UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Een whitepaper analyse van frauduleuze Initial Coin Offerings

Sam Gouwy

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

Prof. dr. Benoit DEPAIRE

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2021
2022

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Een whitepaper analyse van frauduleuze Initial Coin Offerings

Sam Gouwy

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

Prof. dr. Benoit DEPAIRE

Voorwoord

Met deze scriptie, genaamd 'Een whitepaper analyse van frauduleuze Initial Coin Offerings', sluit ik mijn academische opleiding als handelingsingenieur in de beleidsinformatica af. Het betreft een studie over de whitepapers van frauduleuze 'Initial Coin Offerings', welke te situeren zijn binnen het domein van crypto.

Voor aanvang van dit project was mijn kennis over blockchain en crypto slechts beperkt. De actuele gebeurtenissen, in combinatie met de uitgesproken interesse van mijn leeftijdsgenoten, gaven de doorslag om mij verder te verdiepen in dit onderwerp. Bovendien gaf dat mij de opportuniteit om mijn reeds verworven kennis, betreffende economie en data-analyse, te benutten in functie van dit onderzoek.

Het verzamelen en analyseren van de data liep niet altijd even vlekkeloos. Gelukkig stonden er op deze momenten een aantal personen voor mij klaar wat ervoor zorgde dat ik terug met goede moed en een frisse blik aan de slag kon gaan. Ik zou dan ook graag van dit moment gebruik maken om deze mensen te bedanken.

Allereerst heeft prof. dr. Depaire mij gedurende het volledige project ondersteund. Zijn feedback was steeds waardevol, constructief en rechtvaardig. De begeleiding resulteerde in een perfecte combinatie van: sturing, suggesties en de nodige vrijheid.

Daarnaast zou ik ook graag drie collega-studenten in het bijzonder willen bedanken. Dmitri Beloshitskiy, Joris Ganne en Senne Vanbroekhoven hebben ervoor gezorgd dat de afgelopen vijf jaar een zeer vermakelijke tijd was. Het samenwerken aan projecten was altijd aangenaam en leverde bovendien zeer mooie resultaten op.

Ook wil ik mijn vriendin bedanken voor de hulp gedurende deze periode. Wanneer ik het even niet meer zag zitten stond ze steeds klaar met de nodige portie motivatie.

Tot slot maak ik graag van de gelegenheid gebruik om mijn ouders te bedanken. Zij hebben mij gedurende mijn schoolcarrière aangemoedigd, geholpen en op de feiten gewezen wanneer mijn uitstelgedrag de bovenhand nam. Zonder hun zou ik niet staan waar ik vandaag de dag sta.

Veel leesplezier.

Sam Gouwy
Zonhoven, juni 2022

Een whitepaper analyse van frauduleuze Initial Coin Offerings

Sam Gouwy

Bedrijfseconomische wetenschappen, UHasselt, Diepenbeek, 3590, België.

Abstract

De opkomst van cryptomunten gaat gepaard met een groei in populariteit van Initial Coin Offerings (ICO's). Dat is de manier voor crypto-startups om geld in te zamelen voor blockchainprojecten. Tevens is het voor investeerders een manier om vroeg in te stappen en een groter rendement te realiseren. Steeds meer van deze investeerders worden slachtoffer van frauduleuze ICO's. Eén van de instrumenten die gebruikt worden om investeerders te overtuigen van de toekomstplannen van het project is de whitepaper, een document met informatie over het project. Deze studie onderzoekt op welke manier de whitepapers van frauduleuze ICO's al dan niet verschillen van de whitepapers van legitieme ICO's. De studie is uitgevoerd op basis van een dataset van 112 whitepapers van ICO's waarvan de helft is geïdentificeerd als frauduleus. De whitepapers dateren van 2015 tot en met 2021. Met behulp van *Natural Language Processing* en manueel onderzoek worden er bepaalde kenmerken uit de tekst gehaald. Deze kenmerken vormen op hun beurt de basis voor een descriptief en correlationeel onderzoek. De bevindingen van deze studie beschrijven de impact van het toevoegen van usecases, een technische uitleg, de omvang en de toon van de paper. Zo blijkt uit het onderzoek dat het risico op oplichting lager is wanneer het gaat om een lange paper die een technische uitleg en usecases bevat.

Keywords: Initial Coin Offering, ICO, Scam, Whitepaper

1 Inleiding

De bloeiende cryptomarkt trekt ondanks zijn volatiliteit veel nieuwe investeerders aan. Veel privé en publieke investeerders zien de bijna absurde stijgingen van Bitcoin en Ethereum en willen de volgende trein niet missen. Ze beschouwen cryptomunten als langdurige belleging als bescherming tegen inflatie als alternatief voor goud of ze zien gewoonweg een toekomst voor de nieuwe technologie. Deze groeiende interesse zorgde in 2017 en begin 2018 voor een groei in nieuwe Initial Coin Offerings (ICO's), een relatief nieuw fenomeen. Een ICO is het crypto equivalent van een Initial Public Offering. ICO's zijn geschikt voor bedrijven die financiering nodig hebben bij nieuwe blockchainprojecten zoals tradingplatformen, NFT's, online spellen, *smart contracts* en nog vele andere projecten. Bij een ICO zullen bedrijven digitale tokens aanbieden die na voltooiing van het project gebruikt kunnen worden. Het uitgangspunt is dat deze tokens in waarde zullen toenemen wanneer het onderliggende project succesvol is. Ook werpen er als maar meer onderzoeksinstellingen en consultancyfirma's een blik op ICO's als legitieme bedrijfsfinancieringsmethode. Volgens icobench.com, een populair ICO-platform bij investeerders en onderzoekers, is er tot op heden meer dan 27 miljard dollar opgehaald met 5728 ICO's. Als je de volledige cryptomarkt bekijkt, die momenteel een marktwaarde heeft rond de 2 000 miljard Amerikaanse dollar, kan je toch spreken over een gigantische financiële markt. Eén die niet alleen amateur-investeerders, maar ook business angels, venture capitalists en investeringsbanken aanspreekt [1].

Gezien de relatief recente ontwikkeling en adaptatie van Initial Coin Offerings is hier tot op heden slechts een beperkt wettelijk kader rond opgesteld. Het ontbreken van een wettelijk kader in combinatie met de enorme populariteit van cryptomunten is de ideale kweekvijver voor oplichters. Zo publiceerde Statis Group LLC, een adviesbureau uit New York dat investeringsbanken bijstond bij de opkomende cryptomarkt, in 2018 een rapport waaruit blijkt dat men bij 80% van de geïntroduceerde Initial Coin Offerings fraude vaststelde. Er wordt dan gesproken over een ICO scam [2, 3]. In tegenstelling tot IPO's zijn er voor ICO's tot op heden vrijwel geen verplichtingen. Dat zorgt voor asymmetrische informatie, een fenomeen waarbij de ene partij (projectteam) veel meer informatie heeft dan de andere partij (investeerder). Om de investeerder toch te overtuigen om te investeren in het project zullen veel projecten een whitepaper publiceren. Een whitepaper is een document dat het projectteam ter beschikking stelt voor toekomstige investeerders. Hierin wordt zo veel mogelijk informatie over het project uitgeschreven zoals het probleem dat men tracht op te lossen, de oplossing die men wil ontwikkelen, de architectuur van hun tokens, het team, roadmap ... [4, 5].

Een potentiële investeerder zal normaliter de whitepaper bekijken alvorens te beslissen om al dan niet te investeren. Kennis omtrent het onderscheiden van een legitieme en een frauduleuze whitepaper kan voor hen dus een meerwaarde betekenen. Gelet op de recente opkomst van ICO's is het een onderbelicht onderzoeksgebied. Het onderzoek dat reeds is uitgevoerd legt de focus op de

succesfactoren. Daarnaast is er zeer weinig kwalitatief noch kwantitatief onderzoek gedaan naar de inhoud van whitepapers in de context van frauduleuze ICO's [6, 7]. Deze paper beschrijft in welke mate een frauduleuze ICO en een legitieme ICO van elkaar te onderscheiden zijn op basis van hun whitepaper.

Voor dit onderzoek werd er een verkennende literatuurstudie naar de succesfactoren van ICO's uitgevoerd. Deze zal de basis vormen voor de rest van het onderzoek. Het is namelijk interessant om te onderzoeken of deze succesfactoren al dan niet terug te vinden zijn in de scams. Verder is het, ondanks het feit dat ICO een relatief nieuw concept is, belangrijk dat er algemeen aanvaarde definities gebruikt worden binnen dit onderzoek. Deze concepten zullen in sectie 2 beschreven worden op basis van literatuur. Tot slot zal er eveneens ander onderzoek, zoals dat van Zhang et al. (2021) [8], over de impact van de positieve toon van een whitepaper op het succes van de ICO, gebruikt worden als basis voor bepaalde hypotheses. In sectie 3 wordt de methodologie besproken, meer specifiek de data verzameling en de methoden voor de analyses. In sectie 4 zal de exploratieve analyse toegelicht worden. De data zal uiteindelijk een selectieprocedure ondergaan om nadien gebruikt te worden bij het opstellen van een logistisch regressiemodel. Dat model zal een antwoord bieden op de vraag wat juist de verschillen zijn tussen een whitepaper van een frauduleuze en legitieme ICO. Die resultaten zullen in sectie 5 besproken worden. De discussie volgt in sectie 6 en tot slot is de conclusie terug te vinden in sectie 7 .

2 Literatuur

In deze sectie zal een achtergrond gegeven worden van de belangrijke concepten die gebruikt zullen worden doorheen het onderzoek. Ook wordt er op het einde van deze sectie een overzicht gegeven van bevindingen uit onderzoeken naar de succesfactoren van ICO's. De inzichten van die onderzoeken maken een groot deel uit van het vertrekpunt voor dit onderzoek.

2.1 Distributed ledger- en blockchaintechnologie

Alvorens we de centrale concepten van de paper bespreken, namelijk ICO's en de bijhorende whitepaper, is het nodig om *Distributed ledger technologies* (DLT's) en blockchaintechnologie toe te lichten. Onderzoekers aan de universiteit van Cambridge definiëren DLT's als volgt: Een *append-only* keten van cryptografisch met elkaar verbonden 'blokken' van gegevens, onderhouden en bijgewerkt door een gedecentraliseerd netwerk, waarbij netwerkknooppunten worden aangemoedigd door economische prikkels om zich op niet-strategische wijze in te zetten voor het onderhoud en de beveiliging van het systeem. Die structuur zorgt ervoor dat deze gegevens bestand zijn tegen interferentie door externen, censuur, vervalsing, manipulatie, of andere vormen van kwaadwillige handelingen [9]. Diezelfde gegevens zijn georganiseerd in een specifieke structuur die vaak wordt aangeduid als '*global ledger*' of 'globaal grootboek'.

Blockchaintechnologie is gebaseerd op dit protocol. Het is een gedistribueerd, gedecentraliseerd en onveranderlijk grootboek van transacties. Een transactie kan verwijzen naar een monetaire transactie, gebruikt in cryptocurrencies zoals Bitcoin, maar ook naar een gegevenstransactie [10].

2.2 Initial Coin Offering

Een Initial Coin Offering is een mechanisme waarmee nieuwe ondernemingen kapitaal ophalen door middel van het verkopen van tokens aan investeerders. Het is gelijkaardig aan de crowdfundingbenadering. Deze token is een cryptomunt en is bedoeld om een toekomstige functionele eigenschap te hebben in het project van de onderneming (bv. nutsfunctie, recht op eigendom en royalty's). De eerste ICO was Mastercoin in 2013 maar de meest gekende is ongetwijfeld Ethereum. Hierbij slaagde Vitalik Buterin er samen met Charles Hoskinson in om 500 000 dollar in te zamelen voor hun Ether project. Zoals in de inleiding reeds werd aangehaald zien vele mensen crypto als een volwaardige investering. Om grote rendementen te halen op succesvolle projecten ben je er dus best zo vroeg mogelijk bij. Het is dan ook voor investeerders interessant om te investeren in ICO's. De digitale tokens die bij een ICO horen, verlenen bepaalde rechten, zoals het recht op gebruik van de platformdienst of het eigendomsrecht. De tokens kunnen gekocht worden met andere cryptomunten en af en toe ook met reguliere munteenheden, zoals de Amerikaanse dollar. Nadat het project voltooid is zullen de tokens vrij verhandeld kunnen worden, vaak ook op verschillende cryptobeurzen van derde partijen zoals Coinbase en Crypto.com. Het aantal tokens dat gecreëerd wordt is beperkt. Bij een succesvol project zal de waarde van deze tokens ook stijgen [11].

