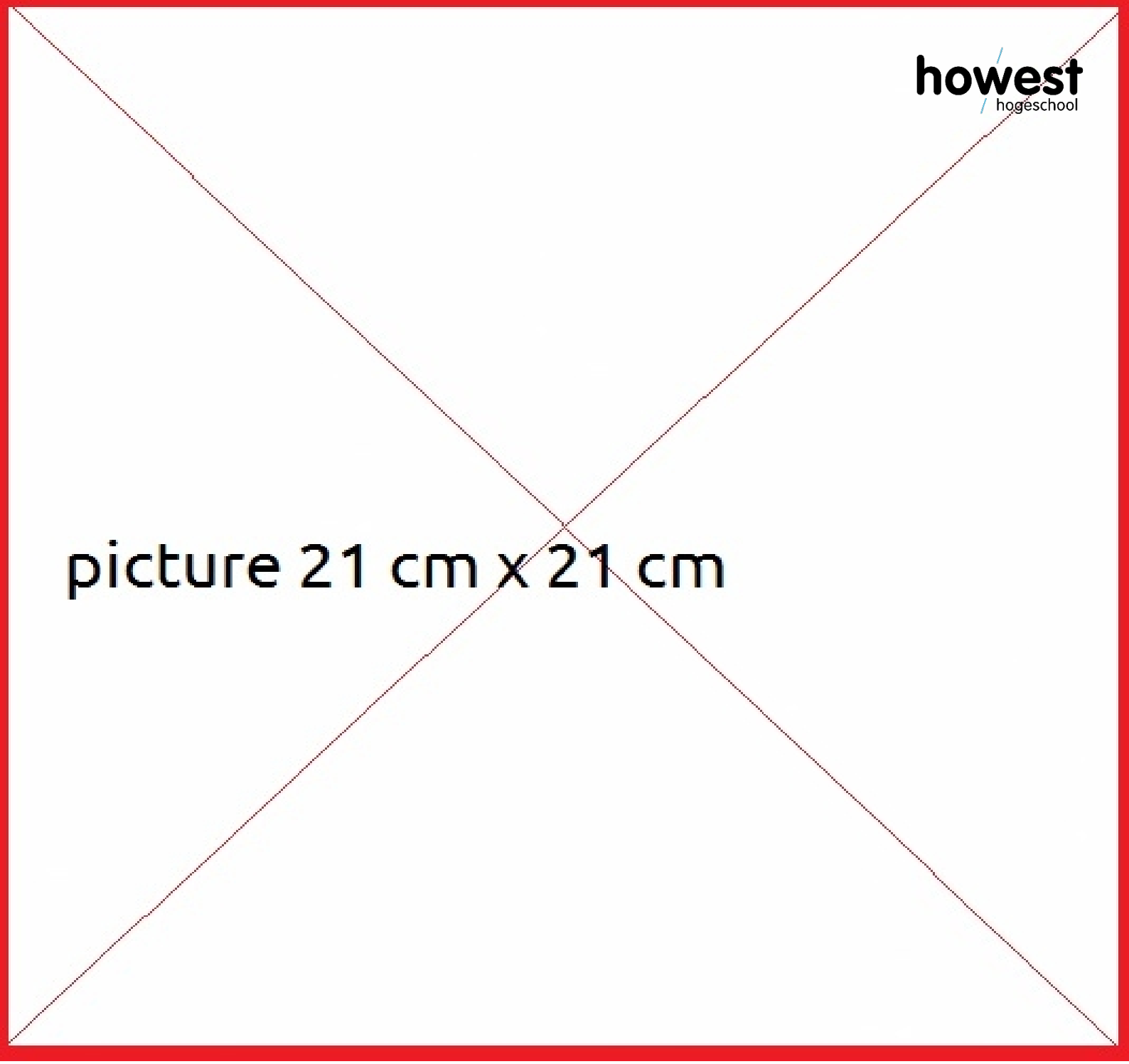

BACHELORPROEF

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

Het woord vooraf omschrijft kort het opzet van de bachelorproef, zonder sterk in detail te gaan. De context van Deze bachelorproef is mijn laatste werk om mijn opleiding Multimedia & Communication Technology met keuzetraject Artificial Intelligence Engineer aan HoWest af te sluiten. 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 een goede 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 die 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 zou graag even 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 welk 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 candle 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 is opgehaald geweest 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

Doel: In de inleiding beschrijf je ook hoe jouw bachelorproef in elkaar steekt. Een krachtige heldere inleiding zorgt ervoor dat je de lezer voor je wint en hij/zij sneller de rest van jouw document zal gaan lezen.

