

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



**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



**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

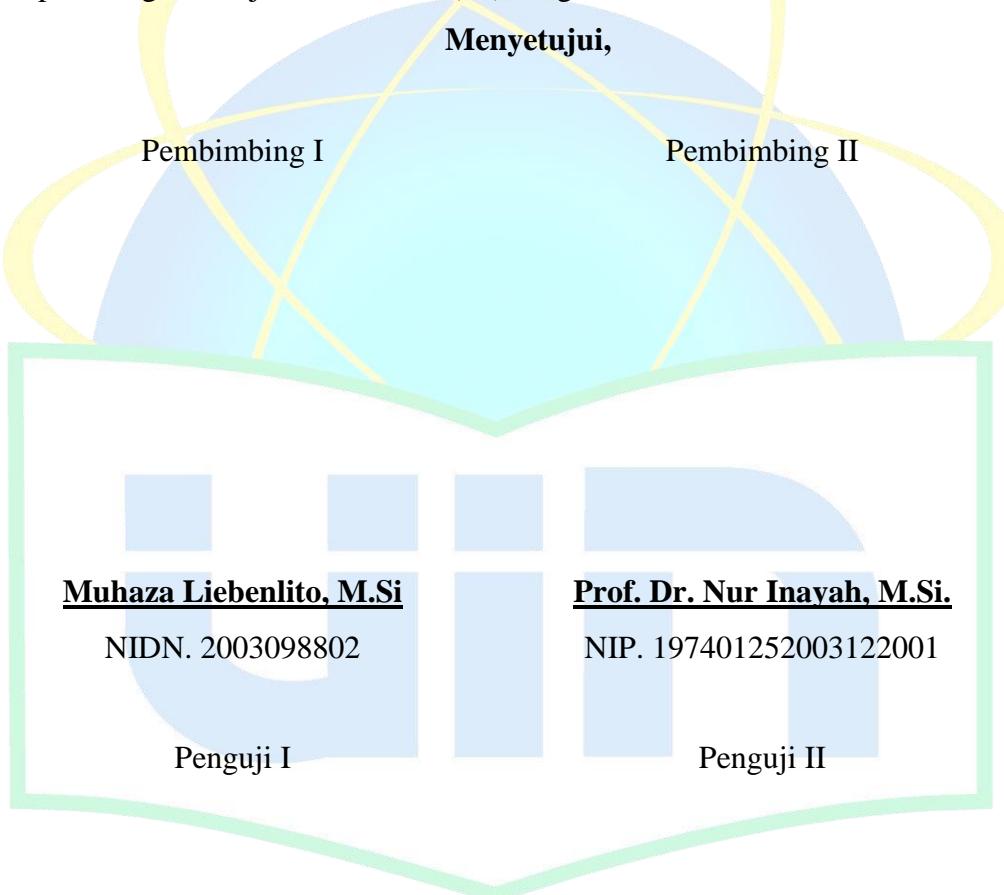
Gilang Islamay Putra Djuharis

NIM. 11190940000055



LEMBAR PENGESAHAN

Skripsi ini berjudul “PREDIKSI HARGA BITCOIN MENGGUNAKAN TRANSFORMER: MENGINTEGRASIKAN HARGA, SENTIMEN, TREND, DAN VOLUME DALAM ANALISIS DERET WAKTU MULTIVARIAT” yang ditulis oleh **Gilang Islamay Putra Djuharis NIM. 11190940000055** telah diujicobakan 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.



XXX

XXX

NIP. XXX

NIP. XXX

Mengetahui,

Dekan Fakultas Sains dan
Teknologi

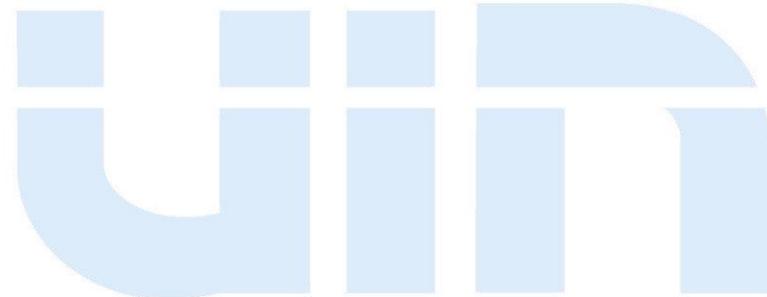
Ketua Program Studi Matematika

Husni Teja Sukmana,
S.T., M.Sc, Ph.D

NIP. 197710302001121003

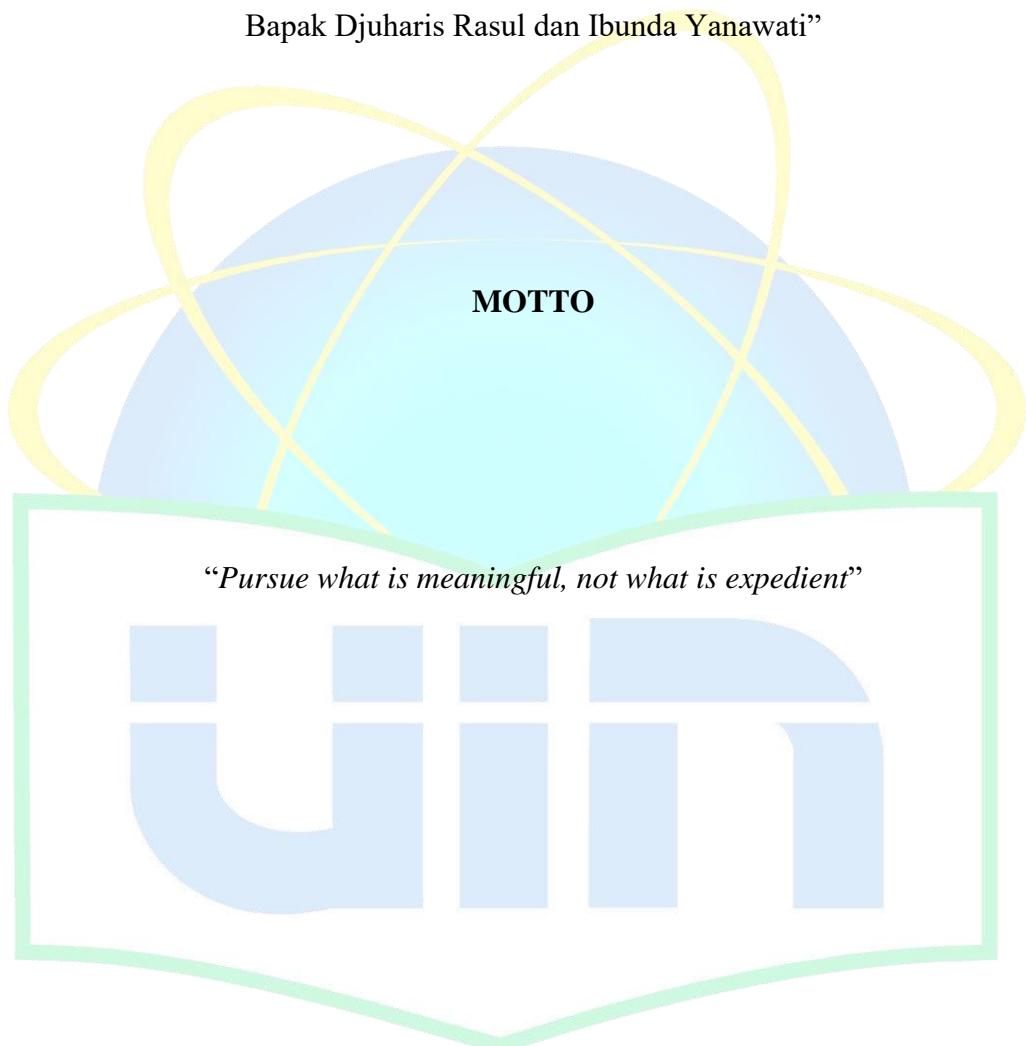
Taufik Edy Sutanto,
M.Sc.Tech.,Ph.D.

NIP. 197905302006041002



PERSEMBAHAN

“Skripsi ini saya persembahkan untuk kedua orang tua saya,
Bapak Djuharis Rasul dan Ibunda Yanawati”



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, TREND, 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 terima kasih 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 memberikan saran, bimbingan, serta masukan yang sangat berharga sehingga penulis dapat menyelesaikan skripsi ini.
4. XXX selaku Pengaji I dan XXX selaku Pengaji 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 sehingga 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 di antara lain Risky Amalia Marharyadi yang membantu melabeli data manual, yang telah memberikan wawasannya, 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 volatilitasnya 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 Transformer yang menggunakan mekanisme *attention* untuk keduanya. Diintegrasikan analisis sentimen, tren, dan volume ke dalam variabel model dan didapatkan model yang paling mutakhir adalah dengan hanya memasukkan variabel sentimen serta tren 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 Transformer, 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
PERSEMBAHAN.....	v
KATA PENGANTAR	vi
ABSTRAK.....	viii
<i>ABSTRACT.....</i>	ix
DAFTAR ISI.....	x
DAFTAR GAMBAR	xii
DAFTAR TABEL.....	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Perumusan Masalah	4
1.3 Batasan Masalah	4
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	5
BAB II LANDASAN TEORI.....	6
2.1 Analisis Sentimen	6
2.2 <i>Recurrent Neural Network (RNN)</i>	6
2.3 Long-Short Term Memory.....	7
2.4 Transformer.....	8
2.5 Bidirectional Encoder Representations Transformers (BERT)	11
2.6 Evaluasi Model	12
BAB III METODOLOGI PENELITIAN	14
3.1 Data Penelitian	14
3.2 Robustly Optimized Bert Pretraining Approach (roBERTa).....	14

