Wat is het meest geschikte AI model voor het forecasten van de koers van cryptomunten aan de hand van open source data

INTERNE PROMOTOR: WOUTER GEVAERT
EXTERNE PROMOTOR: SEBASTIEN PEREZ

ONDERZOEKSVRAAG UITGEVOERD DOOR

STUDENT JOREN VANGOETHEM

VOOR HET BEHALEN VAN DE GRAAD VAN BACHELOR IN DE

MULTIMEDIA & CREATIVE TECHNOLOGIES

HOWEST | 2021-2022

Woord vooraf

Deze bachelorproef sluit aan op mijn research project van in het vorige semester waar ik onderzocht welke neurale netwerk modellen het best geschikt waren om te gebruiken op de crypto markt voor crypto trading.
Voor iets meer dan 3 weken lang heb ik geprobeerd het beste model te vinden en welke data hier voor nodig was om een zo goed mogelijk 'trading model' te maken dat uiteindelijk ook effectief ingezet kan worden op de crypto markt. Ik heb hierbij samengewerkt met Andreas Maerten en achteraf vergeleken we ons resultaat met andere manieren van crypto trading, zowel manueel als algoritmisch traden en ons neurale netwerk had voor en nadelen vergeleken met beide maar hier ga ik later verder op in.
Ik zal het onderzoek, de technische details en de resultaten zo goed mogelijk proberen bespreken in deze bachelorproef. Alsook advies geven voor iemand dat een gelijkaardig onderzoek zou willen uitvoeren of verder bouwen op mijn onderzoek.

Ik zou graag Wouter Gevaert en Marie Dewitte willen bedanken voor hun hulp tijdens mijn onderzoek.

Abstract

Mijn onderzoeksvraag “wat is het meest geschikte model voor het voorspellen van de crypto koers aan de hand van open source data?” heb ik gekozen vanwege mijn eigen interesse naar crypto en ik wou hier graag eens wat dieper op ingaan of het mogelijk was voor een neuraal netwerk om patronen of correlaties te zien in deze data en correcte beslissing te maken.

Mijn onderzoek ging vooral in op welke types van neurale netwerken en welke data ik hiervoor nodig zou hebben. Het werd al snel duidelijk dat LSTM’s de enige goede optie waren voor het voorspellen van time series data. LSTM’s kunnen vanwege hun long term memory (in tegenstelling tot een GRU netwerk dat geen long term memory heeft) betere voorspellingen maken omdat ze rekening kunnen houden met wat er net gebeurd is. Je kan niet echt voorspellen welke richting een crypto munt op zal gaan (stijgen of dalen in waarde) door enkel de laatste prijswaarden te bekijken.
Dus het model en vooral ook nog de opmaak van het model zoals de aantal layers en aantal neurons per layer waren heel belangrijk om een snel maar accuraat model te verkrijgen. Het grote nadeel bij LSTM modellen is dat training heel lang duurt vanwege de grote hoeveelheid berekeningen vergeleken met andere typen modellen. Maar gelukkig was ons model niet extreem groot en viel dit nog redelijk goed mee.

Het andere belangrijke element was natuurlijk de training en test data. De data werd opgehaald met de publieke API van Binance, een van de grootste crypto exchanges ter wereld.
De candle data was echter niet genoeg om een goed model te bekomen dus ik zal nog veel dieper ingaan op wat we allemaal gedaan hebben om onze data zo goed mogelijk te krijgen en de impact hiervan op ons model.

Het leverde een mooi resultaat op maar er is zeker ruimte voor verbetering indien er meer tijd in training en design van het model gestoken kan worden.

Inhoudsopgave

Figurenlijst

Table of Figures

Lijst met afkortingen

ADOSC	Accumulation/Distribution Oscillator
ATR	Average True Range
EMA	Exponential Moving Average
GRU	Gated Recurrent Unit
LSTM	Long Short Term Memory
MACD	Moving Average Convergence Divergence
MFI	Money Flow Index
RNN	Recurrent Neural Network
RSI	Relative Strength Index

Verklarende woordenlijst

bearish	een dalende trend in de prijs van een asset
layer	een laag van neurons in een neuraal netwerk
LSTM	een neural netwerk type waarbij er een long-term memory door alle layers heen gaat
bullish	een stijgende trend in de prijs van een asset
trend reversal	een switch tussen down en up trend
candle	een weergave van de Low, High, Open en Close prijs van een bepaalde tijdsperiode

1 Inleiding

Traden van crypto currencies is nog relatief nieuw en populair onder de jeugd vergeleken met aandelen verhandelen wat vooral bij de iets oudere generaties populair is. Het leek mij een leuk idee om hierop in te spelen op iets wat bij de jeugd populair is. Het is ook iets waar ik zelf actief mee bezig ben.

1.1 aanleiding en inspiratie

hier zal ik wat dieper ingaan op de reden dak ik de onderzoeksvraag “Wat is het meest geschikte AI model voor het forecasten van de koers van cryptomunten aan de hand van open source data?” heb gekozen.

Het idee van een neuraal netwerk dat kan voorspellen wanneer je best kon kopen en verkopen op de crypto markt of op de aandelen markt, leek enorm interessant. Momenteel worden de meeste trades op de crypto en aandelen markt algoritmisch uitgevoerd. Dit wil dus zeggen dat men niet meer manueel kijkt naar de prijzen en hoe goed of slecht bedrijven het doen. Wat men echter wel doet is deze data aan een bepaald algoritme geven en deze zal dan een buy, sell of hold target teruggeven. Het is niet duidelijk hoeveel trades algoritmisch gebeuren maar afhankelijk van de bronnen die je online kan raadplegen ligt dit tussen 60% en 80%. Als algoritmisch traden zo goed werkt waarom dan niet met neurale netwerken, uiteindelijk is dit ook maar een reeks van berekeningen op de input data die dan een buy, sell of hold target predict.

Dit onderzoek lijkt ook technisch zeer interessant om te zien hoe goed neurale netwerken time series data kunnen voorspellen in een anders zeer onvoorspelbare omgeving waar zelfs de meeste mensen vaak fouten maken.

1.2 Deelvragen

Dit zijn de deelvragen die in dit onderzoek ook beantwoord zullen worden.

– is het mogelijk met enkel prijs en volume data een voorspelling te doen van de crypto markt?

– Welk type model heeft het beste resultaat, is het mogelijk reinforcement learning te gebruiken?

– Is het resultaat beter als we trainen per coin in plaats van 1 model voor alles?

1.3 Keuzes

Er zijn enkele redenen dat we dit onderzoek niet op de aandelenmarkt maar crypto markt doen. Bijvoorbeeld de crypto markt is 24/7 actief in tegenstelling tot de vaste uren van de aandelenmarkt die ook enkel op weekdagen open is. De substantieel lagere commissies op crypto currencies zijn ook voordelig en afhankelijk van welke exchange je gebruikt liggen de fees bij crypto meestal tussen 0% en 1% voor aankoop en verkoop. Bij aandelen loopt dit al heel snel hoog op omdat je vaak nog eens een maandelijkse kost hebt voor je markt data (de prijs data die je opvraagt via de API), een ‘maintenance fee’ voor het behouden van je account, commissie op je trades, etc.…

Het wordt al snel duidelijk waarom crypto de betere keuze is hiervoor. Een ander voordeel dat zeker bij het trainen van een neuraal netwerk belangrijk is is dat de data bij crypto exchanges gratis en makkelijk op te vragen is via de API. Voor ons onderzoek hebben we data gebruikt van de Binance Exchange.

1.4 Doelen

Het voornaamste doel van dit onderzoek is een werkend model maken dat goed genoeg kan beslissen wanneer te kopen en te verkopen om te onderzoeken of het wel degelijk mogelijk is om neurale netwerken te gebruiken op time series data in een zeer wisselvallige en onvoorspelbare omgeving. Ook belangrijk is welke data er precies relevant is om tot deze beslissing te komen, is enkel candle data voldoende of zullen we meer nodig hebben zoals indicatoren? Dit zal nog tot in detail onderzocht en getest worden.

Hopelijk kan het model uiteindelijk goed genoeg voorspellen wanneer te kopen en te verkopen dat het kan ingezet worden op de echte crypto markt.

Een ander doel is natuurlijk zeker ook dat het model goed genoeg is om later effectief in te zetten op de crypto markt en zo winstgevende trades te maken.

2 Research

Voor dat we begonnen met het testen of programmeren was het belangrijk om info te verzamelen over de beschikbare exchanges, een platform waarop je crypto munten kan verhandelen, alsook of deze wel een api beschikbaar hadden om voldoende data op te halen voor training.

We hebben hierbij voor de Binance Exchange API gekozen omdat deze gratis toegankelijk is. Er bestaan reeds een aantal libraries voor allerlei programmeertalen om data op te vragen wat dit zeker vergemakkelijkte. Wij kozen voor de Python-binance library te vinden op Hithub.

2.1 data

Het eerste dat we nodig hebben vooraleer we kunnen beginnen met voorspellingen doen is de LOHC data, ook wel candles genoemd. Een candle bevat de Low, Open, High en Close prijs van een bepaalde periode. Deze worden meestal groen en rood weergegeven om aan te duiden welke kant deze opgaat, een groene candle is een candle waarbij de Close hoger ligt dan de Open en omgekeerd bij de rode candle. Ook haalden we per candle het volume op, dit is de hoeveelheid van deze crypto currency die verhandeld is tijdens deze periode. Het volume varieert echter wel van exchange tot exchange omdat ze enkel een zicht hebben op het volume dat binnen hun eigen exchange verhandeld wordt.

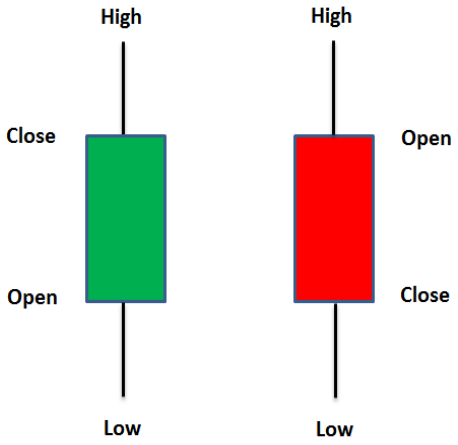


Figure 1: Candle

Voor het onderzoek maken we gebruik van 1 minuut candles, dit wil zeggen dat er tussen de open en de close exact 1 minuut ligt. Dit is de hoogste resolutie die je bij exchanges kan ophalen zodat we zoveel mogelijk data hebben en ook omdat ons model dan hopelijk gebruik kan maken van de kleine schommelingen in prijs op minuten.

We kunnen ook later nog van minuut data de candles comprimeren tot andere candles zoals bijvoorbeeld uurcandles, dit kan gemakkelijk gedaan worden door 60 candles te nemen. Van de eerste candle houd je de Open bij, van de laatste candle de Close en dan van de 60 candles de hoogste High en de laagste Low. Het volume kan je gewoon optellen en dan heb je een uurcandle. Daarom zijn minuutcandles de beste optie omdat je hiervan elke andere lengte kan afleiden.

Voor het ophalen van onze data maken we gebruik van de Binance Exchange API, deze is volledig gratis te gebruiken, je moet wel een account aanmaken op de site en API keys aanmaken onder ‘account management’.

We hebben 15 GB aan minuut candles van ongeveer 360 verschillende crypto currencies opgehaald en deze weggeschreven naar csv bestanden, deze worden later nog verwerkt voor normalisatie en data augmentation.

Het enige probleem nu is dat we geen targets hebben waarop we ons model kunnen trainen dus deze wordt achteraf toegevoegd door onze data door een c++ programma te runnen die dan targets gaat zetten voor buy, sell en hold. Dit programma is in de bijlagen te vinden en word later meer over verteld.

Een andere optie naast labelling was reinforcement learning maar dit had de training nog langer laten duren en dit was nu al een limiterende factor tijdens het onderzoek.

Deze data moet nu natuurlijk nog genormaliseerd worden en we gaan hier ook nog indicators aan toevoegen, er wordt later dieper ingegaan in wat indicators juist zijn en waarom we deze gebruiken. De normalisatie is redelijk simpel, we nemen voor elke candle het percentage verschil met de vorige candle, stel dat de vorige minuut de Close prijs op 100 stond en nu op 101 dan zal de genormaliseerde waarde 0.01 zijn.

Om het lezen, schrijven en totale hoeveelheid data wat in te perken schrijven we niet meer naar CSV maar naar binary files, dit bespaart ons een redelijke hoeveelheid aan opslag en maakt het lezen en schrijven sneller. Het nadeel is echter dat deze files niet meer te lezen zijn voor mensen maar dit is snel opgelost door het even in te laden met een scriptje en weer te geven in terminal voor controle van de data.

2.2 Model

Nu dat we de data hebben en een idee hebben van hoe deze data gestructureerd is kunnen we beginnen met het ontwerpen van een neuraal netwerk waar we onze voorspellingen mee willen uitvoeren.

Hiervoor is er eerst wat onderzoek gedaan naar wat anderen al geprobeerd hebben om voorspellingen te doen bij stock of crypto trading met time series data. [1][2]

2.2.1 LSTM Netwerken

Het werd hier al snel duidelijk dat LSTM netwerken een van de beste opties zijn. Maar wat exact een LSTM netwerk is zal ik hier even proberen verduidelijken.

