на израз (36) можемо показати да се резидуали разматраног модела могу представити као парцијална сума шокова случајне компоненте v_t:

$$e_t^* = \sum_{s=1}^{t} v_s \tag{43}$$

Из претходног записа једноставно је одредити вредности очекивања и варијансе коју би имали резидуали да је посматрана временска серија случајан ход. Ослањајући се на особине математичког очекивања и варијансе, као и на својства процеса бели шум, добијамо:

$$E(e_t^*) = \sum_{s=1}^{t} E(v_s) = 0 \tag{44}$$

и

$$V(e_t^*) = \sum_{s=1}^{t} V(v_s) = t\sigma_v^2 \tag{45}$$

Релација (45) је од великог значаја јер нам даје везу која ће нам послужити да оценимо варијансу случајне компоненте v_t:

$$\sigma_v^2 = \frac{1}{\mathrm{T}} V(e_t^*) \tag{46}$$

Ако је варијанса случајне компоненте v_t једнака 0 тада v_t не би била случајан процес већ константа једнака нули. То би било могуће ако и само ако посматрана временска серија, x_t , није случајан ход. Пошавши од ове идеје, Квиатовски и сарадници (1992) постављају следећи пар хипотеза:

H_0: $\sigma_v^2 = 0$ (серија није случајан ход)

H_1: : $\sigma_v^2 > 0$ (серија је случајан ход)

Уколико је нулта хипотеза тачна, резидуали из модела (41) такође не би били случајан ход, те бисмо имали:

$$x_t = c + b \cdot t + e_t \tag{47}$$

где су e_t резидуали који не садрже јединични корен.

Да бисмо проверили нулту хипотезу искористићемо оцењене резидуале без јединичног корена, e_t, да оценимо варијансу случајне компоненте v_t. За то ћемо употребити релацију (46). Варијанса случајне компоненте v_t је *T* пута мања од варијансе резидуала модела који у себи има јединични корен, $V(e_t^*)$. Као оцену варијансе модела који у себи има јединични корен искористићемо релацију (43). Према њој резидуали који садрже јединичи корен се могу представити као низ парцијалних сума шокова белог шума. Као апроксимацију за поменуте шокове искористићемо оцене резидуала из модела (47). Тада се до оцена резидуала модела који садржи јединични корен може доћи на следећи начин:

$$\widehat{e_t^*} = \sum_{s=1}^{t} \hat{e}_s \tag{48}$$

где су $\hat{e}_s$ оцене резидуала из полазног модела (47), а $\widehat{e_t^*}$ оцене резидуала које у себи садрже случајан корен.

Варијанса резидуала који садрже јединични корен у том случају била би:

$$\hat{V}(e_t^*) = \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{e_t^*} - E(e_t^*)\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{s=1}^{t}\hat{e}_s - 0\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{s=1}^{t}\hat{e}_s\right)^2 \quad (49)$$

Из релација (49) и (46) добијамо оцену варијансе случајне компоненте v_t:

$$\hat{\sigma}_v^2 = \frac{1}{T}\hat{V}(e_t^*) = \frac{1}{T^2}\sum_{t=1}^{T}\left(\sum_{s=1}^{t}\hat{e}_s\right)^2 \quad (50)$$

Да тестирамо да ли је варијанса случајне компоненте v_t довољно мала да би се сматрала нулом упоредићемо је са варијансом модела (47)[62], $\hat{\sigma}_e^2$. Тако добијамо статистику теста:

$$KPSS = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_e^2} \quad (51)$$

Уколико је КПСС статистика теста довољно близу 0, закључили бисмо да је варијанса случајно компоненте v_t 0. То директно имплицира да посматрана временска серија није случајан ход. КПСС статистика теста нема стандардну расподелу, а њен облик зависи од детерминистичких компоненти садржаних у полазном моделу. За одређивање детерминистичких компоненти које модел треба да садржи такође се користи Сток-Вотсонов (1989) тест.

Уколико се КПСС тестом покаже да временска серија логаритмованих цена прати случајан ход тржиште се може сматрати слабо ефикасним. У супротном, можемо закључити да тржиште није слабо ефикасно.

3.11.3 Бартелсов тест случајности

Последица слабе форме тржишне ефикасности је да се цене понашају као случајан ход, док се приноси понашају као бели шум. Претходним тестовима испитивано је само понашање цена. Међутим, до важних закључака о тржишној ефикасности се може доћи и анализом приноса. Један могући приступ је провера њихове случајности. Као што је већ истакнуто, слаба форма тржишне ефикасности подразумева да се цене мењају на случајан начин. То би значило да приноси, односно прирасти логаритмованих цена, представљају случајну секвенцу бројева. Тестирајући да ли су приноси случајна секвенца уједно испитујемо и тржишну ефикасност.

[62] Како је реч о сиромашном моделу (будући дарегресирамо временску серију на две детерминистичке компоненте) коју типично карактерише присуство аутокорелације и/или хетероскедастичности, оцена варијансе модела биће коригована Њуи-Вестовом корекцијом како би се евентуална појава поменутих проблема отклонила.

Постоји више различитих тестова случајности. Аутор дисертације се определио за непараметарску верзију Фон Нојмановог (*von Neumann*) (1941) теста коју предлаже Бартелс (*Bartels*) (1982). Разлог за овакву одлуку крије се у чињеници да је Монте Карло симулацијама показано да Бартелсов тест случајности има далеку већу моћ за испитивање случајности од тестова заснованих на корацима (видети Бартелс 1982). Са друге стране, Бартелсов тест је непараметарски тест и не зависи од претпоставке о нормалности попут Фон Нојмановог (1941) теста.

Фон Нојман је показао да уколико нема систематских правилности у кретању одређене секвенце бројева просек квадрата њене прве диференце биће приближно два пута већи од просечног квадратног одступања. Ослањајући се на ову идеју, Бартелс у свом тесту испитује да ли су рангови временске серије поређани на случајан начин стављајући у однос просечан квадрат прве диференце рангова и варијансу рангова:

$$RVN = \frac{\sum_{t=2}^{T}(R_t - R_{t-1})^2}{\sum_{t=1}^{T}(R_t - \bar{R})^2} \tag{52}$$

где је R_t ранг опсервације реализоване у тренутку t, а $\bar{R}$ просечан ранг.

Бартелс (1982) је показао да статистика теста има нестандардну[63] бета расподелу чија су оба параметра једнака следећем изразу:

$$a = b = \frac{5T(T+1)(T-1)^2}{2(T-2)(5T^2 - 2T - 9)} - \frac{1}{2} \tag{53}$$

где су a и b параметри бета расподеле, а T обим узорка.

За обим узорка већи или једнак 100, Бартелс (1982) показује да се расподела статистике теста може добро апроксимирати нормалном расподелом облика: $\mathcal{N}\left(2, \frac{20}{5T+7}\right)$.

Тест проверава следећи пар хипотеза:

H_0: $RVN = 2$ (серија представља секвенцу случајних бројева)

H_1: : $RVN \neq 2$ (серија није секвенца случајних бројева)

Уколико је RVN статистика теста статистички значајно мања од 2 серија испољава тренд у свом кретању. Уколико је RVN статистика теста статистички значајно већа од 2 серија испољава систематско осцилирање. Оба поменута случаја индиковала би да серија приноса

[63] Не стандардно расподеле се односи на интервал вредности који случајна променљива *RVN* може узети. Будући да *RVN* узима вредности из интервала [0,4] да би се њена расподела свела на стандардну бета расподелу потребно ју је поделити са 4.

није случајна секвенца, већ да је одликује једно од поменута два систематска понашања. Присуство било којег од њих нарушило би слабу форму тржишне ефикасности.

3.11.4 Анализа случајног процеса приноса

Многи аутори испитују слабу форму ефикасности тржишта анализирајући временску серију приноса. Њихов циљ је да покажу да аутокорелациона структура приноса није она коју треба да поседује бели шум. То се углавном остварује на два начина: тестовима аутокорелисаности приноса и показивањем да приноси следе случајан процес који није бели шум. У овој дисертацији ослонићемо се на други поменути приступ. За приносе добијене из серија логаритмованих цена за које је показано да прате случајан ход проверићемо да ли се могу описати као *ARMA*(p,q) модел. Модел је дат изразом испод:

$$x_t = \phi_0 + \sum_{i=1}^{p} \phi_i x_{t-i} + e_t - \sum_{j=1}^{q} \theta_j e_{t-j} \tag{54}$$

где су ϕ_i ауторегресивни параметри, θ_j параметри покретних просека, e_{t-j} помаци резидуала, а ϕ_0 слободни члан. Вредности p и q редом представљају ред ауторегресивне компоненте, односно, компоненте покретних просека.

Да би се модел сматрао адекватним да опише понашање посматране временске серије потребно је да се задовоље два критеријума. Први критеријум је статистичка значајност свих параметра присутних у моделу. Други критеријум захтева подударност резидуала и белог шума. Другим речима, неопходно је да се на крају процеса моделирања добију резидуали који су неаутокорелисани и нормално расподељени. Прво од два захтевана својства резидуала нам гарантује да нема необјашњене аутокорелације (модел у потпуности објашњава кретање серије), док нам друго гарантује да ће статистички тестови имати теоријске расподеле. Као доказ ових особина биће приложени корелограми са пратећим Бокс-Љунговим (*Box-Ljung*) (1978) статистикама и вредности Жак-Бера (*Jarque-Bera*) (1980) статистике. Бокс-Љунгова статистика служи за тестирање присуства аутокорелације реда *k*:

$$\text{H}_0\text{: } \rho_1 = \rho_2 = \cdots = \rho_k = 0$$

$$\text{H}_1\text{: : } \exists i \leq k, \rho_i \neq 0$$

Бокс-Љунгова Q(к) статистика теста има χ^2_{k-p-q} расподелу (будући да сви аутокорелациони коефицијенти теоријски имају $\rho_i \sim \mathcal{N}\left(0, \frac{1}{\text{T}}\right)$ расподелу), где су p и q редови ауторегресивне, односно, копоненте покрентих просека оцењеног модела. Статистика је дата изразом испод:

$$Q(k) = T(T+2) \sum_{i=1}^{k} \frac{\rho_i^2}{T-i} \tag{55}$$

где су ρ_i аутокорелациони коефицијенти резидуала, *k* ред аутокорелације који се испитује, а *T* обим узорка.

Жак-Бера тест се користи за испитивање нормалности расподеле резидуала:

$$\text{H}_0\text{: } e_t \sim \mathcal{N}(0, \sigma^2)$$

$$\text{H}_1\text{: : } e_t \nsim \mathcal{N}(0, \sigma^2)$$

Основна идеја од које тест полази јесте да нормална расподела има специфичан облик звона који карактерише коефицијент асиметрије $\alpha_3 = 0$ и коефицијент спљоштености $\alpha_4 = 3$. Тестирајући да ли емпиријска расподела резидуала поседује ове вредности коефицијената асиметрије и спљоштености фактички проверавамо да ли је она нормална или не. Како су коефицијенти асиметрије и спљоштености и сами нормално расподељени ($\alpha_3 \sim \mathcal{N}\left(0, \frac{6}{T}\right)$ и $\alpha_4 \sim \mathcal{N}\left(3, \frac{24}{T}\right)$) може се показати да ће *JB* статистика теста (56) имати χ_2^2 расподелу.

$$JB = \frac{T}{6}\left(\alpha_3^2 + \frac{(\alpha_4 - 3)^2}{4}\right) \tag{56}$$

Како расподела статистике теста не зависи од спецификације модела ни обима узорка критична вредност која би сугерисала значајно одступање од нормалне расподеле је увек 5.99. Из тог разлога типично се реализована статистика теста наводи без пратеће п-вредности.

Поред наведеног, позабавићемо се и серијама логаритмованих цена за које се показало да не прате случајан ход. За такве серије рад ће покушати да пронађе *ARMA*(p,q) репрезентацију њиховог нивоа. То ће бити још јачи доказ тржишне неефикасности.

4. Модификација оцена Џагадиша и Вуа

Четврта секција посвећена је анализи сентимента. У њој ће прво бити представљен оригинални модел за мерење сентимента Џагадиша и Вуа (2019). Као што ће читалац имати прилику да види главни недостатак овог приступа је што се сентимент појединачних речи не може оценити директно из постављеног модела. Уместо тога, аутори предлажу модификацију која ће у себи садржати грешку мерења. Инспирисана тим проблемом и Јохансеновом (1996) процедуром за оцену параметара из коинтегрисаног *VAR* модела, дисертација ће покушати да пронађе алтернативни начин њиховог оцењивања. Поменути приступ биће изложен у наставку овог поглавља.

4.1 Модел Џагадиша и Вуа

За разлику од конвенционалних приступа у оцени сентимента представљених у прегледу литературе, Џагадиш и Ву (2019) предложили су квантитативни метод његове оцене заснован на линеарној регресији. Велика предност оваквог приступа оцењивању сентимента је то што се узима у обзир разлика у тежини изречених речи из исте групе (позитивних, негативних и неутралних), односно, магнитуда сентимента. Да би то постигли, аутори полазе од појма сентимента текста. Према Џагадишу и Вуу (2019) сентимент текста се дефинише као пондерисани збир релативних фреквенција сваке речи у тексту. То се квантитативно може записати на следећи начин:

$$S_{i,t} = \sum_{j=1}^{J} w_j \frac{f_{i,j,t}}{a_{i,t}} = \sum_{j=1}^{J} w_j rf_{i,j,t} \tag{57}$$

где је: $S_{i,t}$ сентимент i-тог текста ($1 \leq i \leq N$, при чему је N укупан број преузетих текстова) објављеног на дан t ($1 \leq t \leq T$, при чему је T последњи датум у узорку), w_j пондер сентимента[64] за j-ту реч ($1 \leq j \leq J$, при чему је J укупан број јединствених речи у тексту), $f_{i,j,t}$ апсолутна фреквенција j-те речи у i-том тексту објављеном на дан t, $a_{i,t}$ укупан број речи из i-тог текста објављеног на дан t, а $rf_{i,j,t}$ релативна фреквенција j-те речи у i-том тексту објављеном на дан t.

Пондери w_j представљају нивое сентимента појединачних речи. У изворним истраживањима сентимента, ови пондери узимали су само три вредности -1, 0 и 1. Зато се обично каже да се сентимент речи традиционално креће у границама [-1,1]. Насупрот томе, Џагадиш и Ву (2019) дозвољавају да ове вредности слободно варирају дуж реалне осе чиме се узима у обзир да позитивне и негативне речи не морају бити исте тежине. Постоје две мане овако дефинисаног сентимента речи. Прва је да ниво сентимента речи излази из традиционалних граница, што отежава интерпретацију пондера. Примера ради, не може се априори рећи да ли је пондер -5 јако негативан или умерено негативан сентимент. Задржавањем традиционалних граница

[64] Пондери сентимента (енгл. *Sentiment weight*) је оригинални назив који су Џагадиш и Ву (2019) употребили да означе ниво сентимента сваке речи у свом раду.

проблем би нестао, јер би границе служиле као бенчмарк. Примера ради, пондер од -0.85 сугерисао би високо негативан сентимент јер је близу граничне вредности -1. Друга слабост је последица назива који је дат овој променљивој. Иако ове вредности носе назив пондери не морају се сабирати до 1. Сама природа сентимента не дозвољава наметање ограничења да се пондери сентимента речи морају сабирати до 1. Непостојање логичне интуиције за увођење истакнутог ограничења најбоље се може сагледати у одсуству његове интерпретације. Како би се могао објаснити захтев да збир сентимента свих позитивних, негативних и неутралних речи у енглеском (или неко другом) језику (или у неком подскупу речи из датог језика) мора бити једнак 1? Без обзира на то, покушај ових аутора да емпиријски мере сентимент речи узимајући у обзир да све речи из исте категорије сентимента не морају бити исте тежине био значајан корак унапред у постојећој литератури.

Након увођења појма сентимента текста, Џагадиш и Ву (2019) уводе претпоставку да сентимент текстова утиче на економске променљиве (конкретно, на приносе) о којима текстови пишу. Последично, онда је могуће оценити регресиону једначину следећег типа:

$$r_t = \alpha + \beta S_{i,t} + \epsilon_{i,t} \tag{58}$$

где су: алфа и бета регресиони параметри, r_t приноси на финансијску активу о којој текстови пишу, а ϵ_t резидуали.

Међутим, сентимент текстова није априори познат, јер априори не знамо ни вредности пондера сентимента. Да бисмо оценили пондере сентимента, аутори предлажу следећу модификацију једначине (58):

$$r_t = \alpha + \beta\left(\sum_{j=1}^{J} w_j rf_{i,j,t}\right) + \epsilon_{i,t} = \alpha + \sum_{j=1}^{J} \beta w_j rf_{i,j,t} + \epsilon_{i,t} = \alpha + \sum_{j=1}^{J} B_j rf_{i,j,t} + \epsilon_{i,t} \tag{59}$$

где су $B_j = \beta w_j$ пондери сентимента који садрже грешку мерења једнаку утицају сентимента на приносе, β .

Аутори стога предлажу да се уместо праве вредности пондера w_j оцене пондери мерени са грешком уз помоћ једначине (59) класичним вишеструким линеарним регресионим моделом. Ипак, како би се грешка очистила, аутори предлажу да се тако оцењени параметри стандардизују:

$$z_j = \frac{B_j - \bar{B}}{std(B_j)} \tag{60}$$

Као последица шума у подацима, веома често се могу добити оцене мањег броја пондера, z_j, које представљају изузетно екстремне вредности у односу на већину оцењених пондера. Како би се ефекат тог шума изоставио из даље анализе, аутори предлажу винзоризацију (енгл. *winsorization*) оцена. Винзоризација је статистичка техника за повећање прецизности оцена заменом екстремних вредности са оба краја узорка неком вредношћу коју истраживач сматра адекватном. За потребе овог рада, сви стандардизовани пондери чија је апсолутна вредност већа од 3, биће винзоризовани.

4.2 Фриш-Вау-Ловелова Теорема

Како би се дискусија која следи лакше пратила, пожељно је да читалац буде упознат са Фриш-Вау-Ловелововом (Frisch–Waugh–Lovell) теоремом. Реч је о теореми економетријске анализе који носи назив по тројци својих аутора. Теорема решава питање отклањања утицаја релевантног регресора изостављеног из модела. У поставци теореме полази се од регресионог модела са N регресора:

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_N x_{N,i} + \varepsilon_i \tag{61}$$

где је: y_i зависна променљива, $x_{к,i}$ $к$-та објашњавајућа променљива и ε_i случајна грешка.

Претпоставимо да је због потреба истраживања потребно изоставити једну од променљивих из модела, иако је она статистички значајна. Њен изостанак манифестовао би се променом параметара нагиба уз задржане променљиве због погрешне спецификације модела (будући да релевантан регресор недостаје). Фриш-Вау-Ловелова теорема даје нам алат да ту грешку отклонимо. Теорема каже да се грешка може отклонити регресирањем свих променљивих из модела (и зависне и задржаних објашњавајућих) само на изостављену променљиву. У том случају добијамо N регресија – $(N-1)$ за задржане објашњавајуће променљиве и једну за зависну променљиву.

$$y_i = \gamma_1 x_{k,i} + \xi_i \tag{62а}$$

$$\forall r \in (1, N) \wedge r \neq k, \qquad x_{r,i} = \gamma_r x_{k,i} + \varphi_{r,i} \tag{62б}$$

где су: γ_r регресиони параметри, ξ_i случајне грешке добијене у помоћном моделу за зависну променљиву и $\varphi_{r,i}$ случајне грешке добијене у моделу за r-ту објашњавајућу променљиву.

Циљ ових регресија је да се одстрани или очисти утицај изостављене променљиве из свих преосталих променљивих у моделу. Случајне грешке, тј. резидуали, из овако оцењених модела представља ће нове регресиране променљиве из којих је очишћен утицај изостављене променљиве. Фриш-Вау-Ловелова теорема нам гарантује да, уколико оценимо модел у коме се регресира „очишћена“ зависна променљива, ξ_i, на „очишћене“ задржане објашњавајуће променљиве, $\varphi_{r,i}$, оцене параметара нагиба уз задржане регресоре биће исте као оне које су егзистирале у моделу са свих N регресора. Другим речима, резултат мора бити следећи модел:

$$\xi_i = \beta_1 \varphi_{1,i} + \beta_2 \varphi_{2,i} + \cdots + \beta_N \varphi_{N,i} + \varepsilon_i \tag{63}$$

Осврнимо се још једном на суштину теореме. Фриш-Вау-Ловелова теорема нам дозвољава да оценимо модел без регресора за који априори знамо да треба да се нађе у моделу, при чему се то неће одразити на регресионе параметре. То се постиже чишћењем утицаја изостављеног регресора и из зависне променљиве и из свих објашавајућих променљивих пре оцењивања модела. Циљ овакве трансформације је да се поједностави даљи рад, при чему се то неће одразити на финалне резултате.

4.3 Алтернативни приступ оцењивању

У одељку **3.4** указане су предности употребе *TF-IDF*-а у текстуалној анализи. Из тог разлога, прво побољшање које предлаже ова дисертација је употреба *TF-IDF*-а у рачунању сентимента. У складу са тим, потребно је да редефинишимо сентимент из израза (57), тако да се његов обрачун заснива на *TF-IDF*-у:

$$S_i = \sum_{j=1}^{J} w_j TFIDF_{i,j,t} \tag{64}$$

где је: S_i сентимент *i*-тог текста, J укупан број јединствених речи у тексту, w_j пондер сентимента за *j*-ту реч и $TFIDF_{i,j}$ вредност *TFI-DF* статистике *j*-те речи у *i*-том тексту.

Извођење које ће уследити независно је од изложене промене дефиниције сентимента. Другим речима, до истих закључака бисмо дошли и да смо задржали оригиналну дефиницију сентимента из рада Џагадиша и Ву (2019). Промена дефиниције извршена је само зато што се TFIDF сматра примеренијим, јер он мери информациони значај сваке речи, док релативна фреквенција мери њену учесталост. На тај начин сентимент се више не мери само на бази присутности или одсутности одређених речи, већи и на бази њиховог утицаја на читаоца.

Кренимо од модела (58). Заменимо сентимент по дефиницији (64) у полазни модел као што су то учини Џагадиш и Ву (2019) са својом дефиницијом.

$$r_t = \alpha + \beta \left(\sum_{j=1}^{J} w_j TFIDF_{i,j,t} \right) + \epsilon_t \tag{65}$$

Зарад једноставности даљег извођења пожељно је да се извођење настави без константе у моделу. За то је потребно применити Фриш-Вау-Ловелову теорему. Да би се очистио утицај константе, оценимо $(J+1)$ модел у којима ће се редом регресирати приноси и *TF-IDF* за сваку од *J* речи само на константу.

$$r_t = a + y_t \tag{66а}$$

$$\forall j \in (1, J), \qquad TFIDF_{i,j,t} = a + x_{i,j,t} \tag{66б}$$

где је: a константа, $x_{i,j,t}$ *TF-IDF j*-те речи из *i*-тог текста објављеног на дан *t* из којег је очишћен утицај константе, док је y_t принос на дан *t* из којег је очишћен утицај константе.

