**PREDIKSI HARGA BITCOIN MENGGUNAKAN TRANSFORMER: MENGINTEGRASIKAN HARGA, SENTIMEN, TREN, DAN VOLUME DALAM ANALISIS DERET WAKTU MULTIVARIAT**

**SKRIPSI**

**A picture containing logo, graphics, symbol, graphic design

Description automatically generated**

**Gilang Islamay Putra Djuharis**

**11190940000055**

**PROGRAM STUDI MATEMATIKA**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UIN SYARIF HIDAYATULLAH JAKARTA**

**2024 M / 1444 H**

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

Skripsi

Diajukan kepada

Universitas Islam Negeri Syarif Hidayatullah Jakarta

Fakultas Sains dan Teknologi

Untuk Memenuhi Salah Satu Persyaratan dalam

Memperoleh Gelar Sarjana Matematika (S.Mat)

Oleh:

Gilang Islamay Putra Djuharis

11190940000055

**PROGRAM STUDI MATEMATIKA**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UIN SYARIF HIDAYATULLAH JAKARTA**

**2024 M / 1444 H**

PERNYATAAN

DENGAN INI SAYA MENYATAKAN BAHWA SKRIPSI INI BENAR-BENAR HASIL KARYA SENDIRI YANG BELUM PERNAH DIAJUKAN SEBAGAI SKRIPSI ATAU KARYA ILMIAH PADA PERGURUAN TINGGI ATAU LEMBAGA MANAPUN.

Tangerang Selatan, XX Februari 2024

Gilang Islamay Putra Djuharis

NIM. 11190940000055

**LEMBAR PENGESAHAN**

Skripsi ini berjudul “PREDIKSI HARGA BITCOIN MENGGUNAKAN TRANSFORMER: MENGINTEGRASIKAN HARGA, SENTIMEN, TREN, DAN VOLUME DALAM ANALISIS DERET WAKTU MULTIVARIAT” yang ditulis oleh **Gilang Islamay Putra Djuharis NIM. 11190940000055** telah diuji dan dinyatakan lulus dalam sidang Munaqosah Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta pada hari XXX, XX Februari 2024. Skripsi ini telah diterima untuk memenuhi salah satu persyaratan dan memperoleh gelar sarjana strata satu (S1) Program Studi Matematika.

**Menyetujui,**

|  |  |
| --- | --- |
| Pembimbing I  **Muhaza Liebenlito, M.Si**  NIDN. 2003098802 | Pembimbing II  **Prof. Dr. Nur Inayah, M.Si.**  NIP. 197401252003122001 |
| Penguji I  **XXX**  NIP. XXX | Penguji II  **XXX**  NIP. XXX |

**Mengetahui,**

|  |  |
| --- | --- |
| Dekan Fakultas Sains dan Teknologi  **Husni Teja Sukmana, S.T., M.Sc, Ph.D**  NIP. 197710302001121003 | Ketua Program Studi Matematika  **Taufik Edy Sutanto, M.Sc.Tech.,Ph.D.**  NIP. 197905302006041002 |

PERSEMBAHAN

“Skripsi ini saya persembahkan untuk kedua orang tua saya,

Bapak Djuharis Rasul dan Ibunda Yanawati”

**MOTTO**

“*Pursue what is meaningful, not what is expedient*”

KATA PENGANTAR

*Assalamu’alaikum Warahmatullahi Wabarakatuh.*

*Alhamdulillahi rabbil ‘alamin*, puji syukur ke hadirat Allah SWT atas segala Rahmat dan hidayah-Nya sehingga penulis berhasil menyelesaikan skripsi yang berjudul “**PREDIKSI HARGA BITCOIN MENGGUNAKAN TRANSFORMER: MENGINTEGRASIKAN HARGA, SENTIMEN, TREN, DAN VOLUME DALAM ANALISIS DERET WAKTU MULTIVARIAT”**.

Skripsi ini disusun guna memenuhi salah satu syarat kelulusan untuk memperoleh gelar Sarjana Matematika (S.Mat) pada Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah Jakarta. Skripsi ini dapat terselesaikan atas dukungan dari berbagai pihak. Karenanya, penulis bermaksud ingin menyampaikan terimakasih kepada:

1. Bapak Husni Teja Sukmana, S.T., M.Sc, Ph.D. selaku Dekan Fakultas Sains dan Teknologi.
2. Bapak Taufik Edy Sutanto, M.Sc.Tech.,Ph.D. selaku Ketua Program Studi Matematika. Dan ibu Dr. Gustina Elfiyanti, M.Si. selaku Sekretaris Prodi Matematika.
3. Bapak Muhaza Liebenlito, M.Si selaku Pembimbing Skripsi I dan Ibu Prof. Dr. Nur Inayah, M.Si. selaku Dosen Pembimbing II. Yang keduanya selalu meluangkan waktu dalam memberikankan saran, bimbingan, serta masukan yang sangat berharga sehingga penulis dapat menyelesaikan skripsi ini.
4. XXX selaku Penguji I dan XXX selaku Penguji II yang senantiasa memberikan kritik dan saran yang konstruktif dalam proses penyelesaian skripsi ini.
5. Seluruh dosen Program Studi Matematika telah memberikan ilmu yang berharga dan bermanfaat bagi saya.
6. Bapak Djuharis Rasul dan Ibu Yanawati, yang selalu memberikan doa, semangat, dan dukungan selama masa perkuliahan sehinga skripsi ini dapat diselesaikan.
7. Seluruh teman-teman Matematika Angkatan 2019 yang telah menemani segala macam proses perkuliahan dari awal hingga akhir.
8. Serta para sahabat penulis diantara lain Risky Amalia Marhariyadi yang membantu melabeli dataset manual, yang telah memberikan insightnya, Farrel yang telah meminjamkan GTX 1050 untuk mempercepat komputasi, serta Nindi, Sheila, dan teman teman sekalian yang telah memberikan dukungan secara mental.

Penulis menyadari bahwa skripsi ini masih memiliki kekurangan dan belum sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang konstruktif agar skripsi ini dapat lebih baik. Akhirnya, penulis berharap skripsi ini dapat bermanfaat bagi semua orang.

*Wassalamu’alaikum Warahmatullahi Wabarakatuh*

Tangerang Selatan, XX Februari 2024

Penulis

ABSTRAK

Perkembangan teknologi bersandingan langsung dengan berkembangannya mata uang kripto. Dengan bitcoin mengalami kenaikan sebesar 75% hanya dengan jangka satu tahun pada 2020, perkembangan mata uang kripto sudah tidak bisa dipungkiri lagi. Dengan votalitasnya harga mata uang kripto maka terdoronglah pencarian model paling mutakhir untuk meramal harga mata uang kripto. Penelitian kali ini bereksperimen untuk mencari model mutakhir tersebut tidak hanya dari sisi peramalan tetapi juga dalam sisi Pemrosesan Bahasa Alami, dengan model transformers yang menggunakan mekanisme *attention* untuk keduanya. Diintegrasikan analisis sentimen, trend, dan volume ke dalam variabel model dan didapatkan model yang paling mutakhir adalah dengan hanya memasukan variabel sentimen serta trend sehingga mendapatkan model dengan RMSE sebesar 0.0256, sMAPE sebesar 3.0428%, serta MAPE sebesar 3.0565%. Yang mana melampaui model *Long-Short Term Memory* dengan RMSE sebesar 0.0262, sMAPE sebesar 3.0615%, serta MAPE sebesar 3.0887%

