

PREDIKSI HARGA SAHAM SEKTOR ENERGI MENGGUNAKAN METODE SPATIAL TEMPORAL ATTENTION-BASED CONVOLUTIONAL NETWORK BERDASARKAN DATA TEKS DAN NUMERIK

Novita Anggraini, Dwi Arman Prasetya, Trimono

Sains Data, UPN "Veteran" Jawa Timur

Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Indonesia

novitaangrn@gmail.com

ABSTRAK

Perubahan harga saham dipengaruhi oleh berbagai faktor, termasuk data historis harga saham dan sentimen yang terkandung dalam berita keuangan. Penelitian ini bertujuan untuk mengembangkan model prediksi harga saham yang lebih akurat dengan memanfaatkan *Spatial-Temporal Attention-Based Convolutional Network* (STACN). Model ini dirancang untuk menggali hubungan kompleks antara data historis harga saham dan informasi dari berita finansial. Metode yang digunakan melibatkan integrasi *Convolutional Neural Network* (CNN) untuk mengekstraksi fitur dari *thought vectors* berita, *Long Short-Term Memory* (LSTM) untuk menangkap pola temporal dari data harga saham, dan *Spatial-Temporal Attention Network* (STAN) untuk memberikan perhatian pada fitur-fitur yang relevan. Studi kasus dilakukan pada saham sektor energi yang terdaftar di Bursa Efek Indonesia, dengan menggunakan data historis harga saham dan berita dari portal bisnis Indonesia. Hasil eksperimen menunjukkan bahwa model STACN Bi-LSTM menghasilkan akurasi yang lebih tinggi dibandingkan dengan model-model lain seperti LSTM dan Bi-LSTM konvensional, dengan nilai MAE sebesar 24.2776, RMSE 32.9127, dan R^2 0.9365. Temuan ini membuktikan bahwa integrasi analisis spasial-temporal dan mekanisme perhatian efektif dalam meningkatkan akurasi prediksi harga saham.

Kata kunci : *Prediksi harga saham, CNN, LSTM, Attention Mechanism, STACN, Bursa Efek Indonesia*

1. PENDAHULUAN

Pasar keuangan global terus mengalami perubahan yang dipengaruhi oleh faktor ekonomi, politik, dan sosial [1].

Tingginya ketidakpastian dalam pergerakan harga saham mendorong pengembangan metode prediksi yang akurat untuk mendukung keputusan investasi. Prediksi harga saham umumnya dilakukan dengan menganalisis data historis untuk memahami pola pergerakan [2].

Namun, sentimen pasar dan informasi dari berita keuangan juga berperan penting dalam perubahan harga saham [3].

Berita keuangan memberikan informasi penting tentang kondisi ekonomi, kebijakan pemerintah, dan peristiwa global yang dapat mempengaruhi harga saham dalam jangka pendek maupun panjang [4].

Pemahaman terhadap informasi ini menjadi krusial bagi investor dalam mengantisipasi pergerakan pasar.

Meskipun kombinasi data historis dan berita keuangan berpotensi meningkatkan akurasi prediksi, penggabungan kedua sumber data ini menghadirkan tantangan tersendiri. Data historis bersifat numerik dan terstruktur, sedangkan berita keuangan berupa teks tidak terstruktur dengan makna yang kompleks. Berita keuangan sering mengandung istilah teknis, kiasan, dan opini yang sulit dianalisis menggunakan model konvensional [5].

Kompleksitas ini semakin meningkat dengan adanya variasi gaya penulisan dan sudut pandang dari berbagai sumber berita. Oleh karena itu, diperlukan

pendekatan yang dapat mengintegrasikan kedua jenis data ini secara efektif dan memahami konteks informasi secara komprehensif.

Kompleksitas prediksi harga saham yang melibatkan data numerik dan tekstual membutuhkan pendekatan yang dapat mengintegrasikan kedua jenis data tersebut secara efektif [6].

Penelitian ini mengimplementasikan *Spatial-Temporal Attention Based Convolutional Network* (STACN) sebagai solusi untuk tantangan tersebut. Arsitektur STACN mengintegrasikan beberapa komponen *neural network* yang saling melengkapi, yaitu *Convolutional Neural Network* (CNN) yang berperan dalam ekstraksi fitur tekstual dari berita, *Long Short-Term Memory* (LSTM) untuk analisis sekuensial data historis, serta mekanisme *attention* yang mengoptimalkan identifikasi informasi yang signifikan dalam konteks prediksi. Dengan menggabungkan ketiga komponen ini, model diharapkan lebih memahami keterkaitan sentimen berita dan pergerakan harga saham melalui analisis pola spasial berita dan temporal data historis untuk prediksi lebih akurat.

Penelitian ini menguji sejauh mana model STACN bisa memprediksi harga saham dengan memanfaatkan data numerik dan tekstual. Fokusnya adalah pada saham sektor energi yang tercatat di Bursa Efek Indonesia (BEI). Sektor ini dipilih karena dikenal memiliki fluktuasi harga yang cukup tinggi, dipengaruhi oleh berbagai faktor eksternal seperti perubahan harga minyak dunia dan kebijakan energi nasional [7].

Karakteristiknya yang dinamis membuat sektor energi menjadi objek yang menarik untuk menguji seberapa efektif model prediksi ini bekerja. Selain itu, penelitian ini melakukan evaluasi komparatif terhadap performa STACN dengan beberapa model prediksi lainnya, termasuk LSTM konvensional, Bi-LSTM, serta varian STACN yang diintegrasikan dengan LSTM dan Bi-LSTM. Untuk mengukur akurasi prediksi, digunakan metrik *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), *R-squared* (R^2). Dengan evaluasi yang komprehensif ini, diharapkan bisa terlihat kelebihan dan kekurangan masing-masing model. Hasil penelitian ini diharapkan bisa memberikan kontribusi dalam pengembangan model prediksi harga saham yang lebih akurat, terutama dengan menggabungkan data numerik dan data tekstual.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Beberapa penelitian telah mengkaji penggunaan data numerik dan teks berita untuk prediksi harga saham. Salah satu studi mengembangkan *Spatial-Temporal Attention-Based Convolutional Network* (STACN) yang menggabungkan CNN, LSTM, dan mekanisme attention untuk meningkatkan akurasi prediksi. Model ini menunjukkan performa lebih baik dibandingkan model berbasis CNN atau LSTM saja, dengan tingkat kesalahan lebih rendah, yaitu *Mean Squared Error* (MSE) sebesar 0.085 [8].

