Prediksi Harga Dogecoin Berdasarkan Sentimen dari Twitter Menggunakan LSTM

1st Ecky Prasetyo Nugroho
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
eqivalen@student.telkomuniversity.ac.id

2nd Siti Sa'adah
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
sitisaadah@telkomuniversity.ac.id

3rd Farah Afianti
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
farahafi@telkomuniversity.ac.id

***Abstrak*— Dogecoin adalah mata uang kripto yang diciptakan oleh Billy Markus dan Jackson Palmer, tetapi mereka membuat Dogecoin hanya untuk dibuat sebagai bahan candaan di dunia mata uang kripto. Tugas akhir ini menganalisis sentimen dan prediksi terhadap Doge dengan melakukan korelasi antara harga Doge terhadap data yang dikumpulkan dari media social Twitter mengenai Doge. Penelitian ini dilakukan menggunakan pendapat-pendapat yang disampaikan oleh pengguna jejaring sosial yang menggunakan bahasa Inggris. Metode yang digunakan adalah LSTM dengan mengacu pada penelitian-penelitian sebelumnya yang menunjukkan bahwa LSTM memberikan akurasi tertinggi. Data yang digunakan pada penelitian ini adalah harga doge dan tweet pada periode januari-april 2021. Menentukan korelasi antara doge dan tweet dilakukan dengan korelasi pearson dimana hasil korelasi tersebut menentukan korelasi positif, korelasi negatif dan tidak berkorelasi, setelah itu dilakukan prediksi harga doge close dengan LSTM. Harga Doge Close berkorelasi dengan sentimen, namun tidak kuat tidak juga lemah. Tidak ada peningkatan akurasi hasil prediksi dibandingkan pengujian pertama yang dimana pada pengujian pertama nilai RMSE sebesar 0,003 dan pengujian kedua nilai RMSE sebesar 0,008.**

***Kata kunci*— analisis sentimen, LSTM, prediksi, korelasi**

I. PENDAHULUAN

Mata uang kripto adalah media pertukaran tanpa adanya koin atau catatan fisik. Dalam sistem ini transaksi antara mitra dagang dicatat secara online dan diautentikasi oleh pihak ketiga yang dikenal sebagai penambang [10]. Mata uang kripto pada dasarnya dipengaruhi oleh minat dan aksesibilitasnya pada pencarian. Misalnya, jika minat untuk membeli bitcoin di pasar kripto melampaui inventaris penjual, pada saat itu biaya bitcoin akan naik. Di sisi lain, ketika ada lebih banyak dealer bitcoin daripada pembeli, biaya bitcoin akan jatuh ke titik nilai di mana jumlah penjual dan pembeli setara.[11] Mata uang kripto merupakan produk yang paling berfluktuasi di pasar. Seperti gambaran pada periode 31 Agustus 2018 hingga 1 September 2019, volatilitas tiga besar cryptocurrency mencapai 71% untuk Bitcoin, 95% untuk Etherum dan 97% untuk XRP [1].

Pada tanggal 29 Januari 2021, Elon Musk, saat itu orang terkaya di dunia (Klebnikov, 2021), secara tidak terduga mengubah bio akun Twitternya menjadi #bitcoin. Harga Bitcoin naik dari sekitar $ 32.000 menjadi lebih dari $ 38.000 dalam hitungan jam, meningkatkan kapitalisasi pasarnya sebesar $ 111 miliar [2]. Mai dkk. (2018) menunjukkan bahwa pengguna media sosial dengan aktivitas yang relatif lebih rendah mendorong efek pada cryptocurrency, yang masuk akal: tindakan mereka tidak biasa atau tidak terduga [2]. Hal ini jadi pertimbangan bahwa sentimen dari media sosial khususnya Twitter memiliki pengaruh terhadap naik turunnya harga mata uang kripto. Dengan menemukan korelasi antara sentimen yang di Twitter, maka hasil prediksi dengan mengaitkan sentimen Twitter dengan harga mata uang kripto menggunakan metode LSTM akan bisa terlihat [12].

Long Short Term Memory network (LSTM) adalah salah satu jenis Recurrent Neural Network (RNN) yang dapat menangani long-term dependencies. LSTM dapat memperoleh prediksi deret waktu yang lebih baik karena sulit untuk mengadaptasi banyak metode linier klasik untuk masalah rediksi multi-input [3]. LSTM, menurut sifatnya, menggunakan karakteristik temporal dari sinyal deret waktu apapun; oleh karena itu, meramalkan deret waktu keuangan adalah implementasi LSTM yang dipelajari dengan baik dan berhasil [4]. Namun demikian, telah diamati bahwa pergerakan mata uang kripto sangat tidak stabil baru-baru ini dikarenakan pandemi Covid-19 yang menyebabkan resesi ekonomi di berbagai negara, dan juga ada variabel yang tidak terduga untuk mata uang kripto tertentu, seperti tweet dari Elon Musk yang membuat pasar mata uang kripto memiliki volatilitas yang tinggi. Dengan diketahuinya nilai prediksi dari mata uang kripto menggunakan Root Mean Square Error (RMSE) untuk mengevaluasi apakah semua yang terjadi saat ini sangat berdampak untuk stabilitas mata uang kripto berdasarkan nilai RMSE yang didapat nantinya.