**Kata Kunci:** Mata Uang Kripto, Bitcoin, Analisis Sentimen, Model Transformers, Prediksi Harga

*ABSTRACT*

*The development of technology goes hand in hand with the growth of cryptocurrency. With Bitcoin experiencing a 75% increase in just one year in 2020, the development of cryptocurrency cannot be denied. Due to the volatility of cryptocurrency prices, there is a drive to find the most advanced models for predicting cryptocurrency prices. This research experiment explores the search for such advanced models, not only from the forecasting perspective but also in Natural Language Processing, using transformer models with attention mechanisms for both aspects. Sentiment analysis, trends, and volume are integrated into the model variables, and the most advanced model is obtained by incorporating only sentiment and trend variables, resulting in an RMSE of 0.0256, sMAPE of 3.0428%, and MAPE of 3.0565%. Which surpass Long-Short Term Memory’s model with an RMSE of 0.0262, sMAPE of 3.0615%, and MAPE of 3.0887%.*

***Keywords:*** *Cryptocurrency, Bitcoin, Sentiment Analysis, Transformer Models, Price Prediction*

DAFTAR ISI

[PERNYATAAN ii](#_Toc155816221)

[PERSEMBAHAN v](#_Toc155816222)

[KATA PENGANTAR vi](#_Toc155816223)

[ABSTRAK viii](#_Toc155816224)

[*ABSTRACT* ix](#_Toc155816225)

[BAB I PENDAHULUAN 1](#_Toc155816226)

[1.1 Latar Belakang 1](#_Toc155816227)

[1.2 Perumusan Masalah 4](#_Toc155816228)

[1.3 Batasan Masalah 5](#_Toc155816229)

[1.4 Tujuan Penelitian 5](#_Toc155816230)

[1.5 Manfaat Penelitian 5](#_Toc155816231)

[BAB II LANDASAN TEORI 6](#_Toc155816232)

[2.1 Recurrent Neural Network (RNN) 6](#_Toc155816233)

[2.2 Long-Short Term Memory 7](#_Toc155816234)

[2.3 Transformer 7](#_Toc155816235)

[2.4 Bidirectional Encoder Representations Transformers (BERT) 11](#_Toc155816236)

[2.5 Evaluasi Model 12](#_Toc155816237)

[BAB III METODOLOGI PENELITIAN 14](#_Toc155816238)

[3.1 Robustly Optimized Bert Pretraining Approach (roBERTa) 14](#_Toc155816239)

[3.2 Temporal Fusion Transformer 14](#_Toc155816240)

[3.3 Pra Pemrosesan Data Teks 16](#_Toc155816241)

[3.4 Augmentasi Data 19](#_Toc155816242)

[3.5 Nilai Pencilan 21](#_Toc155816243)

[3.6 Multiple Seasonal-Trend decomposition using LOESS 21](#_Toc155816244)

[3.7 Time Lag Plot 22](#_Toc155816245)

[3.8 Transformasi Data Rangkaian Waktu 22](#_Toc155816246)

[BAB IV HASIL DAN PEMBAHASAN 24](#_Toc155816247)

[KESIMPULAN DAN SARAN 35](#_Toc155816248)

DAFTAR GAMBAR

[Gambar 1. Arsitektur *Recurrent Neural Network* 6](#_Toc155816381)

[Gambar 2. Arsitektur *Long-Short Term Memory* 7](#_Toc155816382)

[Gambar 3. Arsitektur Transformer 8](#_Toc155816383)

[Gambar 4. Arsitektur *Attention Head* 10](#_Toc155816384)

[Gambar 5. Arsitektur Kalkulasi *Attention* 10](#_Toc155816385)

[Gambar 6. Arsitektur BERT 11](#_Toc155816386)

[Gambar 7. Arsitektur Temporal Fusion Transformer 15](#_Toc155816387)

[Gambar 8. Mekanisme VSN 16](#_Toc155816388)

[Gambar 9. Jumlah Data Latih 21](#_Toc155816389)

[Gambar 10. Matriks Konfusi dari Model Kandidat 24](#_Toc155816390)

[Gambar 11. Matriks Konfusi Model Terbaik 26](#_Toc155816391)

[Gambar 12. Harga dibandingkan Kovariat 27](#_Toc155816392)

[Gambar 13. *Lag Plot* Harga dibandingkan Kovariat 28](#_Toc155816393)

[Gambar 14. MSTL Harga 29](#_Toc155816394)

[Gambar 15. Harga per Bulan 30](#_Toc155816395)

[Gambar 16. Grafik Nilai Pencilan 31](#_Toc155816396)

[Gambar 17. Data Variabel 32](#_Toc155816397)

[Gambar 18. Grafik Keterjelasan Encoder dan Decoder Model 33](#_Toc155816398)

[Gambar 19. Grafik *Attention* dari model 34](#_Toc155816399)

DAFTAR TABEL

[Tabel 1. Contoh Tweet Spam 17](#_Toc155816507)

[Tabel 2. Akurasi Google Translate 19](#_Toc155816508)

[Tabel 3. Akurasi Model Kandidat 24](#_Toc155816509)

[Tabel 4. Akurasi Model setelah Pelatihan Model 25](#_Toc155816510)

[Tabel 5. Evaluasi Model 26](#_Toc155816511)

[Tabel 6. Tabel Evaluasi Performa tanpa Kovariat 32](#_Toc155816512)

[Tabel 7. Tabel Evaluasi Performa menggunakan Kovariat 32](#_Toc155816513)

[Tabel 8. Keterjelasan Encoder dan Decoder Model 33](#_Toc155816514)

# BAB I PENDAHULUAN

1. Latar Belakang

Perkembangan yang pesat pada bidang ilmu teknologi dan informasi, *Financial Tech* (Fintech) atau teknologi keuangan terus melakukan terobosan baru, mendorong peningkatan dan inovasi model keuangan serta membentuk kembali rantai pasok dan rantai nilai untuk industri keuangan[1]. Sebagai sistem pertukaran mata uang elektronik alternatif yang baru, *cryptocurrency* atau mata uang kripto telah diakui secara luas memiliki konsekuensi signifikan bagi pasar-pasar berkembang dan ekonomi global[2].

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. Salah satu mata uang kripto yang paling populer adalah Bitcoin[2]. 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.Tak hanya itu, banyak sekali hal menarik dari mata uang kripto seperti kemudahan penggunaan, keamanan, dan desentralisasi. Dengan dapat diakses melalui berbagai perangkat ditambah dengan identitas yang aman dan transparan, tetapi pada saat yang bersamaan bersifat anonim dengan setiap transaksi tercatat pada blockchain yang mendasarinya tanpa melibatkan peran perantara seperti bank[3].

Volatilitas harga yang tinggi pada mata uang kripto menjadi salah satu alasan mengapa beberapa investor enggan masuk ke pasar ini[4], 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%. Bahkan pasar saham memiliki mekanisme *Short Sale Price Test Circuit* Breaker, yang dipicu ketika harga mengalami penurunan lebih dari 10% dalam satu hari perdagangan untuk mencegah harga turun lebih banyak lagi[5]. 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 untuk meramalkan data deret waktu menggunakan berbagai teknik dan data. Misalnya, penggunaan data Google Trends dengan model *statistical* dan *Machine Learning* seperti Vector Autoregression dan Random Forest untuk meramalkan pergerakan harga Bitcoin berdasarkan harga dan popularitasnya[5]. Studi tersebut menunjukkan bahwa data pencarian juga dapat menjadi variabel prediktor untuk aset investasi. Selain itu, penggunaan google trend untuk memprediksi penjualan retail, otomotif, rumah, dan tempat wisata juga mencapai kesimpulan serupa[12]. Demikian pula, penelitian tentang hubungan antara harga bitcoin dan trennya juga memberikan hasil positif[2]. Hubungan antara harga dan volume pasar Bitcoin juga telah terbukti memiliki korelasi[10]. Berdasarkan penelitian ini, dapat disimpulkan bahwa popularitas dan volume kripto berkorelasi dengan harga mereka.