De naam LSTM staat voor Long Short Term Memory. Dit is omdat LSTM netwerken een short term memory maar ook een long term memory hebben omdat er een flow van data door alle opeenvolgende neurons gaat waardoor oudere data ook een invloed heeft op de volgende neurons en de uiteindelijke output waarde.

Het voordeel van LSTM's bij time series data is dat er soms nuttige informatie in zowel iets oudere data en de nieuwste data zit en dat je beide nodig hebt voor een correcte voorspelling.

Afbeeldingen in dit hoofdstuk komen uit het artikel van referentie [5].

2.2.2 LSTM Gates

LSTM's gebruiken een serie van 'gates' die bepalen hoe de data verwerkt word. Namelijk de forget gate, input gate en output gate.

Hieronder vind u de legende voor de volgende afbeeldingen over de verschillende gates.

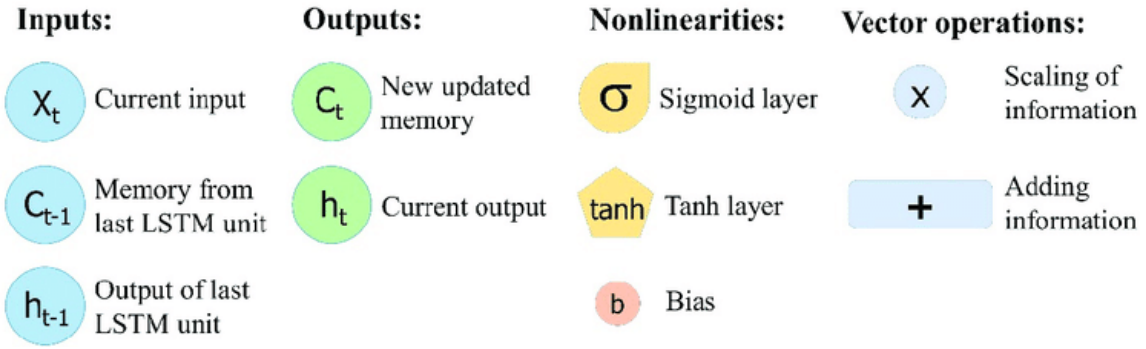


Figure 2: LSTM neuron legend

2.2.3 Forget Gate

Rechts ziet u een weergave van de forget gate.

Deze gate zal bepalen of de nieuwe input data relevant is afhankelijk van de cell state en hidden state door middel van een sigmoid activation. De cell state word doorgegeven van de vorige neuron en is dus het long term memory. De hidden state is de output van de vorige neuron. De input data is de nieuwe data die toegevoegd word als deze relevant genoeg is. Kortom de forget gate bepaald welke delen van het geheugen vergeten mogen worden.

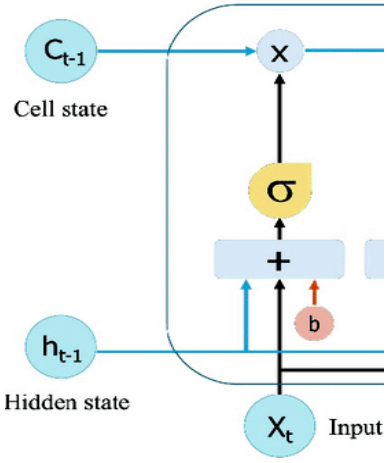


Figure 3: LSTM Forget Gate

2.2.4 Input Gate

De volgende gate word dan de input gate, deze zal bepalen welke gegevens van de input data toegevoegd zullen worden aan de cell state. De input in deze gate is hetzelfde als de input in de forget gate maar hier word er bepaald wat toegevoegd word en niet wat vergeten word.

Er worden hier 2 verschillende activatie functies gebruikt.

De Tanh functie zal de vorige hidden state combineren met de nieuwe input data om een memory update vector te maken. Deze vector bevat de informatie van de input data en hoeveel de cell state moet geupdate worden met deze data. Er word tanh gebruikt omdat de waardes hier tussen -1 en 1 liggen omdat je mogelijks ook de impact van nieuwe data wil verminderen.

De sigmoid activatie functie zal bepalen welke onderdelen van de nieuwe input data effectief relevant genoeg zijn om te onthouden. Het zou dus kunnen dat de tanh functie bepaalde onderdelen een hoge waarde geeft en dat er hier toch bepaald word dat deze minder impact moeten hebben. Hier word sigmoid gebruikt, deze waardes liggen dus tussen 0 en 1 waarbij 0 wil zeggen dat je de data niet wil updaten.

Deze vectoren worden pointwise multiplied met elkaar en deze output vector word dan toegevoegd aan de cell state.

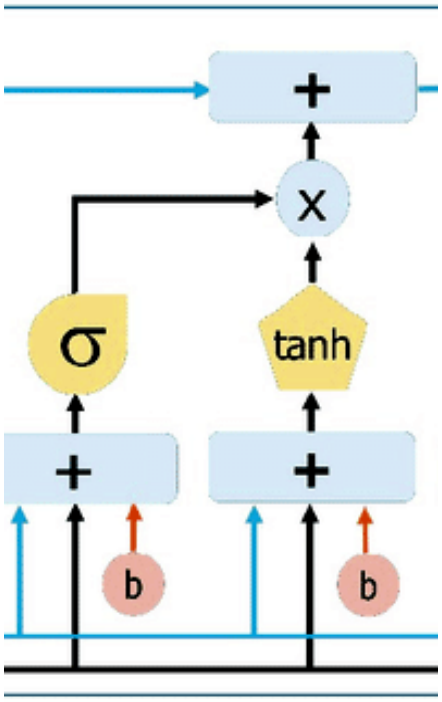


Figure 4: LSTM input gate

2.2.5 Output Gate

De laatste gate is de output gate. Deze zal de nieuwe hidden state bepalen aan de hand van de nieuwe cell state, de vorige hidden state en de nieuwe input data.

De nieuwe cell state word gecombineerd met de input data en vorige hidden state. De cell state word eerst nog door een tanh functie gestuurd om de waarden binnen het bereik van -1 en 1 te forceren.

De hidden state en input data worden door en sigmoid functie gestuurd.

Deze twee vectoren worden dan weer gemultipliceerd met elkaar en dit is dan de nieuwe hidden state.

In de afbeelding rechts ziet u ook de output maar deze word enkel helemaal op het einde gegeven en niet tijdens het doorsturen van data.

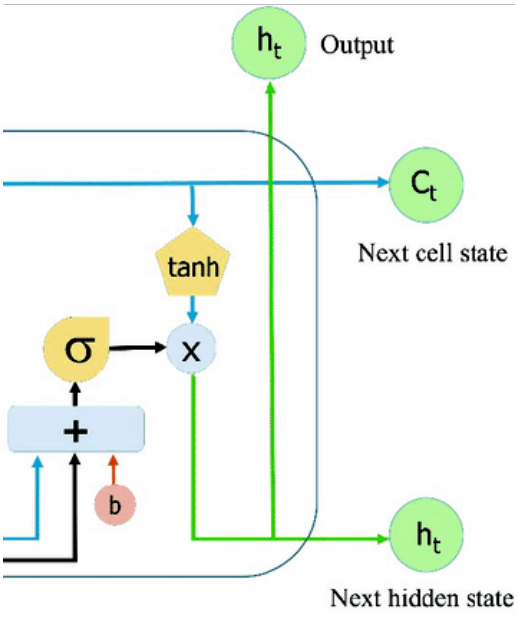


Figure 5: LSTM output gate

2.2.6 Eerste Tests

Een groot nadeel bij LSTM modellen is echter dat training heel lang duurt en krachtige hardware nodig heeft en dit werd ook voor ons al snel duidelijk.

We zijn begonnen met een simpel LSTM netwerk zoals hieronder te zien met 3 LSTM layers, een dense layer en een final output layer met 3 output neurons voor onze buy, sell en hold targets. De frame size, het aantal datapunten die je meegeeft om een prediction uit te voeren, was 240 candles.

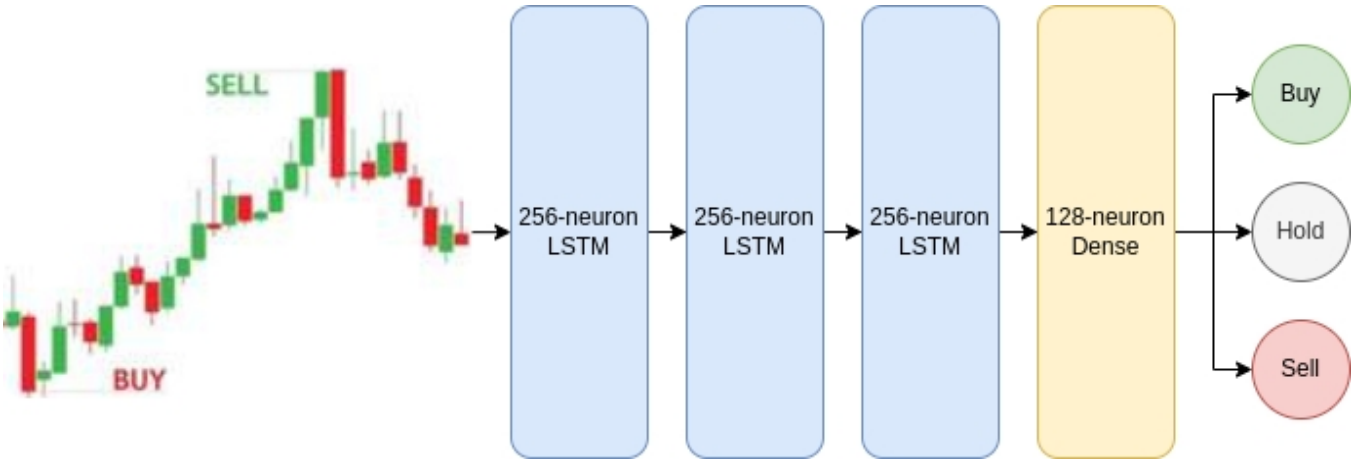


Figure 6: initial model layout

Het was echter snel duidelijk dat het model uit enkel prijs en volume geen goede voorspellingen kon maken maar dit hadden we eigenlijk wel verwacht. Daarom dat ons c++ programma ook direct gemaakt is met indicator berekeningen maar deze werden nog niet toegepast. Het is in theorie mogelijk maar het model zou mogelijks veel groter moeten zijn. De indicators zijn echter berekent met enkel prijs en volume data dus er word geen nieuwe data toegevoegd maar er worden wel al berekeningen op gedaan om bepaalde trends of prijs veranderingen te voorspellen en dit helpt het model enorm omdat het dit direct kan gebruiken en dus zelf minder deze berekeningen moet maken.

Ik zal nu eerst dieper ingaan op wat indicatoren juist zijn, wat het nut er van is en hoe deze berekend worden.

2.3 Indicatoren

Indicators zijn een onderdeel van technische analyse, iets wat investeerders vaak nog manueel doen om de volatiliteit, richting en sterkte van een trend van een stock of coin te bepalen. Het doel van een goede technische analyse is voorspellen wat er in de nabije toekomst zal gebeuren, er zijn zeer veel verschillende indicators met verschillende resultaten. We zullen enkel dieper ingaan op de 6 volgende indicators omdat deze zeer gekend zijn en vaak gebruikt worden en deze ook in ons onderzoek gebruikt zullen worden.

- Accumulation / Distribution Oscillator
- Average True Range
- Bollinger Bands®
- Moving Average Convergence Divergence
- Money Flow Index
- Relative Strength Index

de reden voor het combineren van meerdere indicators is omdat deze indicators andere dingen weergeven of voorspellen [6]. Zo zijn er een aantal types van indicators maar de meest gebruikte zijn volgende:

- momentum
- trend
- oscillator
- volatiliteit

de combinatie van deze verschillende soorten zorgt er voor dat we voldoende variatie hebben en hopelijk genoeg informatie voor ons neuraal netwerk om te leren wanneer het beter zou kopen, verkopen of niets doen.

2.2.1 Accumulation / Distribution Oscillator

De A/D oscillator, ook wel gekend als de chaikin oscillator, is een momentum indicator van de Acummulation/Distribution lijn en niet zo zeer de prijs van de coin zelf. De A/D lijn is een cummulatieve indicator die door middel van volume en prijs de supply en demand probeert te bepalen en hiermee de sterkte van een trend, of deze nu up of down is. Het kan echter ook dat de indicator het omgekeerde voorspelt van de trend op dit moment. Bevoorbeeld: Als de prijs aan het stijgen is maar de indicator daalt is de kans groot dat er een trend reversal aankomt.

Hieronder bevindt zich de formule voor het berekenen van de A/D oscillator.[15]

$$\mathrm{N} = \frac{(\mathrm{Close} - \mathrm{Low}) - (\mathrm{High} - \mathrm{Close})}{\mathrm{High} - \mathrm{Low}}$$

$$\mathrm{M} = \mathrm{N} * \mathrm{Volume\,(Period)}$$

$$\mathrm{ADL} = \mathrm{M\,(Period - 1)} + \mathrm{M\,(Period)}$$

$$\mathrm{CO} = (\text{3-day EMA of ADL}) - (\text{10-day EMA of ADL})$$

where:

$\mathrm{N} = \text{Money flow multiplier}$

$\mathrm{M} = \text{Money flow volume}$

$\mathrm{ADL} = \text{Accumulation distribution line}$

$\mathrm{CO} = \text{Chaikin oscillator}$

Figure 7: A/D oscillator formula

hier ziet u een voorbeeld van de indicator op een grafiek. Bovenaan de candles, in het midden het volume en onderaan de indicator. Het is duidelijk dat het volume een grote rol speelt bij deze indicator.