In de inleiding introduceer je de onderzoeksvraag. Je vermeldt de achtergrond of bestaande situatie. Je licht toe waarom de onderzoeksvraag voor jou/jouw stagebedrijf relevant is. Ook eventuele deelvragen worden nauwgezet omschreven.

De inleiding omschrijft ook de gebruikte onderzoeksmethode. Je legt uit waar, wanneer, met wie en hoe je het onderzoek gaat doen.

Je kunt alvast gebruikmaken van één of meerdere standaardzinnen:
- De data voor dit onderzoek zijn verzameld door…
- Vijf stukken worden onderzocht, die allemaal…
- De onderzoeksgegevens in deze bachelorproef zijn afkomstig uit vier belangrijke bronnen, namelijk…
- Door kwalitatieve methoden te gebruiken probeer ik… uiteen te zetten/uit te lichten.
- De studie is uitgevoerd in de vorm van een enquête, waarbij data zijn verzameld via…
- De methode die in deze studie gebruikt is, is een gemengde aanpak gebaseerd op…

Evaluatiecriteria van dit hoofdstuk:

Onvoldoende: de inleiding geeft géén goed beeld weer waarover deze bachelorproef precies gaat. De onderzoeksvraag komt te weinig of niet naar voor.
Beperkt: de onderzoeksvraag wordt geformuleerd zonder voldoende situering. Het nodigt de lezer niet verder uit om echt verder te lezen.
Volstaat: de inleiding omschrijft de onderzoeksvraag en schetst de context. De opbouw van de bachelorproef is heel beperkt.
Goed: vanuit een concrete probleemstelling wordt de onderzoeksvraag voorgesteld. Het onderzoek wordt kort besproken. De structuur van de bachelorproef krijgt eveneens aandacht
Excellent: de interesse van de lezer wordt onmiddellijk gewekt door te vertrekken vanuit een zinvolle achtergrondschets. Van daaruit wordt de probleemstelling afgeleid en wordt zo de onderzoeksvraag voorgesteld. De student benadrukt ook de nood aan onderzoek. De structuur van de bachelorproef wordt tenslotte toegelicht.

1.1 aanleiding en inspiratie

Het idee van een neuraal netwerk dat kon voorspellen wanneer je best kon kopen en verkopen op de crypto markt of op de aandelen markt, leek enorm interessant en natuurlijk ook zeker financieel gezien aantrekkelijk. De dag van vandaag 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 maar dat deze data aan een bepaald algoritme word gegeven en deze zal dan een buy, sell of hold target teruggeven. Het is niet duidelijk hoeveel trades algoritmisch gebeuren maar afhankelijk van de bron nen 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.

1.2 Keuzes

De reden dat we dit onderzoek niet op de aandelenmarkt maar crypto markt doen is eigenlijk heel simpel. 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 voodelig 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.

1.3 Doelen

Het voornaamste doel van dit onderzoek is een werkend model maken die 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

In dit onderdeel probeer je alle 'theoretische' deelvragen te beantwoorden. Deze deelvragen kun je haast altijd beantwoorden door middel van jouw gevoerde literatuurstudie. Voor elke vraag maak je een aparte paragraaf aan. Vergeet zeker en vast bronvermelding niet!

De situering van de gebruikte technologie(en) mag ook niet ontbreken. Opgelet, jouw publiek zijn IT-experten. Basiszaken voor niet IT'er horen hier niet thuis, wel nieuwe onderzochte zaken die buiten het curriculum van MCT vallen.