Након тога сачувајмо резидуале свих модела. Ово је неопходан корак, јер по Фриш-Вау-Ловеловој теореми они представљају део регресиране променљиве из кога је очишћен утицај константе. На тај начин, када формирамо модел (67) добићемо исте оцене за b и пондере w_j као да смо оцењивали модел (65):

$$y_t = b\left(\sum_{j=1}^{J} w_j x_{i,j,t}\right) + e_t \tag{67}$$

Представимо модел (67) у матричној нотацији:

$$y = bXW + \varepsilon \tag{68}$$

где је: y вектор (димензије N×1) приноса из којег је очишћен утицај константе, X матрица (димензије N×J) вредности *TF-IDF*-а свих јединствених речи из које је очишћен утицај константе, b непознати параметар нагиба (скалар), W вектор (димензије J×1) непознатих пондера сентимента и ε вектор (димензије N×1) случајних грешака (резидуала).

Према једној од полазних претпоставки класичног линеарног регресионог модела, модел ће бити адекватан уколико његови резидуали буду нормално расподељени. Полазећи од те претпоставке формирајмо функцију веродостојности, L:

$$L = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}e_t^2} = (2\pi\sigma^2)^{-\frac{N}{2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{N} e_t^2}$$

односно, у матричној нотацији:

$$L = (2\pi\sigma^2)^{-\frac{N}{2}} \cdot e^{-\frac{1}{2\sigma^2}\varepsilon'\varepsilon} \tag{69}$$

где је σ варијанса резидуала.

С обзиром на то да је функција веродостојности (69) непрактична за рад, типично се у статистичкој литератури прави њена монотона трансформација коришћењем природног логаритма. Монотоно трансформисана функција веродостојности (69) биће практичнија за рад, а њен максимум ће се наћи у истој тачки као и максимум оригиналне функције веродостојности (69). На тај начин, максимизирањем функције (70) добићемо исту оптималну вредност као да смо максимизирали (69), али ће поступак максимизације бити једноставнији.

$$\ln(L) = \ln\left((2\pi\sigma^2)^{-\frac{N}{2}} \cdot e^{-\frac{1}{2\sigma^2}\varepsilon'\varepsilon}\right)$$

$$\ln(L) = \ln\left((2\pi\sigma^2)^{-\frac{N}{2}}\right) + \ln\left(e^{-\frac{1}{2\sigma^2}\varepsilon'\varepsilon}\right)$$

$$\ln(L) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\varepsilon'\varepsilon \tag{70}$$

Заменимо у једначину (70) чему су једнаки резидуали према постављеном моделу (68). Затим средимо добијени израз:

$$\ln(L) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - bXW)'(y - bXW)$$

$$\ln(L) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y'y - 2y'bXW + b^2W'X'XW) \quad (71)$$

Како би се оценили и параметар b и вектор пондера W искористићемо исту идеју троетапне процедуре као у Јохансен (1996). Аутор је своју чувену идеју употребио за извођење оцена за параметре коинтеграције и параметре прилагођавања у коинтегрисаном VAR моделу. Идеја процедуре је следећа: максимизујмо монотоно трансформисану функцију веродостојности у односу на једну од посматране две групе параметара. На тај начин добићемо оцену једне групе параметара у функцији друге. Тако добијену оцену уврстимо у функцију веродостојности, и максимизујмо је по параметрима друге групе. Овим путем добићемо аналитичку оцену друге групе параметара. Након што ту аналитичку оцену уврстимо у израз који смо добили максимизирајући функцију веродостојности по првој групи параметара, добићемо аналитичке оцене за параметре прве групе. Процедуру представљамо у наставку. За почетак максимизујмо трансформисану функцију веродостојности по параметру b. Према Фармаовој (Fermat) теореми, екстремне вредности функције налазе се у нулама њеног првог извода (услов првог реда), те диференцирамо (71) по b:

$$\frac{\partial ln(L)}{\partial b} = -\frac{1}{2\sigma^2}(-2y'XW + 2bW'X'XW) \quad (72)$$

Израз (72) ће бити једнак нули ако и само ако је израз у загради једнак нули, из чега следи:

$$(-2y'XW + 2bW'X'XW) = 0$$

$$2bW'X'XW = 2y'XW$$

$$b = y'XW(W'X'XW)^{-1} \quad (73)$$

Према томе, израз (73) представља оцену за праметар b у функцији од још увек неоцењеног вектора параметара W. Сада када нам је ова оцена позната, уврстимо је у (71):

$$ln(L) = -\frac{N}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y'y - 2y'XW(W'X'XW)^{-1}y'XW + y'XW(W'X'XW)^{-1}W'X'XW(W'X'XW)^{-1}y'XW) \quad (74)$$

Како се у изразу (74) један до другог појављују скалар и његова инверзна вредност, можемо их скратити јер: $(W'X'XW)^{-1}W'X'XW = 1$. Те се израз (74) своди на:

$$ln(L) = -\frac{N}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y'y - 2y'XW(W'X'XW)^{-1}y'XW + y'XW(W'X'XW)^{-1}y'XW)$$

$$ln(L) = -\frac{N}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y'y - y'XW(W'X'XW)^{-1}y'XW) \quad (75)$$

Израз (75) је потребно максимизирати у односу на вектор W, при чему се поменути вектор појављује само у умањиоцу из друге заграде. То значи да се максимизацијом датог умањиоца максимизоваћемо цео израз (75)[65]. Посматрајмо за тренутак умањилац као засебан израз:

$$(max)\ y'XW(W'X'XW)^{-1}y'XW \tag{76}$$

Зарад једноставности даљег излагања уведимо ознаке:

- $S_{yx} = y'X$ – Вектор димензије 1×J
- $S_{xx} = X'X$ – Матрица димензије J×J

Када их уврстимо у претходни израз добијамо:

$$(max)\ S_{yx}W(W'S_{xx}W)^{-1}S_{yx}W \tag{77}$$

Приметимо да су $S_{yx}W$ и $W'S_{xx}W$ скалари[66]. Захваљујући том својству, израз (46) се може написати као:

$$(max)\ \frac{S_{yx}WS_{yx}W}{W'S_{xx}W} = \frac{(S_{yx}W)^2}{W'S_{xx}W} \tag{78}$$

Из претходног записа приметимо је да је функција циља хомогена реда 0:

$$f(k \cdot W) = \frac{(S_{yx}kW)^2}{kW'S_{xx}kW} = \frac{k^2(S_{yx}W)^2}{k^2W'S_{xx}W} = \frac{(S_{yx}W)^2}{W'S_{xx}W} = f(W) \tag{79}$$

Ово својство функције циља за последицу има проблем који је у литератури познат под називом *вишеструко оптимално решење*. Нека је W^*оптимално решење претходног проблема, онда је и $W^{**} = W^* \cdot k$ такође оптимално решење, где је k произвољна коначна ненулта константа. Како би се оптимална вредност за k пронашла, потребно је увести бар једно ограничење.

Претпоставимо да априори знамо колико треба да износи производ садржан у бројиоцу израза (78) (тј. $S_{yx}W$) при оптималном решењу. Означимо ту вредност са r_0. Ову вредност детаљно разматрамо у наредној подсекцији, док је у наставку овог извођења узимамо као дату. Под датим околностима проблем (78) се може модификовати на следећи начин:

---

[65] То се јасно види ако се израз (75) развије. Издвојени умањилац је једина део овог израза који повећава његову вредност:

$$ln(L) = -\frac{N}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}y'y + \frac{1}{2\sigma^2}y'XW(W'X'XW)^{-1}y'XW$$

[66] То се јасно види из димензија свих чиниоца оба производа: $(1 \times J) \cdot (J \times 1) = (1 \times 1)$ и $(1 \times J) \cdot (J \times J) \cdot (J \times 1) = (1 \times 1)$.

$$(min)\ \frac{1}{2}W'S_{xx}W \tag{80}$$
$$s.t. \quad S_{yx}W = r_0$$

Решавање новог проблем захтева формирањем Лагранжове функције:

$$La = \frac{1}{2}W'S_{xx}W - \lambda(S_{yx}W - r_0) \tag{81}$$

где је r_0 претпостављени ниво производа $S_{yx}W$ при оптималном решењу.

Формирајмо услове првог реда на бази оба непозната параметра:

$$\frac{\partial La}{\partial W} = S_{xx}W - \lambda S'_{yx} = 0 \tag{82}$$

$$\frac{\partial La}{\partial \lambda} = S_{yx}W - r_0 = 0 \tag{83}$$

Решавањем једначине (82) по W добијамо:

$$W = \lambda S_{xx}^{-1}S'_{yx} \tag{84}$$

Заменом израза (84) у израз (83) добија се једначина по λ коју треба решити:

$$\lambda S_{yx}S_{xx}^{-1}S'_{yx} = r_0$$
$$\lambda = \frac{r_0}{S_{yx}S_{xx}^{-1}S'_{yx}} \tag{85}$$

На крају, заменимо добијено λ из израза (85) назад у израз (84):

$$W = \frac{r_0}{S_{yx}S_{xx}^{-1}S'_{yx}}S_{xx}^{-1}S'_{yx} = \frac{S_{xx}^{-1}S'_{yx}}{S_{yx}S_{xx}^{-1}S'_{yx}}r_0 = \widetilde{W}r_0 \tag{86}$$

где је: $\widetilde{W} = \frac{S_{xx}^{-1}S'_{yx}}{S_{yx}S_{xx}^{-1}S'_{yx}}$ и представља вектор пондера из којих је грешка мерења делимично отклоњена.

Оптимална вредност за r_0 се лако може добити из ограничења о збиру пондера (видети Марковица (*Markowitz*) 1952). Ипак, у проблему (80) таквог додатног ограничење нема, те се питање третмана r_0 мора решити на другачији начин. Пре него наставимо са излагањем скренимо пажњу на то да је један део грешке већ отклоњен. Подсетимо да Џагадиш и Ву (2019) предлажу оцену производа $B_J = bW$ као апроксимацију за пондере W. Њихове оцене дате су изразом $\hat{B}_J = S_{xx}^{-1}S'_{yx}$, и управо се појављују у изразу (86) до којег смо дошли алтернативним извођењем. Из једначине (84) се може видети да су алтернативне оцене заправо оцене

Џагадиша и Вуа (2019) помножене корективним фактором λ. Корективни фактор је дефинисан изразом (85) и служи за отклањање грешке у мерењу коју њихова процедура прави. Како нам је познат именилац корективног фактора, један део грешке мерења је отклоњен. Проналажењем бројиоца, r_0, грешка ће бити отклоњена у потпуности.

Одређивање оптималне вредности r_0 без увођење нове информације (тј. новог ограничења) је незахвалан задатак. Покажимо најпре да се оцена оптималног r_0 након утврђивања првобитних оцена за W и b (изрази (73) и (86)) не може добити минимизирањем суме квадрата полазног модела. То је последица чињенице да ће се њиховом заменом у полазном моделу утицај r_0 бити елиминисан. Сагледајмо овај проблем кренувши од израза (68). Уврстимо у њега првобитно добијене оцене за W и b:

$$y = bXW + \varepsilon$$

$$y - bXW = \varepsilon$$

$$y - y'XW(W'X'XW)^{-1}XW = \varepsilon$$

$$y - y'X\widetilde{W}r_0(r_0\widetilde{W}'X'X\widetilde{W}r_0)^{-1}X\widetilde{W}r_0 = \varepsilon$$

$$y - y'X\widetilde{W}r_0^2(r_0^2\widetilde{W}'X'X\widetilde{W})^{-1}X\widetilde{W} = \varepsilon$$

$$y - y'X\widetilde{W}r_0^2\frac{1}{r_0^2}(\widetilde{W}'X'X\widetilde{W})^{-1}X\widetilde{W} = \varepsilon$$

$$y - y'X\widetilde{W}(\widetilde{W}'X'X\widetilde{W})^{-1}X\widetilde{W} = \varepsilon \qquad (87)$$

Из израза (87) јасно се види да се утицај r_0 потире, те да нам овај приступ не може помоћи у његовом одређивању[67]. Оптимална вредност r_0 се не може калибрисати употребо унакрсне, односно крос, валидације (енгл. *cross-validation*), будући да није реч о егзогеном параметру.

4.3.1 Оцене пондера сентимента

У наставку анализирамо неколико могућих начина путем којих можемо доћи до оцена пондера сентимента. Пођимо од тога да се проблем оцењивања може превазићи тако што ће се утицај r_0 апроксимирати. Посматрајмо ограничење из модела (80). Заменимо чему је по дефиницији једнака вредност S_{yx}:

$$y'XW = r_0 \qquad (88)$$

У изразу (88) препознајемо дефиницију сентимента $S = XW$, те се ограничење може представити и на следећи начин:

$$y'S = r_0 \qquad (89)$$

[67] Чињеница да ће се скалар r_0 поништити је директна последица хомогености циљне функције. Погледати (79).

Из редефинисаног ограничења видимо да је r_0 заправо пондерисани збир приноса при чему су пондери оцењени нивои сентимента. Под претпоставком да је реч о логаритамским, тј. континуалним приносима, овај збир представљаће кумулативни сентиментом поднерисани принос на крају анализираног периода. Како ова вредност априори није позната, као прокси за r_0 можемо узети кумулативни принос на крају посматраног периода оцењен из узорка. Укључивање кумулативних приноса у сентимент има и лепу економску интерпретацију. Апроксимација би сугерисала да заинтересована јавност формира свој сентимент и на бази кумулативних приноса остварених до тог тренутка. Другим речима, на ставове људи утичу и перформансе финансијске активе.

$$\tilde{r}_0 = cr_T \tag{90}$$

Из претходног проистиче прва предложена оцена за пондере сентимента:

$$\widehat{W} = \widetilde{W}\tilde{r}_0 \tag{91}$$

До оптималне оцене за параметар b можемо доћи на два начина. Интуитиван, али комплекснији пут је замена оцене (91) у израз (73):

$$b = y'XW(W'X'XW)^{-1}$$

$$b = S_{yx}W(W'S_{xx}W)^{-1}$$

$$\tilde{b} = S_{yx}\frac{S_{xx}^{-1}S'_{yx}}{S_{yx}S_{xx}^{-1}S'_{yx}}\tilde{r}_0\left(\tilde{r}_0\frac{S_{yx}S_{xx}^{-1}}{S_{yx}S_{xx}^{-1}S'_{yx}}S_{xx}\frac{S_{xx}^{-1}S'_{yx}}{S_{yx}S_{xx}^{-1}S'_{yx}}\tilde{r}_0\right)^{-1}$$

$$\tilde{b} = \frac{S_{yx}S_{xx}^{-1}S'_{yx}}{S_{yx}S_{xx}^{-1}S'_{yx}}\tilde{r}_0\left(\tilde{r}_0^2\frac{S_{yx}S_{xx}^{-1}S_{xx}S_{xx}^{-1}S'_{yx}}{\left(S_{yx}S_{xx}^{-1}S'_{yx}\right)^2}\right)^{-1}$$

$$\tilde{b} = \tilde{r}_0\left(\frac{\tilde{r}_0^2}{S_{yx}S_{xx}^{-1}S'_{yx}}\right)^{-1}$$

$$\tilde{b} = \tilde{r}_0\cdot\frac{S_{yx}S_{xx}^{-1}S'_{yx}}{\tilde{r}_0^2} = \frac{S_{yx}S_{xx}^{-1}S'_{yx}}{\tilde{r}_0} \tag{92a}$$

Алтернативно, имајући на уму да је Лагранжов множитељ дефинисан изразом (80) корективни фактор којим се отклања утицај грешке (тј. утицај параметра нагиба) из оцена Џагадиша и Вуа (2019), можемо дефинисати следећу релацију:

$$b = \frac{1}{\lambda} \tag{93}$$

Тада имамо:

$$\tilde{b} = \frac{1}{\frac{\tilde{r}_0}{S_{yx}S_{xx}^{-1}S'_{yx}}}$$

$$\tilde{b} = \frac{S_{yx}S_{xx}^{-1}S'_{yx}}{\tilde{r}_0} \tag{93б}$$

Пре него пређемо на други приступ, испитајмо чему су једнаки варијанса случајних грешака и максимум функције веродостојности. Кренимо од израза (75) и уврстимо у њега две већ предложене смене, као и нову смену $S_{yy} = y'y$:

$$ln(L) = -\frac{N}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(S_{yy} - S_{yx}W(W'S_{xx}W)^{-1}S_{yx}W)$$

Заменимо у горњи израз предложену оцену пондера сентимента. Приметимо да смо извођењем израза (95а) показали следеће $S_{yx}\widehat{W} = \tilde{r}_0$ и $(\widehat{W}'X'X\widehat{W})^{-1} = \frac{S_{yx}S_{xx}^{-1}S'_{yx}}{\tilde{r}_0^2}$. Из тога следи:

$$ln(L) = -\frac{N}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(S_{yy} - \tilde{r}_0\frac{S_{yx}S_{xx}^{-1}S'_{yx}}{\tilde{r}_0^2}\tilde{r}_0\right)$$

$$ln(L) = -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(S_{yy} - S_{yx}S_{xx}^{-1}S'_{yx}) \tag{94}$$

Израчунајмо први извод функције веродостојности по параметру σ^2:

$$\frac{\partial ln(L)}{\partial\sigma^2} = -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\left(S_{yy} - S_{yx}S_{xx}^{-1}S'_{yx}\right) = 0$$

$$\frac{1}{2\sigma^4}\left(S_{yy} - S_{yx}S_{xx}^{-1}S'_{yx}\right) = \frac{N}{2}\frac{1}{\sigma^2}$$

$$\left(S_{yy} - S_{yx}S_{xx}^{-1}S'_{yx}\right) = N\sigma^2$$

$$\sigma^2 = \frac{1}{N}\left(S_{yy} - S_{yx}S_{xx}^{-1}S'_{yx}\right) \tag{95}$$

Максимум функције веродостојности ћемо добити када спојимо резултате из израза (95) и (94). Приметимо да максимум функције веродостојности дат изразом (96) остаје исти као максимум функције веродостојности класичног линеарног регресионог модела:

$$ln(L) = -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\left(S_{yy} - S_{yx}S_{xx}^{-1}S'_{yx}\right)$$

$$ln(L) = -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}N\sigma^2$$

$$ln(L) = -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{N}{2} \tag{96}$$

$$ln(L) = -\frac{N}{2}(ln(2\pi) + ln(\sigma^2) + 1)$$

Други приступ оцењивању подразумевао би нормирање. Наиме, због природе сентимента и наслеђа конвенционалног квантификовања истог кроз три нумеричке категорије (–1, 0 и 1), има смисла нормирати пондере тако да се сви они нађу између –1 и 1. То није неуобичајена пракса у економетријској литератури[68]. Штавише, сличну трансформацију предлажу и сами Џагадиш и Ву (2019), који пондере стандардизују не би ли смањили грешку мерења. Овде разматрамо један могућ приступ нормирању и указујемо на то да ће се његовом применом утицај r_0 у потпуности елиминисати. Приступ се ослања на математички образац којим се вредности из једног реалног интервала мапирају у други. Другим речима, оцењене вредности пондера сентимента добијене из израза (86) је потребно мапирати из интервала $[\widehat{w}_{min}, \widehat{w}_{max}]$ у интервал $[-1, 1]$.

$$\overline{w}_j = \frac{1-(-1)}{w_{max} - w_{min}}\left(w_j - w_{min}\right) - 1 = \frac{2\left(w_j - w_{min}\right)}{w_{max} - w_{min}} - 1 \tag{97}$$

где је: w_{max} највећи оцењени пондер, w_{min} најмањи оцењени пондер, w_j пондер сентимента *j*-те речи и $\overline{w}_j$ нормирани пондер сентимента *j*-те речи.

Ново добијени пондери неће садржати утицај r_0. То се лако може видети када у израз (97) уврстимо оцену (96):

$$\overline{w}_j = \frac{2\left(\widetilde{w}_j r_0 - \min\left(\widetilde{W} r_0\right)\right)}{\max\left(\widetilde{W} r_0\right) - \min\left(\widetilde{W} r_0\right)} - 1$$

$$\overline{w}_j = \frac{2\left(\widetilde{w}_j r_0 - \min\left(\widetilde{W}\right) r_0\right)}{\max\left(\widetilde{W}\right) r_0 - \min\left(\widetilde{W}\right) r_0} - 1$$

$$\overline{w}_j = \frac{2\left(\widetilde{w}_j - \widetilde{w}_{min}\right) r_0}{\left(\widetilde{w}_{max} - \widetilde{w}_{min}\right) r_0} - 1$$

$$\forall j \in (1, J), \qquad \overline{w}_j = \frac{2\left(\widetilde{w}_j - \widetilde{w}_{min}\right)}{\widetilde{w}_{max} - \widetilde{w}_{min}} - 1 \tag{98}$$

где је: $\widetilde{w}_{max}$ највећа оцена пондера из вектора $\widetilde{W}$ и $\widetilde{w}_{min}$ најмања оцена пондера из истог вектора.

Израз (98) је лако оценити, будући да не садржи r_0 и да смо приступ оцењивања вектора $\widetilde{W}$ већ установили. Друга предност овог приступа је то што пружа конзистентност са изворном литературом, будући да ће се сви оцењени пондери наћи у интервалу $[-1, 1]$. На овај начин дисертација даје своју подршку идеји успостављања конвенције по којој би се сентимент речи

[68] Као пример наводимо и Јохансен (1996) где се оцене коинтеграционих параметра нормирају тако да параметар уз променљиву од интереса има вредност 1.

мерио на скали од -1 до 1, како би се задржао склад са традиционалним приступима у мерењу сентимента.

Нормиране оцене $\bar{W}$ биће адекватне егзактне оцене вектора W уколико се прихвати конвенција о мерењу сентимента у интервалу $[-1,1]$. Прихватање конвенције сугерисало би да највећа и најмања вредност вектора W на нивоу популације морају бити 1 и –1. Одатле следи:

$$\bar{W} = \frac{2(W - w_{min})}{w_{max} - w_{min}} - 1 = \frac{2\big(W - (-1)\big)}{1 - (-1)} - 1 = \frac{2(W + 1)}{2} - 1 = W \tag{99}$$

Захваљујући датом својству нормиране оцене смемо да убацимо у израз (73) и из њих добијемо нумеричке оцене за параметар нагиба b. Тиме се обезбеђује да кроз процес оцењивања дођемо до обе оцене појединачно (и b и W) без грешке мерења. Дати резултат је утолико значајнији ако се узме у обзир да се добијање обе оцене не може постићи чак ни стандардизацијом коју су предложили Џагадиш и Ву (2019). Нумеричке оцене параметра b биће:

$$\bar{b} = S_{yx}\bar{W}(\bar{W}'S_{xx}\bar{W})^{-1} \tag{100}$$

Трећи приступ оцењивању ослања се на нумеричке методе и оптимизацију са ограничењима. Пођимо од модела (79). У складу са предложеном конвенцијом, посматрани израз можемо обогатити сетом ограничења облика:

$$\forall j \in (1, J), \qquad -1 \leq w_j \leq 1 \tag{101}$$

Постављени проблем није лако решити будући да он подразумева оптимизацију са хиљадама непознатих вредности и ослања се на незасићена ограничења и хомогену функцију циља. Из тог разлога проблем би било могуће решити алгоритмима нумеричке оптимизације, при чему и даље постоји проблем ангажовања велике количине рачунарских ресурса. Овај приступ се додатно може побољшати Бајесовским путем или аналогно Блек-Литерман (*Black-Litterman*) (1991) моделом. На тај начин можемо да комбинујемо[69] резултате добијене из података са експертским претпоставкама о сентименту сваке речи. Другим речима, овакво оцењивање дозвољавало би комбиновање мерења сентимента на бази машинског учења са традиционалним мерењем заснованим на лексикону у једну јединствену оцену. Тиме би се прецизност мерења додатно поправила. Међутим, ово не решава проблем интензивног коришћења рачунарских ресурса, већ га усложњава. Исцрпно ангажовање радних ресурса можемо избећи увођењем неке поједностављене претпоставке о пондерима сентимента у виду засићеног ограничења. Тада би се добијени проблем могао решити по узору на аналитичко решење Марковицевог (1952) проблема.