Figure 8: A/D oscillator example

2.2.2 Average True Range

Deze indicator is een volatiliteit indicator die de volatiliteit van een coin probeert te bepalen. Deze zegt niet echt iets over de richting of sterkte van trend maar kan wel samen met andere indicators een duidelijker beeld geven over de sterkte van een trend. De ATR is een subjectieve indicator en is vrij te interpreteren, er is geen vaste regel voor welke waardes een trend reversal voorspellen.

Hieronder bevindt zich de formule voor het berekenen van de ATR, `Cp` staat voor Previous Close.[17]

$$TR = \text{Max}[(H\ -\ L), \text{Abs}(H\ -\ C_P), \text{Abs}(L\ -\ C_P)]$$

$$ATR = \left(\frac{1}{n}\right)\sum_{(i=1)}^{(n)} TR_i$$

where:

$TR_i = \text{A particular true range}$

$n = \text{The time period employed}$

Figure 9: ATR formula

hier ziet u een voorbeeld van de indicator op een grafiek. Het volume is niet van belang bij deze indicator. Ook kan u zien dat de ATR niet de trend volgt maar stijgt bij grote veranderingen in prijs, dit is omdat het de volatiliteit aanduid en niet de trend.

Figure 10: ATR example

2.2.3 Bollinger Bands

Bollinger bands is ook een volatiliteit indicator maar wordt weergegeven over de candle grafiek en bevat 3 effectieve outputs, een lower, middle en upper band om een duidelijke volatiliteits 'range' aan te duiden. Deze indicator wordt vooral gebruikt om te zien of een coin oversold of overbought is. Als de waarde van de coin dicht bij of over de lower band gaat dan is deze oversold en vice versa voor de upper band. Deze formule maakt gebruik van een standaard afwijking en deze kan zelf gekozen worden maar meestal wordt 2 gebruikt. Een breakout buiten de bands is meestal een duidelijk teken van hoge volatiliteit en wordt meestal als een duidelijk signaal gezien om te kopen of verkopen.

Hieronden bevindt zich de formule om de bollinger bands te berekenen. De uiteindelijke middle band is het average van de upper en lower band en staat niet vermeld in deze formule.[16]

$$\mathrm{BOLU} = \mathrm{MA}(\mathrm{TP}, n) + m * \sigma[\mathrm{TP}, n]$$

$$\mathrm{BOLD} = \mathrm{MA}(\mathrm{TP}, n) - m * \sigma[\mathrm{TP}, n]$$

where:

$\mathrm{BOLU} = \text{Upper Bollinger Band}$

$\mathrm{BOLD} = \text{Lower Bollinger Band}$

$\mathrm{MA} = \text{Moving average}$

$\mathrm{TP\ (typical\ price)} = (\mathrm{High} + \mathrm{Low} + \mathrm{Close}) \div 3$

$n = \text{Number of days in smoothing period (typically 20)}$

$m = \text{Number of standard deviations (typically 2)}$

$\sigma[\mathrm{TP}, n] = \text{Standard Deviation over last } n \text{ periods of TP}$

Figure 11: Bollinger Bands Formula

Hieronder ziet u een voorbeeld van de indicator op een grafiek. Het is duidelijk te zien dat meestal als de candles de upper of lower band aanraken er een trend reversal is. De grootte of duur van de trend reversal is echter niet te bepalen met enkel bollinger bands dus deze zijn soms maar heel klein en van korte duur.

Figure 12: Bollinger Bands Example

2.2.4 Moving Average Convergence Divergence

De MACD is een trend-following momentum indicator die de relatie tussen 2 moving averages van een verschillende lengte weergeeft. De MACD wordt meestal gebruikt met exponential moving averages (EMA) maar andere types kunnen zeker ook gebruikt worden. Een EMA houdt meer rekening met de meer recente data punten minder met oude data punten.

De MACD is een lagging indicator, dit wilt zeggen dat deze eigelijk een beetje achter loopt op wat er eigelijk aan het gebeuren is maar desondanks is dit een vaak gebruikte en nuttige indicator en wordt deze toch gebruikt om trend reversals te voorspellen.

Hieronder bevindt zich de formule voor de MACD.[12]

$$\text{MACD} = \text{12-Period EMA} \; - \; \text{26-Period EMA}$$

Figure 13: MACD formula

Naast deze lijn kan je ook de MACD signal line gebruiken door een EMA te nemen van de MACD. Als je deze dan aftrekt van de effectieve MACD krijg je het MACD historgram te zien op onderstaande grafiek.

De blauwe lijn is de MACD, de oranje lijn is de Signal line en dan zie je ook het histogram in groen en rood.

Het kruisen van de blauwe en oranje lijn wordt gezien als een bullish of bearish crossover afhankelijk van of de blauwe naar boven of naar beneden door de oranje lijn gaat.

Figure 14: MACD example

2.2.5 Money Flow Index

De MFI is een oscillator die gebruik maakt van prijs en volume data om een overbought of oversold signaal weer te geven. De naam Money Flow Index is omdat deze de prijs en volume gebruikt en dit dus eigelijk een berekening is op de hoeveelheid geld die verhandeld wordt. Deze indicator wordt vooral gebruikt voor het voorspellen van een trend reversal. De MFI bevind zich altijd tussen een waarde van 0 en 100, een waarde boven 80 wordt meestal als een overbought signaal gezien en onder 20 als oversold. Als de indicator begint te stijgen tijdens het dalen van de prijs kan dit ook wijzen op een trend reversal.

Hieronder bevindt zich de formule vinden van de MFI.[14]

$$\text{Money Flow Index} = 100 - \frac{100}{1 + \text{Money Flow Ratio}}$$

where:

$$\text{Money Flow Ratio} = \frac{\text{14 Period Positive Money Flow}}{\text{14 Period Negative Money Flow}}$$

$$\text{Raw Money Flow} = \text{Typical Price} * \text{Volume}$$

$$\text{Typical Price} = \frac{\text{High} + \text{Low} + \text{Close}}{3}$$

Figure 15: MFI formula

Op onderstaande grafiek ziet u de MFI en er zijn ook horizontale lijnen getrokken op 80 en 20 om deze overbought en oversold signalen duidelijk te maken. Het is duidelijk te zien dat een groot volume een grote impact kan hebben op de MFI en deze indicator op zich niet heel duidelijk is en meestal samen met andere indicators gebruikt wordt.

Figure 16: MFI example

2.2.6 Relative Strength Index

De RSI is een momentum indicator dat ook wordt gebruikt voor overbought en oversold signalen maar deze gebruikt enkel prijs data en geen volume data. Er zijn meerdere varianten van de RSI maar er zit niet zo een groot verschil tussen de andere varianten dus wij hebben de originele RSI gekozen. De RSI is vooral nuttig in een situatie met hoge volatiliteit omdat deze anders een lange tijd hetzelfde signaal weergeeft als een coin blijft stijgen of dalen.

Ook deze indicator bevindt zich steeds tussen 0 en 100 en de signalen worden vooral gebruikt als deze boven 80 of onder 20 gaat.

Hieronder bevindt zich de formule voor de RSI.[13]

$$RSI_{\text{step one}} = 100 - \left[\frac{100}{1 + \frac{\text{Average gain}}{\text{Average loss}}}\right]$$

$$RSI_{\text{step two}} = 100 - \left[\frac{100}{1 + \frac{(\text{Previous Average Gain} \times 13) + \text{Current Gain}}{((\text{Previous Average Loss} \times 13) + \text{Current Loss})}}\right]$$

Figure 17: RSI formula

Zoals te zien op onderstaande grafiek volgt de RSI duidelijk de prijs maar de hoge en lage pieken buiten de 80 en 20 wijzen toch meestal wel op een trend reversal.

Figure 18: RSI example

2.4 Training

2.4.1 Supervised Learning

Voor het trainen van het model hebben we dus gekozen voor een supervised vorm van training. We hebben onze data op voorhand gelabeled op een manier waarvan we denken het een goed target is voor het model om naartoe te werken, echter heeft dit ook wat nadelen dat hier toegelicht zullen worden.

2.4.2 Nadelen

Een van de nadelen van deze manier van werken is dat de accuracy tijdens training geen duidelijke metric is voor de effectieve winst die het model zou kunnen halen omdat het model in theorie elke keer een 'foute' prediction zou kunnen doen en toch nog goed genoeg zijn om winst te maken.

Bevoorbeeld, in onderstaande afbeelding ziet u groene en rode aanduidingen op de grafiek. Groen staat voor een aankoop en rood voor een verkoop. Deze zijn niet op de meest optimale punten, waar dus de labels zouden staan, maar het is wel duidelijk dat deze trades winstgevend zouden zijn. Toch zou het model in dit geval een lage accuracy gehaald hebben tijdens training desondanks de goede winst. Dit is een nadeel bij onze manier van training die mogelijks verholpen kan worden door reinforcement learning te gebruiken waarbij je eerder rekening zou houden met de totale winst over een periode als reward en niet met vaste labels.

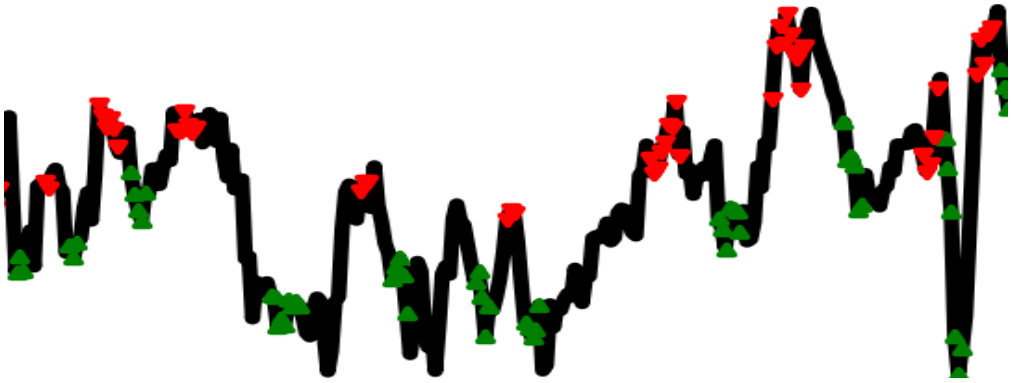


Figure 19: Buy & sell example

Reinforcement learning is zeker een goed alternatief en is ook wel wat info over te vinden waar het toegepast word op trading met LSTM netwerken. [3][4]

Het trainen van LSTM modellen is in het algemeen ook redelijk traag vergeleken met veel andere soorten layers vanwege de grotere hoeveelheid berekeningen dat gedaan worden binnen 1 LSTM neuron. Hieronder ziet u de interne structuur van een LSTM neuron zoals eerder uitgelegd en daaronder een simpele neuron dat wordt gebruikt in Dense layers.

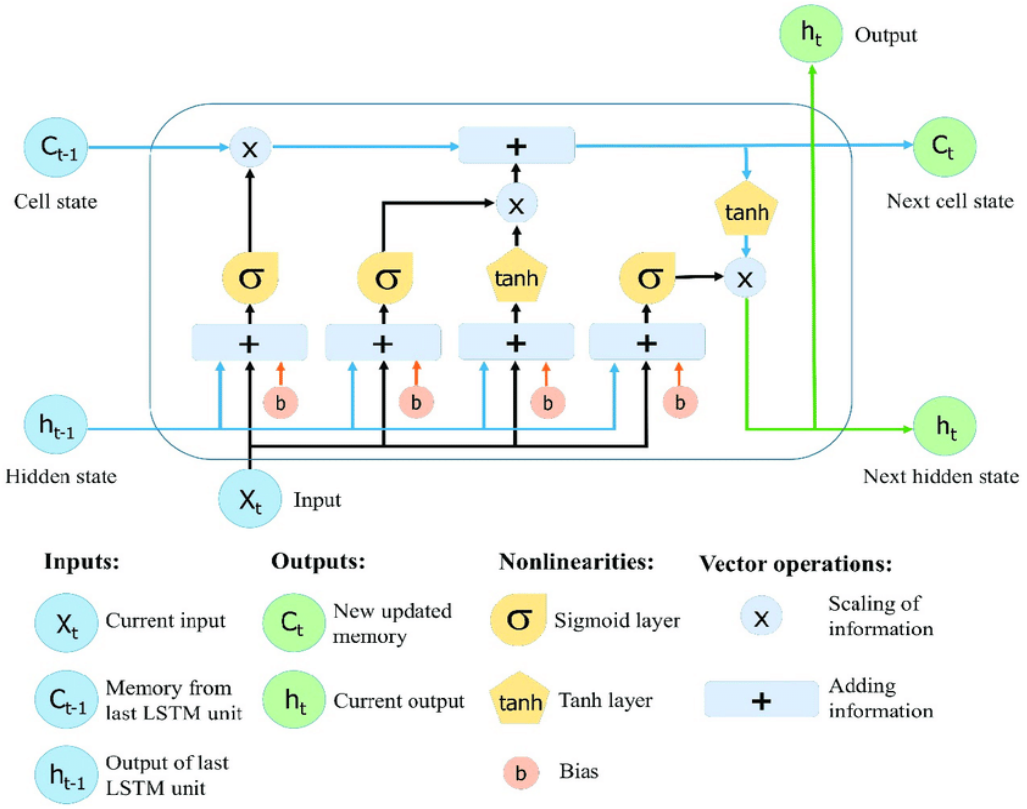


Figure 20: LSTM neuron structure

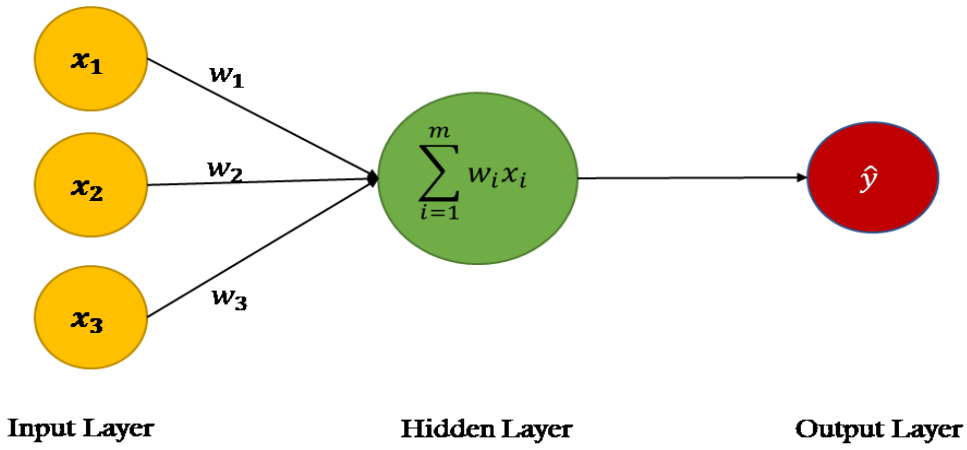


Figure 21: Simple Neuron Structure

2.5 Testing

Een van de belangrijkste dingen voor het bevestigen van een succesvolle training is natuurlijk het testen. Zeker aangezien het bij ons niet direct aan de accuracy tijdens training te zien is of het model effectief winst haalt of niet.