Evaluatiecriteria van dit hoofdstuk:

Onvoldoende: Essentiële theoretische onderdelen worden niet besproken. Geen of foutieve bronvermelding.
Beperkt: de theoretische onderbouw van de bachelorproef is te licht. Technologieën worden niet voldoende toegelicht. Ook de kwaliteit van de bronnen is onvoldoende.
Volstaat: beperkt aantal bronnen werd geraadpleegd. Alle theoretische deelvragen worden correct beantwoord. De lezer heeft voldoende achtergrondinformatie.
Goed: er werden diverse bronnen geraadpleegd zoals (boeken, websites, podcast, online cursussen, ...). Research geeft goed beeld van de actuele beschikbare technologie. Volledige & correcte bronvermelding laten toe bronnen te verifiëren.
Excellent: research is van hoog niveau: verschillende bronnen werden vergeleken en op kwaliteit beoordeeld.

2.1 data

Het eerste dat we nodig hebben voor we kunnen beginnen met voorspellingen doen is de LOHC data, ook wel candles genoemt. 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 verhandelt word.

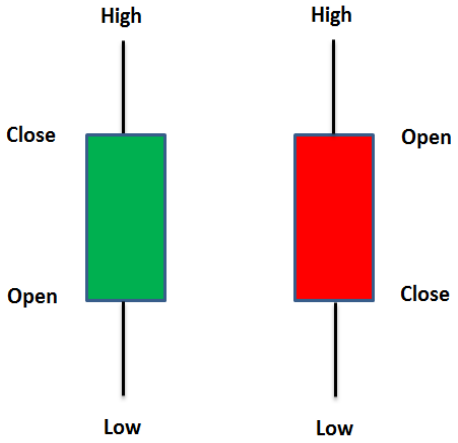

Figure 1: Candle

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 bevoorbeeld 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 word achteraf toegevoegd door onze data door een c++ programma te runnen die dan targets gaat zetten voor buy, sell en hold. Dit programma zal in de bijlagen te vinden zijn.

Een andere optie naast labelling was reinforcement learning maar dit had 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 word 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 begonnen met het ontwerpen van een neuraal netwerk waar we onze voorspellingen mee willen uitvoeren. Na wat onderzoek naar wat anderen al geprobeerd hebben voor voorspellingen op time series data en dan zeker stock en crypto werd het al snel duidelijk dat LSTM de enige optie was. 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 dan 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 ons buy, sell en hold target. De frame size, het aantal datapunten die je meegeeft om een prediction uit te voeren, was 240 candles.

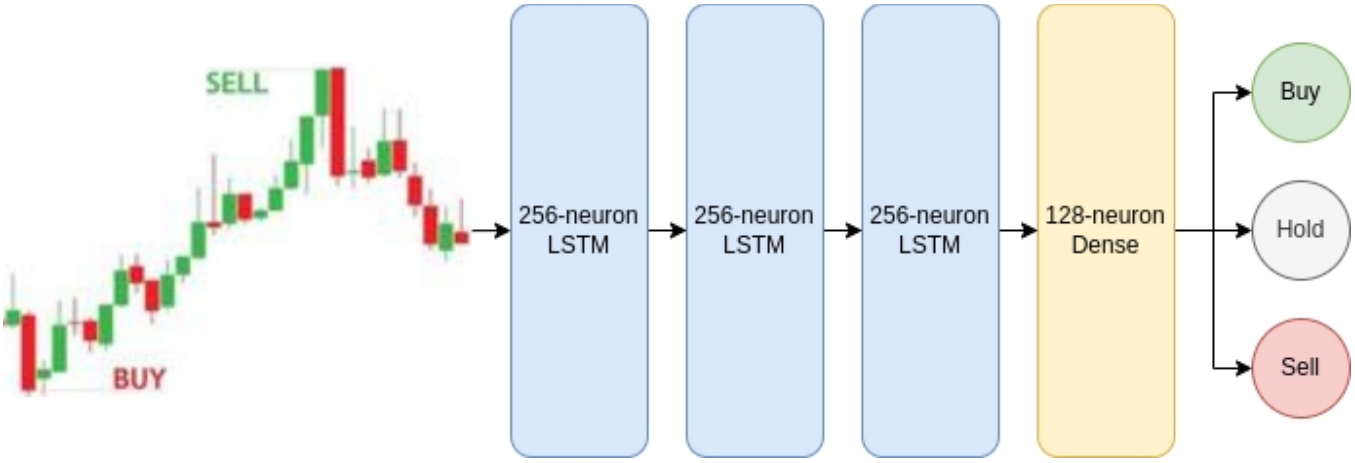

Figure 2: initial model layout

Het was echter snel duidelijk dat het model uit enkel prijsdata geen goede voorspellingen kon doen 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. Ik zal nu eerst dieper ingaan op wat indicatoren juist zijn, wat het nut er van is en hoe deze berekend worden.