ICO's zijn onder te verdelen in 2 grote categorieën: Utility tokens en Security tokens. In bepaalde literatuur zijn er naast die twee nog twee groepen gedefinieerd namelijk: Cryptocurrency tokens en Asset tokens.

- **Utility token**: Dit is bij uitstek de meest populaire categorie die aangeboden wordt via ICO's. Utility tokens geven hun houders in de toekomst toegang tot de diensten of producten van de start-up. Omdat deze tokens niet worden verkocht als beleggingsinstrument vallen ze niet onder de huidige wetgevingen. De tokens hebben op het moment van de uitgave nog geen onderliggende waarde [12].
- **Cryptocurrency token**: Ook Crypto munten worden soms als aparte categorie beschouwd. Dit zijn tokens die als doel hebben gebruikt te worden als betaalmiddel [13]. Ze kunnen beginnen als Utility Token voor een opkomend handelsplatform. Men verwijst regelmatig naar dit soort projecten als een Initial Exchange Offering (IEO) [13]. Dit is ook bij uitstek de bekendste categorie aangezien Bitcoin en Ethereum hieronder gecategoriseerd worden.
- **Security token**: Security tokens (vrij vertaald als 'Effectentokens') verlenen de eigenaars ervan het recht op bepaalde eigendomsrechten van een onderneming. Het is vergelijkbaar met aandelen. De gelijkenissen met een IPO is dan ook zeer groot. Deze tokens kunnen hiernaast ook stemrecht en recht op

dividenden met zich meebrengen. De waarde van dit soort token is dan ook afhankelijk van de waarde van het onderliggend actief. Het onderscheid met de twee andere soorten tokens is dermate significant dat men deze niet ziet als ICO, maar als een afzonderlijk gegeven, namelijk: STO (Security Token Offering) [13]. Gelet op deze aparte status en de controversiële regulering zijn er niet veel succesvolle STO's.

- **Asset token**: Een andere categorie die vaak terugkomt, is een Asset token. Dit zijn tokens die gegenereerd worden en horen bij fysieke activa zoals vastgoed, goud... Vaak zal men deze tokens als subcategorie zien van Security tokens [14]. Een voorbeeld hiervan is het recent gelanceerde platform van de bekende Youtube-ster Logan Paul, genaamd 'Liquid Marketplace'. Hierop 'tokeniseert' hij bepaalde activa. Zo zou hij bijvoorbeeld een schilderij van 500 000 euro kunnen tokenizeren naar 500 000 tokens van elk 1 euro. Hierdoor kunnen meerdere personen een deel van het schilderij 'bezitten'.

Bovenstaande opdeling is de meest voorkomende binnen de literatuur. Dat wil niet zeggen dat dat de enige is en dat alle crypto varianten hieronder thuis te brengen zijn. Een andere zeer populaire categorie zijn NFT's (Non Fungible Tokens). Deze zijn het afgelopen jaar in opmars gekomen. Het zijn unieke tokens die zich vaak in de vorm van kunst op de blockchain bevinden. Aangezien het gaat om één unieke token (non fungible) kan de waarde zeer hoog oplopen. Projectteams die als doel hebben een NFT-project te lanceren zullen dit vaak doen aan de hand van Utility tokens (om deze nadien in te wisselen met de NFT's) [15, 16].

2.3 Whitepaper

Binnen dit onderzoek zullen whitepapers van verschillende crypto projecten onder de loep worden genomen. Zoals eerder reeds vermeld werd is de whitepaper een essentieel onderdeel van een ICO. De term whitepaper is echter niet eigen aan ICO's. Deze bestond al vooraleer er sprake was van crypto en zelfs het internet. De eerste whitepaper is in 1879 in het Duitse Keizerrijk tot stand gekomen waarna de term verspreidde naar andere delen van de wereld. In die tijd werden whitepapers gebruikt voor twee zaken, namelijk: het presenteren van het vast beleid van de regering en, tegelijkertijd, het uitnodigen tot geven van advies over dat beleid [17]. Vanaf dat moment werd de term meer en meer gebruikt voor soortgelijke publicaties. Vanaf 1990 werd de term ook gebruikt door business to business marketeers, bedoeld om de producten of diensten van een specifiek bedrijf te promoten. Meer bepaald werden hier de probleembeschrijvingen en oplossingen in weergegeven, gecombineerd met argumenten waarom het onderliggende projectteam geschikt is voor dit project. Tot slot werd er ook een stappenplan in opgenomen, we spreken dan over een roadmap [18].

Deze onderdelen zijn herkenbaar in whitepapers die vandaag de dag gebruikt worden voor ICO's. Volgens Ofir en Sadeh (2020)[19], die onderzoek deden naar de verschillen tussen ICO's en IPO's, is er bij ICO's veel meer

informatieasymmetrie. Dat is te wijten aan een aantal zaken waaronder het ontbreken van openbaarmakingsverplichtingen en een gebrek aan fundamentele kennis bij investeerders. Een whitepaper zal deze tekortkomingen voor een groot deel proberen op te lossen.

Een whitepaper zal ook vaak de samenstelling van het projectteam omvatten en een overzicht van het aantal tokens dat niet wordt vrijgegeven bij de ICO. Hiervan zal een deel bestemd zijn voor het projectteam. In termen van een IPO of een crowdfunding kan je dit vergelijken met het aandelenbehoud van de oprichters. De literatuur toont aan dat deze gegevens een belangrijke rol spelen in het succes van crowdfunding [20, 21]. Onderzoek dat reeds gedaan is naar succesfactoren suggereert dat de toon van een whitepaper en de gelijkenis met andere whitepapers een impact heeft op het succes van een ICO. Namelijk dat een positieve toon in whitepapers vaak zorgt voor een succesvolle ICO [8, 22].

Heel wat opkomende projectteams bestaan voornamelijk uit blockchain ingenieurs en informatici die relatief weinig tijd spenderen aan marketing. In plaats van een uitgebreide whitepaper stelt men daarom soms enkel een slide deck of een litepaper ter beschikking. Dat vat de inhoud van een whitepaper samen met behoud van de belangrijke punten van het hele document [23]. Technisch gezien wordt er nog steeds gesproken over een whitepaper.

2.4 ICO Scams

Zoal reeds werd vermeld in de inleiding (zie sectie 1) is de opkomende cryptomarkt, meer specifiek de ICO-markt, populair bij oplichters [24]. In deze paper zal een scam gedefinieerd worden als 'een ICO die is opgericht met het doel om potentiële investeerders op te lichten'. In ander onderzoek werd de term scam meermaals gebruikt om een ICO die niet succesvol was aan te duiden. Een niet succesvolle ICO is echter niet noodzakelijk hetzelfde als een frauduleuze ICO, het zou perfect kunnen dat het project faalt ondanks de goede bedoelingen van het projectteam. Heel wat legitieme projecten zullen hun inzameldoel niet altijd halen, in dat geval zullen bepaalde projectteams er voor kiezen de fondsen terug te storten. Het zou ook kunnen dat het doel wel bereikt wordt, maar dat het product / de dienst totaal niet aanslaat en de prijs van de token dus daalt. Dat soort situaties zijn zeer verschillend van frauduleuze projecten waarop dit onderzoek zich focust. Op basis van literatuur en online artikels over effectief frauduleuze ICO projecten zijn er verschillende type scams te definiëren. [25, 26].

- **Ponzifraude**: De klassieke piramide fraude waarbij de eerste winsten betaald worden met de inbreng van nieuwe investeerders. Na een grote populariteit verdwijnt het projectteam.
- **Exit Scam**: Hierbij 'trekt' het projectteam al de ingezamelde fondsen weg uit het project.

- **Kopieer Scam**: Veel crypto projecten zijn opensource met als gevolg dat bepaalde fraudeurs de bestaande code en whitepaper kopiëren en nadien een namaak token op de markt trachten te brengen.
- **De 'Pump and dump' Scam**: Fraudeurs 'pompen' de prijs artificieel op door extreem veel advertenties of door zelf tokens aan elkaar te verkopen aan een hoge prijs waardoor de marktprijs van deze token artificieel hoog staat. Nadien verkopen ze deze tokens aan deze verhoogde prijs en verlaten ze het project.
- **Phishing scams**: Waarbij fraudeurs phishing websites opzetten onder de dekmantel van een ICO. Vaak lokken ze hun slachtoffers via Phishing e-mails wiens contactinformatie voornamelijk verzameld wordt aan de hand van cryptoblogs.

2.5 Succesfactoren

De succesfactoren van ICO's zijn met voorsprong de meest prominentie focus bij onderzoeken over ICO's en whitepapers van ICO's. Zoals aangeven in de introductie zullen deze onderzoeken een basis vormen voor dit onderzoek. Uiteraard wordt er voornamelijk gekeken naar factoren die te maken hebben met de whitepaper. In deze subsectie wordt er een overzicht gegeven van die factoren.

Het eerste thema, de invloed van de toon van een whitepaper op het succes, is al aangeraakt in subsectie 2.3 over whitepapers. Een positieve toon in de whitepaper zou voor een succesvolle ICO zorgen [8]. Het onderzoek van Zhang et al. (2021) maakte gebruik van een sentiment analyse om woorden met een negatieve en positieve toon te identificeren. Op die manier werd de algemene toon van de whitepaper gedefinieerd.

Het blijkt ook dat de aanwezigheid van usecases of een prototype een grote kans op slagen betekent [27, 28]. Het onderzoek dat rond dat onderwerp gebeurde, beschreef alle tokens die een grondige beschrijving hadden van hun usecases of eventueel prototype als 'Utility Token'. De onderliggende data van het onderzoek toont aan dat bepaalde Asset tokens, die hun usecases goed beschreven, vaak door middel van een mock-up of prototype, voor dat onderzoek ook als Utility tokens gezien werden. Om verwarring te vermijden met ander onderzoek wordt er gedurende dit onderzoek gebruik gemaakt van de benaming 'usecases' bij de aanwezigheid van usecases die grondig beschreven worden, vaak aan de hand van mock-ups of een prototype.

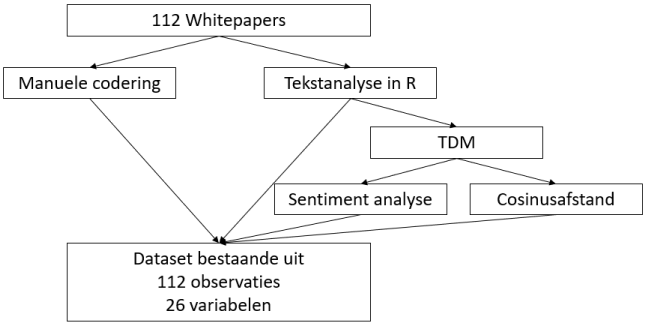
Ander onderzoek beschrijft ook dat de lengte van een whitepaper en de grote van het projectteam een positieve impact hebben [5, 7, 29]. Ook is er onderzoek gedaan naar de aanwezigheid van een roadmap en een juridische disclaimer. De aanwezigheid van deze 2 zaken had echter geen significante impact op het succes [7]. Wel wees onderzoek uit dat een technische uitleg of een link naar GitHub met de technische informatie een positieve impact had op het succes van een ICO [28].

In een onderzoek omtrent de inhoud van whitepapers maakt men gebruik van cosinusgelijkheid [29]. Een afstandsmaat die aangeeft in welke mate documenten op elkaar lijken. Er werd hier echter onderzocht wat de meest voorkomende onderwerpen zijn en niet wat de impact van deze gelijkenis op het succes was. Er zijn echter wel onderzoeken naar de classificatie van documenten die deze maatstaf ook gebruiken [30]. Het onderzoek naar de content van whitepapers is het enige onderzoek dat werd gevonden dat ook onderzoek deed naar frauduleuze ICO's. Dat onderzoek heeft uitgewezen dat wanneer het gaat om een whitepaper die niet lijkt op een professioneel artikel met veel afbeeldingen, de kans op fraude groot is [4, 31]. Er werd hier in deze onderzoeken een algoritme gebruikt om de complexiteit van de paper te onderzoeken.