Penelitian lain menerapkan *Bidirectional LSTM* (Bi-LSTM) dengan analisis sentimen Twitter untuk memprediksi harga saham Bank Central Asia (BBCA). Hasilnya menunjukkan bahwa sentimen publik memiliki pengaruh signifikan terhadap pergerakan harga saham, dengan MSE sebesar 0.094 dan *Root Mean Square Error* (RMSE) sebesar 0.306, yang menunjukkan tingkat akurasi yang lebih tinggi dibandingkan metode konvensional [9].

Studi lainnya menggunakan *Deep Bidirectional LSTM* (DBLSTM) dengan *Word2Vec* untuk memahami hubungan antara berita keuangan dan harga saham di sektor medis. Model ini berhasil meningkatkan akurasi prediksi tren harga saham dengan akurasi klasifikasi sebesar 86%, menunjukkan bahwa pendekatan berbasis teks dan numerik mampu menangkap hubungan kompleks antara berita dan pergerakan pasar [10].

Hasil dari ketiga penelitian di atas menunjukkan bahwa kombinasi data numerik dan teks berita meningkatkan prediksi harga saham. Model berbasis CNN dan LSTM efektif dalam menangkap pola pergerakan harga, terutama ketika dilengkapi dengan mekanisme *attention*.

Penelitian ini bertujuan untuk mengadaptasi pemanfaatan model *deep learning* untuk pengolahan data teks dan numerik untuk saham sektor energi di Bursa Efek Indonesia, dengan memanfaatkan teknik

deep learning yang telah terbukti dalam studi sebelumnya.

2.2. Data Saham Sektor Energi

Sektor Energi mencakup perusahaan yang bergerak di bidang ekstraksi dan distribusi energi, baik yang bersumber dari fosil (minyak bumi, gas alam, batu bara) maupun energi terbarukan (tenaga surya, angin, biomassa) [11].

Perusahaan-perusahaan di sektor ini sangat dipengaruhi oleh harga komoditas energi global, yang berfluktuasi akibat faktor eksternal seperti kebijakan pemerintah, ketegangan geopolitik, dan kondisi ekonomi global. Selain itu, perkembangan teknologi energi terbarukan dan sentimen pasar yang tercermin dalam berita finansial turut mempengaruhi kinerja saham sektor ini [12].

2.3. Thought Vector

Thought vector atau *embedding* vektor, merupakan representasi numerik dari teks dalam ruang vektor berdimensi tinggi yang dapat menangkap makna semantik dari kata, frasa, atau dokumen [13].

Konsep ini telah berkembang signifikan dengan munculnya arsitektur *transformer* yang lebih canggih, menghasilkan representasi kontekstual yang lebih kaya dan dinamis seperti yang ditunjukkan dalam *Large Language Models* (LLMs). Dalam konteks pemrosesan bahasa alami, *thought vector* memungkinkan model untuk memahami hubungan semantik antar kata dan mengukur kemiripan makna menggunakan operasi aljabar vektor. Representasi vektor ini telah terbukti efektif dalam berbagai tugas NLP seperti analisis sentimen, klasifikasi teks, dan mesin penerjemah, karena kemampuannya dalam menangkap konteks dan hubungan semantik yang kompleks dalam bahasa [14].

2.4. CNN

Convolutional Neural Network (CNN) merupakan arsitektur *deep learning* yang dirancang untuk mengekstraksi fitur hierarkis dari data melalui operasi konvolusi, yang telah menunjukkan performa unggul dalam pemrosesan data terstruktur maupun tidak terstruktur. Dalam analisis teks keuangan, CNN dapat mendeteksi pola-pola linguistik lokal dan global melalui filter konvolusinya, memungkinkan ekstraksi fitur semantik yang relevan dengan pergerakan harga saham. Keunggulan CNN terletak pada kemampuannya dalam menangkap dependensi spasial serta arsitekturnya yang dapat diperdalam untuk mempelajari representasi yang lebih abstrak [15].

2.5. LSTM

Long Short-Term Memory (LSTM) adalah arsitektur jaringan saraf yang dirancang untuk memproses dan memprediksi data urutan waktu dengan mempertahankan informasi dari langkah waktu sebelumnya. LSTM efektif dalam menangkap pola temporal dalam data historis harga saham,

memungkinkan model untuk mempertimbangkan pergerakan harga sebelumnya dalam membuat prediksi [16].

2.6. Attention

Mekanisme *attention* adalah teknik dalam jaringan saraf yang memungkinkan model untuk fokus pada bagian tertentu dari input yang lebih relevan untuk tugas tertentu. Dalam konteks prediksi harga saham, mekanisme *attention* dapat digunakan untuk mengidentifikasi informasi penting dari berita keuangan yang memiliki dampak signifikan terhadap pergerakan harga. Dengan menekankan informasi yang relevan, model dapat membuat prediksi yang lebih akurat dan efisien [17].

