Sentiment Analysis on Bitcoin Tweets: A Comparison of RoBERTa, BERTweet and VADER

Group 5

Abstract

This research focuses on comparing Transformer based models BERTweet and RoBERTa, with lexical based model VADER in the task of sentiment classification of tweets related to the topic of Bitcoin. In contrast to RoBERTa, BERTweet is additionally pre-trained on Bitcoin related twitter content. Research findings show that BERTweet and RoBERTa outperform VADER in the task of sentiment classification. More specifically BERTweet moderately outperforms RoBERTa in the classification of tweets expressing positive and negative sentiment, while RoBERTa significantly outperforms in the classification of negative tweets. Furthermore, findings reveal that tweet sentiment – as classified by BERTweet and RoBERTa – show similar patterns in correlation and Granger Causality with bitcoin return and bitcoin price volatility. VADER in contrast shows almost uniformly opposite results, with the exception of finding significant Granger Causality of sentiment and lagged volatility.

1 Introduction

Solving the double spend problem and thus pioneering scarcity in the digital realm, bitcoin has been establishing itself as the first asset class that transcends materiality and exists exclusively in digital form. bitcoin the currency or asset is used as a medium to transfer value between two parties without the involvement of any intermediaries, therefore eliminating counterparty risk. The transaction is carried out on Bitcoin the network, which is based on a decentralized cryptographically secured Blockchain (Nakamoto, 2008). Such structure and the engrained incentive system – making it economically rewarding for new computing power to join and secure the Network – has led Bitcoin to become one of the most secure computer networks in the world (Caselin, 2022).

In 2021 bitcoin hit several historic milestones including reaching a 1 trillion USD market capitalization and becoming the asset to reach this mark the fastest (Nagarajan, 2021).

Compared to traditional equities markets, there is a significantly larger proportion of retail investors in the bitcoin market (Dantes, 2021; Price, 2021). On average retail investors invest in a less systematic and more emotion driven way than institutional investors (Fisher and Statman, 2000), which may be reflected in bitcoins manic price behaviour and high volatility (Byström and Krygier, 2018). Twitter is a crucial medium for the retail investing Bitcoin community, as it is used to exchange opinions on daily market movements and events. Hence, analyzing twitter sentiment can act as a reasonable proxy for the general bitcoin market sentiment, which is of interest as it may give indications regarding future price and return trajectories (Gurrib and Kamalov, 2021).

BERT (Bidirectional Encoder Representations from Transformers) is a language model, which is versatile in its application, has achieved state-of-the-art (SOTA) performance on an array of NLP tasks and through the fine-tuning approach can be easily adapted to the downstream task of sentiment analysis (Devlin et al., 2018).

RoBERTa (Robustly Optimized BERT Pretraining Approach) was able to match or even outperform BERT's SOTA results. RoBERTa's main distinction from BERT is its use of an optimized pretraining procedure (Liu et al., 2019).

BERTweet inherited the architecture of $BERT_{BASE}$, i.e. the basic architectural composition of BERT, while benefiting from the improved pre-training procedure introduced with RoBERTa. BERTweet is pre-trained on a large-scale data set containing english tweets, thus enabling it to outperform RoBERTa in tweet related tasks by focusing on twitter domain specific language in

the pre-training stage (Nguyen et al., 2020).

VADER (Valence Aware Dictionary for sEntiment Reasoning) on the other hand is a lexical and rules based sentiment analyser, which is especially attuned to the task of sentiment analysis in the domain of social media (Hutto and Gilbert, 2014).

The scope of this research entails conducting sentiment analysis on Tweets related to the subject of Bitcoin and comparing the performance of the more refined Transformer based models (i.e. RoBERTa and BERTweet) to the general-purpose model VADER, which is used as base case. While both RoBERTa and BERTweet are fine-tuned to the task of sentiment analysis on bitcoin related tweets, BERTweet is additionally pre-trained on bitcoin related tweets. This research faced constraints regarding computing power, as well as bias that was potentially introduced when assigning sentiment labels to tweets manually. The research scope combined with the outlined constraints led to the following research aim and -questions which are addressed with this study:

How does the performance of Transformer Based Models (RoBERTa and BERTweet) compare to a general purpose model (VADER) in the downstream task of sentiment analysis considering the outlined constraints?

