Prediksi Harga Bitcoin Menggunakan Transformer: Mengintegrasikan Harga, Sentimen, Tren, dan Volume dalam Analisis Deret Waktu Multivariat

Gilang Islamay Putra Djuharis

UIN Syarif Hidayatullah Jakarta

Abstract

Syalalalala

# PENDAHULUAN

Perkembangan teknologi tidak hanya terbatas pada sektor perindustrian, tetapi juga mencakup kemajuan teknologi dalam dunia ekonomi dengan adanya perkembangan mata uang digital yang dikenal sebagai *cryptocurrency* atau mata uang kripto. Mata uang kripto memungkinkan pengembangan sistem yang bekerja secara mandiri tanpa melibatkan peran perantara seperti bank.

Menurut Google Trends, istilah "*cryptocurrency*" mencapai puncak popularitasnya pada bulan Mei 2021. Dengan popularitas mata uang kripto di kalangan masyarakat, banyak investor yang melihat peluang keuntungan. Data dari coinmarketcap juga menunjukkan bahwa harga Bitcoin, salah satu mata uang kripto terpopuler, mengalami kenaikan sebesar 405% selama tahun 2020 dan 161% selama tahun 2021 dengan nilai *market capitalization* atau nilai total dari semua bitcoin yang beredar di pasar paling besar 1.28 trillion dollar pada 9 November 2021.

Salah satu hal menarik bagi para investor adalah pasar mata uang kripto yang buka sepanjang waktu, berbeda dengan pasar Foreign Exchange (Forex) dan saham yang hanya buka pada hari kerja dan memiliki jam tertentu. Hal ini memungkinkan para investor untuk melakukan transaksi jual-beli aset mereka dengan fleksibilitas. Namun, volatilitas harga yang tinggi pada mata uang kripto menjadi salah satu alasan mengapa beberapa investor enggan masuk ke pasar ini, harga Bitcoin sudah tepatnya 59 kali mengalami penurunan lebih dari 10% dalam sehari. Dengan penurunan yang paling besar dialami pada tanggal 13 Maret 2020 yaitu sebesar 35.19%. Oleh karena itu, diperlukan suatu model yang dapat memprediksi harga mata uang kripto di masa depan untuk membantu investor memaksimalkan keuntungan dan meminimalkan kerugian.

Banyak penelitian telah dilakukan mengenai peramalan harga Bitcoin menggunakan berbagai teknik. Sebagai contoh, penelitian yang dilakukan oleh Nasir et al. (2019) menggunakan data Google Trends dan algoritma machine learning seperti Vector Autoregression dan Random Forest untuk meramalkan pergerakan harga Bitcoin berdasarkan harga dan popularitasnya. Penelitian tersebut menunjukkan bahwa data pencarian juga dapat menjadi variabel prediktor untuk aset investasi. Penelitian lain oleh Alghamdi et al. (2022) menemukan hubungan yang kuat antara harga Bitcoin dengan sentimen, dengan Mean Absolute Error (MAE) sebesar 0.245, Mean Square Error (MSE) sebesar 0.2528, dan Root Mean Squared Error (RMSE) sebesar 0.5028. Selain itu, penelitian oleh Gemici dan Polat (2019) menyimpulkan adanya kointegrasi antara harga dan volume.

Melalui penelitian-penelitian tersebut, terlihat bahwa penggunaan teknik-teknik machine learning dan analisis sentimen dapat memberikan kontribusi yang signifikan dalam meramalkan pergerakan harga Bitcoin. Namun, seiring dengan perkembangan teknologi, terutama di bidang kecerdasan buatan, model-model yang lebih canggih dan efektif telah muncul. Salah satu model paling mutakhir per-2023 adalah model transformer dengan menggunakan mekanisme *self-attention* (dijelaskan lebih lanjut pada bab 4), yang juga semakin populer seiring dengan kemunculan Chat GPT (Chat Generative Pre-Trained Transformer) yang menarik perhatian dunia. Menurut Google Trends, kata kunci "Transformer Deep Learning" dan "Transformer Model" mengalami peningkatan popularitas sejak awal tahun 2022 dan mencapai puncaknya pada bulan Maret dan Juni 2023.

Sentimen analisis bukanlah hal yang baru, sebelum ditemukannya transformer telah banyak model sentimen untuk berbagai macam bidang. Dari bidang sosial, kesehatan, maupun politik. Meskipun demikian, masih sedikit model sentiment analysis yang tertuju pada bidang mata uang kripto khususnya menggunakan infrastruktur transformer. beberapa diantaranya adalah cryptobert oleh ElKulako, dimana model ini dilatih menggunakan 3,2 juta post sosial media seperti StockTwits, Telegram, Reddit, serta twitter mengenai mata uang kripto. Dan CryptoBERT oleh kk08.