2.7. STACN

Spatial-Temporal Attention Based Convolutional Network (STACN) adalah model yang menggabungkan CNN, LSTM, dan mekanisme *attention* untuk memprediksi harga saham dengan memanfaatkan data teks dan numerik secara bersamaan. CNN digunakan untuk mengekstraksi fitur dari teks berita, LSTM untuk menangkap pola temporal dalam data harga saham historis, dan mekanisme *attention* untuk menyoroti informasi yang paling relevan dari kedua sumber data tersebut. Pendekatan ini memungkinkan model untuk memahami hubungan kompleks antara berita keuangan dan pergerakan harga saham, sehingga meningkatkan akurasi prediksi [18].

2.8. Metrik Evaluasi

Dalam melakukan evaluasi kinerja model prediktif, penggunaan berbagai metrik evaluasi sangat penting untuk mengukur seberapa baik model dapat melakukan prediksi. Metrik evaluasi yang digunakan adalah *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), dan *R-squared* (R^2) *Score*.

Mean Absolute Error (MAE) merupakan metrik evaluasi yang mengukur rata-rata nilai absolut dari selisih antara nilai prediksi dan nilai aktual.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

Mean Squared Error (MSE) menghitung rata-rata kuadrat selisih antara nilai prediksi dan nilai aktual. MSE memberikan bobot yang lebih besar pada *error* yang besar karena menggunakan kuadrat dalam perhitungannya.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

Root Mean Squared Error (RMSE) merupakan akar kuadrat dari MSE. RMSE memiliki kelebihan dibandingkan MSE karena memiliki satuan yang sama dengan variabel target, sehingga lebih mudah diinterpretasikan.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

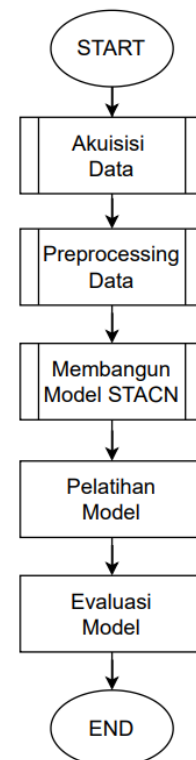
R-squared (R^2) *Score* atau koefisien determinasi mengukur seberapa baik model dapat menjelaskan variasi dalam data. Nilai R^2 berkisar antara 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan model yang lebih baik.

$$R^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \sum_{i=1}^n (y_i - \bar{y})^2 \quad (4)$$

Dalam rumus-rumus di atas, y_i merepresentasikan nilai aktual, $\hat{y}$ merepresentasikan nilai prediksi, $\bar{y}$ merepresentasikan rata-rata nilai aktual, dan n adalah jumlah sampel data.

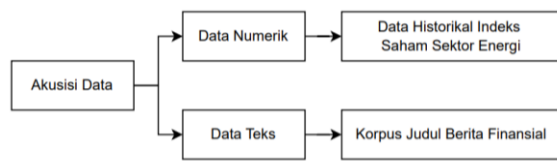
3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan berbasis pembelajaran mesin dengan menerapkan model *Spatial Temporal Attention-based Convolutional Network* (STACN) untuk memprediksi harga saham sektor energi. Alur metodologi penelitian dirancang untuk memastikan setiap tahap pemrosesan data, pelatihan model, serta evaluasi kinerja dilakukan secara sistematis dan terstruktur. Diagram alir metodologi penelitian yang ditampilkan pada Gambar 3.1 menggambarkan tahapan utama yang meliputi pengumpulan dan pengolahan data, pembangunan arsitektur model, pelatihan model, serta evaluasi kinerja model menggunakan berbagai metrik evaluasi.



Gambar 1. Diagram alir metodologi penelitian

3.1. Akuisisi Data

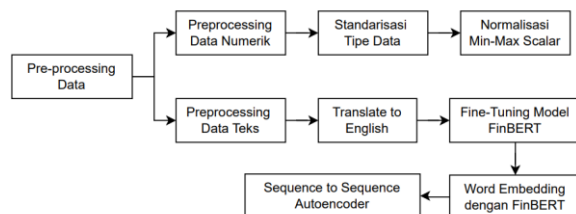


Gambar 2. Diagram alir akuisisi data

Penelitian ini diawali dengan tahap akuisisi data, yang terdiri dari dua jenis data utama, yaitu data numerik dan data teks. Data numerik yang digunakan dalam penelitian ini mencakup informasi historis mengenai indikator teknikal saham, yaitu *Open, High, Low, Close, Volume* (OHLCV). Data tersebut diperoleh dari indeks saham sektor energi di Indonesia, dengan rentang waktu mulai dari 25 Januari 2021 hingga 31 Januari 2025, yang diambil dari laman resmi Indonesia Stock Exchange (IDX).

Sementara itu, data teks dalam penelitian ini berupa berita finansial yang berkaitan dengan sektor energi. Data ini dikumpulkan dalam bentuk judul berita dengan jumlah lima berita per hari, yang bersumber dari portal berita CNBC Indonesia. Rentang waktu pengambilan data teks disesuaikan dengan data numerik, yakni dari 25 Januari 2021 hingga 31 Januari 2025.

3.2. Preprocessing Data



Gambar 3. Diagram alir *preprocessing* data

Tahap praproses data dalam penelitian ini terbagi menjadi dua alur utama, yaitu praproses data numerik dan praproses data tekstual. Kedua alur ini dijalankan secara paralel untuk mempersiapkan data yang akan digunakan dalam model prediksi. Pada alur praproses data numerik, terdapat dua tahapan utama. Tahap pertama adalah standarisasi tipe data, di mana seluruh data numerik diseragamkan tipe datanya untuk memastikan konsistensi dalam pengolahan. Tahap kedua adalah normalisasi menggunakan metode *MinMax Scaler*, yang mengubah rentang nilai data menjadi skala 0 hingga 1.

Sementara itu, praproses data tekstual melibatkan lima tahapan yang lebih kompleks. Tahap pertama dimulai dengan implementasi *FinBERT*, sebuah model bahasa yang telah dilatih khusus untuk domain keuangan. Selanjutnya, dilakukan augmentasi token dengan menambahkan token-token khusus yang berkaitan dengan saham Indonesia untuk meningkatkan relevansi model terhadap konteks lokal.

