Asim Emre Yilmaz

Bacheloroppgave

En prediktiv analyse av Twitter-sentiment på Bitcoins daglige avkastning ved bruk av kunstige nevrale nettverk

Bacheloroppgave i Økonomi og administrasjon
Veileder: Denis M. Becker

April 2022

NTNU
Norges teknisk-naturvitenskapelige universitet.
Fakultet for økonomi
NTNU Handelshøyskolen

NTNU
Kunnskap for en bedre verden

Bacheloroppgave

2022

NTNU

Kunnskap for en bedre verden

Asim Emre Yilmaz

En prediktiv analyse av Twitter-sentiment på Bitcoins daglige avkastning ved bruk av kunstige nevrale nettverk

Bacheloroppgave i Økonomi og administrasjon
Veileder: Denis M. Becker

Bacheloroppgave
April 2022

NTNU
Norges teknisk-naturvitenskapelige universitet.
Fakultet for økonomi
NTNU Handelshøyskolen

NTNU
Kunnskap for en bedre verden

Forord

Denne oppgaven representerer slutten av min bachelorgrad i økonomi og administrasjon ved NTNU Handelshøyskolen. Selv om det mest er preget av den globale pandemien, er jeg glad og stolt over å være en del av NTNU.

Jeg vil takke min veileder, førsteamanuensis ved NTNU, Denis M. Becker, for hans støtte, forståelse og veiledning i denne studien.

Innholdet i denne oppgaven står for forfatterens regning.

28. april 2022, Trondheim

Asim Emre Yilmaz

Sammendrag

Denne oppgaven tar sikte på å analysere korrelasjonen og årsakssammenhengen mellom det sentimentale innholdet i Twitter data og den daglige prisavkastningen til fremtredende kryptovaluta Bitcoin, og å undersøke om Twitter sentimentet kan være en effektiv prediktiv faktor for Bitcoin-prisen. For å nå målene samles de tweetene som er relatert til Bitcoin på en tidsramme på 57 dager og kombineres med daglig tweetvolum og Bitcoins prisavkastning. For videre arbeid trekkes sentimental informasjon ut fra tweet-tekstene, og alle tweetene gis sentiment-score og kategoriseres i en av fire ulike polariteter, nemlig compound, positive, neutral og negative. Ved å bruke disse dataene utføres flere kvantitative analyser med allment anerkjente metoder, arkitekturer og modeller av kunstige nevrale nettverk. Funn av eksperimentene avslører at det ikke er noen statistisk signifikant sammenheng mellom Twitter sentiment og Bitcoins prisavkastning. Videre indikerer resultatene av studien at det ikke er tilstrekkelig bevis til å antyde at Twitter sentiment og daglig tweetvolum inneholder prediktiv informasjon for Bitcoin-prisen.

Abstract

This study aims to analyze the correlation and causality between the sentimental content of Twitter data and the daily price return of prominent cryptocurrency Bitcoin, and to research if the twitter sentiment can be an effective predictive factor for Bitcoin price. To achieve the aims, tweets related to Bitcoin of a timeframe of 57 days are gathered and combined with daily tweet volume and Bitcoin price return. For the further work, sentimental information is extracted from the tweet texts, and all tweets are given sentiment scores and categorized into one of four different polarities, namely, compound, positive, neutral and negative. Using this data, several quantitative analyzes are performed with widely acknowledged methods, architectures, and models of Artificial Neural Networks. Findings of the experiments reveals that there is no statistically significant correlation between Twitter sentiment and Bitcoin price return. Furthermore, results of the study indicates that there's not sufficient evidence to suggest that Twitter sentiment and daily tweet volume contains predictive information for Bitcoin price.

Innholdsfortegnelse

Figurliste

Tabelliste

1 Innledning

Bitcoin (BTC) er sannsynligvis det mest interessante og populære investeringsverktøyet det siste tiåret. Siden den første opprettelsen i januar 2009 av Satoshi Nakamoto (Frankenfield 2021; Nakamoto 2008) har den vokst enormt. Folk var veldig tvilende til Bitcoin de første par årene, men etter noen kontroversielle år fikk Bitcoin seg akseptert blant investorer, markeder og i daglig bruk. I 2018 var det omtrent 22 millioner unike blockchain.com lommebøker, og antallet er mer enn 80 millioner nå (Blockchain.com n.d.). Dette er bare antall på blockchain.com og det er mange flere andre plattformer. Det nøyaktige antallet Bitcoin lommebøker og eiere er ganske umulig å vite, men det antas at det er mer enn 100 millioner unike Bitcoin eiere der 400.000 av dem aktivt bruker Bitcoin daglig (www.buybitcoinworldwide.com n.d.). Samtidig økes markedsverdien på Bitcoin fra ca. 15 milliarder USD til 172 milliarder mellom januar 2017 og januar 2018. Det er nå, i april 2022, ca. 750 milliarder USD med en topp på 1.156 billioner USD i oktober 2021 (Statista, 2022).

Disse tallene viser at Bitcoin er et viktig instrument i finansverdenen, og å forstå Bitcoin er nødvendig. I dag er alle, til en viss grad, kjent med Bitcoins konsept, men likevel er det fordelaktig å beskrive konseptet og strukturen. Bitcoin er en programvare som registrerer transaksjoner på en hovedbok som kalles blockchain. Gruppe av transaksjoner kalles blokker og disse blokkene bekreftes med mye datakraft. Dette konseptet kalles *mining* der brukere kjører sitt komplekse og kraftige system for å løse disse blokkoppgavene og bli belønnet i retur (Floyd 2021). Bitcoin er det første desentraliserte betalingsnettverket som drives av brukerne (Bitcoin.org, n.d). Å være desentralisert betyr at det ikke er noen autoritet bak Bitcoin, og den får sin makt fra menneskene som bruker den. Dette er veldig viktig fordi at å være desentralisert, innebærer at dynamikken til Bitcoin, kan være drevet av sosiale faktorer (Garcia et al. 2014).

Volatilitet er en av nøkkelegenskapene til Bitcoin og en av hovedgrunnene til at den er veldig populær og attraktiv. Det er ingen konstante og visse faktorer som påvirker volatiliteten til Bitcoin. Kristoufek (2015) forklarer dette ved å si at det er logisk at prisdriverne varierer over tid siden Bitcoin har dynamisk natur og raske prissvigninger.

Formålet med denne oppgaven er å analysere forholdet mellom Twitter innlegg som er relatert til Bitcoin og dens pris. Målet er å finne ut om Bitcoin tweeter på noen måte har innflytelse på Bitcoin-prisen. Bitcoin-prisen er ustabil, og det er mange faktorer som kan påvirke den. Som

det er fortalt ovenfor kan det å være desentralisert, gjøre at Bitcoin-prisen påvirkes av sosiale hendelser mer enn andre vanlige investeringsverktøy. Mai et al. (2015) sier at Bitcoin er et sosialt fenomen der prisen er drevet av hva folk tildeler den. For å overvåke sosiale bevegelser er Twitter en av de beste plattformene, og derfor brukes Twitter data i denne oppgaven. Garcia et al (2014) fant at Google Search-data og tweetvolumet til Bitcoin er relatert til Bitcoin prisen. Når ordet «bitcoin» søkes mer på Google, går prisen opp, og når prisen går opp, øker antall Bitcoin tweeter. Dette kan være villedende i denne oppgaven fordi Bitcoin-prisen kan påvirke Twitter mønsteret. Denne oppgaven prøver imidlertid å finne ut om det stikk motsatte er på spikeren. I tillegg til dette brukes Twitter ikke bare av brukere av sosiale medier, men også av investorer og eksperter av finans. Shen, Urquhart og Wang (2019) hevder at Google-søk er gjort av de som ikke vet om BTC, men de som tweeter er kjent med BTC og dens dynamikk. Dessuten sier French (2021) at Covid-19-pandemien har påvirket Twitter basert markedsusikkerhetsindeks (TMU), og TMU har blitt en prominent indikator på BTC-avkastning. TMU er beregnet basert på tweetene med usikkerhet om makroøkonomi, politikk og aksjemarkeder, ikke Bitcoin, men det viser at Twitter kan ha innflytelse på BTC-priser. Disse er hovedgrunnene til at Twitter data er valgt i denne oppgaven.

Oppgaven fortsetter med følgende deler: Del 2: Litteraturgjennomgang der det relaterte tidligere arbeidet med Bitcoin priser og dets påvirkningsfaktorer gjennomgås; Del 3: Data og metoder. Her finnes beskrivelsen av dataene og hvordan de hentes i tillegg til metodene som brukes for å innhente, forberede, behandle og analysere dataene; Del 4: Implementering hvor arbeidet som er utført beskrives, og det nevnes hvordan metodene ble implementert i ulike typer modeller; Del 5: Analyser og empiriske resultater av forskningen; Del 6: Drøfting hvor resultatene gjennomgås og problemstilling forsøkes besvart; Del 7: Denne delen avslutter oppgaven, og er sammensatt av en sammendrag av oppgaven; Del 8: Referanser i en liste.

2 Litteratur

Det er mange og forskjellige typer forskning gjort relatert til Bitcoin og dens volatilitet og pris. I denne oppgaven er det to typer litteratur. Den første er forskning som setter søkelys på å finne determinantene og parameterne som påvirker prissvingninger på Bitcoin og avkastning mens den andre fokuserer på effekten av ulike sosiale medieplattformer og verktøy, spesielt Twitter, på Bitcoins prisavkastning.

Meland og Øyen (2017) prøvde å forklare Bitcoins prissvingninger. Anvendt teoretisk fundament i forskningen er signaleringsteori. Signaleringsteori sier at informasjon er et nøkkelelement i beslutningstaking, og beslutningstakere kan ha to typer informasjon: offentlig og privat. Dette fører til en informasjonsasymmetri. I denne situasjonen med informasjonsasymmetri har et svært begrenset antall personer den private informasjonen, og de fleste av beslutningstakerne bruker offentlig informasjon (Connelly et al. 2011). Med 279 ukentlige observasjoner og 9 forskjellige variabler utviklet de 2 modeller. Resultatene viser at det er en negativ sammenheng mellom Bitcoin volum og Bitcoin-pris som de forklarer med grunnleggende tilbud og etterspørselskonsept. Videre fant de en betydelig positiv sammenheng mellom Googles søkevolum for ordet Bitcoin og prisen. I tillegg er det vist i resultater at positive og negative politiske hendelser og uttalelser har en betydelig sammenheng med Bitcoin-prisen. Melands forskning har to viktige resultater for denne oppgaven. For det første har et sosialt medieverktøy som Google-søk forhold til Bitcoin-prisen. For det andre har den politiske situasjonen innvirkning. Twitter er et sosialt medie der folk søker og prøver å finne informasjon om ting, spesielt kryptovalutaer, og politiske nyheter og bevegelser dukker opp på Twitter på kortest tid med en betydelig mengde folk legger ut sine meninger om nyheter.

