



Sentiment analysis for event-based stock price predictions using bidirectional long short term memory

Okki Setyawan¹, Hilman F.Pardede^{2*}

Universitas Nusa Mandiri¹, Universitas Nusa Mandiri²

Email : 14002429@nusamandiri.ac.id , hilman@nusamandiri.ac.id

Received: February 20, 2022. **Revised:** March 18, 2022. **Accepted:** April 10, 2022. **Issue Period:** Vol.6 No.1 (2022), Pp. 50-58

Abstrak: Pasar saham sudah menjadi peranan yang penting dalam perekonomian dan menarik perhatian investor karena juga menghasilkan dana dan membuat keputusan investasi bagi perusahaan dan investor. hal yang paling utama dilakukan oleh investor sebelum melakukan investasi yaitu melihat informasi tentang pasar modal di berbagai media, baik sosial media dan portal berita yang beredar di internet untuk melihat pergerakan saham yang sering berubah karena fluktuatif. Sentimen analisis yang dilakukan dalam penelitian ini menggunakan dataset publik dari portal dataset di kaggle.com yaitu dataset DJIA (Dow Jones Industrial Average) yang berbentuk peristiwa (timeseries) yang berisikan data kumpulan headline berita dari redditnews.com sebanyak 1990 baris data serta memiliki label kelas untuk naik atau turun, kemudian di training dengan model deep learning yaitu BiLSTM dan LSTM dengan variasi hyperparameter tuning , dan dikomparasikan dengan model lain diantaranya SVM , KNN, Logistic Regression, Naïve Bayes Multinomial. Dari peneltian tersebut didapatkan hasil terbaik dengan model Deep BiLSTM dengan 1 hidden layer , Epoch 20 dan batch size 16 menghasilkan skor Akurasi sebesar 0.8677 dan F1 score sebesar 0.8674 serta AUC sebesar 0.8671, penelitian ini merupakan improvement dari penelitian sebelumnya yang menggunakan model LSTM dengan tingkat akurasi 0.5212, F1 Score 0.6762 dan AUC 0.6211.

Kata kunci: Saham, Deep Learning, BiLSTM, LSTM.

Abstract: The stock market has become an important role in the economy and attracts the attention of investors because it also generates funds and makes investment decisions for companies and investors. The most important thing investors do before investing is to look at information about the capital market in various media, both social media and news portals circulating on the internet to see stock movements that often change due to fluctuations. Sentiment analysis carried out in this study uses a public dataset from the dataset portal at kaggle.com, namely the DJIA (Dow Jones Industrial Average) dataset in the form of events (timeseries) which contains data from a collection of news headlines from redditnews.com as many as 1990 lines of data and has a class label. to go up or down, then trained with deep learning models, namely BiLSTM and LSTM with variations of hyperparameter tuning, and compared with other models including SVM, KNN, Logistic Regression, Naïve Bayes Multinomial. From this research, the best results were obtained with the Deep BiLSTM model with 1 hidden layers, Epoch 20 and batch size 16 resulting in an accuracy score of 0.8677 and F1 score 0.8674 also AUC 0.8671, this study is an improvement from previous research using the LSTM model with an accuracy rate of 0.5212. F1 Score 0.6762 and AUC 0.6211.



DOI: 10.52362/jisicom.v6i1.772

Ciptaan disebarluaskan di bawah [Lisensi Creative Commons Atribusi 4.0 Internasional](https://creativecommons.org/licenses/by/4.0/).



Keywords: Stock, Deep Learning, BiLSTM, LSTM.

I. PENDAHULUAN

Di era revolusi 4.0 ini, jaringan internet dapat digunakan sebagai sarana untuk menyebarluaskan informasi tentang bentuk dan metode investasi baik bagi pendatang baru maupun yang sudah lama berkecimpung di bidang investasi (dipengaruhi oleh platform media sosial). Dalam situasi serba digital, investor kini bisa belajar tentang pasar saham dari mana saja, salah satunya dari media sosial yang sedang populer saat ini, salah satu manfaat media sosial adalah memudahkan masyarakat untuk mendapatkan informasi tentang pergerakan saham. Oleh karena itu investor perlu mengetahui hal tersebut sangat disarankan sebelum berinvestasi [1].