Er is ook heel wat literatuur aanwezig over de succesfactoren bij crowdfunding. Aangezien het hier gaat om een minder recent fenomeen is hier meer literatuur over te vinden. Deze onderzoeken maakten onder andere gebruik van kwalitatieve variabelen om het succes van de crowdfunding te voorspellen. Het gaat hier bijvoorbeeld om professionele uitstraling en professionaliteit van de video [32, 33].

3 Methodologie

Het doel van deze studie is om een antwoord te formuleren op de vraag wat juist de verschillen zijn tussen een whitepaper van een frauduleuze en legitieme ICO. Dat onderzoek is gevoerd op basis van een dataset van whitepapers. Het onderzoek hanteert verschillende methodes om bepaalde, vooraf gedefinieerde, kenmerken uit de whitepapers te halen. De variabelen zijn gebaseerd op de literatuur rond succesfactoren, toegelicht in sectie 2.5. De gehele totstandkoming van de dataset wordt weergegeven in figuur 1.



Figuur 1: Totstandkoming dataset

Deze dataset zal gebruikt worden om relationele testen op uit te voeren om na te gaan welke kenmerken statistisch significant verschillen voor scams en niet-scams. Tot slot zal er in deze sectie ook de logistische regressie besproken

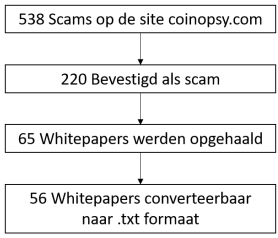
worden. Dat is de methode die deze paper zal hanteren om te kunnen antwoorden op de onderzoeksvraag. Deze meet de relatie tussen de geselecteerde variabelen en de variabele die aangeeft of de ICO een scam is. Voordat dat model opgesteld wordt, wordt er aan *feature selection* gedaan met behulp van 2 algoritmen die de belangrijkste variabelen zullen selecteren van het model. Die algoritmen zullen ook in deze sectie besproken worden.

3.1 Dataverzameling

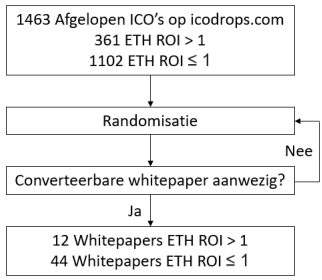
Dit onderzoek had als doel een gebalanceerde dataset samen te stellen die voor de helft bestond uit whitepapers van ICO scams en voor de andere helft uit whitepapers van goede (niet frauduleuze) ICO's. Het gros van de onderzoeken naar succesfactoren wordt gedaan op basis van de website ICObench.com [34], een website die frauduleuze ICO's onmiddellijk verwijdert en dus niet bruikbaar is voor dit onderzoek. Voor de verzameling van ICO scams werd er daarom gebruik gemaakt van de site coinopsy.com. Dit is een database voor uitgestorven cryptotokens die door middel van inzendingen aangevuld wordt. De uitgestorven tokens worden onderverdeeld in een aantal categorieën waaronder *Scam or Other issues*. Binnen dit onderzoek wordt er niet gekeken naar tokens waarbij het overduidelijk om scams gaat zoals 'Ponzicoin' en 'Scamcoin'. Dat waren ICO's die vrij duidelijk aangaven dat het om een piramide systeem ging. Desondanks zijn er een heel aantal mensen die hierin investeren, wellicht met de gedachte dat piramidesystemen voor de eerste investeerders ook winstgevend kunnen zijn. Deze worden niet meegenomen in dit onderzoek. Dat omdat het onderzoek het probleem tracht te bekijken vanuit het standpunt van serieuze investeerders die op zoek zijn naar een ICO om in te investeren en niet als grap of als stunt willen investeren. Daarom zal er ook niet gekeken worden naar tokens die overduidelijk als grap gecreëerd zijn, zoals: Dogecoin, Joke coin, Freecoin, enzovoort.

Na het verwijderen van dit soort ICO's uit de lijst worden de overgebleven tokens manueel onderzocht om te kijken of deze effectief frauduleus zijn. Dit is nodig aangezien het gaat om een opensource database die werkt met inzendingen. Dit werd gedaan door de ICO's te vergelijken met een andere database van scams cryptoscambrokers.com. Daarnaast werd er ook gekeken naar socialmediaposts op Twitter en Reddit alsook naar crypto blogs zoals bitcointalk.org. Op dit soort sites werd vaak een duidelijke beschrijving gegeven over de werking van de scam. Dit soort blogpost werd bovendien ondersteund door screenshots van een oude website. Wat regelmatig voorkwam was dat de sociale media en de website van het project verwijderd waren. Er was ook niet altijd een blogpost of Redditpost over de scam terug te vinden. Een andere tool die gebruikt werd om te onderzoeken of het over een scam ging was de website isthiscoinascam.com. Hierop worden ratings aan bepaalde cryptoprojecten. Wanneer de bovenstaande methoden geen duidelijkheid konden scheppen werd de ICO niet meegenomen.

Het vinden van de whitepapers is echter vrij uitdagend. De sites van deze ICO's waren namelijk vrijwel altijd offline gehaald waardoor het onderzoek



Figuur 2: Verzamelmethode ICO scams



Figuur 3: Verzamelmethode goede ICO's

gebruik diende te maken van het internet archive (web.archive.org), een tool die gebruikt wordt om websites te bezoeken op een bepaalde dag in het verleden. Ondanks de extreem grote databank van deze website was een groot deel van de websites niet terug te vinden. Ook was het niet altijd even vanzelfsprekend om de oude URL terug te vinden die nodig is om het archief te gebruiken. Indien de website wel bestond in het archief ontbrak er heel af en toe een whitepaper wat niet onlogisch is aangezien onderzoek aantoont dat dit bij 20% van de ICO's het geval is [27, 35].

Van de 538 tokens die op coinopsy gemarkeerd zijn als *scam or other issue* waren er 220 die geverifieerd konden worden als een effectieve scam. Daarvan werden er voor 65 tokens een whitepaper teruggevonden. Deze whitepapers werden nadien omgezet van pdf formaat naar text formaat voor verdere analyse. Voor 9 van deze Witepapers was dit echter niet mogelijk omdat het ging om ingescande documenten of niet converteerbare pdf's. Een overzicht wordt gegeven in Figuur 2.

De groep van legitieme ICO's bestaat idealiter uit representatieve ICO's, die zowel succesvol als onsuccesvol zijn. Op basis van de gevonden literatuur zien we dat dat 20% van de ICO's een ETH ROI groter dan 1 heeft [36]. Ethereum return on investment geeft het rendement weer in Ethereum, bijvoorbeeld: bij de ICO kost 1 token 0,05 ETH en nu heeft de token een waarde van 0,15 ETH, dan spreken we over een ETH ROI van 3 aangezien de prijs verdriedubbeld is. Door ETH ROI te gebruiken hebben de algemene fluctuaties in de cryptomarkt een beperkte impact. Dit omdat de prijs van ethereum als standaard wordt gezien die mee fluctueert met de algemene trends. Het komt er dus op neer dat er bij 20% van de gevallen voordeel gehaald wordt door te investeren in de ICO. Bij de andere 80% kon er beter geïnvesteerd worden in Ethereum. De dataset met legitieme ICO's werd daarom ook op die manier samengesteld om de realiteit zo goed mogelijk de weerspiegelen.

De listing website icodrops.com wordt gebruikt om de ETH ROI en de link naar de website van de ICO's te vinden. Op deze site staan 1463 afgelopen

ICO's waarvan er 361 met een ETH ROI groter dan 1. Door middel van randomisatie werden er 12 van deze 361 ICO's geselecteerd en 44 uit de groep waarvan de ROI kleiner of gelijk aan 1 is.

Cryptoprojecten lanceren geregeld geactualiseerde versies van hun whitepaper. Als er geen whitepaper 1.0 op de website te vinden was werd er gebruik gemaakt van het internet archief. Wanneer er geen converteerbare whitepaper te vinden was die dateerde van de ICO werd er een nieuwe willekeurige ICO geselecteerd. Het gehele proces wordt weergegeven in figuur 3.

Bij zowel ICO scams als goede ICO's werden er een aantal variabelen manueel verzameld. Hieronder wordt een overzicht gegeven van die variabelen. De variabelen zijn gebaseerd op de onderzoeken naar succesfactoren zoals beschreven in sectie 2.5. De variabelen Aantal_talen, Slide_deck en Lite_paper zijn niet gebaseerd op eerder onderzoek. Aangezien Slidedecks en litepapers technisch gezien ook whitepapers zijn, zullen deze op exact dezelfde manier geanalyseerd worden. Om deze informatie niet verloren te laten gaan wordt de informatie opgeslagen in de variabelen Slide_deck en Lite_paper. De reeds bestaande literatuur over succesfactoren, besproken in sectie 2, maken ook geen onderscheid met volwaardige whitepapers in hun analyse. Net omdat de grens met volwaardige whitepapers zo onduidelijk is. Heel wat projectteams gebruiken de term litepaper niet terwijl het een zeer korte whitepaper is. Of ze publiceren een zeer kleurrijk artikel dat zeer nauw aansluit bij een slidedeck.

De variabele 'Artikelvorm' is manueel verzameld in tegenstelling tot ander onderzoek waarbij er gedeeltelijk algoritmen gebruikt zijn [4, 31]. Er werd hier gekeken of de whitepaper volgens een officiele standaard (APA, MLA, Chicago,...) is samengesteld. En of het om een professionele paper ging. Bij het verzamelen van deze variabele werd er niet gekeken of het al dan niet om een ICO scam ging om dit zo objectief mogelijk te houden. Een techniek die ook gebruikt werd bij het onderzoek naar succesfactoren voor crowdfunding [32, 33]. De waarde van deze variabele is dus het resultaat van menselijk oordeel van de onderzoeker.

Zoals toegelicht in sectie 2.5 zal er gebruik gemaakt worden van de benaming 'usecases' bij de aanwezigheid van usecases die grondig beschreven worden, vaak aan de hand van mock-ups of een prototype. Dit is een expliciet onderdeel van een whitepaper waarmee bewezen wordt dat er na de ICO een gebruik is voor de tokens [4].

- **Aantal_pagina**: Een numerieke variabele die het aantal pagina's van de whitepaper weergeeft.
- **Aantal_talen**: Een numerieke variabele die beschrijft in hoeveel talen de whitepaper beschikbaar was op het moment van de ICO.
- **Artikelvorm**: Een binaire variabele die aangeeft of de whitepaper in de vorm van een professioneel artikel komt volgens een officiele standaard.
- **ICO**: De naam van de ICO
- **Legal_disclaimer**: Een binaire variabele die aangeeft of de whitepaper een juridische disclaimer bevat.

- **Litepaper**: Een binaire variabele die weergeeft of de ICO enkel een Litepaper ter beschikking stelt. En hier zelf ook zo naar verwijst.
- **Roadmap**: Een binaire variabele die de waarde één heeft indien er een roadmap aanwezig is. Een roadmap is het stappenplan dat het projectteam zal volgen en vaak data bevat van de lancering van het project.
- **Scam**: Een binaire variabele die aangeeft of het om een frauduleuze ICO gaat.
- **Slide_deck**: Een binaire variabele die aangeeft of de whitepaper een slide deck is.
- **Team**: Een numerieke variabele die het aantal teamleden weergeeft die met naam genoemd worden in de whitepaper of op de website.
- **Technische_uitleg**: Een binaire variabele die aangeeft of er een technische uitleg gegeven wordt in de whitepaper, of een link naar GitHub met de technische informatie.
- **Usecases**: Een binaire variabele die aangeeft of de whitepaper usecases of een protype beschikbaar stelt [4].