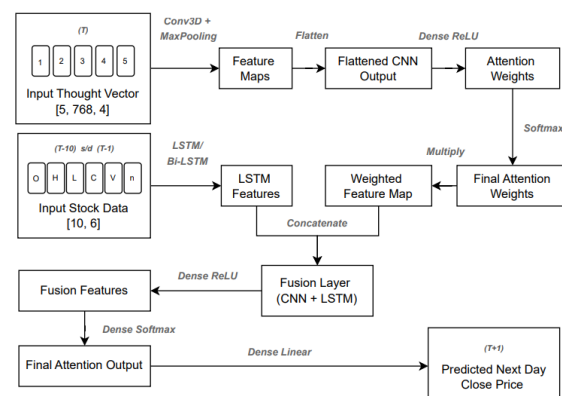
Tahap ketiga adalah *fine-tuning*, di mana model *FinBERT* disesuaikan dengan karakteristik data tekstual yang spesifik untuk pasar saham Indonesia. Setelah itu, dilakukan *inference embedding* untuk mengubah teks menjadi representasi vektor yang dapat diproses oleh model. Tahap terakhir menggunakan *sequence-to-sequence autoencoder* untuk mereduksi dimensi *embedding* dengan tetap mempertahankan informasi penting dari data tekstual.

Hasil dari kedua alur praproses ini akan menghasilkan data yang telah terstandarisasi dan siap digunakan untuk tahap pemodelan selanjutnya.

3.3. Membangun Model STACN

Model STACN (*Spatial-Temporal Attention Convolutional Network*) dikembangkan dengan mengintegrasikan analisis spasial dan temporal melalui dua jalur input paralel. Input yang digunakan terdiri dari dua jenis data. Pertama, *thought vector* lima berita pada hari T dengan dimensi [5, 768, 4]. Kedua, data historis teknikal saham yang mencakup 6 indikator, yaitu *open, high, low, close, volume, dan num_stock*, selama 10 hari sebelum hari T (dari T-10 hingga T-1) dengan dimensi [10, 6].

Jalur pertama memproses *thought vectors* menggunakan *Convolutional 3D (Conv3D)* dan *MaxPooling* untuk menghasilkan *feature maps* yang menangkap pola spasial dari data tekstual. Sementara itu, jalur kedua memproses data historis saham berdimensi [10, 6] menggunakan *Bidirectional LSTM* atau *LSTM* untuk menangkap pola temporal dari rangkaian data. Gambar 4 memberikan gambaran lengkap mengenai arsitektur dan proses pemodelan STACN yang diimplementasikan dalam penelitian ini.



Gambar 4. Diagram alir pembuatan model STACN

Selanjutnya, mekanisme *attention* diterapkan pada output CNN yang telah di-flatten sebuah layer Dense dengan 90 unit dan fungsi aktivasi ReLU, serta transformasi *softmax* 5-dimensi. Hasil pembobotan kemudian digabungkan dengan output LSTM dalam *fusion layer*, dilanjutkan dengan dua layer Dense (100 dan 50 unit) untuk menghasilkan *fusion features*. Prediksi akhir diperoleh melalui *attention mechanism* final dan layer Dense dengan fungsi aktivasi linear

untuk menghasilkan prediksi harga saham hari berikutnya ($T+1$).

3.4. Pelatihan Model

Pelatihan model dilakukan menggunakan arsitektur *Spatial Temporal Attention-based Convolutional Network* (STACN) yang mengombinasikan *Long Short-Term Memory* (LSTM) *bidirectional* dan konvolusi 3D untuk menangkap pola temporal dan spasial dalam data saham sektor energi. Dataset yang digunakan terdiri dari data numerik pasar saham dan data teks dari berita terkait sektor energi yang telah diolah menjadi vektor pemikiran (thought vector). Data dibagi menjadi data latih dan data uji dengan proporsi 85:15. Input model terdiri dari dua bagian, yaitu data berita dengan dimensi [5, 768, 4, 1] dan data numerik pasar saham dengan dimensi [10, 6], di mana *timestep* yang digunakan adalah 10. Proses pelatihan menggunakan *optimizer Adam* dengan *learning rate* sebesar 0.001 dan *loss function Mean Squared Error* (MSE). Model dilatih selama 200 *epoch* dengan *batch size* 8 serta menggunakan mekanisme *early stopping* untuk menghindari *overfitting*.

3.5. Evaluasi Model

Evaluasi model dilakukan dengan mengukur performa prediksi harga saham menggunakan metrik evaluasi seperti *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), *Mean Squared Error* (MSE), koefisien determinasi (R^2), dan *Mean Absolute Percentage Error* (MAPE). Model terbaik dipilih berdasarkan nilai MAE terendah pada data validasi. Selain itu, prediksi model dibandingkan dengan nilai aktual untuk mengamati pola dan akurasi prediksi. Visualisasi hasil prediksi dilakukan untuk melihat sejauh mana model mampu mengikuti tren pergerakan harga saham yang sebenarnya.

4. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan dengan mengikuti metodologi sebagaimana yang ditunjukkan dalam Metodologi Penelitian. Proses dimulai dari pengumpulan data numerik dan teks, pemrosesan data, pembentukan vektor pemikiran (*thought vector*), hingga pelatihan dan evaluasi model menggunakan metode *Spatial Temporal Attention-based Convolutional Network* (STACN). Evaluasi dilakukan untuk mengukur akurasi prediksi harga saham sektor energi berdasarkan berbagai metrik yang relevan.