Harga saham dapat terus naik atau turun karena fluktuasi penawaran dan permintaan. Jika lebih banyak orang ingin membeli saham, harga pasarnya akan naik. Jika lebih banyak orang mencoba menjual saham, harganya akan turun. Penawaran dan permintaan sangat sensitif terhadap berita terbaru [2].

Sebagian besar investor lebih suka membaca berita tentang pasar modal dan saham dari media elektronik di Internet. Berita tersebut berisi saran pembelian saham, perkiraan saham jangka pendek, nilai IHSG, tips investasi dan posisi saham yang cenderung menguat atau melemah (naik atau turun) [3].

Ada banyak faktor yang mempengaruhi trend saham, salah satunya adalah berita harian. Ketika individu merespon trend pasar dengan mengkritik atau mendukung trend pasar secara terbuka setiap hari, berita harian dapat berdampak signifikan terhadap fluktuasi harga saham. Artikel dan platform berita digunakan untuk menyebarkan informasi tersebut kepada publik dan dapat memengaruhi transaksi, yang diperlukan untuk analisis informasi yang mendalam guna mendukung investor sebelum membuat keputusan perdagangan yang bijaksana [4].

Dari beberapa penelitian terkait pasar saham, peneliti mengusulkan untuk melakukan penelitian lebih lanjut dengan Daily News for Stock Market Prediction yang tersedia pada website Kaggle.com menggunakan model Deep Learning dengan algoritma Bidirectional Long Short Term Memory dan word embedding pada data training untuk mendapatkan hasil model yang lebih baik dalam memprediksi terkait pasar saham up ataupun down, karena dengan model algoritma Bidirectional Long Short Term Memory yang dapat menghubungkan dua hidden layer LSTM yang terdapat pada Bidirectional Long Short Term Memory ke hasil keluaran yang sama dengan arah yang berlawanan dari dua hidden layer tersebut. Model yang dibangun pada penelitian ini juga menggunakan variasi dari beberapa layer LSTM dan layer BiLSTM pada arsitektur model. Improvement model dilakukan dengan melihat dari beberapa hasil model algoritma konvensional machine learning yang dilakukan pada penelitian ini dan pada penelitian sebelumnya yang hanya menggunakan algoritma model Deep Learning Long Short Term Memory.

II. METODE DAN MATERI

Penelitian ini menggunakan dataset bersifat publik yang tersedia pada website Kaggle yaitu Daily News for Stock Market [5]. Dataset berisi teks data berita yang dipertimbangkan untuk setiap tanggal hanya 25 peringkat berita utama dan teks data saham Dow Jones Industrial Average (DJIA) yang digunakan untuk membuktikan konsep apakah data berita dapat menentukan saham tersebut up ataupun down. Data berita dengan rentang waktu yang dimiliki dari 08-06-2008 sampai dengan 07-01-2016 dan data saham dalam rentang waktu 08-08-2008 sampai dengan 07-01-2016. Dataset terdiri dari 27 features yaitu feature pertama date, feature kedua label, dan features ketiga sampai dengan features 27 merupakan headlines dari peringkat Top1 hingga Top25. Bagian ini memberikan pedoman bagi penulis tentang elemen-elemen penulisan dan ilustrasi saat menyiapkan naskah.

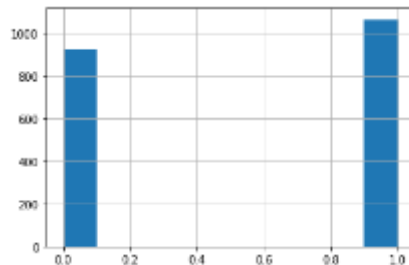
2.1. Eksplorasi Data Analisis

Proses ini melihat dataset secara lebih luas dimulai dari menampilkan dimensi data untuk memperlihatkan jumlah baris dan kolom, melihat tipe data, memeriksa data kosong (null) hingga melihat histogram distribusi label data.