[69] Треба истаћи да идеја комбиновања оцена није ексклузивно везана за нумеричку оптимизацију. Оцене добијене по претходно разматраним приступима можемо објединити (рецимо упросечавањем) са проценама стручњака у циљу добијања прецизнијих процена сентимента. Разлика је само у томе што ће се код нумеричких метода то радити итеративно, док ће се код осталих метода једино финално решење објединити са проценама експерата.

Сумирајмо приказане резултате. Дисертација нуди неколико оцена за пондере сентимента речи. Прво представљено решење, у ознаци $\widetilde{W}$, делимично је отклонило грешку мерења. Друго решење, $\widehat{W}$, је апроксимативно и дозвољава нам да један део грешке мерења апроксимирамо кумулативним приносом. Треће решење се ослања на традиционално мерење сентимента у интервалу $[-1,1]$ и дозвољава нам да кроз нормирање дођемо до адекватних егзактних оцена пондера сентимента $\overline{W}$. Претпоставка о кретању сентимента унутар традиционалног интервала може бити искоришћена за довршавање оптимизације нумеричким путем. Коначно, све оцене се могу додатно унапредити њиховим обједињавањем са другим методима мерења сентимента предложеним у литератури.

Ова секција предложила је побољшања према којима ће се оцењивање и мерење сентимента базирати на *TF-IDF*-у, отклонити грешке у мерењу сентимента речи, а њихов ниво вратити у оквиру традиционалних граница. До краја ове дисертације квалитет предложених побољшања ће се испитати емпиријски. Најпре ће се израз (98) употребити за оцену сентимента речи. Паралелно са тиме оцениће се и сентимент речи по оригиналној процедури Џагадиша и Вуа (2019) дат изразом (60). Тако добијене вредности биће употребљене за мерење сентимента текстова. Поредећи квалитет прогнозе које прави финални модел када се у њему нађу сентимент вести процењен на бази израза (98) са оним које финални модел прави када се сентимент вести процењује на бази израза (60) провериће се квалитет предложених побољшања. Ове теме детаљно дискутујемо у поглављу које следи.

5. Анализа добијених резултата

Пето поглавље представља и анализира емпиријске резултате до којих је истраживање дошло. Кроз приказане резултате биће испитане хипотезе постављене у уводном делу дисертације и испитане перформансе побољшаних оцена пондера сентимента. Као што је раније истакнуто, истраживање је подељено у три етапе. Прва етапа припрема улазних података за даљу анализу. У оквиру ње најпре су преузети одговарајући интернет чланци. Део чланака искоришћен је за оцену пондера сентимента по методологији из поглавља **4.3.1**. Тај корак обезбедиће да смо у стању да измеримо ниво сентимента сваког текста, јер ћемо располагати сентиментом сваке речи. На крају етапе обрачунат је сентимент свих чланака, као и њихова читљивост. У другој етапи, конструисани предиктори (сентимент вести о самој крипто-валути, њихова читљивост и сентимент вести о Биткоину) биће употребљени за оцену помоћног модела (12). На бази добијених резултата анализираће се утицај онлајн чланака на приносе одабраних крипто-валута. Самим тим у другој етапи биће идентификовани релевантни предиктори приноса (неопходни за наредну, трећу, етапу). Коначно, у трећој етапи конструисан је ансамбл алгоритам који предвиђа кретање приноса на бази релевантних предиктора. Добијене грешке прогнозе поређене су са грешкама идентичног алгоритма базираног на улазним подацима добијеним по методологији Џагадиша и Вуа (2019). Због различитих потреба свих етапа, целокупан узорак текстова и приноса подељен је у три подузорка – по један за сваку етапу. У наставку дискутујемо сваку од њих посебно. Поред тога, дискутујемо и поједине алтернативне приступе у моделирању и њихове резултате, као и тестове слабе форме тржишне ефикасности осам одабраних крипто-валута.

5.1 Прва етапа: оцена сентимента и читљивости вести

Први подузорак намењен је за тренинг (односно оцену) пондера сентимента. Подузорком су обухваћени текстови и приноси из периода 01.01.2021 – 31.12.2021. Из датих текстова најпре су на бази алгоритма описаног у одељку **3.3** изрударени вектори речи. Затим су тако добијени вектори трансформисани у математичке векторе информационе значајности речи из текста. Сваки вектор представља један текст, а сваки елемент вектора представља *TF-IDF* једне од укупно J (у овом истраживању $J = 5248$) идентификованих различитих речи. Тако је добијена матрица облика (4). Уз помоћ ње и модификованог вектора приноса описаног изразом (9) оцењени су пондери (98). Описана процедура спроведена је код сваке од осам одабраних крипто-валута појединачно.

У табели **2** издвојено је по 20 најистакнутијих представника сваке од три традиционалне групе речи према сентименту. У питању су заједнички резултати добијени на бази свих осам модела. Издвојене речи имали су вредност пондера сентимента једнаку традиционалним пондерима сваке групе (-1, 0 и 1) или су им биле изразито близу. Другим речима, представљене су речи чији су пондери сентимента били најмањи (негативне), највећи (позитивне) и најмањи по модулу тј. магнитуди (неутралне). Резултати разоткривају једну занимљиву појаву. Процедура је у свакој групи речи према сентименту идентификовала речи из општег енглеског језика као најистакнутије. Такав резултат дошао је као изненађење будући да се очекивало да се међу најутицајнијим речима нађу одређени финансијски термини (попут појмова „банкротство“ или

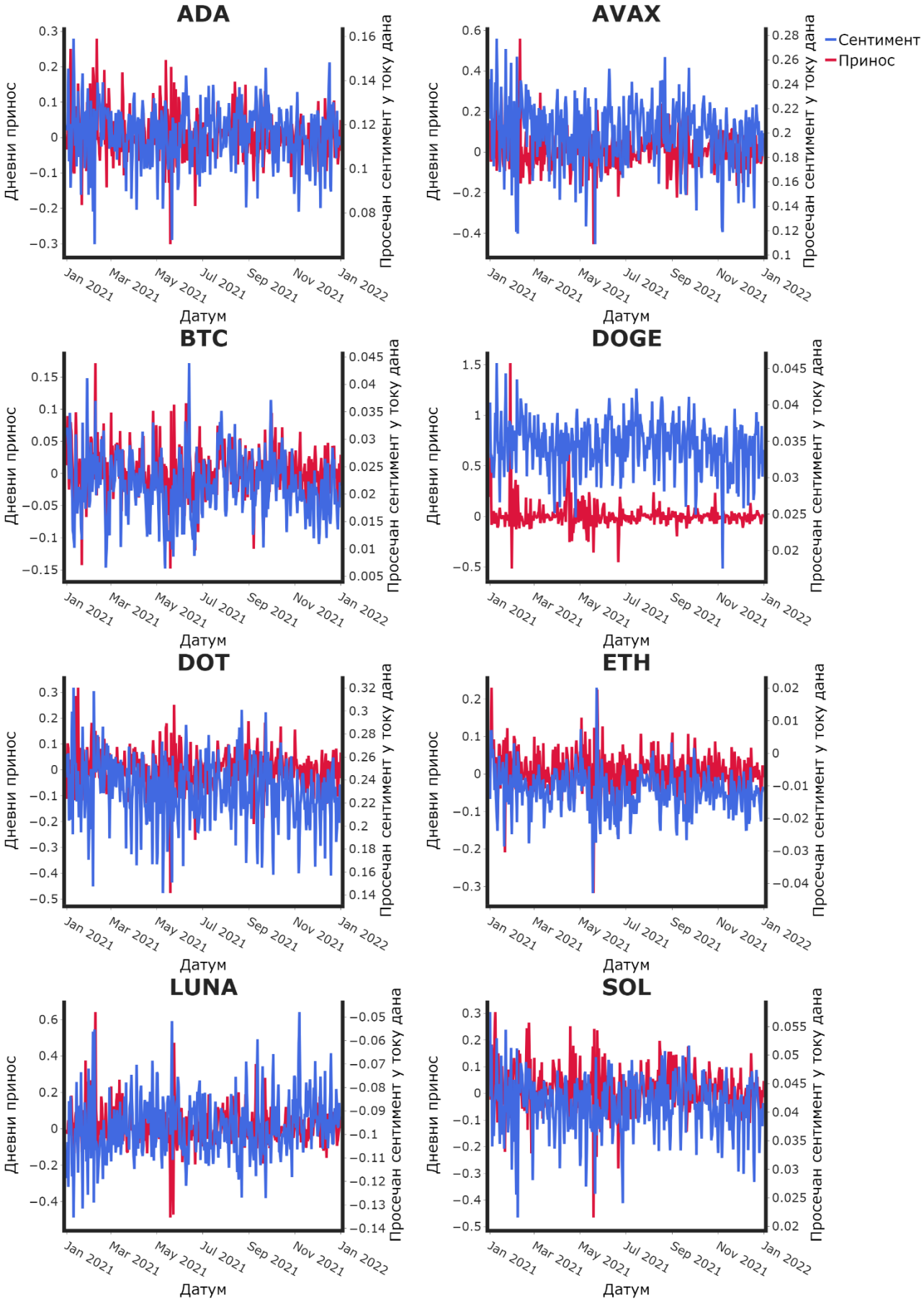
„профит“). Ипак, утицајност општих појмова можемо бранити ставом да заинтересовану јавност не чине само економисти и људи из домена финансија, већ људи различитих профила. Вокабулар који они користе није исти. У таквим околностима само општи појмови заједнички за све могу бити заједнички именилац који утиче на ставове свих актера. Идентификоване речи се углавном подударају са онима које би просечан човек издвојио као позитивне, односно, негативне. Ипак, има и речи које су се у том друштву нашле помало изненађујуће. Пример су речи „трилер“ и „диск“. Природно би било да се ове речи третирају као неутралне, с тога је неизбежно поставити питање да ли су се оне нашле међу најутицајнијим речима због шума у подацима или је њихов утицај заиста тако изразит.

Анализом пондера добијени су још неки занимљиви резултати, те овде издвајамо неке од њих. Имајући у виду широк дијапазон вести које преносе портали у њима се не ретко налазе експлицитни изрази, односно псовке, вулгаризми и непристојан говор. Испоставља се да су експлицитни изрази имали поприлично неутралан утицај на сентимент вести (просек 0,021). То нам говори да сензационалистички простаклук ни мало не дотиче заинтересовану јавност код крипто-валута. Резултати су показали и да је употреба нумеричких вредности у тексту имала јак негативан утицај на сентимент (просек -0,66). Приложени доказ нас доводи до закључка да заинтересована јавност јако негативно реагује на текстуалне вести преплављене бројевима. Занимљиво је и то да спомињање Корона вируса није имало значајнији утицај на сентимент (просечан пондер 0,074), али се може окарактерисати као благо позитиван. Резултат има смисла јер се Корона није негативно одразила на тржишта крипто-валута. Напротив, у време пандемије плаћање дигиталним новцем достиже пик своје популарности.

Табела 2: Приказ по 20 најутицајнијих речи из сваке групе сентимента са њиховим преводима на српском језику

Негативне		**Неутралне**		**Позитивне**	
Изворно	**Превод**	**Изворно**	**Превод**	**Изворно**	**Превод**
abort	абортирати/прекинути	album	албум	amicable	пријатељски
assassin	убица	ammo	муниција	allegory	алегорија
blackmail	уцењивати	autocorrect	самокорекција	beget	изродити
counterintuitive	контрапродуктивно	base	основа	cake	торта
clumsy	смотан	battery	батерија	candy	слаткиш
crook	нитков	cent	цент	cheerful	весео
deforestation	одшумљавати	chemistry	хемија	cherish	неговати
demolition	разарање	circulate	кружити	deregulate	дерегулисати
disease	болест	decode	декодирати	disk	диск
disfavor	немилост	diaspora	дијаспора	divine	божански
fracture	фрактура	ego	его	eulogy	хвалоспев
gullible	лаковеран	email	е-пошта	honey	драга
lawless	безакоње	fade	избледети	laundry	веш
leaderless	безвођство	hat	шешир	innocent	невин
monotony	монотонија	hyperspace	хиперпростор	mercy	милост
mutilate	унаказити	insect	инсект	mask	маска
paranoia	параноја	knee	колено	oasis	оаза
rubbish	ђубре	logistics	логистика	paradox	парадокс
subvert	саботирати	millionaire	милионер	polite	љубазан
thriller	трилер	overlay	преклапање	priest	свештеник

Слика 15: Кретање приноса и сентимента код осам одабраних крипто-валута у 2021. години

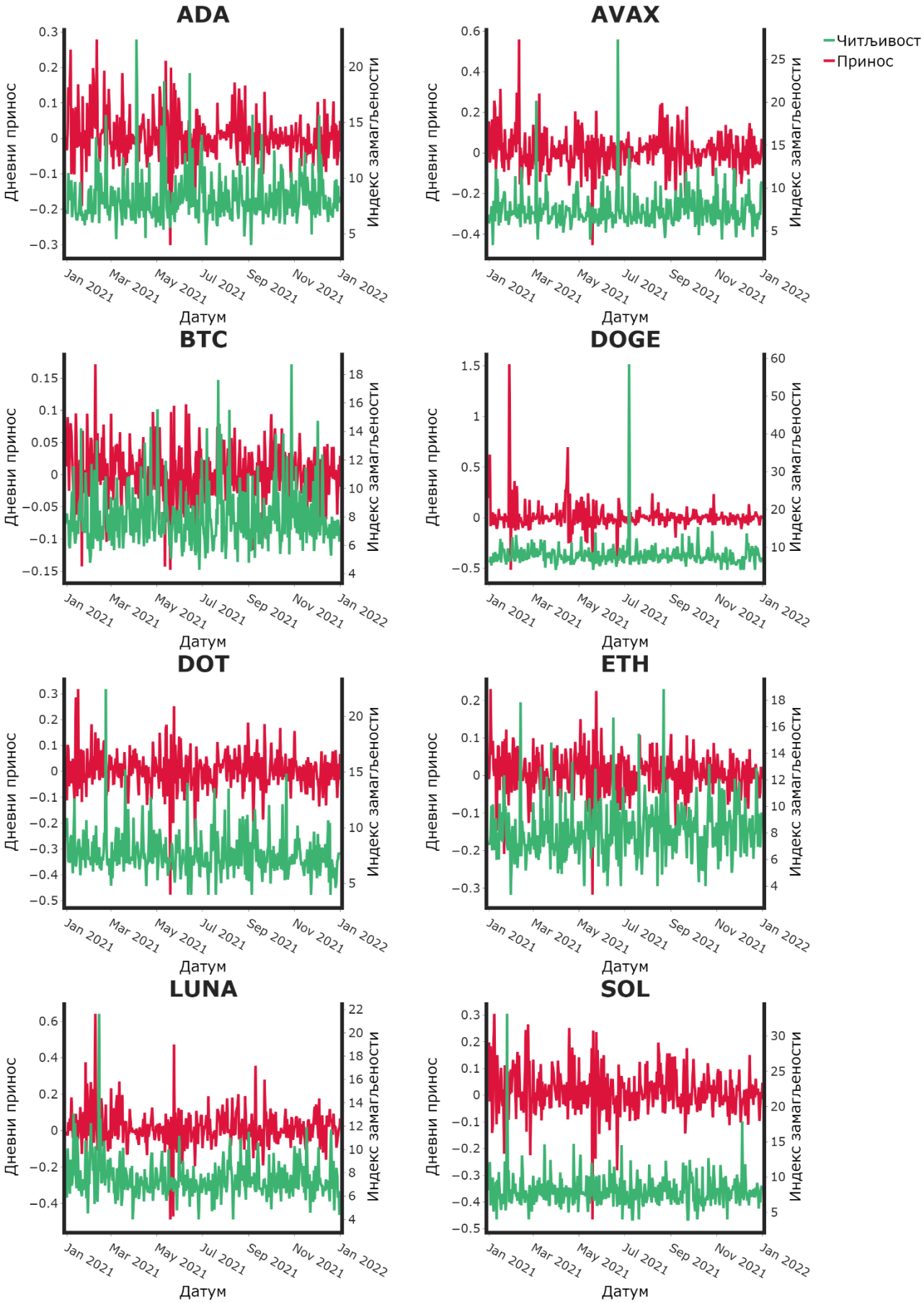


Извор: Приказ аутора

Подсетимо се да се у току једног дана публикује велики број различитих текстуалних вести о крипто-валутама. Све преузете вести се хронолошки могу поређати, будући да је њихово тачно време објављивања познато посетиоцима сајта Крипто Њуз. Захваљујући томе, изрударени вектор сентимената вести можемо посматрати као временску серију чији су подаци доступни на унутар-дневном нивоу (енгл. *intraday*). Кретање сентимента у 2021. години (тј. у првом подузорку) анализирано је графичким путем на слици **15**. Уместо унутар-дневних варијација слика прати просечан ниво сентимента у току сваког дана. Тиме је задржана упоредивост са приносима, чији су подаци доступни на дневном нивоу. Слика показује да просечан дневни сентимент није превише усаглашен са кретањем приноса. Већи степен корелације присутан је само код најпопуларнијих (у погледу броја објављених текстова у току дана) крипто-валута – Биткоина и Итиријума. О поменуте две крипто-валуте у просеку се дневно објави око педесетак вести, док се код свих осталих посматраних валута у просеку објављује десетак вести. То може бити знак да је потребно обезбедити велики број вести у току дана да би се њиховим агрегирањем дошло до адекватне процене дневног сентимента. Осим тога, резултат је у складу са теоријом Ђидовалвија (2003) према којој предвидљивост постоји само у веома кратком року. Са слике се може приметити и то да је код свих анализираних крипто-валута волатилност процењеног сентимента доста мања у односу на волатилност приноса. Приметно је, такође, да је у 2021. години код свих посматраних крипто-валута (изузев *LUNA*), преовладавао благо позитиван сентимент.

Поред сентимента, у овој фази обрачуната је и читљивост, односно замагљеност, сваког текста. Осврт на анализу замагљености обухваћених вести је важан, будући да је јасноћа написаног текста предуслов који мора бити задовољен да би његов сентимент имао потпун утицај на читаоца. Просечна замагљеност вести код свих крипто-валута кретала се између вредности 7 и 8. Подсетимо се да се текстови сматрају лако читљивим, односно, разумљивим обичним људима уколико је њихов индекс замагљености 7 или мањи. То нас наводи на закључак да су текстови у просеку пристојног нивоа читљивости. Томе у прилог говори и чињеница да се типично одступање од просечне замагљености код свих крипто-валута креће око нивоа ±2, те јако мали број текстова пробија ниво изнад којег се текстови сматрају нечитким. Прецизније, удео лако читљивих текстова (оних чији је индекс замагљености 7 или мањи) је већи од 30% код свих крипто-валута. Читљивост је посебно висока код крипто-валута *LUNA* и *AVAX* где удео лако читљивих вести премашује 50% узорка. Удео вести које су тешко разумљиве просечном читаоцу (оних чији је индекс замагљености 12 или већи) знатно је мањи. Код већине посматраних крипто-валута овај удео се креће око 5%. Изузеци су *LUNA* и *AVAX* код којих је овај удео око 2%. У узорку постоје и вести чији је садржај екстремно замагљен. Текст се сматра претешким за читање уколико је његов индекс замагљености већи од 17. То су обично високо стручни текстови чије разумевање захтева добро познавање области о којој се пише. Алтернативно, разумевање текстова овако високог нивоа замагљености захтевало би да читалац има завршен други степен академских студија (мастер, односно, магистратуру) да би се надоместио недостатак експертизе. Код свих крипто-валута удео високо стручних вести је изузетно мали и креће се око 0.5%. Најнечитљивија вест, са нивоом замагљености од чак 58.5, забележена је код крипто-валуте *DOGE*. Замагљеност високо стручних вести била је најмања код крипто-валута *LUNA* и *AVAX*, а кретала се око 20. Све до сада изнесене чињенице указују на пристојан ниво читљивости код свих крипто-валута. Ипак, међу њима посебно су се издвојиле *LUNA* и *AVAX* као најчитљивије, односно Биткоин као најмање читљива.

Слика 16: Однос приноса и читљивости код осам одабраних крипто-валута у 2021. години



Извор: Приказ аутора

Уколико графички упоредимо кретање просечне читљивости и приноса у 2021. години приметићемо низак ниво корелисаности између ове две величине. То је приказано на слици **16**. Поред тога што упућује на потенцијално ниску предиктивну моћ ове променљиве, резултат има још једну важну импликацију. Ниску корелацију између сентимента и приноса не можемо објаснити замагљеношћу текстова будући да је њихова читљивост била сасвим пристојна у посматраном периоду. То нас наводи на закључак да су читаоци у просеку били у стању да разумеју поруке (пристојна читљивост), али је њихова покретачка снага била мала (ниска корелација са приносима). Селекцијом предиктора ћемо се детаљније позабавити у наставку.

5.2 Друга етапа: Анализа утицаја

Други подузорак обухвата период 01.01.2022. – 28.02.2022, односно, прва два месеца 2022. године. Циљ датог подузорка је двојак. Подузорак ће нам првенствено помоћи да сагледамо везе између приноса и техничких показатеља добијених рударењем текста чиме ће се проверити валидност постављених хипотеза. Са друге стране, други подузорак ће послужити и за одређивање релевантних предиктора приноса које треба задржати у финалној етапи истраживања. Потенцијални предиктори које дисертација разматра су: ранији ниво приноса, сентимент вести о посматраној валути, утицај унакрсног сентимента Биткоина и замагљеност објављених вести. Оба циља остварујемо оцењивањем модела (12) и испитивањем значајности његових параметара. Анализа утицаја је спроведена на нивоу значајности од 10%, у складу са уобичајеном праксом анализе временских серија. Добијене резултате разматрамо појединачно и здружено.

За обрачун сентимента текстова из другог подузорка коришћени су пондери оцењени на бази првог подузорка. На тај начин се обезбеђује објективност у мерењу сентимента, будући да се претходно оцењени пондери сада користе на подацима које модел за њихову оцену није видео. Другим речима, пондери нису специјално прилагођени (односно оверфитовани) приносима из другог подузорка захваљујући томе што су два поменута узорка раздвојена. Да би изрударени пондери били адекватни за мерење сентимента у било ком другом подузорку, потребно је било обезбедити довољно вести за учење модела. Само тако ће модел имати довољно знања и информација да произведе генералне оцене пондера сентимента. Из тог разлога први узорак је најдужи. Ово истичемо јер се задржавањем објективности у мерењу сентимента гарантује да ће истраживање бити прави испит за употребну моћ анализе сентимента по модификованој методологији Џагадиша и Вуа (2019).