Shen, Urquhart og Wang (2019) brukte Twitter data og prøvde å finne forholdet til avkastningen, volatiliteten og handelsvolumet til Bitcoin. De benyttet tweetvolumet til Bitcoin, altså antall tweetene, i stedet for tekstinnholdet i tweetene. De studerte at tweetvolumet fra tidligere dager har betydelig innvirkning på volatiliteten, dvs. jo høyere antall tweetene, jo høyere volatilitet. På samme måte har tweetvolumet også betydelig innflytelse på handelsvolumet. De brukte Granger kausalitetstest og oppdaget at de to avhengige variablene har bilateral sammenheng med tweetvolumet. Med andre ord gir det høyere tweetvolumet høyere volatilitet de neste dagene, og den høyere volatiliteten gir høyere antall tweeter

påfølgende dager, og det samme for handelsvolumet. Til slutt viser funnene deres om avkastning at Bitcoins tweetvolumet ikke har noen betydelig innflytelse på Bitcoin-prisen.

Gao, Huang og Wang (2021) lurte på om Twitter data påvirker prisen på Bitcoin og gjorde sentimentanalyseundersøkelser med Twitter data. De samlet finansrelaterte tweeter og laget en timevis sentimentanalyse for å forutsi avkastning og volatilitet. For å lage en sentimentanalyse brukte de Loughran-McDonald (2011) leksikon som har en liste over positive, negative og usikre ord med fokus på finans for å kategorisere tweetene i tre forskjellige følelser: bullish (positiv), bearish (negativ) og null (nøytral).

Regresjonsanalyse viste at antall både bullish og bearish tweeter er i stand til å forutsi volatilitet i henholdsvis positiv eller negativ retning. Dessuten er det funnet at antall bullish sentimenter har en viktig innvirkning på markedsusikkerhet. Når det gjelder forhold til avkastning, er det vist at bullish sentimenter har positiv og statistisk signifikant innflytelse på samtidig pris og negativ innvirkning på fremtidig pris, noe som gir oss en forklaring på volatilitet. På den annen side har bearish tweeter negativ innflytelse på samtidig avkastning, men denne påvirkningen er ikke statistisk signifikant. Selv om de brukte økonomiske tweeter i stedet for Bitcoin tweeter i sin forskning, viser resultatene at Twitter data kan ha en delvis innvirkning på Bitcoins prisavkastning.

Kaminski (2016) utførte en sentimentanalyse som ligner veldig på denne oppgaven og forsøkte å finne sammenhenger og årsakssammenhenger mellom Bitcoins markedsindikatorer og Twitter innlegg som inneholder emosjonelle signaler på Bitcoin. Hans korrelasjonsanalyse viste at Bitcoins handelsvolum er sensitivt for negative tweeter og usikkerhetssignaler. Jo lavere antall negative tweeter, desto høyere er sluttprisen på Bitcoin. Tweetene med positive følelser, imidlertid, ser ikke ut til å ha innflytelse på Bitcoin-prisen alene, men forholdet mellom antall positive og negative tweeter har en positiv korrelasjon med sluttpris. Hvis det er et høyere forhold mellom positive og negative tweeter, er sluttprisen høyere. Han foretok en Granger-årsakstest og fant ut at det ikke er noen statistisk signifikans for emosjonelle signaler som en prediktor for markedet. Omvendt speiler emosjonelle følelser markedet i stedet for å forutsi det. Når handelsvolumet går opp, flyr følelsene høyt på Twitter. Han konkluderte med at Twitter er Bitcoins virtuelle handelsgulv, og det gjenspeiler følelsesmessig dynamikken.

Kaminski (2016) sin forskning viser allerede at Twitter sentiment ikke påvirker Bitcoin-prisen. Imidlertid er det en relativt gammel forskning, og French (2021) hevder at med COVID-19 pandemien har Twitter mye mer innvirkning på kryptovalutapriser. Han bruker Twitter Based Uncertainty Index (TMU), som er forklart ovenfor i introduksjonsdelen, for å se om den har samme effekt på Bitcoin før og under COVID-19 pandemien. Antall daglige inntektsgivende brukere på Twitter har økt med 34 % ved begynnelsen av COVID-19-pandemien, som var den raskeste veksten i denne beregningen i selskapets historie (Hutchinson 2020). French (2021) sier at TMU hadde en liten innvirkning på Bitcoins avkastning og betingede volatilitet, men med pandemien har de blitt mer sensitive for usikkerhetsinnholdet i tweetene, og det styrker TMUs evne til å forutsi dem. Med andre ord, i pre-COVID-19-perioden, var TMU ikke en ledende indikator på avkastning eller volatilitet på avkastning, men i den svært usikre COVID-19-perioden var mengden av folk som twitret om usikkerhet en ledende indikator på både Bitcoins prisavkastning og dens betingede volatilitet.

Akbiyik et al. (2021) har utført en omfattende undersøkelse med over 30 millioner Bitcoin relaterte tweeter og 5 forskjellige maskin læring arkitekturer. Målet var å finne hvilken arkitektur som fungerer best for modellen deres, og om informasjonen hentet fra Twitter har noen innflytelse på prisvolatiliteten for Bitcoin. De brukte NLTK-baserte VADER for å anvende sentimentanalyse på tweet-tekster, og de brukte også tweet-forfatternes metainformasjon som bl.a. følgertellinger og følgetellinger. I evaluering av framføring brukte de 4 beregninger: gjennomsnittlig absolutt prosent feil (MAPE), gjennomsnittlig absolutt feil (MAE), rot middel kvadrat feil (RMSE) og gjennomsnittlig kvadratisk logaritmisk feil (MSLE). I RMSE og MSLE hadde Long Short-Term Memory (LSTM) dyp læringsmodell, som er en modell for Tilbakevendende Nevrale Nettverk (RNN), bedre ytelse enn de andre. Når det gjelder prediksjonsresultatene, dukket det opp et uventet resultat. Det semantiske innholdet i tweetene var ikke veldig informativt om prisvolatiliteten for Bitcoin. Tweet-forfatternes metainformasjon hadde imidlertid en mye mer prediktiv kraft på volatilitet.

3 Data og Metoder

I denne delen beskrives de metodiske måtene som følges i oppgaven. Dessuten presenteres dataene, innsamling av dataene, dataenes egenskaper og hvordan de håndteres. I tillegg er verktøyene, teknikkene og prosessene definert, og forklart hvorfor de er valgt, deres fordeler og ulemper, og hvorfor de passer godt for å løse problemet.

3.1 Datainnsamling

Det finnes tre ulike deler av data som ble samlet inn i oppgaven: Tweet data, tweet volumdata og prisdata.

Tweet data. Dataene som ble brukt, hentet fra twitter.com ved bruk av Twitter sin offisielle API og Tweepy som er et Python-bibliotek spesielt laget for å skrape tweeter. Tweetene som inneholder ordet "bitcoin" eller "btc" som er lagt ut mellom 1. januar 2022 og 26. februar 2022, samlet inn. Datasettet består av tweet-tekster og datoer. Det står 4 985 483 tweeter inkludert retweeter og datoer tweetene er lagt ut på, i datasettet. Retweetene ble ikke slettet siden de antas å være en god indikator på styrken til en spesifikk tweet.

Tweet volumdata. Siden mangler det en del tweeter på bestemte datoer i datasettet, ble antall tweeter i datasettet ikke brukt som tweetvolumvariabel. I stedet samlet de daglige tweetvolumene fra bitinfocharts.com manuelt.

Bitcoin prisdata. Daglige sluttkurser for Bitcoin vs. USD for relatert periode er lastet ned fra Yahoo Finance som en csv-fil. Filen inneholder BTC/USD-kursens daglige sluttkurs, justert sluttkurs og voluminformasjon for relatert tidsramme. Siden avkastningen var denne oppgavens målverdi, ble det beregnet en daglig avkastningsverdi ved bruk av justert sluttkurs ved hjelp av formelen i figur 1. Timevis sluttkurser kunne bli brukt i oppgaven, men det er tenkt at dette er altfor kort for BTC-prisen til å få påvirket. Tvert imot Bitcoins prissvingninger kunne påvirke tweeter på en slik kort periode.

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1$$

r_t = return
P_t = price at point of time
P_{t-1} = price at the preceding point of time

Figur 1: Formell for avkastning.

3.2 Dataforbehandling

Dataforbehandling er nødvendig for både sentimentanalyse og prediksjonsmodeller. Det er viktig fordi rådataene vanligvis inneholder manglende verdier, inkonsistente deler eller støy. Derfor er flere dataforbehandlingsteknikker benyttet for å få dataene i en standard form før bruk, slik at de kan brukes enklere og mer effektivt. Siden dataene ikke inneholder noen kategoriske data, var det ingenting å gjøre med dette. Datarensing og datatransformasjon er gjort til å forberede dataene til videre prosesser.

Datarensing. Rensing av dataene er viktig fordi dataene kan ha unøyaktige eller irrelevante deler. Twitter-tekster kan inneholde store og små bokstaver, URL-er, medielenker, emojier, irrelevante hashtag, tall og tegnsetting. Først ble det kontrollert og fjernet manglende verdier. Deretter ble det fjernet alle nettadressene, mediekoblingene, emojiene, tallene og tegnsetting fra teksten ved hjelp av regex-syntaks. Deretter ble alle bokstaver forvandlet til små bokstaver. Etterpå ble ofte brukte ord, nemlig stoppord, slettet fra tekstene siden de ikke har en vesentlig betydning for sentimentanalysen. Stoppord-pakken til NLTK Corpus-biblioteket er benyttet til å lage en liste over stoppord.

Datatransformasjon. Etter sentimentanalysen og de nyopprettede variablene med sentiment score var datasettet klart for modellen til å tilpasse. For å øke ytelsen til en modell er standardisering eller skalering av dataene avgjørende. Derfor ble dataene skalert ved å bruke MinMaxScaler-funksjonen i Scikit Learn-forbehandlingsbiblioteket før de brukes i modellen. På den måten ble alle verdier skalert mellom -1 og 1.