DOI: 10.52362/jisicom.v6i1.772

Ciptaan disebarluaskan di bawah [Lisensi Creative Commons Atribusi 4.0 Internasional](https://creativecommons.org/licenses/by/4.0/).



Gambar 1. Histogram Distribusi Label

2.2. Preprocessing

Pada proses ini melakukan Penambahan feature headline adalah satu feature yang dihasilkan dari penggabungan semua kolom headline dari kolom Top 1 sampai Top 25 yang berguna untuk diolah kedalam model Deep Neural Network. Pada tahap ini akan diperlihatkan visualisasi data baik keseluruhan maupun perkelas, hal ini berguna untuk melihat posisi data dalam bentuk grafik Pada gambar histogram diatas menggambarkan rentang data dengan panjang data di tiap tiap frekuensi datanya, dengan panjang 1500 kata sebanyak 30 data, panjang 2000 kata sebanyak 90 data, panjang 2500 kata sebanyak 430 data, panjang 3000 kata sebanyak 440 data, panjang data 3500 kata sebanyak 200 data, panjang 4000 kata 30 data dan sisanya 4500 kata sebanyak 10 data. Pembuangan top headline dari top 1 sampai top 25 dilakukan karena sudah tidak digunakan lagi fiturnya, oleh sebab itu dibuang dan menggunakan 1 fitur headline hasil penggabungan top 1 sampai top 25 dan ditambahkan fitur `comment_length` untuk melihat total kata didalam fitur headline.

Tabel I. Hasil Pengolahan PreProcessing

<i>No.</i>	<i>Date</i>	<i>Label</i>	<i>Headline</i>	<i>Comment Length</i>
0	8/8/2008	0	b'Georgia 'downs two Russian warplanes' as cou...	2390
1	8/11/2008	1	b'Why wont America and Nato help us? If they w...	1691
2	8/12/2008	0	b'Remember that adorable 9-year-old who sang a...	2185
3	8/13/2008	0	b' U.S. refuses Israel weapons to attack Iran:...	2076





4	8/14/2008	1	b'All the experts admit that we should legalis...	1859
---	-----------	---	---	------

2.3. Text Preprocessing

Proses yang dilakukan dalam text preprocessing dimulai dari proses perubahan karakter menjadi *lower case*, kemudian melakukan perubahan kata singkat menjadi kata sebenarnya menggunakan *contractions*, menghapus angka atau *remove number*, melakukan *remove single character*, menghapus *punctuation* serta *word tokenize* dan yang terakhir melakukan *stop word*.

Tabel 2. Hasil Text Preprocessing

No.	Headline
0	[georgia, downs, two, russian, warplanes, coun...
1	[america, nato, help, us, help, us, help, iraq...
2	[remember, adorable, yearold, sang, opening, c...
3	[refuses, israel, weapons, attack, iran, repor...
4	[experts, admit, legalise, drugs, war, south, ...

2.4. Train Test Splitting

Hasil pembagian data latih dan data uji yang dilakukan dengan pembagian data latih dengan rentan waktu kurang dari 1 januari 2015 dan data uji lebih dari 31 Desember 2014. Dihasilkan pembagian kelas sebagai berikut :

a. Data Latih

Setelah dilakukan pembagian data , untuk data latih dijabarkan dua kelas, yaitu 0 dan 1, dan didalam data latih kelas 0 terdapat 868 baris data dan kelas 1 terdapat 992 baris data.

Tabel 3. Pembagian Data Latih

Kelas 0	Kelas 1
868	992

b. Data Uji

Setelah dilakukan pembagian data , untuk data uji dijabarkan dua kelas, yaitu 0 dan 1, dan didalam data latih kelas 0 terdapat 186 baris data dan kelas 1 terdapat 192 baris data.