5.2.1 Преглед резултата по крипто-валутама

Резултати оцењених модела (12) за сваку од крипто-валута биће представљени табеларно. Табеле ће садржати информацију о оцењеним параметрима, њиховим п-вредностима, броју опсервација и мерама квалитета оцењеног модела (F статистика и кориговани R^2). Осим тога,

све табеле ће се састојати из две целине. У првој целини биће приказане наведене статистике за пун помоћни модел (12). Након што се из пуног модела отклоне незначајни предиктори добија се редуковани модел. Оцене параметара редукованог модела и његове пратеће статистике биће дате у другој целини свих табела. У наставку анализирамо сваки модел (тј. сваку табелу) појединачно.

Дискусију започињемо са крипто-валутом *ADA*. Оцењени резултати дати су Табелом **3**. Увидом у значајност приказаних оцена закључено је да се читљивост вести није одразила на приносе крипто-валуте *ADA*. Извртавањем овог показатеља из пуног модела добијен је редуковани модел у којем су сви параметри статистички значајни. На почетку 2022. године кретање приноса *ADA* коина било је одређено ставовима инвеститора формираних на бази онлајн вести, будући да су оба показатеља сентимента статистички значајна. Утицај сентимента вести о самој крипто-валути *ADA* био је позитиван и умерен. Значајнији утицај имале су и вести о Биткоину, чије сентимент је демотивисао инвеститоре да улажу у *ADA* коин. Другим речима, вести које говоре позитивно о Биткоину додатно су допринели паду вредности *ADA* коина у очима тржишта. То би могао да буде знак да корисници *ADA* коина Биткоин доживљавају као конкурентну крипот-валуту. Поред наведеног, резултати из Табеле 3 упућују и на благу аутокорелацију приноса (ауторегресивна компонента првог реда је статистички значајна).

Табела 3: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *ADA* (оцене сентимента нормиране на интервал [-1,1])

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.157	0.078	0.087	-1.524	-0.001	0.114	17.34	508
п-вредност	0.0000	0.0299	0.0559	0.0000	0.3473		0.0000	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.153	0.079	0.069	-1.529	-	0.115	22.82	508
п-вредност	0.0000	0.0278	0.0944	0.0000	-		0.0000	

Сасвим другачију слику добијамо када оценимо модел (12) у случају крипто-валуте *AVAX*. Приметимо да у табели 4 ни један од предиктора није статистички значајан. Познато је да присуство ирелевантног регресора може утицати на оцене преосталих регресионих параметара у моделу. Зато је експериментисано са изостављањем и задржавањем различитих комбинација потенцијалних регресора. Није постојала нити једна комбинација (укључујући и случајеве у којима се приноси регресирају само на један предиктор) чији су параметри били статистички значајни. Према томе, ни један од показатеља није имао утицаја на кретање приноса *AVAX* коина. Резултати су посебно занимљиви јер је у питању једна од најмлађих крипто-валута чије је тржиште још увек у развоју, а већ показује неосетљивост на сентимент инвеститора.

Табела 4: Приказ оцењеног пуног помоћног модела за приносе крипто-валуте *AVAX* (оцене сентимента нормиране на интервал [-1,1])

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.013	0.006	0.029	-0.113	0.001	-0.023	0.1381	154
п-вредност	0.7756	0.9352	0.7377	0.5988	0.8066		0.9680	

Резултати добијени оцењивањем помоћног модела над подацима о Биткоину презентовани су табелом **5**. Приметимо да је у пуном моделу регистрована значајност само прве доцње приноса. Ипак, отклањањем ирелевантних регресора из модела ситуација се променила и искристалисао се још један значајан параметар. Реч је о унакрсном сентименту. За испитивање утицаја унакрсног сентимента на Биткоин искоришћен је Итиријум. Будући да је реч о другој највећој, најутицајнијој[70] и најпопуларнијој крипто-валути, природно је очекивати да се ставови инвеститора о њој одразе и на Биткоин. Посматрано из угла корисника Биткоина, негативан предзнак испред сентимента Итиријума сугерише нам конкурентски однос између ове две крипто-валуте. Редуковани модел из табеле **5** нам открива и то да се ставови о Биткоину изнесени у онлајн вестима нису одразили на његове приносе. То, такође, важи и за читљивост поменутих вести. Може се извући закључак да су приноси Биткоина у посматраном периоду били детерминисани само својим претходним вредностима и ставовима инвеститора о Итиријуму.

Табела 5: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *BTC* (оцене сентимента нормиране на интервал [-1,1])

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{ETH,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-0.008	0.158	-0.012	0.218	0.000	0.026	9.095	1210
п-вредност	0.1220	0.0000	0.8513	0.4716	0.9868		0.0000	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{ETH,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-	0.164	-	-0.126	-	0.036	22.33	1210
п-вредност	-	0.0000	-	0.0046	-		0.0000	

Резултати добијени оцењивањем једначине (12) у случају *DOGE* коина дати су табелом 6. Оцене пуног модела упућују на то да константа и сентимент вести о *DOGE* коину нису релевантни предиктори приноса ове крипто-валуте. Уједно, утицај унакрсног сентимента био је на граници значајности. С тога је оцењен редуковани модел у којем фигуришу само статистички значајни параметри. Статистичка значајност читљивости онлајн вести је елемент редукованог модела који одмах скреће пажњу на себе из два разлога. Прво, ово је први модел који је препознао да читљивост вести представља сигнал за инвеститоре који има утицај на кретање приноса. Ипак, други разлог је још упечатљивији. Веза има неочекивани предзнак. Будући да индекс замагљености мери замагљеност текста, што је он већи текст је нејаснији.

[70] Као што је раније истакнуто *ETH* је тржишни лидер по питању бројних иновација на крипто-тржишту, има огромну мрежу корисника и уз Биткоин многима је прва асоцијација на крипто-валуте.

Позитиван предзнак сугерише да што су текстови били мање јаснији инвеститорима, то су они били спремнији више да инвестирају. Једно могуће објашњење за овај наизглед нелогичан резултат је порекло *DOGE* коина. Како је ова крипто-валута настала као шала, један број инвеститора је не доживљава озбиљно. Могуће је да тешко разумљиви текстови (они чије разумевање захтева експертизу у датој области или високо образовање) дају позитиван сигнал инвеститорима јер доприносе да им сама валута делује озбиљније. Подижући њену озбиљност охрабрују инвеститоре да уложе новац у *DOGE* коин чиме се врши притисак на раст приноса. Алтернативно, резултат можемо посматрати као аномалију периода, будући да почетком 2022. године започиње једна од највећих криза у савременој историји. Негативна веза пронађена је у случају сентимента вести о Биткоину. Значајност ове везе сведочи да су корисници *DOGE* коина доживљавали Биткоин као конкурентску крипто-валуту. Коначно, приметимо и то да је забележен статистички значајан утицај ауторегресивне компоненте првог реда на кретање приноса ове крипто-валуте.

Табела 6: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *DOGE* (оцене сентимента нормиране на интервал [-1,1])

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.015	0.130	-0.021	-0.714	0.001	0.020	3.539	493
п-вредност	0.3417	0.0013	0.8584	0.1037	0.0389		0.0074	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-	0.119	-	-0.333	0.002	0.022	4.658	493
п-вредност	-	0.0019	-	0.0100	0.0204		0.0032	

Табела 7 представља оцену модела (12) у случају још једне јако младе крипто-валуте – Полкадота. Из приложеног одмах је приметно да су сви параметри статистички значајни. *DOT* је једина анализирана крипто-валута код које је забележен овакав случај.

Табела 7: Приказ оцењеног пуног помоћног модела за приносе крипто-валуте *DOT* (оцене сентимента нормиране на интервал [-1,1])

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.084	0.202	0.076	-0.339	-0.004	0.081	5.725	215
п-вредност	0.0070	0.0027	0.0397	0.0103	0.0452		0.0002	

Проанализирајмо сваки од предиктора. Пођимо од тога да је сентимент инвеститора битно опредељивао кретање приноса ове крипто-валуте, будући да су оба показатеља базирана на сентименту статистички значајна. Позитиван сентимент из вести о Полкадоту мотивисао је инвеститоре да купују ову крипто-валуту, док су позитивне вести о Биткоину имале супротан ефекат. И у овом случају резултати сугеришу да корисници *DOT*-а доживљавају Биткоин као конкурентску крипто-валуту. Још једном имамо ситуацију у којој је утицај читљивости препознат као релевантан предиктор. За разлику од *DOGE* коина, овога пута добијен је

очекивани предзнак. Негативан предзнак сугерише да што су вести биле јасније просечном читаоцу, то је он био мотивисанији да инвестира у Полкадот, чиме се врши притисак на раст његових приноса. На крају истакнимо и то да су, поред индикатора добијених рударењем текста, на кретање приноса крипто-валуте Полкадот утицале и њихове претходне вредности.

Модел (12) оцењен у случају друге најутицајније крипто-валуте, Итиријум, дат је табелом **8**. Резултати сугеришу да полазни модел треба редуковати, будући да константа, сентимент вести о Итиријуму, као и њихова замагљеност нису статистички значајни регресори. Ово је јако занимљив резултат, будући да се од вестима из области крипто-валута највећи број њих објави управо о Итиријуму. Тим питањем ћемо се детаљније позабавити у наредном одељку. Након изостављања свих ирелевантних регресора, добијен је редуковани модел. У њему поново наилазимо на изненађење у виду позитивног предзнака уз сентимент текстова о Биткоину. Резултат сугерише да су позитивне вести о Биткоину мотивисале инвеститоре да купују Итиријум. То сугерише да корисници Итиријума не доживљавају Биткоин као конкурентску крипто-валуту, узевши у обзир разлике у њиховом функционисању. У таквим околностима могуће је да корисници Итиријума примају позитивне вести о Биткоину као позитиван сигнал за целу класу крипто-валута, а не као знак да је боље прећи на Биткоин. Занимљиво је да ова два тржишна сегмента имају другачији поглед на однос између Биткоина и Итиријума. Док корисници Биткоина доживљавају Итиријум као конкурента (на шта указују резултати из табеле 5), корисници Итиријума немају такав поглед на ове две валуте. Одсуство симетрије у ставовима може бити и последица аномалије анализираног периода, те ово питање захтева даља испитивања. Поред сентимента вести о Биткоину, значајни утицај забележили су константа и ауторегресивна компонента првог реда.

Табела 8: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *ETH* (оцене сентимента нормиране на интервал [-1,1])

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.007	0.219	0.011	0.840	0.001	0.056	27.97	1831
п-вредност	0.1859	0.0000	0.8352	0.0002	0.8200		0.0000	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.008	0.216	-	0.842	-	0.058	55.96	1831
п-вредност	0.0093	0.0000	-	0.0001	-		0.0000	

Наредну разматрамо крипто-валуту *LUNA* чија је оцена модела (12) дата у табели 9. Оцена пуног модела у себи садржи само један статистички значајан параметар. Ипак, постепеним елиминисањем ирелевантних регресора из ње дати параметар је престао да буде значајан, док су се истовремено искристалисали неки други релевантни регресори. У крајњем моделу поново наилазимо на ситуацију у којој су оба показатеља сентимента статистички значајна. То нам говори да су ставови инвеститора формирани на бази онлајн вести одиграли битну улогу у детерминисању кретања приноса ове крипто-валуте. Позитиван предзнак уз сентимент вести о самој крипто-валути указује да су вести о њој имале позитиван утицај на кретање њених приноса, што је и било очекивано. Међутим, још једном неочекивано наилазимо на позитиван

предзнак уз сентимент вести о Биткоину. Дисертација ће понудити исто објашњење овог феномена. Позитиван предзнак вероватно указује на то да корисници крипто-валуте *LUNA* не доживљавају Биткоин као конкурентску валуту.

Табела 9: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *LUNA* (оцене сентимента нормиране на интервал [-1,1])

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.089	0.175	0.283	0.608	-0.001	0.043	3.241	202
п-вредност	0.1220	0.0082	0.1133	0.2368	0.8087		0.0133	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.027	-	0.171	0.315	-	0.045	5.78	202
п-вредност	0.0677	-	0.0095	0.0532	-		0.0036	

На самом крају осврнимо се и на оцену модела (12) за случај *SOL* коина. Оцене пуног и редукованог модела дате су табелом **10**. Из табеле је евидентно да константу, ауторегресивни параметар првог реда и читљивост објављених вести модел није препознао као статистички значајне предикторе приноса. Њиховим изостављањем долазимо до редукованог модела. Оцењени модел се издваја по томе што је једини од свих до сада анализираних оцена имао статистички значајну негативну константу. То нам сугерише да је *SOL* већином остваривао негативне приносе. Изузев тога ситуација је готово идентична као код крипто-валуте *LUNA*. Поред константе, значајни предиктори су још само показатељи базирани на сентименту онлајн вести. То је доказ да су приноси ове крипто-валуте били детерминисани ставовима инвеститора. Поново је предзнак код оба показатеља позитиван, што нас наводи на исте закључке као и у претходном случају.

Табела 10: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *SOL* (оцене сентимента нормиране на интервал [-1,1])

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-0.111	-0.036	0.4725	2.326	-0.002	0.027	4.095	448
п-вредност	0.0021	0.4355	0.0216	0.0051	0.1805		0.0029	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-0.125	-	0.432	2.365	-	0.031	7.021	448
п-вредност	0.0003	-	0.0340	0.0040	-		0.001	

5.2.2 Генерализовани преглед резултата

Кроз здружени преглед свих претходних резултата покушаћемо да добијемо општу слику стања на тржишту крипто-валута и испитамо прве две постављене хипотезе. Будући да су у истраживање укључене крипто-валуте, које су важиле за највеће и најпопуларније у време његовог спровођења, очекивано је да закључке које изведемо на бази њих важе за цело крипто-тржиште. Кренимо од хипотезе H_A. Код четири од анализираних осам крипто-валута установљено је да сентимент вести написаних о њима има статистички значајан утицај на њихове приносе. Занимљиво је да су све четири крипто-валуте (*ADA*, *DOT*, *LUNA* и *SOL*) јако младе, будући да постоје свега неколико година. Једина крипто-валута новијег датума код које није пронађена значајна веза између сентимента вести о њој и приноса је AVAX. Ова веза није пронађена ни код крипто-валута *BTC*, *DOGE* и *ETH*. Приметимо и то да је у сва четири случаја код којих је веза била значајна детектован очекивани предзнак. Према томе, резултати сугеришу да се сентимент информација из онлајн вести или не одражава на приносе крипто-валута или се одражава позитивно. Са друге стране, занимљиво је и то да код две најпознатије крипто-валуте (Биткоин и Итиријум) није препознат статистички значајни утицај сентимента вести. Будући да се о обе крипто-валуте пуно пише, инвеститорима је била доступна велика количина информација. Ипак, све те доступне информације нису могле да објасне кретање приноса у анализираном периоду.

Друга постављена хипотеза тематизује два показатеља – читљивост вести објављених о датој крипто-валути и утицај унакрсног сентимента. Читљивост се испоставила као јако плитак предиктор. Њен утицај је био значајан у свега два од осам анализираних случајева. Могуће објашњење за овај резултат је то да су текстови углавном лако читљиви и да је добар ниво читљивости постојан кроз време. Приликом започињања овог истраживања пошло се од претпоставке да кретање и промене у нивоу читљивости могу давати одређене сигнале инвеститорима. Примера ради, људи углавном компликују своје излагање када желе да слажу, као и када желе да сакрију или ублаже истину. Из тог разлога, уколико би се појаво период у којем се објављују компликоване и нејасне вести инвеститор би могао да закључи да нешто на тржишту није уреду. Емпиријски резултати показали су да се читљивост објављених вести не мења претерано кроз време, те су сигнали које добијамо из ње танки. Изузев плиткости утицаја читљивости објављених вести, ниједан други закључак се не може генерализовати. Једине две крипто-валуте код којих је препознат значајан утицај сигнала добијених из читљивости су *DOGE* и *DOT*. Са једне стране имамо *DOGE* коин који се сврстава у ред старијих крипто-валута, док се на другој страни налази Полкадот, једна од најмлађих крипто-валута. Док се о *DOGE* коину јако пуно пише (што је очекивано, будући да је због свог порекла ова крипто-валута изразито популарна у јавном мњењу), број објављених вести о Полкадоту је много мањи. Према томе животни век и популарност нису фактори који предодређују значајност утицаја читљивости објављених вести. Закључци се не могу генерализовати ни о знаку везе између читљивости и приноса. У случају Полкадота ова веза је била негативна, што упућује на закључак да јасно читљиви текстови дају позитиван сигнал за инвестирање. Насупрот томе, код *DOGE* коина анализа је показала да је веза имала позитиван предзнак. Резултат упућује на то да су нејасни текстови давали позитиван сигнал за инвестирање, што се може објаснити пореклом ове крипто-валуте (о чему се дискутовало у претходном одељку). Још нешто је занимљиво у вези са добијеним резултатима. Читљивост вести није опредељивала приносе две крипто-валуте које су према броју објављених чланака далеко испред осталих. Захваљујући

таквој масовности објављених чланака, Биткоин и Итиријум имају највећу дисперзију тема објављених вести. Очекивано је да осцилације у читљивости њихових вести буду највеће, али ни из тога инвеститори нису могли да добију ваљане инвестиционе сигнале.

Други показатељ обухваћен другом хипотезом је утицај унакрсног сентимента. Питање његове значајности утолико је интересантнији, будући да је ово још увек недовољно истражена тема у постојећој литератури. Унакрсни сентимент се испоставио као јако користан предиктор. Његова значајност није потврђена само у једном од осам анализираних случајева. То нам говори да су инвеститори били склони да на тржишту реагују на нове информације о тржишном лидеру Биткоину, као и да се та чињеница може употребити за предвиђање будућег понашања приноса већине крипто-валута. Иако је постојање ове везе препознато код већине крипто-валута, њен предзнак није био конзистентан. То отвара јако занимљиво питање његове интерпретације. Ова дисертација понудила је једну могућу интерпретацију знака везе између приноса и унакрсног сентимента из угла конкурентности крипто-валута. Негативан предзнак ове везе подразумевао би да корисници посматране крипто-валуте Биткоин доживљавају као њеног употребног конкурента. До тог закључка долазимо јер негативан предзнак подразумева да позитивни ставови о Биткоину врше притисак на пад приноса посматране крипто-валуте. Онда мора бити да Биткоин као валута (пре свега средство плаћања) одвлачи део тражње за посматраном валутом. Уместо да рударе посматрану крипто-валуту или да своје трансакције обављају њом, људи ће се определити за Биткоин због позитивних ставова који круже о њему. Мањак интересовања ће се одразити на пад тражње, а самим тим и на приносе посматране крипто-валуте. Уколико је знак везе позитиван, закључили бисмо да корисници посматране крипто-валуте не доживљавају Биткоин као њеног употребног конкурента. Позитивне вести о Биткоину биле би примљене као позитиван сигнал за целу класу крипто-валута. Самим тим, ове вести би их додатно мотивисале да користе и рударе посматрану валуту, што ће се последично одразити и на њену тражњу и на њену вредност. Код анализираних крипто-валута примећена су оба разматрана случаја. Резултати су показали да корисници крипто-валута *ADA*, *DOGE* и *DOT* доживљавају Биткоин као употребног конкурента. Са друге стране, корисници крипто-валута *ETH*, *LUNA* и *SOL* нису доживљавали Биткоин као свог употребног ривала. Кад је реч о самом Биткоину, утицај унакрсног сентимента испитиван је преко Итиријума. Испоставило се да за разлику од корисника Итиријума, корисници Биткоина Итиријум доживљавају као употребног конкурента. Приметимо да перцепција коју један тржишни сегмент има о другом не мора бити симетрична (о чему сведочи пример Биткоина и Итиријума). У сваком случају, други део друге хипотезе се испоставио као валидан, а утицај унакрсног сентимента је препознат као важан предиктор у већини оцењених релација.

Када све сумирамо видимо да прва хипотеза није била одбачена у половичном броју случајева, да је први део друге хипотезе имао минорне успехе, док други део друге хипотезе није био одбачен у великој већини случајева. Ови резултати бацају сумњу на то да цене крипто-валута ипак не одражавају све јавно доступне информације. Приметно је да се кретање приноса у већини анализираних случајева могло објаснити утицајем које вести имају на ставове корисника крипто-валута. Томе треба додати и чињеницу да је у већини случајева била значајна и ауторегресивна компонента што упућује на могуће одсуство слабе форме тржишне ефикасности. Претходни резултати дају сигнале потенцијалне нарушености тржишне ефикасности на анализираним тржиштима крипто-валута, које је потребно даље испитати. До краја ове дисертације позабавићемо се питањем слабе форме тржишне ефикасности, док ће се

питање полу-јаке форме тржишне ефикасности испитати засебним истраживањем. На крају треба рећи и то да је најмање сигнала о тржишној неефикасности примећено код крипто-валуте *AVAX*, будући да у њеном случају ниједан од предиктора није успео да објасни њене приносе.

Не смемо изгубити из вида да приказани резултати говоре о периоду великих промена и нестабилности. Заоштравања односа на међународној сцени и почетка рата на истоку Европе били су покретачи негативне спирале привредних циклуса на светском нивоу. Последично, у периоду обухваћеном узорком у заинтересованој јавности постепено је завладала паника. Конкретно, у случају Биткоина то је довело до великог пада цене. На почетку анализираног периода цена Биткоина је износила приближно 38 000$, док је на крају овог периода износила приближно 26 000$. Паника се наставила и после узорачког периода што је девастирало тржиште крипто-валута. Том приликом многе крипто-валуте су престале да постоје. То је прва криза светских размера која погађа тржиште крипто-валута. Током светске економске кризе 2008. године још увек су биле у процесу настанка, а трговање њима још није заживело. Док се остатак света од 2019. борио са пандемијом КОВИД19 и њеним последицама, крипто-валуте поново су остале ушушкане. Онлајн трговина и плаћање које је пандемија са собом донела промотивно су деловали на њиховим тржиштима. Овога пута, крипто-валуте нису остале имуне на чари кризе. То даје посебну драж приказаним резултатима будући да сада можемо да сагледамо њихова тржишта у условима трансмисије глобалне кризе. Како анализирани период осликава сам почетак ове кризе, несумњиво је да се један део уочених сигнала неефикасности дугује негативним нетржишним утицајима.

5.3 Трећа етапа: предикција

У претходној етапи одређени су фактори који су опредељивали кретање приноса код сваке од крипто-валута посебно. У трећој етапи идентификовани фактори биће употребљени за предвиђање будућег кретања приноса. За потребе ове етапе формиран је трећи подузорак који обухвата период 01.03.2022. – 23.03.2022. Примарни циљ овог подузорка биће да упореди предиктивну моћ предложене алтернативне процедуре (засноване на побољшаним оценама) са оном коју има оригинална процедура (заснована на оценама Џагадиша и Вуа 2019). Уједно, резултати добијени у овој етапи ће нам помоћи да боље сагледамо емпиријску предвидљивост кретања приноса на бази јавно доступних информација.