3.2 Tekstanalyse in R

Naast manueel geïdentificeerde kenmerken zijn er ook een aantal kenmerken ontgonnen met behulp van *Natural Language Processing* (NLP). Dat zijn algoritmes die grote hoeveelheden teksten kunnen verwerken en vervolgens nauwkeurig informatie en inzichten uit die documenten halen [37]. Zo kunnen er algoritmen bepaalde documenten herkennen en classificeren. In dit onderzoek werden in totaal twee soorten NLP gebruikt: sentimentanalyse en Cosinusafstand. Deze algoritmen en de toepassing op dit onderzoek zullen toegelicht worden in de volgende subsecties.

Om teksten te kunnen analyseren moeten deze eerst uitgezuiverd worden. Dat gebeurd aan de hand van R packages: Tidytext, Tm en Quanteda. De verzameling van teksten voor NLP noemt men ook een 'corpus'. Hieruit werden tekens, cijfers en onnodige witregels verwijderd. Nadien werd de tekst getokenized per woord. Waarna de variabelen **Aantal_woorden** en **Woorden_per_pagina** voor iedere whitepaper berekend kon worden. Vervolgens werden Engelse stopwoorden verwijderd en woorden herleid naar hun stamvorm. Zo werden 'tokens' en 'token' niet langer beschouwd als twee verschillende woorden. De woorden werden nadien verzameld in een term document matrix (TDM), wat in feite een matrix is met als rijen alle unieke woorden en als kolommen de documenten. De cellen geven aan hoe vaak ieder woord voorkomt in elk document. Op basis van het uitgezuiverde corpus en de TDM kunnen analyses uitgevoerd worden.

3.3 Sentimentanalyse

Een sentimentanalyse van een tekst is een vrij bekende techniek. Het is een systematische meting waarbij een gevoel of stemming rondom berichtgeving in kaart gebracht wordt. Het is niet noodzakelijk een NLP methode waarbij het

automatisch via een algoritme en een lexicon gebeurt, het kan ook manueel gebeuren [38]. In dit onderzoek gebruiken we echter een geautomatiseerde versie. Binnen dit onderzoek wordt er gebruik gemaakt van de nrc database, wat een enorme lexicon is aan woorden (14 182 woorden) waarbij 10 categorische variabelen horen. Ieder woord uit het lexicon kan geassocieerd worden aan acht emoties: woede, angst, anticipatie, vertrouwen, verrast, verdriet, vreugde en afkeer en aan één gevoel (positief of negatief) [39, 40]. Het woord 'abortion' kan bijvoorbeeld geassocieerd worden met een negatief gevoel en met de emoties: afgunst, angst en verdriet. De acht emoties zijn gebaseerd op *Plutchik's wheel of emotions*, een bekende methode uit de literatuur om emoties in kaart te brengen [41]. Dat maakt ook dat het nrc lexicon veruit de populairste is voor sentimentanalyses.

Zoals eerder aangehaald is er onderzoek gedaan naar de impact van de toon van de whitepaper op het succes [8, 22]. In kader van dit onderzoek zal deze sentimentanalyse methode dus gebruikt worden om de relatieve emotie en gevoel van de tekst in kaart te brengen. Door de het nrc lexicon te combineren met iedere kolom in de TDM werd er berekend hoeveel woorden geassocieerd konden worden per gevoel en emotie. Op die manier wordt er per document een waarde gegeven aan elke emotie. Wanneer een document 50 woorden bevat die de emotie 'anticipatie' oproepen krijgt anticipatie de waarde 50. Die gegevens zijn omgevormd naar relatieve cijfers zoals aanbevolen bij het gebruik van de nrc database [39]. Dit door de waarde van de individuele emotie te delen door de som van alle emoties. Omdat het doel is om een som van 100% te verkrijgen wordt er niet gedeeld door het totaal aantal woorden die geassocieerd kunnen worden aan een emotie. Woorden kunnen namelijk meerdere emoties hebben. De toon wordt op diezelfde manier berekend. Daaruit vloeien de variabelen **Positief** en **Negatief** die samen 1 vormen (100%) en **Woede**, **Angst** **Anticipatie**, **Vertrouwen**, **Verrast**, **Verdriet**, **Vreugde** en **Afkeer** die op hun beurt ook een som hebben van 1 (100%). De variabele **Afkeer** geeft bijvoorbeeld weer in welke mate de tekst de emotie 'afkeer' oproept. Dat is ook de methode die de ontwikkelaars van de nrc database aanraden [39].

3.4 Cosinusafstand

De cosinusgelijkheid is een populaire maatstaf bij het berekenen van de gelijkenis tussen 2 bestanden. Het is gebaseerd op de cosinusafstand, deze meet de gelijkenis aan de hand van de cosinus van de hoek tussen twee vectoren in een multidimensionale ruimte [42]. Bij het vergelijken van 2 documenten zullen deze documenten afgebeeld worden als vectoren, in dit geval kolommen uit de TDM. Die vector representatie bestaat uit cijfers die weergeven hoe vaak een bepaald woord (de rij in de TDM) voorkomt in een document. De formule voor de cosinusafstand is te zien in formule 1.

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{xy}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{x}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{y}_i)^2}} \tag{1}$$

Stel dat x en y twee vectoren (whitepapers) zijn, waarbij x de vector (x1, x2, ... xn) is en $|| x ||$ de euclidische norm van x is, idem voor y. De euclidische norm van een vector berekent men door alle elementen van die vector te kwadrateren en nadien op te tellen om hier vervolgens de vierkantswortel van te nemen. De berekening van de cosinusafstand zal uiteindelijk een waarde tussen 0 = identiek en 1 = volledig verschillend als uitkomst geven [42].

De cosinusafstand is een veel voorkomende afstandsmaat die voor veel projecten gebruikt word gaande van het vergelijkingen van teksten tot het vergelijken van afbeeldingen [43]. Het voordeel van deze methode is dat het de afstand meet tussen vectoren en dus onafhankelijk is van de lengte van de tekst.

Wanneer documenten vergeleken worden met behulp van hun TDM wordt er aangeraden weinig voorkomende termen te verwijderen. Wanneer dit niet gedaan wordt zullen vrijwel alle afstandswaarden extreem dicht bij 1 liggen. Voor dit onderzoek werden alle termen die in minder dan 10% van de documenten voorkomen, verwijderd [43]. Met de TDM dat hieruit resulteert berekenen we de cosinus afstand tussen ieder document waarbij de rijen van de TDM de whitepapers in vectorformaat zijn. Dit resulteert in een afstandsmatrix. Dat is in dit geval een matrix waarbij de kolommen en rijen bestaan uit de ICO namen. De celwaarden zijn de cosinus afstanden tussen de twee whitepapers op een schaal van nul tot één. Nadien werden de diagonalen omgevormd tot NA waarden zodat de gelijkenis tussen dezelfde whitepapers niet meegerekend werd. Iedere whitepaper heeft dus 111 waarden die de gelijkenis tussen die paper en een andere paper weergeeft.

Voor iedere whitepaper kan men twee waarden achterhalen. Namelijk de gemiddelde gelijkenis met whitepapers van scams en de gemiddelde gelijkenis met whitepapers van goede ICO's, respectievelijk **Cos_dis_scams** en **Cos_dis_goede**. Neem bijvoorbeeld een willekeurige ICO scam, hier wordt het gemiddelde van de 55 cosinusafstanden, tussen die whitepaper en de 55 andere ICO scams, genomen. Idem voor het gemiddelde van de cosinusafstanden met de 56 goede ICO's. De hypothese is namelijk dat wanneer het om een ICO scam gaat, deze meer lijkt op de whitepapers van andere ICO scams. Dit is ook een techniek die gebruikt wordt bij het classificeren van tweets op het politiek spectrum [44].

3.5 Statistische Testen

In de voorgaande secties werd er toegelicht hoe er voor iedere ICO bepaalde gegevens werden verzameld. Deze dataset, bestaande uit 112 ICO's met 26 variabelen, zal gebruikt worden om te onderzoeken wat de significante verschillen zijn tussen ICO scams en goede ICO's. Onze target variabele zal dus altijd **Scam** zijn, een binaire variabele die 1 is bij een scam, en 0 bij een goede ICO. Dit onderzoek maakt gebruik van de t-toets, dat is een statistische toets die wordt gebruikt om de gemiddelden van twee groepen te vergelijken. Meer specifiek wordt er gekeken of er een significant verschil is tussen de twee. Dat wordt

gedaan aan de hand van de t-waarde die berekend wordt volgens formule 2.

$$t = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{1/n + 1/m}} \tag{2}$$

Een grotere t-waarde geeft aan dat het verschil tussen de groepsgemiddelden groter is dan de standaardafwijking, wat duidt op een significanter verschil tussen de groepen [45].

Voor de categorische variabelen wordt de Chi-kwadraattoets gebruikt. Een populaire methode om de statistische significantie te berekenen tussen 2 categorische variabelen [46]. Meer specifiek is het een statistische toets om te evalueren hoe waarschijnlijk het is dat een waargenomen verschil tussen de reeksen door toeval is ontstaan. Het vormt de basis voor een aantal andere methoden zoals de cramer's V maatstaf [47]. Een maatstaf die aangeeft hoe sterk de samenhang is tussen twee groepen (0 is geen samenhang en 1 is perfecte samenhang).

Formule 3 geeft weer hoe de Chi-kwadraat en Cramer's V worden berekend. Waarbij O de geobserveerde waarde is en E de verwachte waarde. Bij de formule van Cramer's V staan I en J voor het aantal rijen en kolommen en n voor de grote van de populatie. Aangezien er voor dit onderzoek 112 observaties zijn en geen missende waarden is n altijd 112. Er worden gedurende dit onderzoek ook enkel binaire variabelen met elkaar vergeleken wat ervoor zorgt dat J en I altijd 2 zijn. Cramer's V maakt namelijk gebruik van een kruistabel [48]. Een tabel die de relatie tussen 2 categorische variabelen weergeeft. Wanneer het dus om een binaire variabele gaat zal die tabel maar uit 2 rijen en 2 kolommen bestaan.

$$\chi^2 = \sum \frac{(O-E)^2}{E};\ V = \sqrt{\frac{\chi^2/n}{min(I,J)-1}} = \sqrt{\frac{\chi^2}{n}} \tag{3}$$

Zowel de t toets als de Chi-kwadraattoets zullen een p-waarde genereren. De p-waarde (p-value) is een getal tussen 0 en 1, waarmee je bepaalt of een steekproefuitkomst statistisch significant is. Wanneer de p-waarde kleiner is dan het gekozen significantieniveau kun je stellen dat de gevonden uitkomst extreem genoeg is om je nulhypothese te verwerpen. Stel dat je een significantie niveau van 0.05% hanteert dan zijn de p-waarden die hier onder vallen statistisch significant.

3.6 Feature selectie

De dataset waar we mee werken bevat naast de target variabele scam, 25 variabelen, waarvan er 1 de naam van de ICO is. We hebben dus 24 variabelen die een potentiële voorspelbare factor hebben. Zoals iedere dataset zijn er bepaalde variabelen die een hogere predictieve waarde hebben. We selecteren best de meest belangrijke variabelen om het model niet te complex te maken. Ook is het selecteren van variabelen belangrijk om overfitting te voorkomen.

Dat is wanneer het model te nauw of precies overeenkomt met een reeks gegevens. Gelet op de beperkte dataset die gebruikt wordt in deze studie is dat geen onbelangrijk risico. Feature selectie gebeurt typisch niet enkel op basis van domeinkennis. Dat zou immers kunnen leiden tot een eventuele bias. Om die reden wordt er gebruik gemaakt van 2 algoritmen om aan feature selectie te doen.

Voor de categorische variabelen is er gebruik gemaakt van het ExtraTrees-Classifier algoritme in Python [49]. Dat is een algoritme dat bouwt op de kracht achter het mechanisme van de beslissingsbomen. Het algoritme maakt verschillende belsissingsbomen en vergelijkt vervolgens de prestaties van al deze modellen. Op basis van die vergelijking bepaalt het algoritme welke categorische variabelen het belangrijkste zijn.

Voor de continue variabelen werd er gebruik gemaakt van het Boruta pakket in R [50]. Dat is een algoritme dat bouwt op de kracht achter het mechanisme van de random forest methode. Burata vertrekt vanuit de originele dataset X en creëert een andere dataset door de waarden van elk kenmerk willekeurig te herschikken. Deze herschikte kenmerken worden schaduw kenmerken genoemd. Op dit punt wordt het schaduw dataframe aan het oorspronkelijke dataframe gekoppeld om een nieuw dataframe te verkrijgen, dat tweemaal het aantal kolommen van X heeft.