How does the performance of a further pre-trained Transformer Based Model (BERTweet) compare to an off-the shelf Transformer Based Model (RoBERTa) in the downstream task of sentiment analysis considering the outlined constraints?

Design a task investigating the empirical relationship found in the bitcoin market or traditional equities market between sentiment and another financial metric (i.e. return, volatility). Does model detection of such relationship or lack thereof may help to further distinguish model sentiment performances?

In the context of this research, it is found that the implemented BERTweet model and RoBERTa model outperform VADER in the task of sentiment classification for Bitcoin related tweets. BERTweet moderately outperforms RoBERTa in the classification of tweets expressing positive and negative sentiment, while RoBERTa significantly outperforms in the classification of negative tweets. Furthermore, findings reveal that tweet sentiment – as classified by BERTweet and RoBERTa – show similar patterns in correlation and Granger Causality with bitcoin return and bitcoin price volatility. A moderate negative correlation, on a 10% significance level, between the sentiments as classified by RoBERTa and the next day bitcoin price volatility can be observed. VADER in contrast shows almost uniformly opposite results to the BERT models, with the exception of finding significant Granger Causality of sentiment and lagged volatility.

Furthermore, findings reveal that tweet sentiment – as classified by BERTweet and RoBERTa – show similar patterns in correlation and Granger Causality with bitcoin return and price volatility.

2 Literature Review

2.1 From Seq2Seq Models to BERT

In 2014 (Ilya et al., 2014) introduced a Sequence-to-Sequence model, which was based on an encoder-decoder architecture. The encoder processes the text according to the sequence of characters and compresses it into vector representation. Subsequently the prediction of the text sequence is carried out character wise by the decoder.

In late 2014 (Bahdanau et al., 2014) introduced an attention mechanism to the decoder. This allows the decoder to indicate to parts in the text input that the encoder should pay attention to, and thus diminishing the amount of information the encoder would have to compress into a fixed-length vector.

In 2017 (Vaswani et al., 2017) introduced a *Transformer Model* which is an attention model with a simplified architecture abandoning recurrence and convolution in its use of neural networks, while being primarily based on *attention* mechanisms. More specifically, Transformer Models substitute recurrent layers, which form the structure of encoders and decoders, with *multi-headed self-attention*. This architectural improvement leads to a reduction in layer complexity and maximizes parallelization of computations, hence reducing the need of computational power to train the model.

Based on the Transformer Model introduced by (Vaswani et al., 2017), (Devlin et al., 2018) introduced the revolutionary language model BERT whose architecture resembles a stacked version of the encoders used in the Transformer model.

2.2 Sentiment Analysis and Bitcoin

Sentiment Analysis has found its popularity in the field of finance, where it is prevalently used for forecasting financial metrics such as market

return or volatility. (Deveikyte et al., 2020) use VADER to conduct sentiment analysis on Twitter data and financial news sources to examine the correlation between sentiment, stock volatility and stock returns, as described in *5.2*.

The Research by (Araci, 2019) has a stronger language model focus, investigating the sentiment analyzing performance of FinBERT, a BERT based model further pre-trained on a text corpus specific to the domain of finance. The research shows that FinBERT is able to outperform SOTA results in different NLP tasks.

Sentiment metrics also play a significant role in the crypto asset market, where social media sentiment has become a part of market indicators such as the Crypto Fear and Greed Index (Wood, 2022).

(López-Cabarcos et al., 2021) investigate how the SP 500 and social media investor sentiment as extracted by Stanford Core NLP influenced Bitcoin volatility. Their findings reveal that social media sentiment have a statistically significant effect on bitcoin volatility.

(Pant et al., 2018) investigate how twitter sentiment can add to the prediction of bitcoin price. The study finds that there is a moderate correlation between the rise of negative sentiment and a fall in bitcoin price.