3.3 Sentimentanalyse

'Naturlig språkbehandling (NLP) refererer til grenen av informatikk – og mer spesifikt grenen av kunstig intelligens eller AI – som er opptatt av å gi datamaskiner muligheten til å forstå tekst og talte ord på omtrent samme måte som mennesker kan' (IBM Cloud Education 2020b, para. 1). Det finnes flere NLP-teknikker og sentimentanalyse er en av dem. Sentimentanalyse er definert som 'prosessen med beregningsmessig identifisere og kategorisere meninger uttrykt i et tekststykke, spesielt for å avgjøre om forfatterens holdning til et bestemt emne, produkt osv. er positiv, negativ eller nøytral' (Lexico Dictionaries | English n.d., para. 1). Prosessen er i utgangspunktet å analysere forfatterens sentiment i teksten og prøve å finne noen subjektive kvaliteter som følelser, sinne, sarkasme, lykke, mistenksomhet osv. Sentimentanalyse brukes på diverse områder og sosiale medier er en av dem. Sentimentanalyse i sosiale medier er mye

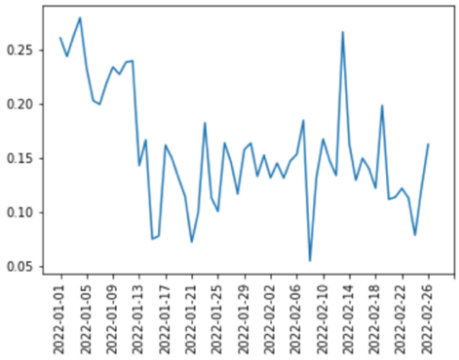
brukt for å forstå hva kundenes meninger om en merkevare er. Imidlertid har det i det siste vært veldig populært å bruke sentimentanalysen på sosiale medier for å komme med forutsigelser om virksomheter, aksjer og valutaer.

Sentimentanalyser kan utføres via maskinlæring, leksikoner eller en blanding av de to. Maskinlæring er mer krevende og mindre presis. For en mer nøyaktig og lettere sentimentanalyse er et sentiment leksikon nødvendig, og det finnes mange av dem. Noen av leksikonene er utviklet spesielt for bestemte arbeidsfelt. I denne oppgaven ble Valence Aware Dictionary and sEntiment Reasoner (VADER) brukt. 'VADER er et leksikon og regelbasert sentimentanalyseverktøy som er spesifikt tilpasset følelser uttrykt i sosiale medier' (Hutto og Gilbert 2014, para. 1). VADER inneholder relativt sett et mindre antall ord enn andre leksikoner; presisjonen er imidlertid imponerende siden den er laget manuelt. Leksikoner som er laget manuelt mye bedre enn de andre fordi de tvetydige ordene eksluderes og de bruker mindre, men sterkere ord (Taboada et al. 2011). Den kan håndtere forkortelser, slanger, uttrykksikoner og emojier som vanligvis finnes i sosiale medier (Hutto og Gilbert 2014). Siden den ikke krever opplæring, er den vanligvis mye raskere enn andre maskinlæringsalgoritmer.

Tekstene blir gitt sentiment score i henhold til deres følelsesintensitet. Score typene kalles polaritet og disse polaritetene i VADER er klassifisert som negative, nøytrale, positive og sammensatte. Negative, nøytrale og sammensatte polariteter har verdi mellom 0 og 1, og sammensatte polaritet beregnes ved å summere de tre og normalisere summen mellom -1 og 1 der -1 er absolutt negativ og 1 er absolutt positiv. I oppgaven ble det anvendt sentimentanalyse med VADER på alle tweetene i datasettet, og polaritet scorene er lagt til datasettet. Deretter ble datasettet omformet til et datoindeksert datasett der gjennomsnittsverdiene for polaritetene etter dato brukes. En prøve fra datasettet er vist i figur 2, og en graf over snittverdiene for sammensatte polaritet etter dato er vist i figur 3.

date	compound	neg	neu	pos
2022-01-01	0.260918	0.029950	0.774254	0.187575
2022-01-02	0.244012	0.024564	0.802802	0.164214
2022-01-03	0.262708	0.023801	0.795802	0.172564
2022-01-04	0.279713	0.023188	0.782762	0.186131
2022-01-05	0.232762	0.028359	0.795028	0.167938

Figur 2: En prøve av snittverdiene til sentiment polaritetene etter dato.



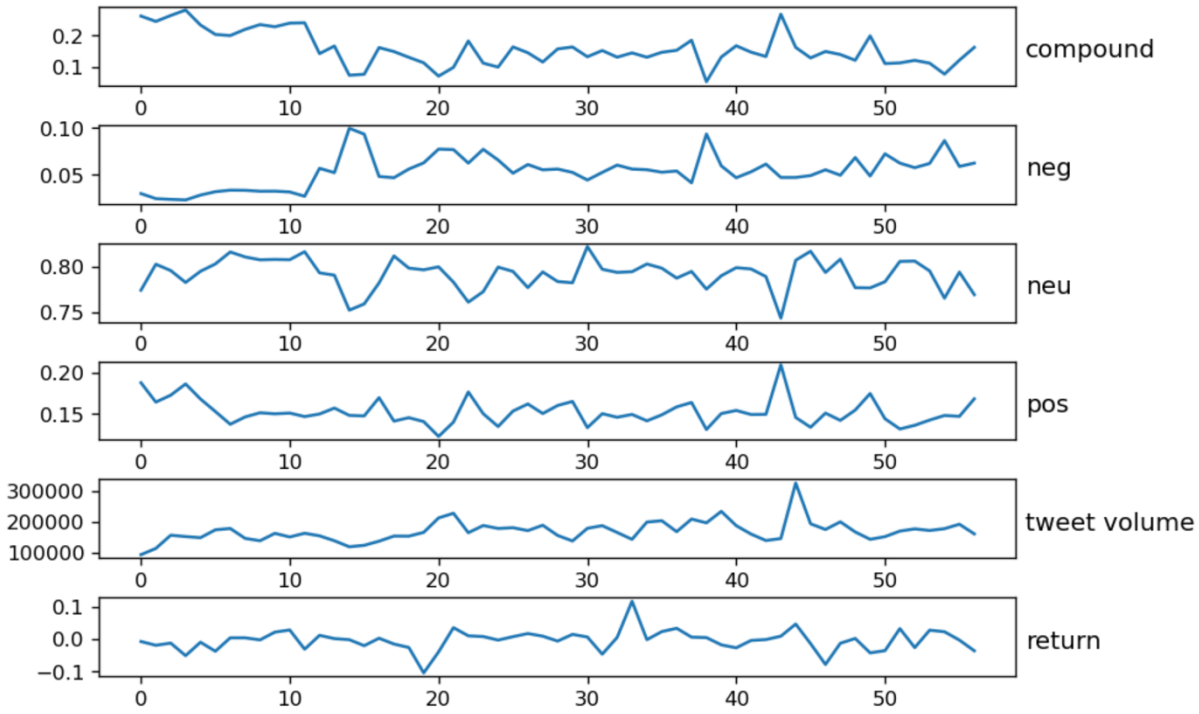
Figur 3: Kurven av snittverdier etter dato.

3.4 Datasettbeskrivelse

Datasettet etter sentimentanalyse består av 4 variabler og datoer som indeks. Indekskolonnen er datoer i formatet åååå-mm-dd. Den starter fra 2022-01-01 og slutter 2022-02-26.

Variabelen *Compound* er gjennomsnittlig sammensatte sentimentscore for dagen, og det er mellom -1 og 1. *Neg* representerer gjennomsnittlig negativ sentimentscore mens *Pos* er gjennomsnittlig positiv sentimentscore, og begge er mellom 0 og 1. *Neu* gir gjennomsnittet av nøytraliteten til tweetene som er lagt ut på en dag og er også mellom 0 og 1.

To nye variabler, *tweet volume* og *return*, er lagt til datasettet. *Tweet volume* er antall daglig tweeter som legges ut. *Return* står for daglig avkastning av BTC/USD-kurs. Variabler er vist i figur 4.

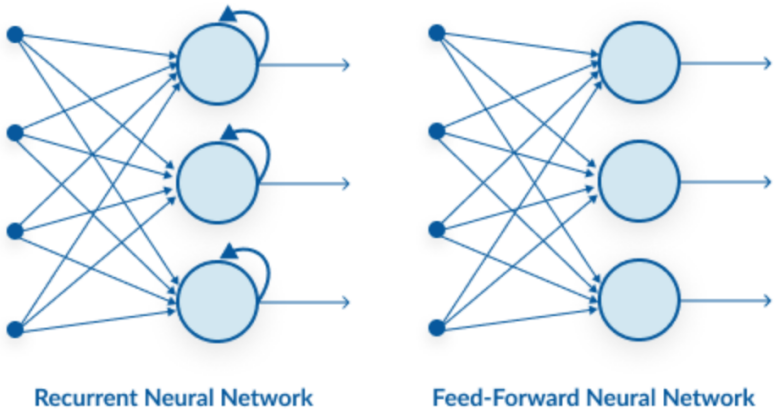


Figur 4: Utforskende dataanalyse.

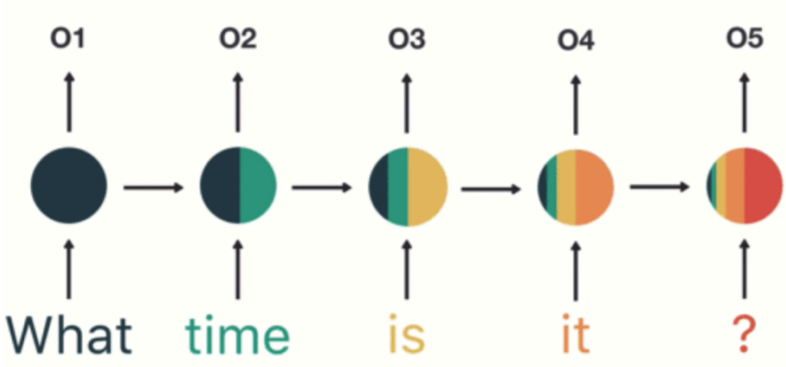
3.5 Tilbakevendende nevrale nettverk

Det finnes forskjellige typer maskinlæringsteknikker. De fleste av dem er anvendelige på flere felt. 'Kunstige nevrale nettverk (ANN) er en undergruppe av maskinlæring og er kjernen i dyplæringsalgoritmer. Navnet og strukturen deres er inspirert av den menneskelige hjernen som etterligner måten biologiske nevroner signaliserer til hverandre' (IBM Cloud Education 2020a, para. 1). ANN-er har evnen til å lære og modellere ikke-lineære og komplekse relasjoner, noe som er veldig viktig siden i det virkelige liv er forholdene mellom faktorene stort sett ikke-lineære og komplekse. De er også anerkjente når det gjelder å generalisere usett data ved å bruke inndataene (Basheer og Hajmeer 2000). Sist, men ikke minst, har ANN-er optimale resultater når dataene er heteroskedastiske (Russo, Madani og Rinaldi 2020). Heteroskedastisitet er at variansen til serien endres konstant over tid. Dataene og forholdet i denne oppgaven er kjent for volatilitet og kompleksitet. Derfor er ANN-er et av de beste alternativene å jobbe med.