Nadien wordt de random forest methode toegepast op de samengestelde dataset. Nadien worden de originele kenmerken vergeleken met hun drempel. Deze drempel is gedefinieerd als de impact van diens schaduw kenmerk. Wanneer de impact van een origineel kenmerk hoger is dan deze drempel, wordt dit een 'hit' genoemd. Het idee is dat een kenmerk alleen nuttig is als het in staat is het beter te doen dan het herschikte kenmerk.

Dat proces wordt een aantal keer herhaald om de impactvolle kenmerken te identificeren. In dit onderzoek zullen er 500 iteraties gebeuren. Wanneer een variabele dan bijvoorbeeld 450 keer als 'hit' werd bestempeld zal het gecategoriseerd worden als een sterke variabele.

Omdat deze algoritmes niet enkel kijken naar de relatie tussen een feature en de target-variabele behoort dit algoritme tot de zogenaamde w*rapper-type feature selection algorithms* [51] . Dit betekent dat een machine learning algoritme, zoals bijvoorbeeld decision trees of random forest, gebruik wordt om features te selecteren. Veel data-analysten verkiezen zulke methoden omdat dit soort algoritmes de relaties tussen de verschillende features mee in rekening neemt.

3.7 Logistische regressie

Het doel van logistische regressie is het voorspellen van een binaire afhankelijke variabele op basis van meerdere onafhankelijke variabelen (voorspellers) [52]. In ons geval zal die binaire variabele de variabele 'scam' zijn. Het is een gelijkaardige techniek als multipele lineaire regressie waarbij de afhankelijke variabele een lineaire combinatie is van de voorspellers. De logistische regressie berekend de kans op één van beide mogelijkheden van de binaire afhankelijke

variabele op basis van de voorspellers volgens formule 4. Een makkelijkere wiskundig benadering is formule 5.

$$P(Y_i = 1 \| X_i) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)} \quad (4)$$

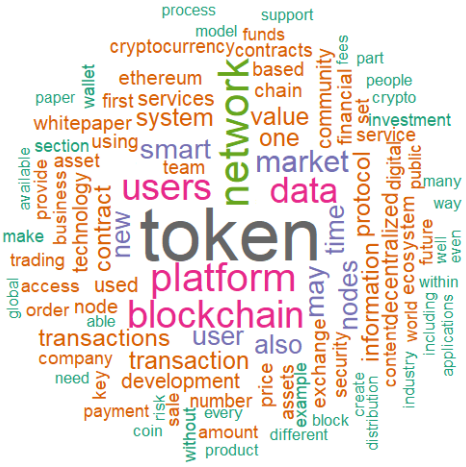
$$f(y_i) = ln \frac{P(y_i)}{1 - P(y_i)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \quad (5)$$

In de formules is y de voorspellende variabele, in ons geval 'scam' β_0 is hier een constante en X_i de onafhankelijke variabelen (voorspellers). β_i zijn de logistische regressiecoëfficiënten. Ze zijn de parameters die horen bij iedere variabele om tot een voorspelling te komen. De parameters worden berekend door middel van de *maximum likelihood estimation* [53]. Een methode waarbij de *likelihood* functie van de geobserveerde data gemaximaliseerd wordt.

4 Exploratieve data-analyse

In deze sectie wordt de verkende data-analyse van het onderzoek beschreven. In de voorgaande Sectie zijn alle variabelen reeds toegelicht, daarom wordt dat in deze sectie achterwege gelaten. Voor extra toelichting is er als bijlage A1 een overzicht te vinden.

Figuur 4 is een woordenwolk van alle woorden die voorkomen in de whitepapers. Token is veruit het meest voorkomende woord dat tevens in iedere whitepaper voorkomt. Netwerk, platform, blockchain, data en users zijn de andere veel voorkomende woorden. Wanneer de woordenwolk wat beter bekeken wordt, krijgt men een goed beeld van de boodschap die projectteams willen brengen. Zo zijn woorden zoals smart, security en value populair.



Figuur 4: Woordenwolk voor alle whitepapers

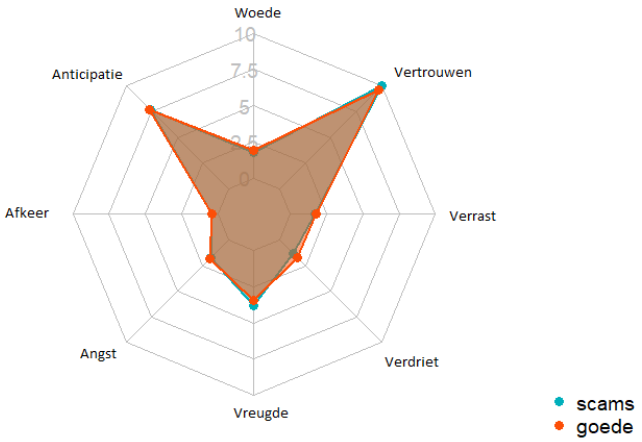
De dataset die gebruikt wordt voor de verdere analyse is toegelicht in tabel 1. De tabel geeft weer wat het algemeen gemiddelde, minimum en maximum is per variabele. Ook geeft de kolom 'Meth' de gebruikte methode weer die gebruikt is om de waardes te identificeren (Sentiment, Tekstanalyse in R, Manueel en Cosinusafstand). De kolommen S en G zijn respectievelijk de gemiddelden voor de ICO scams en voor de goede ICO's. Hier zijn al een aantal interessante zaken uit vast te stellen.

Tabel 1: Een samenvatting van de gehele dataset

Variabele	Meth	Gem	Min	Max	S	G
Woede	Sent	0,0643	0,0055	0,1485	0,0624	0,0663
Anticipatie	Sent	0,2702	0,1640	0,4421	0,2696	0,2708
Afkeer	Sent	0,0143	0	0,0607	0,0141	0,0146
Angst	Sent	0,0626	0	0,1526	0,0614	0,0638
Vreugde	Sent	0,1275	0,0397	0,2492	0,1345	0,1205
Verdriet	Sent	0,0540	0	0,1932	0,0461	0,0620
Verrast	Sent	0,0617	0,0176	0,2232	0,0596	0,0638
Vertrouwen	Sent	0,3452	0,1881	0,5336	0,3521	0,3382
Negatief	Sent	0,2738	0,1077	0,5535	0,2589	0,2887
Positief	Sent	0,7262	0,4465	0,8923	0,7412	0,7113
Aantal_woorden	R	5.845	332	42.255	4.548	7.142
Aantal_talen	M	1,81	1	8	1,55	2,07
Roadmap	M	0,90	0	1	0,88	0,93
Team	M	7,21	0	47	5,29	9,14
Aantal_pagina	M	28,69	5	136	25,30	32,07
Artikelvorm	M	0,26	0	1	0,13	0,39
Technische_uitleg	M	0,52	0	1	0,29	0,75
Litepaper	M	0,07	0	1	0,05	0,09
Slide_deck	M	0,08	0	1	0,11	0,05
Usecases	M	0,61	0	1	0,39	0,82
Legal_disclaimer	M	0,30	0	1	0,30	0,29
Scam	M	0,50	0	1	1	0
Woorden_per_pagina	R	193,42	41,50	585,53	171,90	214,93
Cos_dis_goede	Cos	0,729	0,613	0,866	0,732	0,726
Cos_dis_scams	Cos	0,72	0,591	0,874	0,715	0,732

90% van de ICO's heeft een roadmap aangeboden en 30% heeft een juridische disclaimer in zijn whitepaper staan. In totaal zijn er maar 8 ICO's die enkel een litepaper aanboden en 9 die enkel een slidedeck aanboden. De gemiddelde whitepaper van een scam bestaat uit 4548 woorden en 25 pagina's, die van een legitieme uit it 7142 woorden en 32 pagina's. Er zijn nog een aantal inhoudelijke verschillen op te merken. Legitieme ICO's bevatten vaker usecases en hebben vaker een professioneel artikel als whitepaper. Legitieme ICO's bieden hun whitepaper gemiddeld in 2 talen aan terwijl het gemiddelde voor scams 1,5 is.

De variabelen van de sentimentanalyse dienen echter visueel voorgesteld te worden. In figuur 5 worden ze voorgesteld in een radar-grafiek volgens het wiel van emoties van Plutchik [41].



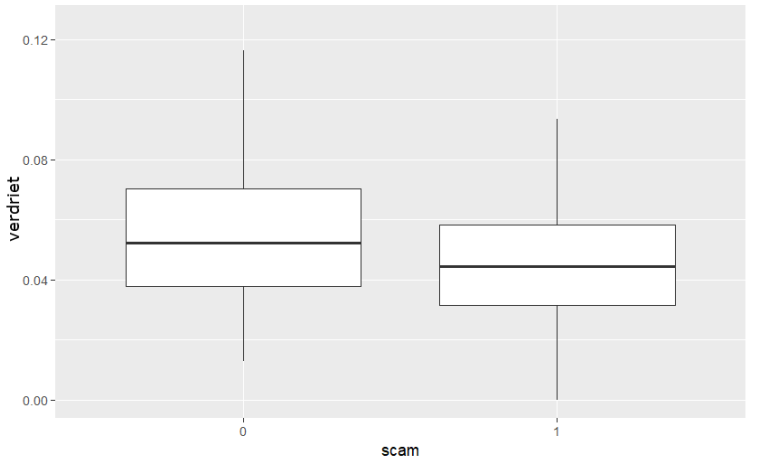
Figuur 5: Sentimentanalyse volgens Plutchik [41]

Alhoewel er 2 grafieken zijn weergegeven, namelijk de sentiment analyse voor scams en goede ICO's, zijn er nauwelijks verschillen te zien met het blote oog. Er kan wel afgeleid worden dat de meeste whitepapers veel woorden gebruiken die geassocieerd worden met vertrouwen en anticipatie. Wat overeenkomt met de literatuur die te vinden is over de opstelling van whitepapers [4]. Whitepapers zijn bedoeld om de investeerder te doen vertrouwen in het project en duidelijk te maken wat de toekomstplannen zijn voor het project. De emoties anticipatie en vertrouwen zijn hier direct aan te linken.

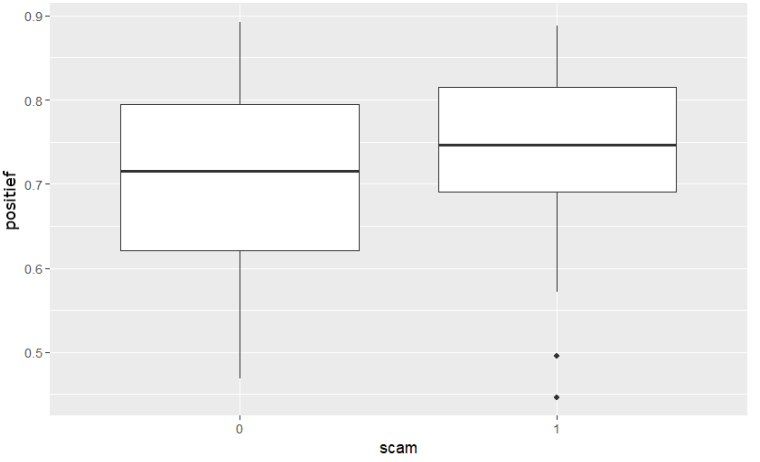
Als we iedere emotie apart bekijken is er bij verdriet wel een verschil merkbaar. Figuur 6 laat zien dat whitepapers van scams relatief minder verdrietig taalgebruik hebben. Meer specifiek, dat whitepapers van scams een kleinere relatieve hoeveelheid aan woorden hebben die gelinkt kunnen worden aan de emotie verdriet. Een mogelijke verklaring zou kunnen zijn dat oplichters serieuzer willen overkomen en dus minder van deze woorden gebruiken.