Walaupun ElKulako melatihnya dengan data training yang besar, tetapi ketika model di uji dengan data sentimen bitcoin yang sudah dilabeli secara manual. Model hanya mendapatkan akurasi sebesar 46%. Sedangkan model oleh kk08 mendapatkan akurasi sebesar 79%, akan tetapi modelnya hanya mendeteksi dua label yaitu positif dan negatif. Sehingga model menurun menjadi 50% saat mengkalkulasikan sentimen netral.

Beberapa penelitian, seperti penelitian oleh Eray Gemici dan Müslüm Polat tentang hubungan antara harga dan volume di pasar Bitcoin menunjukkan adanya korelasi antara harga dan volume. Serta penelitian oleh Ettredge et al. yang menggunakan data pencarian web untuk memprediksi statistik macroeconomic, menemukan hubungan antara popularitas suatu topik dengan penjualan rumah dan statistik macroeconomic. Selain itu, penelitian yang dilakukan oleh Choi dan Varian dalam penelitian mereka yang berjudul "Predicting the Present with Google Trends". Berdasarkan penelitian-penelitian tersebut, dapat disimpulkan bahwa popularitas dan volume mata uang kripto memiliki korelasi dengan harga. Dalam penelitian ini, akan ditambahkan variabel sentimen media sosial untuk mengembangkan model prediksi yang lebih baru.

Dalam sisi analisis deret waktu multivariat, Mehtab dan Sen (2020) menemukan bahwa penggunaan model multivariat dalam memprediksi harga saham memberikan hasil yang lebih baik dibandingkan dengan pendekatan univariat. Hanus et al. (2022) menerapkan analisis multivariat dalam prediksi mata uang kripto dengan membandingkan tiga pendekatan menggunakan recurrent neural network (RNN), yaitu Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), dan The Gated Recurrent Unit (GRU). Mereka menggunakan lima variabel, yaitu harga tutup, harga buka, harga tertinggi, harga terendah, dan volume dari lima mata uang kripto, termasuk Bitcoin, Ethereum, Cardano, Tether, dan Binance Coin. Hasil penelitian menunjukkan bahwa Bi-LSTM dan GRU memiliki performa yang serupa dengan rata-rata Mean Absolute Percentage Error (MAPE) sebesar 0.0465712 untuk Bi-LSTM dan 0.0446512 untuk GRU, sedangkan LSTM memiliki MAPE sebesar 0.0529916. Meskipun LSTM lebih unggul dalam performa pada dataset USDT dan BNB, namun LSTM memiliki variansi yang lebih besar dibandingkan dengan Bi-LSTM dan GRU.

Model transformer telah digunakan dalam berbagai bidang, tidak hanya untuk menentukan skor sentimen dari kalimat, tetapi juga untuk peramalan data deret waktu. Zhao et al. (2022) melakukan penelitian mengenai kemampuan model transformer dalam meramalkan harga Bitcoin dan Ethereum menggunakan analisis sentimen, dan membandingkannya dengan model LSTM. Hasil penelitian menunjukkan bahwa model LSTM lebih unggul dibandingkan dengan model transformer, dengan model transformer dilatih menggunakan data Bitcoin memiliki MSE sebesar 0.00037, MAPE sebesar 0.05816, dan MAE sebesar 0.01432. Dibandingkan dengan LSTM yang memiliki MSE sebesar 0.00032, MAPE sebesar 0.04613, dan MAE sebesar 0.01346. Penelitian ini juga menemukan bahwa model yang dilatih dengan data Bitcoin memiliki peningkatan dalam memprediksi harga Ethereum dibandingkan dengan model yang dilatih dengan data Ethereum itu sendiri.

Terinspirasi oleh penelitian-penelitian yang telah disebutkan, penelitian ini bertujuan untuk menjelajahi lebih lanjut penggunaan model transformer dalam memprediksi harga Bitcoin dengan mempertimbangkan analisis sentimen dan popularitasnya menggunakan data dari Twitter, Reddit, dan Google Trends. Mengintegrasikan transformer dalam menganalisis sentimen hingga memprediksikan harga bitcoin itu sendiri dengan variabel variabel yang diperoleh. Adapun dikarenakan ada limitasiperangkat keras maka digunakan google colab untuk menjalankan programnya dan tokenisasi *natural language processing* dilimitasi sampai 256.