ANN-er består av noder og lag. Noder er som kunstige nevroner. De er koblet til andre noder og har tilhørende vekter og terskler. Data kommer først til inndatalaget, og vektene tildeles. Ved hjelp av vektene, bestemmes betydningen av variablene. Vekter multipliseres med innganger og summeres deretter for å gå gjennom en aktiveringsfunksjon som bestemmer utgangen. Hvis utgangsverdien overskrider terskelen, aktiveres noden for å sende dataene til neste lag (IBM Cloud Education 2020a). I denne opppgaven er det brukt en spesiell type ANN som kalles Tilbakevendende nevrale vettverk (RNN). RNN er en bedre versjon for tidsserieanalyser takket være sin tilbakevendende tilkobling på skjulte lag. Sløyfen sørger for at sekvensiell informasjon fanges opp i inndataene (Aravindpai 2020). RNN blir mer og mer viktig i Naturlig språkbehandling (NLP), særlig innen maskinoversettelse og talegjenkjenning. Imidlertid har RNN ikke bare vært en viktig del av NLP, men også har blitt brukt i økende grad på tidsserieanalyser i det siste (Li, Huang og Lu 2021). En alvorlig ulempe med nevrale nettverk er forsvinnende gradientproblem. Problemet med forsvinnende gradient er i utgangspunktet at trening blir veldig vanskelig ettersom flere lag med visse aktiveringsfunksjoner legges til siden gradienten til tapsfunksjonen nærmer seg null (Wang 2019). Figur 5 viser forskjellen mellom et RNN og et vanlig Feed-forward nevrale nettverk. Et eksempel på hvordan inngangsverdiene er representert i forskjellige tidstrinn for en tekst, er illustrert i figur 6. Som det sees, bestemmes utgangen ved hvert tidstrinn av ikke bare gjeldende, men også av de tidligere inngangene.



Figur 5: Tilbakevendende nevrale nettverksstruktur (Barman, 2020b).



Figur 6: Et eksempel på betydning av sekvensiell data på LSTM (Barman, 2020a).

Det finnes ulike typer RNN-arkitekturer med ulike spesialiseringer. Long Short-Term Memory (LSTM) er en RNN-arkitektur som er spesielt designet for å løse problemet med forsvinnende gradient. Standard RNN-strukturer kan ikke bygge bro over mer enn 5-10 tidstrinn mens LSTM kan lære å bygge bro over mer enn 1000 diskrete tidstrinn (Staudemeyer og Morris 2019). LSTM-nettverk har celler kalt minneblokker i stedet for klassiske nevroner. Disse blokkene har komponenter som gjør dem smartere og holder et minne om en sekvens. LSTM blokker består av tre porter, altså inngangsport, utgangsport og glem-port. Glem-porten bestemmer hvor mye av dataen beholdes, og tilsvarende nullstiller glem-porten blokken. Inngangsporten bestemmer mengden av data som skal beholdes slik at unødvendig data forhindres mens utgangsporten bestemmer mengden av data som overføres som utdata (Øyen 2018). Denne strukturen gjør at LSTM lærer de viktige delene av sekvensen og glemmer de mindre viktige for å gi mer presis nøyaktighet. Dette gjør LSTM til en av de beste praksisene for forutsigelse og prediksjon av tidsserier.

4 Implementering

I denne delen skal implementering av metoder og modeller beskrives, og anvendelser av ulike typer prosedyrer vil bli nevnt.

Både bivariat og multivariat tidsserieanalyser ble utført og resultatene ble analysert og sammenlignet. Før dataene brukes i modellene ble datasettet omformet som et overvåket læringsproblem ved å bruke Shift-funksjonen i Python. Ved hjelp av denne funksjonen opprettes nye variabler som etterslep-observasjoner av gjeldende variabler. Det er rett og slett å kopiere kolonnen og bevege den vertikalt. På denne måten gis en sekvens av inndata, som ikke er stasjonær, til modellen for å bli trent (Brownlee 2017). Figur 7 viser et eksempel på omforming med Shift-funksjonen. Her er det to variabler var1 (*compound*) og var2 (*return*). På det første tidstrinnet er det bare verdier av variabler på tidspunkt *t*. Neste trinn er det verdier ved *t* og verdier ved *t-1*. Ved det tredje tidspunktet står det verdier ved *t*, *t-1* og *t-2*.

date	var1(t-2)	var2(t-2)	var1(t-1)	var2(t-1)	var1(t)	var2(t)
2022-01-01	NaN	NaN	NaN	NaN	0.260918	-0.007163
2022-01-02	NaN	NaN	0.260918	-0.007163	0.244012	-0.018737
2022-01-03	0.260918	-0.007163	0.244012	-0.018737	0.262708	-0.012066
2022-01-04	0.244012	-0.018737	0.262708	-0.012066	0.279713	-0.050734
2022-01-05	0.262708	-0.012066	0.279713	-0.050734	0.232762	-0.009366
2022-01-06	0.279713	-0.050734	0.232762	-0.009366	0.203087	-0.037141
2022-01-07	0.232762	-0.009366	0.203087	-0.037141	0.199685	0.004236
2022-01-08	0.203087	-0.037141	0.199685	0.004236	0.219161	0.004257
2022-01-09	0.199685	0.004236	0.219161	0.004257	0.234118	-0.002155
2022-01-10	0.219161	0.004257	0.234118	-0.002155	0.227352	0.021869

Figur 7: Omformet overvåket læringsdatasett.

4.1 Bivariat tidsserieanalyse

Bivariat tidsserieanalyse består av 1 uavhengig variabel og 1 avhengig variabel som er målvariabelen. Det ble utført individuelle bivariate analyser for enkelte variabler. Målvariabelen er *return* på tidspunktet *t* i hver analyse. Uavhengige variabler for hver analyse er *compound*, *neg*, *pos* og *tweet volume*.

4.2 Multivariat tidsserieanalyse

Multivariat tidsserieanalyse er satt sammen av flere uavhengige variabler og 1 avhengig variabel. Multivariat analyse ble utført inkludert tidligere verdier av målvariabel som uavhengige variabler. Uavhengige variabler i analysen er *compound*, *neg*, *neu*, *pos* og *tweet volume* og *return* (tidligere observasjoner på *t-n*) og målvariabelen er *return* på tidspunkt *t*.

4.3 Model og evaluering

Modellen som er bygget er Sequential modellen fra Keras bibliotek. Inndatalag er et LSTM-lag. Aktiveringsfunksjonen er tanh som er standard aktiveringsfunksjon for LSTM. Den tilbakevendende aktiveringsfunksjonen som er aktiveringsfunksjonen som brukes for det tilbakevendende trinnet, er sigmoid. Tapsfunksjonen er valgt mellom gjennomsnittlig absolutt feil (MAE) og gjennomsnittlig kvadratfeil (MSE), og optimalisereren er adam. Skjulte lag er også LSTM-lag, og de har samme parametere som inndatalaget. Utgangen genereres av et dense lag med 1 celle.

Overtilpasning er en viktig feil for en modell siden modellen lærer seg detaljene og støyen i et datasett. Dette kan påvirke ytelsen til modellen negativt i andre datasett. Derfor, for å forhindre overtilpasning, ble dataene tilpasset til modellen med tidlig stopp ved bruk av Earlystopping. Earlystopping-funksjonen i Keras-biblioteket overvåker en gitt beregning og stopper treningen når beregningen slutter å forbedre seg.

For å evaluere ytelsen av modeller og resultatene til eksperimentene er rot middel kvadrat feil (RMSE) benyttet. Predikerte verdier ble opprettet og invertert fra skalerte verdier. Deretter ble det beregnet RMSE-verdi for predikerte og observerte målverdier ved hjelp av mean_squared_error-funksjonen i Sci-kit-læringsbiblioteket. Formelen til RMSE er gitt i figur 8. Til slutt ble det plottet en graf over de predikerte og observerte verdiene for å se resultatet visuelt.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

i = variabel i
y_i = observert verdi
$\hat{y}_i$ = predikert verdi
n = antall observasjoner

Figur 8: Formell for RMSE.

5 Analyse

I denne delen introduseres empirisk bevis. Funnene og resultatene av observasjonene og eksperimenteringen blir analysert og evaluert i ulike perspektiver. De kvantitative resultatene er presentert med figurer og tabeller for å gi en dyptgående oversikt og gi et detaljert innblikk i arbeidet som er utført.