Bukti sangat mendukung hipotesis bahwa tweet memang menyampaikan informasi yang relevan dengan return harga[4]. Data latih yang digunakan dalam penelitian ini adalah harga close Doge dan tweet dalam bahasa Inggris yang terkait dengan Doge pada periode Januari hingga April 2021. Korelasi antara sentimen Twitter dan harga Doge akan diukur menggunakan korelasi Pearson, sedangkan metode prediksi yang digunakan adalah Long Short-Term Memory (LSTM). Namun apakah ada korelasi antara Doge dan sentiment twitter yang bernilai nilai *compound* seperti pada penelitian sebelumnya [17][18] dan apakah hasil prediksi harga akan sebaik penelitian sebelumnya [1]. Hasil penelitian ini diharapkan dapat memberikan pemahaman yang lebih baik tentang pengaruh sentimen Twitter terhadap pergerakan harga Doge, serta menghasilkan metode prediksi yang lebih akurat dibandingkan dengan penelitian sebelumnya [13].

Tujuan penelitian yang dilakukan adalah untuk memprediksi dan mengetahui keterkaitan harga dari Doge dengan sentimen yang didapatkan dari media jejaring sosial twitter.

II. KAJIAN TEORI

A. Comparative study

Pada studi kasus [2] dikatakan bahwa terdapat dua tindakan Elon Musk yang menghasilkan peningkatan besar dalam volume perdagangan dan pengembalian abnormal positif yang besar dan signifikan. Tindakan pertama yaitu dukungan terhadap bitcoin dan yang kedua dukungan terhadap dogecoin walaupun disebutkan di [2] hanya sebuah candaan. Pada studi kasus [6] juga dikatakan dampak tweet terhadap likuiditas dalam waktu yang tepat efek positif tweet langsung terjadi dalam lima hingga 10 menit ke depan, dan berlangsung sekitar satu jam. Bukti empiris dari penelitian [6] menunjukkan bahwa perhatian investor aktif dapat secara signifikan meningkatkan likuiditas Bitcoin secara real time, yang meningkatkan efisiensi harga di pasar Bitcoin.

Pada studi kasus [1] dikatakan bahwa prediksi menggunakan Support Vector Machine (SVM) memiliki error yang lebih kecil dibandingkan K-Nearest neighbors (KNN), tetapi SVM memiliki waktu eksekusi yang lebih cepat dibandingkan LSTM, jadi SVM bisa digunakan untuk memonitor langsung stabilitas keuangan jika terdapat batasan dalam hal sumber daya, namun jika ingin mendapatkan error yang kecil disarankan menggunakan LSTM. Long Short-Term Memory adalah varian dari RNN yang dapat mempelajari dependensi jangka panjang. LSTM memiliki struktur yang mirip dengan RNN, tetapi unit berulang memiliki struktur yang relatif berbeda. Tidak seperti memiliki lapisan neural network tunggal, mereka memiliki empat lapisan yang berinteraksi satu sama lain. [5] Disebutkan juga dalam penelitian [7] bahwa RNN memiliki kesulitan dalam hal ketergantungan jangka panjang karena terdapat Vanishing Gradient yaitu bobot yang sebelumnya akan berkurang seiring waktu langkah sehingga hanya keluaran terbaru yang mempengaruhi prediksi.

B. Valence Aware Dictionary and sEntiment Reasoner (VADER)

VADER (Valence Aware Dictionary and sEntiment Reasoner) adalah alat analisis sentimen berbasis leksikon dan aturan yang secara khusus selaras dengan sentimen yang ada di media sosial. VADER menggunakan lebih sedikit sumber daya dibandingkan dengan model Pembelajaran Mesin karena tidak memerlukan data pelatihan dalam jumlah besar. Memeriksa lebih lanjut skor F1 (akurasi klasifikasi), VADER (0,96) mengungguli penilai manusia individu (0,84) dalam memberi label sentimen tweet dengan benar ke dalam kelas positif, netral, atau negatif. Alasan di balik ini adalah bahwa VADER peka terhadap Polaritas (apakah sentimennya positif atau negatif) [13]. VADER mengandalkan kamus yang memetakan kata-kata dan banyak fitur leksikal lainnya yang umum untuk sentimen.

Compound adalah nilai sentimen keseluruhan suatu teks yang dianalisis menggunakan algoritma VADER (Valence Aware Dictionary and sEntiment Reasoner). Nilai *Compound* dinyatakan dalam rentang -1 hingga 1, di mana nilai positif menunjukkan sentimen positif, nilai negatif menunjukkan sentimen negatif, dan nilai 0 menunjukkan netral. *Compound* menggabungkan informasi dari nilai sentimen positif, negatif, dan netral dalam satu angka tunggal yang menggambarkan sentimen keseluruhan dari teks yang dianalisis. Nilai *Compound* sangat berguna dalam analisis sentimen teks karena memberikan informasi yang lebih kaya dan lebih mudah diinterpretasikan daripada hanya menghitung jumlah kata positif dan negatif dalam sebuah teks.

Skor *compound* dihitung dengan menjumlahkan skor valensi setiap kata dalam leksikon, disesuaikan dengan aturan, dan kemudian dinormalisasi menjadi antara -1 (negatif paling ekstrim) dan +1 (positif paling ekstrim). Contoh skor valensi dari beberapa teks:

Positif: "okay" adalah 0,9 "good" adalah 1,9, dan "great" adalah 3,1

Negatif: "horrible" adalah –2,5, dan "sucks" dan "sux" keduanya –1,5

Skor *compound* dapat dihitung menggunakan rumus :

$$x = \frac{x}{\sqrt{x^2 + \alpha}} \qquad (1)$$

x = jumlah skor kata penyusun

α = Konstanta normalisasi (nilai default adalah 15)

TABEL 1.
Contoh Perhitungan Compound score VADER

Teks	Jumlah Skor Valensi (x)	Skor Compound
okay its good	0,9+0+1,9 = 2,8	0,58
okay its horrible	0,9+0+(-2,5) = -1,6	-0,38

C. Correlation Pearson

Correlation Pearson adalah suatu rumus yang digunakan untuk mencari hubungan antara variabel bebas dan variabel terikat [16]. Misalnya dua variabel x dan y yang dimaksud dengan korelasi positif/negatif tidak berkorelasi. Ini mengacu pada aturan berikut berikut.