(Abdali and Hoskins, 2021) investigate the predictability of bitcoin price movements based on twitter sentiment related to the subject of Bitcoin. Using SVMs' and BERT as sentiment classifiers the authors conclude that price can be predicted using sentiment with reasonably high accuracy.

The contribution of this research to existing literature will be the following:

A comparison regarding the task of sentiment classification on Bitcoin specific Twitter content with SOTA models RoBERTa and BERTweet as proposed by (Liu et al., 2019) and (Nguyen et al., 2020) as well as VADER as proposed by (Hutto and Gilbert, 2014). The novelty comes from the fact that BERTweet has additionally been pre-trained on Bitcoin related twitter content.

An investigation of the empirical relationships as found by (Deveikyte et al., 2020) and (Pant et al., 2018) with each of the aforementioned models, and inspect if similar relationships can be found.

3 Data and Labelling Procedure

3.1 Twitter Kaggle Data Sets

Two data sets are used for this research that are publicly accessible on the KAGGLE platform and comply with the research ethics guidelines from University College London. The data sets contain tweets that mainly address the subject of Bitcoin.

3.1.1 Long-Time Frame (LTF) Dataset

(Kash, 2022) data set spans over a "long" time period (Feb. 2021 to Feb. 2022) and was relatively large in size (i.e. 1.15GB and 2.5m tweets).

Due to limitations in GPU, 20k tweets were randomly sampled to facilitate usage. 2300 of those tweets were then sampled at random and labelled according to the outlined approach. The labelled tweets were used for fine-tuning BERTweet and RoBERTa to the task of sentiment analysis, while the outstanding non-labelled tweets were used to pre-train BERTweet. The labelled tweets were imbalanced in terms of distribution of sentiments (i.e c. 45% positive, c. 44% neutral, c. 11% negative). This imbalanced expression of opinion may be based on the flat price trajectory during this time period, but may also come from labelling bias.

The entire period between 17th to 30th of June 2021 was extracted from the 2.5m tweets to generate a continuous data set (Time Continues (TC) Data Set). This data set was used for *Task 2* (section *5.2*), where continuity in tweets was crucial.

3.1.2 Short-Time Frame (STF) Dataset

(Gulsen, 2021) data set covers a short-time period (i.e. Feb. 2021 to mid. March 2021) and is used for *Task 1* (section *5.1*). Due to time-constraint only 946 tweets were labelled. A similar imbalance in sentiment was noted where approximately 40%, 53% and 7% were labelled of positive, neutral, and negative sentiment, respectively.

3.2 Labelling Procedure

Tweets are labelled numerically with labels -1, 0, 1, indicating *negative*, *neutral*, and *positive* sentiment. To appropriately label tweets according to these sentiments Refinitiv's MarketPsych framework (Refinitiv) was utilized to further separate tweets into four categories: *Emotions*, *Market Fundamentals*, *Innovation* and *Risk*.

In tweets belonging to the category of *Emotions* users express joy or sadness regarding bitcoin, while *Market Fundamentals* tweets discuss

topics like past or future price movements as well as bitcoin adoption trajectories. *Innovation* tweets generally discuss forks and innovation in crypto currencies. Risk tweets discuss topics, which are seen as potential threats to bitcoin adoption such as SEC regulation or governmental concerns regarding environmental impact of bitcoin mining.

Classifying tweets according to these four categories gives an appropriate framework which facilitates the assignment of sentiment classes.

As bitcoin is part of the Crypto-Asset-Ecosystem, the topic of Bitcoin and matters related to other crypto projects, are intertwined. Therefore, in all data sets one can find tweets that express opinions that do not directly relate to the topic of Bitcoin – these are marked with a neutral label.

4 Model Implementations

4.1 RoBERTa and BERTweet

4.1.1 Pre-processing

Before pre-training raw tweets were pre-processed. In case of BERTweet the pre-processing procedure as described by (Nguyen et al., 2020) was followed:

Identify English tweets using fastText and tokenize using "TweetTokenizer". Convert emojis into text strings using the *emoji* package. Normalize tweets, including the conversion of user mentions and URL links to special tokens. Filter out retweets and tweets with less than 10 and more than 64 tokens.