5.1 Data

Det rå datasettet inneholder 4 985 483 tweeter og datoene som disse tweetene er lagt ut på. Tidsrammen er 57 dager totalt. Antall unike tweeter er annerledes fra det totale antallet tweeter og er 2 932 719 slik det er vist i figur 9. Det er fordi datasettet inneholder ikke bare tweeter, men også retweeter. En retweet er i utgangspunktet å legge ut noens tweet på din egen profil. Figur 10 viser et utvalg av 10 tilfeldige data av tweeter og datoer fra datasettet. RT i begynnelsen av tweetene indikerer at det er en retweet. Det er omtaler, emojier, URL-er, tegnsetting, stoppord og tall å rense i forbehandlingen. Figur 11 viser de samme tweetene etter dataforbehandling. Som det er sett, er teksten mye enklere etter rengjøringsoperasjoner.

	tweet	date
count	4985483	4985483
unique	2932719	57
top	Promote it on @iconictraderss	2022-01-06
freq	23353	201474

Figur 9: Deskriptive verdier av råvariabler.

tweet	date
I am claiming my free Lightning sats from @_bitcoiner's amazing faucet! ⚡\n\n@boltcoiner unlock me a1a6dee5-7e96-4b85-af3b-b4ed86ba707a\n \n#Bitcoin #BTC #LN #LightningNetwork #boltcoiner\nhttps://t.co/lbzmvpcriW	2022-01-08
RT @BTC_Archive: #Bitcoin vouchers at the store in Zurich, Switzerland 🇨🇭 https://t.co/21rxn7IsWQ	2022-01-18
Stop using Bitcoin SV ! Try to buy 23887 $PKN instead !!! #PKN #pokmi #PokenArmy #poken #PokenArmyProgram	2022-02-01
RT @hoseins51979372: @BluesparrowC @BlueSparrowETH Huuuuge News!!!\nNew Utility Income Breakdown in Blue… CEX,\nHaving our own platform, trad…	2022-01-31
RT @deficonnect: Soon to come...\n\nDELTA7 our very own metaverse..\n\n80% complete and almost ready for launch...\n\nThe future is indeed bright…	2022-01-13
RT @BitcoinMagazine: "#Bitcoin is the great uniter." — US Senator 🇺🇸 https://t.co/pPOXI3jEGo	2022-02-11
I found #Bitcoin this year. Along with it I found a group of incredible people who care about fairness, who won't accept anything less than truth, and who understand that sound money will have a profound positive impact on all societies on Earth. Thank you 🙏Happy New Year!	2022-01-01
I believe this is a faithful project.The projector has a lot of attractions so hopefully the project will be better in the future and will be the best.\nI love it..\n@Noyon31350067\n@RobiulH19575815\n@MDAlomMia11,\n#Airdrops #BNB #BTC #cryptocurrency #ETH	2022-01-05
RT @CarlBMenger: #Bitcoin fixes this. https://t.co/vUR0ZBOsWr	2022-02-11
RT @KitcoNewsNOW: Whenever the #Fed tightens into a slowing economy, risk assets suffer. The downside target for #Bitcoin? $26k says @Macro…	2022-02-22

Figur 10: En prøve av rå tweettekster.

tweet	date
claiming free lightning sat bitcoiners amazing faucet boltcoiner unlock aadeeebafbbedbaabitcoin btc ln lightningnetwork	2022-01-08
bitcoin voucher store zurich switzerland	2022-01-18
stop using bitcoin sv try buy pkn instead pkn pokmi pokenarmy poken pokenarmyprogram	2022-02-01
bluesparrowc bluesparroweth huuuuge newsnew utility income breakdown blue cexhaving platform trad	2022-01-31
soon comedelta metaverse complete almost ready launchthe future indeed bright	2022-01-13
bitcoin great uniter u senator	2022-02-11
found bitcoin year along found group incredible people care fairness wont accept anything le truth understand sound money profound positive impact society earth thank happy new year	2022-01-01
believe faithful projectthe projector lot attraction hopefully project better future besti love itnoyonrobiulhmdalommiaairdrops bnb btc cryptocurrency eth	2022-01-05
bitcoin fix	2022-02-11
whenever fed tightens slowing economy risk asset suffer downside target bitcoin k say macro	2022-02-22

Figur 11: Prøven av tweettekster etter forbehandling.

5.2 Sentimentanalyse

Sentimentanalyse ble anvendt på både rå og forbehandlede datasett for å se om det er en signifikant forskjell mellom de to. I figur 12 er sammensatte, negative, nøytrale og positive polariteter av prøve tweetene i rådatasettet vist. Figur 13, derimot, viser scorene av de samme tweetene fra forbehandlet datasett. Det finnes noen forskjeller mellom de to. Ifølge figurene rå tekster gir sterkere resultater enn forbehandlede. Forskjellen i *compound* er mindre enn de andre polaritetene. *Neu* er større i rå tekster generelt, og det kan skyldes meningsløse støyende ord.

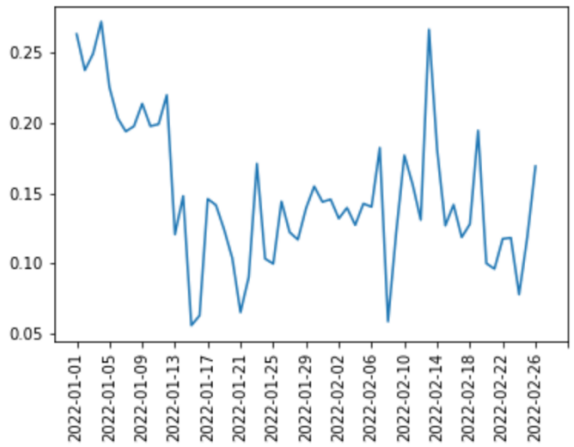
tweet	date	compound	neg	neu	pos
I am claiming my free Lightning sats from @_bitcoiner's amazing faucet! ⚡\n\n@boltcoiner unlock me a1a6dee5-7e96-4b85-af3b-b4ed86ba707a\n\n#Bitcoin #BTC #LN #LightningNetwork #boltcoiner\nhttps://t.co/lbzmvpcriW	2022-01-08	0.8122	0.000	0.709	0.291
RT @BTC_Archive: #Bitcoin vouchers at the store in Zurich, Switzerland 🇨🇭 https://t.co/21rxn7lsWQ	2022-01-18	0.0000	0.000	1.000	0.000
Stop using Bitcoin SV ! Try to buy 23887 $PKN instead !!! #PKN #pokmi #PokenArmy #poken #PokenArmyProgram	2022-02-01	-0.5216	0.183	0.817	0.000
RT @hoseins51979372: @BluesparrowC @BlueSparrowETH Huuuuge News!!!\nNew Utility Income Breakdown in Blue… CEX,\nHaving our own platform, trad…	2022-01-31	0.0000	0.000	1.000	0.000
RT @deficonnect: Soon to come...\n\nDELTA7 our very own metaverse..\n\n80% complete and almost ready for launch...\n\nThe future is indeed bright…	2022-01-13	0.2975	0.000	0.905	0.095
RT @BitcoinMagazine: “#Bitcoin is the great uniter.” — US Senator 🇺🇸 https://t.co/pPOXl3jEGo	2022-02-11	0.6249	0.000	0.709	0.291
I found #Bitcoin this year. Along with it I found a group of incredible people who care about fairness, who won't accept anything less than truth, and who understand that sound money will have a profound positive impact on all societies on Earth. Thank you 🙏Happy New Year!	2022-01-01	0.9219	0.000	0.733	0.267
I believe this is a faithful project.The projector has a lot of attractions so hopefully the project will be better in the future and will be the best.\nI love it..\n@Noyon31350067\n@RobiulH19575815\n@MDAlomMia11,\n#Airdrops #BNB #BTC #cryptocurrency #ETH	2022-01-05	0.9661	0.000	0.586	0.414
RT @CarlBMenger: #Bitcoin fixes this. https://t.co/vUR0ZBOsWr	2022-02-11	0.0000	0.000	1.000	0.000
RT @KitcoNewsNOW: Whenever the #Fed tightens into a slowing economy, risk assets suffer. The downside target for #Bitcoin? $26k says @Macro…	2022-02-22	-0.7096	0.300	0.632	0.067

Figur 12: Prøven av rå tweettekster med sentiment scorer.

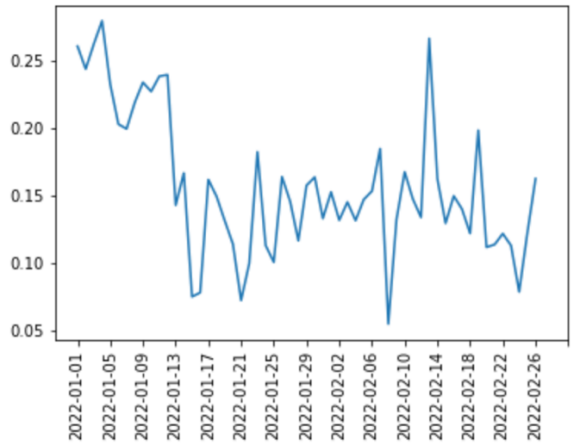
tweet	date	compound	neg	neu	pos
claiming free lightning sat bitcoiners amazing faucet boltcoiner unlock aadeeebafbbedbaabitcoin btc ln lightningnetwork	2022-01-08	0.7964	0.000	0.608	0.392
bitcoin voucher store zurich switzerland	2022-01-18	0.0000	0.000	1.000	0.000
stop using bitcoin sv try buy pkn instead pkn pokmi pokenarmy poken pokenarmyprogram	2022-02-01	-0.2960	0.155	0.845	0.000
bluesparrowc bluesparroweth huuuuge newsnew utility income breakdown blue cexhaving platform trad	2022-01-31	0.0000	0.000	1.000	0.000
soon comedelta metaverse complete almost ready launchthe future indeed bright	2022-01-13	0.6258	0.000	0.610	0.390
bitcoin great uniter u senator	2022-02-11	0.6249	0.000	0.423	0.577
found bitcoin year along found group incredible people care fairness wont accept anything le truth understand sound money profound positive impact society earth thank happy new year	2022-01-01	0.9204	0.057	0.546	0.398
believe faithful projectthe projector lot attraction hopefully project better future besti love itnoyonrobiulhmdalommiaairdrops bnb btc cryptocurrency eth	2022-01-05	0.9403	0.000	0.433	0.567
bitcoin fix	2022-02-11	0.0000	0.000	1.000	0.000
whenever fed tightens slowing economy risk asset suffer downside target bitcoin k say macro	2022-02-22	-0.6249	0.398	0.471	0.131

Figur 13: Prøven av tweet tekstene etter forbehandlingen samt sentiment scorene.

Figurene 14 og 15 viser kurver over gjennomsnittlig sammensatte sentiment scorer etter dato for rå og forbehandlede tekster. Når tallene er observert, ser man at det ikke er en signifikant forskjell mellom sammensatte polariteter av rå og forhåndsbehandlede tekstdata.



Figur 14: Compound etter dato på rådata.



Figur 15: Compound etter dato på forbehandlet data.

Siden det ikke er stor forskjell mellom rå og forhåndsbehandlede data og forbehandling er alt i alt en nyttig operasjon, foretrekkes de forbehandlede dataene for resten av oppgaven.