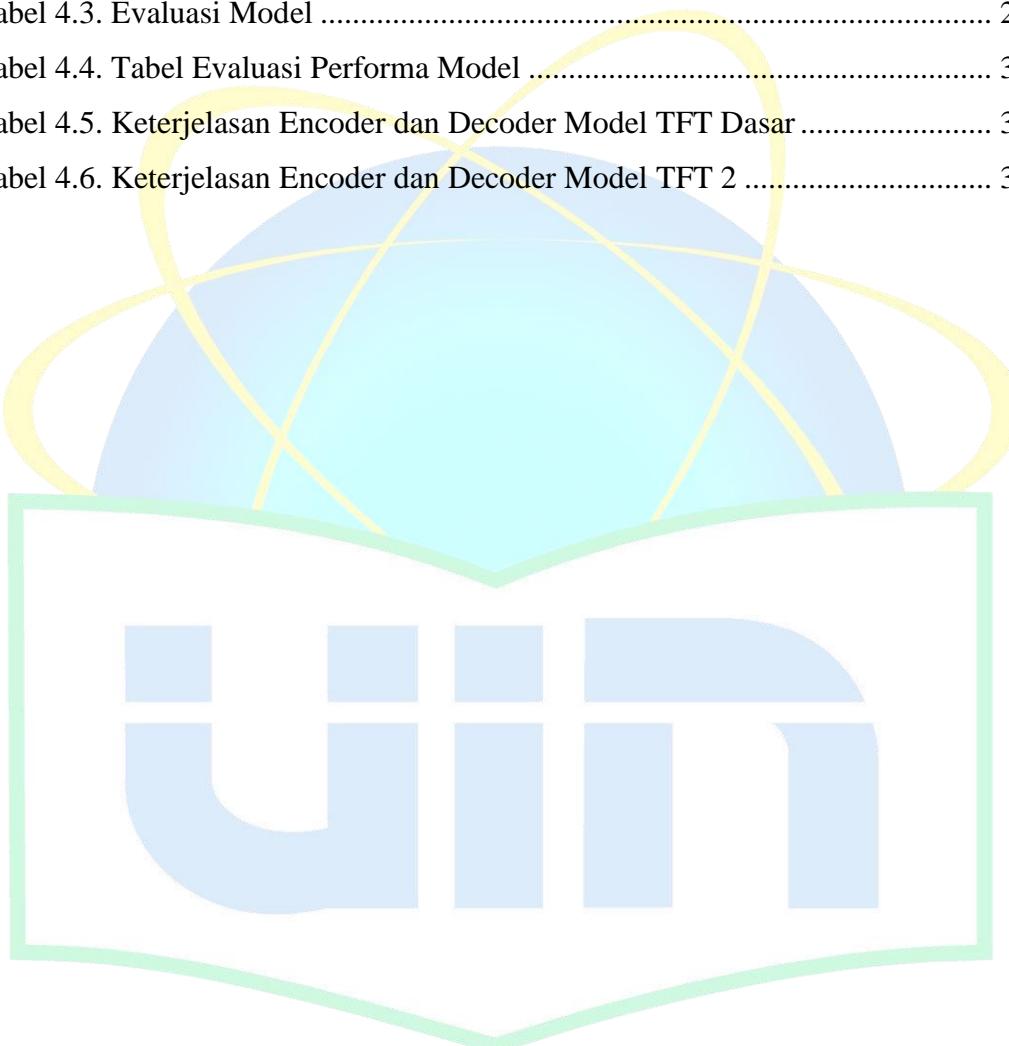
3.3 Temporal Fusion Transformer	15
3.4 Pra Pemrosesan Data Teks	16
3.5 Augmentasi Data.....	19
3.6 Nilai Pencilan.....	21
3.7 Multiple Seasonal-Trend decomposition using LOESS	22
<i>3.8 Time Lag Plot (TLP).....</i>	22
3.9 Transformasi Data Rangkaian Waktu.....	22
3.10 Tahapan Penelitian.....	23
BAB IV HASIL DAN PEMBAHASAN	26
4.1 Analisis Sentimen	26
4.1.1 Pembangunan Model Analisis Sentimen	27
4.1.2 Hasil Model Analisis Sentimen	27
4.2 Analisis Data Eksplorasi	29
4.2.1 Time Lag Plot (TLP).....	30
4.2.2 Multiple Seasonal-Trend decomposition using LOESS (MSTL) ..	30
4.2.3 <i>Outlier</i> (Nilai Pencilan).....	32
4.2.4 Tes Stasioner	33
4.3 Peramalan Deret Waktu	33
4.3.1 TFT Model Dasar.....	35
4.3.2 Model TFT tanpa Volume dan Sentimen.....	36
BAB V KESIMPULAN DAN SARAN	39
BAB VI DAFTAR PUSTAKA.....	40

DAFTAR GAMBAR

Gambar 2.1 Arsitektur <i>Recurrent Neural Network</i>	6
Gambar 2.2. Arsitektur <i>Long-Short Term Memory</i>	7
Gambar 2.3. Arsitektur Transformer.....	9
Gambar 2.4. Arsitektur <i>Attention Head</i>	10
Gambar 2.5. Arsitektur Kalkulasi <i>Attention</i>	10
Gambar 2.6. Arsitektur BERT	11
Gambar 3.1. Arsitektur Temporal Fusion Transformer	15
Gambar 3.2. Mekanisme VSN	16
Gambar 3.3. Jumlah Data Latih	21
Gambar 4.1. Matriks Konfusi dari Model Kandidat	26
Gambar 4.2. Matriks Konfusi Model Terbaik.....	28
Gambar 4.3. Harga dibandingkan Variabel penjelas	29
Gambar 4.4. <i>Lag Plot</i> Harga dibandingkan Variabel penjelas.....	30
Gambar 4.5. MSTL Harga	31
Gambar 4.6. Harga per Bulan	32
Gambar 4.7. Grafik Nilai Penculan.....	33
Gambar 4.8. Data Variabel.....	34
Gambar 4.9. Grafik Keterjelasan Encoder dan Decoder Model TFT Dasar	35
Gambar 4.10. Grafik <i>Attention</i> dari Model TFT Dasar.....	36
Gambar 4.11. Grafik Keterjelasan Encoder dan Decoder Model TFT 2	37
Gambar 4.12. Grafik <i>Attention</i> dari Model TFT 2	38

DAFTAR TABEL

Tabel 3.1. Contoh Tweet Spam.....	17
Tabel 3.2. Akurasi Google Translate	19
Tabel 4.1. Akurasi Model Kandidat.....	26
Tabel 4.2. Akurasi Model setelah Pelatihan Model	27
Tabel 4.3. Evaluasi Model	28
Tabel 4.4. Tabel Evaluasi Performa Model	34
Tabel 4.5. Keterjelasan Encoder dan Decoder Model TFT Dasar	35
Tabel 4.6. Keterjelasan Encoder dan Decoder Model TFT 2	37



BAB I

PENDAHULUAN

1.1 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 Mei 2021. Dengan popularitas mata uang kripto di masyarakat, banyak investor melihat peluang keuntungan. Salah satu mata uang kripto yang paling populer adalah Bitcoin[2]. Data dari coingecko juga menunjukkan bahwa harga Bitcoin, salah satu mata uang kripto yang paling populer, meningkat 405% pada tahun 2020 dan 161% pada tahun 2021, dengan kapitalisasi pasar mencapai 1,28 triliun dolar pada 9 November 2021. Bukan hanya itu, mata uang kripto menawarkan berbagai fitur yang menarik seperti kemudahan penggunaan, keamanan, dan desentralisasi. Mereka dapat diakses melalui berbagai perangkat, menyediakan identitas yang aman dan transparan. Pada saat yang sama, transaksi anonim, dicatat pada blockchain yang mendasarinya tanpa melibatkan perantara seperti bank[3].

Volatilitas harga mata uang kripto yang tinggi merupakan salah satu alasan mengapa sebagian investor enggan memasukinya[4]. Harga Bitcoin telah mengalami penurunan lebih dari 10% tepatnya 59 kali dalam satu hari. Penurunan yang paling signifikan terjadi pada 13 Maret 2020, dengan penurunan 35,19%. Dalam konteks pasar saham, perlu dicatat bahwa harga Bitcoin menunjukkan volatilitas yang signifikan. Bahkan pasar saham menggabungkan Circuit Breaker Mekanisme *Short Sale Price Test Circuit Breaker*, yang dipicu ketika harga mengalami penurunan melebihi 10% dalam satu hari perdagangan[5]. Oleh karena

itu, diperlukan model yang dapat memprediksi harga mata uang kripto di masa mendatang untuk membantu investor dalam memaksimalkan keuntungan dan meminimalkan kerugian.

Banyak penelitian telah dilakukan untuk memprediksi deret waktu menggunakan berbagai model dan data. Misalnya, penggunaan data Google Trends dengan model statistik serta *Machine learning* seperti *Vector Autoregression* (VAR) dan *Random Forest* untuk memperkirakan pergerakan harga Bitcoin berdasarkan harga dan popularitasnya[6]. Penelitian ini menunjukkan bahwa data tren juga dapat menjadi variabel penjelas untuk bitcoin. Selanjutnya, penggunaan Google Trends untuk memprediksi penjualan ritel, penjualan otomotif, real estat, dan tujuan wisata juga mencapai kesimpulan yang sama[7]. Demikian pula, penelitian tentang hubungan antara harga Bitcoin dan trennya juga telah menghasilkan hasil yang positif[2]. Hubungan antara harga dan volume pasar Bitcoin juga terbukti memiliki korelasi[8]. Berdasarkan penelitian ini, dapat disimpulkan bahwa popularitas dan volume mata uang kripto berkorelasi dengan harganya.

Dalam ranah analisis deret waktu *multivariate* (menggunakan lebih dari 2 variabel penjelas), ditemukan bahwa menggunakan model *multivariate* untuk memprediksi harga saham menghasilkan hasil yang lebih baik daripada pendekatan *univariate* (menggunakan kurang dari 2 variabel prediktor)[9]. Analisis *multivariate* telah diterapkan untuk memperkirakan mata uang kripto juga, dengan membandingkan tiga pendekatan menggunakan *Recurrent Neural Network* (RNN) seperti *Long Short-Term Memory* (LSTM), *Bidirectional LSTM* (Bi-LSTM), dan *Gated Recurrent Unit* (GRU). Menggunakan lima variabel: harga penutupan, harga pembukaan, harga tertinggi, harga terendah, dan volume untuk lima mata uang kripto, termasuk Bitcoin, Ethereum, Cardano, Tether, dan Binance Coin. Hasil penelitian menunjukkan bahwa Bi-LSTM dan GRU memiliki kinerja 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 mengungguli dalam set data USDT dan BNB akan tetapi memiliki varians yang lebih tinggi dibandingkan dengan Bi-LSTM dan

GRU[10]. Hubungan antara harga Bitcoin dan sentimen juga telah digunakan untuk memprediksi 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[11].

Penelitian-penelitian tersebut melihatkan 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 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.

Model yang digunakan Chat GPT berasal dari makalah ilmiah berjudul *Attention is All You Need* yang ditulis oleh A. Vaswani, dkk. dari Google pada tahun 2017. Mereka mengusulkan model terbaru yang merupakan peningkatan dari model berbasis *recurrent* untuk *Natural language Processing* (NLP) dengan nama Transformer menggunakan mekanisme *attention*[12]. Dengan arsitekturnya, Transformer dapat memahami bahasa yang melampaui model lain sebelumnya dalam berbagai tolok ukur seperti terjemahan mesin. Salah satu contoh model yang menggunakan Transformer adalah *Bidirectional Encoder Representations from Transformers* (BERT) yang telah digunakan untuk penerjemahan teks, klasifikasi teks, dan kasus penggunaan lainnya[13].

Salah satu subbidang klasifikasi teks adalah analisis sentimen, sebelum penemuan Transformer ada banyak model sentimen yang menggunakan mekanisme *recurrent* untuk berbagai bidang seperti aspek sosial, kesehatan, dan politik. Namun masih sedikit model analisis sentimen yang secara khusus pada bidang mata uang kripto, terutama menggunakan arsitektur Transformer. Beberapa contohnya termasuk CryptoBERT oleh ElKulako, yang dilatih menggunakan 3,2

juta teks media sosial dari platform seperti StockTwits, Telegram, Reddit, dan Twitter tentang cryptocurrency[14]. Yang lainnya adalah CryptoBERT oleh kk08.

Model Transformer telah digunakan di berbagai bidang, tidak hanya untuk menentukan skor sentimen dari kalimat tetapi juga untuk meramal data deret waktu. Misalnya, kemampuan model Transformer untuk memprediksi harga Bitcoin dan Ethereum menggunakan analisis sentimen. Pada tahun 2020, makalah ilmiah lain berjudul "A Transformer-Based Framework for *Multivariate* Time Series Representation Learning" mengusulkan penggunaan arsitektur yang sama untuk peramalan deret waktu dan menemukan bahwa model Transformer mengungguli model lain (Roket, LSTM, XGBoost, dll)[15].

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 menggunakan data dari Twitter serta Reddit, dan popularitasnya berdasarkan 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 yang mana proses memecah teks menjadi token berisikan kata-kata[16] dilimitasi hanya sampai 256 token.

1.2 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?