Dalam ranah analisis deret waktu multivariat, ditemukan bahwa menggunakan model multivariat untuk meramalkan harga saham memberikan hasil yang lebih baik dibanding pendekatan univariat[13]. Analisis multivariat juga telah diterapkan untuk meramalkan kripto dengan membandingkan tiga pendekatan menggunakan *Recurrent Neural Network* (RNN): *Long Short-Term Memory* (LSTM), *Bidirectional LSTM* (Bi-LSTM), dan *Gated Recurrent Unit* (GRU). Dengan menggunakan lima variabel: harga penutupan, harga pembukaan, harga tertinggi, harga terendah, dan volume untuk lima kripto, termasuk Bitcoin, Ethereum, Cardano, Tether, dan Binance Coin. Hasilnya menunjukkan bahwa Bi-LSTM dan GRU memiliki performa serupa dengan *Mean Absolute Percentage Error* (MAPE) rata-rata sebesar 0,0465712 untuk Bi-LSTM dan 0,0446512 untuk GRU, sementara LSTM memiliki MAPE sebesar 0,0529916. Meskipun LSTM lebih unggul dalam dataset USDT dan BNB, namun memiliki varian yang lebih tinggi dibandingkan dengan Bi-LSTM dan GRU[14]. Hubungan antara harga Bitcoin dan sentimen juga digunakan untuk meramalkan harga Bitcoin, dengan *Mean Absolute Error* (MAE) sebesar 0,245, *Mean Square Error* (MSE) sebesar 0,2528, dan *Root Mean Squared Error* (RMSE) sebesar 0,5028[6].

Penelitian-penelitian tersebut melihatkan bahwa penggunaan teknik-teknik *machine learning* ataupun *deep 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* 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.

"*Attention is All You Need*" adalah judul sebuah makalah ilmiah yang disajikan oleh A. Vaswani, dkk dari Google pada tahun 2017. Mereka mengusulkan model baru yang merupakan perbaikan atas model berbasis rekuren untuk *Natural Language Processing* (NLP) menggunakan mekanisme *self-attention*, yang disebut Transformer[7]. Dengan arsitekturnya, transformer dapat memahami bahasa melebihi model-model sebelumnya dalam berbagai tolak ukur (misalnya, Terjemahan Mesin, Parsial Konstituen Bahasa Inggris). Salah satu model yang menggunakan transformer adalah *Bidirectional Encoder Representations from Transformers* (BERT) yang telah digunakan untuk terjemahan teks, klasifikasi teks, dan penggunaan lainnya.

Salah satu subbidang klasifikasi teks adalah analisis sentimen. Sebelum ditemukannya transformer, banyak model sentimen menggunakan mekanisme rekuren untuk berbagai bidang, termasuk aspek sosial, kesehatan, dan politik. Namun, masih sedikit model analisis sentimen yang ditargetkan khusus pada ranah kripto apalagi dengan menggunakan infrastruktur transformer. Beberapa contohnya termasuk CryptoBERT oleh ElKulako, yang dilatih menggunakan 3,2 juta kiriman media sosial dari platform seperti StockTwits, Telegram, Reddit, dan Twitter tentang kripto[9]. Dan model lainnya adalah CryptoBERT oleh kk08.

Model transformer juga digunakan berbagai bidang selain klasifikasi teks, tetapi juga untuk digunakan untuk meramalkan data deret waktu. Contohnya penelitian tentang kemampuan model transformer untuk meramalkan harga Bitcoin dan Ethereum menggunakan analisis sentimen, Membandingkannya dengan model LSTM, hasilnya menunjukkan bahwa model LSTM lebih unggul daripada model transformer. Model transformer yang dilatih menggunakan data Bitcoin memiliki MSE sebesar 0,00037, MAPE sebesar 0,05816, dan MAE sebesar 0,01432, sementara LSTM memiliki MSE sebesar 0,00032, MAPE sebesar 0,04613, dan MAE sebesar 0,01346. Studi ini juga menemukan bahwa model yang dilatih dengan data Bitcoin menunjukkan peningkatan dalam meramalkan harga Ethereum dibandingkan dengan model yang dilatih dengan data Ethereum saja[15]. Pada tahun 2020, makalah ilmiah lain berjudul "*A Transformer-Based Framework for Multivariate Time Series Representation Learning*" mengusulkan penggunaan arsitektur yang sama untuk meramalkan deret waktu, dan menemukan bahwa model transformer lebih unggul dibandingkan dengan model lain (Rocket, Long-Short Term Memory, XGBoost, dll.)[8]

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 limitasi perangkat keras maka digunakan google colab untuk menjalankan programnya dan tokenisasi NLP dilimitasi sampai 256 token.

1. Perumusan Masalah

Adapun masalah yang akan dipecahkan dalam penelitian kali ini adalah sebagai berikut:

1. Bagaimana performa model transformer dalam mengklasifikasikan sentimen sebuah teks dengan tema bitcoin?
2. Bagaimana performa model transformer dalam peramalan deret waktu?
3. Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah:

1. Data teks didapatkan dari hasil scraping sub-reddit r/bitcoin dengan judul “*Daily Discussion*” dan data X (dahulunya twitter) berasal dari kaggle yang merupakan hasil scraping. Sehingga data mulai dari 03/12/2017 hingga 30/06/2023.
2. Pengambilan token pada bagian pemrosesan bahasa alami dilimitasi maksimal 256.
3. Tujuan Penelitian

Tujuan dalam penelitian ini adalah:

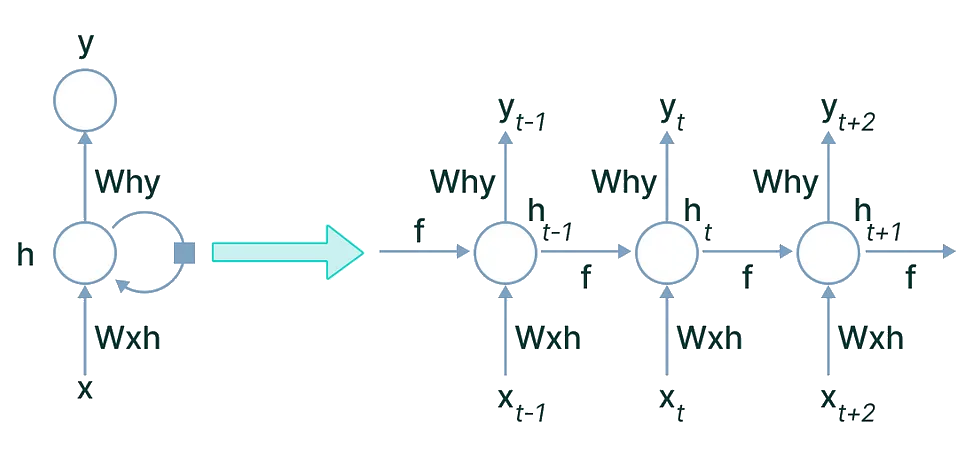
1. Membuat model klasifikasi sentimen terkait dengan topik bitcoin
2. Membangun model peramalan deret waktu terhadap harga bitcoin dengan mengintegrasikan variabel-variabel yang relevan.
3. Manfaat Penelitian

Manfaat yang akan didapat dari penelitian ini adalah mendapatkan model klasifikasi sentimen dan juga peramalan deret waktu harga bitcoin yang mutakhir.