In the case of RoBERTa additionally all letters were converted to lower case, only duplicate tweets were filtered out, while retweets were given a special token and hashtags before words were removed.

4.1.2 Architecture, Pre-Training and Fine-Tuning

Both Transformer models completed two training stages for task specific implementation:

During ***pre-training*** a model learns to capture the concept of language and the context in which specific language is used (Devlin et al., 2018).

The BERTweet model has already concluded this pre-training phase using 80 GB of 850 million English tweets (Nguyen et al., 2020), but to generate further performance improvements BERTweet was additionally pre-trained on c. 17k tweets focused on the topic of Bitcoin.

On the other side, RoBERTa was not further pre-trained and hence in that respect can be seen more as an off-the shelf model. RoBERTa is implemented with its original pre-training as carried out by (Liu et al., 2019) using 160 GB of English text coming from different language corpora.

During ***fine-tuning*** the models are geared towards the task of sentiment analysis, more specifically classifying tweets according to the sentiment they express towards Bitcoin.

The intricacy of BERTweet is that it inherited the architecture of $BERT_{BASE}$, but shares the pre-training procedure of RoBERTa (Nguyen et al., 2020), which consists of ***Masked Language Modelling (MSM)*** with dynamic masking:

Dynamically mask 15% of the tokens of an input sequence and let the model predict the masked tokens. Dynamic masking implies that the masking pattern changes each time an input sequence is fed to the model (Liu et al., 2019).

The further pre-training specifications or design choices slightly vary between RoBERTa and BERTweet (Nguyen et al., 2020; Liu et al., 2019):

Model Input Format: In the case of RoBERTa pre-training inputs are full sentences with length not exceeding 512 tokens and eventually coming from different documents.

In the case of BERTweet the pre-training length of each sequence does not exceed 128 tokens.

Pre-Training Time and Batch Size: RoBERTa uses a batch size of 8k, which ameliorates accuracy when carrying out sentiment analysis as end task, while also improving perplexity regarding the masked language modelling objective. The step size amounts to 500k.

BERTweet uses a slightly smaller batch size of 7k, following the optimization model *Adam* as proposed by (Kingma and Ba, 2014), while training steps significantly exceed those of RoBERTa with 950k training steps.

Both models are pre-trained significantly longer than BERT where pre-training involves a batch size of 256 sequences over 1m training steps (Devlin et al., 2018).

Text Encoding: For RoBERTa, as guided by (Radford et al., 2018), pre-training is carried out with byte-level BPE vocabulary of 50k sub words, while the input was not additionally tokenized or

pre-processed. The pre-processing step mentioned in the beginning of this section is implemented on the data for the fine-tuning part of the model.

For pre-training, in the case of BERTweet, byte-level vocabulary of 64k sub words is used, while tokenization is carried out as described in the beginning of this section. To further the performance of the model 50 additional tokens are introduced for the scope of this research. These words are either suggested as important in the context of Bitcoin (bitcoin.org, 2022) or found to be readily used in the Bitcoin related tweets.

Further details regarding the intricacies of how the MLM pre-training procedure is conducted and how it interplays with each of the models' structures can be explained as follows.

Before the masked language sequences are fed into the models, token embeddings for each word are created with the use of *BPE* and *fastBPE*, which hold 50k and 64k sub-word units and are used for RoBERTa and BERTweet, respectively. Token embeddings are also paired with Segment and Positional embeddings, to create a final embedding vector which also holds information regarding the ordering of the input sequence (Liu et al., 2019; Nguyen et al., 2020). Such vectors are then used as input for BERTweet and RoBERTa.

As mentioned, the structure pertinent to BERT based models are the stacked encoder cells, where each encoder combines self-attention heads, layer normalization (i.e. to alleviate the problem of an internal covariate shift) with a feed forward neural network (Halthor, 2020). Besides their commonality the main difference between RoBERTa and BERTweet is that they inherited their architectural features from $BERT_{LARGE}$ and $BERT_{BASE}$, respectively . These differ as follows:

$BERT_{BASE}$: Number of layers or Encoder blocks (L) = 12; Number of self-attention heads within encoder blocks (A) = 12, Embedding dimension for word vector representation i.e. Hidden size (H) =768, Total Parameters = 110M (Devlin et al., 2018).