1.3 Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah:

1. Data teks didapatkan dari hasil *web scrapping* subreddit r/bitcoin dengan judul “*Daily Discussion*” dan data X (dahulunya Twitter) berasal dari kaggle yang merupakan hasil *web scrapping*. Sehingga data mulai dari 03/12/2017 hingga 30/06/2023.
2. Pengambilan token pada bagian pemrosesan bahasa alami dilimitasi maksimal 256.

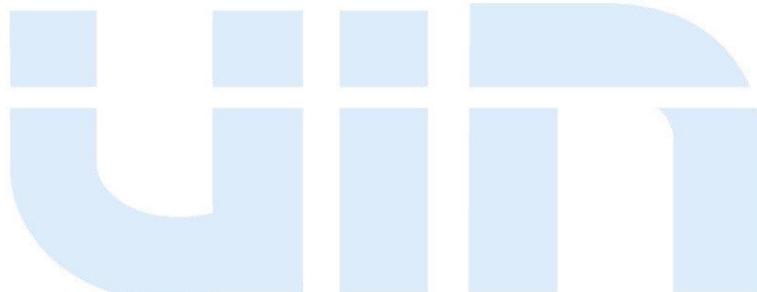
1.4 Tujuan Penelitian

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.

1.5 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

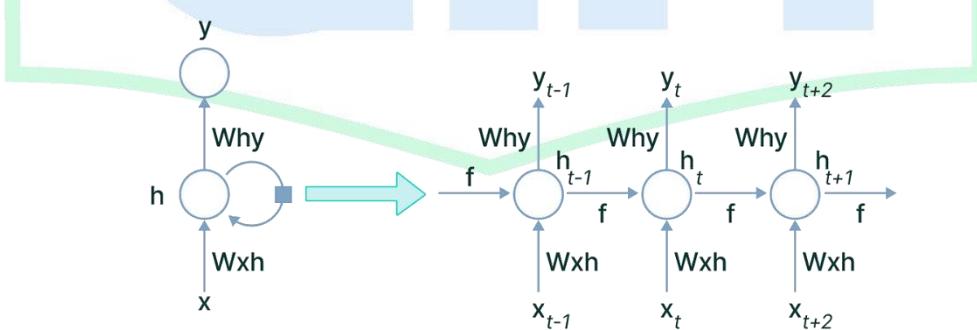
LANDASAN TEORI

2.1 Analisis Sentimen

Analisis Sentimen adalah studi komputasional terhadap pendapat, sikap, dan emosi orang terhadap suatu entitas[17]. Sebelum ditemukannya Transformer, Metode analisis sentimen yang umum digunakan antara lain metode *rule-based* seperti vader[18], *machine learning* seperti Naïve Bayes[19], dan *deep learning* seperti *Long-Short Term Memory*[20]. Namun demikian, seluruh pendekatan tersebut mengalami kesulitan dalam memahami bias dan sindiran karena tidak dapat menginterpretasikan konteks dan tujuan dari suatu kalimat. Beda hal dengan model Transformer yang menempatkan *attention* dan *self-attention* antara satu kalimat dengan kalimat lainnya, sehingga model dapat mengerti konteks dan tujuan dari suatu kalimat tersebut. Dalam mendeteksi sarkasme, model Transformer terbukti lebih unggul dibandingkan dengan model LSTM[21].

2.2 Recurrent Neural Network (RNN)

Recurrent Neural Network adalah arsitektur yang pada umumnya berstruktur berkala, dimana memiliki tujuan untuk mendeteksi pola dalam suatu urutan data[22]. Keunggulan dari model RNN sendiri dibandingkan model lainnya seperti *feedforward networks* adalah dengan adanya siklus yang memberikan informasi ke diri sendiri.



Gambar 2.1 Arsitektur Recurrent Neural Network

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

1. Input Layer

Fitur input untuk setiap t-waktu

2. Hidden State

Menyimpan informasi mengenai input sebelumnya dan diperbarui setiap langkah.

3. Recurrent Connection

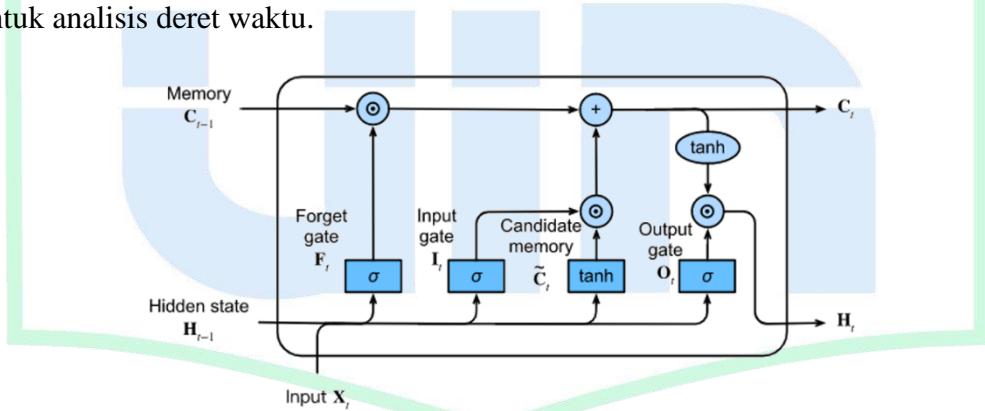
Dimana setiap t-waktu, *hidden state* dari waktu t-1 digunakan untuk kombinasi dengan input saat itu untuk mengeluarkan hasil dan juga memperbarui *hidden state*.

4. Output Layer

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

2.3 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 variabel, karena itu model ini cocok untuk analisis deret waktu.



Gambar 2.2. Arsitektur Long-Short Term Memory

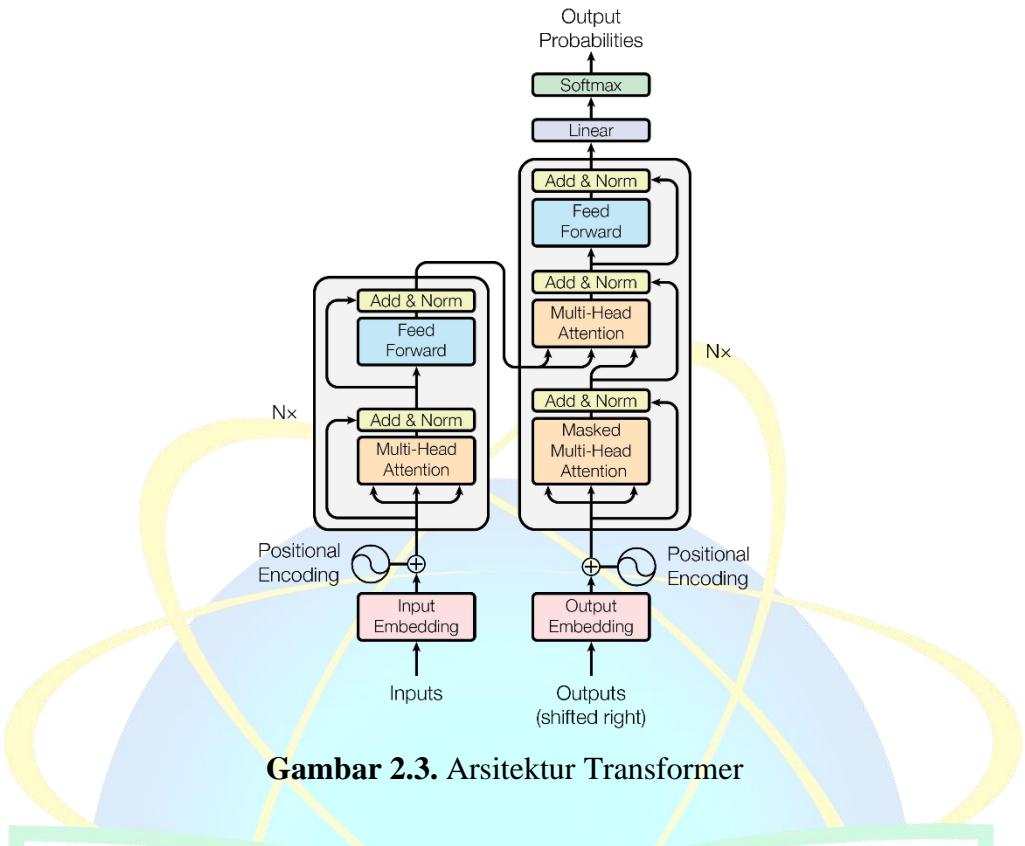
Mengikuti arsitektur RNN pada umumnya yaitu dengan struktur rantai, setiap *cell* dalam LSTM memiliki 4 gerbang. Yaitu yang pertama adalah *Forget Gate* dimana dilakukan fungsi sigmoid untuk mendeterminasi seberapa banyak yang harus dilupakan pada ingatan jangka panjang (*Long-Term Memory*). Kedua yaitu

Input Gate dengan fungsi sigmoid yaitu berapa banyak yang harus diingat serta yang ketiga *Candidate Memory* adalah apa saja yang harus diingat. Yang terakhir adalah *output gate*, dimana pada gerbang ini dilakukan fungsi sigmoid untuk mendeterminasi hasil untuk *cell* selanjutnya.[23]

2.4 Transformer

Model RNN yang sudah ada sejak tahun 1986 yang lalu dikembangkan menjadi 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 paralelisasi pada model. Dan juga model RNN rentan akan *The Vanishing Gradient Problem*[24]

Transformer merupakan model terbaru yang merupakan peningkatan dari model berbasis *recurrent* untuk *Natural language Processing* (NLP) 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.[12]



Gambar 2.3. Arsitektur Transformer

1. *Positional Encoding*

Sebelum input masuk ke dalam 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 input yang dimasukan menggunakan fungsi sin dan cos berikut.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (1)$$

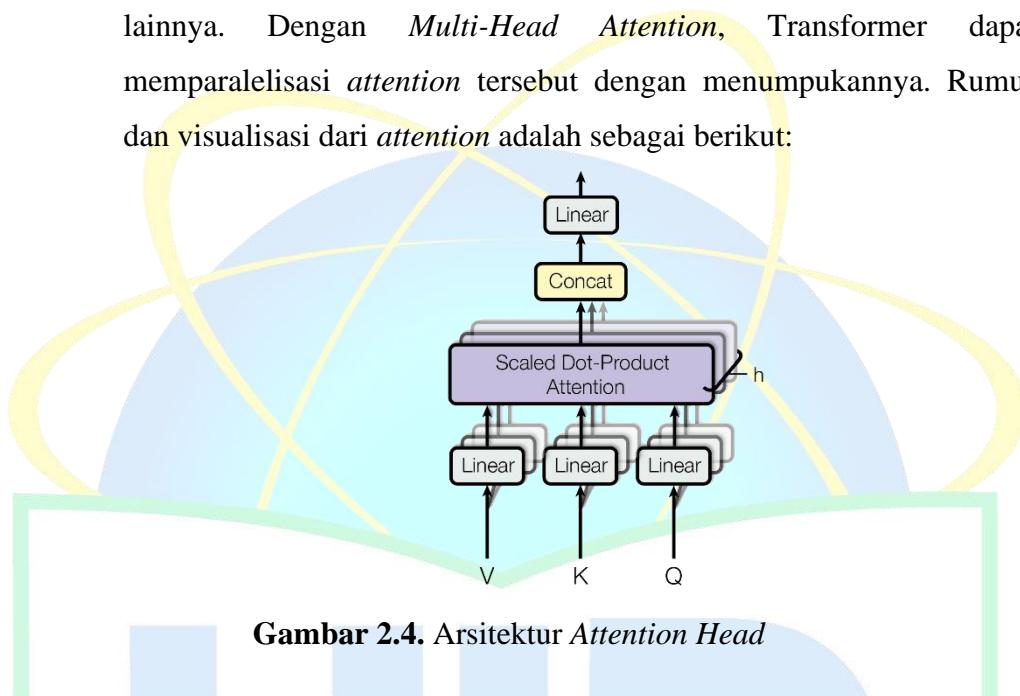
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (2)$$

Dimana pos adalah posisi dan i adalah dimensi. Karena sin dan cos merupakan fungsi periodik, penggunaan sin dan cos dalam *positional*

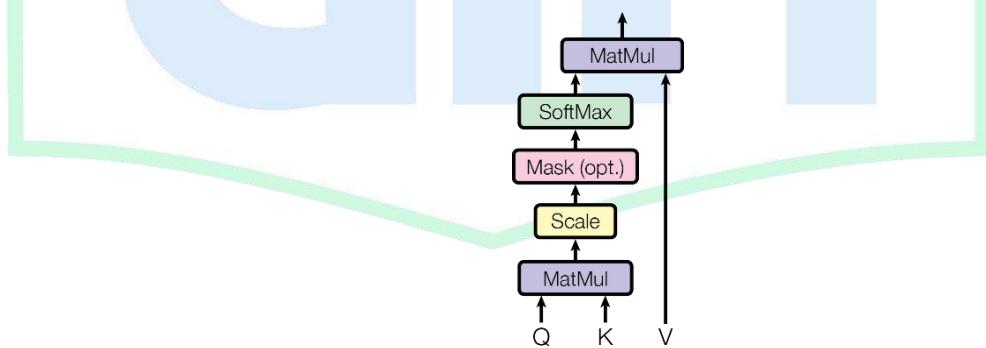
encoding dilakukan untuk membantu model dalam menangkap hubungan antara data meskipun jaraknya jauh.