We proberen telkens meerdere verschillende coins te testen en de resultaten te vergelijken met elk model dat we al gemaakt hadden. Zo was uiteindelijk een 10 layer LSTM model gemiddeld het beste over alle data. We waren hierbij vooral geïnteresseerd in de winst per trade omdat deze hoog genoeg moet liggen om nog winstgevend te zijn na aftrek van trading fees en we zouden deze er al kunnen in verwerken in de berekening maar dit is afhankelijk van de exchange die je zou gebruiken dus dit hebben we niet gedaan.

Als deze hoog genoeg lag dan werd er gekeken naar de totaal behaalde winst over een bepaalde periode, we lieten het model op een deel van onze dataset voorspellingen maken om dan te kijken hoe goed deze het gemiddeld doet en zo altijd het beste model eruit kiezen en kijken wat hier anders aan is om to verder te optimaliseren.

Ook werden er van elk model meerdere plots gemaakt met de voorspellingen om te kijken hoe het model deze resultaten behaalde.

2.6 Extra Verbeteringen

2.6.1 Piramidding

Piramidding is een trading strategie waarmee je winst kan optimaliseren en verlies minimaliseren. Je gaat door middel van meerdere buy orders je average aankoop punt verlagen om de winst bij een stijging in prijs te verhogen en het verlies bij een minder grote stijging te verlagen. Er hangt echter ook een iets groter risico aan vast in het geval dat het toch blijft dalen want dan zit er redelijk wat kapitaal in.

Bevoorbeeld:

Je gaat bij je eerste aankoop slechts een deel van je kapitaal in een buy order zetten. Als de waarde direct begint te stijgen en je verkoopt heb je direct winst maar dit is natuurlijk niet altijd makkelijk.

Als de waarde echter zou dalen kan je nog een deel van je kapitaal er in steken. Hierdoor is je gemiddeld aankooppunt ergens tussen de 2, afhankelijk van de verdeling van kapitaal over de 2 orders. Dit kan je blijven herhalen zolang de waarde daalt en zolang je kapitaal hebt dat je er in wil steken.

Stel dat de waarde dan toch stijgt heb je meer winst dan als je had gewacht en enkel je eerste buy order had gehad. Het is niet zo goed als al je kapitaal in het laagste punt pas er in steken maar dit is moeilijk te voorspellen wanner dit juist is. Zo kan je ook je verlies minimaliseren als je een buy order boven en een buy order onder de waarde van je sell punt hebt gezet. Je zou enkel verkopen in zo een situatie als je verwacht dat de waarde weer sterk zou gaan dalen maar dit is natuurlijk ook altijd een beetje een gok.

Figure 22: Piramidding example

Na wat testing met piramidding zagen we dat dit in bijna alle gevallen voor een verhoogde winst zorgde. Je kan zelf bepalen hoe vaak je het model wilt laten piramidden want je zet elke natuurlijk meer geld in waardoor het risico ook stijgt.

2.6.2 Reinforcement Learning

Onze trainingsmethode maakt gebruik van labelled data, dit heeft voordelen maar ook nadelen.

De training zal sneller verlopen dan bij reinforcement learning maar is ook heel gelimiteerd tot hoe goed je labels zijn, als deze niet optimaal zijn dan zal je neuraal netwerk ook nooit beter worden dan die labels. En zoals eerder bij training al besproken kan je ook bij goede labels toch een lage accuracy hebben ookal doet je model het wel goed.

Maar eerst even, wat is reinforcement learning juist?

Bij reinforcement learning ga je modellen training op basis van hun beslissingen en de reward die je aan deze beslissingen geeft. De agent, het neuraal netwerk, zal proberen de environment te leren kennen door trial & error. Elke beslissing krijgt dan een reward afhankelijk van het resultaat van deze handeling. De agent zal proberen deze rewards te maximaliseren. Hoe de rewards berekent worden is volledig zelf te bepalen maar is zeer belangrijk, een slechte rewards functie heeft slechte resultaten net zoals slechte labelled data bij supervised learning.

Reinforcement learning zou een grote verbetering kunnen zijn in de uiteindelijke performance van het model maar training zou enorm veel langer duren, wat nu al redelijk lang duurt omdat het model bestaat uit LSTM layers. Desondanks is dit zeker iets om te verkennen in verder onderzoek naar deze toepassing van neurale netwerken op crypto trading.

2.6.3 Automation van training & testing

Een limiterende factor van ons onderzoek was tijd, dit zou wel verbeterd kunnen worden als training en testing geautomatiseerd kon worden. Momenteel werd training elke keer manuel gestart met parameters waarvan we dachten dat het beter zou zijn, dit duurde dan telkens wel enkele uren om te trainen. Testing van modellen achteraf duude ook altijd wel even omdat je een groot deel van data moet testen voor een representatieve weergave van hoe goed een model is. Je kan niet enkel testen op cryptomunten die bijna alleen maar stijgen en dan de beste nemen om achteraf te zien dat die op een downtrend enorm veel verlies maakt.

Als deze workflow automatisch zou gebeuren, van training tot testing met dan een duidelijk test resultaat achteraf zou dit heel wat tijd besparen om meer onderzoek te kunnen doen of andere dingen te testen.

2.6.4 Explainable AI

Een verdere verbetering is Explainable AI. Hiermee proberen we te begrijpen wat het model doet, welke data het relevant vind voor de uitkomst en welke niet. Hiermee zouden we onze indicators kunnen verbeteren omdat bepaalde indicators mogelijk niet relevant zijn. Dit word vaak toegepast op CNN's omdat je gemakkelijk een heatmap over de originele image kan leggen om aan te duiden welke

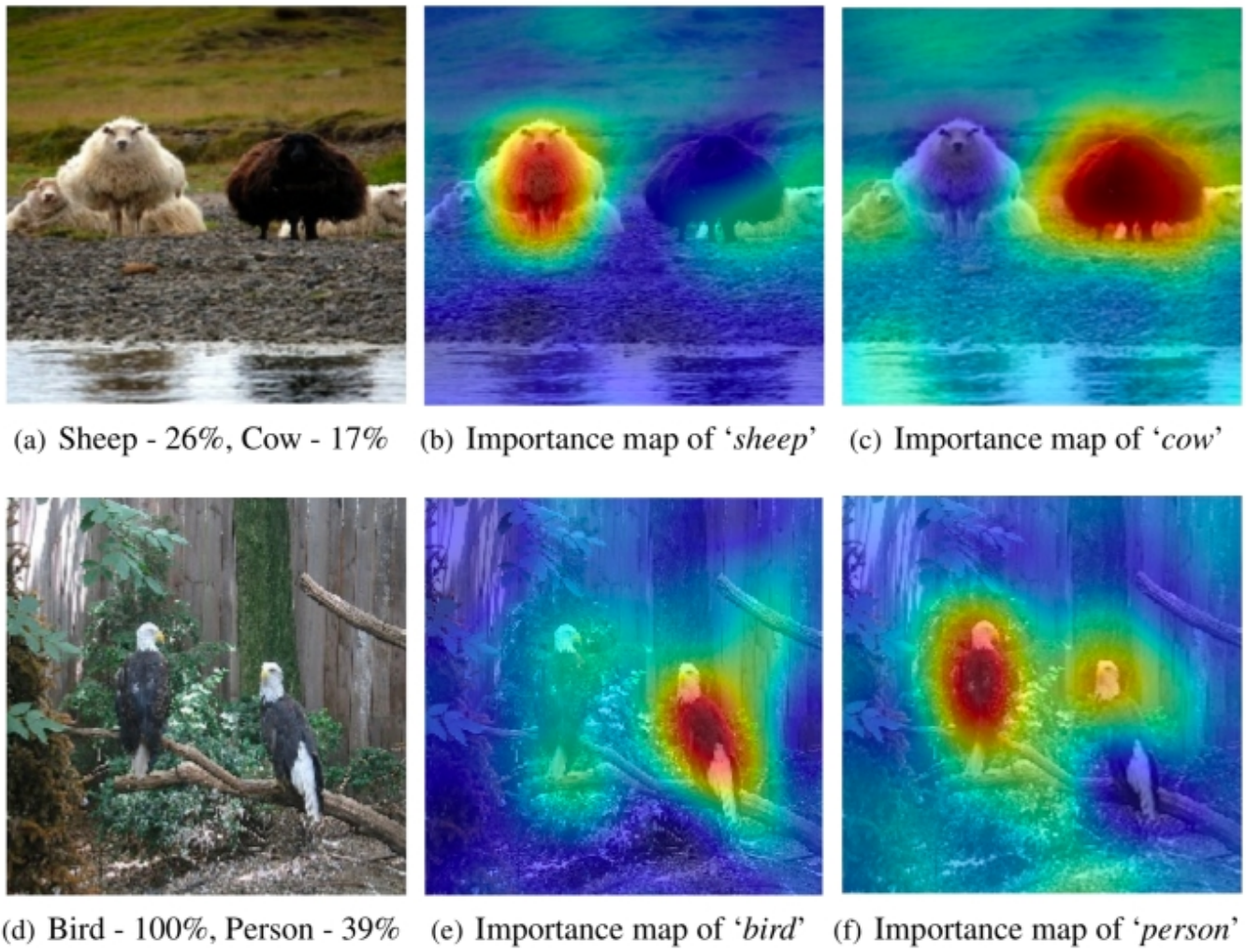


Figure 23: Explainable AI for CNN

pixels belangrijk waren en welke niet. Hieronder een voorbeeld hiervan.

3 Technisch onderzoek

3.1 software, tools en programmeertalen

Er word vooral gebruik gemaakt van Python en een deel C++ voor preprocessing van de data. We hebben C++ gekozen voor de preprocessing omdat dit een groot verschil maakte in de snelheid en dit was belangrijk omdat er toch een 15GB aan CSV files verwerkt moest worden.

De belangrijkste libraries die we gebruikt hebben zijn:

- Tensorflow & Keras
- Pandas
- Numpy
- TaLib (c++ maar bestaat ook voor python) [9]
- Matplotlib
- python-binance (versie 1.0.12)

3.2 structuur en workflow

Hieronder vind u een representatie over hoe we van raw data naar trained neural netwerk gaan. We beginnen eerst met de data collectie met een simpel python script en deze word naar CSV files geschreven.

Dan zijn er 2 mogelijkheden, stel dat je de data wilt rescalen naar bevoorbeeld uurcandles dan kan je dit doen met ons ticker time rescale programma. Dit is geschreven in C++ en zal binary files wegschrijven in plaats van CSV. Dit is voor 2 redenen, de hoeveelheid opslag die we nodig hadden voor de binary files vergeleken met de csv files was ongeveer 30% minder. Snelheid van lezen en schrijven in binary was ook veel sneller dan csv.

Als je je data niet wilt rescalen kan je de csv files ook rechtstreeks in onze data preprocessor verwerken. Deze is verantwoordelijk voor de indicatoren, scaling en labelling van alle data. Dit word dan weer naar binary geschreven om dan in te lezen in ons training script voor neurale netwerken te trainen.

Als we een model getrained hebben gaan we deze testen op onze data om te zien of deze het goed doet en of er bepaalde patronen zijn die goed of slecht zijn. We willen namelijk een model dat ook in een downtrend het goed doet, ookal wilt dit zeggen dat deze mogelijks in een downtrend helemaal niets doet. Dit is nog steeds beter dan verlies maken.
We maken dan met matplotlib grafieken om te kijken waar de buy en sell predicties zijn en wat de winst is vergeleken met onze andere modellen.

