# 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 dunia perindustrian, tetapi juga berdampingan dengan kemajuan teknologi dalam dunia perekonomian, khususnya dalam transformasi dari uang fisik ke bentuk digital. *Cryptocurrency* atau mata uang kripto telah memajukan konsep ini dengan menghilangkan ketergantungan pada peran perbankan tradisional, memungkinkan sistem operasi mandiri tanpa intervensi pihak ketiga.

Google Trends mencatat bahwa kata kunci "cryptocurrency" mencapai puncak popularitas pada bulan Mei 2021. Dengan tingginya minat masyarakat terhadap mata uang kripto, banyak investor melihat peluang di pasar ini. Grafik dari CoinMarketCap juga menggambarkan lonjakan harga cryptocurrency terkemuka seperti Bitcoin, yang mengalami kenaikan sebesar 405% selama tahun 2020 dan 161% selama tahun 2021.

Pasar mata uang kripto yang buka setiap saat adalah salah satu aspek menarik juga bagi para investor, berbeda dengan pasar *Foreign Exchange* (Forex) atau pasar saham yang terbatas pada jam kerja. Kehadiran pasar yang buka 24/7 memberikan fleksibilitas bagi investor untuk melakukan perdagangan kapan saja.

Harga mata uang kripto yang melonjak serta kemampuan pasar untuk beroperasi sepanjang waktu memang faktor yang sangat menarik bagi para investor. Akan tetapi, volatilitasatau harga yang naik turun secara drastis merupakan salah satu alasan investor memilih untuk menjauh dan takut akan masuk ke dalam pasar mata uang kripto. Apalagi harga bitcoin sempat menurun dalam satu hari sebesar 39.5% pada 12 Maret 2021. Maka dari itu sangat dibutuhkan suatu model yang dapat memprediksi harga suatu kripto di masa depan untuk memaksimalkan keuntungan dan meminimalisir kerugian.

Penelitian yang dilakukan oleh Choi dan Varian dengan judul *Predicting the Present with Google Trends* serta Ettredge, dkk dengan judul *using web-based search data to predict macroeconomic statistics* mendapatkan kesimpulan bahwa kepopuleran atas suatu topik memiliki hubungan terhadap penjualan rumah bahkan sampai statistik macroeconomic. Ditambah dengan penelitian yang dilakukan oleh Eray Gemici dan Müslüm Polat mengenai *relationship between price and volume in the bitcoin market* yang menyimpulkan bahwa harga dan volume memiliki korelasi. Berangkat dari ketiga penelitian itu maka kepopuleran suatu mata uang kripto dan volumenya memiliki korelasi dengan harganya itu sendiri. Pada penelitian ini ditambahkan variabel harga serta sentimen sosial media untuk mendapatkan model terbarunya.

Kepopuleran model transformer sendiri pun semakin melonjak bersamaan dengan munculnya ChatGPT, atau Chat Generative Pre-Trained Transformer yang berhasil mengambil perhatian dunia dengan badai. Google trends menyatakan bahwa kata kunci “Transformer Deep Learning” dan “Transformer Model” mulai mengalami lonjakan dari awal tahun 2022 dan masing - masing mencapai puncak popularitas pada bulan Maret dan Juni tahun 2023.

Penelitian terkait peramalan bitcoin menggunakan berbagai macam teknik telah banyak dilakukan. Seperti pe nelitian yang dilakukan oleh Nasir, M. A., dkk (2019) yang menggunakan data google trends dan algoritma *machine learning* seperti *Vector Autoregression* serta *Random Forest* untuk meramal pergerakan harga Bitcoin berdasarkan harga dan kepopulerannya. Studi menunjukan bahwa banyak pencarian juga dapat menjadi variabel peramal untuk aset investasi. Bukan hanya dengan google trends, Alghamdi S., dkk (2022) menemukan relasi yang sangat kuat antara harga dengan sentimen dengan *Mean Absolute Error* (MAE) sebesar 0.245, *Mean Square Error* (MSE) 0.2528, dan *Root Mean Squared Error* (RMSE) 0.5028. Ditambah dengan Gemici, E. dan Polat, M (2019) yang menyimpulkan bahwa harga dan volume memiliki kointegrasi.