Tabel 4. Pembagian Data Uji

Kelas 0	Kelas 1
186	192





2.5. *Word Embedding*

Word embedding proses membuat vektorisasi dari kata menjadi angka dengan salah satu fungsi encoder dimana hasil dari pengolahan encodingnya menjadi seperti berikut ini :

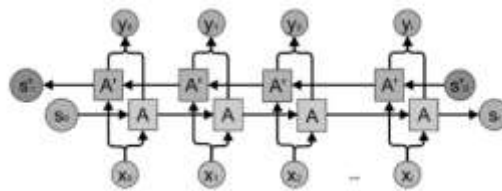
Tabel 5. Hasil Pengolahan Word Embedding

Text	Vector
abusing	0.9806552529335022
murdering	0.9793219566345215
minor	0.9777423739433289
hanged	0.9731205105781555
adultery	0.9714149832725525
repeatedly	0.970407247543335
rapist	0.9695200324058533
nurse	0.9691618084907532
assaulted	0.9691005945205688

2.6. BiLSTM

BiLSTM (*Bidirectional Long Short Term Memory*) adalah perpanjangan dari model LSTM Ada dua lapisan yang prosesnya saling berlawanan. Model ini sangat cocok untuk mengidentifikasi pola dalam kalimat, karena setiap kata dalam dokumen diproses secara berurutan, karena jika Anda mempelajari setiap kata secara berurutan, Anda dapat memahami tweet. Lapisan di bawahnya bersifat maju (*forward*), yaitu pemahaman dan pengolahan dari kata pertama hingga kata terakhir, sedangkan lapisan di atasnya bersifat mundur (*backward*), yaitu dari kata terakhir hingga kata pertama Memahami dan menangani dengan. Dengan dua lapisan yang berlawanan arah ini, model dapat memahami dan melihat melalui kata sebelumnya dan kata pemandu, sehingga membuat proses pembelajaran lebih dalam, yang berdampak pada pemahaman model yang lebih baik tentang konteks dalam tweet gambar [6]. BiLSTM ekstensi dari model LSTM di mana dua LSTM diterapkan pada data input. BiLSTM diterapkan pada urutan input (*forward layer*) untuk layer pertama dan layer keduanya merupakan bentuk kebalikan dari urutan input dimasukkan ke dalam model LSTM (*backward layer*). Penereapan BiLSTM mengarah untuk meningkatkan *learning* dalam jangka panjang dan dengan demikian akan meningkatkan akurasi model [7] BiLSTM. Sebagai varian dari LSTM [6], BiLSTM dapat menangkap konteks informasi yang lebih komprehensif dan korelasi hubungan antar konteks. Dua jaringan LSTM, satu dengan arah maju dan yang lainnya dengan arah mundur, terhubung ke lapisan output yang sama. Keduanya adalah dilatih dengan urutan data yang sama, ada tiga gerbang, yaitu, gerbang input, gerbang lupa, dan gerbang keluaran, dalam satuan LSTM [8]

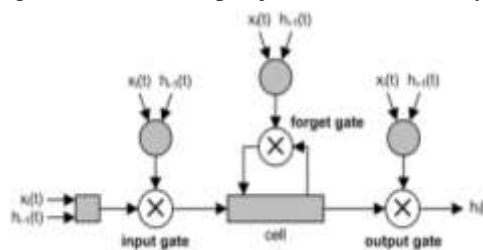




Gambar 2. Arsitektur BiLSTM

2.7. LSTM

LSTM (*Long Short Term Memory*) adalah arsitektur jaringan saraf berulang spesifik yang dirancang untuk memodelkan urutan temporal. LSTM memiliki dependensi jarak jauh yang membuat LSTM lebih akurat daripada RNN konvensional. Tidak seperti RNN, LSTM berisi unit khusus yang disebut blok memori di lapisan tersembunyi berulang. Blok memori berisi sel memori dengan koneksi sendiri yang menyimpan status temporal jaringan di samping unit perkalian khusus yang disebut gerbang untuk mengontrol aliran informasi. Selain itu, arsitektur LSTM modern berisi koneksi lubang intip dari sel internalnya ke gerbang sel yang sama untuk mempelajari waktu keluaran yang tepat [9].