2.3 Indicators

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 gaan 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. 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, ookwel 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 cummulative indicator dat 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 vind u de formule voor het berekenen van de A/D oscillator

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

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

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

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

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 3: 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 in deze indicator.

2.2.2 Average True Range

Figure 4: A/D oscillator example

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 vind u de formule voor het berekenen van de ATR, `Cp` staat voor Previous Close

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

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

where:

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

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

Figure 5: 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 6: ATR example

2.2.3 Bollinger Bands

Bollinger bands is ook een volatiliteit indicator maar word 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 word 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 word 2 gebruikt. Een breakout buiten de bands is meestal een duidelijk teken van hoge volatiliteit en word meestal als een duidelijk signaal gezien om te kopen of verkopen.

Hieronden vind u 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.

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

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

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 7: Bollinger Bands Formula

hier 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 8: Bollinger Bands Example

2.2.4 Moving Average Convergence Divergence

De MACD is een trend-following momentum indicator dat de relatie tussen 2 moving averages van een verschillende lengte weergeeft. De MACD word meestal gebruikt met exponential moving averages (EMA) maar andere types kunnen zeker ook gebruikt worden. Een EMA houd 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 word deze toch gebruikt om trend reversals te voorspellen.

Hieronder vind u de formule voor de MACD.

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

Figure 9: 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 word gezien als een bullish of bearish crossover afhankelijk van of de blauwe naar boven of naar beneden door de oranje lijn gaat.

Figure 10: MACD example

2.2.5 Money Flow Index

De MFI is een oscillator dat 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 word. Deze indicator word 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 word 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 kan u de formule vinden van de MFI.

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

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 11: 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 word.

Figure 12: MFI example

2.2.6 Relative Strength Index

De RSI is een momentum indicator dat ook word 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 bevind zich steeds tussen 0 en 100 en de signalen worden vooral gebruikt als deze boven 80 of onder 20 gaat.

Hieronder vind u de formule voor de RSI.

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

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

Figure 13: 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 14: RSI example

2.4 Training

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.

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 ookal ziet dit er wel goed uit. 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