Korelasi positif : variabel x Ketika tumbuh, variabel y Juga, keadaan meningkat dengan saling mengunci dalam arah yang sama

Korelasi negatif : variabel x Ketika tumbuh, variabel y sebaliknya Keadaan terhubung arah dan berkurang

Tidak ada korelasi : variabel x Bahkan jika tumbuh, variabel y Adalah keadaan yang tidak berubah saat dikaitkan dengannya

Dalam koefisien korelasi Standarisasi jenis normalisasi skala dengan perhitungan di atas 1～0～-1

TABEL 2.
Rentang koefisien korelasi dan arah korelasi

Koefisien korelasi	Arah korelasi
0.7 ～ 1.0	Korelasi positif yang kuat
0.4 ～ 0.7	Korelasi positif

$0.2 \sim 0.4$	Korelasi positif lemah
$-0.2 \sim 0.2$	tidak ada korelasi
$-0.4 \sim -0.2$	Korelasi negatif lemah
$-0.7 \sim -0.4$	Korelasi negatif
$-1.0 \sim -0.7$	Korelasi negatif yang kuat

Rumus koefisien korelasi pearson antara variabel x dan y adalah

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \quad (2)$$

dimana r_{xy} adalah koefisien korelasi

Koefisien korelasi dapat bernilai positif maupun negatif hal ini seperti yang ada pada tabel yang dimana setiap nilai dari koefisien korelasi tertentu memiliki arah korelasi yang berbeda.

TABEL 3.
contoh perhitungan koefisien korelasi pearson

x	y	xy	x^2	y^2
1	2	2	1	4
5	4	20	25	16
4	6	24	16	36
2	4	8	4	16
3	2	6	9	4
15	18	60	55	75

Dari tabel 3 dapat diketahui nilai-nilai dari komponen pada rumus koefisien korelasi Pearson sebagai berikut:

$$n = 5$$
$$\sum_{i=1}^{n} x_i = 15$$
$$\sum_{i=1}^{n} y_i = 18$$
$$\sum_{i=1}^{n} x_i y_i = 60$$
$$\sum_{i=1}^{n} x_i^2 = 55$$
$$\sum_{i=1}^{n} y_i^2 = 76$$

Kemudian nilai-nilai tersebut dimasukan ke rumus koefisien korelasi pearson, sehingga dapat dihitung nilai dari koefisien korelasi Pearson.

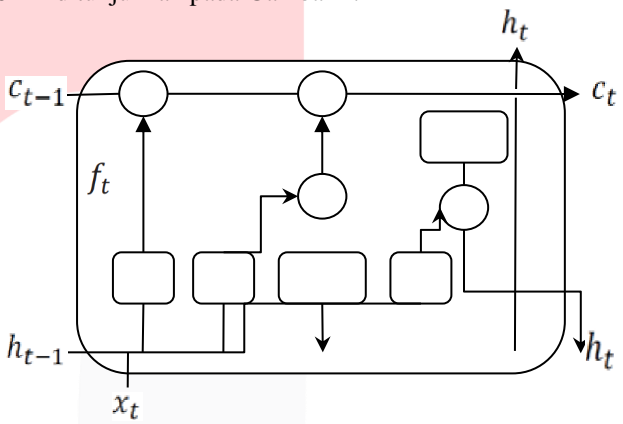
$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r_{xy} = \frac{(5)(60) - (15)(18)}{\sqrt{(5)(55) - (15)^2}\sqrt{(5)(76) - (18)^2}}$$
$$r_{xy} = \frac{300 - 270}{(7{,}07)(7{,}48)}$$
$$r_{xy} = 0{,}57$$

Nilai koefisien korelasi Pearson adalah 0,57 yang artinya terdapat hubungan korelasi positif antara variabel x dan variabel y

D. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) adalah jenis RNN yang yang mampu menyelesaikan masalah ketergantungan jangka panjang [6]. Secara umum, sel LSTM terdiri dari tiga proses utama yaitu Forget, Update, dan Output. Model sel LSTM ditunjukkan pada Gambar 1.



GAMBAR 1.
Struktur Unit LSTM

$w_f = Weight$
$h_{t-1} = Output\ from\ the\ previous\ time\ stamp$
$x_t = New\ Input$
$b_f = Bias$

Langkah pertama dalam LSTM adalah mengidentifikasi informasi yang tidak diperlukan dan akan dibuang dari status sel. Keputusan ini dibuat oleh lapisan sigmoid yang disebut lapisan forget.

$$f_t = \sigma(w_f[h_{t-1}, X_t] + b_f) \quad (3)$$

Langkah selanjutnya adalah memutuskan, informasi baru apa yang akan kita simpan dalam status sel. Keseluruhan proses ini terdiri dari langkah-langkah berikut. Lapisan sigmoid yang disebut "lapisan input" memutuskan nilai mana yang akan diperbarui. Selanjutnya, lapisan tanh membuat vektor nilai kandidat baru, yang dapat ditambahkan ke status.