# DATA

## Pengambilan Data

Petentuan baru yang dikeluarkan oleh twitter mengenai daftar biaya terbaru melimitasi ruang gerak data, maka dari itu data cuitan twitter dari 05/02/2021 hingga 09/01/2023 diambil dari Kaggle untuk penelitian kali ini. Dan *scraping* sosial media reddit dilakukan dengan mengambil komen yang terdapat pada daily discussion di r/bitcoin, ini dilakukan guna mengisi kekosongan data dan meluaskan cangkupan data sentimen analisis itu sendiri menjadi 03/12/2017 hingga 30/06/2023. Serta data historis bitcoin diambil melalui API yang disediakan oleh binance. Data yang didapatkan serta digunakan sebagai berikut:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Twitter | | Reddit | | Bitcoin | | Google Trends | |
| user\_name | ✕ |  |  | Open Time | ✕ |  |  |
| user\_location | ✕ | Title | ✕ | Open | ✕ |  |  |
| user\_description | ✕ | Score | ✕ | High | ✕ |  |  |
| user\_created | ✕ | Comments | ✓ | Low | ✕ |  |  |
| user\_followers | ✕ | Author | ✕ | Close | ✓ |  |  |
| user\_friends | ✕ | Flair | ✕ | Volume | ✓ | Day | ✓ |
| user\_favourites | ✕ | Post ID | ✕ | Close Time | ✓ | Bitcoin | ✓ |
| user\_verified | ✕ | URL | ✕ | Quote Asses Volume | ✕ |  |  |
| date | ✓ | Timestamp | ✓ | Number of Trades | ✕ |  |  |
| text | ✓ | Awards | ✕ | Taker Buy Base Asset Volume | ✕ |  |  |
| hastags | ✕ | Subreddit | ✕ | Taker Buy Quote Asset Volume | ✕ |  |  |
| source | ✕ | Text | ✕ | Ignore | ✕ |  |  |
| is\_retweet | ✕ |  |  |  |  |  |  |

Dimana ✓ = digunakan, dan ✕ = tidak digunakan. Dan data google trends diambil secara manual dalam jangka tiap enam bulan dari 03/12/2017 hingga 30/06/2023 lalu data dinormalisasi dengan mencocokan tanggal yang berkesinambungan. Ini dilakukan guna mengambil data harian google trends.

## Persiapan Data

Dipanjan Sarkar dalam bukunya yang berjudul Text Analytics with Python menyatakan bahwa untuk mendapatkan hasil yang optimal dalam analisis teks dibutuhkan beberapa langkah. Yaitu tokenisasi teks dan normalisasi teks. Dimana tokenisasi melakukan pemisahan kalimat menjadi kata kata lalu dinormalisasi dengan penghapusan karakter spesial (contoh: #, @), mengubah seluruh kalimat menjadi huruf kecil, penghapusan *stopwords* (contoh: yang, di), dan lain lain.

Penelitian kali ini mengaplikasikan pra pemrosesan data tersebut dengan menghapus link di kalimat (jika ada), mengubah seluruh kalimat untuk menjadi huruf kecil, menghapus tanda baca, tokenisasi, menghapus *stopwords*, menghapus nomor dan karakter spesial, lemmatisasi, dan menghapus spasi yang tidak dibutuhkan.

Mata uang kripto yang melonjak kepopularitasannya juga menyebabkan banyaknya tweet yang bersifat spam dan dibuat oleh robot. Contoh dari tweet yang bersifat spam adalah sebagai berikut:

|  |  |
| --- | --- |
| **date** | **text** |
| **2021-07-03 20:53:51** | The $BTC price is at $34655.23 right now. 🔴 Compared to the last tweet, the price has dropped by $72.77 (-0.21%). 🟢 In the last 24 hours the price has increased by $1488.36 (4.49%).  #Bitcoin #BTC https://t.co/5wDhcU31UB |
| **2021-10-19 09:16:20** | 🐳🚨 Bitcoin Whale Alert: [ TX: 6e2f93751abd08b63306d5351da016187927ac712f442d6709d23d10b01e83c8 ]-[ ADDR: 1DuhtLa8TtCC547WfSNLHZvT91b9PQQeDD ]-[ #BTC: 16.68212472 ]-[ BLOCK\_DATE: 2021-10-19 10:37:49 ] #btc\_whale\_alert #bitcoin- BTC\_Whale\_Alert |
| **2021-06-21 07:43:39** | WhaleTrades: 📈💰$2,500,000 #bitcoin LONGED @$32,805.8359 [21/06/21 07:34:55] 🏤BitMEX | $XBTUSD  💬I’d take a fast nickel over a slow dime - buyerofblood |
| **2022-04-16 15:59:12** | 💰 Market Cap. Swap 💰 What would one #Ecash $XEC cost if it had the market capitalization of #Bitcoin  $BTC? One #XEC would be worth $0.0401409. 🤑 💪 🚀 | 👁️ More: https://t.co/fSCkKGmhcK… https://t.co/rmygNBMIhr |
| **2021-02-10 23:10:02** | Bitcoin BTC Current Price: $45,161.73 1 Hour: 1.21% | 24 Hours: -4.05% | 7 Days: 21.40% #btc #bitcoin |

## Label Manual

Pelabelan data manual telah dilakukan untuk melatih model, dengan total 3.241 data yang telah diberi label secara manual dengan sentimen negatif terdapat sebanyak 488 data, sentimen netral sebanyak 1.177 data, dan sentimen positif sebanyak 1.516 data. Selanjutnya, untuk memperoleh data tambahan, diambil 562 data dari website SurgeAI, dengan distribusi sentimen negatif sebanyak 260 data dan sentimen positif sebanyak 302 data.