I figur 16 er gjennomsnitt- og standardavvikverdiene til polaritetene vist. Gjennomsnittlig verdi av sammensetningen av alle tweetene i datasettet er 0,171267. Det betyr at tweetene som er relatert til Bitcoin har en svak positiv orientering. Negativ polaritet har et gjennomsnitt på 0,031121 som er veldig lavt. Samtidig ser man i figur 17 at antall tweeter med ikke-null negativ score er mindre enn halvparten av positive tweeter. Det vil si at antall negativt sentiment tweeter er mye mindre enn positivt sentiment tweeter. Antall tweeter med ikke-null-sammensetning er 2 829 890. Det vil si at det er mer enn 2 millioner nøytrale tweeter i datasettet.

	count	mean	std
compound	4985483.0	0.171267	0.381196
neg	4985483.0	0.031121	0.074590
neu	4985483.0	0.867535	0.154587
pos	4985483.0	0.100524	0.140442

Figur 16: Statistiske beskrivelse av sentiment variablene.

```
dfx[dfx['compound']!=0].shape
```

```
(2829890, 6)
```

```
dfx[dfx['pos']!=0].shape
```

```
(2377350, 6)
```

```
dfx[dfx['neg']!=0].shape
```

```
(1015090, 6)
```

Figur 17: Antall ikke-null observasjoner av variabler.

5.3 Korrelasjonsanalyse

Sentiment data, tweetvolumdata og avkastningsdata er samlet i et tidsseriedatasett etter dato. (Se figur 18) Variablene *compound*, *neg*, *neu* og *pos* er gjennomsnittsverdien av alle tweeter for hver dag. Med andre ord, er *compound* variabelen for 1. januar 2022, som er 0,260918, et gjennomsnitt av alle 146 467 sammensatte scorer av tweeter som ble lagt ut den dagen.

Variabelen *return* viser avkastningen på tidstrinn $t+1$, dvs. forskjellen mellom sluttkurs på $t+1$ og t. Det er plassert på samme dag som uavhengige variabler. Siden det er målvariabelen, har den ingen effekt på treningsprosessen. Dessuten, siden målet er å finne om det er en sammenheng mellom variabler og avkastning, er det lettere å holde oversikt og sammenligne når det er på samme dag.

	date	compound	neg	neu	pos	tweet volume	return
0	2022-01-01	0.260918	0.029950	0.774254	0.187575	93876	-0.007163
1	2022-01-02	0.244012	0.024564	0.802802	0.164214	114214	-0.018737
2	2022-01-03	0.262708	0.023801	0.795802	0.172564	157281	-0.012066
3	2022-01-04	0.279713	0.023188	0.782762	0.186131	152831	-0.050734
4	2022-01-05	0.232762	0.028359	0.795028	0.167938	149229	-0.009366

Figur 18: Tidsserie formet datasett med alle variabler.

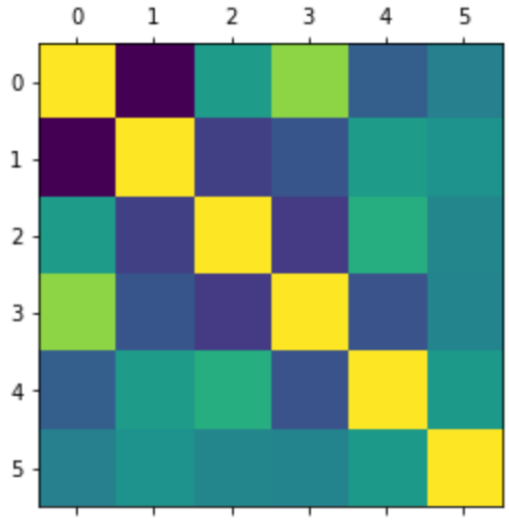
I figur 19, finnes korrelasjonsmatrisen til variablene, og figur 20 viser varmekartet til korrelasjonene. Korrelasjonsverdier mellom sentiment score variabler er ikke overraskende. Det eneste interessante med disse er at *neg* åpenbart har mer skalareffekt på *compound* enn *pos*. Dette kan forklares ved å si at tekstene med positive følelser inneholder mer nøytrale følelser enn tekstene med negative følelser.

Det er også verdt å merke at det er en svak positiv korrelasjon mellom *neg* og *tweet volume* mens *pos* og *compound* er moderat korrelert med *tweet volume*. Dette vil si at tweetvolumet er større når det er en negativ trend på tweetene mens tweetvolumet er mye mindre når tweetene er positive, eller når sammensetningen er positiv. Det er allerede nevnt ovenfor at *compound* har en positiv verdi i gjennomsnitt. Så det er ikke overraskende at *compound* har samme effekt som *pos* på *tweet volume*.

Når det gjelder målvariabelen *return*, har den nesten ingen korrelasjon med de andre variablene. Den har bare en svak korrelasjon med *tweet volume*, men den er veldig lav. De andre verdiene er veldig nær 0 som ikke representerer noen korrelasjon. Derfor er det ikke mulig å si at avkastning har noen bemerkelsesverdig korrelasjon med de andre variablene.

	compound	neg	neu	pos	tweet volume	return
compound	1.000000	-0.892580	0.144254	0.681905	-0.321074	-0.065144
neg	-0.892580	1.000000	-0.535232	-0.387079	0.146638	0.075039
neu	0.144254	-0.535232	1.000000	-0.556332	0.285564	-0.024541
pos	0.681905	-0.387079	-0.556332	1.000000	-0.405083	-0.038138
tweet volume	-0.321074	0.146638	0.285564	-0.405083	1.000000	0.128816
return	-0.065144	0.075039	-0.024541	-0.038138	0.128816	1.000000

Figur 19: Korrelasjonsmatrise.



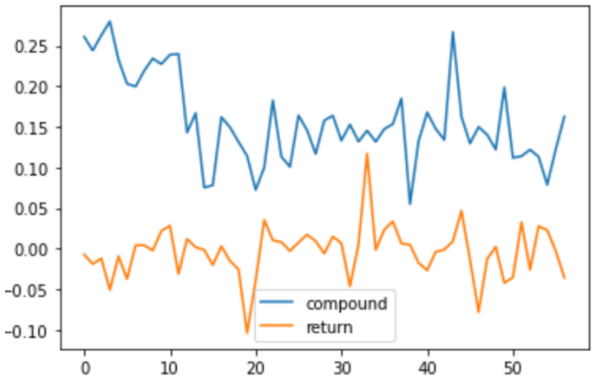
Figur 20: Varmekart av korrelasjoner.

P-verdier for korrelasjonene mellom uavhengige variabler og målvariabel er beregnet individuelt og representert i figur 21. Ingen av korrelasjonene er statistisk signifikante siden ingen av p-verdiene er mindre enn 0,05. Tvert imot er p-verdiene svært høye. Dette betyr at det er vanskelig å antyde at variabler har effekt på målvariabelen. Det kan være en effekt, men effektstørrelsen kan være for liten, prøvestørrelsen kan være for liten, eller volatiliteten kan være for stor til å bli oppdaget av hypotesetesten.

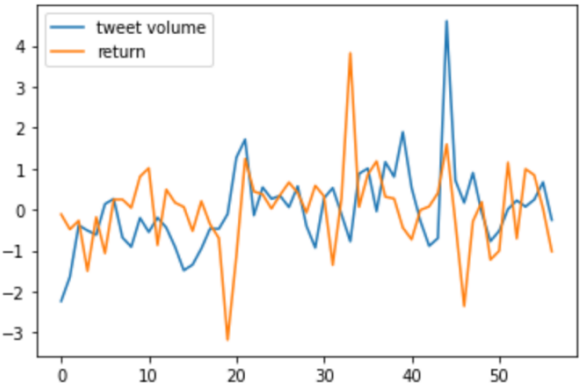
```
                    P-values

compound vs return     :  0.6302007469615262
neg vs return          :  0.5790552628360236
neu vs return          :  0.8562059031754261
pos vs return          :  0.7782060313239408
tweet volume vs return :  0.33958923521086537
```

Figur 21: De kalkulerte p-verdiene til korrelasjoner av variabler.

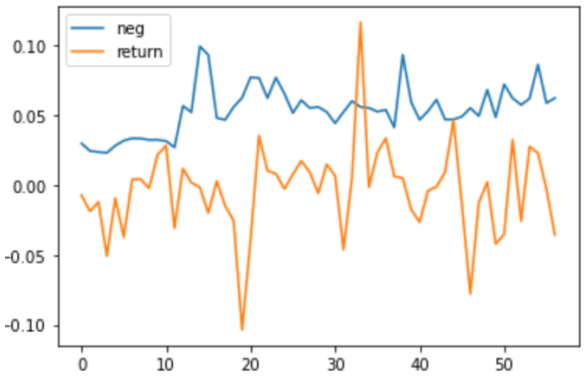
Kurvene nedenfor viser verdier av variabelpar for hele perioden. Figur 23 har de skalerte verdiene mens de andre figurene (Figur 22, Figur 24 og Figur 25) har reelle verdier. Som det er konkludert fra korrelasjonsverdiene, her ser man på diagrammene at det ikke er mulig å finne en nevneverdig korrelasjon mellom gitte variabler og avkastning.



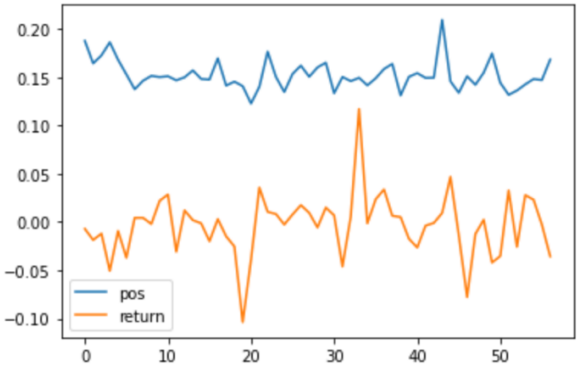
Figur 22: Verdier av compound og return over tid.



Figur 23: Skalerte verdier av tweet volume og return over tid.



Figur 24: Verdier av neg og return over tid.



Figur 25: Verdier av pos og return over tid.

5.4 Bivariat LSTM analyse

Bivariat analyse ble utført for variablene *compound*, *neg*, *pos* og *tweet volume* mot *return* som er den avhengige variabelen. Fire individuelle bivariate analyser ble utført med gitte variabler for å se om variablene hver for seg har informasjon som gjør at modellen kan gi bemerkelsesverdige prediksjoner. Sammenligningen av modellene er vist i tabell 1 nedenfor. Hyperparameteroptimaliseringen er gjort manuelt. Ulike kombinasjoner av antall noder og lag ble testet i hver modell, og alle modellene har gitt best resultat med samme node- og lagkombinasjon. Hver modell har et inndatalag med 10 noder, ett skjult lag med 30 noder og et dense lag som utgangslag med 1 node.