2. 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:



Scaled Dot-Product Attention dapat dijelaskan lebih lanjut dari gambar 2.5.



Gambar 2.5. Arsitektur Kalkulasi Attention

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (3)$$

dimana Q adalah Query, D_k adalah key dari dimensi, dan V adalah value dari dimensi.

3. Add & Norm

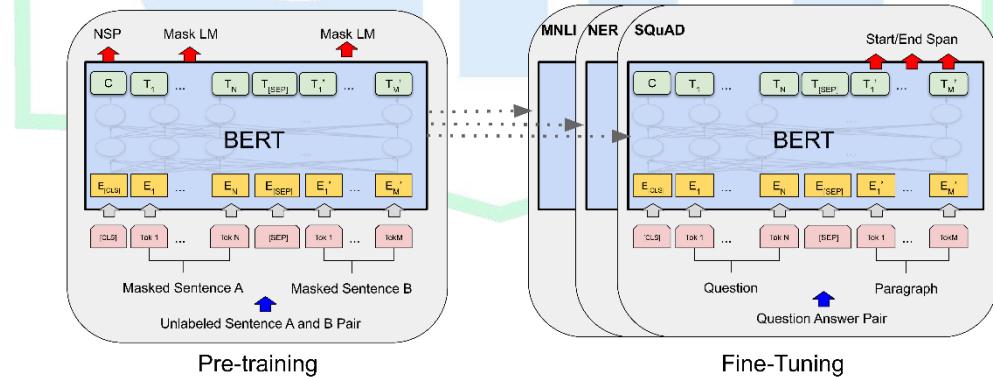
Hasil lalu ditambahkan dengan data yang sebelumnya dan dinormalisasikan. Ini dilakukan guna menangkap informasi terbaru sembari tidak melupakan informasi yang telah ada dalam data itu sendiri, dan dinormalisasi untuk mendapatkan data yang lebih konvergen.

4. 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.

2.5 Bidirectional Encoder Representations Transformers (BERT)

BERT merupakan salah satu model yang lahir dari arsitektur Transformer, yang mana menggunakan beberapa lapisan encoder dari arsitektur Transformer[13]



Gambar 2.6. 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.

2. *Masked Language Model*

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

2.6 Evaluasi Model

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

$$\text{Akurasi} = \frac{\text{Total prediksi label yang benar}}{\text{Total prediksi}} \quad (4)$$

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

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

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

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

Evaluasi peramalan peramalan deret waktu digunakan *Root Mean Square error* (RMSE), *Mean Absolute Error* (MAE), serta *Mean Absolute Percentage Error* (MAPE) untuk evaluasinya[25].

$$RMSE = \sqrt{\sum_{t=1}^n \frac{(A_t - F_t)^2}{n}} \quad (8)$$

$$MAE = \sum_{t=1}^n \frac{|A_t - F_t|}{n} \quad (9)$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|A_t - F_t|}{|A_t|} \quad (10)$$

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

BAB III

METODOLOGI PENELITIAN

Penelitian ini mengadopsi pendekatan eksperimental 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 menganalisis kinerja model terhadap variabel-variabel yang digunakan.

3.1 Data Penelitian

Data yang digunakan pada penelitian merupakan penggabungan antara data primer dan sekunder. Pertama data primer teks reddit didapatkan melalui *web scrapping* (pengambil informasi dari suatu web)[26]. Dan data sekunder teks Twitter didapatkan melalui situs web kaggle, harga serta volume bitcoin melalui API CoinGecko, dan tren melalui situs web Google Trends.

Data teks Twitter mulai dari 05 Februari 2021 hingga 09 Januari 2023, akan tetapi karena ada beberapa hari yang kosong dan untuk memperbanyak data maka dilakukan *web scrapping* dari situs web reddit dengan subreddit r/bitcoin dengan judul “*Daily Discussion*”. Sehingga pada akhirnya data yang digunakan mulai dari 03 Desember 2017 hingga 30 Juni 2023.

3.2 Robustly Optimized Bert Pretraining Approach (roBERTa)

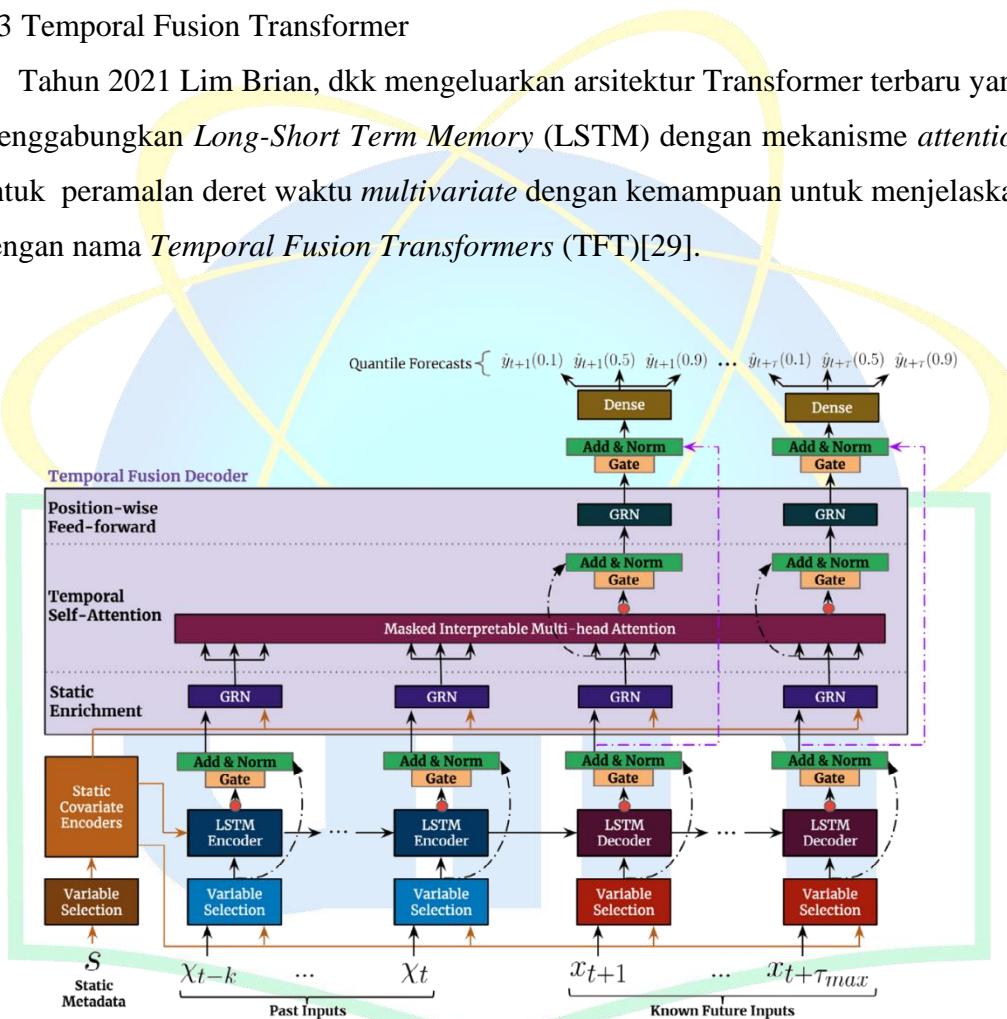
roBERTa merupakan versi BERT yang lebih dikembangkan lagi, ditemukan bahwa BERT memiliki kekurangan yaitu model tidak terlatih secara optimal[27]. Walaupun menggunakan arsitektur yang sama, roBERTa berhasil mendapatkan performa yang lebih baik dibandingkan BERT dengan 4 perubahan.

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[13]. Akan tetapi, beberapa penelitian berargumen bahwa justru keberadaan NSP mengurangi performa model itu sendiri[28]. Maka dilakukan penghapusan objektif NSP untuk model roBERTa. Dengan perubahan ini roBERTa mendapatkan performa yang lebih baik dibandingkan BERT orisinal.

3.3 Temporal Fusion Transformer

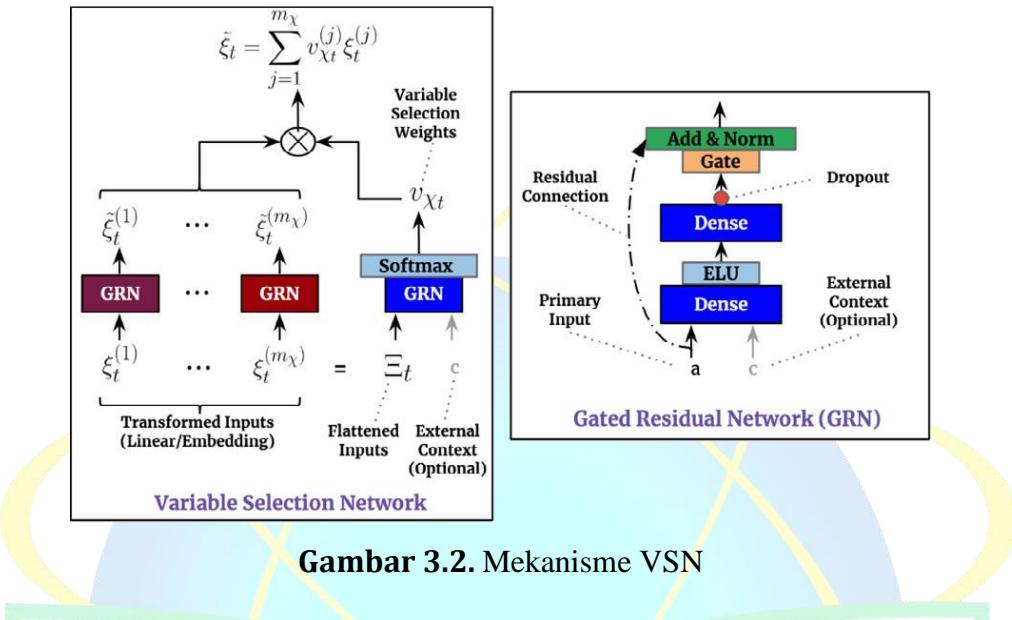
Tahun 2021 Lim Brian, dkk mengeluarkan arsitektur Transformer terbaru yang menggabungkan *Long-Short Term Memory* (LSTM) dengan mekanisme *attention* untuk peramalan deret waktu *multivariate* dengan kemampuan untuk menjelaskan dengan nama *Temporal Fusion Transformers* (TFT)[29].