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (4)$$
$$C\sim_t = tanh(w_c[h_{t-1}, x_t] + b_c) \quad (5)$$

Sekarang, status sel lama akan diperbarui, C_{t-1}, ke status sel baru C_t. Pertama kalikan keadaan lama (C_{t-1}) dengan f_t, Melupakan hal-hal yang sudah diputuskan untuk dilupakan sebelumnya. Kemudian, menambahkan $i_t * C\sim_t$. Ini adalah nilai kandidat baru, yang diskalakan dengan seberapa banyak memutuskan untuk memperbarui setiap nilai status sel.

$$C_t = f_t * C_{t-1} + i_t * C\sim_t \qquad (6)$$

Terakhir akan menjalankan lapisan sigmoid yang memutuskan bagian mana dari status sel yang akan dikeluaran. Kemudian, menempatkan status sel (mendorong nilai menjadi antara -1 dan 1) dan mengalikannya dengan output dari gerbang sigmoid, sehingga hanya menampilkan bagian yang diputuskan.

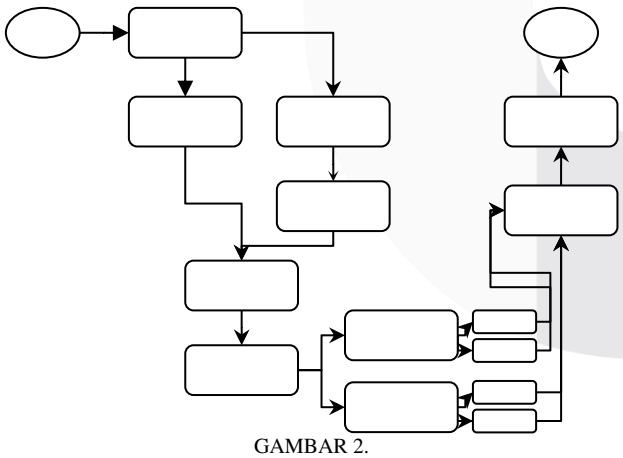
E. Root Mean Square Error (RMSE)

Root mean square error (RMSE) telah digunakan sebagai metrik statistik standar untuk mengukur kinerja model dalam studi meteorologi, kualitas udara, dan penelitian iklim. Mean absolute error (MAE) adalah pengukur lainnya yang banyak dipakai untuk evaluasi model. Kedua model tersebut telah digunakan selama bertahun-tahun, tidak ada konsensus tentang metrik yang paling sesuai untuk kesalahan model. [8] Persamaan umum RMSE yaitu :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)^2} \qquad (7)$$

Aspek penting dari metrik kesalahan yang digunakan untuk evaluasi model adalah kemampuannya untuk membedakan hasil model. Ukuran yang benar-benar diskriminatif yang menghasilkan variasi yang lebih tinggi dalam metrik kinerja modelnya di antara berbagai pengaturan hasil model seringkali lebih diinginkan. Dalam hal ini, MAE dapat dipengaruhi oleh sejumlah besar nilai kesalahan rata-rata tanpa cukup mencerminkan beberapa kesalahan besar. Memberikan bobot yang lebih tinggi pada kondisi yang tidak menguntungkan, RMSE biasanya lebih baik dalam mengungkapkan kontras kinerja model.

III. METODE



GAMBAR 2.
Flowchart Perancangan Sistem

A. Mengumpulkan data mengenai tweet doge dan harga doge/usd

Data input yang digunakan adalah data historis dari Data yang dikumpulkan berupa data harga dari DOGE/USD dari tanggal 1 Januari 2021 sampai dengan tanggal 30 April 2021 yang didapatkan dari situs Yahoo Finance, dan Tweet dengan hastag #doge atau #dogecoin dari tanggal 1 Januari 2021 sampai dengan tanggal 30 April 2021. Dalam dataset terdapat berbagai data yang dapat dilihat pada tabel 4 sebagai sampel dari dataset Doge, dan tabel 5 sebagai sampel dari dataset Tweet Doge.

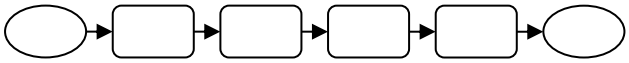
TABEL 4.
Contoh Sampel Data Doge

Date	Open	High	Low	Close	Adj Close	Volume
2021-01-01	4.681	5.685	4.615	5.685	5.685	228961515
2021-01-02	5.686	13.698	5.584	10.615	10.615	3421562680

TABEL 5.
Contoh Sampel Data Tweet Doge

created_at	tweet	likes_count
2021-01-01 06:34:14+00:00	$DOGE on the rise. You know what that means 😉	104.0
2021-04-28 23:50:40+00:00	$doge up https://t.co/kHqFhSZxG4	4785.0

B. Preprocessing Data



GAMBAR 3.
Flowchart Preprocessing Data

1. Data Selection

Tahapan yang dilakukan pada preprocessing data yang pertama adalah memfilter data yang tidak digunakan serta mendapatkan data yang memiliki like lebih dari 100 dengan cara mengambil data tweet doge yang likes_count lebih dari 100, hal ini perlu dilakukan untuk mengurangi tweet doge yang tidak populer. Untuk perbandingannya dapat dilihat sebagai berikut:

```
RangeIndex: 3670033 entries, 0 to 3670032
Data columns (total 34 columns):
 #  Column          Dtype
---  ------          -----
 0  Unnamed: 0      object
 1  id              object
 2  conversation_id  object
 3  created_at      object
 4  user_id         object
 5  username        object
 6  name            object
 7  place           object
 8  tweet           object
 9  language        object
 10 mentions        object
 11 urls            object
 12 photos          object
 13 replies_count   float64
 14 retweets_count  object
 15 likes_count     float64
 16 hashtags        object
 17 cashtags        object
 18 link            object
 19 retweet         object
 20 quote_url       object
 21 video           object
 22 thumbnail       object
 23 near            float64
 24 geo             float64
 25 source          float64
 26 user_rt_id      float64
 27 user_rt         float64
 28 retweet_id      float64
 29 reply_to        object
 30 retweet_date    float64
 31 translate       float64
 32 trans_src       float64
 33 trans_dest      float64
dtypes: float64(12), object(22)
memory usage: 952.0+ MB
```