BAB II  
LANDASAN TEORI

## Recurrent Neural Network (RNN)

Recurrent Neural Network adalah arsitektur yang pada umumnya berstruktur berkala, dimana memiliki tujuan untuk mendeteksi pola dalam suatu urutan data[17]. Keunggulan dari model RNN sendiri dibandingkan model lainnya seperti feedfoward networks adalah dengan adanya siklus yang memberikan informasi ke diri sendiri.



Gambar 1. Arsitektur *Recurrent Neural Network*

Arsitektur dari model RNN juga dapat dijelaskan, yaitu sebagai berikut:

1. Input Layer

Fitur input untuk setiap t-waktu

1. Hidden State

Menyimpan informasi mengenai input sebelumnya dan diperbarui setiap langkah.

1. Recurrent Connection

Dimana setiap t-waktu, *hidden state* dari waktu t-1 digunakan untuk kombinasi dengan input saat itu untuk mengeluarkan output dan juga memperbarui *hidden state*

1. Output Layer

Menghasilkan hasil di langkah tersebut berdasarkan input saat itu dan *hidden state*.

## Long-Short Term Memory

Salah satu model dasar (*baseline*) yang digunakan adalah *Long-Short Term Memory*. Tipe *Recurrent Neural Network* (RNN) ini berkapabilitas untuk mempelajari korelasi jangka panjang antara variable, karena itu model ini cocok untuk analisis deret waktu.



Gambar 2. Arsitektur *Long-Short Term Memory*

Mengikuti arsitektur RNN pada umumnya yaitu dengan struktur rantai, setiap *cell* dalam LSTM memiliki 4 gerbang. Yaitu yang pertama adalah *Forget Gate*, dimana dilakukan fungsi sigmoid untuk mendeterminasi seberapa banyak yang harus dilupakan pada ingatan jangka panjang (*Long-Term Memory*). Kedua Input Gate dengan fungsi sigmoid yaitu berapa banyak yang harus diingat serta yang ketiga *Candidate Memory* adalah apa saja yang harus diingat. Yang terakhir ada output gate, dimana pada gerbang ini dilakukan fungsi sigmoid untuk mendeterminasi output untuk cell selanjutnya.[18]

## 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 yang mana menyebabkan absennya paraparalelisasi pada model. Dimana model RNN juga rentan akan *The Vanishing Gradient Problem*[19]

*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.[8]



Gambar 3. Arsitektur Transformer

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.

|  |  |
| --- | --- |
|  | (1) |

|  |  |
| --- | --- |
|  | (2) |

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:



Gambar 4. Arsitektur *Attention Head*

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



Gambar 5. Arsitektur Kalkulasi *Attention*

|  |  |
| --- | --- |
|  | (3) |

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

## Bidirectional Encoder Representations Transformers (BERT)

BERT merupakan salah satu model yang lahir dari arsitektur transformer, yang mana menggunakan beberapa lapisan encoder dari arsitektur transformer[20]



Gambar 6. Arsitektur BERT

Inti sari dari pelatihan BERT sendiri adalah dengan menggunakan *Next Sentence Prediction* (NSP) dan *Masked Language Model* (MLM) dalam tahap pelatihan model.

1. *Next Sentence Prediction*

Untuk melatih model dalam memahami hubungan antara dua kalimat, model diberikan dua kalimat dan dilatih untuk memprediksi apakah kalimat kedua merupakan kelanjutan dari kalimat pertama. Dengan proporsi 50% kalimat kedua merupakan benar kelanjutan dari kalimat pertama, dan 50% tidak.

1. *Masked Language Model*

Untuk melatih model untuk bersifat *bidirectional* atau dua arah, 15% dari token dilakukan *masking* atau disembuyikan dari model untuk melatihnya mengisi kekosongan token. Dan untuk mencega bias, dari 15% token yang disembunyikan dimasukan token acak sebanyak 10%, tidak diubah sebanyak 10%.

## Evaluasi Model

Evaluasi dari kedua model juga penting untuk dilakukan, guna membandingkan hasil dari model dengan model yang sudah ada. Evaluasi dari model analisis sentiment menggunakan Akurasi, *F-1 Score*, *Precision*, dan *Recall*. Dengan akurasi dilakukan menggunakan rumus

|  |  |
| --- | --- |
|  | (4) |

Selanjutnya dengan *Precision* serta *Recall* menggunakan rumus sebagai berikut

|  |  |
| --- | --- |
|  | (5) |

|  |  |
| --- | --- |
|  | (6) |

dimana TP adalah *True Positive*, FP adalah *False Positive*, dan FN adalah *False Negative*. Dan yang terakhir adalah *F-1 Score*

|  |  |
| --- | --- |
|  | (7) |

Evaluasi peramalan peramalan deret waktu digunakan *Root Mean Square error* (RMSE), *Mean Absolute Percentage Error* (MAPE), serta *Symmetric Mean Absolute Percentage Error* (sMAPE) untuk evaluasinya.

|  |  |
| --- | --- |
|  | (8) |

|  |  |
| --- | --- |
|  | (9) |

|  |  |
| --- | --- |
|  | (10) |

dimana adalah nilai sebenarnya pada waktu t, adalah nilai prediksi pada waktu t, dan n adalah total data yang diprediksi.

# BAB III METODOLOGI PENELITIAN

Penelitian ini mengadopsi pendekatan ekperimental dengan menginvestiasi kemampuan model Transformer dalam menganalisis sentimen dalam topik Bitcoin dan memprediksi harga bitcoin dengan mengintegrasikan faktor-faktor kunci seperti harga, sentimen, tren, serta volume. Pendekatan eksperimental dipilih karena memungkinkan pengujian secara langsung terhadap data waktu nyata untuk menganalisa kinerja model terhadap variabel-variabel yang digunakan.

### Robustly Optimized Bert Pretraining Approach (roBERTa)

roBERTa merupakan versi BERT yang lebih dikembangkan lagi, ditemukan bahwa BERT memiliki kekurangan yaitu model tidak terlatih secara optimal[21]. Walaupun menggunakan arsitektur yang sama, roBERTa berhasil mendapatkan performa yang lebih baik dibandingkan BERT dengan 4 perubahan.

1. Penghapusan Objektif NSP
2. Pelatihan model lebih lama, dengan data yang lebih besar
3. Pelatihan dengan data yang lebih panjang
4. Pengubahan teknik penyembunyian token

Ditemukan bahwa penghapusan NSP mengurangi performa dari model itu[20]. Akan tetapi, beberapa penelitian berargumen bahwa keberadaan NSP justru mengurangi performa model itu sendiri[22]. Maka dilakukan penghapusan objektif NSP untuk model roBERTa. Dengan perubahan ini roBERTa mendapatkan performa yang lebih baik dibandingan BERT orisinal.