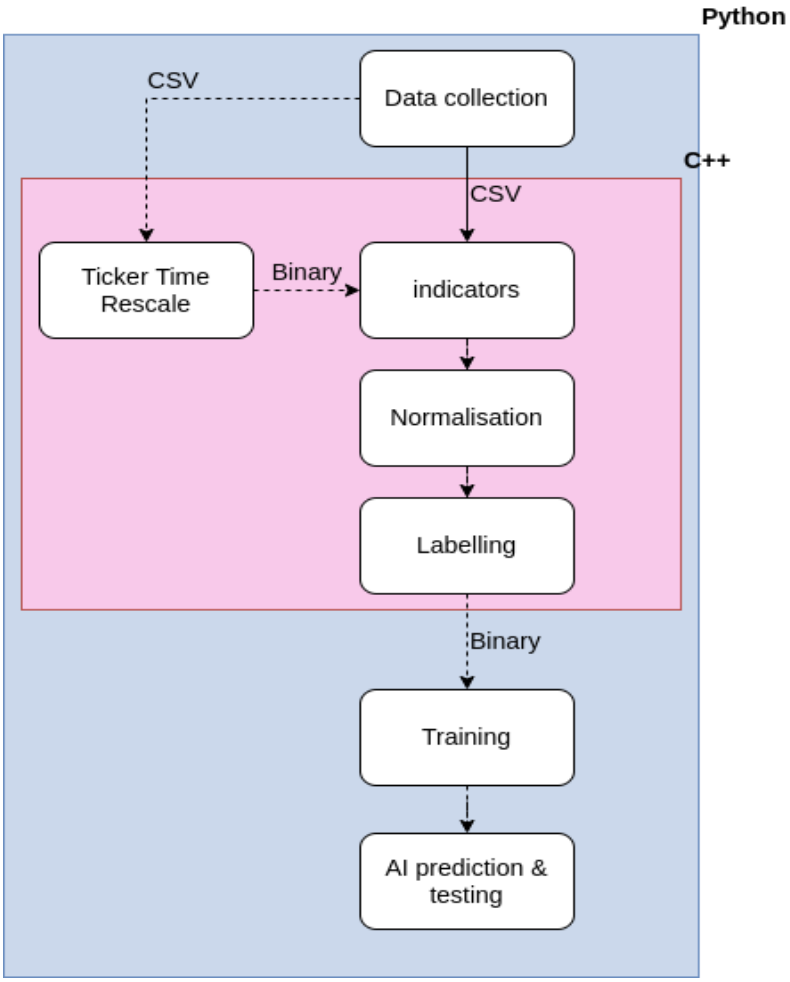


Figure 24: Data flow

3.3 data processing

Onze data processing workflow bevat 3 delen

- labelling
- indicator calculation
- normalization

Labelling

In het labelling deel van onze data preprocessing gaan we targets toevoegen aan onze data om achteraf ons model op te trainen, de manier waarop deze labels berekend worden is zeer belangrijk want dit zal een grote invloed hebben op het uiteindelijke model.

Voor onze berekening kijken we naar een stuk data en overlopen 1 voor 1 elke candle. We houden hiervan de laagste bij die we tegenkomen en gaan dan verder. Dan zolang het blijft stijgen houden we de hoogste candle ook bij tot het terug begint te dalen. Als de stijging van de laagste tot de hoogste meer is dan 1% (dit is voldoende om winst te maken, ook met realistische trading fees) dan zetten we een buy target op die laagste candle en een sell target op de hoogste en beginnen we terug opnieuw tot de hele dataset overlopen is. Hieronder bevinden zich een aantal stukken code waar dit gebeurd.

Dit zijn de variabelen die gebruikt worden in de verdere stukken code.

Hier bepalen we ook de min_change, deze variabele is om te bepalen of we willen dat een sell target ten minste 1%, 2%, … boven een buy target staat. We kunnen dit hoger zetten om het model te forceren op langere termijn te traden en minder de kleine schommelingen.

Hieronder bevind zich een kort stuk code dat zal bepalen of een candle stijgt of daalt door te checken of de close meer is dan 0 of niet (deze zijn reeds gescaled tussen -1 en 1).

deze worden dan aan lijsten toegevoegd om verder te gebruiken voor labelling.

```
std::vector<double> cum_down;
std::vector<double> cum_up;

std::vector<double> cum_down_buy;
std::vector<double> cum_up_buy;

std::vector<double> cum_down_sell;
std::vector<double> cum_up_sell;

double min_change = 0.01; // 1%

double last_max = 0;
bool allow_buy = false;
int last_max_index = 0;

double last_min = 0;
bool allow_sell = false;
int last_min_index = 0;
```

Figure 25: Labelling variables

```
if (candle->m_close > 0)
{
    double prev = cum_up.size() ? cum_up[cum_up.size() - 1] : 0;

    cum_up.push_back(prev + candle->m_close);
    cum_up_buy.push_back(prev + candle->m_close);
    cum_up_sell.push_back(prev + candle->m_close);
}
else
{
    double prev = cum_down.size() ? cum_down[cum_down.size() - 1] : 0;

    cum_down.push_back(prev + fabs(candle->m_close));
    cum_down_buy.push_back(prev + fabs(candle->m_close));
    cum_down_sell.push_back(prev + fabs(candle->m_close));
}
```

Figure 26: Labelling cumulative candle lists

Hier bevind zich het grootste deel van de labelling. We gaan hier telkens over elke candle loopen, er word dan gechecked of deze stijgt of daalt net zoals hierboven.

Dan word er bepaald of de stijging vergeleken met de vorige buy voldoende is om een sell te plaatsen. Vanaf dan gaan we kijken tot waar het blijft stijgen.

Als het stopt met stijgen en de daling hierna is ook groter dan onze op voorhand bepaalde minimum change, dan word er een sell order gezet op deze piek. Voor de buy targets gebeurd hetzelfde maar omgekeerd.

```
for (size_t i = 0; i < candles->size(); i++)
{
    std::unique_ptr<candle>& candle = candles->at(i);

    if (cum_up.size() && cum_down.size())
    {
        double prev_up = cum_up[cum_up.size() - 1];
        double prev_down = cum_down[cum_down.size() - 1];

        double prev_up_buy = cum_up_buy[cum_up_buy.size() - 1];
        double prev_down_buy = cum_down_buy[cum_down_buy.size() - 1];

        double prev_up_sell = cum_up_sell[cum_up_sell.size() - 1];
        double prev_down_sell = cum_down_sell[cum_down_sell.size() - 1];

        if (candle->m_close > 0)
        {
            if (prev_up - prev_down > min_change)
            {
                if (prev_up_sell - prev_down_sell > last_max)
                {
                    last_max_index = i;
                    last_max = prev_up_sell - prev_down_sell;
                    allow_buy = true;
                    cum_up_buy.clear();
                    cum_down_buy.clear();
                    cum_up.clear();
                    cum_down.clear();
                }
                if (allow_sell)
                {
                    candles->at(last_min_index)->m_target = eTarget::SELL;
                    last_min_index = 0;
                    last_min = min_change;
                    allow_sell = false;
                    cum_up_buy.clear();
                    cum_down_buy.clear();
                }
            }
        }
        else if (candle->m_close < 0)
        {
            if (prev_down - prev_up > min_change)
            {
                if (allow_buy)
                {
                    candles->at(last_max_index)->m_target = eTarget::BUY;
                    last_max_index = 0;
                    last_max = min_change;
                    allow_buy = false;
                    cum_down_sell.clear();
                    cum_up_sell.clear();
                }
                if (prev_down_buy - prev_up_buy > last_min)
                {
                    last_min_index = i;
                    last_min = prev_down_buy - prev_up_buy;
                    allow_sell = true;
                    cum_down_sell.clear();
                    cum_up_sell.clear();
                    cum_up.clear();
                    cum_down.clear();
                }
            }

        }
    }
```

Figure 27: Labelling calculation

Het resultaat ziet er dan zo uit, afhankelijk van de min_change zijn de targets enkel op grote prijsverschillen of ook op kleine prijsverschillen.

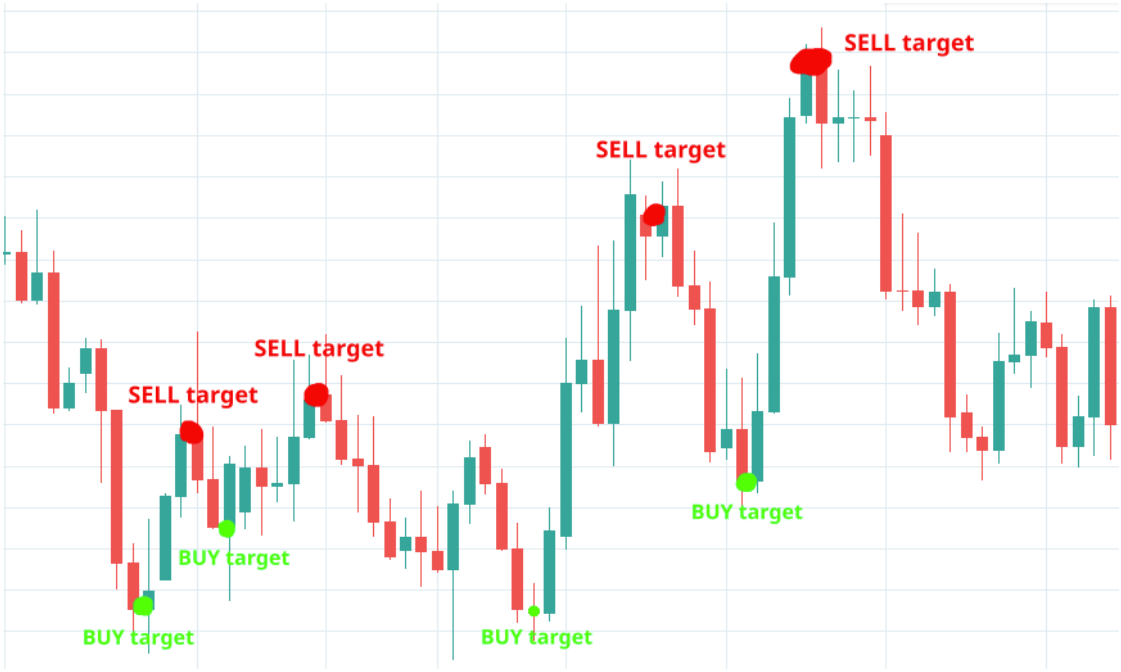


Figure 28: labelling example

Indicators

Voor onze indicators maken we gebruik van een heel bekende library TA-lib, opgestart als hobby project in 1999 door Mario Fortier, deze is gelicenseerd onder de BSD license. Dit laat het gebruik toe in open-source en commerciele producten.

Door TA-lib is de indicator berekening een heel simpel process. Je moet enkel de data meegeven en de parameters van die specifieke indicator. Afhankelijk van de indicator is dit 1 parameter maar dit kunnen er ook 4 zijn of mogelijks meer.

Om even te schetsen hoe dit eruit ziet in code. Hieronder bevind zich een van onze indicator berekeningen, namelijk de MACD. Deze indicator verwacht 3 parameters. De fast, slow en signal period. Deze bepalen met hoeveel historische gegevens de Moving averages en het signal berekent worden. Ta-lib geeft je de data terug in vooraf gedefinieerd variabelen.

```
void calculate_macd(const size_t fast_period = 12, const size_t slow_period = 26, const
size_t signal_period = 9)
{
    double* tmp_macd = new double[m_alloc_size];
    double* tmp_macd_signal = new double[m_alloc_size];
    double* tmp_macd_hist = new double[m_alloc_size];

    int beginIdx, endIdx;
    TA_MACD(0, m_alloc_size, m_close, fast_period, slow_period, signal_period, &beginIdx,
    &endIdx, tmp_macd, tmp_macd_signal, tmp_macd_hist);

    for (size_t i = beginIdx; i < endIdx; i++)
    {
        const std::unique_ptr<candle>& candle = m_candles->at(i);

        candle->m_macd = tmp_macd[i];
        candle->m_macd_signal = tmp_macd_signal[i];
        candle->m_macd_hist = tmp_macd_hist[i];
    }

    delete[] tmp_macd;
    delete[] tmp_macd_signal;
    delete[] tmp_macd_hist;
}
```

Figure 29: MACD calculation TA-lib

Scalen

Een heel belangrijk deel van data processing bij neurale netwerken is het correct scalen van de data. Als de data niet goed gescaled is kan je model moeilijker de correcte weights en biases vinden.

Je model moet zich trainen met de gegevens die binnenkomen, maar als deze waardes op compleet andere schalen zitten. bijvoorbeeld Bitcoin heeft een prijs van rond de 30000 euro en Dogecoin 0.30 euro. Als je model getrained is met Bitcoin data dan zal deze de prijswaardes van Dogecoin minder relevant beschouwen ookal is dit niet het geval.
De manier van scaling is afhankelijk van de data die je gebruikt, wij gebruiken percentage change van het ene datapunt naar het volgende. Dit is de meest logische vorm van scaling voor onze data. Zo blijft dit gelijk over alle coins en niet afhankelijk van de maximum en minimum waardes.

De meeste waardes scalen we door het procentuele verschil met de vorige candle te berekenen. Bij de MFI en RSI indicators delen we deze gewoon door 100 omdat deze al op een schaal van 0 tot 100 staan.