Раздвајањем узорка за одабир предиктора од узорка за предвиђање обезбеђује се објективност резултата, јер се успешност предвиђања проверава над подацима које модел није видео. Одабир ове две целине се мора обавити пажљиво, како не бисмо дошли у ситуацију у којој издвојени предиктори нису били детерминанте кретања приноса у узорку за предвиђање. Зато је важно да ови узорци временски нису превише удаљени један од другог. На тај начин не оставља се простор у којем би било времена за значајније измене у факторима који су опредељивали кретање приноса. Осим тога, битно је да ни један од два узорка не обухвата превише дугачак временски хоризонт. Што су узорци дужи, то ће између њих бити мање сличности, јер се економске околности мењају протоком времена. Последично, уочени фактори у једном од њих вероватно не би играли значајну улогу и у оном другом. Узевши у

обзир претходно изнете ставове, као и то да је циљ предвидети приносе у марту, одлучено је да узорак из друге етапе обухвата период јануар-фебруар 2022. Узорак је довољно дугачак да обезбеди адекватан број текстова за анализу предиктора, а опет довољно кратак тако да није превише одаљен од марта, за који вршимо предвиђање. Ни сам мартовски узорак није превише дугачак. Захваљујући томе подаци с краја овог узорка нису превише удаљени од узорка за анализу предкитора. Према томе, будући да су оба узорка кратка и временски блиска, имамо разлога да верујемо да ће кретање приноса у оба узорка бити опредељено истим факторима. Аргументација о адекватности се може појачати анализом друштвених околности у оба периода. Главно обележје трећег подузорка је свакако рана фаза сукоба на истоку Европе. Истовремено, други подузорак обухвата предратни период и прве дане рата. Овај период одликују предратна психоза, тензије и паника што је најприближније условима самог рата. Ослањајући се на то, можемо претпоставити да су у оба периода приноси били детерминисани истим факторима.

5.3.1 Поређење прогноза

За потребе предвиђања података из трећег подузорка алгоритам описан у одељку **3.8.1** примењен је два пута. Најпре је то урађено по алтернативној методологији, коју заступа ова дисертација (у ознаци А), а затим по методологији Џагадиша и Вуа (2019) (у ознаци *JW*). Код оба модела коришћени су статистички значајни предиктори идентификовани у другој етапи истраживања (секција **5.2.1**). Међутим, у случају крипто-валуте *AVAX* није било статистички значајних предиктора. У складу са традицијом машинског учења, у тим околностима за ову крипто-валуту задржан је пун модел оцењен у табели 4. Захваљујући тој одлуци нећемо изгубити један узорак за проверу квалитета прогнозе.

Табела 11: *RMSE* алтернативног (оцене сентимента нормиране на интервал [-1,1]) и оригиналног *JW* модела код осам одабраних крипто-валута

Валута	**RMSE$_A$**	**RMSE$_{JW}$**
ADA	0.044351	0.047979
AVAX	0.041549	0.043248
BTC	0.016031	0.016519
DOGE	0.028968	0.029042
DOT	0.040575	0.041923
ETH	0.012619	0.013257
LUNA	0.057144	0.060289
SOL	0.038751	0.03893

За почетак представимо тачкасте оцене корена из средње квадратне грешке прогнозе два модела. Вредности ове статистике дате су табелом **11**. Приметимо да су разлике између нивоа ових статистика углавном на трећој децимали, као и да су у свих осам случајева код алтернативног модела забележене ниже грешке прогнозе. Највећа разлика препозната је код крипто-валуте *ADA*, а најмања код крипто-валуте *DOGE*. Изложени резултати нам указују на умерено побољшање квалитета прогнозе код алтернативног модела. Иако су приказани

резултати охрабрујући, да би закључак смели да генерализујемо, неизоставно је спровести статистичко тестирање.

Анализирајмо најпре резултате предложеног омнибус Диболд-Маријановог теста. За потребе тестирања формирана је структурна серија org_l као разлика средње квадратних грешака алтернативног и оригиналног модела. Након њеног регресирања на константу добијена је следећа оцена јединог регресионог параметра $c = -0.0001$. Након примене Њуи-Вестове корекције обрачуната је статистика теста[71] $ODM = -2.102$. Будући да је статистици теста придружена п-вредност 0.036, на нивоу значјности од 5% нулта хипотеза се одбацује. Према томе, закључујемо да између модела постоји статистички значајна разлика у погледу квалитета прогнозе. Како је оцењена константа негативна, закључујемо да је алтернативни модел понудио прецизније прогнозе од оригиналног као што су тачкасте оцене и сугерисале. Како резултати омнибус Диеболд-Маријанов тест говоре у прилог првом предложеном побољшању, треће постављена хипотеза би била испуњена. То нас наводи на закључак да је могуће подићи квалитет прогнозе одабиром квалитетнијих мера сентимента. Међутим, како је тест осетљив на број узорака, закључке треба проверити и МекКракеновим (2000) тестом.

Анализа је настављена спровођењем МекКракеновог (2000) теста. За потребе тестирања дефинисана је нова векторска временска серија f_t. Као што је описано у одељку **3.9.5**, она представља разлику квадрата грешака прогнозе алтернативног и оригиналног модела код свих осам крипто-валута. У очекивању да алтернативни модел понуди прецизније прогнозе, тестом проверавамо да ли је очекивана вредност векторске временске серије f_t негативна. У табели **12** дата је оцена осмодимензионе статистике теста χ из израза (26). Она је важан корак у тестирању јер се из ње изводи финална статистика теста. Осим тога, знаци њених елемената могу дати сигнал који модел је бољи. Приметимо да су сви елементи векторске статистике негативни. То указује да је у свих осам случајева први (алтернативни) модел дао ниже грешке прогнозе од другог (оригиналног) модела. На тај начин, још једном потврђујемо претходно изнету тезу да ће тест потврдити или оповргнути да алтернативни модел генерално даје ниже грешке прогнозе. Након тога је према изразу (32) обрачуната једно-димензиона статистика теста $\chi_8^2 = 10.51$. Критична вредност Хи квадрат расподеле са 8 степени слободе на нивоу значајности од 5% износи 15.51, док на нивоу значајности од 10% износи 13.36. Резултат сугерише да у погледу квалитета прогнозе не постоји статистички значајна разлика између два модела. Према томе, можемо закључити да уочена побољшања нису била довољно велика да бисмо на основу њих смели да тврдимо да алтернативни модел предвиђа значајно боље. Строго формално гледано, овај тест одбацује трећу хипотезу. Без обзира на то, не треба изгубити из вида да минорна побољшања ипак постоје (мањи корени из средње квадратне грешке прогнозе и негативне вредности у 8-димензионој статистици).

Табела 12: Приказ помоћне осмодимензионе статистике χ Мек-Кракеновог теста (оцене сентимента нормиране на интервал [-1,1])

	ADA	AVAX	BTC	DOGE	DOT	ETH	LUNA	SOL
оцена	-1.1752	-0.9772	-1.6146	-0.0249	-1.5139	-1.2457	-1.3086	-0.0736

[71] За ознаку статистике теста искоришћена је скраћеница *ODM* – Омнибус Диболд Маријанов.

5.4 Остали резултати

У овој секцији осврнућемо се на још два резултата. Прво ћемо се позабавити питањем квалитета другог предложеног побољшања у мерењу сентимента речи које се базира на апроксимацији са кумулативним приносима (видети израз (91)). У те сврхе поново ће бити оцењена једначина (12) за сваку крипто-валуту посебно, при чему ће се за мерење сентимента користити израз (91) уместо израза (98). Тиме ће се проверити до каквих би се закључка дошло да је сентимент обрачунат на другачији начин, као и да ли ово побољшање може да да значајно квалитетније прогнозе. Након тога, проверићемо како ће се на моделе одразити увођење информације о типу текста добијене из кластеризације методом *K*-средњих вредности. На крају, истраживање се завршава провером слабе форме тржишне ефикасности код осам одабраних крипто-валута.

5.4.1 Анализа базирана на апроксимативним оценама

Другачије оцењивање сентимента довело би до идентификовања другачијих предиктора, али и до другачијих прогноза. Из тог разлога, за сваку крипто-валуту поново је оцењена једначина (12) са новим мерама сентимента. У наставку су представљени добијени резултати. Код већине крипто-валута добијени су идентични закључци. Ипак, код три крипто-валуте закључци су благо модификовани. Из тог разлога нећемо се детаљније освртати на моделе крипто-валута који сугеришу исте предикторе и њихове везе. Уместо тога позабавићемо се само моделима који сугеришу другачије закључке од претходно изнетих.

Кренимо од крипто-валуте *ADA*. Оцене пуног и редукованог модела дате су табелом **13**. Из табеле се јасно види да у случају ове крипто-валуте закључци остају непромењени.

Табела 13: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *ADA* (апроксимативне оцене сентимента)

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.172	0.084	0.014	-0.245	-0.001	0.136	20.91	508
п-вредност	0.0000	0.0181	0.0512	0.0000	0.3258		0.0000	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.167	0.085	0.011	-0.246	-	0.136	27.56	508
п-вредност	0.0000	0.0171	0.0904	0.0000	-		0.0000	

Закључци се нису променили ни у случају крипто-валуте *AVAX*, те поново имамо ситуацију у којој ни један од предиктора није статистички значајан (видети табелу **14**).

Табела 14: Приказ оцењеног пуног помоћног модела за приносе крипто-валуте *AVAX* (апроксимативне оцене сентимента)

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.024	0.009	0.006	-0.041	0.001	-0.021	0.2263	154
п-вредност	0.5785	0.9018	0.7673	0.4180	0.7988		0.9230	

Табела 15: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *BTC* (апроксимативне оцене сентимента)

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{ETH,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-0.030	0.165	-0.008	0.039	0.000	0.031	10.75	1210
п-вредност	0.0514	0.0000	0.1140	0.0326	0.5719		0.0000	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{ETH,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-0.028	0.165	-0.007	0.039	-	0.032	14.24	1210
п-вредност	0.0446	0.0000	0.0940	0.0342	-		0.0000	

Нови резултати за Биткоин дати су табелом **15**. Код ове крипто-валуте уочене су највеће промене. Лако је уочљиво да су константа и сентимент самог Биткоина овога пута статистички значајни предиктори. Ова промена са собом је донела један резултат који у великој мери скреће пажњу на себе. У оцењеном моделу добијен је негативан предзнак уз сентимент Биткоина. Резултат сугерише да се позитивизам у објављеним вестима негативно одражавао на приносе у посматраном периоду. Насупрот томе, очекивано је да позитивне вести позитивно утичу на приносе финансијске активе о којој су објављене. Подсетимо да ово није јединствен случај у постојећој литератури. Негативну везу између сентимента и приноса пронашао је и рад Шумахера и сарадника (2012), али и Дамјановића и Дреновака (2023). Поменути аутори свој резултат објаснили су као аномалију периода у којем је спроведено њихово истраживање. Исту аргументацију понудиће и ова дисертација будући да је у фокусу истраживања и овога пута један преткризни период[72]. Већ је истакнуто да су други подузорак обележили предратна паника и први дани рата, као и да су се њихови ефекти значајно дестабилизовали тржиште крипто-валута. У ово кризно време, чак, ни позитивне вести нису успеле да подигну морал тржишта, већ су имале контра-ефекат. Поред тога, неочекивани предзнак пронађен је и у случају сентимента код крипто-валуте Итиријум. Овога пута знак уз унакрсни сентимент је сада позитиван. Резултати сада сугеришу да корисници Биткоина, ипак, не доживљавају Итиријум као свог употребног конкурента. Такође, могуће је да је и овај резултат само аномалија анализираног периода.

[72] Период обухваћен истраживањем Шумахера и сарадника (2012) непосредно претходи Светској финансијској кризи из 2007. године.

Модел добијен у случају *DOGE* коина дат је табелом **16**. И у овом случају добијени резултати остају исти, те се закључци нису променили у односу на раније анализирани модел.

Табела 16: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *DOGE* (апроксимативне оцене сентимента)

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-0.001	0.116	0.001	-0.012	0.001	0.015	2.905	493
п-вредност	0.9567	0.0041	0.9211	0.5751	0.0572		0.0214	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-	0.116	-	-0.013	0.001	0.019	4.118	493
п-вредност	-	0.0025	-	0.0248	0.0438		0.0067	

Промена нема ни код крипто-валуте Полкадот. Из табеле **17** поново се може уочити да су сви предиктори статистички значајни, те и у овом случају закључци остају исти.

Табела 17: Приказ оцењеног пуног помоћног модела за приносе крипто-валуте *DOT* (апроксимативне оцене сентимента)

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.083	0.204	0.021	-0.093	-0.004	0.08	5.663	215
п-вредност	0.0078	0.0025	0.0398	0.0115	0.0456		0.0002	

Табела 18: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *ETH* (апроксимативне оцене сентимента

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.069	0.230	0.009	-0.102	0.000	0.067	33.74	1831
п-вредност	0.0000	0.0000	0.0710	0.0000	0.7108		0.0000	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.068	0.230	0.008	-0.102	-	0.069	44.96	1831
п-вредност	0.0000	0.0000	0.0695	0.0000	-		0.0000	

Поред Биткоина, промене су примећене и у случају Итиријума. У односу на табелу **8** приметне су две промене. У табели **18** препознат је још један статистички значајан предиктор. У питању је сентимент о самој крипто-валути. То нам говори да су приноси Итиријума, ипак, били одређени ставовима инвеститора формираних на бази онлајн вести. Приметимо и то да је, у складу са очекивањима, добијен позитиван предзнак везе. Друга промена односи се на предзнак уз сентимент Биткоина. Овога пута веза између приноса и унакрсног сентимента

препозната је као негативна. То нас наводи на закључак да корисници Итиријума доживљавају Биткоин као свог употребног конкурента.

Последња крипто валута код које су уочене промене је *LUNA*. Нови резултати приказани су у табели **19**. За разлику од Биткоина и Итиријума, код којих је препозната значајност нових предиктора, овога пута је присутна другачија ситуација. Наиме, приметимо да константа више није статистички значајан предиктор. Постепеним редуковањем модела испоставило се да су једини значајни предиктори сентимент инвеститора о самој крипто-валути и унакрсни сентимент. У редукованом моделу примећујемо још једну промену. Добијен је негативан предзнак уз сентимент саме крипто-валуте као у случају Биткоина. Будући да услед кризе поверења позитиван глас није успео да допре до инвеститора на прави начин, већ је имао контра-ефекат, и овога пута резултат посматрамо као аномалију периода.

Табела 19: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *LUNA* (апроксимативне оцене сентимента

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	0.077	0.174	-0.040	0.074	-0.001	0.006	1.623	202
п-вредност	0.1690	0.0083	0.1006	0.3039	0.8571		0.2	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-	-	-0.038	0.037	-	0.041	3.16	202
п-вредност	-	-	0.0842	0.0763	-		0.0152	

На крају представимо и модел за крипто-валуту *SOL*, дат у табели **20**. На бази приказаних резултата долазимо до истих закључака као код иницијалне анализе. Другим речима, ни у овом случају није било промена.

Табела 20: Приказ оцењеног пуног и редукованог помоћног модела за приносе крипто-валуте *SOL* (апроксимативне оцене сентимента

Пун помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-0.109	-0.043	0.029	0.127	-0.002	0.020	3.276	448
п-вредност	0.0060	0.3526	0.0476	0.0146	0.1352		0.0116	
Редуковани помоћни модел								
	c	r_{t-1}	$S_{k,i,t-1}$	$\bar{S}_{BTC,t-1}$	FI	Кор. R^2	F	Бр. опс.
оцена	-0.121	-	0.024	0.126	-	0.021	5.039	448
п-вредност	0.0020	-	0.0879	0.0148	-		0.007	

Осврнимо се сада на генерализовани преглед добијених резултата. Промене у начину оцењивања сентимента могу да доведу до значајних промена у закључцима. Такав случај је забележен код три од осам посматраних крипто-валута. Пођимо од тога да је промена мере сентимента условила да он буде препознат као значајан предиктор код још две крипто-валуте. Посебно је занимљиво то што су у питању две највеће и најпопуларније крипто-валуте, Итиријум и Биткоин. Резултати базирани на првој мери сентимента (нормираним оценама) сугерисали су да су све информације о ове две крипто-валуте већ укључени у цену. Такав резултат се може бранити популарношћу ових валута. Будући да је о ове две валуте доступно пуно вести из различитих извора, оне брзо допиру до заинтересоване јавности и не могу се искористити за стицање профита. Насупрот томе, резултати базирани на другој мери (апроксимативним оценама) указују на супротан закључак. Објашњење резултата и у овом случају темељимо на популарности двеју валута. Уколико је у току дана инвеститорима доступна (пре)велика количина информација, они не постижу све да их сагледају и обраде, те цене неће одражавати све јавно доступне информације. Два добијена резултата сведоче да је анализа веома осетљива на избор мере сентимента. Поред тога, сентимент вести о самој крипто-валути је сада препознат као значајан предиктор у шест од осам анализираних случајева, те прва хипотеза у већини случајева не би била одбачена. Ипак, то није једина последица промене мере сентимента. У приказаним резултатима приметне су и промене једног броја предзнака. Код две крипто-валуте (*LUNA* и *BTC*) пронађен је негативан предзнак уз сентимент вести о њима, што је приписано преткризним условима. Такође, предзнаци су промењени и код унакрсног сентимента у једначинама за Биткоин и Итиријум. Чињеница да се закључци о везама могу драстично разликовати појачава потребу за проналажењем адекватне мере сентимента. Што се тиче унакрсног сентимента, он је и даље препознат као релевантан фактор у седам од осам посматраних случајева. Дакле, закључци о другој хипотези су исти. Ипак, генерално гледајући, применом апроксимативних оцена сентимент инвеститора (из обе групе вести) се показао као веома важан предиктор. Заправо, приметимо да закључци зависе од квалитета саме мере сентимента. Уколико је тако, могло би се рећи да постоји одређена предвидљивост тржишта крипто-валута, а да бисмо је уочили све што нам је неопходно су боље мере сентимента. Што је мера прецизнија, то ћемо већу предвидљивост наћи. Коначно, приметимо и да су закључци о читљивости текстова и значајности ауторегресивних компоненти остали не промењени.

Користећи апроксимативне оцене сентимента поново је покренут алгоритам предвиђања приноса описан у одељку **3.8.1**. Овога пута коришћени су статистички значајни предиктори идентификовани у овом одељку (табеле **13**-**20**). Као и код претходног поређења, у случају крипто-валуте *AVAX* задржани су сви предиктори. Добијене тачкасте оцене корена из средње квадратне грешке прогнозе приказане су табелом **21**. Приметимо да је код готово свих крипто-валута алтернативни модел забележио боље резултате. Једини изузетак пронађен је код крипто-валуте *LUNA*, где је регистровано да алтернативни модел даје за нијансу лошије резултате. Додатно, апсолутне разлике између грешака прогнозе су нешто мање него у претходном случају. Најмања апсолутна разлика забележена је код крипто-валуте *ADA*, а највећа код *AVAX*.

Табела 21: *RMSE* за алтернативни (апроксимативне оцене сентимента) и оригинални *JW* модел код осам одабраних крипто-валута

	$\mathbf{RMSE_A}$	$\mathbf{RMSE_{JW}}$
ADA	0.04782	0.04798
AVAX	0.0399	0.04325
BTC	0.01605	0.01652
DOGE	0.02887	0.02904
DOT	0.04119	0.04192
ETH	0.01288	0.01325
LUNA	0.05984	0.05938
SOL	0.03813	0.03893

Као и у претходном случају, прво је спроведен омнибус Диеболд-Маријанов тест. Након обрачуна разлика између средњих квадратних грешака (тј. након конструисања структурне серије org_l) исте су регресиране на константу. Оцењена вредност је износила $c = -0.00005$. Приметимо да је оцењена константе била мања у апсолутном износу у односу на претходни случај. Упркос томе, добијена је већа вредност статистике теста $ODM = -2.194$, а њој придружена п-вредност износила је 0.028. Из добијених резултата закључујемо да је и друго предложено побољшање дало мање грешке прогнозе од оригиналног приступа, тј. да је предвиђање на бази алтернативног приступа прецизније. И овај резултат говори у прилог трећој постављеној хипотези, али остаје да се оно провери и МекКракеновим (2000) тестом.

Зарад спровођења МекКракеновог (2000) теста још једном је креирана векторска временска серија f_t као разлика квадрата грешака прогнозе алтернативног и оригиналног модела код свих осам крипто-валута. И овога пута тестом проверавамо да ли је очекивана вредност векторске временске серије f_t негативна, очекујући да алтернативни модел понуди прецизније прогнозе. Табелом **22** приказана је оцена осмодимензионе статистике теста χ из израза (31). Сви елементи векторске статистике су негативни, изузев оног који се односи на крипто-валуту LUNA. Тим резултатом још једном је потврђена теза да ће тест проверити да ли алтернативни модел генерално даје ниже грешке прогнозе од оригиналног. Обрачуната једно-димензиона статистика теста износила је $\chi_8^2 = 20.74$. Будући да је реализована вредност статистике теста већа од критичне вредности на нивоу значајности од 5%, резултати сугеришу да постоји статистички значајна разлика између прогноза два модела. Из тога можемо закључити да је друго предложено побољшање пружило довољно доказа на основу којих се може тврдити да алтернативни модел предвиђа значајно боље. Овога пута треће постављена хипотеза је потврђена и МекКракеновим (2000) тестом. Такав резултат говори у прилог ставу да се адекватним одабиром мера сентимента може повећати предвидљивост приноса.

Табела 22: Приказ помоћне осмодимензионе статистике χ Мек-Кракеновог теста (апроксимативне оцене сентимента)

	ADA	**AVAX**	**BTC**	**DOGE**	**DOT**	**ETH**	**LUNA**	**SOL**
оцена	-0.57495	-0.83032	-3.64091	-0.52062	-0.84277	-1.21108	0.61046	-1.90792

5.4.2 Кратак осврт на утицај врсте текста

Поред промене у мерењу сентимента, истраживање је проверило да ли ће се модел побољшати укључивањем једне нове информације у анализу. У питању је врста објављене вести. Као што је истакнуто у уводној секцији, о крипто-валутама се објављују најразличитије вести. Вести о берзанским дешавањима, практичној примени крипто-валута, инвестиционим саветима на бази техничке анализе, новој технологији и сл. само су неке од тема о којима се може читати. Узевши у обзир њихову разноврсност, пошло се од претпоставке да објаве вести различитог типа неће имати исти утицај на ставове и економске одлуке људи, а самим тим ни на приносе. Уколико је изнети став тачан и врста објављене вести заслужује да се нађе међу испитиваним потенцијалним предикторима. Врста вести је одређена уз помоћ кластеризације методом *K*-средњих вредности, како би се избегло да истраживач самостално лабелише врсту сваке од преузетих вести. На тај начин истраживање је значајно убрзано, будући да сво оптерећење сноси машина, док истраживач не мора да прочита преузете вести. Текстови су најпре претворени у математичке векторе речи, а затим су груписани у кластере према својој сличности. Идеја је да сваки кластер осликава једну врсту текста.