### Temporal Fusion Transformer

Tahun 2021 Lim Brian, dkk mengeluarkan arsitektur transformer terbaru yang menggabungkan *Long-Short Term Memory* (LSTM) dengan mekanisme *attention* untuk peramalan deret waktu multivariat dengan kemampuan untuk menjelaskan dengan nama *Temporal Fusion Transformers* (TFT)[23].



Gambar 7. Arsitektur Temporal Fusion Transformer

Model TFT dibandingkan dengan model lainnya, mendapatkan performa yang bagus, dikarenakan penggunaan arsitektur LSTM pada Encoder serta Decodernya sehingga dapat menangkap korelasi jangka pendek ataupun panjang dari suatu kovariat, ditambah dengan *Multi-Head Attention* sehingga model dapat menimbang kovariat yang mana yang harus lebih difokuskan pada waktu tertentu untuk melakukan prediksi. TFT juga menerapkan mekanisme gerbang dan *Variable Selection Network* (VSN) untuk meminimalisir kontribusi variabel yang tidak berperan signifikan saat pelatihan model.



Gambar 8. Mekanisme VSN

TFT juga mengedepankan *explainability* atau kejelasan dari model itu sendiri, dibandingkan dengan model lain yang bersifat *black-box* dimana kejelasan dari perhitungan model tidak dapat dilihat yang menimbulkan masalah praktis dan etis[24]. TFT menyelesaikan masalah itu dengan mekanisme *attention*, TFT bisa memberikan penjelasan tentang dinamika temporal serta menganalisa relasi global sementara dari seluruh data (Seperti musim dan *lag effect*s)[23].

### Pra Pemrosesan Data Teks

Pengumpulan data dilakukan dengan mengambil data twitter melalui Kaggle dan *scraping* sosial media reddit dilakukan dengan mengambil komen yang terdapat pada daily discussion di r/bitcoin, 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. Serta data google trends diambil secara manual dalam jangka tiap 270 hari dari 03/12/2017 hingga 30/06/2023 lalu data dinormalisasi dengan mencocokan tanggal yang berkesinambungan. Ini dilakukan guna mengambil data harian google trends.

Hasil yang optimal dalam analisis teks dibutuhkan beberapa langkah. Yaitu tokenisasi teks dan normalisasi teks.[25] 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[26]. Pengaplikasian pra pemrosesan data dilakukan 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.

Pra pemrosesan juga dilakukan dengan menghapus tweet yang bersifat spam. 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:

Tabel 1. Contoh Tweet Spam

|  |  |
| --- | --- |
| ***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 |

Pelabelan data manual juga dilakukan untuk melatih model analisis sentimen, 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 demikian, jumlah total data sebanyak 3.803 dapat diperbanyak lebih lanjut melalui augmentasi data.

Data analisis sentimen telah dibersihkan dan diaugmentasi lalu dilanjutkan dengan membangun model analisis sentimennya. Dikarenakan membangun model dari awal membutuhkan kekuatan komputasi serta data yang sangat besar, maka dilakukan *fine tuning* dari model yang sudah ada. Dengan memilih model yang paling besar akurasinya terhadap data yang telah dilabeli manual, pelabelan data teks dilakukan serta diambil rata rata tiap harinya dengan rumus

|  |  |
| --- | --- |
|  | (11) |

dimana adalah skor sentimen dari teks dalam interval satu hari, dan adalah banyak data teks dalam interval tersebut.[27]

### Augmentasi Data

Augmentasi data banyak dilakukan pada computer vision[28], akan tetapi pada bidang NLP masih sedikit yang membahas[29]. Back-translation adalah teknik augmentasi data yang sangat efektif untuk terjemahan mesin neural[30], yang mana pendekatan ini menjanjikan karena memiliki kemampuan mempertahankan label yang baik secara alami dan kemampuan memparafrase yang sangat berharga[31]. Back-Translation dilakukan dengan teks asli dikembalikan ke dalam bahasa aslinya setelah dua kali terjemahan. Teks asli S1 diterjemahkan ke dalam bahasa lain (seperti Bahasa Indonesia) sebagai S2, dan kemudian diartikan kembali ke dalam bahasa aslinya sebagai S3.[32]

*Back-Translation* pada data teks untuk mengklasifikasikan suatu kalimat juga sudah pernah dilakukan. Augmentasi data menghasilkan peningkatan sekitar 5% pada bobot rata-rata F1 makro.[33] Berdasarkan semua data eksperimental, sebagian besar metrik model setelah augmentasi data menggunakan *Back-Translation* mengalami peningkatan.

Layanan API Google Translate digunakan pada penelitian kali ini untuk melakukan *Back-Translation*, dengan berdasarkan akurasi dari *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[34]

Tabel 2. Akurasi Google Translate

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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. Setelah augmentasi, *undersampling* atau mengambil jumlah data sebanyak data yang paling minoritas dilakukan agar pelatihan model tidak condong ke satu sentimen. Maka didapatkan data latih sebagai berikut



Gambar 9. Jumlah Data Latih

Teknik augmentasi data juga diusulkan oleh Wei dan Zou dengan judul *Easy Data Augmentation* (EDA). Diterapkan operasi penggantian sinonim, penyisipan, pertukaran, dan penghapusan secara acak.[35] Analisis akurasi model paling akurat setelah data diagumentasi menggunakan EDA juga dilakukan pada penelitian kali ini.

### Nilai Pencilan

Nilai pencilan dalam data rangkaian waktu khususnya harga bitcoin yang memang memiliki sifat tidak stabil harus ditangani dengan hati hati, oleh karena itu metode Isolation Forest (iForest) digunakan dalam pendeteksian nilai pencilan pada data ini. Penghapusan sekitar 10% dari nilai pencilan meningkatkan performa model untuk kebanyakan machine learning[36], maka pengaruh nilai pencilan terhadap kinerja modelpun dipelajari dengan tidak menghapus dan menghapus sekitar 10% dari nilai pencilan. Teknik yang digunakan untuk menggantikan nilai pencilan adalah *Moving Average* atau rata rata berjalan.[37]

|  |  |
| --- | --- |
|  | (12) |

dimana adalah pengganti titik observasi, adalah titik observasi, dan n adalah besar jendela observasi.[38]

### 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.[39] Dengan memberikan dekomposisi aditif dari data rangkaian waktu. Diberikan dimana t adalah observasi pada waktu ke-t, maka dekomposisi aditifnya dapat didefiniskan sebagai berikut:

|  |  |
| --- | --- |
|  | (13) |

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

|  |  |
| --- | --- |
|  | (14) |

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.

### Transformasi Data Rangkaian Waktu

Tes Augmented Dickey-Fuller[40] juga dilakukan untuk mengatasi fluktuasi dan volatilitas tinggi harga serta volume bitcoin, data yang tidak stasioner diubah menjadi stasioner dengan menerapkan *detrending* atau penghapusan tren.[41] Teknik *detrending* yang dilakukan pada penelitian kali ini adalah *differencing transformation* atau dengan menghasilkan deret waktu baru di mana nilai baru pada waktu dikalkulasikan dengan perbedaan diantara observasi orisinil dan observasi pada langkah waktu sebelumnya. Rumusnya dapat dijelaskan sebagai berikut.

|  |  |
| --- | --- |
|  | (15) |

Normalisasi data juga dilakukan guna meningkatkan performa model[41], digunakan normalisasi Min-Max sehingga nilai nilai di dalam data dipetakan dalam rentang (0, 1) dengan rumus sebagai berikut

|  |  |
| --- | --- |
|  | (16) |

dimana adalah nilai minimal dari y, dan adalah nilai maksimal dari y. Dan untuk menghindari kebocoran pada data test, maka *fitting* skala data hanya diambil dari data latih.