Meskipun sudah demikian, jumlah total data sebanyak 3.803 dapat diperbanyak lebih lanjut melalui teknik augmentasi. Salah satu metode augmentasi yang digunakan adalah dengan menerjemahkan teks ke suatu bahasa tertentu dan kemudian menerjemahkannya kembali ke bahasa asal. Menggunakan layanan API Google Translate, dengan berdasarkan *An Updated Evaluation of Google Translate Accuracy* oleh Milam Aiken yang menganalisis akurasi google translate dengan 50. Dan pada penelitian kali ini diambil 20 bahasa paling akurat, yaitu adalah sebagai berikut

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **languages** | **bleu2** | **bleu3** | **bleuMean** |
| **0** | Italian | 100 | 90 | 95 |
| **1** | French | 89 | 88 | 88,5 |
| **2** | Swedish | 85 | 86 | 85,5 |
| **3** | Danish | 84 | 82 | 83 |
| **4** | Portuguese | 75 | 91 | 83 |
| **5** | Indonesian | 81 | 82 | 81,5 |
| **6** | Polish | 79 | 84 | 81,5 |
| **7** | Croatian | 83 | 77 | 80 |
| **8** | Bulgarian | 79 | 80 | 79,5 |
| **9** | Finnish | 82 | 77 | 79,5 |
| **10** | Norwegian | 83 | 75 | 79 |
| **11** | Russian | 74 | 84 | 79 |
| **12** | Spanish | 78 | 80 | 79 |
| **13** | Dutch | 71 | 84 | 77,5 |
| **14** | Afrikaans | 71 | 83 | 77 |
| **15** | German | 72 | 81 | 76,5 |
| **16** | Slovak | 68 | 83 | 75,5 |
| **17** | Czech | 64 | 86 | 75 |
| **18** | Latvian | 73 | 77 | 75 |
| **19** | Albanian | 70 | 80 | 75 |

Dimana blue2 adalah translasi dari bahasa inggris ke bahasa tersebut, dan blue3 adalah translasi dari bahasa asal ke bahasa inggris.

Augmentasi juga dilakukan menggunakan model translasi yang terdapat di huggingface dengan model model yang telah dilatih oleh HelsinkiNLP yaitu Universitas Helsinki yang terdapat pada Finland. Model translasi yang diambil antara lain adalah Chinese, Spanyol, Russia, Jepang, Jerman, Perancis, Italia, dan Indonesia. Ini dilakukan guna memperluas cakupan dan membandingkan model yang diaugmentasikan dengan Google Translate dan HelsinkiNLP.

# METODOLOGI

### Transformer

Model RNN yang sudah ada sejak tahun 1986 dan LSTM yang ditemukan pada tahun 1997 sama sama memiliki satu kekurangan, yaitu mereka berdua memiliki waktu yang lama dalam pelatihan modelnya. Secara mereka adalah model *recurrent* yang mana melakukan pelatihan secara berurutan satu bersatu dan tidak mendukung paraparalelisasi.

*Attention is All You Need*, adalah judul karya ilmiah diberikan oleh delapan orang dari google. Mereka mengusulkan model terbaru yang merupakan peningkatan dari model berbasis *recurrent* untuk *Natural language Processing* (NLP) dengan nama Transformer dengan menggunakan mekanisme *attention*. Dengan mekanisme ini transformer dapat melihat hubungan antara satu kata dengan kata yang lainnya, dan paralelisasi dapat dilakukan karena mengkalkulasikan hubungan suatu kata dengan kata lainnya tidak perlu mengetahui nilai kata lainnya. *Positional encoding* juga diperkenalkan untuk mengetahui posisi suatu kata relatif dengan kata lainnya. Arsitektur yang menggunakan encoder dan decoder juga membuat paralelisasi dapat dilakukan untuk lebih jauh mempercepat pelatihan data dan membuat model ini superior dibandingkan model RNN lainnya.



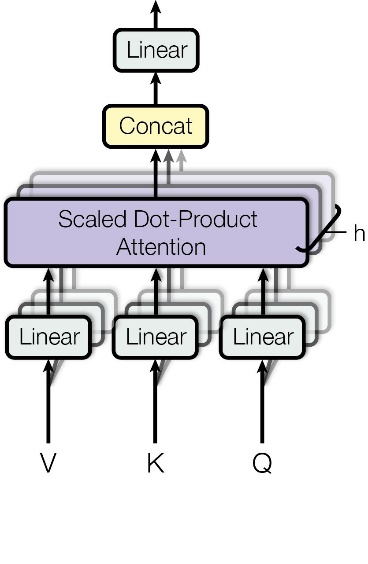
1. *Positional Encoding*

Sebelum model dimasukan kedalam encoder, karena transformer bukanlah model *recurrent* maka tidak diketahui jarak antara satu kata relatif dengan kata lainnya, maka dilakukan *positional encoding* terlebih dahulu untuk mendeterminasi posisi kata kata yang dimasukan menggunakan fungsi sin dan cos berikut.