$BERT_{LARGE}$: L = 24; H = 1024; A= 16; Total parameters = 340M (Devlin et al., 2018).

Both models output simultaneously generated word vectors which are of the same size. These word vectors are then passed on to a fully connected layer, which have as many neurons as the number of sub word units that are present in the models' vocabulary (Halthor, 2020), i.e. 50k and 64k in the case of RoBERTa and BERTweet, respectively. Applying a SoftMax activation layer enables comparing the distribution of each of the masked words to their ground truth (i.e. the true words in a one-hot encoding form) and allows the training of the models by minimizing the cross-entropy loss (Halthor, 2020).

After completion of pre-training the models are fine-tuned to the task of sentiment analysis.

Before fine-tuning the output layer of each of the models is substituted with a layer that can produce the appropriate output to master the given task – in this case a layer that returns numerical values of -1, 0 and 1 representing the different classes of sentiment.

Fine-tuning then becomes a supervised learning task where tweets are classified by the model and compared to the ground truth (i.e. human rules-based annotated tweets). Model parameters are initialized with parameter values from pre-training (Halthor, 2020). During fine-tuning these are only slightly tweaked, but final layer parameters are randomly initialized and redefined completely through the minimization of a loss function (i.e cross-entropy loss). As model parameters are changed only to a small extent, fine-tuning is significantly less time intensive than pre-training, which also led to the choice to pre-train only one model i.e. BERTweet (Halthor, 2020).

4.2 VADER

VADER is a rules based sentiment analyzer that relies on lexical and syntactical features to express sentiment intensity of a given textual body (Hutto and Gilbert, 2014). The appeal to use VADER for this research is multifaceted:

In contrast to the BERT based models VADER is computationally efficient and does not have to undergo long pre-training or fine-tuning procedures but can simply be implemented using the open-source pre-built library (Hutto and Gilbert, 2014).

VADER is geared towards the task of sentiment analysis with a special focus on syntactics used in social media such as Twitter (Hutto and Gilbert, 2014).

VADER's performance rivals or even outperforms other ML approaches for general sentiment tasks (Hutto and Gilbert, 2014) but also tasks specifically involving tweets using Bitcoin domain specific lan-

guage (Preisler et al., 2019) and hence seems to be a good model to use as a base case.

VADER's compound sentiment scores are normalized to -1, 0 and 1, in accordance with thresholds as defined by (Hutto and Gilbert, 2014).

5 Experiment

5.1 Task 1: Evaluating Sentiment Performance

To evaluate the performance of BERTweet, RoBERTa and VADER, sentiment predictions on the STF data set (as described in *3.1.2*) are performed. The corresponding tweets are manually labelled, as described in *3.2*, in order to assess the *Accuracy* and the *F1-Score* for each label type. The *Accuracy* is defined as,

$$\frac{TP + TN}{FP + FN + TP + TN}$$

i.e. the proportion of examples that were correctly classified (Huilgol, 2019). *TP*, *TN*, *FP* and *FN* abbreviate True Positive, True Negative, False Postive and False Negative, respectively. The *F1-Score* is a value between 0 and 1, where a *F1-Score* closer to 1 corresponds to a better model performance (Zeya, 2021).

As the data set is imbalanced and the proportion of negative sentiments in the STF data set is only 7%, the *accuracy* can be misleading. In that case, a high proportion of TN labels can lead to a high *accuracy*, which does not reflect the model performance properly. The *F1-Score* is a common measure of the quality of a classifier and is defined as

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

where the *Precision* is the ratio of TP predictions among all positive prediction ($\frac{TP}{FP+TP}$) and the *Recall* is the ratio of TP predictions among all true predictions ($\frac{TP}{FN+TP}$) (Huilgol, 2019).