Gambar 3. Arsitektur LSTM

2.8. Non Deep Learning

Arsitektur non deep learning yang digunakan pada penelitian ini terdiri dari KNN, SVM, Logistic Regression, Naïve Bayes Multinomial, dan Random Forest, tujuan di sertakan metode ini untuk menguji seberapa besar performa model Deep Learning yang diujikan terhadap dataset yang digunakan .

III. PEMBAHASAN DAN HASIL

3.1. Hasil Arsitektur BiLSTM

Pada model BiLSTM (Bidirectional Long Short Term Memor) yang dilakukan dari berbagai variasi parameter mulaidari variasi *Epoch* mulai dari 0.00005, 0.00003 dan 0.00002 serta *Epoch* sebesar 20 dan *batch size* 16 dan variasi layer 1 sampai 5 didapatkan hasil yang optimal pada variasi *Learning Rate* 0.00005 serta *Epoch* 20, *batch size* 16 dan layer LSTM 4 dengan *neuron* 256 mendapatkan nilai *Accuracy* sebesar 0.8677 dimana nilai ini merupakan nilai tertinggi dari variasi hyperparameter tuning yang dilakukan dalam penelitian dengan model BiLSTM

Tabel 6. Hasil Pengolahan Arsitektur BiLSTM

Layer	Learning Rate	Epoch	Batch Size	Accuracy	F1 Score	AUC
1	0.00005	20	16	0.8677	0.8674	0.8671
1	0.00003	20	16	0.8518	0.8515	0.8513





1	0.00002	20	16	0.7936	0.7936	0.7939
2	0.00005	20	16	0.8412	0.8411	0.8418
2	0.00003	20	16	0.8201	0.8201	0.8203
2	0.00002	20	16	0.7989	0.7988	0.7987
3	0.00005	20	16	0.8174	0.8172	0.8182
3	0.00003	20	16	0.8465	0.8465	0.8468
3	0.00002	20	16	0.8280	0.8291	0.8294
4	0.00005	20	16	0.8280	0.8295	0.8298
4	0.00003	20	16	0.8439	0.8438	0.8444
4	0.00002	20	16	0.8253	0.8253	0.8254
5	0.00005	20	16	0.8253	0.8253	0.8257
5	0.00003	20	16	0.8280	0.8276	0.8289
5	0.00002	20	16	0.8253	0.8263	0.8274

Dari pemaparan hasil pada table diatas didapatkan bahwa eksperimen dengan model BILSTM menunjukkan semakin banyaknya jumlah layer BILSTM dan learning rate yang semakin besar tidak menentukan besarnya nilai matriks evaluasinya dari accuracy, f1 score hingga AUC. namun jumlah layer yang lebih sedikit dan diimbangi nilai Learning Rate yang tinggi menghasilkan Accuracy yang besar, eksperimen ini meningkatkan *Accuracy* dari baseline yang semula bernilai 0.5212, F1 Score 0.6762 serta AUC 0.6211 menjadi 0.8677 Accuracy dan F1 Score sebesar 0.8674 serta AUC sebesar 0.8671. terlihat hubungan antar layer yang semakin banyak dengan optimizer ADAM tidak membuat tingkat akurasi meningkat.

3.2. Hasil Arsitektur LSTM

Pada model LSTM (*Long Short Term Memory*) yang dilakukan dari berbagai variasi parameter mulaidari variasi *Epoch* mulai dari 0.00005, 0.00003 dan 0.00002 serta *Epoch* sebesar 20 dan *batch size* 16 dan variasi layer 1 sampai 5 didapatkan hasil yang optimal pada variasi *Learning Rate* 0.00005 serta *Epoch* 20, *batch size* 16 dan 5 layer LSTM dengan *neuron* 256 mendapatkan nilai *Accuracy* sebesar 0.8544 dimana nilai ini merupakan nilai tertinggi dari variasi hyperparameter tuning yang dilakukan dalam penelitian dengan model LSTM.