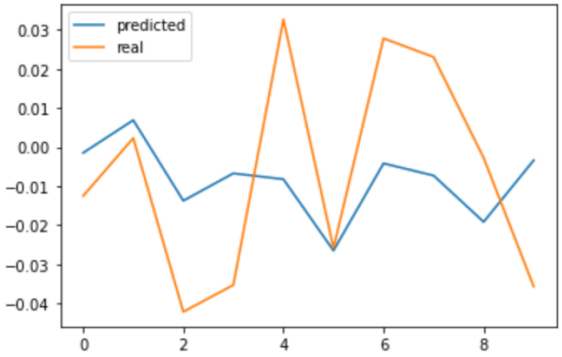
n_in er antall etterslep-observasjoner for variablene. Den gir hvor mange tidstrinn variablene går tilbake i tid. Et datasett med to variabler med *n_in* = 2 vil se ut som i figur 7. *Patience* er antall epoker uten forbedring i gitt tapsfunksjon. Dette er en parameter for Keras sin EarlyStopping-funksjon som stopper treningen hvis tapsfunksjonen ikke reduseres i gitt antall epoker.

Tapsfunksjonen velges mellom gjennomsnittlig absolutt feil (mae) og gjennomsnittlig kvadratfeil (mse). Rot middel kvadratiske feil (RMSE) verdier beregnes ved å bruke reelle *return* verdier og predikerte *return* verdier for å kunne evaluere modeller og foreta sammenligning. Det er ikke en viktig forskjell mellom RMSE-verdiene mens analysen mellom *compound* og *return* har minst RMSE. Beregnede RMSE-scorer er ikke lave nok, men dette er akseptabelt fordi det er naturlig for enkelte variabler å ikke innebære prediktiv informasjon om en kompleks variabel som prisavkastning.

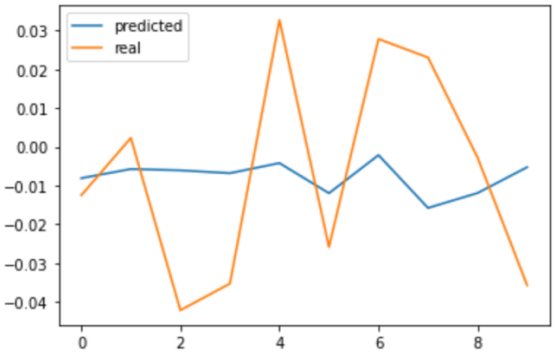
Figurene nedenfor tabell 1 viser resultatene til de bivariate analysene. I figur 26 står kurven over predikerte og reelle testverdier av *compound* vs. *return* bivariat analyse. Det kan utledes fra figuren at forbindelsen i seg selv ikke er veldig god til å forutsi volatilitetens omfang, men den er flink til å forutsi retningene til bevegelser til en viss grad. *Compound* og *return* har med andre ord samme trend oppover eller nedover generelt. På den annen side er de andre variablene veldig dårlige til å forutsi både verdier, volatilitet og retning. Figurene 27, 28 og 29 viser at disse variablene hver for seg har ingen eller svært liten likhet på avkastningen. Deres forutsigelser er spredt i nærheten av sentrum og dermed er RMSE ikke mye høyere enn *compound* vs. *return*.

variables	RMSE	n_in	patience	loss func.
compound vs. return	0.026	6	5	mse
neg vs. return	0.027	5	5	mse
pos vs. return	0.029	6	5	mse
tweet volume vs. return	0.028	5	3	mse

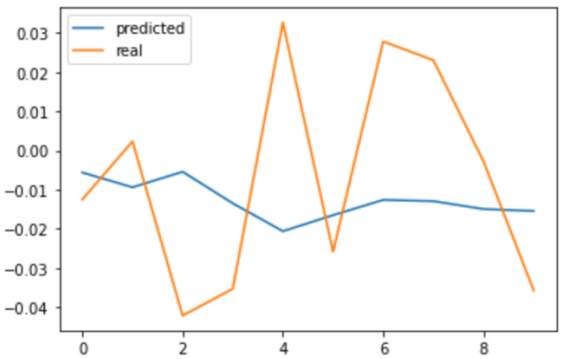
Tabell 1: RMSE verdier og optimale parametere i bivariate analyser.



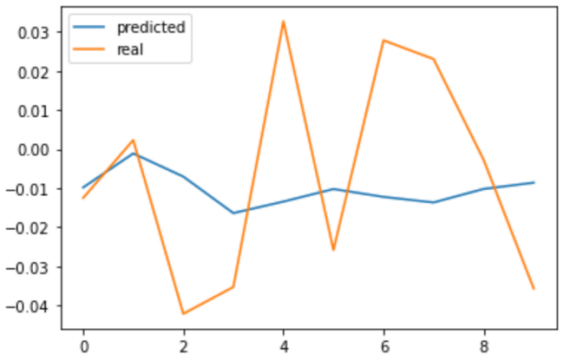
Figur 26: Bivariat analyse på compound og return.



Figur 27: Bivariat analyse på neg og return.



Figur 28: Bivariat analyse på pos og return.



Figur 29: Bivariat analyse på tweet volume og return.

5.5 Multivariat LSTM analyse

I multivariat analyse ble alle variablene som er nevnt ovenfor i datasettbeskrivelse brukt, inkludert selve *return* (ved *t-n*), og målvariabelen *return* er avkastningen på tidspunktet *t*. Modellinnstillingen er gjort manuelt også i multivariat analyse. Ulike kombinasjoner av antall noder og lag ble testet. I motsetning til bivariat analyse, fungerte lite antall lag ikke veldig effektivt for den multivariate analysen. Den beste resultat har blitt oppnådd med et inngangslag med 100 noder, 6 skjulte lag med henholdsvis 150, 300, 600, 300, 100, 50 noder, og et dense lag med 1 node som utgangslag.

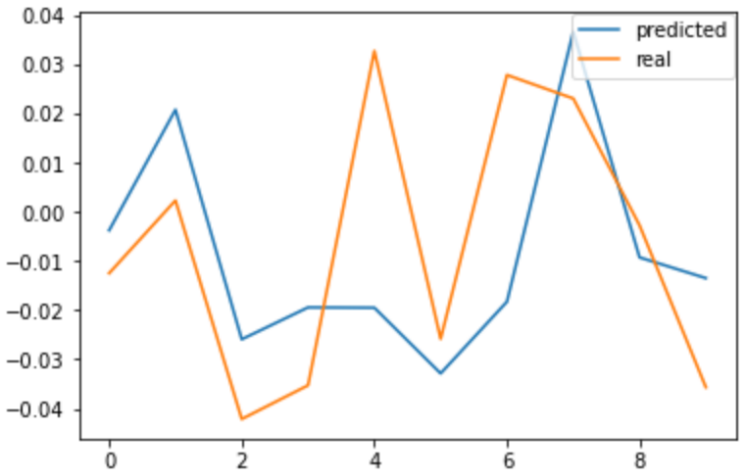
n_in er den samme parameteren som *n_in* i bivariat analyse og bestemmer antall etterslep av observasjoner som skal opprettes. I figur 30 representeres et utvalg av datasettet etter at lag er opprettet. Den beste *n_in*-verdien for den multivariate analysen ble observert som 4. Optimal *patience*-verdi for tidlig stopp av treningen oppdages som 8. Både mae og mse ble testet og mse er bestemt som tapsfunksjon.

date	var1(t-4)	var2(t-4)	var3(t-4)	var4(t-4)	var5(t-4)	var6(t-4)	var1(t-3)	var2(t-3)	var3(t-3)	var4(t-3)	...	var3(t-1)	var4(t-1)
2022-01-05	0.260918	0.029950	0.774254	0.187575	93876.0	-0.007163	0.244012	0.024564	0.802802	0.164214	...	0.782762	0.186131
2022-01-06	0.244012	0.024564	0.802802	0.164214	114214.0	-0.018737	0.262708	0.023801	0.795802	0.172564	...	0.795028	0.167938
2022-01-07	0.262708	0.023801	0.795802	0.172564	157281.0	-0.012066	0.279713	0.023188	0.782762	0.186131	...	0.803127	0.152794
2022-01-08	0.279713	0.023188	0.782762	0.186131	152831.0	-0.050734	0.232762	0.028359	0.795028	0.167938	...	0.816283	0.137383
2022-01-09	0.232762	0.028359	0.795028	0.167938	149229.0	-0.009366	0.203087	0.031973	0.803127	0.152794	...	0.810828	0.146254

Figur 30: Overvåket datasett med alle variabler for multivariat analyse.

Predikerte og reelle verdier for avkastning er brukt til å kalkulere RMSE-verdien som beregnes som 0,026. Det er de samme RMSE-score som bivariat analyse mellom *compound* og *return*.

Det vil si at det ikke er noen ytterligere forbedring i modellen med de andre variablene når det gjelder RMSE. Når figur 31 er analysert, ser man tydelig at den nye modellen er bedre når det gjelder å gi mer volatile resultater. Det gir også god informasjon om retningen for endringen. Modellen er imidlertid ikke bra nok til å forutsi verdiene.



Figur 31: Predikerte og reelle verdier av return etter multivariat analyse.

6 Drøfting

Hensikten med denne oppgaven var å analysere forholdet mellom sentimentet til Twitter data og avkastningen av Bitcoin-prisen, og å se om noen påvirkning av Twitter sentimentet eksisterer på Bitcoins prisavkastning. Resultatene av korrelasjonsanalysen viser tre ting. For det første har Twitter sentiment en ekstremt lav korrelasjon med prisavkastningen på Bitcoin. For det andre har daglig tweetvolum relativt større korrelasjon med avkastning på Bitcoin-prisen. Sist, men det viktigste, er ingen av disse korrelasjonene statistisk signifikante. Derfor er det ingen bevis for å hevde noen sammenheng mellom ovennevnte variabler.