Due to the imbalance of the data set, the *F1-Score* is the preferred measure to evaluate the performance of the three models.

5.2 Task 2: Sentiment, Return and Volatility

A correlation between sentiments in tweets and the price movement on the stock market is found in (Pagolu et al., 2016). Furthermore, previous research by (Pant et al., 2018) observe a moderate Pearson correlation between a rise in negative sentiments and a fall in bitcoin prices the next day. (Deveikyte et al., 2020) report a strong and significant negative correlation between positive sentiments and the stock market volatility observed the next day. The Pearson correlation is a measure of the linear relationship between two continuous variables (Kent-State-University, 2022).

In this analysis a daily average sentiment score of the predictions of each model was built. As the daily return of bitcoin is a function of the bitcoin price, the Pearson correlation of the daily average sentiment score and the next day's return was determined. Furthermore, the correlation between sentiments and next day volatility is analysed. In light of task 1 results, correlation coefficients that align with those of previous research, may give further insights about the performance of a model.

As an additional analysis, Granger Causality Tests are conducted to assess the causality between hourly sentiment averages and lagged hourly bitcoin return and volatility. Also, continuous Bitcoin tweet data on a daily basis is limited and therefore this analysis tries to make use of average sentiments on an hourly basis.

The aim of the Granger Causality Test is to evaluate whether a predictability of one time series X (sentiment) on another time series Y (return/volatility) exists. Hence, the Granger Causality Test is used to check whether sentiments can help to forecast volatility as well as returns on hourly lags. (Bartolucci et al., 2020). In this analysis the Granger Causality test from the Python library *statsmodels* is used. The null hypothesis (H_0) is that the sentiments of Bitcoin related tweets does not Granger-cause the return or volatility of Bitcoin. The applied test returns a F-test statistic and the corresponding p-value. If the p-value is below a defined significance level, H_0 can be rejected on the defined significance level (statsmodels).

6 Results

6.1 Task 1

As exhibited by *Table 1* both BERTweet (BT) and RoBERTa (RT) obtain significantly higher F1-scores than VADER (VD) for every sentiment label. This means that the transformer based models significantly outperform VADER in the sentiment classification of Bitcoin related tweets that express negative, neutral and positive sentiments. The accuracy of BERTweet and RoBERTa is also significantly higher when compared to VADER, which is

Table 1: Model F1-score and Accuracy for Task 1

		BT	RT	VD
F1	-1	0.309	0.400	0.220
	0	0.724	0.690	0.511
	+1	0.677	0.657	0.411
Acc.		0.658	0.649	0.442

however less meaningful due to the imbalance of tweets sentiment in the data set.

Furthermore, RoBERTa drastically outperforms BERTweet in the classification of negative tweets. However, BERTweet appears to have a better performance when the task is to classify tweets that express neutral or positive sentiment. It is to note that the outperformance of BERTweet over RoBERTa, for the classification of neutral and positive sentiment tweets is comparatively moderate when compared to the outperformance of RoBERTa over BERTweet in the classification of negative sentiment tweet.

6.2 Task 2

Table 2: Pearson Corr. Coeff. of Sent. – Ret./Vola.

Next Day	BT	RT	VD
Ret.	0.326	0.332	-0.226
Vola.	-0.345	-0.463*	0.204

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

As exhibited by *Table 2* the correlation found between sentiment, as classified by BERTweet and RoBERTa, and next day bitcoin return indicates to a weak positive relationship. In contrast a weak negative correlation between these two variables was found, when VADER was used as sentiment classifier. However, all results are statistically insignificant.

Interestingly, a similar pattern of discrepancy of correlation results for sentiment and next day bitcoin volatility can be found when sentiment is predicted by the BERT based models compared to when sentiment is predicted by VADER.

In the case when sentiment prediction is carried out by RoBERTa and BERTweet a weak and moderate negative correlation between twitter sentiment and bitcoin volatility is found, respectively.

On the other hand, in the case of sentiment classification carried out by VADER, a weak positive correlation was found for this pair of variables.

Only the correlation between RoBERTa classified sentiment and next day bitcoin volatility is statistically significant at a 10% significance level.