# BAB IV HASIL DAN PEMBAHASAN

#### Analisis Sentimen

Menjelajahi website Huggingface, ditemukan 6 kandidat model untuk sentimen analisis dengan 2 diantara untuk topik khusus bitcoin. Eksperimen-pun dilakukan untuk mengambil model yang terbaik untuk dilakukan pelatihan dengan menguji ketiga model dengan data yang sudah dilabeli manual. Berikut merupakan confusion matrix serta akurasi model kandidat



Gambar 10. Matriks Konfusi dari Model Kandidat

Tabel 3. Akurasi Model Kandidat

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 (diwarnai biru). Maka dari itu 4 model ini-pun dilatih lebih lanjut dengan data latih. Hasil dari ke-4 model tersebut setelah dilatih dengan *hyperparameter* paling optimal adalah sebagai berikut

###### 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.

###### Hasil Model Analisis Sentimen

Hasil dari ke-4 model tersebut setelah dilatih dengan *hyperparameter* paling optimal adalah sebagai berikut:

Tabel 4. Akurasi Model setelah Pelatihan Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **murni** | **GT** | **GTH** | **T** |
| bertweet-base-sentiment-analysis | 0.626667 | 0. 666666 | 0. 666666 | 0.656667 |
| Cryptobert | 0.666666 | 0.673333 | 0.676667 | 0.646667 |
| twitter-roberta-base-sentiment-latest | 0.653333 | 0.65 | 0.646667 | 0.64 |
| twitter-xlm-roberta-base-sentiment | 0.606667 | 0.596667 | 0.616667 | 0.62 |

Dataset murni atau yang tidak melewati augmentasi data mendapatkan akurasi paling besar pada model cryptobert, akan tetapi saat dilakukan augmentasi data peningkatan sebesar 1% pada data yang diaugmentasi dengan GTH. Dan ketika dataset murni dan GTH dilakukan *Easy Data Augmentation* menggunakan model cryptobert tidak didapatkan perkembangan dalam akurasinya. Maka dipilih model yang paling akurat GTH dengan model Cryptobert yang memiliki *hyperparameter* learning\_rate 2.16491656079216e-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. (Model dapat diakses melalui <https://huggingface.co/AfterRain007/cryptobertRefined>) Berikut merupakan confusion matrix dan hasil evaluasinya



Gambar 11. Matriks Konfusi Model Terbaik

Tabel 5. Evaluasi Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |

Setelah mendapatkan model analisis sentimen, dilakukan pelabelan dengan model kepada data teks dari twitter serta reddit dan diambil nilai rata rata sentimennya tiap hari dan dinormalisasikan.

#### Analisis Data Eksplorasi

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.



Gambar 12. Harga dibandingkan Kovariat

Sekilas, 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. Akan tetapi diperlukan analisis mendalam untuk mendeterminasi terkait korelasi antara variabel dengan TLP dan musimannya menggunakan MSTL.

1. Time Lag Plot (TLP)



Gambar 13. *Lag Plot* Harga dibandingkan Kovariat

Harga terhadap tren dan sentimen dapat dilihat dari grafik di atas bahwa keduanya memiliki sifat autokorelasi positif yang tinggi, serta harga terhadap volume memiliki autokorelasi yang bersifat moderat. Dan dari ketiga grafik di atas dapat diketahui bahwa variabel variabel memiliki tendensi nilai pencilan (dijelajahi lebih lanjut pada bagian *iForest*)

1. *Multiple Seasonal-Trend decomposition using LOESS* (MSTL)

Lalu untuk tren musiman dilakukan MSTL, dikarenakan data yang digunakan bersifat harian maka ditetapkan 2 jangka waktu yaitu per-bulan (Seasonal\_30) dan per-tahun (Seasonal\_365).



Gambar 14. MSTL Harga

Dapat dilihat dari grafik di atas, 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.



Gambar 15. Harga per 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. Perlu diingat bahwa tanggal 29 Februari rata rata harga turun drastis dikarenakan tanggal 29 hanya sekali dilewati (yaitu pada tahun kabisat).

1. Data Pencilan

Sebelum dilakukan peramalan deret waktu, yang terakhir dilakukan adalah penanganan nilai pencilan.



Gambar 16. Grafik Nilai Pencilan

Berdasarkan iForest, terdapat 204 outlier berdasarkan 4 variabel yang ada, penanganan outlier dilakukan dengan mengambil 10% outlier yang memiliki score paling tinggi dan dilakukan penanganan dengan menggantikan nilai outlier menggunakan rata rata berjalan.

1. Tes Stasioner

Tes Augmented Dickey-Fuller dilakukan dan menunjukan bahwa data bitcoin yaitu harga dan volume merupakan data yang tidak stasioner, sehingga dilakukan *differencing transformation* untuk mengatasinya.

#### Peramalan Deret Waktu

Data deret waktu harga bitcoin dibagi menjadi dua, yaitu 80% data latih dan 20% data uji untuk memvalidasi peramalan harga bitcoin. Setelah data sudah selesai melewati semua tahap pemrosesan awal dan dilakukan normalisasi Min-Max, maka didapatkan data sebagai berikut



Gambar 17. Data Variabel

Dengan menggunakan beberapa model yang bersifat *Deep Learning*, didapatkan hasil metrik evaluasi sebagai berikut:

Tabel 6. Tabel Evaluasi Performa tanpa Kovariat

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **sMAPE** | **MAE** | **MAPE** | **Waktu Pelatihan (s)** | **Waktu Prediksi (s)** |
| **Long-Short Term Memory** |  |  |  |  |  |
| **Gated Recurrent Unit** |  |  |  |  |  |
| **Temporal Convolutional Network** |  |  |  |  |  |
| **Temporal Fusion Transformer** |  |  |  |  |  |

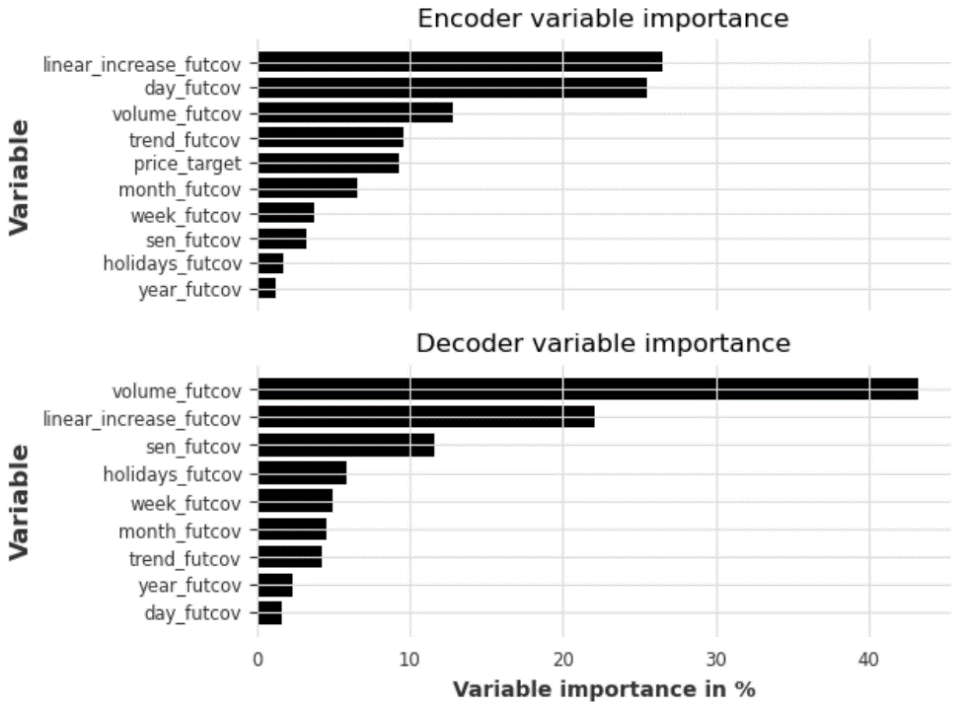
Tabel 7. Tabel Evaluasi Performa menggunakan Kovariat