Vervolgens wordt er gekeken naar de toon van de paper, die volgens de literatuur een impact heeft op het succes van de ICO [8]. Meer specifiek het relatief aantal positieve woorden in de whitepaper zoals berekend in sectie 3.3. De toon wordt afgebeeld door de boxplot in figuur 7. Uit tabel 1 valt ook af te leiden dat er voor goede ICO's 71% van de woorden, die geassocieerd kunnen worden bij een toon, positieve woorden zijn. Voor scams is dit 74% wat op zich wel opmerkelijk is. Dat kan te maken hebben met het feit dat oplichters vaak een positievere toon aannemen om mensen te overtuigen. Ook kan er geconcludeerd worden dat whitepapers over het algemeen een positieve toon hebben.

Een analyse van de afstandsvariabelen levert de twee dichtheidsgrafieken in figuur 8 op. De linkse grafiek toont aan dat het gros van de whitepapers van legitieme ICO's (rood op de grafiek) links ligt. Hoe dichter bij nul, hoe meer

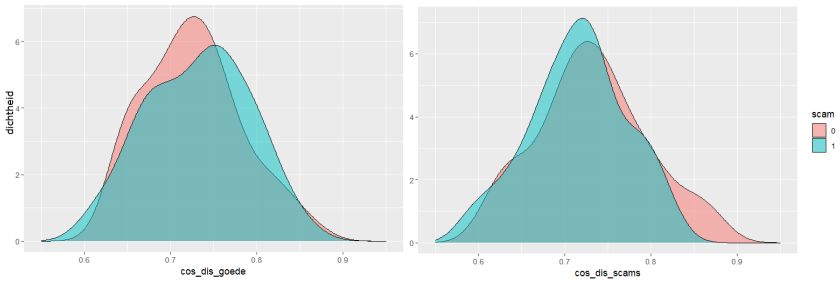


Figuur 6: Percentage aantal woorden die gelinkt worden aan de emotie "verdriet"



Figuur 7: Percentage positieve woorden

deze whitepapers lijken op de whitepapers van legitieme ICO's. Ter herinnering, voor een legitieme ICO wordt de waarde van 'Cos_dis_goede' berekend door het gemiddelde van de cosinusafstanden te nemen tussen de whitepaper en de 55 whitepapers van de andere legitieme ICO's. De whitepapers worden dus niet met zichzelf vergeleken. Hetzelfde zien we op de rechter grafiek. De whitepapers van ICO scams lijken meer op andere whitepapers van ICO scams, meer dan de whitepapers van legitieme ICO's.



Figuur 8: Gelijkenis met ICO scams en niet-scams

5 Resultaten

Het doel van het onderzoek is om significante verschillen bloot te leggen. Dit is gedaan aan de hand van een aantal methoden, die reeds toegelicht zijn in sectie 3. In deze sectie zullen we de resultaten van deze methoden bespreken beginnende bij de statistische testen.

5.1 Statistische testen

Van de 24 potentieel significante variabelen zijn er 17 continue en 7 categorische. Uitkomsten voor de t-toetsen voor de continue variabelen zijn weergegeven in tabel 2. Hieruit kunnen we afleiden dat het percentage aan woorden in de whitepaper die geassocieerd kunnen worden aan verdriet statistisch significant is op het 1% niveau. Het aantal woorden, de grootte van het projectteam en het aantal woorden per pagina zijn statistisch significant op het 5% niveau. Al deze variabelen hebben een grote positieve t-waarde, wat in dit geval wil zeggen dat de gemiddeldes van deze waarden voor niet-scams significant hoger zijn dan voor scams. De whitepaper van een legitieme ICO bestaat bijvoorbeeld gemiddeld uit meer pagina's en woorden. Legitieme ICO's zullen hun whitepaper gemiddeld in meerdere talen aanbieden en meer team leden bij naam benoemen.

Een ander geven dat opvalt is dat de variabelen 'cos_dis_goede' en 'cos_dis_scams' niet statistisch significant zijn. Men kan dus niet met zekerheid zeggen dat er een verband is met de cosinusafstand tussen een whitepaper met die van frauduleuze ICO's en legitieme ICO's. Op zich is dit niet geheel opmerkelijk aangezien de gekende scams die onderzocht werden uiteraard ook "succesvol" waren. Ze sloegen erin om mensen te overtuigen van hun project anders werd er nooit geld ingezameld en was de scam nooit geregistreerd geweest. Tot slot wijst deze test ook uit dat het aantal pagina's en de variabele 'vreugde' significant zijn op het 10% niveau.

Een andere zaak die af te lezen valt is dat 'Negatief' en 'Positief' dezelfde absolute t-waarde (en p-waarde) hebben. Perfect logisch aangezien die twee variabelen samen 100% vormen en er dus een perfecte samenhang is tussen de 2. Dus als het ene significant is moet het andere ook significant zijn. Volgens

Tabel 2: Resultaten T-test

	variabelen	t-waarde	p-waarde
1	aantal_pagina	1,811	0,073*
2	aantal_talen	2,290	0,024**
3	aantal_woorden	2,443	0,016**
4	afkeer	0,231	0,818
5	angst	0,415	0,679
6	anticipatie	0,121	0,904
7	cos_dis_goede	- 0,557	0,579
8	cos_dis_scams	1,457	0,148
9	negatief	1,522	0,131
10	positief	- 1,522	0,131
11	team	2,543	0,013**
12	verdriet	2,941	0,004***
13	verrast	0,782	0,436
14	vertrouwen	- 1,127	0,262
15	vreugde	- 1,914	0,058*
16	woede	0,725	0,470
17	woorden_per_pagina	2,314	0,023**

Info: *p<0,1; **p<0,05; ***p<0,01

die redenering zal er dus ook een samenhang zijn tussen de 8 emotie-variabelen al is dit niet direct af te leiden uit te tabel.

De categorische variabelen zijn getoetst aan de hand van de Chi-kwadraattoets en de bijhorende effectgrootte, berekend volgens de Cramer's V methode. De resultaten hiervan zijn terug te vinden in tabel 3

Tabel 3: Resultaten Chi-kwadraattest en Cramer's V

	variabelen	χ^2	p-waarde	V
1	artikelvorm	10,469	0,001***	0,306
2	legal_disclaimer	0,042	0,836	0,02
3	litepaper	0,538	0,463	0,069
4	roadmap	0,907	0,341	0,09
5	slide_deck	1,087	0,297	0,099
6	technische_uitleg	24,174	8,8e-07***	0,465
7	usecases	21,561	3,4e-06***	0,439

Info: *p<0,1; **p<0,05; ***p<0,01

Uit die tabel is af te leiden dat de variabelen 'artikelvorm', 'technische_uitleg' en 'usecases' significant zijn met een p waarde kleiner dan 0,001. Cramer's V is bij ze alle drie ook hoger dan 0,3 wat wordt gezien als een gemiddelde samenhang met de scam variabele. Technische_uitleg en usecases hebben een Cramer's V van tussen de 0,4 en 0,5 wat een vrij sterke samenhang wilt zeggen. Wat opvalt is dat het hebben van een juridische disclaimer totaal niet significant is.

5.2 Feature selectie

Zoals reeds vermeld werd in subsectie 3.6 zal er, voordat er een logistisch regressiemodel opgesteld wordt, eerst aan feature selectie gedaan worden. De resultaten van de statistische toetsen zijn een zeer goede basis om mee te beginnen. In dit onderzoek wordt er echter ook geopteerd om 2 algoritmen toe te passen. ExtraTreesClassifier voor de categorische variabelen en Boruta voor de continue. Volgens ExtraTreesClassifier is 'usecases' de belangrijkste variabele gevolgd door 'technische_uitleg' en 'artikelvorm'. Dat komt perfect overeen met de variabelen die volgens de statistische toets significant bleken te zijn.

Volgens Boruta zijn de volgende variabelen belangrijk: 'aantal_talen', 'aantal_woorden', 'aantal_pagina', 'team en verdriet, in de volgorde van belangrijk naar minder belangrijk. Opvallend is dat zowel 'vreugde' als het aantal woorden per pagina niet geselecteerd werden door dit algoritme. Dat terwijl er wel degelijk een significant verschil is tussen het aantal woorden per pagina voor whitepapers van scams en van niet-scams. Niet geheel verrassend want de t-waarde zegt weinig over de voorspelbare kracht van een variabele. De gevisualiseerde output van de algoritmes zijn terug te vinden als bijlage A1 en A2.

5.3 Logistisch regressiemodel

Om te kunnen antwoorden op de vraag in welke mate je een frauduleuze ICO kan onderscheiden van een niet frauduleuze ICO op basis van zijn whitepaper zijn er drie logistische regressiemodellen opgesteld. Deze meten de relatie tussen de afhankelijke variabele scam en een aantal onafhankelijke variabelen. De uitkomsten voor de drie modellen zijn af te lezen uit tabel 4.

Het eerste model is opgesteld door alle significante variabelen op te nemen. Dit resulteert in een model waarbij er vier van de 10 onafhankelijke variabelen statistisch significant zijn. Van die vier zijn verdriet, usecases en technische_uitleg statistisch significant op het 5% niveau. Het aantal talen is statistisch significant op het 10% niveau. Al deze variabelen hebben een negatieve beta. Voor 'aantal_talen' wil dat bijvoorbeeld zeggen dat in hoe meer talen de whitepaper aangeboden wordt, hoe kleiner de kans dat het om een scam gaat. Deze regressie toont aan dat hoe groter de relatieve hoeveelheid aan woorden die gelinkt kunnen worden aan de emotie 'verdriet' hoe kleiner de kans dat het om een scam gaat. Ook is de aanwezigheid van usecases en een technische uitleg een significante indicator dat het gaat om een legitieme ICO.

Het tweede model is samengesteld op basis van de feature selectie algoritmen. Op zich verschilt dit model niet veel met het 1ste model. Wel is het aantal talen niet langer statistisch significant. Wanneer er enkel de statistisch significante variabelen gebruikt worden om een logistisch regressiemodel op te stellen resulteert dat in model 3.

Tabel 4: Resultaten logistische regressie

	Afhankelijke variabele:		
	scam		
	(1)	(2)	(3)
aantal_talen	-0,362*	-0,307	
	(0,210)	(0,195)	
aantal_woorden	-0,0001	0,00001	
	(0,0001)	(0,0001)	
aantal_pagina	0,022		
	(0,035)		
team	-0,022	-0,023	
	(0,037)	(0,036)	
verdriet	-24,998**	-18,145*	-14,964*
	(11,205)	(9,435)	(8,361)
vreugde	-10,217		
	(7,705)		
woorden_per_pagina	0,00002	-0,002	
	(0,005)	(0,003)	
usecases	-1,252**	-1,039**	-1,210**
	(0,559)	(0,527)	(0,507)
technische_uitleg	-1,458**	-1,264**	-1,386***
	(0,624)	(0,588)	(0,484)
artikelvorm	0,259	0,164	
	(0,669)	(0,660)	
Constant	4,727**	3,298***	2,276***
	(2,056)	(0,917)	(0,579)
Observaties	112	112	112
Log Likelihood	-56,893	-58,043	-59,959
Akaike Inf. Crit.	135,786	134,086	127,918
Info:		*p<0,1; **p<0,05; ***p<0,01	

$$\begin{aligned} f(scam_i) =& 2,276 - 14,964 * Verdriet_i \\ & - 1,21 * Usecases_i \\ & - 1,386 * Technische\_uitleg_i \end{aligned} \tag{6}$$

De regressiemodellen kunnen weergegeven worden in een formule zoals uitgelegd in sectie 3.7. Dit is gedaan voor de resultaten van model 3 in formule 6. De constante is 2,276, met als onafhankelijke variabelen: 'Verdriet', 'Usecases' en

'Technische_uitleg' die respectievelijk de coëfficiënten -14,964, -1,21 en -1,386 hebben. Het hebben van een grondige beschrijving van de usecases zal de vergelijking dus doen dalen met 1,21. Aangezien 'Usecases' en 'Technische_uitleg' binaire variabelen zijn kunnen we hun coëfficiënten zo interpreteren.

Met behulp van de formule wordt de interpretatie gemakkelijker. De variabele verdriet heeft bijvoorbeeld een vrij grote coëfficiënt. Dat valt te verklaren door het feit dat de waarden van 'verdriet' zijn zeer klein zijn (gemiddeld 0,054) en dus een verandering van 0.10 een groot, negatief effect zal hebben.