Table 3: P-values of Granger Causality between Model Sentiment and Time Shifted Return/Volatility

Type	TS	BT	RT	VD
Ret.	1h	0.001^{***}	0.003^{***}	0.553
	2h	0.002^{***}	0.011^{**}	0.784
	3h	0.004^{***}	0.015^{**}	0.908
Vola.	1h	0.000^{***}	0.002^{***}	0.000^{***}
	2h	0.001^{***}	0.018^{**}	0.009^{***}
	3h	0.003^{***}	0.045^{**}	0.016^{**}

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

When conducting Granger Causality tests between hourly sentiment averages classified by BERTweet and RoBERTa and hourly returns, there is evidence on a 1% significance level that H_0 can be rejected for a 1-hour, 2-hour and 3-hour lag. Hence, there is evidence that the sentiments classified by BERTweet and RoBERTa Granger-cause the return.

In contrast, for the sentiments classified by VADER, there is not enough evidence that H_0 can be rejected for any displayed lag.

In the case where Granger Causality tests are undertaken between hourly sentiment averages (as classified by all three models) and volatility, H_0 can be rejected for all time lags on a 5% significance level.

Only when sentiment is classified by BERTweet, there is evidence on a 1% significance level that sentiment Granger-causes future bitcoin price volatility on all time lags.

6.3 Discussion of Results

As indicated by the obtained F1-scores, the transformer based models significantly outperform VADER in the task of sentiment classification of bitcoin related tweets.

However, it is to note that even though BERTweet is a tweet specific model and was further pre-trained on bitcoin specific tweets, it only marginally outperforms RoBERTa in the classification of positive and neutral tweets but underperforms in the classification of negative tweets.

This ambiguous performance of BERTweet may be explained as follows. During the further pre-

training of BERTweet, computational constraints inhibited the training to no more than three epochs, which may have limited BERTweet's performance potential compared to RoBERTa. Through availability of further computing power, future research may be able to capture BERTweets full potential and create a Bitcoin specific sentiment analyzer model.

Furthermore, it is noticeable that tweet sentiment – as classified by the better performing Transformer models – also shows similar patterns in correlation and Granger Causality with bitcoin return and volatility. Whereas correlation and Granger Causality of the financial metrics with twitter sentiment – as classified by worse performing VADER – show almost uniformly opposite results, with the exception of Granger Causality of sentiment and lagged volatility.

Additionally, it is of notice that when sentiment is classified with RoBERTa, findings for the bitcoin market appear to be analogous to findings for the traditional equities market (Deveikyte et al., 2020). (Deveikyte et al., 2020) finds a strong negative Pearson correlation between positive sentiment and next day stock volatility, while this research finds moderate negative correlation between sentiment and next day bitcoin price volatility. (Deveikyte et al., 2020) findings are significant at a 1% level, on the other side findings of this research are less statistically significant, i.e. significant at a 10% level. Sentiment as classified by BERTweet show correlation results that yield in a similar direction, which however cannot be classified as statistically significant. While (Deveikyte et al., 2020) utilized twitter sentiment of one month and aggregated it to an overall sentiment score, including news articles collected over a period of 4 months, this analysis only uses sentiment data of 14 days to reach this result. Hence, similarities in correlation of sentiment and next day volatility between the equities market and the bitcoin market would have to be verified in further research with sentiment ranging over a longer period of time.

The Pearson correlations between twitter sentiment, as classified by all three models, and the bitcoin return has not been found to be significant in this experiment. Therefore

Thus with this research , the correlation between twitter sentiments and stock market price movements, as found in (Pagolu et al., 2016), cannot be confirmed to be congruent in the bitcoin market.

Furthermore, there is evidence for Granger Causality between the sentiment, as classified by BERTweet and RoBERTa, and hourly shifted bitcoin returns, while sentiments classified by VADER do not show this relationship. However, the sentiments, as classified by all three models, seem to Granger-cause hourly shifted bitcoin price volatility.