GAMBAR 4.
Isi dari data sebelum pemilihan data

```
Int64Index: 25076 entries, 10 to 3670017
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   created_at  25076 non-null  object
 1   tweet       25076 non-null  object
dtypes: object(2)
memory usage: 587.7+ KB
```

GAMBAR 5.
Isi dari data setelah pemilihan data

Dilihat dari gambar 4 dan gambar 5 data yang sebelumnya berjumlah 3670032 menjadi 25076 setelah pemilihan data yang jumlah likenya lebih dari 100, serta dua kolom yang nantinya akan digunakan. Untuk kolom created_at berisi data tanggal dari tweet, sedangkan kolom tweet berisi teks dari tweet tersebut.

2. Noise Cleaner

Tahapan kedua yang dilakukan di preprocessing data adalah membersihkan teks dari noise, seperti emoticon, link, dan simbol-simbol. Hal ini perlu dilakukan untuk meningkatkan akurasi dari penghitungan sentimen. Fokus dalam tahap ini adalah menghapus emoticon, link, dan simbol yang ada dalam teks dengan contoh sebagai berikut

TABEL 6.
Contoh sampel data sebelum dan sesudah dibersihkan

Sebelum	Setelah
#DOGE Move Right Now 🐕✅🚀 Follow The best Channel	DOGE Move Right Now Follow The best Channel
$DOGE on the rise. You know what that means 😉	DOGE on the rise You know what that means

3. VADER sentiment score

Tahapan ketiga yang dilakukan di preprocessing data adalah penghitungan score sentiment menggunakan VADER. Data dari tweet doge yang didapatkan dari tahapan sebelumnya akan dilakukan kalkulasi untuk menentukan nilai dari sentimen dari tweet tersebut.

TABEL 7.
Contoh sampel data yang sudah dihitung score sentimen

Tweet	Compound
for every like I get I will buy 100 worth of doge	0.5267
I think youve been suffering from depression	-0.7783

Pada tahap penghitungan score sentimen menggunakan VADER untuk data tweet doge, nilai *compound* yang dihasilkan akan menjadi nilai sentimen yang dipakai. Nilai *compound* menggambarkan sentimen keseluruhan dari teks yang dianalisis, sehingga dapat memberikan gambaran yang lebih jelas mengenai sentimen dari kumpulan tweet doge yang dianalisis.

4. Resample Data

Tahapan keempat yang dilakukan di preprocessing data adalah resample data, kenapa hal ini diperlukan ?. Tahapan ini diperlukan untuk menyesuaikan jumlah data yang ada di harga Doge, dikarenakan data yang ada di harga Doge berdasarkan harian maka data yang ada di tweet harus mengikuti. Data yang ada di tweet Doge didapatkan 24 jam yang dimana perdetik permenit maupun perjamnya ada data yang baru. Dalam tahap ini akan dilakukan resample data tweet Doge berdasarkan hari. Terdapat dua jenis data yang di resample, *Sum score* atau jumlah skor merupakan penjumlahan skor sentiment dari setiap tweet pada suatu periode waktu tertentu. Sedangkan *mean score* atau rata-rata skor merupakan hasil dari penghitungan rata-rata skor sentiment dari setiap tweet pada suatu periode waktu tertentu. Dengan demikian, perbedaan antara *sum score* dan *mean score* dapat mempengaruhi hasil analisis sentimen pada data tweet Doge. Data yang dihitung adalah data *compound* yang dapat dilihat pada tabel 7.

Dalam tahap resample, data akan disesuaikan dengan rentang waktu yang digunakan dalam analisis harga Doge. Hal ini bertujuan agar informasi yang terkandung dalam data tweet dapat dipetakan dengan lebih akurat pada pergerakan harga Doge. Dalam kasus ini, data tweet Doge per jam akan diubah menjadi data tweet Doge per hari dengan dua jenis resample yaitu resample dengan jumlah dan resample dengan rata-rata, sesuai dengan preferensi penggunaan data. Dengan resample ini, data tweet Doge akan menjadi lebih mudah dibaca dan dianalisis secara bersamaan dengan data harga Doge.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 118 entries, 0 to 117
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Date        118 non-null    datetime64[ns]
 1   mean score  118 non-null    float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 2.0 KB
```

GAMBAR 6.
Isi dari data compound setelah di resample menggunakan rata-rata

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 118 entries, 0 to 117
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       118 non-null    datetime64[ns]
 1   sum score  118 non-null    float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 2.0 KB
```

GAMBAR 7.
Isi dari data compound setelah di resample menggunakan jumlah

Jika dilihat dari gambar 5 jumlah data sebelum di resample adalah 25076 data, setelah di resample berdasarkan hari dapat dilihat dari gambar 6 dan gambar 7 jumlah data sesudah di resample adalah 118 data.

C. Merge Data

Merge data bertujuan untuk menggabungkan data dari harga Doge dan tweet Doge yang sudah melalui preprocessing data, hal ini perlu dilakukan untuk mengetahui korelasi diantara kedua data tersebut. Terdapat dua jenis data yang di merge, yang pertama adalah resample dari tweet Doge yang dijumlah total, yang kedua adalah resample dari tweet Doge yang dirata-rata.