6 Discussie

In het kader van deze studie werd op basis van 112 whitepapers van Initial Coin Offerings onderzocht of er significante verschillen te constateren zijn tussen whitepapers van frauduleuze en legitieme ICO's. Er werden zowel manueel als via *natural language processing* bepaalde kenmerken uit de whitepapers gehaald om deze vervolgens aan een regressie- en statistische analyse te onderwerpen. In deze sectie wordt er toelichting gegeven bij de bevindingen uit die analyses, hoe deze de onderzoeksvraag beantwoorden en waarom deze studie een meerwaarde kan betekenen. Tot slot wordt er besproken hoe dit onderzoek precies bijdraagt aan de literatuur en welke aanbevelingen er zijn voor toekomstig onderzoek.

Uit onze analyses bleek dat de factoren, die legitieme whitepapers van frauduleuze whitepapers onderscheiden, in lijn liggen met de factoren die een ICO al dan niet succesvol maken. Zo is bijvoorbeeld de gemiddelde omvang van de papers van ICO scams significant kleiner dan die van legitieme ICO's. Hetzelfde kan gezegd worden over de grootte van het projectteam. ICO scams hebben gemiddeld een kleiner projectteam dan niet-scams [5, 7, 27, 29]. Daarnaast is gebleken dat de aanwezigheid van een technische uitleg (of een link naar GitHub), net zoals de aanwezigheid van usecases, vaker het geval is bij whitepapers van legitieme ICO's. Ook worden deze whitepapers vaker in de vorm van een professioneel artikel aangeboden [27, 28]. Een andere bevinding is dat bij Legitieme ICO's de whitepapers gemiddeld in meer talen beschikbaar zijn dan bij frauduleuze ICO's.

Vervolgens is aan de oppervlakte gekomen dat legitieme ICO's hun whitepaper vaker als een formeel artikel aanbieden. Dit bevestigt de hypothese die gebaseerd is op het onderzoek over succesfactoren [4, 31]. Een kritische bemerking hierbij is dat deze data manueel, en dus op basis van menselijk oordeel, werd verzameld. Er is hiervoor gebruik gemaakt van officiële standaarden voor wetenschappelijke artikels zoals APA, springer, elsavier ... . Bij de verzameling van de variabele werd er niet vermeld of het al dan niet om een scam ging om bias te vermijden. Dat is dezelfde methode die gehanteerd is bij een aantal studies naar de succesfactoren bij crowdfunding [32, 33]. Toch blijft het een subjectieve methode aangezien het rust op menselijk oordeel. Voor toekomstig onderzoek kan het interessant zijn om met een tekstverwerkend algoritme te werken.

De hypothese omtrent de sentiment analyse werd origineel opgesteld op basis van het onderzoek van Zhang et al. (2021) [8]. Dit onderzoek vond geen significante samenhang tussen de toon van een whitepaper en het al dan niet zijn van een scam. Wel zijn er een aantal andere interessante resultaten uit onze sentimentanalyse voortgekomen. Zo maakte de exploratieve data-analyse duidelijk dat whitepapers, van zowel frauduleuze als legitieme ICO's, voornamelijk anticipatie en vertrouwen uitstralen. Wat ook geconcludeerd kan worden is dat whitepapers van scams een kleinere, relatieve hoeveelheid aan woorden gebruiken die gelinkt kunnen worden aan de emotie verdriet. Dat kan mogelijks te maken hebben met het feit dat oplichters vaak een serieuzere toon aannemen om mensen te overtuigen.

Daarnaast zijn er binnen deze studie whitepapers met elkaar vergeleken aan de hand van de cosinusafstand. Vertrekkende vanuit de onderstelling dat whitepapers van ICO scams gelijkaardig zijn, is onderzocht of de gemiddelde waargenomen cosinusafstand tussen een whitepaper en de whitepapers van ICO scams, verschilt voor scams en niet-scams. Hoewel er inderdaad een klein verschil merkbaar is tussen frauduleuze en legitieme whitepapers is dit verschil niet statistisch significant. Dit is niet vreemd aangezien de scams die onderzocht werden uiteraard ook 'succesvol' waren. Ze sloegen er namelijk in om mensen te overtuigen van hun project, anders werd er nooit geld ingezameld en zou de scam nooit geregistreerd zijn.

Met deze studie is getracht een zekere fundering te leggen voor toekomstig onderzoek. Het onderzoek dat reeds bestaat over zowel ICO's als whitepapers van ICO's gaat voornamelijk over de succesfactoren. Er werd één artikel gevonden dat een deel van zijn onderzoek wijde aan frauduleuze ICO's [4]. Met dit onderzoek is er geprobeerd dit gat te dichten en een zekere fundering te leggen voor toekomstig onderzoek. Maar ook voor potentiële investeerders kunnen de bevindingen van deze studie interessant zijn. Het onderzoek heeft namelijk aspecten uitgelicht waar investeerders op kunnen letten bij het lezen van een whitepaper, zoals de aanwezigheid van usecases, een technische uitleg en het aantal talen waarin de whitepaper wordt aangeboden. De bevinding over de emotie 'verdriet' is mogelijks interessant voor potentiële investeerders maar kan wel inspirerend zijn in het licht van toekomstig onderzoek.

Tot slot zijn er nog enkele kanttekeningen te maken. Het fenomeen van ICO's is een relatief nieuw concept. Dat brengt met zich mee dat de hoeveelheid van papers die peer-reviewed zijn eerder klein is. Een deel van de literatuurstudie is dan ook uit noodzaak gedaan op basis van informatieve websites. Daarnaast zorgde de beperkte hoeveelheid data er voor dat meer geavanceerde machine learning methoden niet in aanmerking komen. Bij een grotere hoeveelheid data zouden deze algoritmen een correctere voorspelling kunnen maken.

7 Conclusie

Door grote hoeveelheid aan fraude en asymmetrische informatie op de ICO-markt is de whitepaper een onmisbaar onderdeel van ICO's geworden. In deze studie werden de whitepapers van frauduleuze en legitieme ICO's geanalyseerd en vergeleken. Het doel was dan ook om significante verschillen tussen beiden te ontdekken opdat frauduleuze ICO's in de toekomst onderscheiden kunnen worden van legitieme ICO's op basis van hun whitepapers. Door middel van het toepassen van verschillende methoden is het onderzoek erin geslaagd enkele besluiten te formuleren.

Whitepapers van ICO scams en legitieme ICO's hebben een vrij gelijkaardige toon. Desalniettemin gebruiken de whitepapers van frauduleuze ICO's relatief minder woorden die de emotie 'verdriet' oproepen. Daarnaast zijn er nog een aantal andere kenmerken die de whitepaper van een frauduleuze ICO van een legitieme ICO onderscheiden. Zo zijn de whitepapers van legitieme ICO's vaker voorzien van een technische uitleg en uitgebreide beschrijving van hun usecases. Ook komen ze vaker voor in de vorm van een professioneel artikel. Verder bevatten ze meer pagina's, meer woorden per pagina en is de totale omvang woorden groter. Ook het aantal talen waarin legitieme ICO's hun whitepaper aanbieden is aanzienlijk hoger. De analyse van de cosinusafstanden laat zien dat de gelijkenis tussen de whitepaper van een legitieme of frauduleuze ICO met andere scams niet significant verschilt. Tot slot zijn er in deze studie drie logistische regressiemodellen opgesteld die aan de hand van een whitepaper kunnen voorspellen of het al dan niet om een scam zou gaan. Zo heeft de aanwezigheid van een technische uitleg en usecases een positief effect op de kans dat het om een legitieme ICO gaat. Ook werd aangetoond dat het relatief aantal woorden dat gelinkt kan worden met de emotie verdriet een impact heeft. Namelijk hoe hoger het percentage, hoe kleiner de kans dat het hier om een scam gaat.

Referenties

[1] Biggs, et al.: Blockchain: Revolutionizing the global supply chain by building trust and transparency. Rutgers University: Camden, NJ, USA (2017)

[2] Dowlat, S.: Cryptoasset market coverage initiation: Network creation, 30 (2018)

[3] Yadav, M.: Exploring signals for investing in an initial coin offering (ico). Available at SSRN 3037106 (2017)

[4] Florysiak, D., Schandlbauer, A.: The information content of ico white papers (2018). https://doi.org/10.2139/ssrn.3265007

[5] Karimov, B., Wójcik, P.: Identification of scams in initial coin offerings with machine learning. Frontiers in Artificial Intelligence **4** (2021)

[6] Bellavitis, et al.: A comprehensive review of the global development of initial coin offerings (ICOs) and their regulation **15**, 00213 (2021). https://doi.org/10.1016/j.jbvi.2020.e00213

[7] Ahmad, et al.: What determines initial coin offering success: a cross-country study **0**(0), 1–24 (2021). https://doi.org/10.1080/10438599.2021.1982712. Accessed 2022-05-05

[8] Zhang, et al.: Positive tone and initial coin offering. Accounting & Finance (2021)

[9] Rauchs, M., Glidden, A., Gordon, B., Pieters, G.C., Recanatini, M., Rostand, F., Vagneur, K., Zhang, B.Z.: Distributed ledger technology systems: A conceptual framework. verkrijgbaar op SSRN 3230013 (2018)

[10] Shafagh, H., Burkhalter, L., Hithnawi, A., Duquennoy, S.: Towards blockchain-based auditable storage and sharing of iot data. In: Proceedings of the 2017 on Cloud Computing Security Workshop, pp. 45–50 (2017)

[11] Li, J., Mann, W.: Initial coin offering and platform building. SSRN Electronic Journal, 1–56 (2018)

[12] ICO Basics - the Difference Between Security Tokens and Utility Tokens. https://www.cointelligence.com/content/ico-basics-security-tokens-vs-utility-tokens/ Accessed 2022-03-28

[13] PWC: ICO / STO Report. Technical Report 5, PricewaterhouseCoopers (2019). https://www.pwc.ch/en/insights/fs/5th-ico-sto-report.html

[14] Sunyaev, A., Kannengießer, N., Beck, R., Treiblmaier, H., Lacity, M., Kranz, J., Fridgen, G., Spankowski, U., Luckow, A.: Token economy. Business & Information Systems Engineering **63**(4), 457–478 (2021)

[15] Wang, Q., Li, R., Wang, Q., Chen, S.: Non-fungible token (nft): Overview, evaluation, opportunities and challenges. arXiv preprint arXiv:2105.07447 (2021)

[16] Rachita: Securities Tokens Vs Utility Tokens Vs NFTs - How Different Are They? https://www.solulab.com/securities-tokens-vs-utility-tokens-vs-nfts-how-different-are-they/ Accessed 2022-03-28

[17] Pemberton, J.E.: British Official Publications (Second Edition), Second edition edn., pp. 57–73 (1973)

[18] Graham, G.: White Papers For Dummies. Wiley, Hoboken, NJ, Verenigde Staten (2013)

[19] Ofir, M., Sadeh, I.: Ico vs. ipo: Empirical findings, information asymmetry, and the appropriate regulatory framework. Vand. J. Transnat'l L. **53**, 525 (2020)

[20] Vismara, S.: Equity retention and social network theory in equity crowdfunding. Small Business Economics **46**(4), 579–590 (2016)

[21] Mollick, E.: The dynamics of crowdfunding: An exploratory study. Journal of business venturing **29**(1), 1–16 (2014)

[22] Chuanjie, F., Koh, A., Griffin, P.: Automated theme search in ICO whitepapers **1**(4), 140–158. https://doi.org/10.3905/jfds.2019.1.011. Publisher: Institutional Investor Journals Umbrella. Accessed 2021-11-18

[23] Definition of Litepaper Applied to Blockchain / Crypto. https://www.meetbunch.com/terms/litepaper Accessed 2022-05-29

[24] Zetzsche, D.A., Buckley, R.P., Arner, D.W., Föhr, L.: The ico gold rush: It's a scam, it's a bubble, it's a super challenge for regulators. University of Luxembourg Law Working Paper (11), 17–83 (2017)

[25] Did You Fall for It? 13 ICO Scams that Fooled Thousands. https://cointelegraph.com/news/did-you-fall-for-it-13-ico-scams-that-fooled-thousands Accessed 2022-03-19