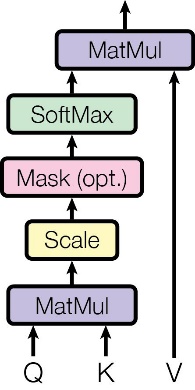
Dimana pos adalah posisi dan i adalah dimensi. Karena sin dan cos merupakan fungsi periodik, penggunaan sin dan cos dalam *positional encoding* dilakukan untuk membantu model dalam menangkap hubungan antara data meskipun jaraknya jauh.

1. *Multi-Head Attention*

Sama halnya seperti *attention* yang sudah dijelaskan, *Multi-Head Attention* memberikan nilai kedalam suatu kata relatif dengan posisi kata lainnya. Dengan *Multi-Head Attention*, transformer dapat memparalelisasi *attention* tersebut dengan menumpukannya. Rumus dan visualisasi dari *attention* adalah sebagai berikut:



Dimana *Scaled Dot-Product Attention* dapat dijelaskan lebih lanjut dari gambar berikut



Dimana Q adalah Query, Dk adalah key dari dimensi, dan V adalah value dari dimensi.

1. *Add & Norm*

Output lalu ditambahkan dengan data yang sebelumnya dan dinormalisasikan. Ini dilakukan guna menangkap informasi terbaru sembari tidak melupakan informasi yang telah ada dalam data itu sendiri, dan dinormalisasi untuk mendapatkan data yang lebih konvergen.

1. *Feed Foward*

Data yang telah melewati *Add & Norm* lalu melalui *Feed Foward Neural Network* dengan aktivasi ReLU

Ini dilakukan untuk mendapatkan komponen hubungan yang kompleks dan non-linear dari data itu sendiri.

### *Multiple Seasonal-Trend decomposition using LOESS*

MSTL adalah algoritma dekomposisi musiman-tren yang tangguh dan akurat yang dirancang untuk menangkap berbagai pola musiman dalam suatu deret waktu. Dengan memberikan dekomposisi aditif dari data rangkaian waktu. Diberikan dimana t adalah observasi pada waktu ke-t, maka dekomposisi aditifnya dapat didefiniskan sebagai berikut:

dimana ,, masing masing melambangkan musiman, tren, dan sisa dari observasi. Dan menjadi jikalau memiliki lebih dari satu musim maka dekomposisi aditifnya menjadi

dimana n adalah banyak musim yang dimiliki oleh

### Time Lag Plot

*Time Lag Plot* atau TLP memeriksa apakah data bersifat random atau tidak, dimana grafik poin digambarkan di pada grafik 2 dimensi (x-y) dan x ditetapkan pada waktu dan y ditetapkan pada waktu dimana n adalah besar *lag* yang ditetapkan.

### *Evaluasi Model*

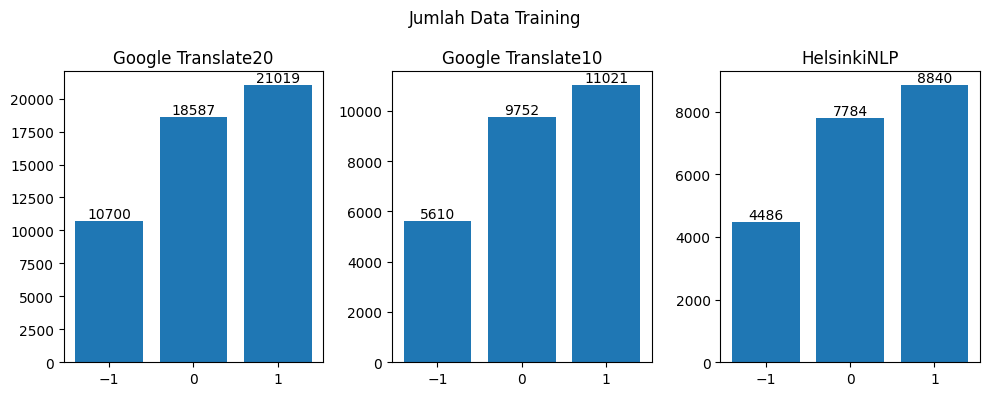
Evaluasi Model dilakukan guna membandingkan performa dari satu model dengan model lainnya. Pada data rangkaian waktu, seringkali digunakan dua evaluasi. Yaitu *Mean Absolute Percentage Error* (MAPE) dan *Symetric Mean Absolute Percentage Error* (sMAPE) dimana kedua evaluasi ini mengukur *Error* atau galat yang mana adalah jarak antara nilai sebenarnya dibandingkan dengan nilai yang di prediksi. Dimana rumus MAPE dan sMAPE adalah