Hieronder bevind zich het stuk code waar dit gebeurd, deze functie bevind zich in onze candle struct waar alle data per candle in zit.

```
void normalize(candle* other)
{
    // don't update timestamp since we're on the newest candle of the two
    auto calc = [](double a1, double a2) { return a1 == 0 ? a1 : (a2 - a1) / a1; };

    this->m_open = calc(this->m_open, other->m_open);
    this->m_close = calc(this->m_close, other->m_close);
    this->m_high = calc(this->m_high, other->m_high);
    this->m_low = calc(this->m_low, other->m_low);
    this->m_volume = calc(this->m_volume, other->m_volume);

    this->m_adosc = calc(this->m_adosc, other->m_adosc);
    this->m_atr = calc(this->m_atr, other->m_atr);

    this->m_macd = calc(this->m_macd, other->m_macd);
    this->m_macd_hist = calc(this->m_macd_hist, other->m_macd_hist);
    this->m_macd_signal = calc(this->m_macd_signal, other->m_macd_signal);

    this->m_upper_band = calc(this->m_upper_band, other->m_upper_band);
    this->m_middle_band = calc(this->m_middle_band, other->m_middle_band);
    this->m_lower_band = calc(this->m_lower_band, other->m_lower_band);

    this->m_mfi /= 100;
    this->m_rsi /= 100;

    this->m_difference_lowhigh = calc(other->m_low, other->m_high);
    this->m_difference_openclose = calc(other->m_open, other->m_close);
}
```

Figure 30: Normalize candle code

3.4 Model opbouw

Het model ging een type RNN worden maar hoe of wat we exact zouden gebruiken was nog niet zo duidelijk. Na wat research zijn we dan voor LSTM netwerken gegaan. Hierin hebben we wat geexperimenteerd met de grootte van layers, aantal layers, training time en learning rate. Uiteindelijk hadden we een redelijk goed 3 Layer LSTM model, maar we wouden natuurlijk ook weten of groter beter zou zijn en dit was in ons geval slechts deels het geval. Het beste model was uiteindelijk een 10 layer LSTM model. deze was hooguit een beetje beter dan het 3 layer model en heel soms zelfs minder goed maar algemeen iets beter.

Het is bij LSTM layers uitzonderlijk dat er meer dan 3 layers gebruikt worden, het is goed mogelijk dat met meer training of wat veranderingen aan de hyperparameters we een even goed of beter 3 layer model kunnen bekomen. Uit onderzoek blijkt ook dat grotere modellen zeker niet altijd beter zijn, vooral bij NLP is dit opvallend. [7][8]

4 Reflectie

Een bachelorproef is in wezen een kritische reflectie op een vraag uit het praktijkveld. Ze levert een bijdrage aan de praktijk. Je zult dus een antwoord moeten formuleren op jouw onderzoeksvraag.

Wees eerlijk: indien jouw onderzoek (nog) niet het gewenste resultaat gaf, vermeld je dit ook.

Een kritische reflectie is onderbouwd en gebaseerd op contacten met betrouwbare bronnen. Met wie kun je aftoetsen? Jouw stagebedrijf, gespecialiseerde communities, contacten uit het werkveld, lectoren…

Een kritische reflectie betekent dat je je baseert op jouw onderzoek en dat vergelijkt met bevindingen uit de praktijk. Je zult dus op zoek moeten gaan naar analoge onderzoeken/resultaten/praktijkervaringen en jouw bevindingen met hen aftoetsen. Stellen zij dezelfde problemen vast? Hebben ze een andere visie? Kunnen ze jou een andere insteek geven?

Een kritische reflectie is dus niet hetzelfde als kritiek geven op een bepaalde situatie uit de praktijk. Evenmin het ventileren van je persoonlijke meningen over de situatie of het probleem uit de praktijk.

Beantwoord daarom gedetailleerd volgende vragen. Vermeld steeds de bronnen/bedrijven/contactpersonen.

- *Wat zijn de sterke en zwakke punten van het resultaat uit jouw researchproject?*
- *Is 'het projectresultaat' (incl. methodiek) bruikbaar in de bedrijfswereld?*
- *Wat zijn de mogelijke implementatiehindernissen voor een bedrijf?*
- *Wat is de meerwaarde voor het bedrijf?*
- *Welke alternatieven/suggesties geven bedrijven en/of community?*
- *Is er een maatschappelijke/economische/socio-economische meerwaarde aanwezig?*
- *Wat zijn jouw suggesties voor een (eventueel) vervolgonderzoek?*

Gebruik hiervoor verschillende onderdelen.

Dit hoofdstuk is heel belangrijk, vandaar de vereiste om minimum 3 à 4 pagina's hieraan te spenderen.

Evaluatiecriteria van dit hoofdstuk:

Onvoldoende: de reflectie over het resultaat ontbreekt volledig of is ondermaats (geen gegronde motivering,...)
Beperkt: de student heeft enkel aan zelfreflectie gedaan. Motivering is aanwezig.
Volstaat: de onderzoeksresultaten werden kritisch geëvalueerd: naast zelfreflectie is er beperkte input van externen.
Goed: de reflectie baseert zich op contacten met verschillende externen. Daardoor is de reflectie zeer waardevol en bruikbaar voor student en lezer.
Excellent: door contacten met externen uit verschillende achtergronden/disciplines voelt de student zeer goed aan wat in het werkveld leeft. Er is niet alleen aandacht voor technische alternatieven, suggesties, ... maar ook niet-technische relevante aspecten.

4.1 Resultaat

Na het onderzoek was het resultaat zowel beter dan verwacht maar op bepaalde gebieden ook teleurstellend. De resultaten zijn ook niet echt representatief van wat je zou kunnen halen bij echte crypto trading, hier komen nog fees bij kijken natuurlijk en de kleine variaties in prijs tijdens het predicten en versturen van je orders. Het model kan in de meeste situaties wel correct inschatten wanneer het best zou aankopen of verkopen wat nog wel indrukwekkend is gezien de onvoorspelbaarheid van de crypto markt, onze korte onderzoeksperiode en beperkte training.

Zolang er voldoende volatiliteit in de prijs van een coin zit kan het model zeer goed de aankopen en verkopen predicten en in een uptrend is dit zeker geen probleem maar in een downtrend werkt dit niet altijd even goed. Het model gaat hier vaak blijven kopen en verkopen op momenten waar het eigelijk verlies maakt. Dit is te wijten aan de reden dat ons model tijdens training geen besef heeft van winst, bij reinforcement learning zou je hogere rewards kunnen toekennen aan acties met hogere winst en zo mogelijks het model aanleren dat dit beter is.

Het model haalde op een 2 maanden tijd bij veel coins meer dan 150% winst, zo een grote winst op een korte periode ligt aan de volatiliteit van crypto, hierdoor kan het model soms zelfs in een downtrend genoeg kleine trades maken met voldoende winst. Iemand die manueel koopt en verkoopt kan zeker ook deze winst halen en meer dus het outperformed goede traders zeker nog niet maar voor een automatisch systeem is dit toch al zeer goed. Het grootste voordeel is dat je zelf niet constant de prijs in de gaten moet houden zoals traders wel dagelijks doen.

Er kan zeker nog verbeterd worden want momenteel koopt het model veel te snel aan, het zou beter meestal nog even wachten dus mogelijks met wat extra training of het toepassen van een andere trainingsmethode zoals reinforcement learning kan dit wel beter worden. Ook hebben we nog niet kunnen experimenteren met Explainable AI, een manier om te onderzoeken welke inputs het meeste invloed hebben op de voorspellingen van het model. Dit zou ons een beter inzicht kunnen geven in het effectieve nut van de indicators of prijsdata.

Als we even kijken naar hoe goed onze modellen het doen op een deel van onze dataset.

Hieronder eerst de resultaten van het 10 layer LSTM model en dan het 3 layer LSTM model.

Average % per hour: 0.6742 %

Average % per trade: 0.0594 %

Average # trades per hour: 11.34

Average % per hour: 1.1511 %

Average % per trade: 0.0553 %

Average # trades per hour: 20.81

Je zou hier heel snel denken dat het 3 layer model betere resultaten zou halen omdat het per uur een hoger percentage winst haalt maar dit is eigenlijk een vertekent beeld. Hier zijn namelijk nog geen trading fees in verwerkt en als je dan kijkt naar het average percentage per trade ga je meer overhouden per trade bij het 10 layer model. We hebben ook langere testen uitgevoerd waarbij er wel fees werden verrekent en daarbij kan je dan duidelijk zien dat je % per trade hoog genoeg moet zijn en dat dit uiteindelijk een grotere rol speelt. Als dit niet het geval is gaan je positieve trades niet genoeg opbrengen om de occasionele negatieve trade op te vangen.

Als we hieronder ook even kijken naar een plot waarop je de sell en buy predictions ziet is het al snel duidelijk dat het model nog veel te leren heeft.

In het algemeen snapt het ongeveer wel dat het moet kopen na een downtrend en verkopen na een uptrend maar dit gebeurt vaak veel te snel. Voor dat de prijs voldoende gedaald of gestegen is word er vaak al een buy of sell uitgevoerd.

Figure 31: Model Predictions

4.2 Sterke en zwakkere punten

Sterke punten

Dankzij de kleine candle size en snelheid van het model kan deze op een zeer korte periode al winst maken en heeft deze geen enorm sterke computer nodig om het model te runnen. Het getrainde model gebruikt ongeveer 1 GB aan VRAM op een gpu maar er zal eerder richting 1.5GB nodig zijn als je de data meerekent.

De data die gebruikt is komt overeen met de data van andere exchanges, dus je zou deze bot op de Binance exchange kunnen gebruiken waarop deze getrained is maar ook op bijvoorbeeld Crypto.com en dit zou in theorie even goed moeten werken. Je moet echter wel rekening houden met het volume. Op Binance word vooral USDT en BUSD gebruikt en hier heb je een groter volume en dus een betere weergave van de aankopen en verkopen. Op Crypto.com daarintegen wordt er veel USDC gebruikt en is hier het volume hoger. Bij een te laag volume kan dit een negatieve impact hebben op de berekening van de indicators en op het neurale netwerk.

Zwakke punten

Het model maakt enorm veel kleine trades maar dit is nadelig in situaties met hoge trading fees omdat je dan veel verliest aan fees en dit moeilijk kan omhoog halen met de kleine winsten die er te halen zijn. Het model zou eigenlijk iets minder moeten aankopen en de aankopen die het wel doet op de lagere punten doen.

Dit is mogelijks te verhelpen door de minimum change bij het labellen van de data te verhogen waardoor het neuraal netwerk leert om grotere trades te doen in plaats van op elke kleine schommeling.

Vergeleken met Algoritmisch traden heb je hiervoor wel ook een NVIDIA gpu nodig dus er is een iets hogere kost om het model op te zetten voor constant gebruik.

4.3 Bruikbaarheid en implementatie

Het model is in huidige toestand wel bruikbaar maar na berekeningen van de uiteindelijke winst met fees ligt dit al heel wat lager en soms mogelijks zelfs negatief. Dit is afhankelijk van de coin waarop getest word en hoe hoog de fees liggen op het gekozen exchange. Dit is deels ook omdat het model vaak nog te vroeg aankoopt waardoor het heel wat potentiele winst laat liggen.

Ik denk echter dat dit zeker te verhelpen is via verbeteringen aan de layout van het model of meer training en dat dit dan een zeer goed model kan worden.

De implementatie van dit model is relatief gemakkelijk, er zijn veel crypto exchanges die publieke API’s aanbieden en ook python libraries waardoor je makkelijk dit model met tensorflow zou kunnen gebruiken op real-time data die je opvraagt via de api. Binance en Crypto.com zijn 2 van de bekendste crypto exchanges met een goede API.

4.4 Alternatieven

Een alternatief dat zeer vaak wordt toegepast op zowel aandelen en crypto is natuurlijk algoritmisch traden. Dit maakt meestal gebruik van indicators, heel vaak dezelfde die ons model kreeg als inputs, om dan hier van af te leiden of de markt in een down of up trend zit en te proberen voorspellen wanneer dit gaat omdraaien. Dit is zeker een goede optie voor een meer voorspelbare trading bot waarbij je beter snapt wat de bot juist doet en dit gemakkelijk zelf kan aanpassen. Ik heb dit zelf in het verleden ook reeds gebruikt en kan bevestigen dat dit relatief simpel is om te maken en wel degelijk winstgevend kan zijn.

4.5 Meerwaarde

Dit onderzoek zal een beperkte meerwaarde bieden aan de maatschappij maar kan voor economische redenen wel interessant zijn voor zowel individuen als bedrijven die hun kapitaal willen investeren maar zelf niet de tijd, kennis of zin hebben om manueel de markt te volgen en trades te maken wanneer nodig. Een gelijkaardig systeem kan zeker gebruikt worden voor iemand die aandelen wilt verhandelen.

Het is niet mogelijk om dit commercieel aan te bieden als een service of toch zeker niet waarbij er slechts 1 neuraal netwerk word gebruikt. Je zou dan met een enorm kapitaal aankopen en verkopen doen, afhankelijk van hoeveel mensen het gebruiken en wat hun kapitaal is natuurlijk. Dit zou dan voor prijs manipulatie zorgen, vooral bij coins met een kleinere market capitalisatie.
Fictief voorbeeld: de totale waarde van alle Dogecoins is 100000 euro en je koopt er voor 10000 euro, dat is 10% van het totale aanbod en dit zou voor een enorme stijging in prijs zorgen.
Dit zou negatieve effecten kunnen hebben op je eigen winsten.