TABEL 6.
Contoh sampel merge data berdasarkan rata-rata score sentimen

Date	Open	High	Low	Close	Volume	mean score
2021-01-01	0.004681	0.005685	0.004615	0.005685	228961515	0.142970
2021-01-05	0.009767	0.010219	0.008972	0.009920	687256067	0.072770

TABEL 7.
Contoh sampel merge data berdasarkan jumlah score sentimen

Date	Open	High	Low	Close	Volume	sum score
2021-01-01	0.004681	0.005685	0.004615	0.005685	228961515	1.4297
2021-01-05	0.009767	0.010219	0.008972	0.009920	687256067	0.7277

D. Correlation

Setelah data dari harga Doge dan tweet Doge yang skornya sudah dihitung menggunakan VADER dan di resample menggunakan sum dan mean digabungkan, selanjutnya data yang sudah digabungkan di korelasikan untuk mengetahui keterkaitan harga Doge dan sentimen tweet Doge yang didapatkan dengan menggunakan korelasi pearson. Dengan contoh sebagai berikut.

GAMBAR 8.
Data Merge harga Doge dengan skor resample rata-rata sentimen

Dilihat dari gambar 8 dapat disimpulkan bahwa Open, High, Low, Close, Volume dari doge dan *mean score* tidak berkorelasi. Hal ini dapat dilihat dari nilai tertinggi dari korelasi antara harga doge dan skor resample yang di rata-rata begitu rendah untuk nilai tertingginya dengan nilai 0,084 , sedangkan untuk nilai terendahnya -0,2. Dilihat dari tabel 2 nilai tersebut masuk ke kategori tidak ada korelasi.

GAMBAR 9.
Korelasi data merge harga Doge dengan skor resample jumlah sentimen

Dilihat dari gambar 9 dapat disimpulkan bahwa Open, High, Low, Close, Volume dari doge dan *sum score* berkorelasi dengan harga doge, namun korelasi tertinggi ada pada volume transaksi bukan pada harga. Korelasi tertinggi antara doge dan *sum score* yaitu volume transaksi dengan *sum score* dengan nilai 0,92.Dilihat dari tabel 2 nilai tersebut masuk kategori korelasi positif yang kuat. Namun untuk doge sendiri mulai dari Open, High, Low, Close nilai korelasinya kurang dari 0,8.Dilihat dari tabel 2 nilai yang kurang dari 0,7 masuk ke kategori korelasi positif, namun nilainya tidak kuat. Volume transaksi berkorelasi dengan sentimen karena volume transaksi mencerminkan tingkat permintaan dan penawaran instrumen yang dalam hal ini adalah mata uang kripto. Sentimen pada gilirannya mengukur kepercayaan dan keyakinan pelaku pasar terhadap prospek suatu instrumen finansial atau pasar secara keseluruhan, namun hal ini tidak mempengaruhi harga secara absolut [14][15]. ini juga dapat dilihat dari korelasi volume transaksi dengan harga Doge open, high, low, dan close yang kurang dari 0,8.

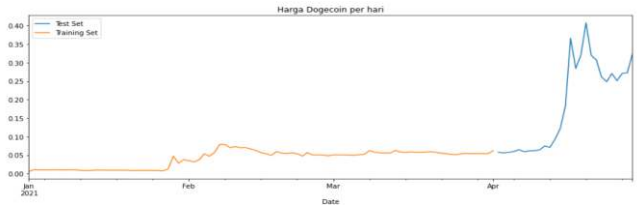
Pada penelitian ini berfokus pada prediksi harga, dilihat dari hasil korelasi antara harga Doge dan *sum score* berkorelasi maka doge bisa diprediksi dengan *sum score*.

E. Pengujian

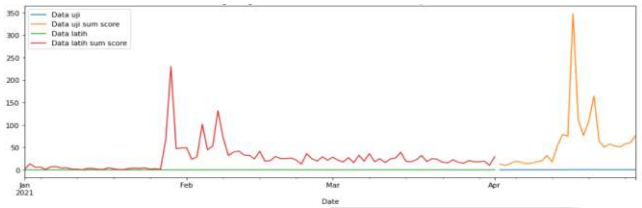
Pada tahap ini pengujian dibagi menjadi dua yang dimana untuk pengujian pertama data yang digunakan hanya satu data yaitu doge Close tanpa ada data tambahan dari sentimen twitter, sedangkan untuk pengujian kedua data yang digunakan menggunakan dua data yaitu doge Close dan *sum score* dari sentimen twitter.

F. Data latih dan data uji

Pada tahap ini dilakukan pembagian data harga close doge untuk pengujian pertama, sedangkan untuk pengujian kedua dilakukan pembagian data harga close doge dan *sum score* untuk melakukan pengujian terhadap algoritma LSTM. Dimana data dibagi menjadi data latih dan data uji. Pembagian data latih dan data uji dilakukan berdasarkan tanggal yang dimana data latih menggunakan data terhitung dari tanggal 01-01-2021 sampai dengan tanggal 01-04-2021, sedangkan data uji terhitung dari tanggal 02-04-2021 sampai dengan tanggal 28-04-2021.