Analisis menggunakan deret waktu multivariat juga telah banyak dilakukan sebelumnya, Mehtab S. dan Sen J (2020) menemukan bahwa dalam memprediksi harga saham penggunaan multivariat terbukti lebih unggul dibandingkan univariat dengan membangun model yang lebih akurat. Hanus, S., dkk (2022) mengaplikasikan multivariat ke 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). Mengaplikasikan lima variabel yaitu harga tutup, harga buka, harga tertinggi, harga terendah, dan volume dari lima mata uang kripto yaitu Bitcoin, Ethereum, Cardano, Tether, serta Binance Coin. Mereka menemukan bahwa Bi-LSTM dan GRU memiliki performa yang serupa dalam rata rata *Mean Absolute Percentage Error* (MAPE) yaitu 0.0465712 untuk Bi-LSTM dan 0.0446512 untuk GRU, sedangkan LSTM memiliki rata rata MAPE sebesar 0.0529916. Walaupun LSTM unggul dalam performa di dataset USDT dan BNB, akan tetapi LSTM memiliki variansi yang lebih besar dibandingkan Bi-LSTM dan GRU

Model Transformer sendiri sudah mulai banyak digunakan bukan hanya untuk menentukan sentiment score dari suatu kalimat, akan tetapi untuk peramalan data deret waktu. Zhao H., dkk (2022) meneliti kapabilitas model transformer dalam meramal bitcoin dan ethereum menggunakan sentiment analysis dan membandingkannya dengan model LSTM. Mereka menemukan bahwa model LSTM lebih unggul dibandingkan model transformer, dengan model transformer dilatih menggunakan data bitcoin memiliki MSE 0.00037, MAPE 0.05816, dan MAE 0.01432. Dibandingkan dengan LSTM yang memiliki MSE 0.00032, MAPE 0.04613, dan MAE 0.01346. Dan yang membuat penelitian ini menarik adalah mereka menemukan bahwa model yang dilatih dengan data bitcoin memiliki memiliki peningkatan dalam memprediksi harga ethereum dibandingkan dengan model yang dilatih dengan data ethereum itu sendiri.

Terinspirasi dari penelitian penelitian yang telah disebutkan, penelitian ini diadakan untuk mengeksplorasi lebih jauh mengenai model transformer dalam memprediksi harga bitcoin dengan analisis sentimen serta kepopulerannya menggunakan twitter, reddit, dan google trends.

## Metodologi Penelitian

### Data Penelitian

### Data yang akan digunakan dalam penelitian kali ini berasal dari beberapa sumber, yaitu twitter, thread reddit daily discussion r/bitcoin, google trends, serta investing.com. Ketentuan 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 serta *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 tahun 2018 hingga 30/06/2023. Lalu data historis bitcoin diambil dari investing.com. Dan yang terakhir data google trends diambil secara manual dalam jangka tiap enam bulan dari tahun 2018 hingga 30/06/2023.

### Tahapan Penelitian



### Persiapan Data

Data text dari twitter dan reddit dipersatukan terlebih dahulu untuk dilakukan pemrosesan awal analisis sentimen. 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.

Data historis bitcoin hanya menggunakan kolom Close sebagai harga tutup di hari itu, Volume sebagai banyak pertukaran bitcoin di hari itu, serta Close Time sebagai waktu harga tutup di hari itu. Dan yang terakhir data google trends dinormalisasikan dengan mencocokan tanggal yang berkesinambungan. Ini dilakukan guna mengambil data harian google trends. Akhirnya 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.

### Sentimen Analisis

Bidirectional Encoder Representations from Transformer (BERT) menyediakan banyak model untuk mengklasifikasi sentimen dengan data teks, salah satunya adalah *twitter-XLM-roBERTa-base for Sentiment Analysis* dimana model yang dikerucutkan untuk analisis sentimen melalui data twitter dan menggunakan enam Bahasa yaitu Arab, Inggris, Prancis, Jerman, Hindi, Italia, Portugal, dan Spanyol. Serta mengklasifikasikan sentiment sebuah teks menjadi positif, netral, dan negatif. Berdasarkan alasan tersebut maka digunakan model BERT tersebut untuk melaksanakan analisis sentimen pada penelitian kali ini. Setelah data teks dikalkulasikan satu persatu menggunakan model yang telah dipilih, lalu diambil menjadi rata rata sentimennya perhari dan disatukan menjadi satu himpunan data besar berisikan tanggal, sentimen, harga bitcoin, volume bitcoin, serta trends dengan banyak 2035 data poin.