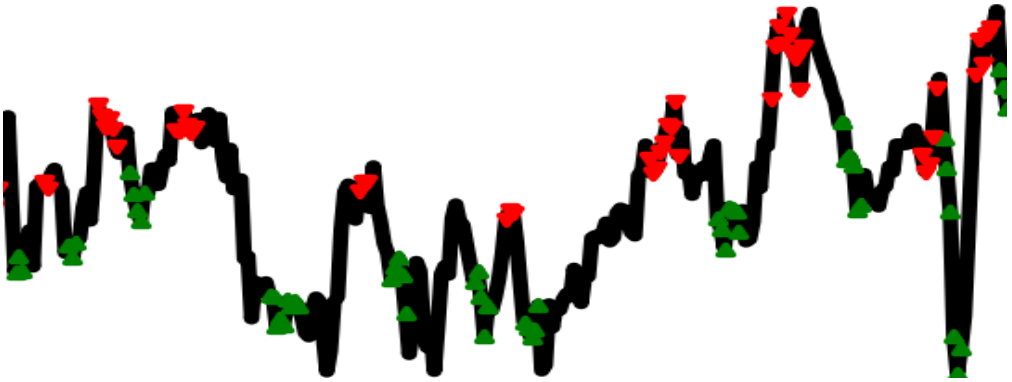

Figure 15: Buy & sell example

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

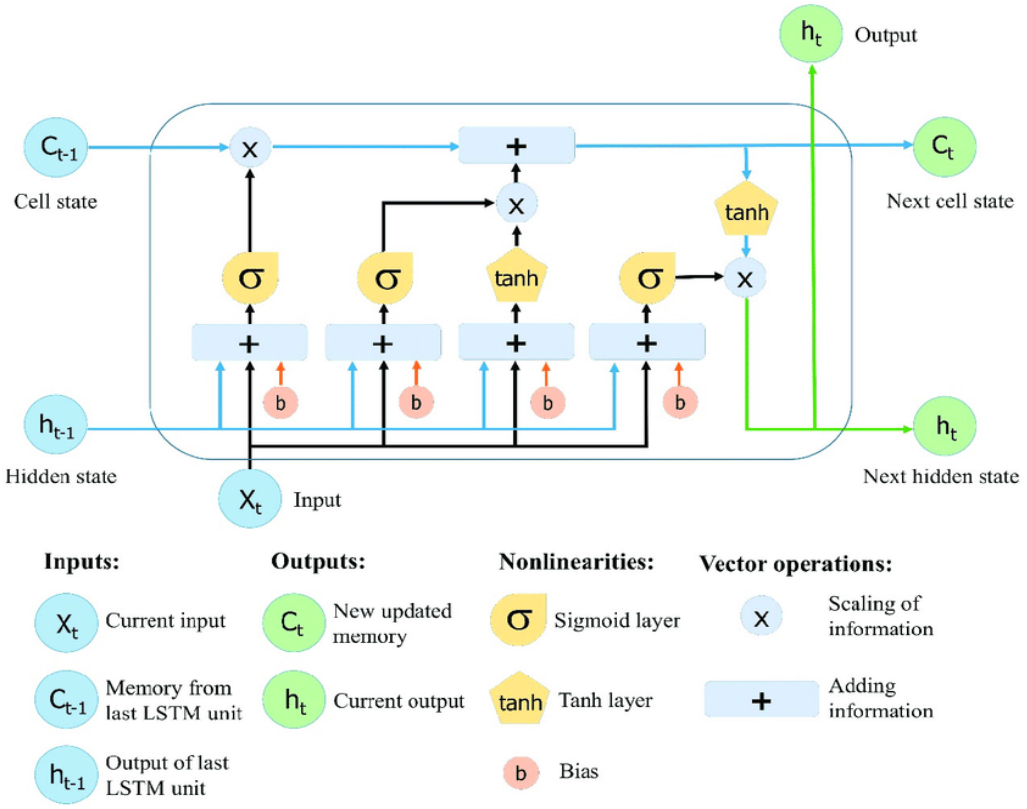

Figure 16: LSTM neuron structure

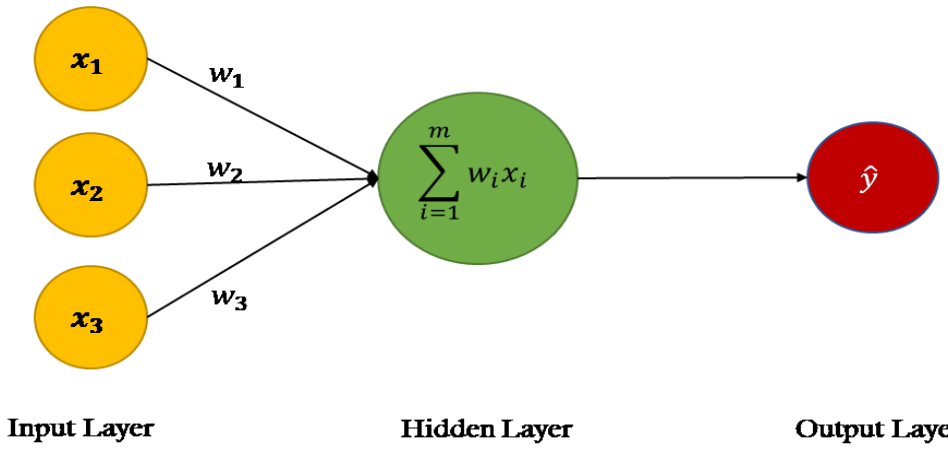

Figure 17: Simple Neuron Structure

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.

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 predicties 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.

2.6 Extra verbeteringen

Piramidding

Reinforcement Learning

Automation van training & testing

3 Technisch onderzoek

In dit hoofdstuk beschrijf je jouw gevoerde onderzoek uit de module Researchproject. Dit hoofdstuk mag verder onderverdeeld worden.