4.1. Akuisisi Data

Penelitian ini mengumpulkan dua jenis data, yaitu data numerik dan data teks, dalam rentang waktu yang sama, yakni dari 25 Januari 2021 hingga 31 Januari 2025. Tabel 1 menggambarkan data numerik yang diperoleh, terdiri dari 974 baris data yang merepresentasikan hari aktif bursa selama periode tersebut. Setiap baris data mencakup 6 indikator teknikal saham, yaitu *open*, *high*, *low*, *close*, jumlah emiten dalam sektor, dan *volume*.

Tabel 1. Tabel data numerik

date	open	high	low	close	n_stock	vol
25/01/2021	803	803	780	786	66	1376189590
26/01/2021	786	794	764	770	66	1946080885
27/01/2021	770	778	754	770	66	2038294143
...	...	...	...	...	...	...
30/01/2025	2878	2895	2844	2869	90	2247141030
31/01/2025	2869	2893	2868	2890	90	2011277420

Sementara itu, Tabel 2 menggambarkan data teks berupa judul berita yang dikumpulkan dengan rata-rata 5 berita per hari. Dengan total 974 hari aktif, jumlah berita yang terkumpul adalah 4.870 berita.

Tabel 2. Tabel data teks

date	title_news
25/01/2021	['Harga batubara diprediksi naik simak rekomendasi saham Bukit Asam PTBA ini', ... 'Harga minyak turun untuk sesi kedua lockdown Covid19 menekan permintaan BBM']
26/01/2021	['Harga Minyak Kedelai Naik Harga CPO Ikutan Dapat Berkah', ... 'Pelan-pelan Harga Batu Bara Siap Cicipi Level US 90ton Lagi']
27/01/2021	['Naik 20 PTBA Produksi Batu Bara Tembus 30 Juta Ton di 2021', ... 'PGN Bakal Bangun 50 Ribu Jargas di 2021 Ada yang Minat']
...	...
30/01/2025	['IHSG Terjun 136 ke 7068 di Sesi I Kamis 301 AMMN AKRA ANTM Top Losers LQ45', ... 'Harga Minyak Naik Tipis pada Kamis 301 Setelah Tumbang']
31/01/2025	['Ini Kata PLN Soal Pasar Karbon PLTU Wajib Beli Kredit Karbon', ... 'Harga Minyak Dunia Melemah Jumat 311 Brent ke US 7670 dan WTI ke US 7262']

4.2. Preprocessing Data

Tahap *preprocessing* data dalam penelitian ini bertujuan untuk mempersiapkan data tekstual dan mengintegrasikannya dengan data historis saham. Proses ini diawali dengan pengumpulan judul berita, yang kemudian diterjemahkan ke dalam bahasa Inggris seperti yang tercantum pada Tabel 3, sebelum dikonversi menjadi vektor menggunakan model *FinBERT*. Sebelum digunakan, model *FinBERT* terlebih dahulu *difine-tuning* menggunakan korpus berita finansial yang telah dikumpulkan dalam penelitian ini, sehingga model dapat lebih memahami konteks dan terminologi yang relevan dengan analisis saham.

Tabel 3. Judul berita terjemahan

date	title_news_translated
25/01/2021	['Coal prices are predicted to rise, check out this stock recommendation for Bukit Asam PTBA', ... 'Oil prices drop for the second session,

date	title_news_translated
	Covid-19 lockdown suppresses fuel demand']
26/01/2021	['Soybean Oil Prices Rise, Palm Oil Prices Also Benefit', ... 'Slowly, Coal Prices Are Ready to Test the US \$90/ton Level Again']
27/01/2021	['PTBA Coal Production Rises by 20%, Reaches 30 Million Tons in 2021', ... 'PGN to Build 50 Thousand Gas Networks in 2021, Anyone Interested?']
...	...
30/01/2025	['IHSG Drops 136 to 7068 in Session I on Thursday, 301 AMMN AKRA ANTM Top Losers LQ45', ... 'Oil Prices Rise Slightly on Thursday 301 After Declining']
31/01/2025	['This is What PLN Says About the Carbon Market, Power Plants Must Buy Carbon Credits', ... 'Global Oil Prices Weaken on Friday 311, Brent to US \$76.70 and WTI to US \$72.62']

Penyesuaian model *FinBERT* dilakukan dalam dua tahap *fine-tuning*. Tahap pertama bertujuan untuk menyesuaikan model dasar *FinBERT* dengan karakteristik berita saham yang telah dikumpulkan. Pada tahap ini, dilakukan penambahan token baru untuk mengakomodasi terminologi spesifik dalam domain finansial, seperti nama emiten di sektor energi dan istilah pasar saham Indonesia, misalnya IHSG, BI-Rate, dan lainnya. Dalam proses transformasi teks menjadi *word embedding*, setiap judul berita diproses sehingga panjangnya diseragamkan menjadi 25 token. Selanjutnya, judul-judul tersebut dikonversi menjadi *embedding*, menghasilkan representasi numerik berdimensi 25×768 untuk setiap judul berita. Tabel 4 menggambarkan contoh *embedding* yang dihasilkan, yang merepresentasikan lima judul berita dalam satu hari dalam bentuk vektor berdimensi $5 \times 25 \times 768$, menggunakan model *FinBERT* yang telah disesuaikan.

Tabel 4. Tabel word embedding judul berita

date	title_news_translated
25/01/2021	[[[-0.56455, -0.04825, ... -0.049553, 0.00715] ... [0.11924, -0.18943, ... -0.125667, 0.38238]]
26/01/2021	[[[-0.53965, -0.01065, ... -0.03516, -0.00949] ... [0.12866, -0.18381, ... -0.12021, 0.37724]]
27/01/2021	[[[-0.56493, -0.08860, ... -0.03965, -0.01813] ... [0.12673, -0.18798, ... -0.12481, 0.375697]]
...	...
30/01/2025	[[[-0.56561, -0.08471, ... -0.02707, -0.00198] ... [0.11878, -0.18409, ... -0.13056, 0.38518]]
31/01/2025	[[[-0.46028, -0.06694, ... -0.05958, -0.01508] ... [0.12943, -0.18229, ... -0.12201, 0.37991]]

Dataset yang digunakan dalam penelitian ini terdiri dari 974 baris, dengan setiap baris berisi lima judul berita yang diterbitkan pada satu hari bursa.