1. Pembangunan Model

Tahap terakhir dalam penelitian ini adalah pembangunan model untuk memprediksi naik turunnya harga bitcoin. Transformer adalah salah satu model terobosan terbaru yang keluar di dunia machine learning dengan menggunakan mekanisme *self-attention* sehingga transformer dapat mencapai paralelisasi. Walaupun demikian, untuk melatih model transformer membutuhkan banyak waktu dikarenakan model yang kompleksnya. Oleh karena itu digunakan model yang lebih cepat lagi yaitu Informer. Sebagaimana dijelaskan oleh Haoyi Zhou, dkk (2020) bahwa informer menggunakan mekanisme dari Vanilla Transformer tetapi daripada menggunakan *self-attention*, informer menggunakan *ProbSparse Self-Attention* yang kompleksitas waktunya adalah *O(L log L)* dibandingkan Vanilla Transformer yang memiliki kompleksitas waktu sebesar

## HASIL DAN PEMBAHASAN

1. Pengolahan Data

Data yang pertama diolah adalah data teks dari twitter serta reddit, kedua data teks disatukan terlebih dahulu menjadi satu file csv untuk dilakukan prapemrosesan. Dikarenakan data teks yang didapatkan adalah data teks yang tidak terstruktur, maka dilakukan tahap prapemrosesan data teks yang terdiri dari penghapusan URLS dan entitas HTML, penyelarasan huruf kecil, tokenisasi, penghapusan tanda baca, kata kata penghubung, karakter spesial, serta angka, dan lematisasi.

\*Before after prapemrosesan data\*

Setelah data teks melewati tahap prapemrosesan data, selanjutnya dilakukan sentimen analisis pada tiap data teks dan diambil rata ratanya untuk melihat sentimen masyarakat internet pada hari tersebut.

1. Evaluasi Model Analisis Sentimen

Sebelum data teks dilakukan sentimen analisis menggunakan BERT, model di evaluasi terlebih dahulu performanya menggunakan dataset yang sudah di label secara manual dengan banyak data sebesar 3900 data poin yang memiliki nilai sentimen yang tersebar rata sebanyak 1300 tiap nilai sentimennya. Didapatkan hasil performa dari berbagai macam model transformer yang digunakan sebagai berikut:

\*Performa Transformer\*

Dikarenakan model dapat dibilang belum memuaskan, maka harus dilakukan *fine-tune model* atau melatih model untuk lebih tertuju lagi terhadap topik yang dibahas yaitu bitcoin.

1. Pelatihan Model Sentimen Analisis

Data yang sudah dilabeli secara manual digunakan untuk melatih model untuk sentimen analisis. Rasio yang digunakan adalah sebanyak 3000 data digunakan sebagai training dan 900 data digunakan sebagai testing. Lalu juga menambahkan kosa kata istilah dalam topik bitcoin seperti hodl, hash, miner, dan lain sebagainya guna menambah performa bitcoin. Dalam tahap ini juga dilakukan grid search guna memaksimalkan performa model itu sendiri.

1. Analisa Prediksi Model

Setelah model dilatih, prediksi dapat dilakukan dengan menyembunyikan beberapa data point terakhir untuk membandingkan hasil model dan data yang asli. Berikut grafik yang diperoleh dari model transformer

\*Grafik model transformer\*

1. Evaluasi Model Transformer untuk Data Time Series

Guna mengevaluasi model transformer, hasil evaluasinya dibandingkan dengan model model multivariat lainnya. Berikut hasil performanya dibandingkan dengan model lainnya

\*Tabel perbadingan MSE, RMSE, dan MAE dari beberapa model\*

## KESIMPULAN DAN SARAN