Dimana masing masing dan didefinisikan sebagai nilai aktual dan nilai prediksi untuk jumlah data observasi. Perbedaan dari MAPE dengan sMAPE sendiri berdiri pada

# HASIL DAN PEMBAHASAN

#### Analisis Sentimen

Data latih yang tersedia setelah melakukan augmentasi adalah sebagai berikut:



Akan tetapi, diperlukan data yang seimbang agar pelatihan model tidak condong ke satu sentimen. Maka dilakukan *undersampling* atau mengambil jumlah data sebanyak data yang paling minoritas. Maka didapatkan jumlah data latih sebanyak berikut



Setelah semua data telah dibersihkan, pertama yang dilakukan adalah membangun model analisis sentimennya. Dikarenakan membangun model dari awal membutuhkan kekuatan komputasi yang besar, maka dilakukan *fine tuning* dari model yang sudah ada. Dengan memilih model yang paling besar akurasinya terhadap data yang telah dilabeli manual. Berikut merupakan confusion matrix dan juga akurasi kandidat model



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Acc Negative Sentiment | Acc Neutral Sentiment | Acc Positive Sentiment | Overall Accuracy |
| twitter-roberta-base-sentiment-latest | 0,59 | 0,72 | 0,27 | 0,526666667 |
| bertweet-base-sentiment-analysis | 0,53 | 0,75 | 0,24 | 0,506666667 |
| finbert | 0,05 | 0,06 | 0,91 | 0,34 |
| twitter-xlm-roberta-base-sentiment | 0,52 | 0,7 | 0,3 | 0,506666667 |
| cryptobert | 0,37 | 0,56 | 0,59 | 0,506666667 |
| twitter-xlm-roberta-bitcoin-sentiment | 0,47 | 0,51 | 0,38 | 0,453333333 |

Berdasarkan data yang terdapat dalam *confusion matrix* dan tabel di atas, dapat diketahui terdapat 4 kandidat model dengan tingkat akurasi total terbesar (diwarna biru terang).

##### Pembangunan Model Analisis Sentimen

*grid search* dilakukan dengan 3 hyperparameter, yaitu weight\_decay dengan rentang 0 sampai 0.3, learning\_rate dengan rentang 1e-5 sampai 5e-5, serta warmup\_steps antara 0, 250, dan 500. Setelah *grid search*, pelatihan ulang dilakukan kembali untuk 20 model yang memiliki akurasi terbaik dengan rentang menyimpanan model setiap 100 steps untuk mendapat model yang absolut terbaik.

##### Hasil Model Analisis Sentimen

Akurasi paling akurat yang didapatkan setelah melakukan grid search diperoleh model cryptobert yang dibuat oleh ElKulako dengan pelatihan hyperparameter learning rate sebesar 2.164916560792168e-05, per\_device\_eval\_batch\_size sebesar 32, per\_device\_train\_batch\_size sebesar 16, warmup steps sebesar 500, dan weight decay sebesar 0.1295835055926347. Serta pelatihan dilakukan sebanyak 600 steps (0.57 epoch). Didapatkan hasil evaluasi sebagai berikut:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **eval\_accuracy** | **eval\_f1** | **eval\_loss** | **eval\_precision** | **eval\_recall** |
| 0,676666667 | 0,676594863 | 0,858020842 | 0,678267182 | 0,676666667 |

#### Analisis Data Eksplorasi

###### Data Sentimen Analisis

Matriks konfusi dari data uji menggunakan model yang telah dilatih adalah sebagai berikut



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sentiment Score** | **Precision** | **Recall** | **F1-Score** | **n** |
| **-1** | 0,73 | 0,76 | 0,75 | 100 |
| **0** | 0,72 | 0,48 | 0,56 | 100 |
| **1** | 0,63 | 0,83 | 0,72 | 100 |
| **Accuracy** | 0,68 | | | 300 |
| **Macro Avg** | 0,69 | 0,68 | 0,67 | 300 |
| **Weighted Avg** | 0,69 | 0,68 | 0,67 | 300 |

###### Data Rangkaian Waktu

1. Harga dibandingkan Variabel Lainnya

Dari dataset yang telah diolah, perbandingan dapat dilakukan setelah melakukan standarisasi data. Standarisasi data dilakukan dengan menormalkan nilai dari setiap variabel ke dalam rentang 0 hingga 1. Hal ini bertujuan untuk mempermudah proses perbandingan.



Dari tabel diatas dapat dilihat harga dan sentiment tidak memiliki korelasi sama sekali dari tahun ke tahun. Akan tetapi beda hal dengan tren serta volume, dari tahun ke tahun tren sangat amat berkolerasi dengan harga. Tren dan volume yang naik seringkali bersandingan dengan harga bitcoin yang naik, akan tetapi bukan berarti tren dan volume yang turun bersandingan dengan harga yang turun. Sering kali Tren dan Volume naik bersandingan dengan harga yang turun.