Gambar 3.1. 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 variabel penjelas, ditambah dengan *Multi-Head Attention* sehingga model dapat menimbang variabel penjelas 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.



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[30]. 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 effects*)[29].

3.4 Pra Pemrosesan Data Tekstual

Pengumpulan data dilakukan dengan mengambil data Twitter melalui Kaggle dan *web scraping* sosial media reddit dilakukan dengan mengambil komen yang terdapat pada judul halaman “daily discussion” di r/bitcoin, guna mengisi kekosongan data dan meluaskan cakupan data sentimen analisis itu sendiri menjadi 03/12/2017 hingga 30/06/2023. Serta data historis bitcoin diambil melalui API yang disediakan oleh CoinGecko. 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.[31] 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[32]. Pengaplikasian pra pemrosesan data dilakukan dengan menghapus tautan 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 kepopuleritasannya juga menyebabkan banyaknya tweet yang bersifat spam dan dibuat oleh robot. Maka pra pemrosesan juga dilakukan dengan menghapus tweet yang bersifat spam, yang mana contoh dari tweet yang bersifat spam adalah sebagai berikut:

Tabel 3.1. Contoh Tweet Spam

<i>date</i>	<i>Text</i>
2021-07-03 20:53:51	<p>The \$BTC price is at \$34655.23 right now.</p> <ul style="list-style-type: none">➊ Compared to the last tweet, the price has dropped by \$72.77 (-0.21%).➋ 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	❶ ❷ Bitcoin Whale Alert: [TX: 6e2f93751abd08b63306d5351da016187927ac712f442d67 09d23d10b01e83c8]-[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: <input checked="" type="checkbox"/> ⚡ \$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/fSCKGmhcK... 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 laman web 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

$$\sum_{i=1}^n \frac{s(x_i)}{n} \quad (11)$$

dimana $s(x_i)$ adalah skor sentimen dari teks x_i dalam interval satu hari, dan n adalah banyak data teks dalam interval tersebut.[33]

3.5 Augmentasi Data

Augmentasi data banyak dilakukan pada computer vision[34], akan tetapi pada bidang NLP masih sedikit yang membahas[35]. *Back-Translation* adalah teknik augmentasi data yang sangat efektif untuk terjemahan mesin neural[36], yang mana pendekatan ini menjanjikan karena memiliki kemampuan mempertahankan label yang baik secara alami dan kemampuan parafrase yang sangat berharga[37]. *Back-Translation* dilakukan dengan teks asli dikembalikan ke dalam bahasa aslinya setelah dua kali terjemahan. Teks asli S1 diterjemahkan ke dalam bahasa lain sebagai S2, dan kemudian diartikan kembali ke dalam bahasa aslinya sebagai S3.[38]

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[39]. 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 50 bahasa di google translate. Dan pada penelitian kali ini diambil 20 bahasa paling akurat[40]

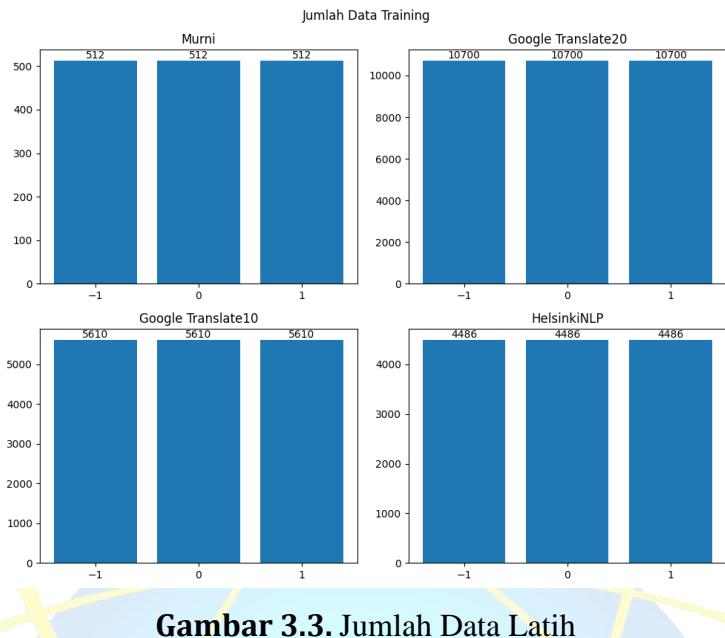
Tabel 3.2. Akurasi Google Translate

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 Hugging Face 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 3.3. Jumlah Data Latih

dimana Google Translate10 (GTH) adalah dengan augmentasi 10 bahasa paling akurat di Google Translate, dan Google Translate20 (GT) adalah dengan augmentasi 20 bahasa paling akurat di Google Translate

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.[41] Analisis akurasi model paling akurat setelah data diagumentasi menggunakan EDA juga dilakukan pada penelitian kali ini.

3.6 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 pendekripsi nilai pencilan pada data ini. Penghapusan sekitar 10% dari nilai pencilan meningkatkan performa model untuk kebanyakan *machine learning*[42], teknik yang digunakan untuk menggantikan nilai pencilan adalah *Moving Average* atau rata-rata berjalan[43].

$$y'_t = \frac{\sum_{i=0}^{n-1} y_{t-i}}{n} \quad (12)$$

dimana y'_t adalah pengganti titik observasi, y_t adalah titik observasi, dan n adalah besar jendela observasi[44]. Metode Box-Plot juga dilakukan sebagai metode dasar perbandingan dengan metode iForest.

3.7 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[45]. Dengan memberikan dekomposisi aditif dari data rangkaian waktu. Diberikan X_t dimana t adalah observasi pada waktu ke-t, maka dekomposisinya dapat didefinisikan sebagai berikut:

$$X_t = \hat{S}_t + \hat{T}_t + \hat{R}_t \quad (13)$$

dimana \hat{S}_t , \hat{T}_t , \hat{R}_t masing-masing melambangkan musiman, tren, dan sisa dari observasi. Dan menjadi jika memiliki lebih dari satu musim maka dekomposisinya menjadi

$$X_t = \hat{S}_t^1 + \hat{S}_t^2 + \dots + \hat{S}_t^n + \hat{T}_t + \hat{R}_t \quad (14)$$

dimana n adalah banyak musim yang dimiliki oleh X_t

3.8 Time Lag Plot (TLP)

TLP memeriksa apakah data bersifat random atau tidak, dimana grafik poin digambarkan pada grafik 2 dimensi (x-y) dan x ditetapkan pada waktu $ke - i$ dan y ditetapkan pada waktu $ke - (i + n)$ dimana n adalah besar lag yang ditetapkan.

3.9 Transformasi Data Rangkaian Waktu

Tes Augmented Dickey-Fuller[46] dilakukan untuk mengatasi fluktuasi dan volatilitas tinggi harga serta volume bitcoin, data yang tidak stasioner diubah menjadi stasioner dengan menerapkan *detrending* atau penghapusan tren.[47] Teknik *detrending* yang dilakukan pada penelitian kali ini adalah *differencing transformation* atau dengan menghasilkan deret waktu baru di mana nilai baru y'_{t_i} pada waktu t_i dikalkulasikan dengan perbedaan diantara observasi orisinal dan observasi $y_{t_{i-1}}$ pada langkah waktu sebelumnya. Rumusnya dapat dijelaskan sebagai berikut.

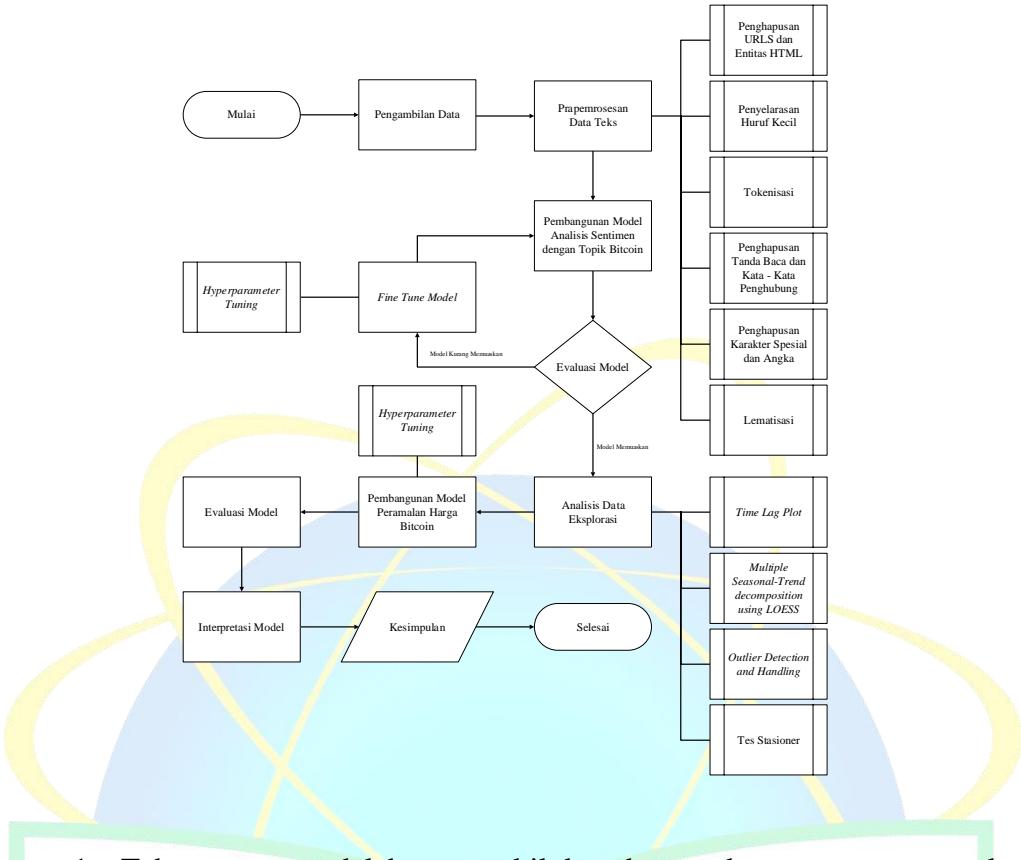
$$y'_{t_i} = y_{t_i} - y_{t_{i-1}} \quad (15)$$

Normalisasi data juga dilakukan guna meningkatkan performa model[47], digunakan normalisasi Min-Max sehingga nilai nilai di dalam data dipetakan dalam rentang (0, 1) dengan rumus sebagai berikut

$$y_{t_i} = \frac{y_{t_i} - y_{\min}}{y_{\max} - y_{\min}} \quad (16)$$

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

3.10 Tahapan Penelitian



1. Tahap pertama adalah pengambil data dengan data pertama merupakan data teks dari www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets format CSV, data berisikan 13 kolom yang mana diambil hanya kolom *date* dan *text*. Data kedua merupakan hasil *web scrapping* dari halaman web reddit dengan subreddit r/bitcoin dengan tema perbincangan “Daily Discussion”, dengan kedua data teks digabungkan maka didapatkan data teks mengenai topik Bitcoin dari 3 Desember 2017 hingga 30 Juni 2023 dengan total sebanyak 5.585.180 data yang akan diolah. Data ketiga adalah data harian google trends, dimana diambil data tiap tahun untuk mendapatkan data hariannya. Dan data terakhir adalah harga dan volume bitcoin yang diambil menggunakan API CoinGecko.
2. Data teks lalu diambil secara acak untuk dilakukan pelabelan secara manual dan didapatkan teks dengan label negatif sebanyak 717, netral sebanyak 1371, dan positif sebanyak 1332 yang mana akan diperbanyak lebih lanjut dengan augmentasi data.