[26] Intelligenthq: ICO Scams: How to Identify and Avoid Them? - IntelligentHQ. Section: Alternative Finance. https://www.intelligenthq.com/

ico-scams-how-to-identify-and-avoid-them/ Accessed 2021-11-23

[27] Panin, A., Kemell, K.-K., Hara, V.: Initial coin offering (ico) as a fundraising strategy: a multiple case study on success factors. In: International Conference on Software Business, pp. 237–251 (2019). Springer

[28] Liu, Y., Sheng, J., Wang, W.: Technology and Cryptocurrency Valuation: Evidence from Machine Learning. https://doi.org/10.2139/ssrn.3577208. https://papers.ssrn.com/abstract=3577208 Accessed 2022-02-23

[29] Di Dio, D., Tam, N.T.: On leveraging deep learning models to predict the success of icos (2019)

[30] Dinu, L.P., Ionescu, R.-T.: A rank-based approach of cosine similarity with applications in automatic classification. In: 2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, pp. 260–264 (2012). https://doi.org/10.1109/SYNASC.2012.24

[31] Zhang, S., Aerts, W., Lu, L., Pan, H.: Readability of token whitepaper and ICO first-day return **180**, 58–61. https://doi.org/10.1016/j.econlet.2019.04.010. Accessed 2021-11-18

[32] Suppe, F.: The structure of a scientific paper. Philosophy of Science **65**(3), 381–405 (1998)

[33] Koch, J.-A., Siering, M.: Crowdfunding success factors: The characteristics of successfully funded projects on crowdfunding platforms (2015)

[34] ICOs Rated by Experts. https://icobench.com/ Accessed 2021-11-21

[35] Burns, L., Moro, A.: What makes an ico successful? an investigation of the role of ico characteristics, team quality and market sentiment. An Investigation of the Role of ICO Characteristics, Team Quality and Market Sentiment (September 27, 2018) (2018)

[36] Samieifar, S., Baur, D.G.: Read me if you can! an analysis of ICO white papers **38**, 101427. https://doi.org/10.1016/j.frl.2020.101427. Accessed 2021-11-18

[37] Chowdhary, K.R.: Natural Language Processing, pp. 603–649. Springer, New Delhi (2020). https://doi.org/10.1007/978-81-322-3972-7_19. https://doi.org/10.1007/978-81-322-3972-7$_1$9

[38] Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM **56**(4), 82–89 (2013). https://doi.org/10.1145/2436256.2436274

[39] Mohammad, S., Turney, P.: Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26–34. Association for Computational Linguistics, Los Angeles, CA (2010). https://aclanthology.org/W10-0204

[40] Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence **29**(3), 436–465 (2013)

[41] Plutchik, R.: The Emotions. Amsterdam University Press, Amsterdam, Nederland (1991)

[42] Getting to know your data. In: Han, J., Kamber, M., Pei, J. (eds.) Data Mining (Third Edition), Third edition edn. The Morgan Kaufmann Series in Data Management Systems, pp. 39–82. Morgan Kaufmann, Boston (2012)

[43] Li, B., Han, L.: Distance weighted cosine similarity measure for text classification. In: International Conference on Intelligent Data Engineering and Automated Learning, pp. 611–618 (2013). Springer

[44] Joshi, A., Bhattacharyya, P., Carman, M.: Political issue extraction model: A novel hierarchical topic model that uses tweets by political and non-political authors. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 82–90 (2016)

[45] Bevans, R.: An Introduction to T-tests. https://www.scribbr.com/statistics/t-test/ Accessed 2022-05-30

[46] Tallarida, R.J., Murray, R.B.: Chi-Square Test, pp. 140–142. Springer, New York, NY (1987). https://doi.org/10.1007/978-1-4612-4974-0_43. https://doi.org/10.1007/978-1-4612-4974-0$_4$3

[47] Wu, B., Zhang, L., Zhao, Y.: Feature selection via cramer's v-test discretization for remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing **52**(5), 2593–2606 (2013)

[48] Bartlett, M.S.: Contingency table interactions. Supplement to the Journal of the Royal Statistical Society **2**(2), 248–252 (1935)

[49] GeeksforGeeks: Extra Tree Classifier for Feature Selection. https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/ Accessed 2022-05-20

[50] Kursa, M.B., Rudnicki, W.R.: Feature selection with the boruta package.

Journal of statistical software **36**, 1–13 (2010)

[51] Sutha, K., Tamilselvi, J.J.: A review of feature selection algorithms for data mining techniques. International Journal on Computer Science and Engineering **7**(6), 63 (2015)

[52] Seber, G., Lee, A.: Linear Regression Analysis. John Wiley & Sons, Hoboken, NJ, Verenigde Staten (2003)

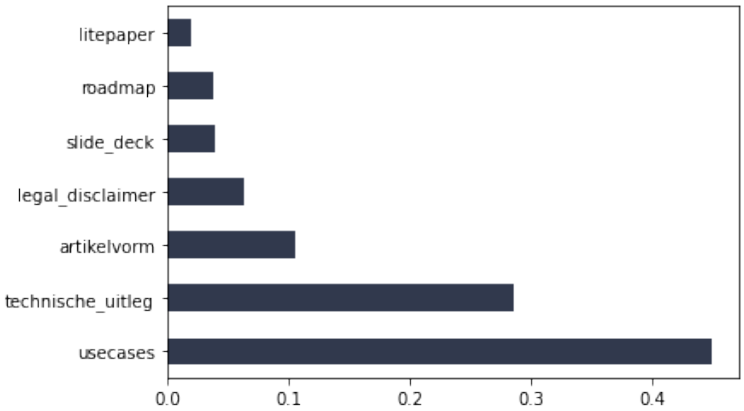
[53] Myung, I.J.: Tutorial on maximum likelihood estimation. Journal of mathematical Psychology **47**(1), 90–100 (2003)

Bijlage A Tabellen en figuren

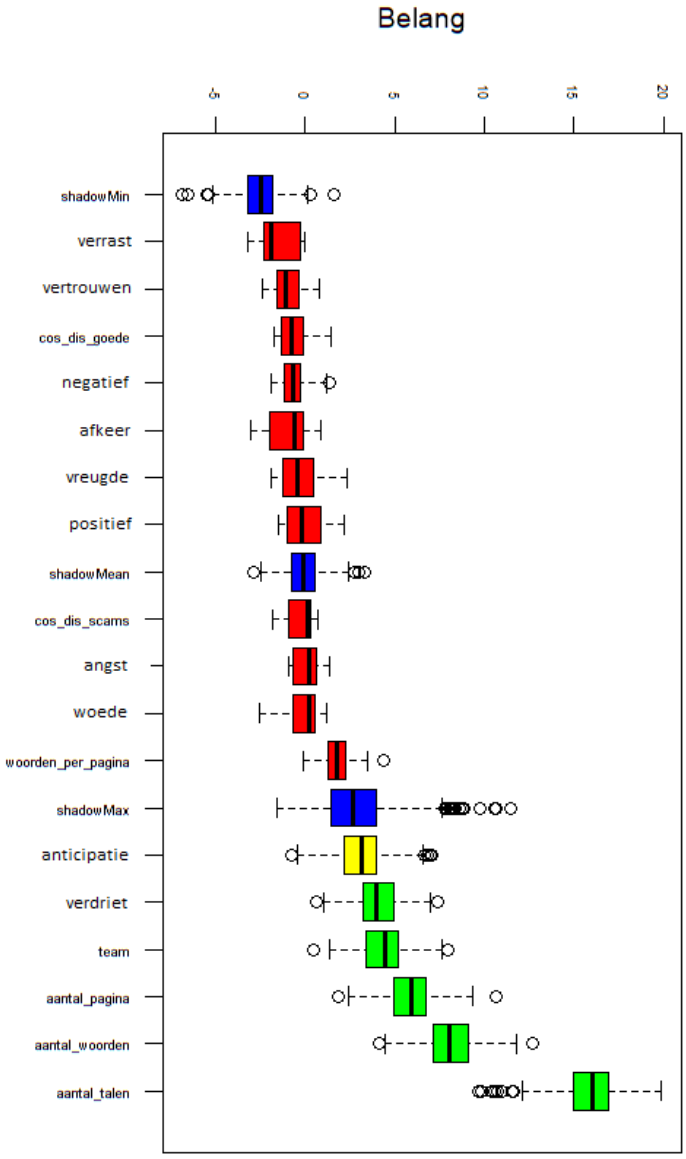
Tabel A1: Beschrijving van de variabelen

Variabele	Beschrijving
ICO	Naam van de ICO.
Positief	Een waarde tussen 0 en 1 die de positieve toon van de whitepaper aangeeft.
Negatief	Een waarde tussen 0 en 1 die de negatieve toon van de whitepaper aangeeft.
Woede	Een waarde tussen 0 en 1 die aangeeft in welke mate de whitepaper de emotie 'woede' oproept.
Angst	Een waarde tussen 0 en 1 die aangeeft in welke mate de whitepaper de emotie 'angst' oproept.
Anticipatie	Een waarde tussen 0 en 1 die aangeeft in welke mate de whitepaper de emotie 'anticipatie' oproept. 3.3
Vertrouwen	Een waarde tussen 0 en 1 die aangeeft in welke mate de whitepaper de emotie 'vertrouwen' oproept.
Verrast	Een waarde tussen 0 en 1 die aangeeft in welke mate de whitepaper de emotie 'verrast' oproept.
Verdriet	Een waarde tussen 0 en 1 die aangeeft in welke mate de whitepaper de emotie 'verdriet' oproept.
Vreugde	Een waarde tussen 0 en 1 die aangeeft in welke mate de whitepaper de emotie 'vreugde' oproept.
Afkeer	Een waarde tussen 0 en 1 die aangeeft in welke mate de whitepaper de emotie 'afkeer' oproept.
Aantal_woorden	Een numerieke variabele die het aantal woorden in een whitepaper aangeeft.
Aantal_talen	Een numerieke variabele die beschrijft in hoeveel talen de whitepaper beschikbaar was op het moment van de ICO.
Roadmap	Een binaire variabele die de waarde één heeft indien er een roadmap aanwezig is. Een roadmap is het stappenplan dat het projectteam zal volgen en vaak data bevat van de lancering van het project.
Team	Een numerieke variabele die het aantal teamleden weergeeft die met naam genoemd worden in de whitepaper of op de website.
Aantal_pagina	Een numerieke variabele die het aantal pagina's van de whitepaper weergeeft.

Variabele	Beschrijving
Artikelvorm	Een binaire variabele die aangeeft of de whitepaper in de vorm van een professioneel artikel komt volgens een officïele standaard (APA, MLA, Chicago,...).
Technische_uitleg	Een binaire variabele die aangeeft of er een technische uitleg gegeven wordt in de whitepaper, of een link naar GitHub met de technische informatie.
Litepaper	Een binaire variabele die weergeeft of de ICO enkel een Litepaper ter beschikking stelt. Een Litepaper is simpelweg een verkorte versie van de whitepaper en worden expliciet litepaper genoemd door het ontwikkelingsteam.
Slide_deck	Een binaire variabele die aangeeft of de whitepaper een slide deck is.
Usecases	Een binaire variabele die aangeeft of de whitepaper usecases beschrijft. Dit is een expliciet onderdeel van een whitepaper waarmee bewezen wordt dat er na de ICO een gebruik is voor de tokens.
Legal_disclaimer	Een binaire variabele die aangeeft of de whitepaper een juridische disclaimer bevat.
Scam	Een binaire variabele die aangeeft of het om een frauduleuze ICO gaat.
Woorden_per_pagina	Een numerieke variabele die aangeeft hoeveel pagina's de whitepapar bevat
Cos_dis_goede	Een numerieke variabale die aangeeft in welke mate de whitepaper van de ICO lijkt op de whitepapers van legitieme ICO's. Dit is berekend op basis van de cosinus afstand tussen de 2 variabelen op een schaal van 0 tot 1. (1 = absoluut geen gelijkenis en 0 = Perfect gelijk)
Cos_dis_scams	Een numerieke variabale die aangeeft in welke mate de whitepaper van de ICO lijkt op de whitepapers van frauduleuze ICO's. (zie Cos_dis_goede)



Figuur A1: De belangrijkste variabelen volgens ExtraTreesClassifier



Figuur A2: De belangrijkste variabelen volgens Boruta