1. Multiple Seasonal-Trend Decomposition using LOESS



Dikarenakan data bersifat harian, maka ditetapkan 2 jangka waktu yaitu per bulan (Seasonal\_30) dan per tahun (Seasonal\_365). Komponen residual menunjukkan bahwa ada jumlah variasi yang signifikan yang tidak dapat dijelaskan dalam harga Bitcoin itu sendiri melainkan disebabkan oleh berbagai faktor lain. Dapat dilihat juga dari grafik per tahun, terdapat tendensi lebih tinggi pada bulan Desember dan rendah pada bulan Juni. Lalu untuk grafik per bulan sulit untuk diuraikan karena ukuran terlalu kecil, maka dari itu grafik bulanan dipecahkan menjadi 12 untuk tiap bulan.



Dapat dilihat pula harga bitcoin seringkali tidak memiliki tendensi untuk naik ataupun turun berdasarkan bulannya, akan tetapi pada Juli dan Oktober memiliki tendensi untuk naik. Dan pada bulan September, serta November harga memiliki tendensi untuk turun.

1. *Time* *Lag Plot* (TLP)

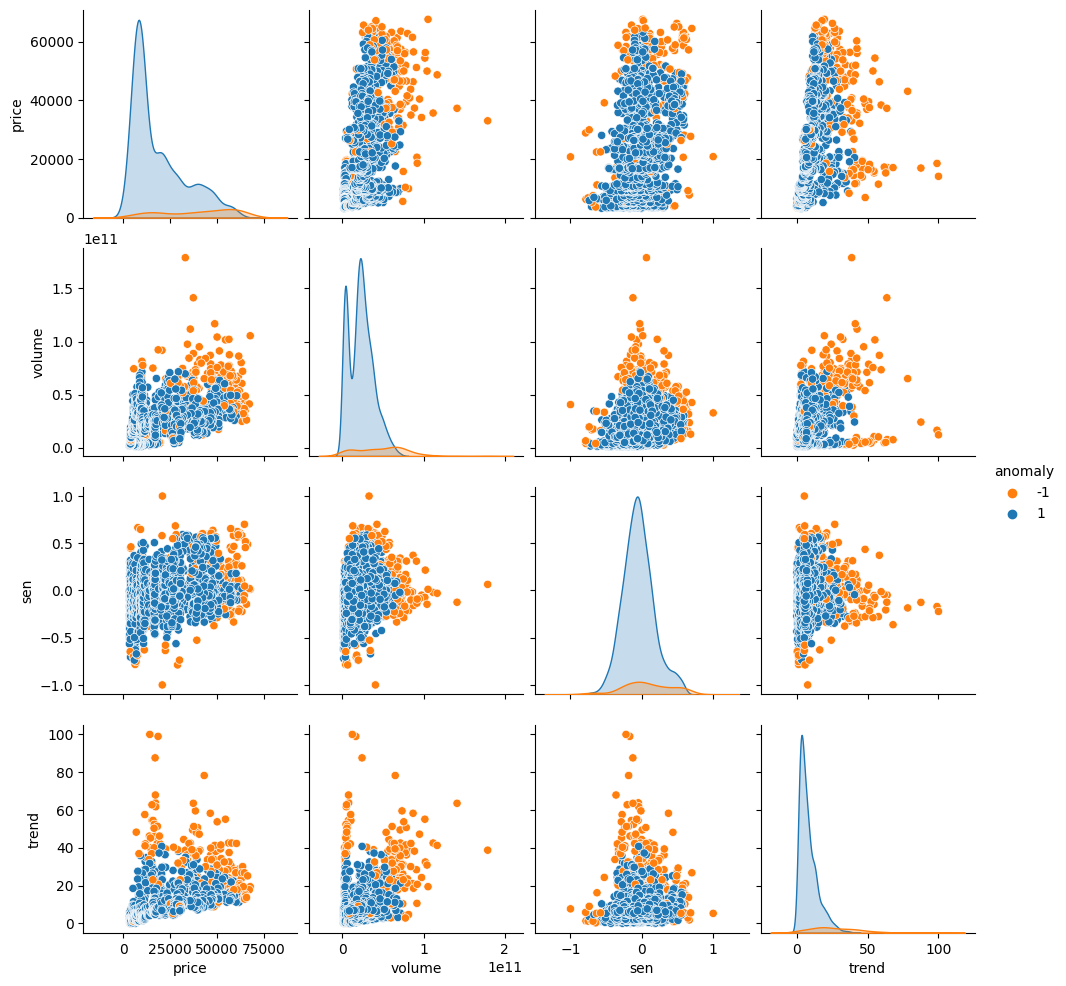
Berikut merupakan TLP harga dibandingkan dengan variabel lainnya, menggunakan ,



Harga terhadap tren dan sentimen dapat dilihat dari grafik di atas bahwa keduanya memiliki sifat autokorelasi positif yang sangat amat tinggi. Serta harga terhadap volume memiliki sifat autokorelasi yang moderat. Dan dapat dilihat dari tiga grafik di atas ada tendensi memiliki *outlier*, maka dari itu diperlukan langkah lebih lanjut untuk mendeteksi serta menangani *outlier*