7 Conclusion

This research focuses on comparing Transformer based models BERTweet and RoBERTa, with lexical based model VADER in the down stream task of sentiment analysis on tweets related to the topic of Bitcoin. In particular BERTweet is developed as a bitcoin specific language model and hence is further pre-trained on Bitcoin related tweets.

In the experiment, BERTweet and RoBERTa outperform VADER in the classification of tweets expressing positive, neutral and negative sentiment. BERTweet marginally outperforms RoBERTa in the classification of neutral and positive tweets, while RoBERTa more significantly outperforms BERTweet in the classification of negative tweets.

When sentiment is classified with RoBERTa, correlation regarding sentiment and next day volatility for the bitcoin market appear to be analogous to findings for the traditional equities market as exhibited by the work of (Deveikyte et al., 2020). This result would have to be verified by further research.

It is also found that tweet sentiment – as classified by the Transformer models – show similar patterns in correlation and Granger Causality with bitcoin return/volatility.

The manual labelling procedure applied in this research project may introduce high bias. Further research may focus on defining an universally applicable labelling method for Bitcoin related tweets. As this research attempts with *Task 2*, further research could explore strategies regarding the evaluation of model performance without having to rely on labelling.

References

Sara Abdali and Ben Hoskins. 2021. Twitter sentiment analysis for bitcoin price prediction.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Ben-

gio. 2014. Neural machine translation by jointly learning to align and translate.

Silvia Bartolucci, Giuseppe Destefanis, Marco Ortu, Nicolas Uras, and Roberto Tonelli. 2020. The butterfly "affect": impact of development practices on cryptocurrency prices. *EPJ Data Science*.

bitcoin.org. 2022. Some bitcoin words you might hear.

Hans Byström and Dominika Krygier. 2018. What drives bitcoin volatility? Technical report, The Knut Wicksell Centre for Financial Studies, Lund, Sweden.

Ben Caselin. 2022. Why bitcoin mining is a matter of national security.

Damanick Dantes. 2021. Bitcoin 1q retail flow exceeding institutional investment: Jpmorgan strategist.

Justina Deveikyte, Helyette Geman, Carlo Piccari, and Alessandro Provetti. 2020. A sentiment analysis approach to the prediction of market volatility.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Kenneth L. Fisher and Meir Statman. 2000. Investor sentiment and stock returns. *Financial Analysts Journal*, 56(2):16–23.

Furkan Gulsen. 2021. Bitcoin sentiment analysis.

I. Gurrib and F. Kamalov. 2021. Predicting bitcoin price movements using sentiment analysis: a machine learning approach. *Studies in Economics and Finance*, 56(2):16–23.

Ajay Halthor. 2020. Bert neural network.

Purva Huilgol. 2019. Accuracy vs. f1-score.

C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Sutskever Ilya, Vinyals Oriol, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Kash. 2022. Bitcoin tweets.

Kent-State-University. 2022. Spss tutorials: Pearson correlation.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

M. Ángeles López-Cabarcos, Ada M. Pérez-Pico, Juan Piñeiro-Chousa, and Aleksandar Šević. 2021. Bitcoin volatility, stock market and investor sentiment. are they connected? *Finance Research Letters*, 38:216–225.

Shalini Nagarajan. 2021. Bitcoin hit the $1 trillion market cap milestone twice as fast as amazon and three times faster than apple, data shows.

Satoshi Nakamoto. 2008. Bitcoin: A peer-to-peer electronic cash system.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets.

Pagolu, Challa, Panda, and Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements.

Dibakar Raj Pant, Prasanga Neupane, Anuj Poudel, Pokhreland Anup Kumar, and Bishnu Kumar Lama. 2018. Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. IEEE.

Bernhard Preisler, Margot Mieskes, and Christoph Becker. 2019. Bitcoin value and sentiment expressed in tweets.

Mike Price. 2021. Bitcoin 1q retail flow exceeding institutional investment: Jpmorgan strategist.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Refinitiv. [link].

statsmodels. statsmodels tsa stattools grangercausalitytests.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Jackson Wood. 2022. The crypto fear and greed index, explained.

Zeya. 2021. Essential things you need to know about f1-score.