Volgende zaken worden verwacht:
- Beschrijving van de ontwikkelde software
- Opbouw applicatie: structuur/opbouw van het resultaat
- Achterliggende technologieën
- Overwonnen moeilijkheden/problemen
- …

Gebruik waar mogelijk visuele elementen: afbeeldingen/figuren/tabellen, zodat jouw tekst vlot leesbaar blijft.

Evaluatiecriteria van dit hoofdstuk:

Onvoldoende: de lezer kan moeilijk of niet achterhalen wat de student precies technisch gerealiseerd heeft.
Beperkt: het resultaat wordt beschreven; toch blijven heel wat zaken onduidelijk.
Volstaat: duidelijke omschrijving van de geleverde onderzoeksresultaten met minimum aan motivering.
Goed: dit hoofdstuk geeft naast de opbouw van het resultaat ook de verschillende keuzes met motivering weer. Gekozen technologieën en methodologieën worden eveneens overzichtelijk toegelicht.
Excellent: dit onderdeel laat de lezer op een vlotte manier toe het ganse onderzoeksproces te reconstrueren. Een kritische analyse van tussentijdse resultaten toont aan dat het onderzoeksproces continue bijgestuurd werd.

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++)
- Matplotlib
- python-binance (versie 1.0.12)

3.2 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 berekent 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 ziet u een voorbeld van hoe deze targets gezet worden