Оптималан број кластера одређен је на бази дијаграма одрона, на начин описан у одељку **3.10**. Дијаграми одрона код сваке од осам крипто-валута приказани су на слици **17**. У свим приказаним случајевима одрон је доста благ, те се оптималан број кластера не може лако одредити. Упркос томе, се може рећи да је оптимални број кластера у свим случајевима између 3 и 5. Из тог разлога експериментисано је са бројем кластера, те је анализа спроведена са 3, 4, односно 5 кластера. Као крајњи резултат код свих осам крипто-валута добијена је дискретна случајна променљива која за сваки текст показује ком кластеру припада, односно, о којој од *K* врсти текста је реч.

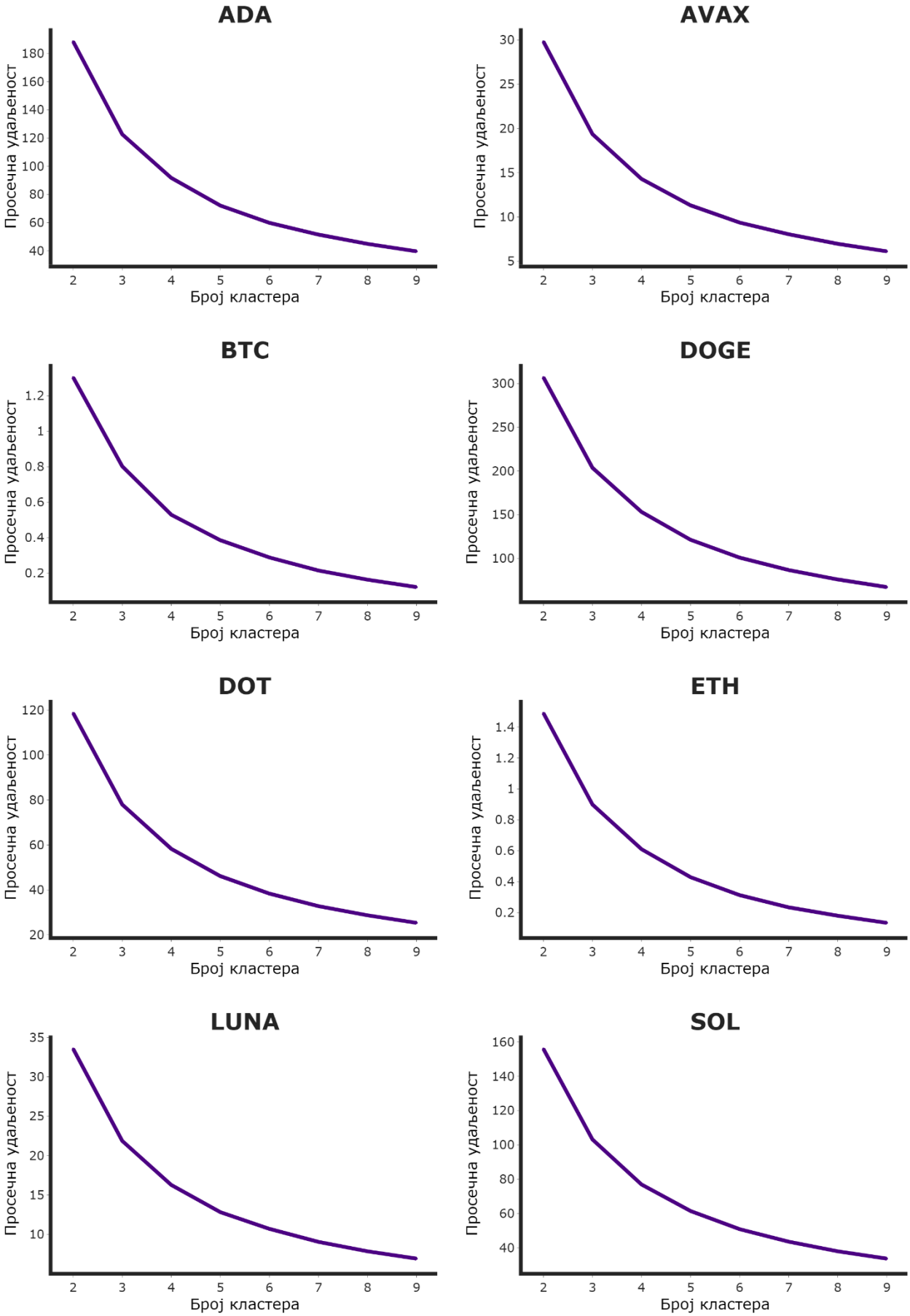
Како регресиона анализа претпоставља непрекидне предикторе, неопходно је моделирати врсте текста (категорије из добијене дискретне променљиве) преко вештачких променљивих. Вештачка променљива (енгл. *dummy variable*) је индикатор променљива која узима вредност 1 уколико текст припада датој категорији и 0 у осталим случајевима. Како би се избегла савршена мултиколинарност (проблем у економетријској литератури познат под називом *замка вештачких променљивих*) уводи се једна вештачка променљива мање од броја категорија. Тако је формиран модификовани модел (12) који укључује вештачке променљиве:

$$r_{k,t} = c + \beta_1 r_{k,t-1} + \beta_2 S_{k,i,t-1} + \beta_3 \bar{S}_{BTC,t-1} + \beta_4 FI_{k,i,t-1} + \gamma_1 I_1 + \cdots + \gamma_{M-1} I_{M-1} + \varepsilon_{k,i} \quad (102)$$

где је: I_m вештачка променљива која моделира m-ту врсту текста, а γ_m параметар модела који стоји уз њу.

Резултати добијени након оцењивања претходне једначине нису подржали увођење ове информације у модел. Ни једна од вештачких променљивих није била статистички значајна, те се врста текста испоставила као јако лош предиктор приноса. То нас наводи на закључак да априори познавање врсте објављених текста не може да нам помогне у предвиђању приноса. Додатно, не може се рећи да вести одређеног типа имају већи, односно, мањи утицај на инвеститоре од осталих врста вести.

Слика 17: Дијаграми одрона код осам одабраних крипто-валута



Извор: Приказ аутора

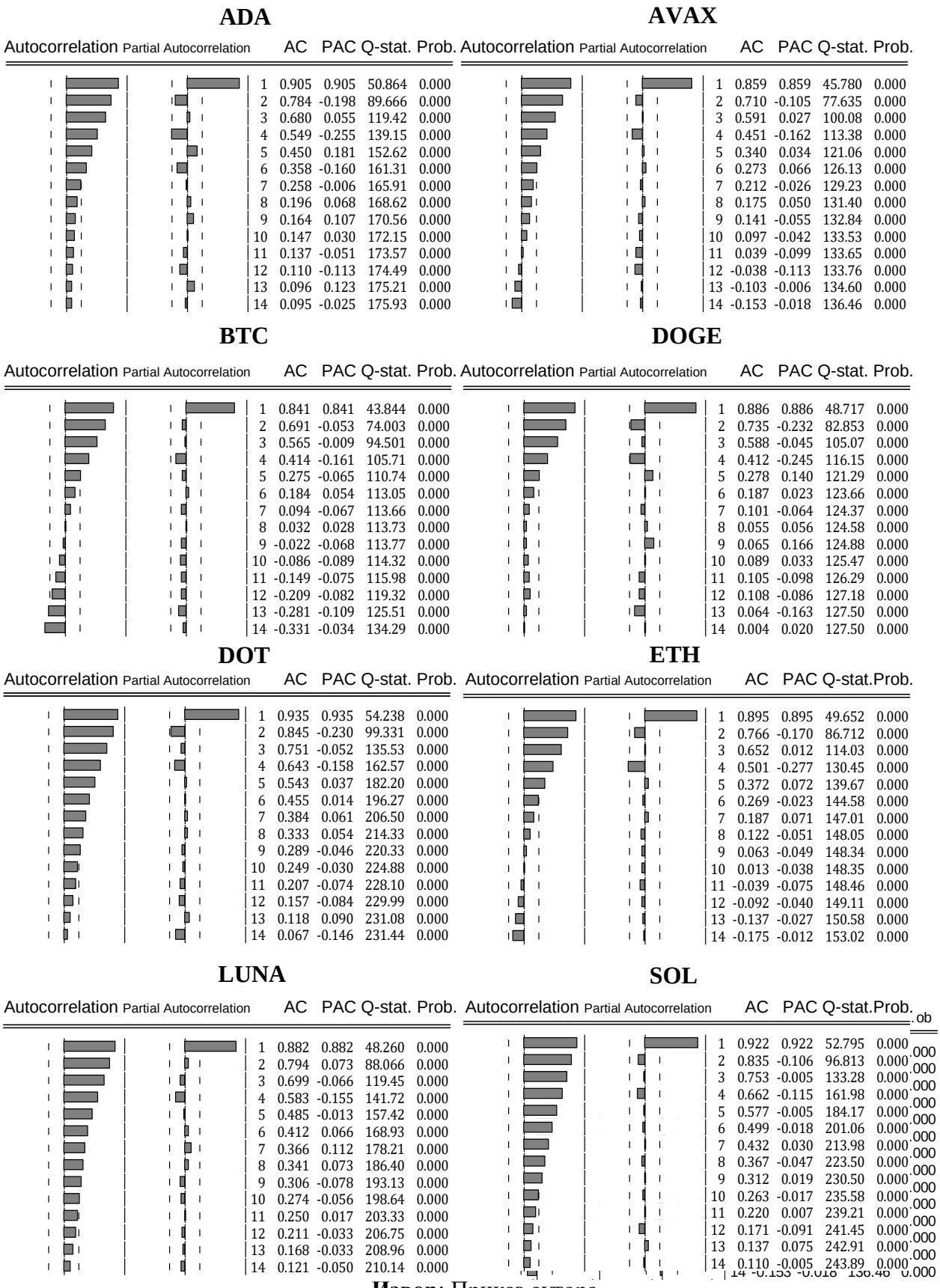
Како је покушај побољшања модела додавањем информације о врсти вести био безуспешан, поменути приступ је напуштен у овом истраживању. Ипак, једно могуће објашњење за овакав резултат је да информација о врсти вести поред информације о сентименту вести делује сувишно. Из тог разлога можда би промена у приступу моделирању дала другачије резултате. С тим у вези, неким наредним истраживањем може се проверити да ли ће обрачун сентимента по категоријама дати боље резултате. У том случају математички вектори вести би се најпре поделили у категорије према својој врсти. Након тога би се за сваку групу вектора посебно применила процедура за оцену сентимента (попут оне спроведене у првој етапи истраживања описаној у одељку **5.4.2**). На тај начин имали бисмо различите процене сентимента за различите врсте текста, што би био алтернативни приступ да се провери да ли одређене групе вести имају већу, односно, мању моћ у предвиђању приноса крипто-валута. Код таквог истраживања треба водити рачуна да се на сваки кластер текстова примени иста процедура оцењивања сентимента зарад очувања једнообразности истраживања. У супротном разлике у нивоу сентимента појединачних речи не би се само дуговале тематици текста (односно подобласти у којој се речи користе), већ и методама њиховог обрачуна.

5.4.3 Слаба форма тржишне ефикасности

У светлу доказа из постојеће литературе познато је да ефикасност тржишта крипто-валута варира кроз време. При томе, много су чешћи периоди током којих тржишта крипто-валута испољавају обрасце неефикасности него ефикасности. Истакнути резултати указују на потребу да се питање слабе форме тржишне ефикасности крипто-валута мора повремено изнова испитати. То је мотивисало аутора дисертације да се и сам позабави питањем слабе форме тржишне ефикасности. Посебну драж овом истраживању даје чињеница да се у његовом фокусу налази један специфичан период. У питању је период зачетка светске економске кризе која се трансмитује на тржиште крипто-валута, што питање ефикасности чини још релевантнијим. Читалац не треба да изгуби из вида да су се закључци из постојеће литературе односили на период историјске стабилности тржишта крипто-валута. Као такви послужиће као добар бенчмарк за поређење ефикасности у кризном и стабилном периоду. Ипак, очекивано је да закључци буду исти, тј. да тестови слабе форме тржишне ефикасности представљени у методолошкој секцији овог рада покажу да су анализирана тржишта претежно неефикасна. Иако су се тржишта крипто-валута развијала великом брзином током времена, што је допринело подизању њихове ефикасности, посматрани период је много нестабилнији од периода у којима су истраживања спровођена до сада. Такве околности могу условити да све неефикасности и слабости ових тржишта дођу до изражаја. Додатна потврда за овакве сумње је сигнал добијен из помоћне регресије. Код чак 5 крипто-валута пронађена је значајна ауторегресивна компонента (*ADA*, *BTC*, *DOGE*, *DOT* и *ETH*) што слути на неслучајно и предвидиво понашање њихових приноса. У овом делу истраживања посматрани су дневни подаци о логаритмованим ценама и логаритамским приносима из другог подузорка[73] (да подсетимо, реч је о периоду: 01.01.2022. – 28.02.2022.). Будући да се трансакције о крипто-валутама обављају свакога дана без изутетка, узорком је обухваћено 59 опсервација за сваку од осам крипто-валута. На слици **18** приказани су корелограми нивоа логаритмованих цена.

[73] Будући да је други подузорак предодређен за анализу утицаја.

Слика 18: Корелограми логаритмованих цена осам одабраних крипто-валута



Извор: Приказ аутора

Корелограми крипто-валута *LUNA* и *SOL* изгледају као корелограми случајног хода. Са друге стране, код корелограма крипто-валута *AVAX*, *BTC*, *DOGE* и *ETH* то није случај. Њихове корелограме одликује брзо опадање нивоа обичне аутокорелационе функције, умерена висина првог парцијалног аутокорелационог коефицијента, и значајност још по неког парцијалног аутокорелационог коефицијента. Код крипто-валута *DOT* и *ADA* тешко је пресудити пре покретања формалних статистичких тестова. Иако њихова обична аутокорелациона функција опада споро, при чему је само први парцијални аутокорелациони коефицијент значајан и висок, постоји још пар парцијалних аутокорелационих коефицијената чији се ниво налази близу статистичке значајности. Ипак, не треба сметнути с ума да су обична и парцијална аутокорелациона функција пристрасне оцене правих вредности обичних и парцијалних аутокорелационих коефицијеанта на нивоу популације (видети Младеновић и Нојковић 2018). Последично, могуће је да је њихова висина последица само шума у подацима, али се то мора проверити формалним статистичким тестовима.

На почетку формалне статистичке анализе представићемо резултате Сток-Вотсоновог (1989) теста. Резултати су дати табелом **23**. Из приложених резултата се недвосмислено види да константа није значајна ни у једној од регресија. Све тачкасте оцењене вредности су веома близу 0. Једина валута код које је оцена константе иоле већа је SOL, али то и даље није довољно да се може сматрати значајном. У складу са приказаним можемо закључити да тренд не треба да се нађе ни у једној од осам спецификација за АДФ и КПСС теста. Будући да све цене осцилирају око ненулте средње вредности, користиће се спецификација која од детерминистичких компоненти садржи само константу.

Табела 23: Резултати Сток-Вотсоновог теста за осам посматраних крипто-валута

	ADA	**AVAX**	**BTC**	**DOGE**	**DOT**	**ETH**	**LUNA**	**SOL**
c	-0.005	-0.004	-0.001	-0.004	-0.005	-0.004	0.001	-0.009
п-вредност	0.4621	0.6084	0.8104	0.4851	0.4534	0.5493	0.9186	0.2708

Табела 24: Резултати АДФ теста јединичног корена за осам посматраних крипто-валута

	ADA	**AVAX**	**BTC**	**DOGE**	**DOT**	**ETH**	**LUNA**	**SOL**
$\boldsymbol{\varphi_1}$	-0.078	-0.174	-0.195	-0.165	-0.097	-0.147	-0.096	-0.072
$\boldsymbol{\tau_\mu}$	-1.413	-2.808*	-2.766*	-2.720*	-2.607*	-2.919**	-1.847	-2.084
p	4	3	3	3	3	3	3	0
Критичне вредности τ_μ:	Ниво значајности 1%:				-3.548			
	Ниво значајности 5%:				-2.912			
	Ниво значајности 10%:				-2.594			

Табелом **24** представљени су резултати АДФ теста. Звездицом су назначене стационарне серије на нивоу значајности од 10%, а са две звездице стационарне серије на нивоу значајности од 5%. Приметимо да је АДФ тест потврдио да се 5 серија не понаша као случајан ход (будући да не садрже једининични корен), као и да се оне углавном подударају са онима код којих су примећени сигнали неефикасности током претходног моделирања. Приметимо и то да

логаритмоване цене крипто-валуте *ADA* ипак нису случајан ход, док су логаритмоване цене крипто-валуте *DOT* на граници значајности.

Резултати КПСС теста представљени су у Табели **25**. Једном звездицом означене су серије логаритмованих цена које се не понашају као случајан ход на нивоу значајности 5%, док су са две звездице означене оне код којих то можемо закључити на нивоу значајности 10%[74]. Тест је потврдио стационарност само код три крипто-валуте: *AVAX*, *BTC* и *ETH*. Међутим, приметимо да ако ниво значајности спустимо на 1% њима би се придружила и серија логаритмованих цена крипто-валуте *DOGE*.

Табела 25: Резултати КПСС теста јединичног корена за осам посматраних крипто-валута

	ADA	**AVAX**	**BTC**	**DOGE**	**DOT**	**ETH**	**LUNA**	**SOL**
КПСС	0.723	0.305**	0.249**	0.671	0.729	0.445*	0.512	0.754
Критичне вредности КПСС:	Ниво значајности 1%:				0.739			
	Ниво значајности 5%:				0.463			
	Ниво значајности 10%:				0.347			

Сагледавши све наведене доказе можемо да извучемо закључке о понашању анализираних серија логаритмованих цена. Будући да су и анализа корелограма и два формална статистичка теста дале исте закључке за крипто-валуте *AVAX*, *BTC* и *ETH* можемо да тврдимо да се њихово кретање не може описати као случајан ход. Такође, тестови су сагласни да цене крипто-валуте *ADA*, *LUNA* и *SOL* могу да се опишу као случајан ход. Међутим, када је реч о крипто-валутама *DOGE* и *DOT* нема консензуса. Анализирајмо оба случаја појединачно. У случају крипто-валуте *DOT* АДФ тест је на граници значајности, КПСС тест чврсто одбацује стационарност њених цена, док корелограм изразито изгледа као корелограм случајног хода. У светлу датих околности одлучено је да логаритмоване цене крипто-валуте *DOT* ипак садрже јединични корен. Када говоримо о другој спорној серији приметимо да њен корелограм апсолутно не изгледа као корелограм случајног хода. Томе треба додати и да је АДФ тестом потврђено да серија није случајан ход, као и да се исти закључак може добити КПСС статистиком на нивоу значајности од 1%. Последично, можемо закључити да се серија логаритмованих цена крипто-валуте *DOGE* не може описати као случајан ход. Сумирајући резултате тестова јединичног корена можемо да закључимо да тржишта четири крипто-валуте (*AVAX*, *BTC*, *DOGE* и *ETH*) нису слабо ефикасна. Њихове цене не садрже јединични корен, те се не могу описати као случајан ход и крећу се на предвидив начин. Када је реч о преостале четири крипто-валуте (*ADA*, *DOT*, *LUNA*, *SOL*) присуство јединичног корена је потврђено. Међутим, пре доношења закључка о ефикасности потребно је да проверимо понашање њихових приноса.

Закључке о слабој форми тржишне ефикасности ћемо обогатити анализом случајности приноса. Резултати провере случјаности приноса Бартелсовим (1982) тестом дати су Табелом **26** за свих осам крипто-валута. Звездицом су означене валуте чије промене приноса нису биле случајне, тј. непредвидиве. Неслучајно понашање приноса детектовано је код две крипто-валуте: *BTC* и *DOGE*. Реализоване вредности *RVN* статистике теста су у оба случаја мање од 2

[74] Како су хипотезе код КПСС изокренуте, значајност од 10% је сада убедљивији доказ стационарности.

индикујући да постоје трендови у кретању приноса. Ови закључци додатно оспоравају слабу форму тржишне ефикасности BTC-а и умањују закључке о слабој форми тржишне ефикасности *DOT*-а.

Табела 26: Резултати Бартелсовог теста случајности код осам посматраних крипто-валута

	ADA	**AVAX**	**BTC**	**DOGE**	**DOT**	**ETH**	**LUNA**	**SOL**
RVN	1.813	1.955	1.667*	1.827	1.650*	1.780	1.814	1.961
п-вредност	0.2364	0.4312	0.0988	0.2529	0.0877	0.1990	0.2368	0.4402

Финални тест слабе форме тржишне ефикасности је проналажење случајних процеса које могу описати логаритмоване цене, односно принос анализираних крипто-валута. На тај начин потврдићемо да они не следе теоријске процесе које би морали да следе да је тржиште заиста слабо ефикасно. У случају логаритмованих цена у питању је случајан ход, док је то бели шум у случају приноса. Табела **27** сумира резултате. У њој су приказане моделиране серије, оцењене спецификације, информације о редукованости[75], Жак-Бера (1980) и Бокс-Љунг (1978) статистике. Као што се може видети из табела датих у прилогу ове дисертације, све доцње укључене у модел су статистички значајне. Поред тога како су све Жак-Бера (1980) статистике мање од 5.99 евидентно је да су резидуали свих оцењених модела нормално расподељени. Приказане Бокс-Љунгове (1978) статистике сведоче да нема аутокорелације резидуала до 10. реда (придружене им п-вреднсоти су веће од 0.1). То се детаљно може проверити и увидом у корелограме резидуала приказане сликом **19**. Са слике је јасно да се сви аутокорелациони коефицијенти (и парцијални и обични) налазе унутар дозвољених граница. Према томе, изнети докази сведоче о економетријској стабилности оцењених модела. Захваљујући томе можемо да видимо процесе које прате логаритмоване цене *AVAX*-а, *BTC*-а, *DOGE*-а и *ETH*-а, као и приноси крипто-валута *ADA, DOT* и *LUNA*. Из приказаног може се закључити да њихова тржишта нису у потпуности слабо ефикасна јер постоји предвидљивост приноса и/или цена. Једина крипто-валута за коју дисертација није успела да пронађе адекватан модел је *SOL*. Тестови јединичног корена једногласно су закључили да њене логаритмоване цене прате случајан ход, док са корелограма приказаног на слици **19** види да се њени приноси понашају као бели шум. Додатно, тест случајности је потврдио да су промене приноса случајне и да нема систематичног кретања. У складу са тим можемо да закључимо да приказани тестови нису могли да оповргну слабу форму тржишне ефикасности.