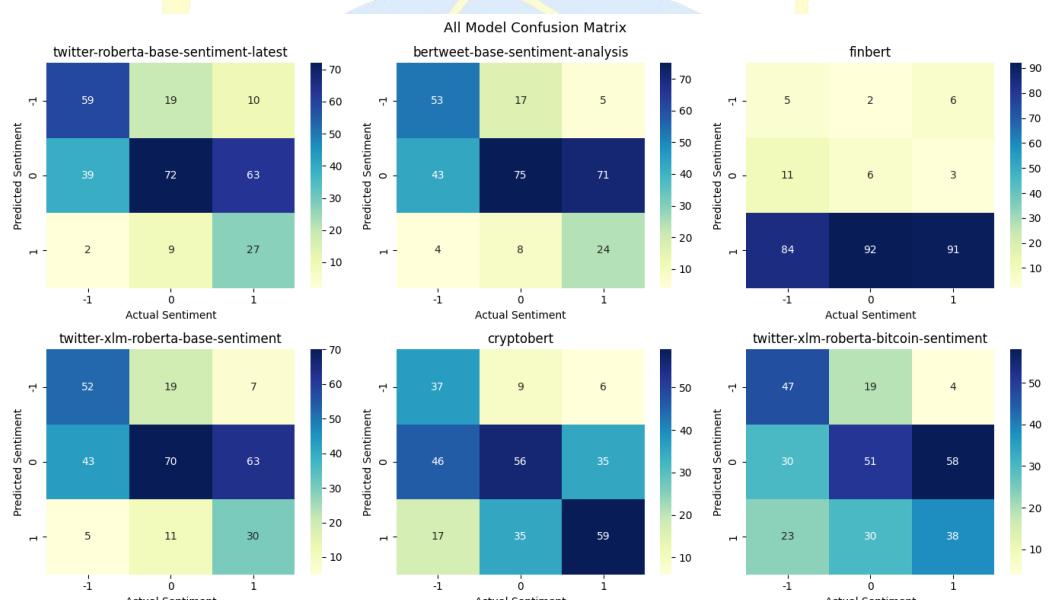
3. Data teks selanjutnya melewati proses prapemrosesan (pembersihan) yaitu dengan penghapusan URLS dan Entitas HTML, mengubah seluruh huruf menjadi huruf kecil, tokenisasi, penghapusan tanda baca dan kata – kata penghubung, penghapusan karakter spesial dan angka, lalu lematisasi.
4. Setelah data teks dibersihkan, augmentasi data dilakukan dengan teknik *Back-Translation* dan dilakukan pembersihan kembali dengan diakhiri *undersampling*. Sehingga data yang digunakan yang berawal dari 2.151 data menjadi 3 dataset lainnya dengan jumlah data 13.458, 16.380, dan 31.100.
5. Pembangunan model analisis sentimen dilakukan berdasarkan 5 model kandidat yang memiliki akurasi paling besar saat dites terhadap data yang sudah dilabeli manual. Dan *Hyperparameter Tuning* dilakukan dengan menggunakan modul optuna untuk mendapatkan model dengan akurasi yang paling maksimal. Setelah model paling akurat didapatkan, pelabelan pada semua data teks dilakukan menggunakan model untuk mendapatkan rata rata skala sentimen tiap harinya.
6. Analisis data eksplorasi dilakukan guna mengenali dan melakukan prapemrosesan data, dengan beberapa tes yang dilakukan adalah *Time Lag Plot*, *Multiple Seasonal-Trend decomposition using LOESS*, deteksi serta penanganan data pencilan, dan tes stasioner.
7. Pembangunan model peramalan harga bitcoin dilakukan dengan empat model untuk membandingkan satu model dengan model lainnya. Menggunakan modul optuna untuk *Hyperparameter Tuning*, dibandingkan model Long-Short Term Memory, Gated Recurrent Unit, Temporal Convolutional Networks, dan Temporal Fusion Transformer.
8. Interpretasi dari model Temporal Fusion Transformer dilakukan dengan melihat *explainability* dari model itu sendiri. Yaitu melihat skala kepentingan variabel dalam encoder serta decoder, dan rata rata *attention* pada saat pelatihan dan prediksi.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Analisis Sentimen

Menjelajahi laman web Hugging Face, ditemukan 6 kandidat model untuk sentimen analisis dengan satu diantaranya khusus untuk topik bitcoin. Eksperimen 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 4.1. Matriks Konfusi dari Model Kandidat

Tabel 4.1. 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 gambar 4.1. dan tabel 4.1. maka dapat diketahui terdapat 4 kandidat model dengan tingkat akurasi total terbesar (ditebal). 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

4.1.1 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.

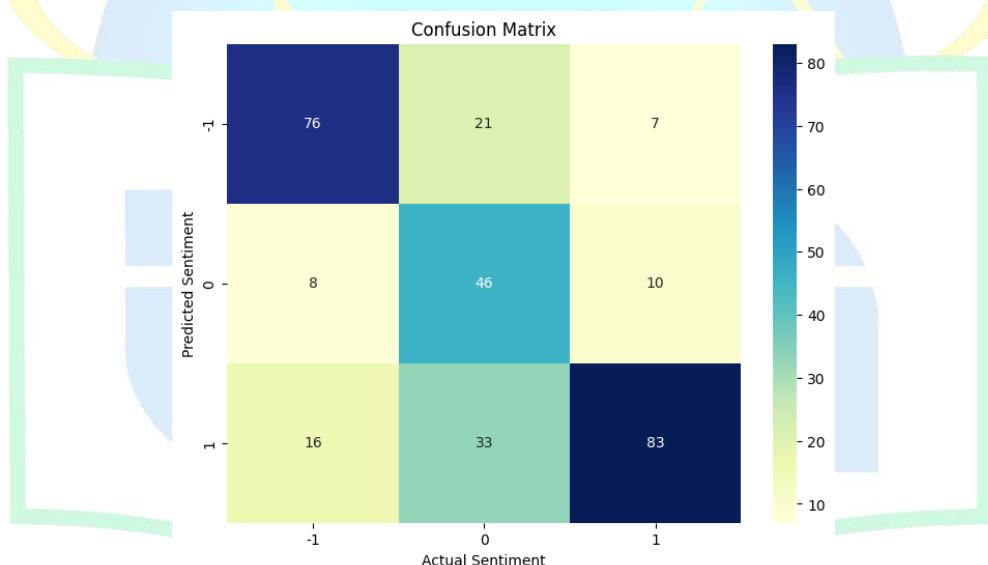
4.1.2 Hasil Model Analisis Sentimen

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

Tabel 4.2. 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://Hugging Face.co/AfterRain007/cryptobertRefined](https://HuggingFace.co/AfterRain007/cryptobertRefined)) Berikut merupakan confusion matrix dan hasil evaluasinya



Gambar 4.2. Matriks Konfusi Model Terbaik

Tabel 4.3. Evaluasi Model

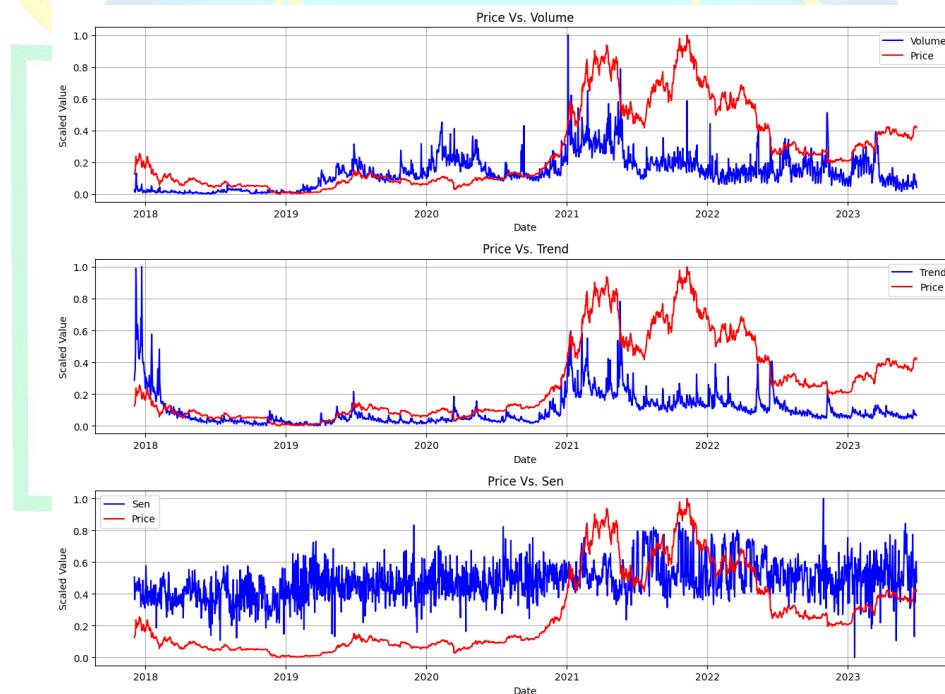
Sentiment Score	Precision	Recall	F1-Score	n
-1	0,73	0,76	0,75	100
0	0,72	0,48	0,56	100

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.

4.2 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.

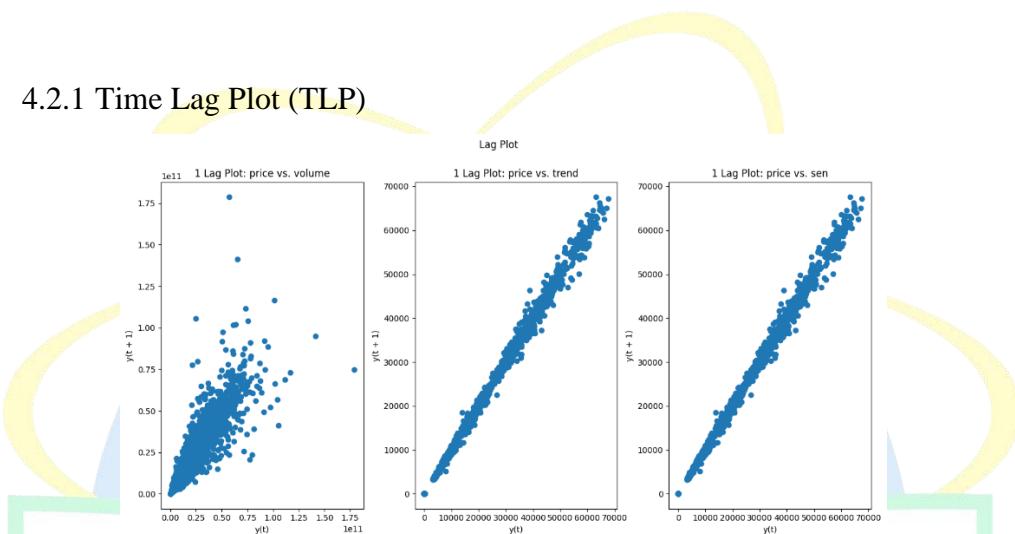


Gambar 4.3. Harga dibandingkan Variabel penjelas

Sekilas, dari tabel diatas dapat dilihat harga dan sentimen tidak memiliki korelasi sama sekali dari tahun ke tahun. Akan tetapi beda hal dengan tren serta

volume, dari tahun ke tahun tren sangat amat berkorelasi 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.