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **sMAPE** | **MAE** | **MAPE** | **Waktu Pelatihan (s)** | **Waktu Prediksi (s)** |
| **Long-Short Term Memory** |  |  |  |  |  |
| **Gated Recurrent Unit** |  |  |  |  |  |
| **Temporal Convolutional Network** |  |  |  |  |  |
| **Temporal Fusion Transformer** |  |  |  |  |  |

Berdasarkan tabel 7, dapat disimpulkan bahwa model Temporal Fusion Transformers sendiri sudah mendapatkan hasil yang sangat memuaskan yaitu memiliki RMSE sebesar 0.0265. Akan tetapi, dengan memasukan kovariat didapatkan model yang lebih mutakhir yaitu dengan peningkatan dari semua metrik evaluasi. Dan saat *cross-validation* dilakukan untuk model TFT juga menghasilkan sMAPE sebesar X, MAE sebesar X, dan MAPE sebesar X.

#### Explainability

Dari model Temporal Fusion Transformer, dapat dihasilkan juga penjelasan dari model itu sendiri.

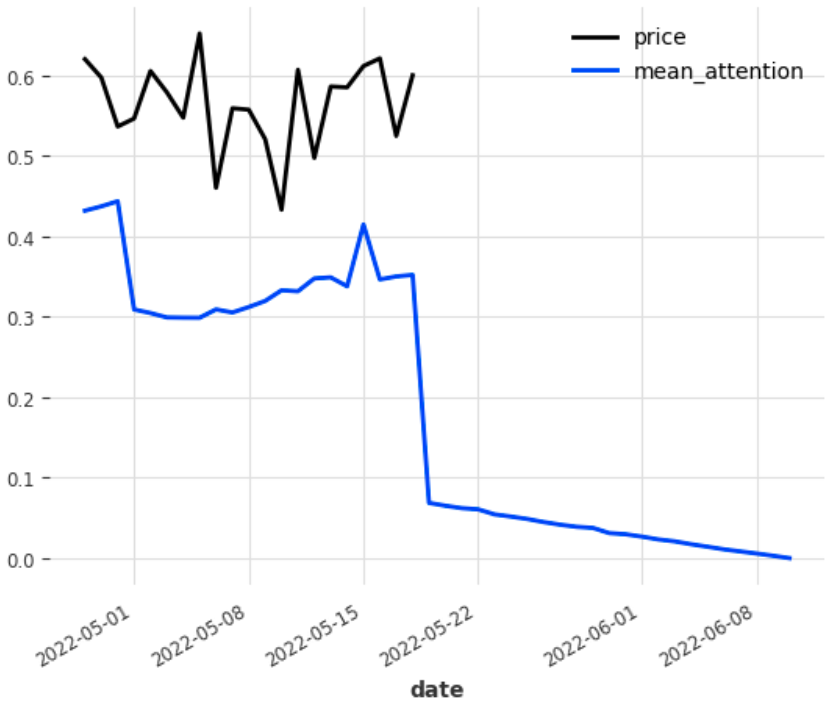


Gambar 18. Grafik Keterjelasan Encoder dan Decoder Model

Tabel 8. Keterjelasan Encoder dan Decoder Model

|  |  |  |
| --- | --- | --- |
| **type** | **encoder** | **decoder** |
| **year\_futcov** | 1,20 | 2,30 |
| **holidays\_futcov** | 1,70 | 5,80 |
| **sen\_futcov** | 3,20 | 11,60 |
| **week\_futcov** | 3,70 | 4,90 |
| **month\_futcov** | 6,50 | 4,50 |
| **price\_target** | 9,30 | - |
| **trend\_futcov** | 9,60 | 4,20 |
| **volume\_futcov** | 12,80 | 43,20 |
| **day\_futcov** | 25,50 | 1,60 |
| **linear\_increase\_futcov** | 26,50 | 22,10 |

Variabel variabel yang penting di dalam Encoder terlihat terdistribusi secara rata dari varibel harga keatas sembari menghiraukan variabel tahun, libur, sentimen, bulan, serta minggu. Sedangkan untuk Decoder sendiri sangat bergantung dengan variabel volume dengan nilai kepentingan lebih dari 40%, dengan variabel pendukung yaitu linear\_increase dan sentimen.



Gambar 19. Grafik *Attention* dari model

Atensi dari model terhadap kovariat dapat dilihat berdasarkan gambar X. Atensi naik pada saat training tepatnya di titik -3, sehingga dapat disimpulkan bahwa model menggunakan perhatian yang besar pada 3 titik sebelum peramalan dilakukan.

# 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 1464 data untuk melatihmodel cryptobert mendapatkan peningkatan akurasi sebesar 0.16% dan ketika dilakukan augmentasi data dengan *back-translation* menggunakan 10 bahasa yang memiliki BLEU score terbesar oleh google translate, didapatkan peningkatan lagi sebesar 0.01%. Oleh karena itu disimpulkan data latih yang banyak bukan berarti akan menghasilkan model yang terbaik akan tetapi kualitas dari data latih itu juga harus diperhatikan, dan juga *Easy Data Augmentation* dilakukan akan tetapi tidak membuahkan peningkatan akurasi.

Pengintegrasian sentimen, volume, dan trend kedalam model prediksi harga bitcoin mendapatkan peningkatan hasil dibandingkan hanya menintegrasikan kovariat waktu, dari RMSE sebesar 0,0264 menjadi 0,0260. Dan juga model TFT lebih unggul dibandingkan dengan model LSTM, dengan LSTM mendapatkan hasil RMSE terbaik sebesar 0,0262 dan TFT terbaik sebesar 0,0256.

Saran untuk penelitian selanjutnya adalah dengan memperbesar token dari model yang digunakan. 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.

Data yang digunakan untuk meramal deret waktu juga dapat diperluas. Keterbatasan saat ini menyebabkan penelitian hanya memanfaatkan data dari 03/12/2017 hingga 29/06/2023, sedangkan Bitcoin telah tersedia sejak tahun 2009.

[1] L. Zhao, “The function and impact of cryptocurrency and data technology in the context of financial technology: introduction to the issue,” *Financ. Innov.*, vol. 7, no. 1, 2021, doi: 10.1186/s40854-021-00301-w.

[2] A. H. Al-Nefaie and T. H. H. Aldhyani, “Bitcoin Price Forecasting and Trading: Data Analytics Approaches,” *Electron.*, vol. 11, no. 24, 2022, doi: 10.3390/electronics11244088.

[3] U. Rahardja, Q. Aini, E. Purnamaharahap, and R. Raihan, “GOOD, BAD AND DARK BITCOIN: A Systematic Literature Review,” *Aptisi Trans. Technopreneursh.*, vol. 3, no. 2, pp. 1–5, 2021, doi: 10.34306/att.v3i2.175.