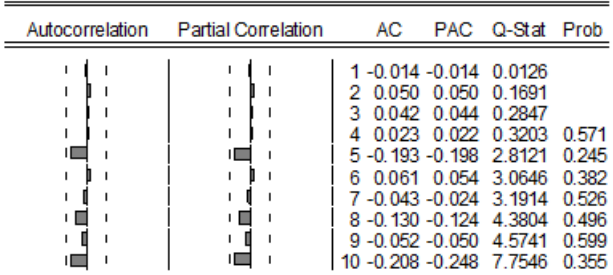
Табела 27: Спецификације оцењених модела код одабраних крипто-валута (сумарни преглед табела из прилога)

	ADA	**AVAX**	**BTC**	**DOGE**	**DOT**	**ETH**	**LUNA**
Серија	r	ln(P)	ln(P)	ln(P)	r	ln(P)	r
Спецификација	MA(7)	AR(1)	ARMA(1,13)	ARMA(1,5)	MA(7)	ARMA(4,5)	AR(3)
Редуковани	Да	Не	Да	Да	Да	Да	Да
JB	2.278	1.724	4.567	0.397	1.384	4.171	0.146
Q(10)	7.754	8.872	2.201	6.756	5.578	3.581	2.856
п-вредност (Q)	0.355	0.449	0.698	0.455	0.472	0.733	0.970

[75] Модел који садржи све доцње до реда p односно q назива се пуни модел (енгл. *full model*). Модел који изоставља један број доцњи због њихове статистичке незначајности назива се редуковани (енгл. *reduced model*).

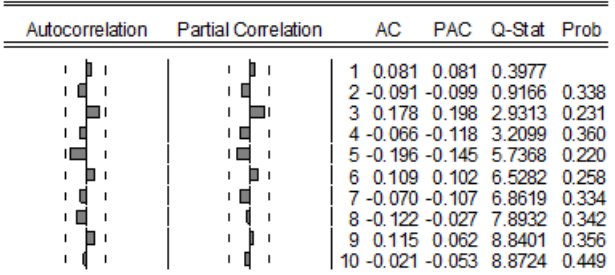
Слика 19: Корелограми резидуала оцењених модела из Табеле 27 и корелограм приноса крипто-валуте *SOL*

ADA (резидуали)



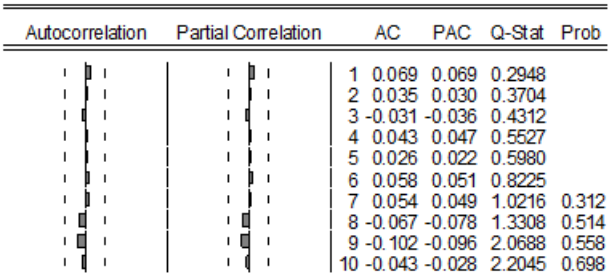
	AC	PAC	Q-Stat	Prob
1	-0.014	-0.014	0.0126	
2	0.050	0.050	0.1691	
3	0.042	0.044	0.2847	
4	0.023	0.022	0.3203	0.571
5	-0.193	-0.198	2.8121	0.245
6	0.061	0.054	3.0646	0.382
7	-0.043	-0.024	3.1914	0.526
8	-0.130	-0.124	4.3804	0.496
9	-0.052	-0.050	4.5741	0.599
10	-0.208	-0.248	7.7546	0.355

AVAX (резидуали)



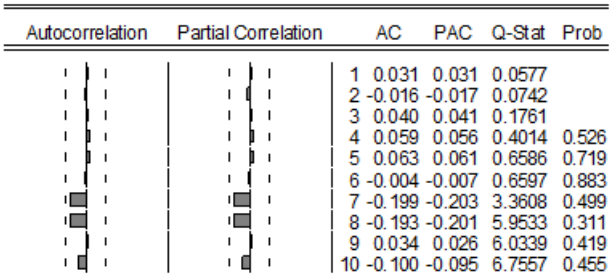
	AC	PAC	Q-Stat	Prob
1	0.081	0.081	0.3977	
2	-0.091	-0.099	0.9166	0.338
3	0.178	0.198	2.9313	0.231
4	-0.066	-0.118	3.2099	0.360
5	-0.196	-0.145	5.7368	0.220
6	0.109	0.102	6.5282	0.258
7	-0.070	-0.107	6.8619	0.334
8	-0.122	-0.027	7.8932	0.342
9	0.115	0.062	8.8401	0.356
10	-0.021	-0.053	8.8724	0.449

BTC (резидуали)



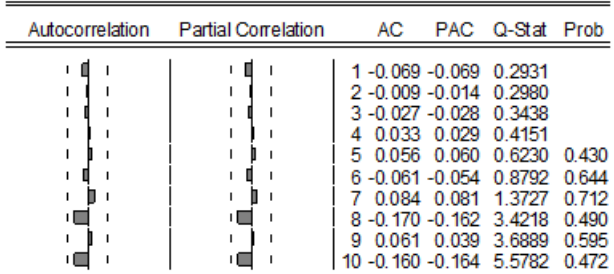
	AC	PAC	Q-Stat	Prob
1	0.069	0.069	0.2948	
2	0.035	0.030	0.3704	
3	-0.031	-0.036	0.4312	
4	0.043	0.047	0.5527	
5	0.026	0.022	0.5980	
6	0.058	0.051	0.8225	
7	0.054	0.049	1.0216	0.312
8	-0.067	-0.078	1.3308	0.514
9	-0.102	-0.096	2.0688	0.558
10	-0.043	-0.028	2.2045	0.698

DOGE (резидуали)



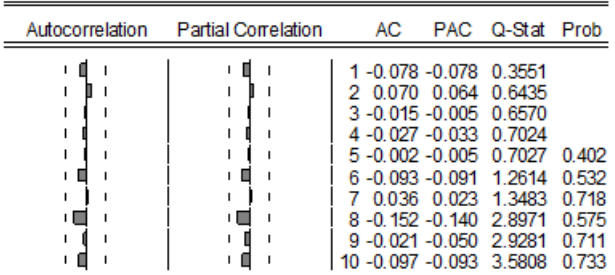
	AC	PAC	Q-Stat	Prob
1	0.031	0.031	0.0577	
2	-0.016	-0.017	0.0742	
3	0.040	0.041	0.1761	
4	0.059	0.056	0.4014	0.526
5	0.063	0.061	0.6586	0.719
6	-0.004	-0.007	0.6597	0.883
7	-0.199	-0.203	3.3608	0.499
8	-0.193	-0.201	5.9533	0.311
9	0.034	0.026	6.0339	0.419
10	-0.100	-0.095	6.7557	0.455

DOT (резидуали)



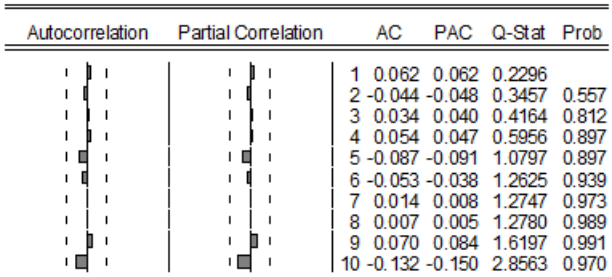
	AC	PAC	Q-Stat	Prob
1	-0.069	-0.069	0.2931	
2	-0.009	-0.014	0.2980	
3	-0.027	-0.028	0.3438	
4	0.033	0.029	0.4151	
5	0.056	0.060	0.6230	0.430
6	-0.061	-0.054	0.8792	0.644
7	0.084	0.081	1.3727	0.712
8	-0.170	-0.162	3.4218	0.490
9	0.061	0.039	3.6889	0.595
10	-0.160	-0.164	5.5782	0.472

ETH (резидуали)



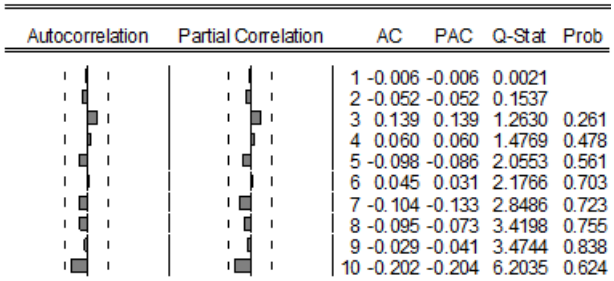
	AC	PAC	Q-Stat	Prob
1	-0.078	-0.078	0.3551	
2	0.070	0.064	0.6435	
3	-0.015	-0.005	0.6570	
4	-0.027	-0.033	0.7024	
5	-0.002	-0.005	0.7027	0.402
6	-0.093	-0.091	1.2614	0.532
7	0.036	0.023	1.3483	0.718
8	-0.152	-0.140	2.8971	0.575
9	-0.021	-0.050	2.9281	0.711
10	-0.097	-0.093	3.5808	0.733

LUNA (резидуали)



	AC	PAC	Q-Stat	Prob
1	0.062	0.062	0.2296	
2	-0.044	-0.048	0.3457	0.557
3	0.034	0.040	0.4164	0.812
4	0.054	0.047	0.5956	0.897
5	-0.087	-0.091	1.0797	0.897
6	-0.053	-0.038	1.2625	0.939
7	0.014	0.008	1.2747	0.973
8	0.007	0.005	1.2780	0.989
9	0.070	0.084	1.6197	0.991
10	-0.132	-0.150	2.8563	0.970

SOL (приноси)



	AC	PAC	Q-Stat	Prob
1	-0.006	-0.006	0.0021	
2	-0.052	-0.052	0.1537	
3	0.139	0.139	1.2630	0.261
4	0.060	0.060	1.4769	0.478
5	-0.098	-0.086	2.0553	0.561
6	0.045	0.031	2.1766	0.703
7	-0.104	-0.133	2.8486	0.723
8	-0.095	-0.073	3.4198	0.755
9	-0.029	-0.041	3.4744	0.838
10	-0.202	-0.204	6.2035	0.624

Извор: Приказ аутора

Помоћни модели препознали су ауторегресивну компоненту првог реда као значајан предиктор. Ипак, у њима је она фигурисала у трансформисаном облику како би се прилагодила фреквенцијама објављивања чланака (видети израз (9)). С тим у вези, за сам крај проверимо и да ли би ауторегресивне компоненте првог реда била значајна и на нетрансформисаним подацима. Резултати су дати табелом 28. Табела показују да ar(1) модел није добра спецификација за описивање индивидуалног кретања приноса. То је било и очекивано будући

да подаци упућују на другачију спецификацију (видети табелу 27). Иако је ar(1) модел популаран и једноставан начин за проверу слабе форме тржишне ефикасности, у овом случају он није довољан да се уоче сигнали у кретању за које из претходног излагања евидентно видимо да постоје. Модел занемарује могућност постојања аутокорелисаности вишег реда, док у обзир узима само ниво парцијалне (али не и обичне) аутокорелације. Из тог разлога исправније је ослонити се на резултате добијене из сложеније ARMA спецификације (дате таблом 27). За сличан приступ су се определили и раније истакнути Паламалаи и сарадници (2021) и Банди и Вилди (2019). Кад посматрамо резултате оцене ауторегресивног модела првог реда, занимљиво је истаћи и да је ауторегресивни коефицијент првог реда код крипто-валуте DOT близу границе значајности, као и да је он највећи међу оцењеним коефицијентима посматраним валутама. Будући да је код ове крипто-валуте ниво парцијалне аутокорелације првог реда око 21%, могуће је да би се у даљем истраживању и испоставио као статистички значајан.

Табела 28: Оцењени ауторегресивни коефицијенти првог реда осам одабраних крипто-валута

	ADA	**AVAX**	**BTC**	**DOGE**	**DOT**	**ETH**	**LUNA**	**SOL**
ar(1)	0.099	0.073	0.090	0.128	0.208	0.140	0.049	0.075
п-вредност	0.4709	0.5908	0.5540	0.3499	0.1171	0.3151	0.7329	0.5942

Сагледавши све претходно изнете доказе, можемо закључити да они у највећој мери подржавају закључке до којих је дошла постојећа литература. За крај, сумирајмо најважније од њих. Цене четири од осам посматраних крипто-валута не садрже јединични корен, те се као стационарна серија не могу сматрати случајним ходом. Кретање приноса код три од четири крипто-валуте код чијих је цена пронађен јединични корен можемо описати ARMA спецификацијом, те и у њиховом случају може говорити о нарушености слабе форме тржишне ефикасности. Ове налазе додатно потврђују тестови случајности који су код две крипто-валуте детектовале неслучајно кретање приноса. Једина крипто-валута код које слаба форма хипотезе о тржишној ефикасности не може бити одбачена на бази прикупљених доказа у анализираном периоду је SOL.

6. Закључак

Истраживање спроведено за потребе ове дисертације покушало је да да одговор на више питања. Пошавши од модела Џагадиша и Вуа (2019), истражене су могућности његовог унапређења. Модел поменутих аутора садржи грешку мерења сентимента, чијим би се отклањањем добиле прецизније процене сентимената инвеститора. Крајњи циљ добијања прецизнијих процена сентимента инвеститора је да се провери могу ли нам оне помоћи да прецизније предвидимо кретање приноса. Вођена истакнутим циљевима и инспирисана процедуром коју је предложио Јохансен (1996), дисертација је предложила алтернативни приступ оцењивању пондера сентимента. Процедура је очистила један део грешке мерења сентимента. Остатак грешке се може апроксимирати уз помоћ кумулативних приноса или прецизно оценити увођењем претпоставки, односно, нових информација. Претпоставке се у том случају задају као ограничења у процесу оптимизације. Дисертација је понудила једно такво решење претпоставивши да се сентимент речи креће у традиционалним границама, тј. у интервалу [-1,1].

Како би се испитала употребна моћ нових оцена, спроведена је троетапна процедура. Предложена процедура је пажљиво дизајнирана тако да у обзир узме све аспекте истраживања. Целокупан узорак је подељен у три подузорка, по један за сваку етапу. Први подузорак је намењен оцени пондера и највећи је од три подузорка. Његова величина обезбеђује да алгоритам има довољно информација за учење приликом оцене пондера сентимента. Након ове етапе, истраживач може да искористи оцењене пондере за мерење сентимента било ког текста. Захваљујући томе, из вести се могу изрударити потенцијални предиктори приноса. Други подузорак намењен је идентификовању релевантних предиктора приноса које треба користити за предвиђање приноса из трећег подузорка. На тај начин предвиђања се врше над подацима које модел није упознао приликом анализе предиктора. Други и трећи подузорак одабрани су тако да њихови подаци нису превише удаљени једни од других[76]. Такав дизајн обезбедио је да утицаји уочени у другом подузорку егзистира и у трећем подузорку. На тај начин је оправдана употреба идентификованих предиктора из другог подузорка за предвиђање у трећем подузорку.

Резултати добијени на бази другог подузорка омогућили су проверу две од три постављене хипотезе. Прва хипотеза претпоставља значајан утицај сентимента вести о самој крипто-валути на њене приносе. Друга хипотеза састојала се од два дела. Први део претпоставља значајан утицај читљивости објављених вести на приносе, док други део претпоставља да на приносе утиче и унакрсни сентимент. Резултати су разматрани посебно и за оцене са апроксимацијом остатка грешке (у даљем тексту апроксимативне оцене) и за оцене добијене уз претпоставку о традиционалним границама пондера (у даљем тексту егзактне нормиране оцене). У случају егзактних оцена прва хипотеза није одбачена у половичном броју случајева (4/8). Први део друге хипотезе је у највећој мери одбачен (2/8). То није био случај са другим делом ове хипотезе који није одбачен у већини случајева (7/8). Примена апроксимативних оцена пружила је сличне резултате. Промене су примећене у три од осам анализираних случајева. Овога пута прва хипотеза није одбачена у већини случајева (6/8). Утицај сентимента

[76] Предугачак други подузорак обухватио би факторе који нису утицајни у трећем подузорку, док би предугачак трећи узорак обухватио промену фактора, те би захтевао вишеструку анализу релевантних предиктора.

о сопственим вестима сада је препознат и код ETH-а и BTC-а. Промене су примећене и у знаку веза. У два случаја добијен је нелогичан предзнак, што се може објаснити као аномалија периода за који је анализа спроведена. Када је реч о оба дела друге хипотезе закључци су остали непромењени. Када код оба типа оцена здружено погледамо резултате о обе анализиране врсте сентимента (сопствени сентимент и унакрсни сентимент), може се рећи да је сентимент вести користан индикатор за анализу крипто-валута. То нарочито важи за унакрсни сентимент, којем постојећа литература поклања недовољно пажње. Поред тога, резултати сведоче и да је анализа јако осетљива на избор мере сентимента (тј. методе оцене пондера). Приметно је да се закључци у појединим случајевима могу значајно разликовати, уколико се на њих примене различите мере сентимента. То не треба да чуди, будући да је већина анализа у финансијама (попут портфолио оптимизације) веома осетљива на избор улазних података. Према томе, избор адекватне мере сентимента је веома важно питање у финансијском рударењу текста, на које је ова дисертација понудила своје одговоре.

Приказани резултати су сигнали потенцијалне тржишне неефикасности тржишта крипто-валута. Резултати показују да је у анализираном периоду постојала веза између информација добијених рударењем текста и приноса крипто-валута. Поред тога ауторегресивна компонента је била значајна у већини случајева (5/8). Сви резултати упућују на закључак да, ипак, нису све јавно доступне информације укључене у цену крипто-валута. Ослањајући се на доказе из постојеће литературе, према којима ефикасност тржишта варира кроз време, дисертација је спровела формалне тестове слабе форме тржишне ефикасности. Тестови су показали да су крипто-валуте *AVAX, BTC, DOGE* и *ETH* изразито неефикасне, будући да њихове цене нису случајан ход (не поседују јединични корен) и да се могу описати као *ARMA*(p,q) процес. Тестови су показали да цене крипто-валута *ADA, DOT* и *LUNA* поседују јединични корен, али да се њихови приноси не понашају као бели шум, већ се могу описати као *ARMA*(p,q) процес. То ове крипто-валуте чини умерено неефикасним. Коначно у случају крипто-валуте SOL ни један тест хипотезе о слабој форми тржишне ефикасности није успео да је обори. Међу посматраним крипто-валутама најмање ефикасним се показао Биткоин који је пао на сваком тесту неефикасности. Резултатима посебну драж даје чињеница да су добијени у време зачетка светске економске кризе која се трансмитовала и на тржиште крипто-валута. Из тог разлога несумњиво је да се један део неефикасности дугује негативним нетржишним утицајима. У овом раду изостала је формална провера хипотезе о полу-јакој форми тржишне ефикасности, што ће се испитати засебним истраживањем.

Последња од постављених хипотеза претпоставила је да ће нове мере сентимента генерисати побољшане предикторе који ће дати прецизније прогнозе кретања приноса. Како би се валидност хипотезе проверила конструисан је ансамбл алгоритам који обрачунава предикторе по оригиналној и иновираној методологији. Алгоритам би затим правио предикције приноса за наредни дан на бази свих доступних вести из текућег дана. Добијене грешке су најпре биле упоређене тачкастим оценама корена из средње квадратне грешке прогнозе. И апроксимативне и егзактне нормиране оцене су понудиле мање грешке прогнозе од оцена Џагадиша и Вуа (2019). Зарад генерализације закључака спроведена су два статистичка теста. Први је омнибус Диеболд-Маријанов тест предложен у овој дисертацији, док је други МекКракенов (2000) тест. Први тест је у оба случаја сугерисао да иновиране оцене сентимента дају боље прогнозе од оригиналних. Насупрот томе, МекКракенов (2000) тест препознао је статистички значајну разлику само код апроксимативних оцена. Иако статистичка значајност није потврђена код

егзактних нормираних оцена, неспорно је да минорно побољшање, ипак, постоји будући да су тачкасте мере квалитета прогнозе мање у свим анализираним случајевима. Према резултатима статистичких тестова апроксимативне оцене су се показале бољим од егзактних нормираних оцена (будући да су оба статистичка теста препознала њихову значајност).

Допринос резултата дисертације је вишеструк. Испитан је утицај показатеља добијених из рударења текста на приносе крипто-валута током специфичног кризног периода. Ту се посебно издваја испитивање утицаја унакрсног сентимента који је још увек недовољно истражен у постојећој литератури, а показао се као значајан предиктор приноса. Коначно, дисертација је испитала и хипотезу о слабој форми тржишне ефикасности осам одабраних крипто-валута током посматраног кризног периода. Поред наведених емпиријских резултата, дисертација је понудила и иновиране оцене сентимента речи за које се испоставило да помажу у предвиђању будућег кретања приноса. Уједно, у дисертацији је представљена и модификована верзија Диеболд-Маријановог теста за потребе вишеузорачког тестирања, тзв. Омнибус Диеболд-Маријанов тест. У истраживању је примењен низ програмских процедура и алгоритама, и развијени су помоћни алати за подршку рударењу текста који су специфично прилагођени анализи крипто-валута.

7. Литература

Aalborg, H. A., Molnár, P. & de Vries, J. (2019). What can explain the price, volatility and trading volume of Bitcoin? *Finance Research Letters, Elsevier*. **29**(C): 255-265.

Adedoku, A. (2019). Bitcoin-Altcoin price synchronization hypothesis: Evidence from recent data. *Journal of Finance and Economics*. **7**(4) 137-147.

Adhami, S., Giudici, G. & Martinazzi, S. (2018). Why Do Businesses Go Crypto? An Empirical Analysis of Initial Coin Offerings. *Journal of Economics and Business*. **100**(C): 64-75.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing and Management.* **39**(1): 45–65.

Amdouni, A. (2021). Investor Sentiment Measurement Techniques: A Review of Literature. *Journal of Business and Management Review*. **11**(2).

Anamika, A. (2022). Do news headlines matter in the cryptocurrency market? *Applied Economics*. **54**(54): 6322-6338.

Antweiler, W. & Frank, M. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*. **59**(3): 1259-1294.

Audrino, F. & Tetereva, A. (2019). Sentiment spillover effects for US and European companies. *Journal of Banking & Finance*. **106**(1): 542-567.

Baker, M. & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *Journal of Finance*. **61**(4): 1645-1680.

Baker, M. & Wurgler, J. (2007). Investor Sentiment in the Stock Market. *Journal of Economic Perspectives*, **21**(2): 129-152.

Bandopadhyaya, A. & Jones, A. (2006). Measuring investor sentiment in equity markets. *Journal of Asset Management.* **7**, 208–215.

Bandopadhyaya, A., & Jones, A. L. (2008). Measures of Investor Sentiment: A Comparative Analysis Put-Call Ratio Vs. Volatility Index. *Journal of Business & Economics Research (JBER)*. **6**(8).

Barberis, N., Shleifer, A. & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*. **49**(3): 307–343.

Bariviera, A. F. (2017). The inefficiency of Bitcoin revisited: A dynamic approach. *Economics Letters*. **161**: 1-4.

Banea, C., Mihalcea, R. & Wiebe, J. (2018). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC'08). Marrakech, Morocco. European Language Resources Association (ELRA).

Bartels, R. (1982). The Rank Version of von Neumann's Ratio Test for Randomness. *Journal of the American Statistical Association*. **77**(377): 40–46.

Bartlett, J. (2019). The £4bn OneCoin scam: how crypto-queen Dr Ruja Ignatova duped ordinary people out of billions — then went missing. *The Times*. (https://www.thetimes.co.uk/

article/the-4bn-onecoin-scam-how-crypto-queen-dr-ruja-ignatova-duped-ordinary-people-out-of-billions-then-went-missing-trqpr52pq - посећен 27.11.2022.)

Bekaert, G., Hoerova, M. and Lo Duca, M. (2013). Risk, Uncertainty and Monetary Policy. *Journal of Monetary Economics*. **60**(7): 771–88.

Bernardi, M., Catania, L. & Petrella, L. (2017). Are News Important to Predict the Value-at-Risk? *European Journal of Finance*. **23**(6): 535–572.