Både bivariate og multivariate analyser ble utført for å forutsi Bitcoins prisavkastning. Funnene viser at Twitter sentiment kan ha en innvirkning på å forutsi retningen til de daglige prisendringene til en viss grad. Imidlertid er verken Twitter sentiment eller daglig tweetvolum eller kombinasjonen av de to i stand til å forutsi prisavkastningen til Bitcoin i henhold til resultatene. Det er flere mulige forklaringer på dette resultatet. En av dem kan være mangelen på tilstrekkelig lengde på perioden dataene tilhører. Antall tweeter er stort, men det er tweeter fra en relativt kort tidsramme. En annen forklaring ville være mangelen på støttedata. At om en tweet er retweet eller ikke, kunne ha blitt inkludert i datasettet som en dummy-variabel for eksempel. Videre, i tråd med arbeidet til Akbiyik et al. (2021), kunne tweet metadata ha vært et godt støttedata for sentimentet. Modeller og teknikker som var til nytte kan være en annen årsak til resultatet selv om de mest aksepterte metodikkene er brukt i oppgaven.

Sammenligner man resultatene med eldre studier, må det påpekes at resultatene samsvarer med tidligere funn av Kaminski (2016). De fant også at Twitter sentiment ikke har noen betydelig innflytelse på Bitcoins prisavkastning. På den annen side indikerer deres forskning at negativt sentiment har større innvirkning på prisavkastningen på Bitcoin. I motsetning til funnene deres viser funnene i denne oppgaven at positivt sentiment har klart mer effekt på Bitcoins prisavkastning. Denne inkonsekvensen kan skyldes endringen i trendene og populariteten til Bitcoin over tid.

Alt i alt var målet med oppgaven å analysere forholdet mellom Twitter sentiment og Bitcoins prisavkastning, og finne ut om Bitcoins pris kunne forutsies ved hjelp av Twitter sentiment. Resultatene av oppgaven viser at selve Twitter sentimentet, uten andre beregninger eller

informasjon om tweet eller forfatteren, og tweetvolumet nesten ikke har noen innvirkning på prediksjonen om daglig avkastning på Bitcoin-prisen.

7 Avslutning

I denne oppgaven var målet todelt: (i) å analysere forholdet og korrelasjonen mellom det sentimentale innholdet i de Bitcoin relaterte innleggene fra mikrobloggplattformen Twitter og den daglige avkastningen av prisen på den ledende kryptovalutaen Bitcoin; og (ii) å undersøke om den daglige avkastningen på Bitcoin-prisen kan forutsies ved å bruke Twitter sentiment data.

For å nå målene ble det samlet en god mengde Bitcoin relaterte Twitter innlegg, nemlig tweeter, for en viss tidsperiode ved bruk av offisielle Twitter API. De innsamlede dataene ble forbehandlet ved beskjæring og refaktorering slik at hele dataen hadde samme format og kunne på beste måte hjelpe analysen. Deretter ble det utført en sentimentanalyse for å trekke ut de subjektive meningene fra tweettekstene. Som et resultat av sentimentanalysen ble sentiment scorer av forskjellige aspekter gitt til dataene, og disse scorene ble kombinert for å lage et samlet sentiment data. Disse dataene ble støttet med voluminformasjonen til Bitcoin relaterte tweeter. Deretter ble Bitcoins prisavkastning beregnet ved å bruke daglige justerte sluttkurs av Bitcoin-pris, og det ble gjennomført en korrelasjonsanalyse på disse dataene. Dessuten er en viden brukt tidsserieprediksjonsarkitektur av tilbakevendende nevrale nettverk, Long Short-Term Memory, benyttet for å forutsi Bitcoins prisavkastning ved hjelp av Twitter sentiment og daglig tweetvolum.

Resultatene var i stor grad i samsvar med det tidligere arbeidet som ble gjort på feltet: Twitter-sentimentet i seg selv er ikke korrelert til Bitcoins prisavkastning og er ikke nok til å gi presis informasjon om forutsigelsen på Bitcoins prisavkastning. I tillegg ble det funnet at *compound* som er en samlet beregning av sentiment, og positiv sentiment informasjon, gir en bedre forståelse av prisbevegelsen til Bitcoin enn den negative sentimentet. Dessuten gir Twitter sentiment mening om retningen til prissvingningene til Bitcoin til en viss grad.

For følgende arbeider kunne det vært oppnås en bedre presisjon av prisprediksjon på Bitcoin ved å: (i) bruke et større datasett med hensyn til tidsintervall; (ii) inkludere støtteinformasjon til Twitter sentiment som retweetinformasjon, forfatterinformasjon og tweet metadata; og (iii) legge til følelser fra andre sosiale medieplattformer som Facebook, Reddit osv.

8 Referanseliste

Akbiyik, M.E., Erkul, M., Kaempf, K., Vasiliauskaite, V. and Antulov-Fantulin, N. (2021). Ask 'Who', Not 'What': Bitcoin Volatility Forecasting with Twitter Data. [online] Available at: <https://arxiv.org/abs/2110.14317> [Accessed 26 Apr. 2022].

Aravindpai (2020). *ANN vs CNN vs RNN | Types of Neural Networks*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning> [Accessed 26 Apr. 2022].

Barman, B. (2020a). *Parameter Sharing*. <https://biplabbarman097.medium.com/what-is-rnns-why-is-this-hype-e77071f0d8ba>.

Barman, B. (2020). *Recurrent Neural Network Structure*. <https://biplabbarman097.medium.com/what-is-rnns-why-is-this-hype-e77071f0d8ba>.

Basheer, I.A. and Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, [online] 43(1), pp.3–31. Available at: <https://www.sciencedirect.com/science/article/pii/S0167701200002013> [Accessed 26 Apr. 2022].

Bitcoin.org. (n.d.). *FAQ - Bitcoin*. [online] Available at: <https://bitcoin.org/en/faq#general>.

Blockchain.com. (n.d.). *Blockchain Wallet Users*. [online] Available at: <https://www.blockchain.com/charts/my-wallet-n-users>.

Brownlee, J. (2017). *How to Convert a Time Series to a Supervised Learning Problem in Python*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/>.

Connelly, B.L., Certo, S.T., Ireland, R.D. and Reutzel, C.R. (2011). Signaling Theory: A Review and Assessment. *Journal of Management*, 37(1), pp.39–67.

Floyd, D. (2021). *How Bitcoin Works*. [online] Investopedia. Available at: <https://www.investopedia.com/news/how-bitcoin-works> [Accessed 26 Apr. 2022].

Frankenfield, J. (2021). *Bitcoin*. [online] Investopedia. Available at: <https://www.investopedia.com/terms/b/bitcoin.asp>.

French, J.J. (2021). #Bitcoin, #COVID-19: Twitter-Based Uncertainty and Bitcoin Before and during the Pandemic. *International Journal of Financial Studies*, 9(2), p.28.

Gao, X., Huang, W. and Wang, H. (2021). Financial Twitter Sentiment on Bitcoin Return and High-Frequency Volatility. *Virtual Economics*, 4(1), pp.7–18.

Garcia, D., Tessone, C.J., Mavrodiev, P. and Perony, N. (2014). The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of The Royal Society Interface*, 11(99), p.20140623.

Hutchinson, A. (2020). *Twitter Adds More Users in Q2, but Sees Revenue Decline 19%.* [online] Social Media Today. Available at: <https://www.socialmediatoday.com/news/twitter-adds-more-users-in-q2-but-sees-revenue-decline-19/582222/>.

Hutto, C.J. and Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. [online] Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at: <https://github.com/cjhutto/vaderSentiment> [Accessed 26 Apr. 2022].

IBM Cloud Education (2020a). *What are Neural Networks?* [online] www.ibm.com. Available at: <https://www.ibm.com/cloud/learn/neural-networks>.

IBM Cloud Education (2020b). *What is Natural Language Processing?* [online] www.ibm.com. Available at: <https://www.ibm.com/cloud/learn/natural-language-processing>.

KAMINSKI, J.C. (2016). Nowcasting the Bitcoin Market with Twitter Signals. [online] Available at: <https://arxiv.org/abs/1406.7577> [Accessed 26 Apr. 2022].

Kristoufek, L. (2015). What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis. *PLOS ONE*, [online] 10(4), p.e0123923. Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123923>.

Lexico Dictionaries | English. (n.d.). *Sentiment Analysis | Meaning of Sentiment Analysis by Lexico*. [online] Available at: <https://www.lexico.com/definition/sentiment_analysis>.

Li, S., Huang, H. and Lu, W. (2021). A Neural Networks Based Method for Multivariate Time-Series Forecasting. *IEEE Access*, 9, pp.63915–63924.

LOUGHRAN, T. and MCDONALD, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), pp.35–65.

Mai, F., Bai, Q., Shan, Z., Wang, X. (Shane) and Chiang, R.H.L. (2015). From Bitcoin to Big Coin: The Impacts of Social Media on Bitcoin Performance. *SSRN Electronic Journal*. [online] Available at: <http://www.fmaconferences.org/Orlando/Papers/Bitcoin_FMA.pdf>.

Meland, M. and Øyen, V. (2017). Explaining Bitcoin's price fluctuations. [online] Available at: <https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2489827/Gruppe%2018_15966917_15965836.pdf?sequence=1&isAllowed=y>.

Nakamoto, S. (2008) Bitcoin: A Peer-to-Peer Electronic Cash System. <https://bitcoin.org/bitcoin.pdf>.

Øyen, S. (2018). Forecasting Multivariate Time Series Data Using Neural Networks. [online] Available at: <https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2559922/18579_FULLTEXT.pdf> [Accessed 26 Apr. 2022].

Russo, C., Madani, K. and Rinaldi, A.M. (2020). Knowledge Acquisition and Design Using Semantics and Perception: A Case Study for Autonomous Robots. *Neural Processing Letters*, 53(5).

Shen, D., Urquhart, A. and Wang, P. (2019). Does twitter predict Bitcoin? *Economics Letters*, [online] 174, pp.118–122. Available at: <https://www.sciencedirect.com/science/article/pii/S0165176518304634>.

Statista. (2022). *Bitcoin market cap 2013-2022*. [online] Available at: <https://www.statista.com/statistics/377382/bitcoin-market-capitalization> [Accessed 26 Apr. 2022].

Staudemeyer, R.C. and Morris, E.R. (2019). Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks. [online] Available at: <https://arxiv.org/pdf/1909.09586.pdf> [Accessed 26 Apr. 2022].

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, [online] 37(2), pp.267–307. Available at: <https://dl.acm.org/citation.cfm?id=2000518>.

Wang, C. (2019). *The Vanishing Gradient Problem*. [online] Medium. Available at: https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484.

www.buybitcoinworldwide.com. (n.d.). *How Many People Own, Hold & Use Bitcoins? (2022)*. [online] Available at: <https://www.buybitcoinworldwide.com/how-many-bitcoin-users> [Accessed 26 Apr. 2022].