Untuk memperoleh representasi numerik yang kaya secara semantik, setiap judul berita dikonversi menjadi *embedding* menggunakan model *FinBERT* yang telah disesuaikan. Dalam satu hari, dipilih lima judul berita, masing-masing dikodekan ke dalam vektor berdimensi $5 \times 25 \times 768$, di mana angka lima menunjukkan jumlah berita harian yang digunakan, 25 merupakan jumlah token maksimum dalam satu judul berita, dan 768 merepresentasikan dimensi *embedding* untuk setiap token.

Embedding yang telah diperoleh selanjutnya diproses menggunakan *sequence-to-sequence autoencoder* berbasis *Long Short-Term Memory (LSTM)* dengan empat lapisan untuk mengekstraksi *thought vector*. Dalam model ini, setiap kata dalam judul berita diperlakukan sebagai *timestep* dan diproses secara sekuensial dalam lapisan *encoder*. Representasi dari *hidden states* pada setiap lapisan LSTM dikumpulkan dan digunakan sebagai *thought vector*, yang berfungsi sebagai abstraksi semantik dari judul berita.

Thought vector yang dihasilkan selanjutnya digunakan sebagai *initial hidden state* dalam *decoder*, yang bertugas merekonstruksi kembali judul berita berdasarkan representasi tersebut. Pendekatan ini memungkinkan model untuk menangkap dan mengompresi informasi semantik secara lebih efisien. Optimasi model dilakukan menggunakan algoritma *Adaptive Moment Estimation (ADAM)*, dengan judul berita asli sebagai label pelatihan.

Model *autoencoder* yang telah dilatih diterapkan pada setiap *embedding* judul berita finansial melalui proses inferensi. Setiap *embedding* dikonversi menjadi matriks berdimensi 4×768 , yang merepresentasikan *thought vector* hasil ekstraksi. Sebagaimana ditunjukkan pada Tabel 5, dengan lima judul berita per hari, dalam satu hari bursa akan terbentuk matriks berdimensi $5 \times 4 \times 768$. Angka lima menunjukkan jumlah berita harian yang digunakan, angka empat merepresentasikan jumlah vektor yang dihasilkan oleh *autoencoder*, dan 768 merupakan dimensi *embedding* dari setiap vektor. Representasi ini bertujuan untuk menangkap informasi semantik secara lebih padat dan bermakna, sehingga dapat digunakan sebagai input model STACN.

Tabel 5. Tabel thought vector judul berita

date	thought_vector
25/01/2021	[[[9.5079727e-08 ... 5.2500120e-09] [5.9594505e-04 ... -1.4862095e-03] [-5.7549119e-02 ... 3.0091063e-03] [2.1013623e-10 ... -2.3088778e-09]] ...
26/01/2021	[[[5.70410862e-07 ... 5.93684263e-07] [-1.84623024e-03 ... 5.36025725e-02] [-3.65296699e-04 ... 6.31877780e-02] [8.28723401e-08 ... -1.28420169e-07]] ...
27/01/2021	[[[-4.73826475e-07 ... 8.48272350e-07] [-1.17138145e-03 ... -2.16182787e-03]]

date	thought_vector
	[-2.90140688e-01 ... 1.49002969e-01] [1.08405729e-09 ... -7.92406070e-07]]
...	...
30/01/2025	[[[-3.4218438e-08 ... 4.4111911e-08] [-4.7999981e-04 ... -2.2782929e-02] [-3.3611365e-02 ... 1.0718027e-01] [3.9652509e-12 ... -3.9938663e-11]] ...
31/01/2025	[[[5.73033390e-07 ... 6.35148538e-08] [1.42512326e-05 ... -1.60857313e-03] [-3.16329807e-01 ... -1.55190515e-04] [4.80680001e-07 ... -4.26394763e-07]] ...

4.3. Membangun Model STACN

Pada tahap pembuatan model, data yang menjadi input adalah data berita dan data historis saham yang telah melalui tahap praproses sebelumnya. Arsitektur ini dirancang untuk menangkap hubungan kompleks antara berita dan pergerakan saham dengan mengoptimalkan fitur spasial dari data teks menggunakan CNN dan pola temporal dari data numerik menggunakan LSTM.

Data berita yang telah dikonversi menjadi representasi vektor (*thought vector*) dengan dimensi (*batch_size*, 5, 768, 4, 1) diproses menggunakan arsitektur *Convolutional Neural Network* (CNN) berbasis 3D *convolution*. Pemrosesan dimulai dengan lapisan *Conv3D*, yang mengubah dimensi fitur dari 768 menjadi 192 dengan 128 filter. Lapisan ini diikuti oleh *MaxPooling3D*, yang bertujuan untuk mereduksi dimensi spasial dan mempertahankan fitur yang paling informatif. Proses konvolusi dan *pooling* ini dilakukan secara bertahap hingga ukuran fitur berkurang menjadi (5, 1, 4, 128), yang kemudian diratakan (*flatten*) sebelum diteruskan ke lapisan *fully connected*.

Sementara itu, data historis saham yang terdiri dari fitur numerik dengan dimensi (*batch_size*, 10, 6) diproses menggunakan arsitektur *Long Short-Term Memory* (LSTM). Model pertama-tama melewati data ke dalam *layer* pertama dengan 256 unit, diikuti dengan lapisan *dropout* untuk mencegah *overfitting*. Kemudian, hasil dari *layer* pertama dimasukkan ke dalam *layer* kedua dengan 64 unit untuk menangkap pola temporal dalam data saham.