4.2.1 Time Lag Plot (TLP)

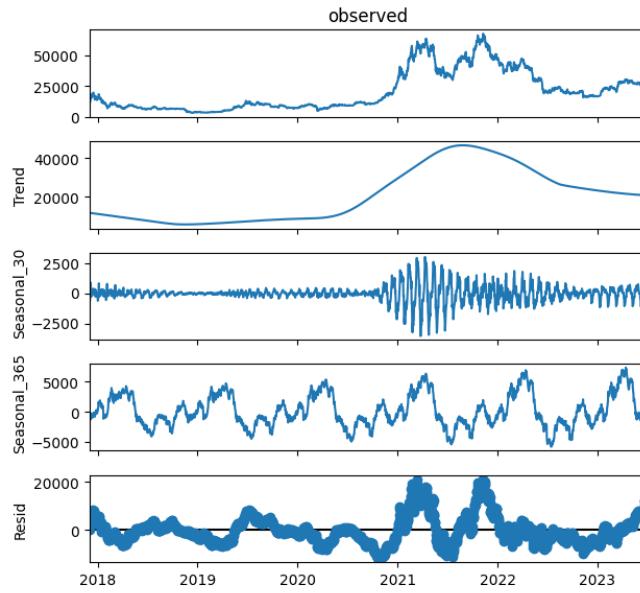


Gambar 4.4. Lag Plot Harga dibandingkan Variabel penjelas

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 penculan (dijelajahi lebih lanjut pada bagian *isolation forest*)

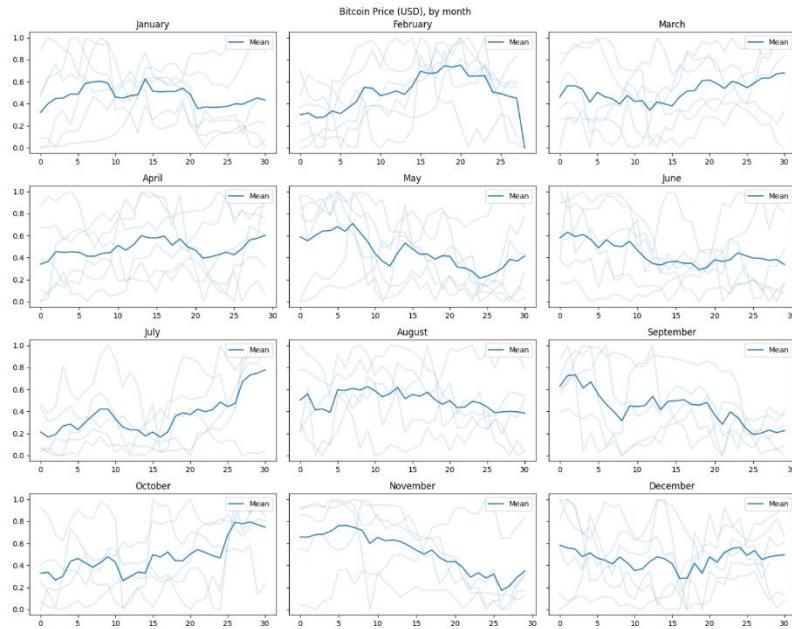
4.2.2 Multiple Seasonal-Trend decomposition using LOESS (MSTL)

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 4.5. 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.

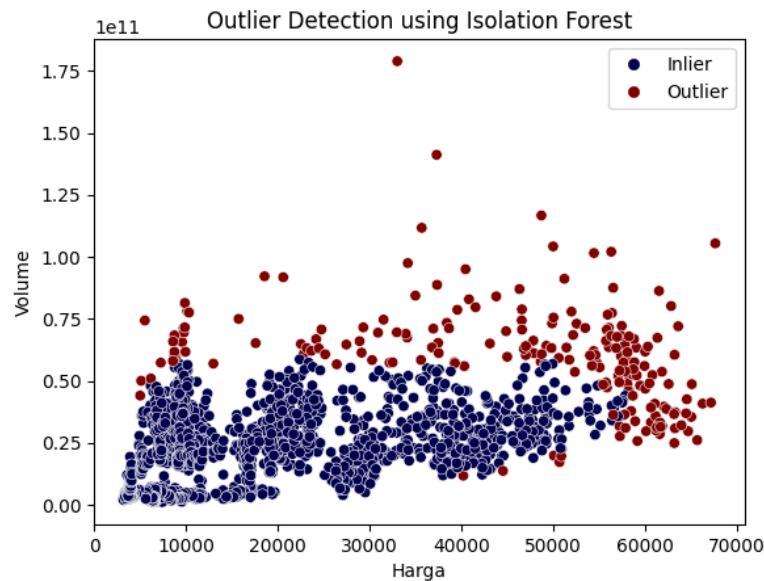


Gambar 4.6. Harga per Bulan

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).

4.2.3 *Outlier* (Nilai Penculan)

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



Gambar 4.7. Grafik Nilai Pencilan

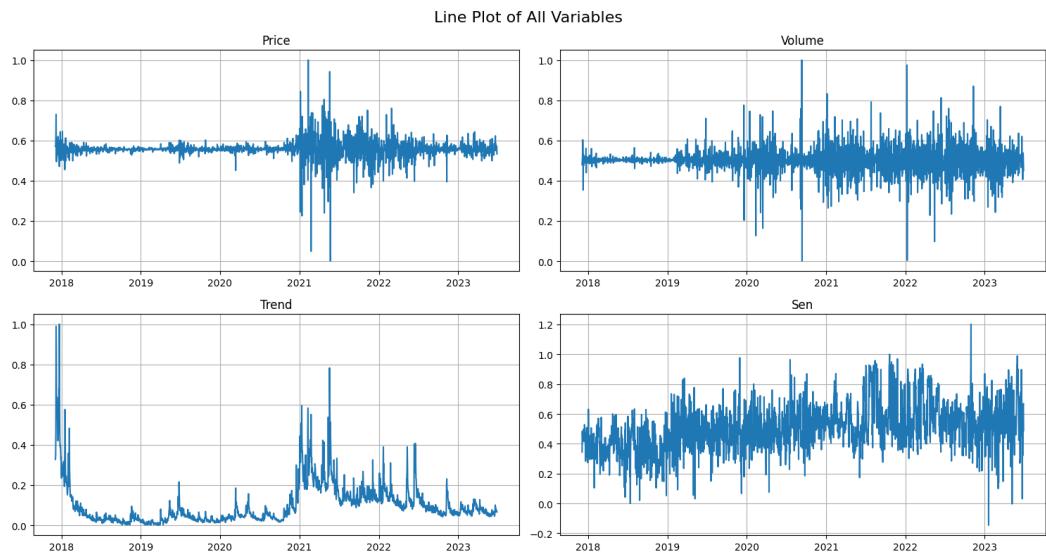
Berdasarkan Isolation Forest, terdapat 204 nilai pencilan berdasarkan 4 variabel yang ada, penanganan nilai pencilan dilakukan dengan mengambil 10% nilai pencilan yang memiliki score paling tinggi dan dilakukan penanganan dengan menggantikan nilai pencilan menggunakan rata rata berjalan. Dan penanganan nilai pencilan dilakukan juga menggunakan teknik *box plot* untuk eksperimen.

4.2.4 Tes Stasioner

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

4.3 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 4.8. Data Variabel

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

Tabel 4.4. Tabel Evaluasi Performa Model

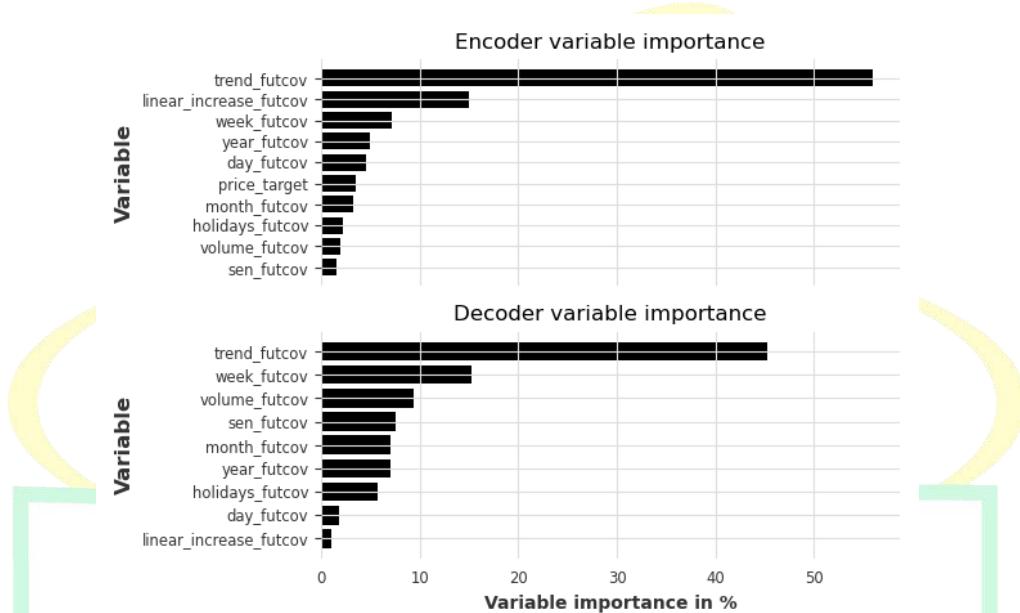
Model	Moving Average			Box Plot			Waktu (s)
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	
TFT	0,0261	0,0177	3,0973	0,0463	0,0312	6,7257	166,14
TFT_{w/o}	0,0263	0,0176	3,0737	0,0479	0,0316	6,9818	
TCN	0,0263	0,0175	3,0627	0,0477	0,0322	7,0084	46,94
TCN_{w/o}	0,0262	0,0172	3,0166	0,048	0,0318	7,0042	
GRU	0,0262	0,0175	3,0764	0,0475	0,0316	6,9442	77,98
GRU_{w/o}	0,0263	0,0175	3,0678	0,0479	0,0318	7,025	
LSTM	0,0262	0,0174	3,0532	0,0477	0,0314	6,9306	74,05
LSTM_{w/o}	0,0264	0,0174	3,0469	0,0479	0,0317	7,0063	

Berdasarkan tabel 4.4. (bold nilai terkecil), dapat disimpulkan bahwa berdasarkan evaluasi yang telah dilakukan, model Temporal Fusion Transformers lebih mutakhir dibandingkan model *deep learning* lainnya seperti LSTM, GRU, dan TCN. Dan model juga mengalami peningkatan saat dimasukan variabel penjelas

seperti sentimen, volume, dan tren. Dan saat dilakukan crossvalidation, model TFT mendapatkan score yang tidak jauh berbeda. Dengan nilai RMSE sebesar 0.031110, MAE sebesar 0.020376, dan MAPE sebesar 3.555097.