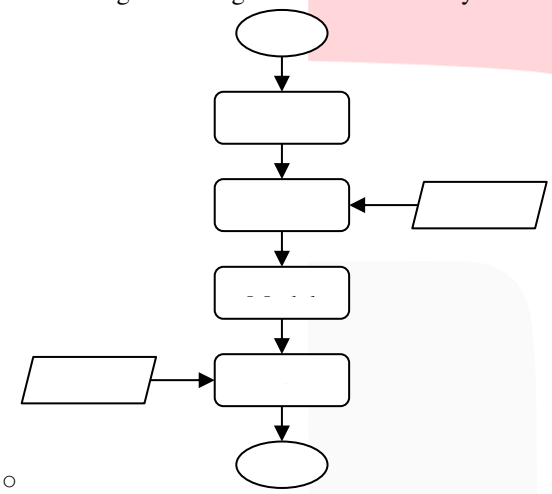


GAMBAR 10.
Visualisasi data doge close pada pengujian pertama



GAMBAR 11.
Visualisasi data doge close dan sum score pada pengujian kedua

G. Pembangunan Long Short-Term Memory



GAMBAR 12.
Flowchart Pembangunan Long Short-Term Memory

Pada gambar 12 menunjukan flowchart dari pembangunan LSTM dimulai dari pendefinisian parameter LSTM, proses learning yang menggunakan data latih, kemudian pemodelan menggunakan LSTM, setelah itu evaluasi menggunakan data uji dan di analisis menggunakan RMSE.

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

Pada tahap pengujian pertama dimana seluruh parameter dari LSTM diatur ke nilai standarnya didapatkan bahwa performa dari LSTM diukur dengan evaluasi pemisahan data latih dan data uji dengan data latih dari tanggal 01-01-2021 sampai dengan tanggal 01-04-2021 dan data uji dari tanggal 02-04-2021 sampai dengan tanggal 28-04-2021.Pengujian dilakukan dua kali, pengujian pertama hanya menggunakan data doge Close, sedangkan pengujian kedua menggunakan doge Close dan *sum score*. Hasil pengujian dapat dilihat sebagai berikut :

TABEL 9.
Hasil pengujian pertama

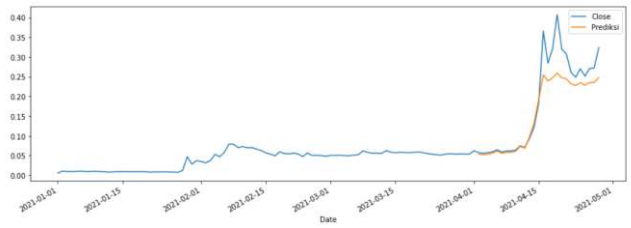
Date	Close Price	Prediction
02-04-2021	0,05766	0.05604
03-04-2021	0,05580	0.05399
04-04-2021	0,05740	0.05575
05-04-2021	0,05969	0.05831
06-04-2021	0,06445	0.06378
07-04-2021	0,05902	0.05756
08-04-2021	0,06146	0.06032
09-04-2021	0,06168	0.06057
10-04-2021	0,06384	0.06306
11-04-2021	0,07464	0.07601
12-04-2021	0,07076	0.07128
13-04-2021	0,09344	0.09956
14-04-2021	0,12151	0.13332
15-04-2021	0,18220	0.18669
16-04-2021	0,03658	0.23992
17-04-2021	0,28417	0.22703
18-04-2021	0,32047	0.23393
19-04-2021	0,40731	0.24374
20-04-2021	0,00319	0.23377
21-04-2021	0,30692	0.23163
22-04-2021	0,26096	0.22118
23-04-2021	0,24850	0.21743
24-04-2021	0,27021	0.22368
25-04-2021	0,25111	0.21826
26-04-2021	0,27067	0.22379
27-04-2021	0,27218	0.22418
28-04-2021	0,32368	0.23444
RMSE	0,003	

TABEL 10.
Hasil pengujian kedua

Date	Close Price	Prediction
02-04-2021	0,05766	0.05443
03-04-2021	0,05580	0.05229
04-04-2021	0,05740	0.05422
05-04-2021	0,05969	0.05689
06-04-2021	0,06445	0.06149
07-04-2021	0,05902	0.05581
08-04-2021	0,06146	0.05832
09-04-2021	0,06168	0.05878
10-04-2021	0,06384	0.06108
11-04-2021	0,07464	0.07259
12-04-2021	0,07076	0.06780
13-04-2021	0,09344	0.09128
14-04-2021	0,12151	0.11387
15-04-2021	0,18220	0.14481
16-04-2021	0,03658	0.18428
17-04-2021	0,28417	0.17231
18-04-2021	0,32047	0.17791
19-04-2021	0,40731	0.18749
20-04-2021	0,00319	0.17815
21-04-2021	0,30692	0.17584
22-04-2021	0,26096	0.16737
23-04-2021	0,24850	0.16465
24-04-2021	0,27021	0.16930
25-04-2021	0,25111	0.16519
26-04-2021	0,27067	0.16943
27-04-2021	0,27218	0.16974
28-04-2021	0,32368	0.17836
RMSE	0,008	

Dilihat dari tabel 9 dan tabel 10 Hasil menunjukkan bahwa pengujian pertama memiliki performa yang baik dibandingkan pengujian kedua dengan nilai RMSE 0,003 untuk pengujian pertama dan nilai RMSE 0,008 untuk pengujian kedua.