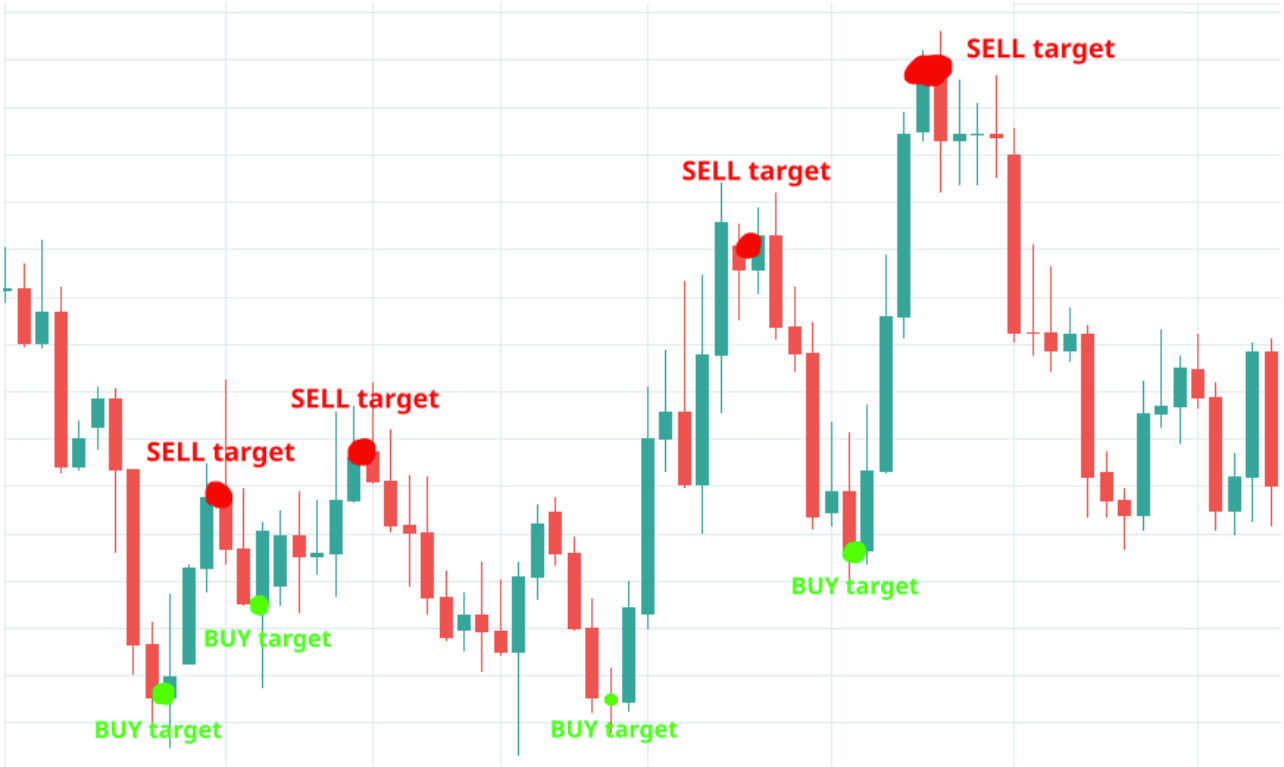

Figure 18: labelling example

Indicators

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

Normalization

Een heel belangrijk deel van data processing bij neurale netwerken is het correct normaliseren van de data. Als de data niet goed genormaliseerd is kan je model moeilijker de correcte weights en biases vinden en is het mogelijk dat je met exploderende gradients te maken hebt bij heel grote of heel kleine getallen.

De meeste waardes normaliseerde we door het procentuele verschil met de vorige candle te berekenen. Bij de MFI en RSI indicators deelden we deze gewoon door 100 omdat deze al op een schaal van 0 tot 100 staan wat ideaal is voor neurale netwerken.

3.3 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.

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 beter dan verwacht maar 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 onze korte onderzoeksperiode en training. Bij een willekeurige coin haalde het model op ongeveer 2 maanden een totale winst van 262%, zo een grote winst op een korte periode ligt aan de volatiliteit van crypto, hierdoor kan het model zelfs in een downtrend genoeg kleine trades maken met voldoende winst.

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.

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 getrainede model gebruikt minder dan 1 GB aan VRAM op een gpu.

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 bevoorbeeld Crypto.com en dit zou in theorie even goed moeten werken.

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 eigelijk iets minder moeten aankopen en de aankopen die het wel doet op de lagere punten doen.

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 deels ook omdat het model vaak nog te vroeg aankoopt waardoor het heel wat potentiele winst laat liggen.

Ik denk echter zeker dat dit te verhelpen is met 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 word 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 eigelijk doet en dit gemakkelijk zelf kan aanpassen.

4.5 Meerwaarde

Dit onderzoek zal een beperkte meerwaarde bieden aan de maarschappij 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.

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 zelfs omdat de data beter is.

5 Advies

Een **advies** houdt concrete aanbevelingen voor het werkveld in. Je kunt ingaan op:

- de bruikbaarheid en toepasbaarheid van je vooropgestelde oplossingen.
- welke concrete aanbevelingen het werkveld volgens jou kan ondernemen op basis van jouw onderzoeksresultaten.
- welk stappenplan het werkveld hierbij zou kunnen gebruiken.
- welke tools je voor het werkveld ontwikkeld hebt.
- andere relevante adviezen voor het werkveld, gebaseerd op je onderzoek.

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

Evaluatiecriteria van dit hoofdstuk:

Onvoldoende: het advies ontbreekt of is ondermaats van kwaliteit: eigen interpretaties, herhaling van informatie uit vorige hoofdstuk(ken),...
Beperkt: het advies is aanwezig maar is te weinig onderbouwd. De link met het gevoerde onderzoek en/of reflectie met externen is afwezig. Hierdoor is het advies weinig bruikbaar.
Volstaat: het advies bouwt duidelijk verder op eigen onderzoeksresultaten en is daardoor voldoende onderbouwd.
Goed: naast een onderbouwd advies worden ook andere relevante alternatieven/suggesties geformuleerd.
Excellent: het advies bestaat ook uit een concreet stappenplan gebaseerd op eigen onderzoekservaringen en contacten met externen.

6 Conclusie

In je conclusie beantwoord je definitief jouw onderzoeksvraag. Dit is één van de belangrijkste onderdelen van jouw bachelorproef en wordt altijd gelezen.

Voer geen nieuwe elementen aan in je conclusie. Je conclusie is immers een beknopte weergave van wat je reeds eerder schreef. Integreer in de conclusie ook de belangrijkste elementen uit jouw reflectie en jouw advies.

Kortom, in dit onderdeel horen dus enkel de belangrijkste zaken thuis: zaken waaruit de lezer kan leren.

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.

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.

7 Literatuurlijst

Alle bronnen waarvan je gebruikmaakt, zet je in de literatuurlijst. Je gebruikt hiervoor de IEEE-stijl.

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).