Black, F. & Litterman, R. (1991). Asset Allocation Combining Investor Views with Market Equilibrium. *Journal of Fixed Income*. **1**(2): 7-18.

Bodie, Z., Kane, A. & Marcus, A. (2007). *Essentials of investments*. Sixth edition. New York: The McGraw-Hill Company INC

Bollen, J., Mao, H. & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*. **2**(1): 1-8.

Bonato, M., Gkillas, K., Gupta, R. & Pierdzioch, C. (2020). Investor Happiness and Predictability of the Realized Volatility of Oil Price. *Sustainability*. **12**(10): 1-11.

Breitinger, C., Gipp, B. & Langer, S. (2015). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*. **17**(4): 305–338.

Brown, G. & Cliff, M. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance*. **11**(1): 1-27.

Bundi, N. & Wildi, M. (2019). Bitcoin and market-(in)efficiency: a systematic time series approach, *Digital Finance – Springer*. **1**(1): 47-65.

Burghardt, M. (2011). Investor Sentiment Construction. *Retail Investor Sentiment and Behavior*. *Gabler*. 35–68.

Butler, K. C. & Malaikah, S. J. (1992). Efficiency and inefficiency in thinly traded stock markets: Kuwait and Saudi Arabia. *Journal of Banking & Finance*. **16**(1): 197-210.

Cankaya, S., Alp, E.A. & Findikçi, M. (2019). News Sentiment and Cryptocurrency Volatility. *Contributions to Economics: Blockchain Economics and Financial Market Innovation. Springer*. **0**: 115-140.

Caporale, G., Spagnolo, F., & Spagnolo, N. (2018). Macro News and Bond Yield Spreads in the Euro Area. *European Journal of Finance*. **24**(2): 114–134.

Cohen, L., Malloy, C. & Nguyen, Q. (2020). Lazy Prices. *The Journal of Finance*. **75**(3): 1371-1415.

Cheah, E. T., Mishra, T., Parhi, M. & Zhang, Z. (2018). Long memory interdependency and inefficiency in Bitcoin markets. *Economics Letters.* **167**: 18–25.

Chen, H., Shan, L. and Wang, C. (2012). Investment Sentiment in Finance Market. *Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)*. Chongqing, China. Atlantis Press: 3325-3332.

Christoffersen, P.F. (2012). *Elements of Financial Risk Management*. (Second Edition). Academic Press, London, UK.

Ciaian, P., Rajcaniova, M. & Kancs, D. (2016). The economics of Bitcoin price formation. *Applied Economics*. **48**(19): 1799-1815.

Ciupa, K. (2019). *Cryptocurrencies: opportunities, risks and challenges for anti-corruption compliance systems*. OECD Global Anti-Corruption and Integrity Forum. France, Paris

Corbet, S., Lucey, B. & Yarovaya, L. (2018). Datestamping the Bitcoin and Ethereum Bubbles. *Finance Research Letters* **26**(1): 81–88.

Corbet, S., Larkin, C., Lucey, B., Meegan, A. & Yarovaya, L. (2020). The impact of macroeconomic news on Bitcoin returns. *The European Journal of Finance*. **26**(14): 1396-1416

Damjanović, A. & Drenovak, M. (2023). Da li su sve tekstualne vesti samo šum za investitore? – Uticaj onlajn tekstova na prinose Bitkoina. *Ekonomske teme*. **61**(2): 121-144

Day, M.Y. & Lee, C.C. (2016). Deep learning for financial sentiment analysis on finance news providers. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. San Francisco, CA, USA. 1127-1134

De Long, J. B., Shleifer, A., Summers, L. H. & Waldmann, R. J. (1990). Positive feedback investment strategies and destabilizing rational speculation. *The Journal of Finance*. **45**(2): 379–395.

Demeterfi, K., Derman, E., Kamal, M. and J. Zou. (1999). More Then You Ever Wanted To Know About Volatility Swaps. The Journal of Derivatives. 6(4): 9-32

Demetrescu, M., Hanck, C. & Kruse, R. (2015). *Fixed-b Asymptotics for t-Statistics in the Presence of Time-Varying Volatility*. Vfs Annual Conference: Economic Development – Theory and Policy, Munster, Germany.

Diebold, F. & Mariano, S. (1991). Comparing Predictive Accuracy: An Asymptotic Test. *Journal of Business & Economics Statistics*. **13**(3): 253-623

Dickey, D. A. & Fuller, W. A. (1979): Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association.* **74**(366): 427-431.

Edwards, T. and Preston, H. (2017). *Reading VIX®: Does VIX Predict Future Volatility*?. S&P Dow Jones Indices. (https://www.spglobal.com/spdji/en/documents/research/research-reading-vix-does-vix-predict-future-volatility.pdf - посећен 20.12.2023.)

Engelberg, J. & Parsons, C. (2011). The Causal Impact of Media in Financial Markets. *The Journal of Finance*. **66**(1): 67–97

Entrop, O., Frijns, B. & Seruset, M. (2020). The determinants of price discovery on bitcoin markets. *The Journal of Futures Market*. **40**(5): 816-837

Fama, E. F. (1963). Mandelbrot and the Stable Paretian Hypothesis. *The Journal of Business*. **36**(4): 420–429.

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business*. **38**(1): 34–105.

Frazier, K., Ingram, R. & MackTennyson, B. (1984). A Methodology for the Analysis of Narrative Accounting Disclosures. *Journal of Accounting Research*. **22**(1): 318-331

Fry, J. & Cheah, J. (2016). Negative bubbles and shocks in cryptocurrency markets. *International Review of Financial Analysis*. **47**(C): 343-352

Gallagher, L. A. & Taylor, M. P. (2002). Permanent and Temporary Components of Stock Prices: Evidence from Assessing Macroeconomic Shocks. *Southern Economic Journal*. **69**(2): 345–362.

Galton, F. (1907). Vox Populi. *Nature*. **75**: 450–451.

Gidofalvi, G. & Elkan, C. (2003). *Using news articles to predict stock price movements*. Technical Report. Department of Computer Science and Engineering - University of California. San Diego.

Gomber, P., Koch, J.A. & Siering, M. (2017). Digital Finance and FinTech: current research and future research directions. *Journal of Business Economics.* **87**(5). 537–580.

Griffin, J. and Shams, A. (2018). Manipulation in the VIX?. *The Review of Financial Studies*. **31**(4): 1377-1417.

Gunning, R. (1952). The Technique of Clear Writing. *Journal of Applied Mathematics and Physics, McGraw-Hill*. **5**(6): 36–37.

Heinig, S. & Nanda, A. (2018) Measuring sentiment in real estate – a comparison study. *Journal of Property Investment & Finance*. **36**(3): 248-258.

Henry, E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written? *The Journal of Business Communication (1973)*, **45**(4), 363–407.

Hestla-Barnhart, A. (2015). Fixing the VIX: An Indicator to Beat Fear. *Journal of Technical Analysis*. **57**: 30-38.

Ho, K.Y., Shi Y. & Zhang Z. (2020). News and return volatility of Chinese bank stocks. *International Review of Economics & Finance*. **69**(C): 1095-1105.

Jarque, C. M. & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*. 6 (3): 255–259.

Jegadeesh, N. & Wu, D. (2019). Word power: A new approach for content analysis. *Journal of Financial Economics*. **110**(3): 712-729.

Johansen, S. (1996). *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press, Oxford, UK.

LeBaron, B., Arthur, W.B. & Palmer, R. (1999). Time series properties of an artificial stock market. *Journal of Economic Dynamics and Control*. **23**(9–10): 1487–1516.

Lamon, C., Nielsen, E. & Redondo, E. (2017). Cryptocurrency Price Prediction Using News and Social Media Sentiment. *SMU Data Science Review*. **1**(3): 1-22.

Li, F. (2006). Annual Report readability, current earnings and earnings persistence. Working paper. University of Michigan, Ann Arbor.

Ljung, G. M. & Box, G. E. P. (1978). On a Measure of a Lack of Fit in Time Series Models. *Biometrika*. **65**(2): 297–303.

Lo, A. & MacKinlay, A. C. (1988). Stock Market Prices Do Not Follow Random Walks: Evidence From a Simple Specification Test. *The Review of Financial Studies*. **1**(1): 41-66.

Lo, A. (2004). The adaptive markets hypothesis *The Journal of Portfolio Management (30th Anniversary eddition)*. **30**(5): 15-29.

Lo, A. (2012). Adaptive markets and the new world order. *Financial Analysts Journal*. **68**(2): 18-29.

Loughran, T. & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance.* **66**(1), 35–65.

Kaplanski, G. & Levy, H. (2010). Sentiment and stock prices: The case of aviation disasters. *Journal of Financial Economics*, **95**(2):174-201.

Karalevicius, V., Degrandea, N. & De Weerdt, J. (2017). Using sentiment analysis to predict interday Bitcoin price Movements. *The Journal of Risk Finance*. **19**: 56-75.

Kavussanos, M. & Dockery, E. (2001). A multivariate test for stock market efficiency: the case of ASE. *Applied Financial Economics*. **11**(5): 573-579.

Kim, M. J. & Park, S. Y. (2023) Testing for market efficiency in cryptocurrencies: evidence from a non-linear conditional quantile framework. *Applied Economics Letters*. **30**(16): 2245-2251.

Kraaijeveld, O. & De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*. **65**(101188).

Krishnamurthy, S. (2021). Deriving market signals from the term structure of VIX. *Journal of Technical Analysis*. **69**: 32-65.

Kristoufek, L. (2013). Bitcoin meets google trends and Wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific Reports.* **3**(3415)

Kristoufek, L. (2018) On Bitcoin markets (in)efficiency and its evolution. *Physica A: Statistical Mechanics and its Applications.* **503**(1): 257–262.

Kumar, A. & Ajaz, T. (2019). Co-movement in crypto-currency markets: evidences from wavelt analysis. *Finance Innovations*. **5**(33).

Kumar, A. S., Nagaraju, T. & Ajaz, T. (2020). On the informational efficiency of the cryptocurrency market. *IUP Journal of Applied Economics*. **19**(1): 47–56.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *Journal of Econometrics*. **54**(1-3), 159–178.

Mai, F., Bai, Q., Shan, Z., Wang, X. S. & Chiang, R. H. (2015). From Bitcoin to Big Coin: The Impacts of Social Media on Bitcoin Performance. *SSRN Electronic Journal*. 1-16.

Mao, H., Counts, S. and Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. *Papers 1112.1051, arXiv.org*.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*. **7**(1): 77-91.

McCracken, M.W. (2000). Robust out-of-sample inference. Journal of Econometrics. **99**(2000): 195-223.

Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D. & Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press, Cambridge, Massachusetts, United States.

Mladenović, Z. i Nojković, A. (2018). *Primenjena analiza vremenskih serija*. Centar za izdavačku delatnost, Ekonomski fakultet Univerziteta u Beogradu, Beograd, Srbija.

Mladenović, Z. i Petrović, P. (2018). *Uvod u ekonometriju*. Centar za izdavačku delatnost, Ekonomski fakultet Univerziteta u Beogradu, Beograd, Srbija.

Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*. White Paper.

Nathan, A., Galbraith, G. & Grimberg, J. (2021). *Top of Mind: Crypto – A New Asset Class* (Issue 98). Goldman Sachs. URL: https://www.goldmansachs.com/insights/pages/crypto-a-new-asset-class-f/report.pdf

Newey, W. & West, K. (1987a). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*. **55**(3): 703–708

Newey, W. & West, K. (1987b). Hypothesis testing with efficient method of moments estimation. *International economic review*. **28**(3): 777-787.

Obaid, K., and Pukthuanthong, K. (2022). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. *Journal of Financial Economics*. **144**: 273–97.

Ozadyin, O. (2021). *Bitcoin's Lagged Effect on Altcoins: A short-term research*. Press Academia Procedia, Istanbul, Turkey: 10-13.

Palamalai, S., Kumar, K. & Maity, B. (2021). Testing the random walk hypothesis for leading cryptocurrencies. *Borsa Istanbul Review*. **21**(3). 256-268.

Peterson, R. L. (2016). *Trading on sentiment: The power of minds over markets*. Hoboken: Wiley

Petrović Lj. (2015). *Teorijska statistika – Teorija statističkog zaključivanja*. (3. Izdanje). Centar za izdavačku delatnost, Ekonomski fakultet Univerziteta u Beogradu, Beograd, Srbija.

Phillips, R. C. & Gorse, D. (2017). Predicting cryptocurrency price bubbles using social media data & epidemic modelling. *IEEE Symposium Series on Computational Intelligence* (SSCI). 1-7.

Polasik, M., Piotrowska, A. I., Wisniewski, T. P., Kotkowski, R., Lightfoot, G. (2015). Price fluctuations and the use of bitcoin: An empirical inquiry. *International Journal of Electronic Commerce.* **20**(1): 9–49.

Prasad, S., Mohapatra, S., Ramizur-Rahman, M. and Puniyani, A. (2023). Investor Sentiment Index: A Systematic Review. *International Journal of Financial Studies*. **11**(6).

Qasem, M., Thulasiram, R. & Thulasiram, P. (2015). Twitter sentiment classification using machine learning techniques for stock markets. *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Kochi, India. 834-840.

Qian, B. & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*. **26**, 25–33.

Renault, T. (2020). Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digit Finance.* **2**: 1–13.

Rognone, L., Hyde S. & Zhang S. S. (2020). News sentiment in the cryptocurrency market: An empirical comparison with Forex. *International Review of Financial Analysis*. **69**(101462).

Сајт: *blockchain.com*. Посећен 15.11.2022.

Сајт: *coinmarketcap.com*. Посећен 15.11.2022.

Сајт: *cryptonews.net*. Посећен 23.03.2022.

Сајт: *merriam-webster.com*. Посећен 22.08.2022.

Samuelson, P. (1965). Proof That Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review Spring*. **6**: 41-49.

Sapkota, N. (2022). News-based sentiment and bitcoin volatility. *International Review of Financial Analysis*. **82**(2022).

Sapkota, N. & Grobys, K. (2023). Fear sells: On the sentiment deceptions and fundraising success of initial coin offerings. *Journal of International Financial Markets, Institutions and Money*. **83**(101716).

Sarkodie, S.A., Ahmed, M.Y. & Owusu, P.A. (2022). COVID-19 pandemic improves market signals of cryptocurrencies – evidence from Bitcoin, Bitcoin Cash, Ethereum, and Litecoin. *Finance Research Letters*. **44**(102049).

Şaşmaz, E. & Tek, F. (2021). Tweet Sentiment Analysis for Cryptocurrencies. *6th International Conference on Computer Science and Engineering (UBMK)*. 613-618.

Schumaker, R., Zhang, Y., Huang, C. N. & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*. **53**(2012): 458-464.

Sensoy, A. (2019). The inefficiency of Bitcoin revisited: A high-frequency analysis with alternative currencies. *Finance Research Letters*. **28**: 68-73.

Sherman, A., Javani, F., Zhang, H. & Golaszewski, E. (2019). On the Origins and Variations of Blockchain Technologies. *IEEE Security Privacy*. **17** (1): 72–77.

Shmueli, G., Bruce, P., Gedeck, P. & Patel, N. (2020). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python*. Wiley, Hoboken, USA.

Sovbetov, Y. (2018). Factors Influencing Cryptocurrency Prices: Evidence from Bitcoin, Ethereum, Dash, Litecoin, and Monero. *Journal of Economics and Financial Analysis*. **2**(2): 1-27.

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*. **28**: 11–21.

Stock, J. H. & Watson, M. W. (1989): Interpreting the Evidence on Money Income Causality, *Journal of Econometrics.* **40**(1): 161-181.

Swiss Financial Market Supervisory Authority [FINMA] (2018). *Guidelines for enquiries regarding the regulatory framework for initial coin offerings*. (https://www.iosco.org/library/ico-statements/Switzerland%20-%20FINMA%20-%20ICO%20Guidelines.pdf – посећен 28.11.2022).

Tetlock P. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*. **62**(3): 1139-1168.

Tsay, R. (2006). Analysis of Financial Time Series. (Second Edition). Wiley, Hoboken, U.S.

Urquhart, A. (2016). The inefficiency of Bitcoin. *Economics Letters*. **148**(2016): 80-82.

Vo, A. D., Nguyen Q. P. & Ock C. Y. (2019). Sentiment Analysis of News for Effective Cryptocurrency Price Prediction. *International Journal of Knowledge Engineering*. **5**(2).

von Neumann, J. (1941) Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *Annals of Mathematical Statistics*. **12**(4): 367-395.

Weiss, M., Indurkhya, N. & Zhang, T. (2015). *Fundamentals of Predictive Text Mining*. (2nd ed.). Texts in Computer Science, Springer, London, UK.

Xia, Q. (2022). Advances in Research on Investor Sentiment and Stock Returns. *Proceedings of the 2022 2nd International Conference on Economic Development and Business Culture (ICEDBC 2022)*. Dali, China. 791-795.

Yang, S., Mo, S. & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*. **15**(10): 1637-1656.

Yi, E., Yang, B., Jeong, M., Sohn, S. & Ahn, K. (2023). Market efficiency of cryptocurrency: evidence from the Bitcoin market. *Scientific Reports.* **13**(4789).

8. Прилози

Прилог 1: Оцењени редуковани MA(7) модел приноса крипто-валуте *ADA*

Variable	Coefficient	Std.Error	Statistic	Prob.
c	-0.005	0.001	-5.658	0.000
MA(2)	-0.328	0.126	-2.598	0.012
MA(4)	-0.361	0.128	-2.811	0.007
MA(7)	-0.423	0.127	-3.324	0.002
Adjusted R-squared	0.208	Akaike info criterion		-3.278
S.E. of regression	0.045	Schwarz criterion		-3.137
Log likelihood	100.703	Hannan-Quinn criter.		-3.223

Прилог 2: Оцењени пуни AR(1) модел логаритмованих цена крипто-валуте *AVAX*

Variable	Coefficient	Std.Error	Statistic	Prob.
c	4.360	0.059	74.027	0.000
AR(1)	0.860	0.055	15.716	0.000
Adjusted R-squared	0.812	Akaike info criterion		-2.723
S.E. of regression	0.061	Schwarz criterion		-2.651
Log likelihood	80.954	Hannan-Quinn criter.		-2.695

Прилог 3: Оцењени редуковани *ARMA*(1,13) модел логаритмованих цена крипто-валуте *BTC*

Variable	Coefficient	Std.Error	Statistic	Prob.
c	10.599	0.027	389.284	0.000
AR(1)	0.865	0.054	16.000	0.000
MA(2)	-0.332	0.148	-2.241	0.029
MA(5)	-0.496	0.145	-3.430	0.001
MA(7)	-0.363	0.140	-2.589	0.013
MA(11)	-0.272	0.130	-2.092	0.042
MA(13)	-0.311	0.148	-2.107	0.040
Adjusted R-squared	0.816	Akaike info criterion		-3.988
S.E. of regression	0.031	Schwarz criterion		-3.739
Log likelihood	122.653	Hannan-Quinn criter.		-3.891

Прилог 4: Оцењени редуковани *ARMA*(1,5) модел логаритмованих цена крипто-валуте *DOGE*

Variable	Coefficient	Std.Error	Statistic	Prob.
c	-2.027	0.138	-14.633	0.000
AR(1)	0.968	0.038	25.459	0.000
MA(4)	-0.465	0.109	-4.278	0.000
MA(5)	-0.440	0.109	-4.028	0.000
Adjusted R-squared	0.863	Akaike info criterion		-3.725
S.E. of regression	0.036	Schwarz criterion		-3.583
Log likelihood	112.025	Hannan-Quinn criter.		-3.670

Прилог 5: Оцењени редуковани MA(7) модел приноса крипто-валуте *DOT*

Variable	Coefficient	Std.Error	Statistic	Prob.
MA(1)	0.204	0.099	2.060	0.044
MA(3)	0.177	0.099	1.777	0.081
MA(5)	-0.386	0.102	-3.773	0.000
MA(7)	-0.438	0.103	-4.244	0.000
Adjusted R-squared	0.123	Akaike info criterion		-3.222
S.E. of regression	0.047	Schwarz criterion		-3.081
Log likelihood	99.042	Hannan-Quinn criter.		-3.167

Прилог 6: Оцењени редуковани *ARMA*(4,5) модел логаритмованих цена крипто-валуте *ETH*

Variable	Coefficient	Std.Error	Statistic	Prob.
c	7.955	0.011	745.648	0.000
AR(1)	1.129	0.059	19.185	0.000
AR(4)	-0.224	0.049	-4.602	0.000
MA(2)	-0.716	0.086	-8.286	0.000
MA(5)	-0.246	0.087	-2.829	0.007
Adjusted R-squared	0.859	Akaike info criterion		-3.551
S.E. of regression	0.039	Schwarz criterion		-3.368
Log likelihood	102.642	Hannan-Quinn criter.		-3.480

Прилог 7: Оцењени редуковани AR(3) модел приноса крипто-валуте *LUNA*

Variable	Coefficient	Std.Error	Statistic	Prob.
AR(3)	0.329	0.142	2.325	0.024
Adjusted R-squared	0.089	Akaike info criterion		-2.382
S.E. of regression	0.073	Schwarz criterion		-2.346
Log likelihood	67.699	Hannan-Quinn criter.		-2.368

9. Биографија аутора

Александар Дамјановић рођен је у Приштини, 1996. године. Због рата који је задесио његову отаџбину, Александар је приморан да избегне на подручје југоисточне Србије где завршава прва четири разреда основне школе. Године 2007-е Александар се сели у Београд где завршава основну и средњу школу, као и Економски факултет Универзитета у Београду (основне и мастер академске студије). На основним студијама Александар је дипломирао на студијском програму за Статистику, информатику и квантитативне финансије (модул Статистика), док је у оквиру Међународног мастер-програма за квантитативне финансије (енгл. *International master in quantitative finance – IMQF*) мастерирао са темом „Трговање волатилношћу“. И током основних и мастер академских студија Александар се истакао као најбољи студент у генерацији у оквиру свог студијског програма. Од септембра 2019. године био је запослен као демонстратор на Економском факултету у Београду где је наставу држао у оквиру предмета: Операциона истраживања и Основи статистичке анализе. Од 2021. године ради као асистент на Рачунарском факултету где наставу држи у оквиру неколико предмета, док од 2022. наставу изводи и на Банкарској академији. На трећем степену академских студија, Александар је јануара 2022. као први у генерацији завршио са свим испитним обавезама са просеком 10,00. Током својих студија и истраживачког рада, Александар се посебно заинтересовао за програмирање и моделирање као и примену истих у сфери финансија. У време писања ове тезе Александар је објавио два академска научна рада.

10. Потписане изјаве аутора

образац изјаве о ауторству

Изјава о ауторству

Име и презиме аутора: _____

Број индекса: _____

Изјављујем

да је докторска дисертација под насловом

_____

_____

• резултат сопственог истраживачког рада;

• да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;

• да су резултати коректно наведени и

• да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, _____

_____

образац изјаве о истоветности штампане и електронске верзије докторског рада

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора _____

Број индекса _____

Студијски програм _____

Наслов рада _____

Ментор _____

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, _____

_____

образац изјаве о коришћењу

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

_____

_____

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.

Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, _____

_____