Dit kan deels verholpen worden door het kapitaal te verdelen over een grote hoeveelheid coins maar als je geen limiet zet op je inzet kan dit negatieve gevolgen hebben.

4.6 Vervolgonderzoek

Het is vooral belangrijk om meer onderzoek te doen naar welke layers en neuron aantallen het beste werken en hoeveel training er juist nodig is. Zoals onderzoek naar meerdere soorten modellen reeds bewees is het niet altijd beter om een groter model te gebruiken maar kan een kleiner model vaak betere performance halen omdat deze beter opgebouwd of getrained zijn en soms omdat de data beter is.

Een deel dat ik zeker wil verkennen is Reinforcement Learning [10], ik ben reeds gestart met het maken van een eigen OpenAI gym environment voor crypto trading die te vinden is op Github. Ik ben er van overtuigd dat dit goede resultaten kan opleveren als ik mijn reward functies op punt krijg. Een Mede student Tuur Vanhoutte heeft reeds wat meer ervaring met Reinforcement learning vanuit zijn research project en heeft dit toegepast met LibTorch, de c++ versie van PyTorch. Hij ondervond enorme verbeteringen in training snelheid in c++ dus dit zou ik zeker ook verder willen verkennen.

Verder denk ik dat Explainable AI zeer belangrijk is voor elk onderzoek met neurale netwerken, jammer genoeg had ik hiervoor niet voldoende tijd om dit uit te werken maar ga dit zeker nog verder onderzoeken.

4.7 Feedback van externen

Ik heb aan een aantal mensen wat vragen gesteld en feedback gevraagd over mijn onderzoek. Ik zal hier de feedback overlopen van Nick Langens, Computer Science student aan de KU leuven dat ook onderzoek deed naar trading maar dan met reinforcement learning en machine learning in plaats van deep learning.
Yente De Wael, master student Computer Science aan de VUB, crypto trader in zijn vrije tijd en heeft ook ervaring met algoritmisch traden.

1. wat is uw ervaring met trading, zowel met en zonder AI?

Nick heeft zowel ervaring met manueel en algoritmisch traden en uit zijn ervaring heeft het nieuws soms een grotere effect op de prijzen dan de effectief observeerbare patronen die je kan afleiden uit prijsdata. Daarom is het moeilijk om een onderzoek uit te voeren op trading want het implementeren van real-time nieuws data is niet betrouwbaar en moeilijk te implementeren.

2. vind u dat ik mijn onderzoek correct heb aangepakt of had u dingen anders gedaan en indien ja, wat?

De aanpak met neurale netwerken vond hij vooral interessant maar hij mist een beetje de quantitatieve vergelijkingen met andere methodes zoals machine learning, algoritmisch traden en andere model soorten.

Ik vind dit zeker terechte feedback, het zou beter zijn als we wat meer vergelijkingen konden toevoegen maar voor algoritmisch en machine learning zou nog eens makkelijk en maand onderzoek nodig zijn. Zeker bij algoritmisch omdat je dan zelf manueel de parameters moet aanpassen van je indicators en beslissingen voor kopen en verkopen.

3. denkt u dat reinforcement learning betere resulaten zou opleveren dan supervised learning?

Volgens nick is dit moeilijk correct te doen omdat je model een zo realistisch mogelijke environment nodig heeft om effectief correct te leren wat te doen. Indien er fouten zitten in de environment, hoe klein dan ook, zal het model deze leren en misbruiken indien mogelijk. Het is mogelijk maar je moet goed opletten dat alles zo juist mogelijk is.

4. heeft u ervaring met indicators en indien ja, wat denkt u van de combinatie van gebruikte indicators en heeft u hier feedback op?

Nick heeft een beperkte ervaring met het gebruiken en combineren van indicators en had hier geen feedback over.

5. heeft u ervaring met LSTM modellen en wat denkt u van de manier dat ik LSTM heb toegevoegd?

Nick heeft niet genoeg ervaring hiermee om een oordeel te vellen over wat beter is in welke situatie.

5 Advies

5.1 Introductie

In dit onderdeel zou ik graag iedereen die wilt verder bouwen op dit onderzoek of een gelijkaardig onderzoek zou willen starten helpen. Alle dingen die ik anders zou doen als ik opnieuw zou beginnen en de dingen die ik zelf ook nog wil uitwerken om het resultaat te verbeteren zullen hier behandeld worden.

5.2 Risico

Automated trading komt natuurlijk met voor en nadelen. [11] het is belangrijk deze goed af te wegen en de mogelijke problemen of nadelen te minimaliseren.

Ik zou traden met echt geld enkel en alleen doen als je je bewust bent van alle risico's. Er bestaat zeker een kans dat je al je inzet verliest. Je kan dit natuurlijk ook zonder geld onderzoeken op historische data of real time maar dan zonder de effectieve trades uit te voeren, hiervoor kan je gebruik maken van Binance met een Demo account of Alpaca API hun Paper Trading api.

5.3 Voor Wie Is Dit?

Voor AI onderzoekers, hobby traders of eender wie dat zich wil verdiepen in zowel trading en neurale netwerken. Een voorkennis van trading is aangeraden, zeker op vlak van indicators maar dit is zeker niet moeilijk aan te leren en er is veel info te vinden online over hoe je juist indicators moet gebruiken.

5.4 Model

LSTM modellen zijn het meest populair bij trading modellen en dit zou ik ook aanraden maar het is zeker mogelijk dit probleem op andere manieren aan te pakken. Wij hebben zeker geen optimaal model in dit onderzoek en ik raad aan zelf te testen wat best werkt en wat niet.

5.5 Data

Voor het verzamelen van data raad ik de Binance Exchange api aan. deze is stabiel, snel, gratis te gebruiken en omdat Binance al een aantal jaar bestaat vind je hier voldoende historische data voor een grote hoeveelheid coins.

De indicators de je wil toevoegen, of niet toevoegen, is volledig te kiezen. Ik koos voor indicators die ik reeds ken en gebruikt heb omdat ik weet dat deze een goed beeld kunnen geven op wanneer je zou moeten kopen en verkopen. Zonder indicators zou het model het veel moeilijker kunnen hebben met het voorspellen van correcte buy en sell targets.

5.6 Aanbevelingen

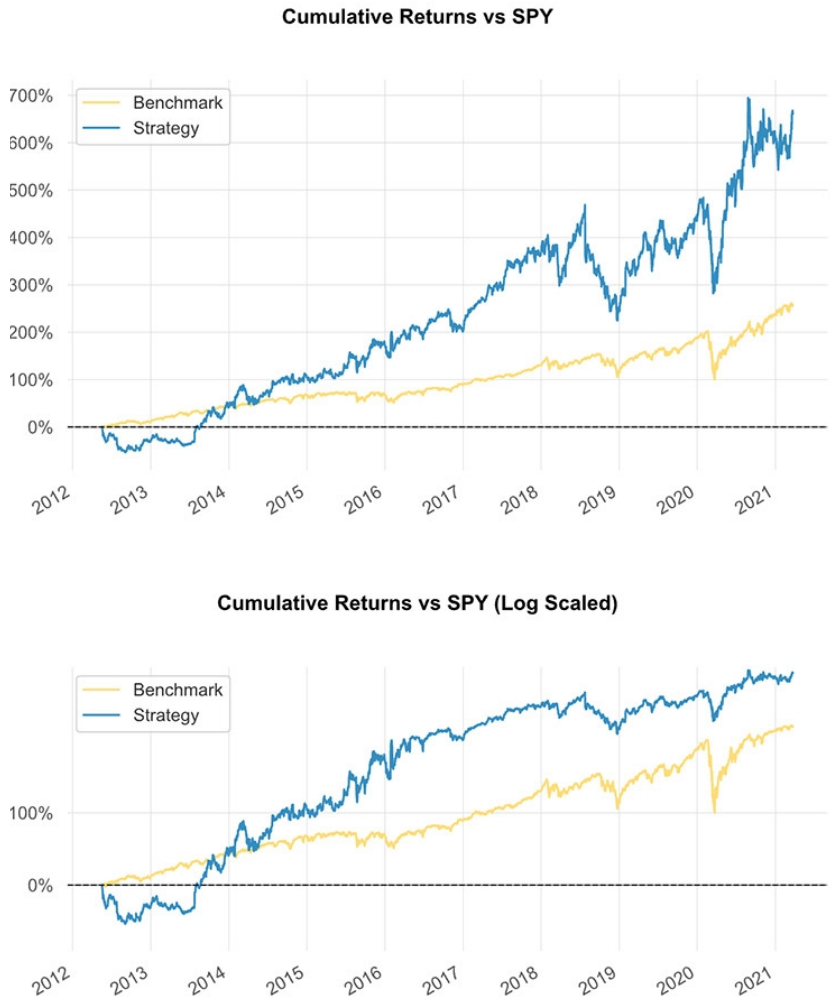
Indien mogelijk zou ik reinforcement learning willen aanraden. De problemen die zich voordoen door labeled training toe te passen kunnen hiermee verholpen worden. Hier is natuurlijk wel wat meer tijd voor nodig voor de training.

Een van de belangrijkste uitbreidingen zou Explainable AI zijn. Er werd eerder al vermeld waarom dit zo belangrijk is. Dit was nog niet toegepast in dit onderzoek waardoor we eigenlijk meer gokte naar en correcte model layout en welke data te gebruiken en dit is niet aan te raden. Door een vorm van Explainable AI toe te passen zou je indicators die niet relevant zijn kunnen laten vallen of andere indicators toevoegen die wel een meerwaarde bieden.

Indien je effectief zou willen traden met je ontwikkeld neuraal netwerk raad ik het Binance exchange platform aan, deze heeft de laagste fees voor kleine trades.
Indien je een groot kapitaal hebt en dus een redelijk hoog maandelijks trade volume kan behalen is het crypto.com platform mogelijks beter omdat je hierop minder fees betaald als je trade volume hoger ligt.

Om testing te verbeteren en duidelijkere vergelijkingen te kunnen maken tussen modellen raad ik de library Quantstats aan. Hiermee kan je de performance van je model veel beter onderzoeken vanwege de grote lijst aan grafieken en parameters die voor je berekent worden. Zo word onder andere de drawdown, expected returns, profit factor en nog veel andere dingen berekent en weergegeven op grafieken.

Hieronder bevind zich een voorbeeld van een deel van een Quantstats report.



Key Performance Metrics

Metric	Strategy	Benchmark
Risk-Free Rate	0.0%	0.0%
Time in Market	100.0%	100.0%
Cumulative Return	660.21%	255.34%
CAGR%	25.75%	15.4%
Sharpe	0.8	0.95
Sortino	1.21	1.32
Sortino/√2	0.86	0.93
Max Drawdown	-53.62%	-33.72%
Longest DD Days	532	272
Volatility (ann.)	37.32%	16.6%
R^2	0.23	0.23
Calmar	0.48	0.46
Skew	0.95	-0.7
Kurtosis	18.55	17.87
Expected Daily %	0.09%	0.06%
Expected Monthly %	1.91%	1.19%
Expected Yearly %	22.49%	13.52%
Kelly Criterion	4.86%	6.4%
Risk of Ruin	0.0%	0.0%
Daily Value-at-Risk	-3.75%	-1.66%
Expected Shortfall (cVaR)	-3.75%	-1.66%
Gain/Pain Ratio	0.17	0.21
Gain/Pain (1M)	0.97	1.55

Figure 32: Quantstats report

5.7 Tips

Indien mogelijk raad ik het zeker aan je data in binair formaat op te slaan omdat dit de lees en schrijf snelheid verbeterd alsook de hoeveelheid opslag verminderd.

Automatiseer best je training en testing zodat dit minder manueel werk vergt, dit had ik niet gedaan van het begin en hierdoor was het vaak manueel stukjes code aanpassen om het juiste model te testen etc. en test ook zeker op voldoende data, bepaalde coins reageren volledig anders omdat deze minder of meer volatiel zijn of eerder up of down trends hebben.

Piramidding is een simpele toevoeging in het geval dat je echt wil gaan traden en kan je winst enorm vergroten, dit mag zeker niet vergeten worden.

En tot slot, focus je niet op totale winst als je nog geen fees in je berekeningen hebt staan. Focus je dan eerder op winst per trade. Enkel als je fees er ook in zitten is je totale wint percentage realistisch.

6 Conclusie

Evaluatiecriteria van dit hoofdstuk:

Onvoldoende: het besluit bevat geen antwoord op de onderzoeksvraag, is weinig zeggend of bevat plots nieuwe (niet onderbouwde) informatie.
Beperkt: het besluit bevat enkel een antwoord op de onderzoeksvraag zonder daarbij de belangrijkste zaken uit reflectie en advies daarbij te betrekken.
Volstaat: de onderzoeksvraag wordt correct beantwoord waarbij duidelijk verwezen wordt naar informatie uit de onderdelen reflectie en/of advies.
Goed: de belangrijkste elementen uit de voorbije onderdelen worden kernachtig samengevat. Van daaruit wordt tenslotte de onderzoeksvraag beantwoord.
Excellent: naast het onderbouwd beantwoorden van de onderzoeksvraag wordt de lezer getriggerd om zelf verder onderzoek over het thema te voeren. Suggesties worden hierbij aangeleverd.