4.3.1 TFT Model Dasar

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



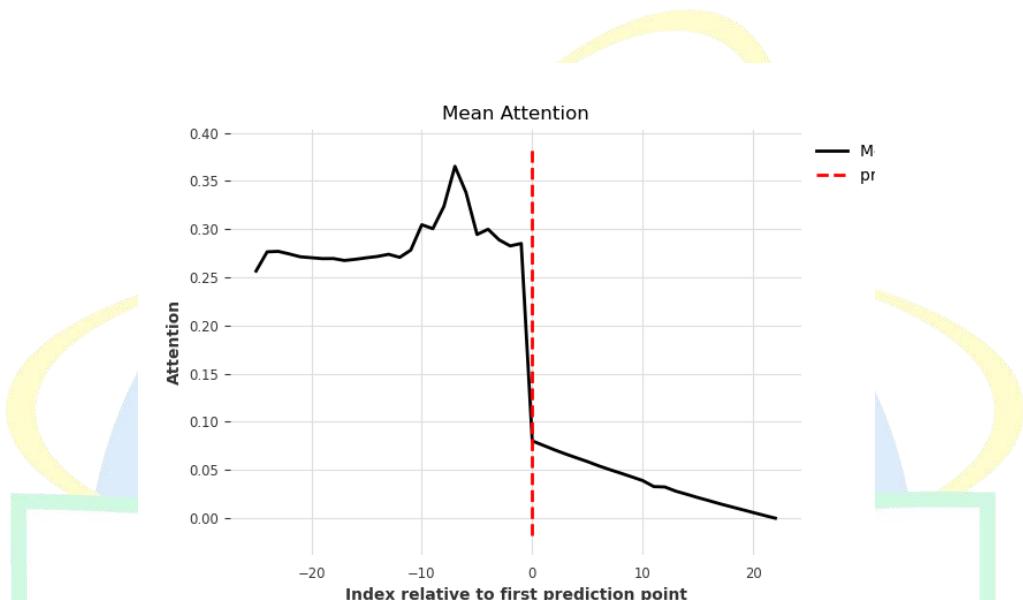
Gambar 4.9. Grafik Keterjelasan Encoder dan Decoder Model TFT Dasar

Tabel 4.5. Keterjelasan Encoder dan Decoder Model TFT Dasar

type	encoder	decoder
sen_futcov	0,02	0,08
volume_futcov	0,02	0,09
holidays_futcov	0,02	0,06
month_futcov	0,03	0,07
price_target	0,04	-
day_futcov	0,05	0,02
year_futcov	0,05	0,07
week_futcov	0,07	0,15
linear_increase_futcov	0,15	0,01

trend_futcov	0,56	0,45
--------------	------	------

Variabel variabel yang paling penting pada encoder adalah tren dan linear_increase dengan signifikansi lebih dari 10%, bahkan tren mencapai lebih dari 50%. Pada bagian decoder, trend dan week adalah variabel yang lebih penting dimana trend lebih dari 40% dan week lebih dari 10%.

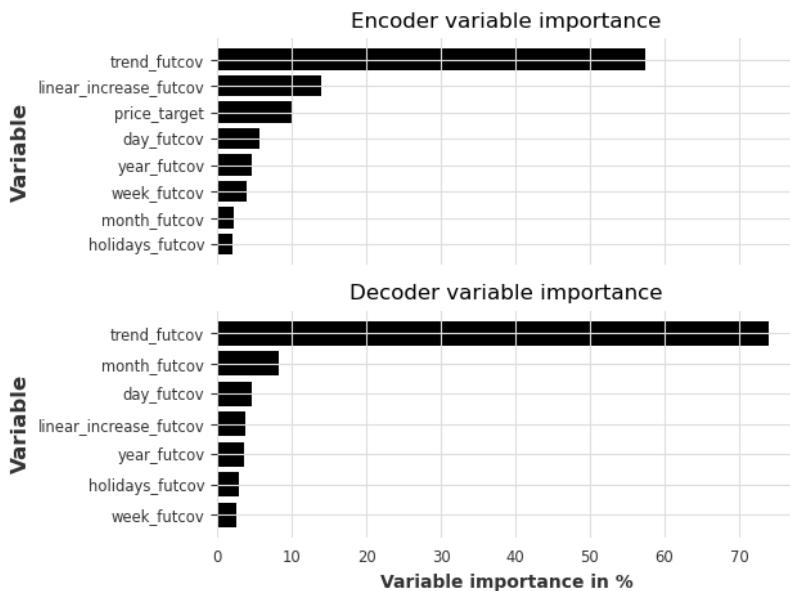


Gambar 4.10. Grafik Attention dari Model TFT Dasar

Perhatian dari model terhadap variabel penjelas dapat dilihat berdasarkan gambar 4.10. Perhatian naik pada saat training tepatnya di titik -7, sehingga dapat disimpulkan bahwa model menggunakan perhatian yang besar pada 7 titik sebelum peramalan dilakukan untuk melakukan peramalan.

4.3.2 Model TFT tanpa Volume dan Sentimen

Ditemukan bahwa dalam sisi encoder terdapat volume dan sentimen yang memiliki kepentingan paling kecil dalam memprediksi harga. Maka variabel volume dan sentimen dihapuskan lalu dilakukan *hyperparameter tuning* kembali untuk mendapatkan model yang terbaik (model TFT 2). Dan didapatkan model yang lebih baik dengan hasil RMSE sebesar 0.025863, MAE sebesar 0.017389, dan MAPE sebesar 3.033683.

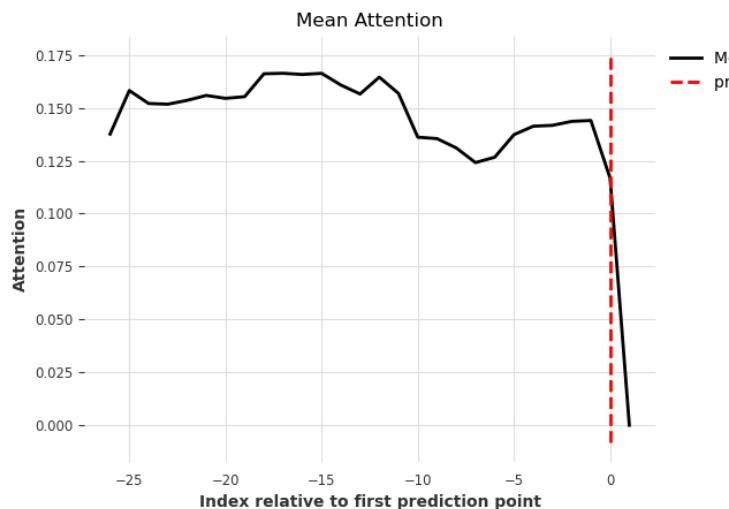


Gambar 4.11. Grafik Keterjelasan Encoder dan Decoder Model TFT 2

Tabel 4.6. Keterjelasan Encoder dan Decoder Model TFT 2

type	encoder	decoder
holidays_futcov	0,02	0,03
month_futcov	0,02	0,08
week_futcov	0,04	0,03
year_futcov	0,05	0,04
day_futcov	0,06	0,05
price_target	0,10	-
linear_increase_futcov	0,14	0,04
trend_futcov	0,58	0,74
holidays_futcov	0,02	0,03
month_futcov	0,02	0,08

Sama halnya seperti model dasar, variabel variabel yang paling penting pada encoder adalah tren dan linear_increase. Akan tetapi, kali ini ada harga itu sendiri yang memiliki tingkat kepentingan 10%. Lalu pada bagian decoder, didominasi oleh tren dengan kepentingan sebesar 74%.



Gambar 4.12. Grafik Attention dari Model TFT 2

Perhatian dari model hampir selalu konstan, akan tetapi mengalami penurunan dari titik -12 hingga -7 yang mana mengartikan pada titik titik tersebut tidak penting untuk memprediksi harga kedepannya. Dan mode mulai naik kembali yang mana mengartikan titik setelahnya mengalami peningkatan kepentingan.

Dikarenakan adanya kemiripan dengan data sinyal atau frekuensi, eksperimen dilakukan dengan menganggap data sentimen serupa. Dengan melakukan *smoothing* atau penghalusan data menggunakan STL. Didapatkan hasil evaluasi metrik yang tidak jauh beda dengan model TFT dasar, yaitu dengan nilai evaluasi RMSE 0.0259, MAE 0.0174, dan MAPE 3.0482.

BAB V

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 melatih model 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 membutuhkan peningkatan akurasi.

Pengintegrasian sentimen, volume, dan trend kedalam model prediksi harga bitcoin mendapatkan peningkatan hasil dibandingkan hanya menintegrasikan variabel penjelas, dari RMSE sebesar 0,0263 menjadi 0,0261. Dan juga model TFT lebih unggul dibandingkan dengan model *deep learning* lainnya (TCN, GRU, dan LSTM)

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.

BAB VI

DAFTAR PUSTAKA

- [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-Nefaei 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. Purnamaharap, 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] A. Arratia and A. X. López-Barrantes, “Do Google Trends forecast bitcoins? Stylized facts and statistical evidence,” *J. Bank. Financ. Technol.*, pp. 1–12, 2021, doi: 10.1007/s42786-021-00027-4.
- [7] 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.
- [8] 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.
- [9] S. Mehtab and J. Sen, “Stock Price Prediction Using Convolutional Neural Networks on a Multivariate Timeseries,” 2020, doi: 10.36227/techrxiv.15088734.v1.
- [10] S. Hansun, A. Wicaksana, and A. Q. M. Khalil, “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.
- [11] 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.
 - [12] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.
 - [13] 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.
 - [14] 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.
 - [15] 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.
 - [16] V. S and J. R, “Text Mining: open Source Tokenization Tools – An Analysis,” *Adv. Comput. Intell. An Int. J.*, vol. 3, no. 1, pp. 37–47, 2016, doi: 10.5121/acii.2016.3104.
 - [17] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.
 - [18] C. J. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-based Model for,” *Eighth Int. AAAI Conf. Weblogs Soc. Media*, pp. 216–225, 2014.
 - [19] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, “Sentiment Analysis of Review Datasets Using Naïve Bayes‘ and K-NN Classifier,” *Int. J. Inf. Eng. Electron. Bus.*, vol. 8, no. 4, pp. 54–62, 2016, doi: 10.5815/ijieeb.2016.04.07.

- [20] J. Wang, L. C. Yu, K. R. Lai, and X. Zhang, “Dimensional sentiment analysis using a regional CNN-LSTM model,” *54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Short Pap.*, pp. 225–230, 2016, doi: 10.18653/v1/p16-2037.
- [21] A. Avvaru, S. Vobilisetty, and R. Mamidi, “Detecting Sarcasm in Conversation Context Using Transformer-Based Models,” no. 2017, pp. 98–103, 2020, doi: 10.18653/v1/2020.figlang-1.15.
- [22] 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>
- [23] 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.
- [24] 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.
- [25] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Clim. Res.*, vol. 30, no. 1, pp. 79–82, 2005, doi: 10.3354/cr030079.
- [26] B. Zhao, “Encyclopedia of Big Data,” *Encycl. Big Data*, no. May 2017, 2020, doi: 10.1007/978-3-319-32001-4.
- [27] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” no. 1, 2019.
- [28] G. Lample and A. Conneau, “Cross-lingual Language Model Pretraining,” 2018.
- [29] B. Lim, S. Arik, 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.
- [30] 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.
- [31] D. Sarkar, *Text Analytics with Python*, vol. 32, no. 1. 2016. doi: 10.1140/epja/i2006-10279-1.
 - [32] R. Feldman and J. Sanger, *The Text Mining Handbook*. 2006. doi: 10.1017/cbo9780511546914.
 - [33] 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.
 - [34] 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.
 - [35] 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.
 - [36] 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.
 - [37] 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.
 - [38] 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.
 - [39] 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.
 - [40] M. Aiken, “An Updated Evaluation of Google Translate Accuracy,” *Stud. Linguist. Lit.*, vol. 3, no. 3, p. p253, 2019, doi: 10.22158/sll.v3n3p253.
 - [41] 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.
- [42] 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.
- [43] 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>
- [44] 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.
- [45] 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.
- [46] 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.
- [47] 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.