Setelah ekstraksi fitur dari kedua sumber data selesai, hasil *output* CNN dan LSTM digabungkan (*concatenate*) dalam lapisan *fully connected*. Proses ini menghasilkan representasi gabungan yang mencerminkan hubungan antara sentimen berita dan pergerakan harga saham. Lapisan *fully connected* yang digunakan terdiri dari beberapa lapisan *Dense*, dengan jumlah *neuron* yang menurun secara bertahap, yaitu dari 100, 50, hingga 5 *neuron*. Untuk memperkuat hubungan antar fitur, dilakukan operasi *Multiply* pada salah satu cabang sebelum akhirnya kembali digabungkan dengan fitur dari jalur LSTM. Hasil akhir dari kombinasi ini kemudian diproses dalam lapisan

Dense terakhir dengan 32 *neuron* sebelum menghasilkan output akhir berupa prediksi harga saham dalam bentuk satu nilai numerik.

4.4. Pelatihan Model

Setelah model STACN berhasil dibangun, tahap selanjutnya adalah pelatihan model menggunakan data yang telah diproses sebelumnya. Dataset dibagi menjadi *training set* (85%) dan *testing set* (15%), dengan data pelatihan digunakan untuk mengoptimalkan bobot model dan data uji untuk mengukur kinerjanya.

Untuk melatih model, digunakan fungsi *loss Mean Squared Error* (MSE), karena tugas utama model adalah prediksi harga saham, yang merupakan regresi numerik. *Optimizer Adam* dengan *learning rate* awal sebesar 0.001 digunakan untuk mempercepat konvergensi model. Selain itu, digunakan *batch size* sebesar 32 dan jumlah *epoch* sebanyak 500, dengan *early stopping* diterapkan untuk menghentikan pelatihan jika model tidak menunjukkan peningkatan signifikan dalam beberapa *epoch* berturut-turut.

Selama proses pelatihan, *callback* untuk menyimpan bobot terbaik (*best model checkpoint*) juga diterapkan untuk memastikan model yang disimpan adalah model dengan performa terbaik berdasarkan validasi *loss* terendah. Penelitian ini mencakup empat skenario pelatihan yang diuji. Ringkasan dari skenario percobaan tersebut disajikan dalam Tabel 6.

Tabel 6. Skenario percobaan pemodelan

No	Data Numerik	Data Teks	Skema Percobaan
1.	Ya	Tidak	LSTM
2.	Ya	Tidak	Bidirectional LSTM
3.	Ya	Ya	STACN dengan LSTM
4.	Ya	Ya	STACN dengan Bidirectional LSTM

4.5. Evaluasi Model

Evaluasi model dilakukan untuk mengukur kemampuan *Spatial-Temporal Attention-Based Convolutional Network* (STACN) dalam memprediksi harga saham berdasarkan kombinasi data berita dan numerik. Pengujian dilakukan menggunakan data uji atau *testing set*, dengan kinerja model dievaluasi menggunakan metrik *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), dan *R-squared* (R^2) *Score*.

Tabel 7. Tabel perbandingan evaluasi model

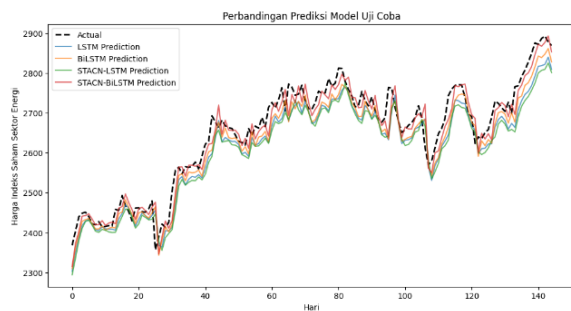
Model	MAE	RMSE	MSE	R^2
LSTM	39.53	47.53	2259.59	0.86
Bi-LSTM	31.99	40.13	1610.44	0.90
STACN LSTM	37.53	45.55	2074.67	0.87
STACN Bi-LSTM	24.27	32.91	1083.24	0.94

Hasil evaluasi menunjukkan bahwa STACN dengan arsitektur Bi-LSTM memiliki kinerja yang lebih baik dibandingkan dengan model LSTM dan Bi-LSTM konvensional. Berdasarkan Tabel 7, hasil

evaluasi menunjukkan bahwa STACN dengan arsitektur Bi-LSTM memiliki kinerja yang paling baik dalam memprediksi harga saham dibandingkan model lainnya. Model ini menghasilkan MAE sebesar 24.2776, RMSE sebesar 32.9127, dan MSE sebesar 1083.2434, dengan R^2 mencapai 0.9365, menunjukkan kemampuan yang lebih baik dalam menangkap pola pergerakan harga saham.

Sebagai perbandingan, model LSTM konvensional memiliki MAE sebesar 39.5287 dan RMSE sebesar 47.5352, dengan MSE sebesar 2259.5952 dan R^2 hanya sebesar 0.8676, yang menunjukkan bahwa model ini memiliki tingkat kesalahan yang lebih tinggi dibandingkan STACN Bi-LSTM. Sementara itu, model Bi-LSTM konvensional dan STACN LSTM juga menunjukkan performa yang lebih rendah dibandingkan STACN Bi-LSTM, dengan MAE masing-masing sebesar 31.9858 dan 37.5253, serta RMSE sebesar 40.1303 dan 45.5486.

Performa masing-masing model juga dapat diamati melalui grafik pemodelan yang ditampilkan pada Gambar 5.