Dit onderzoek was zeer interessant, zowel om te zien wat AI kan op vlak van time-series prediction in een zeer onvoorspelbare omgeving maar ook vooral omdat ik zelf wel geinteresseerd ben in crypto trading. Door de beperkte tijd van dit onderzoek heb ik het nog niet volledig af kunnen krijgen zoals ik gewild had maar ik ga hier zeker op verder bouwen na deze bachelorproef. Tijdens het onderzoek zijn er redelijk wat problemen en mogelijke verbeteringen duidelijk geworden en kan ik deze wat bijsturen om een nog beter model te kunnen maken dat hopelijk effectief ingezet kan worden in een echt trading platform. Desondanks dat het winstgevend zou zijn is het nog net niet betrouwbaar genoeg om in te zetten met echt geld, in een langdurige downtrend zou het wel eens genoeg grote fouten kunnen maken waardoor je mogelijks heel wat geld verliest. Het is dus zeer belangrijk dat het model beter leert hoe het moet reageren in een downtrend voor dat het echt ingezet kan worden.

Dus om eens terug te komen op de onderzoeksvraag. Wat is het meest geschikte AI model voor het forecasten van de koers van cryptomunten aan de hand van open source data?

Een LSTM netwerk lijkt hier de beste optie, de open source data moet wel wat uitgebreid worden met wat indicator berekeningen maar dit is relatief simpel door Ta-Lib. En hoewel een LSTM netwerk hiervoor wel goed geschikt is, is labelled training dat niet. Unsupervised learning lijkt mij de volgende stap naar een beter resultaat.

Ik raad iedereen met wat interesse in trading en AI zeker aan om deze manier van trading te verkennen, of algoritmisch traden. Niet alleen voor de mogelijkse winst maar voor de kennis die je onderweg bijleert. Zeker aangezien je geld op een spaarrekening niets opbrengt en erger nog, waarde verliest is het interessant om nieuwe manieren van investeren te onderzoeken.

7 Literatuurlijst

[1] Pedro Lara-Ben´ıtez, Manuel Carranza-Garc´ıa, Jos´e C. Riquelme, 2021, An Experimental Review on Deep Learning Architectures for Time Series Forecasting. Available: https://arxiv.org/pdf/2103.12057.pdf

[2] Jingyi Shen & M. Omair Shafiq, 28 Aug, 2020, Short-term stock market price trend prediction using a comprehensive deep learning system. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6

[3] Armando Vieira, Sep 29, 2019, Trading Through Reinforcement Learning using LSTM Neural Networks, Available: https://medium.com/@Lidinwise/trading-through-reinforcement-learning-using-lstm-neural-networks-6ffbb1f5e4a5

[4] Bruce Yang, Aug 25, 2020, Deep Reinforcement Learning for Automated Stock Trading, Available: https://towardsdatascience.com/deep-reinforcement-learning-for-automated-stock-trading-f1dad0126a02

[5] Rian Dolphin, Oct 21, 2020, LSTM Networks | A Detailed Explanation, Available: https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9

[6] Team Choice, Dec 21, 2021, Best Combination of Technical Indicators for Intraday Trading, Available: https://choiceindia.com/blog/best-combination-of-technical-indicators-for-intraday-trading/

[7] Alberto Romero, A New AI Trend: Chinchilla (70B) Greatly Outperforms GPT-3 (175B) and Gopher (280B), Available: https://towardsdatascience.com/a-new-ai-trend-chinchilla-70b-greatly-outperforms-gpt-3-175b-and-gopher-280b-408b9b4510

[8] Edd Gent, DeepMind's New AI With a Memory Outperforms Algorithms 25 Times Its Size, Available: https://singularityhub.com/2021/12/20/biggers-not-always-better-deepminds-new-language-ai-is-small-but-mighty/

[9] TicTacTec LLC, Ta-lib, Available: https://www.ta-lib.org

[10] Daniel Johnson, Reinforcement Learning: What is, Algorithms, Types & Examples, Available: https://www.guru99.com/reinforcement-learning-tutorial.html

[11] Jean Folger, 4 march, 2021, Automated Trading Systems: The Pros and Cons, Available: https://www.investopedia.com/articles/trading/11/automated-trading-systems.asp

[12] Jason Fernando, Moving Average Convergence Divengence (MACD), Available: https://www.investopedia.com/terms/m/macd.asp

[13] Jason Fernando, Relative Strength Index (RSI), Available: https://www.investopedia.com/terms/r/rsi.asp

[14] Cory Mitchell, Money Flow Index – MFI Definition and Uses, Available: https://www.investopedia.com/terms/m/mfi.asp

[15] James Chen, Chaikin Oscillator, Available:
https://www.investopedia.com/terms/c/chaikinoscillator.asp

[16] Adam Hayes, Bollinger Band®, Available:
https://www.investopedia.com/terms/b/bollingerbands.asp

[17] Adam Hayes, Average True Range (ATR), Available:
https://www.investopedia.com/terms/a/atr.asp

[18] Ben Dickson, June 15, 2020, The cas for self-explainable AI,
https://bdtechtalks.com/2020/06/15/self-explainable-artificial-intelligence/

8 Bijlages

In jouw bachelorproef zelf staan enkel kernzaken. Veel documenten die je wel gebruikt hebt, maar niet direct in jouw bachelorproef hoeven te staan, voeg je als bijlage toe.

Indien documenten bijdragen aan jouw onderzoek moet je ze opnemen in de bijlage, zodat men kan controleren hoe je onderzoek is uitgevoerd en waar het op is gebaseerd. Veel voorkomende bijlageonderdelen zijn: interview(vragen), tabellen en analyses, gedetailleerde technische gegevens, code, enz.

In dit onderdeel voeg je ook

- jouw verslag van bijgewoonde sessies uit de module researchproject toe;
- jouw handleidingen uit de module researchproject (installatiehandleiding & gebruikershandleiding).

8.1 Verslag Computer Crime unit

Dinsdag 11 January 2022.

Francis Nolf van de federale politie – computer crime unit kwam een presentatie geven over hoe ze te werk gaan en wat ze zoal doen om computer crime tegen te gaan. De CCU is ontstaan in 2001 na het samenvoegen van 2 aparte teams na de samenvoeging van alle politie departementen. De CCU is opgedeeld in 2 delen. De federale CCU en de regionale CCU. Deze hebben onder zich nog verschillende kleinere teams gespecialiseerd in bepaalde dingen.
Ik zal even verder uitleggen wat de verschillende teams doen.

Zo is er bijvoorbeeld een OSINT team (Open Source Intelligence) dat zich focust op het verzamelen en verwerken van publieke informatie dat beschikbaar is op het internet. Onder andere een bepaalde locatie onderzoeken via Google maps voordat er een interventie op deze locatie gebeurd zodat ze zich zo goed mogelijk kunnen voorbereiden hierop zonder ter plaatsen te moeten gaan rondkijken en zo mogelijks hun geplande interventie verklappen.

Het volgende team specialiseert zich in het afluisteren van telefoons en het decrypten van deze communicatie tussen 2 personen. Zo zal dit team ook proberen de locatie van onder andere drugs dealers te ontdekken gebaseerd op de communicatie patronen van de smartphones.

Er is ook een team dat onderzoek doet naar alles wat te maken heeft met data opslag op allerlei soorten apparaten. De meest voor de hand liggende zijn dan smartphones, computers, etc. Maar er zijn heel veel apparaten die iets minder voor de hand liggen dat ook vaak gebruikt worden zoals spelconsoles zoals een XBOX, routers, access points, etc. Elk apparaat dat data kan bevatten kan nuttig zijn in een onderzoek.

Het laatste en ook nieuwste team is verantwoordelijke voor Hacking. Dit team is nog maar recent in werking getreden na een verandering in de wet waardoor het nu toegestaan is dat de federale politie effectief hacking en exploiting kan gaan toepassen om informatie te verzamelen van verdachten. Dit word meestal gedaan met hacking specifieke besturingssystemen zoals Kali Linux. Dit besturingssysteem komt met een hele reeks aan programma's specifiek om te hacken.

Na een introductie van de CCU ging de presentatie verder in op de details van hoe onderzoeken gevoerd worden. Een van de voorbeelden bijvoorbeeld ging over de manier van omgaan met apparaten waarvan de waarde van groot belang is. Ze proberen geen schade aan te brengen aan het originele apparaat en gaan dus ook niet de originele harde schijf gebruiken maar nemen een schijf kopie die bit per bit een exacte kopie is van de originele schijf. Hier worden ook hashes van genomen om later te kunnen aantonen dat deze wel degelijk exact hetzelfde zijn en niet aangepast.

Verder ging het ook over de tools die ze gebruiken, zowel publieke en private tools, binnen de ccu voor de extractie van data van apparaten. Er zijn veel programma's reeds ingebouwd in Linux zoals DD, DCFLDD,.. om drives te kopiëren op byte-level. Er komen natuurlijk ook veel wetten kijken bij alles rond data verzameling en hoelang ze deze moeten bijhouden. Deze zijn zeer strikt en hier mag geen fout gemaakt worden. Er is ook een groot verschil tussen de wetten omtrent data verzameling en opslag over verschillende landen. In België zal de data na een onderzoek voor 6 maanden lang bijgehouden worden om later terug op te komen in verder onderzoek, maar in Duitsland is het verboden om data bij te houden als het onderzoek is afgerond.

Het is ook enkel toegestaan deze data op te vragen als er toestemming word gegeven door een procureur of magistraat. Onder geen enkele andere voorwaarde mag de data opgevraagd worden.

Een sysadmin van een bedrijf kan ook als medeplichtige beschouwd worden als deze weigert mee te werken met de politie in een onderzoek waarbij er mogelijks data nodig is van bedrijfsservers.

Naast de Belgische wetten zijn er ook internationale wetten waardoor het opvragen van data over buitenlandse verdachten verplicht via het ministerie van buitenlandse zaken moet opgevraagd worden en dit kan heel lang duren van weken tot jaren. Er zijn bepaalde uitzonderingen zoals facebook, outlook,…
deze bedrijven bieden rechtstreeks contact met interne teams voor het opvragen van informatie. Het is nog steeds nodig dat dit goedgekeurd word door een procureur of magistraat.

Na de uitleg over het gebruik en opvragen van data was er ook een belangrijk deel over het in beslag nemen van fysieke apparaten. Het is enkel voor de politie toegestaan om apparaten in beslag te nemen en zelfs binnen de politie enkel degene met een hogere rang. Na het in beslag nemen van een apparaat proberen ze alle signalen te blokkeren om het apparaat in een oorspronkelijke staat te bewaren. Dit is moeilijk omdat veel apparaten altijd een soort van verbinding hebben zoals 4G, bluetooth, cell service, etc. en om hier tegen in te gaan gebruiken ze speciale zakken die werken als een kooi van Faraday. Dit is omdat bewijs dat gevonden word anders ongeldig beschouwd kan worden als het apparaat verbonden geraakt met het internet omdat er dan mogelijks aanpassingen aan zijn gebeurd.

Voor bepaalde apps zoals messenger en whatsapp moet er weer toegang verleend worden door een magistraat en enkel als de verdachte toestemming geeft hiervoor.

Ook het afluisteren van apparaten moet toegestaan worden door een magistraat en elk afgeluisterd bericht moet een rapport van op papier gezet worden en naar de magistraat gestuurd worden elke 5 dagen. Daarbovenop kan het soms gebeuren dat een advocaat vraagt om elk gesprek (niet alleen die dat als relevant worden beschouwd voor het onderzoek) uit te typen.

Het laatste deel van de sessie ging over de bedreigingen die momenteel populair zijn en dan gaat het bijvoorbeeld over ransomware, hierbij word je data op je apparaat geëncrypteerd en moet je betalen om deze terug te laten decrypten. Recent worden er ook heel veel scams gedaan, in de meeste gevallen doen de scammers zich voor als iemand anders of een bedrijf.

Deze zaken zijn moeilijk op te lossen omdat deze scammers meestal uit andere landen komen wat het onderzoek moeilijk maakt door de vele regels opgelegd door de wet.

Bij deze scams horen vaak ‘money mules’, ze proberen je te doen geloven dat je geld kan verdienen zonder echt te moeten werken, dit is een tussenpersoon waarnaar het geld initieel verstuurd word, deze houden dan ook een deel van het geld en storten de rest weer verder door naar een andere rekening om het moeilijker te maken om te traceren naar wie het geld eigenlijk gaat.

Het gevaar hierbij is dus dat deze money mules, dat vaak eigenlijk ook mensen zijn die slachtoffer zijn van een scam want ze weten vaak niet welk geld ze eigenlijk doorstorten, ook medeplichtig beschouwd worden voor de wet en dus ook een straf kunnen krijgen.

Tot slot werden er nog een paar problemen besproken die het moeilijk maken om deze misdaden op te lossen. Onderzoekers mogen bijvoorbeeld geen acties van een verdachte uitlokken omdat dit ongeldig word beschouwd in een rechtszaak. Een 2de probleem is dat peer-to-peer netwerken niet onderzocht mogen worden en dus het delen van bestanden en dergelijke niet onderzocht kan worden. Ook het toestaan van legal hacking, dus waarbij de onderzoekers zelf de verdachte gaan proberen hacken om te onderzoeken, zou het onderzoek helpen.

De presentatie van Francis Nolf was zeer informatief en leuk gebracht en zo zien we eens hoe cyber crime word aangepakt in België.

8.2 Handleiding Researchproject