1. *Outlier*

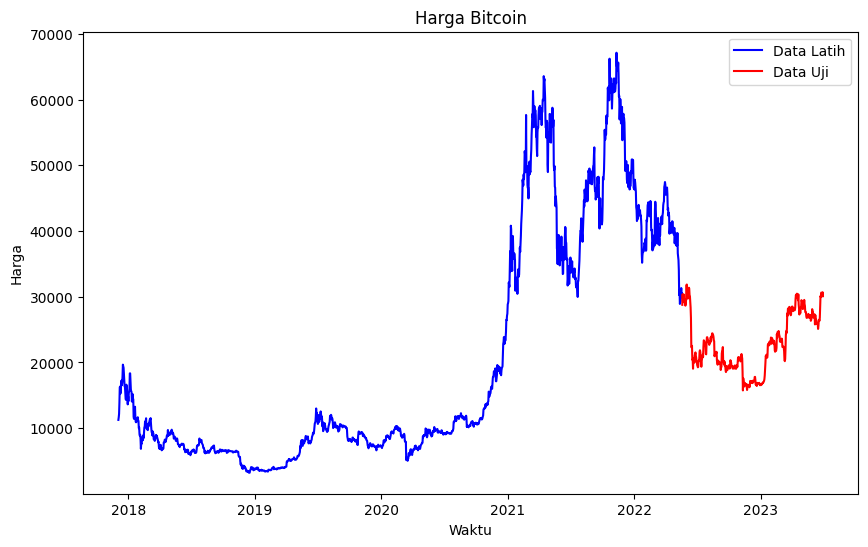
*Outlier* dalam data rangkaian waktu khususnya harga bitcoin yang memiliki sifat tidak stabil harus ditangani dengan hati hati, oleh karena itu metode Isolation Forest digunakan dalam pendeteksian *outlier* pada data ini. Pengaruh outlier terhadap kinerja modelpun dipelajari dengan tidak menghapus dan menghapus 10% dari outlier.



Berdasarkan Isolation Forest, terdapat 204 outlier berdasarkan 4 variabel yang ada, penanganan outlier dilakukan dengan mengambil 24 outlier yang memiliki score paling tinggi dan di transformasi dengan menggantikan nilai outlier menggunakan rolling average dengan window sebesar 24

#### Peramalan Deret Waktu

Data deret waktu harga bitcoin dibagi menjadi dua, yaitu 80% data latih dan 20% data uji untuk memvalidasi peramalan harga bitcoin.



1. Pembangunan Model Peramalan Deret Waktu

*grid search* mendapatkan model paling akurat dengan *hyperparameter* input\_chunk\_length sebesar, output\_chunk\_length sebesar, d\_model sebesar, nhead sebesar, num\_encoder\_layers sebesar, num\_decoder\_layers sebesar, dim\_feedforward sebesar, dan dropout sebesar.

1. Hasil Model Peramalan Deret Waktu

Akurasi paling akurat yang didapatkan setelah melakukan grid search dengan data validasi sebesar dan data uji coba sebesar . Dengan prediksi harga dibandingkan dengan harga asli dalam data uji coba sebagai berikut

\*Gambar grapik time series bang\*

# KESIMPULAN DAN SARAN

Meledaknya mata uang kripto ke masyarakat umum meningkatkan urgensi untuk pembuatan model yang kokoh dan akurat. Dengan menggunakan model paling mutakhir yaitu transformers dalam *natural language processing* dan *Time Series Forecasting*, dilakukan beberapa eksperimen.

Ditemukan bahwa menggunakan XXXX data untuk *fine tuning* model NLP yang sudah ada di Hugging Face didapatkan F1-Score sebesar XXXX. Dan untuk *Time Series Forecasting* dilakukan empat eksperimen, dan ditemukan bahwa model yang paling akurat adalah dengan

Adapun saran untuk penelitian selanjutnya dalam sisi *Natural Language Processing* adalah dengan menggunakan model yang memiliki maksimum token lebih tinggi. Dikarenakan X (dahulunya twitter) yang dahulu melimitasi sebanyak 280 karakter, per 9 Februari 2023 sudah bisa melebihi dari 4000 karakter. Ditambah data reddit dan sumber sumber lain memiliki maksimal karakter yang tidak terhingga. Serta perbandingan model translasi dengan Google Translate dapat dilakukan menggunakan bahasa yang sama untuk mengevaluasi model itu sendiri.

Penggunaan model Transformer dalam peramalan deret waktu yang lebih mutakhir juga dianjurkan. Dikarenakan model yang digunakan pada penelitian ini masih naif, yaitu tidak memanfaatkan hasil dari model itu sendiri untuk menjadi argumen konsiderasi dalam memprediksi hasil selanjutnya.