[4] D. Higdon, J. Nelson, and J. Ibarra, JuanAbraham, “Cryptocurrency price prediction using tweet volumes and sentiment analysis,” *SMU Data Sci. Rev.*, vol. 1, no. 3, p. 1, 2018.

[5] U.S. Security and Exchange Commission, “Division of Market Regulation.” [Online]. Available: https://www.sec.gov/divisions/marketreg/mrfaqregsho1204.htm

[6] M. A. Nasir, T. L. D. Huynh, S. P. Nguyen, and D. Duong, “Forecasting cryptocurrency returns and volume using search engines,” *Financ. Innov.*, vol. 5, no. 1, 2019, doi: 10.1186/s40854-018-0119-8.

[7] S. Alghamdi, S. Alqethami, T. Alsubait, and H. Alhakami, “Cryptocurrency Price Prediction using Forecasting and Sentiment Analysis,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 10, pp. 891–900, 2022, doi: 10.14569/IJACSA.2022.01310105.

[8] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[9] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, “A Transformer-based Framework for Multivariate Time Series Representation Learning,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 2114–2124, 2021, doi: 10.1145/3447548.3467401.

[10] M. Kulakowski, F. Frasincar, and E. Cambria, “Sentiment Classification of Cryptocurrency-Related Social Media Posts,” *IEEE Intell. Syst.*, vol. 38, no. 4, pp. 5–9, 2023, doi: 10.1109/MIS.2023.3283170.

[11] E. Gemici and M. Polat, “Relationship between price and volume in the Bitcoin market,” *J. Risk Financ.*, vol. 20, no. 5, pp. 435–444, 2019, doi: 10.1108/JRF-07-2018-0111.

[12] M. Ettredge, J. Gerdes, and G. Karuga, “Using web-based search data to predict macroeconomic statistics,” *Commun. ACM*, vol. 48, no. 11, pp. 87–92, 2005, doi: 10.1145/1096000.1096010.

[13] H. Choi and H. Varian, “Predicting the Present with Google Trends,” *Econ. Rec.*, vol. 88, no. SUPPL.1, pp. 2–9, 2012, doi: 10.1111/j.1475-4932.2012.00809.x.

[14] S. Mehtab and J. Sen, “Stock Price Prediction Using Convolutional Neural Networks on a Multivariate Timeseries,” 2020, doi: 10.36227/techrxiv.15088734.v1.

[15] S. Hansun, A. Wicaksana, and A. Q. M. Khaliq, “Multivariate cryptocurrency prediction: comparative analysis of three recurrent neural networks approaches,” *J. Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00601-7.

[16] Z. Hu, Y. Zhao, and M. Khushi, “A survey of forex and stock price prediction using deep learning,” *Appl. Syst. Innov.*, vol. 4, no. 1, pp. 1–30, 2021, doi: 10.3390/ASI4010009.

[17] R. M. Schmidt, “Recurrent Neural Networks (RNNs): A gentle Introduction and Overview,” no. 1, pp. 1–16, 2019, [Online]. Available: http://arxiv.org/abs/1912.05911

[18] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[19] S. Hochreiter, “The Vanishing Gradient Problem During Learning Recurrent Neural Nets And Problem Solutions,” *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 6, no. 2, pp. 107–116, 1998.

[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[21] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” no. 1, 2019.

[22] G. Lample and A. Conneau, “Cross-lingual Language Model Pretraining,” 2018.

[23] B. Lim, S. Arık, N. Loeff, and T. Pfister, “Temporal Fusion Transformers for interpretable multi-horizon time series forecasting,” *Int. J. Forecast.*, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[24] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Comput. Surv.*, vol. 51, no. 5, 2018, doi: 10.1145/3236009.

[25] D. Sarkar, *Text Analytics with Python*, vol. 32, no. 1. 2016. doi: 10.1140/epja/i2006-10279-1.

[26] R. Feldman and J. Sanger, *The Text Mining Handbook*. 2006. doi: 10.1017/cbo9780511546914.

[27] M. Frohmann, M. Karner, S. Khudoyan, R. Wagner, and M. Schedl, “Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting,” *Big Data Cogn. Comput.*, vol. 7, no. 3, 2023, doi: 10.3390/bdcc7030137.

[28] and G. E. H. A. Krizhevsky, I. Sutskever, “ImageNet Classification with Deep Convolutional Neural Networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[29] W. Wang, B. Li, D. Feng, A. Zhang, and S. Wan, “The OL-DAWE Model: Tweet Polarity Sentiment Analysis with Data Augmentation,” *IEEE Access*, vol. 8, pp. 40118–40128, 2020, doi: 10.1109/ACCESS.2020.2976196.

[30] S. Edunov, M. Ott, M. Auli, and D. Grangier, “Understanding back-translation at scale,” *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 489–500, 2018, doi: 10.18653/v1/d18-1045.

[31] M. Bayer, M. A. Kaufhold, and C. Reuter, “A Survey on Data Augmentation for Text Classification,” *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–44, 2022, doi: 10.1145/3544558.

[32] J. Ma and L. Li, “Data Augmentation for Chinese Text Classification Using Back-Translation,” *J. Phys. Conf. Ser.*, vol. 1651, no. 1, 2020, doi: 10.1088/1742-6596/1651/1/012039.

[33] S. T. Aroyehun and A. Gelbukh, “Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling,” *COLING 2018 - 1st Work. Trolling, Aggress. Cyberbullying, TRAC 2018 - Proc. Work.*, pp. 90–97, 2018.

[34] M. Aiken, “An Updated Evaluation of Google Translate Accuracy,” *Stud. Linguist. Lit.*, vol. 3, no. 3, p. p253, 2019, doi: 10.22158/sll.v3n3p253.

[35] J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 6382–6388, 2019, doi: 10.18653/v1/d19-1670.

[36] M. Mudassir, S. Bennbaia, D. Unal, and M. Hammoudeh, “Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach,” *Neural Comput. Appl.*, vol. 6, 2020, doi: 10.1007/s00521-020-05129-6.

[37] C. Jeenanunta, K. D. Abeyrathna, and M. H. M. R. S. Dilhani, “Time Series Outlier Detection for Short-Term Electricity Load Demand Forecasting | INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET),” vol. 2, no. 1, pp. 37–50, 2018, [Online]. Available: https://ph02.tci-thaijo.org/index.php/isjet/article/view/175908

[38] F. R. Johnston, J. E. Boyland, M. Meadows, and E. Shale, “Some Properties of a Simple Moving Average when Applied to Forecasting a Time Series,” *J. Oper. Res. Soc.*, vol. 50, no. 12, p. 1267, 1999, doi: 10.2307/3010636.

[39] K. Bandara, R. Hyndman, and C. Bergmeir, “MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns,” *Int. J. Oper. Res.*, vol. 1, no. 1, p. 1, 2022, doi: 10.1504/ijor.2022.10048281.

[40] D. D. A and Wayne A. Fuller, “Likelihood ratio statistics for autoregressive time series with a unit root,” *Econometrica*, vol. 49, no. 4, pp. 1057–1072, 1981.

[41] K. Murray, A. Rossi, D. Carraro, and A. Visentin, “On Forecasting Cryptocurrency Prices: A Comparison of Machine Learning, Deep Learning, and Ensembles,” *Forecasting*, vol. 5, no. 1, pp. 196–209, 2023, doi: 10.3390/forecast5010010.