Gambar 5. Perbandingan prediksi data test model uji coba

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian tentang prediksi harga saham menggunakan model STACN Bi-LSTM, dapat disimpulkan bahwa model ini menunjukkan kinerja yang lebih unggul dibandingkan model LSTM dan Bi-LSTM konvensional. Dengan memanfaatkan CNN untuk ekstraksi fitur *thought vector*, Bi-LSTM untuk menangkap pola temporal dua arah, serta mekanisme perhatian berbasis STAN, model STACN Bi-LSTM berhasil meningkatkan akurasi prediksi harga saham. Hasil pengujian menunjukkan performa yang sangat baik dengan MAE 24.2776, RMSE 32.9127, dan R^2 0.9365, yang lebih tinggi dibandingkan model lainnya. Meskipun demikian, penelitian ini mengakui bahwa masih ada ruang untuk meningkatkan kinerja model. Penggunaan data yang lebih bervariasi seperti mengintegrasikan data *real-time* dan sumber informasi tambahan seperti laporan keuangan resmi atau sentimen media sosial dapat memperkaya *dataset* dan meningkatkan kemampuan generalisasi model.

DAFTAR PUSTAKA

- [1] R. Nurjanah, "FAKTOR EKONOMI MAKRO DOMESTIK DAN GLOBAL TERHADAP IHSG DOMESTIC AND GLOBAL MACRO ECONOMIC ON THE JAKARTA COMPOSITE INDEX," *Tax and Business Journal*, vol. 308, no. 2, 2023.
- [2] M. Girma dan O. Drosis, "Exploring the effects of diverse historical stock price data on the accuracy of stock price prediction models," *J Emerg Investig*, hlm., 2024, doi: 10.59720/23-144.
- [3] R. Ren, D. Wu, dan T. Liu, "Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine," *IEEE Syst J*, vol. 13, hlm. 760–770, 2019, doi: 10.1109/JSYST.2018.2794462.
- [4] R. Basilone, "The Impact of News on Stock Market Investors," *MedRN: Interdisciplinary Coronavirus & Infectious Disease Related Research (Topic)*, hlm., 2021, doi: 10.2139/ssrn.3908662.
- [5] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. Chitkushev, dan D. Trajanov, "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers," *IEEE Access*, vol. 8, hlm. 131662–131682, 2020, doi: 10.1109/ACCESS.2020.3009626.
- [6] C. T. Lin, Y. K. Wang, P. L. Huang, Y. Shi, dan Y. C. Chang, "Spatial-temporal attention-based convolutional network with text and numerical information for stock price prediction," *Neural Comput Appl*, vol. 34, no. 17, hlm. 14387–14395, Sep 2022, doi: 10.1007/s00521-022-07234-0.
- [7] S. Sektor, E. M. V. E. C. M. Melati, P. Silvianti, dan F. Afendi, "Analysis Of Stock Market, Mining Commodity, Exchange Rate, And Energy Sector Stock Index Using Vector Error Correction Model," *Indonesian Journal of Statistics and Its Applications*, hlm., 2023, doi: 10.29244/ijsa.v7i1p44-55.
- [8] C. T. Lin, Y. K. Wang, P. L. Huang, Y. Shi, dan Y. C. Chang, "Spatial-temporal attention-based convolutional network with text and numerical information for stock price prediction," *Neural Comput Appl*, vol. 34, no. 17, hlm. 14387–14395, Sep 2022, doi: 10.1007/s00521-022-07234-0.
- [9] N. Afrianto, D. H. Fudholi, dan S. Rani, "Prediksi Harga Saham Menggunakan BiLSTM dengan Faktor Sentimen Publik," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 1, hlm. 41–46, Feb 2022, doi: 10.29207/resti.v6i1.3676.
- [10] Y. Ren, F. Liao, dan Y. Gong, "Impact of News on the Trend of Stock Price Change: an Analysis based on the Deep Bidirectional LSTM Model," dalam *Procedia Computer Science*, Elsevier B.V., 2020, hlm. 128–140. doi: 10.1016/j.procs.2020.06.068.

- [11] Bursa Efek Indonesia, "Panduan IDX Industrial Classification," *Lampiran Pengumuman BEI No.: Peng-00007/BEI.POP/01-2021*, vol. Versi 1.1, Jan 2021.
- [12] S. Sektor, E. M. V. E. C. M. Melati, P. Silvianti, dan F. Afendi, "Analysis Of Stock Market, Mining Commodity, Exchange Rate, And Energy Sector Stock Index Using Vector Error Correction Model," *Indonesian Journal of Statistics and Its Applications*, hlm., 2023, doi: 10.29244/ijsa.v7i1p44-55.
- [13] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, dan Y. Liu, "RoFormer: Enhanced Transformer with Rotary Position Embedding," Apr 2021, [Daring]. Tersedia pada: <http://arxiv.org/abs/2104.09864>
- [14] B. Min dkk., "Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey," *ACM Comput. Surv.*, vol. 56, no. 2, Sep 2023, doi: 10.1145/3605943.
- [15] K. S. Varshitha, C. G. Kumari, M. Hasvitha, S. Fiza, A. K, dan V. Rachapudi, "Natural Language Processing using Convolutional Neural Network," dalam *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 2023, hlm. 362–367. doi: 10.1109/ICCMC56507.2023.10083608.
- [16] P. Babu, C. Reddy, K. Jashwanth, S. U. Rahim, dan L. Ali, "STOCK PRICE PREDICTION USING LSTM," *IJARCCCE*, hlm., 2024, doi: 10.17148/ijarcce.2024.134132.
- [17] G. Brauwers dan F. Frasincar, "A General Survey on Attention Mechanisms in Deep Learning," *IEEE Trans Knowl Data Eng*, vol. 35, hlm. 3279–3298, 2022, doi: 10.1109/TKDE.2021.3126456.
- [18] C. T. Lin, Y. K. Wang, P. L. Huang, Y. Shi, dan Y. C. Chang, "Spatial-temporal attention-based convolutional network with text and numerical information for stock price prediction," *Neural Comput Appl*, vol. 34, no. 17, hlm. 14387–14395, Sep 2022, doi: 10.1007/s00521-022-07234-0.