GAMBAR 13.
Visualisasi pengujian pertama



GAMBAR 14.
Visualisasi pengujian kedua

B. Analisis Hasil Pengujian

Berdasarkan penelitian yang dilakukan, pengujian pertama dan pengujian kedua memiliki sedikit perbedaan hyperparameter di input_shape. Pada pengujian pertama menggunakan input_shape=(1,1) yang menentukan bentuk input data pada model sebagai 1 baris dan 1 kolom. Ini menunjukkan bahwa setiap satu input yang diterima oleh model memiliki 1 nilai yang berisi yang berisi data doge Close , sedangkan pengujian kedua menggunakan input_shape=(2,1) yang menentukan bentuk input data pada model sebagai 2 baris dan 1 kolom. Ini menunjukkan bahwa setiap satu input yang diterima oleh model memiliki 2 nilai yang berisi data doge Close dan *sum score*. Selain itu, semua hyperparameter yang digunakan pengujian pertama dan pengujian kedua sama. Hyperparameter pengujian pertama dan pengujian kedua berupa tiga layer LSTM dengan jumlah unit masing-masing 64, 32, dan 16. Fungsi aktivasi yang digunakan pada ketiga layer LSTM adalah "tanh". Selain itu, terdapat tiga layer dropout dengan nilai 0.2 pada setiap layer LSTM. Model juga memiliki satu layer dense dengan fungsi aktivasi "relu". Loss function yang digunakan masih menggunakan mean squared error dan optimizer yang digunakan adalah Adam. Model dilatih dengan 100 epochs dan batch size 50. Dengan menggunakan lebih dari satu layer LSTM, model ini dapat menangkap lebih banyak informasi dari data dengan volatilitas yang tinggi. Penggunaan fungsi aktivasi "tanh" pada layer LSTM dapat mengurangi efek vanishing gradient pada model, sehingga model dapat lebih efektif belajar dari data. Dropout layer juga ditambahkan pada setiap layer LSTM untuk mengurangi overfitting. Pada tabel 9 dan tabel 10 dapat dilihat bahwa RMSE dari pengujian pertama sebesar 0.003 dan pengujian kedua sebesar 0.008 yang mana pada pengujian kedua tidak terdapat peningkatan pada akurasi hasil prediksi.

V. KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan maka didapatkan kesimpulan sebagai berikut :

Doge Close dan *sum score* berkorelasi, namun tidak kuat tidak juga lemah.

Pada pengujian kedua tidak terdapat peningkatan pada akurasi hasil prediksi dibandingkan pengujian pertama yang dimana pada pengujian pertama nilai RMSE sebesar 0,003 dan pengujian kedua nilai RMSE sebesar 0,008.

Dapat disimpulkan bahwa bahwa sentimen kurang berpengaruh pada peningkatan akurasi hasil prediksi, namun harus diperhatikan bahwa hal ini mungkin saja tergantung pada beberapa faktor lain seperti representasi data sentimen, jumlah data yang tersedia, dan metode pengolahan data sentimen yang digunakan. Oleh karena itu, perlu dilakukan analisis yang lebih mendalam untuk memastikan bahwa sentimen memang memiliki pengaruh yang signifikan pada peningkatan akurasi hasil prediksi.

Untuk penelitian yang akan datang, model yang dibuat akan lebih baik jika data prediksi yang digunakan menggunakan lebih banyak dan menggunakan tweet yang berbahasa Indonesia.

REFERENSI

[1] Saadah, Siti, and AA Ahmad Whafa. "Monitoring Financial Stability Based on Prediction of Cryptocurrencies Price Using Intelligent Algorithm." 2020 International Conference on Data Science and Its Applications (ICoDSA). IEEE, 2020.

[2] Ante, Lennart. "How Elon Musk's Twitter Activity Moves Cryptocurrency Markets." Available at SSRN 3778844 (2021).

[3] Chen, Wei, et al. "Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants." International Journal of Forecasting 37.1 (2021): 28-43.

[4] Sezer, Omer Berat, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019." Applied Soft Computing 90 (2020): 106181.

[5] Patel, Mohil Maheshkumar, et al. "A deep learning-based cryptocurrency price prediction scheme for financial institutions." *Journal of Information Security and Applications* 55 (2020): 102583.

[6] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[7] Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult." IEEE transactions on neural networks 5.2 (1994): 157-166.

[8] Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature." Geoscientific model development 7.3 (2014): 1247-1250.

[9] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," in 3rd International Conference for Learning Representations, San Diego, 2015.

[10] Makeusof.com. (2021, 29 April). What Is a Decentralized Cryptocurrency Exchange (DEX)?. https://www.makeuseof.com/what-is-a-decentralized-cryptocurrency-exchange-dex/

[11] Financemagnates.com. (2020, 22 Mei). Interest-Bearing Crypto Accounts: A 'Gateway' for New Crypto Users?. https://www.financemagnates.com/cryptocurrency/news/interest-bearing-crypto-accounts-a-gateway-for-new-crypto-users/

[12] Huppmann, Sebastian. (2019). Bitcoin Price Prediction using Sentiment Analysis of Twitter Data. 10.13140/RG.2.2.32123.13605.

[13] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225.

[14] Pan, Y., Hou, L., & Pan, X. (2022). Interplay between stock trading volume, policy, and investor sentiment: A multifractal approach. Physica A: Statistical Mechanics and its Applications, 603, 127706.

[15] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock price prediction using news sentiment analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 205-208). IEEE.

[16] Sedgwick, P. (2012). Pearson's correlation coefficient. Bmj, 345.

[17] Edgari, E., Thiojaya, J., & Qomariyah, N. N. (2022, March). The Impact of Twitter Sentiment Analysis on Bitcoin Price during COVID-19 with XGBoost. In 2022 5th International Conference on Computing and Informatics (ICCI) (pp. 337-342). IEEE.

[18] Pano, T., & Kashef, R. (2020). A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. Big Data and Cognitive Computing, 4(4), 33.