Tabel 7. Hasil Pengolahan Arsitektur LSTM

Layer	Learning Rate	Epoch	Batch Size	Accuracy	F1 Score	AUC
-------	---------------	-------	------------	----------	----------	-----





1	0.00005	20	16	0.8492	0.8491	0.8492
1	0.00003	20	16	0.8333	0.8333	0.8335
1	0.00002	20	16	0.8121	0.8121	0.8121
2	0.00005	20	16	0.8333	0.8332	0.8339
2	0.00003	20	16	0.8359	0.8359	0.8362
2	0.00002	20	16	0.8227	0.8223	0.8236
3	0.00005	20	16	0.8465	0.8465	0.8468
3	0.00003	20	16	0.8412	0.8412	0.8413
3	0.00002	20	16	0.8571	0.8585	0.8584
4	0.00005	20	16	0.8439	0.8438	0.8439
4	0.00003	20	16	0.8280	0.8280	0.8282
4	0.00002	20	16	0.8253	0.8253	0.8256
5	0.00005	20	16	0.8544	0.8544	0.8550
5	0.00003	20	16	0.8280	0.8277	0.8288
5	0.00002	20	16	0.8306	0.8306	0.8309

Dalam eksperimen ini menunjukkan bahwa posisi layer yang semakin banyak perlu diimbangi learning rate yang tinggi agar hasil akurasi meningkat , eksperimen ini meningkatkan *Accuracy* dari yang semula bernilai 0.5212, F1 Score 0.6762 serta AUC 0.6211 menjadi 0.8544 *Accuracy* dan F1 Score sebesar 0.8544 serta AUC sebesar 0.8550.

3.3. Hasil Arsitektur Non Deep Learning

hasil dari eksperimen dari arsitektur non deep learning yang terdiri dari KNN, SVM , Random Forest, Naïve Bayes serta Logistic Regression nilai *Accuracy* dari eksperimen tertinggi ada pada algoritma Naïve Bayes dengan tipe Multinomial sebesar 0.5385 dan selanjutnya pada Random Forest sebesar 0.5257 ,pada KNN sebesar 0.5226 dan terakhir pada Logistic Regression sebesar 0.4994

Tabel 8. Hasil Model Non Deep Learning

Model	Accuracy
KNN	0.5226
SVM	0.5367





Random Forest	0.5257
Naïve Bayes Multinomial	0.5385
Logistic Regression	0.4652

IV. KESIMPULAN

Dari hasil penelitian yang sudah dilakukan, setelah dilakukan komparasi model konvensional machine learning dan model deep learning diantaranya model algoritma BiLSTM dan LSTM untuk deep learning dengan variasi tuning 1-5 Layer LSTM dan BiLSTM pada 256 Neuron, Layer Dense 128 Neuron, Dropout Layer 0.2, variasi nilai Learning Rate 0.00005, 0.00003, dan 0.00002, Epoch 20, Batch Size 16, Optimizer ADAM serta Activation RELU. Dimana model algoritma KNN, SVM, Logistic Regression, random forest, serta Naïve Bayes Multinomial untuk algoritma konvensional. Didapatkan hasil terbaik pada model dengan algoritma BiLSTM dengan 1 Layer Accuracy sebesar 0.8677 F1 Score 0.8674 serta AUC sebesar 0.8671 pada learning rate 0.00005. Dilanjutkan hasil terbaik LSTM mendapatkan nilai Accuracy sebesar 0.8671, F1 Score 0.8585 serta AUC sebesar 0.8584 dengan 3 Layer LSTM dengan learning rate 0.00002. dan nilai tertinggi dari komparasi model konvensional yaitu Naïve Bayes Multinomial yang mendapatkan Accuracy sebesar 0.5385. Pada model deep Learning dapat disimpulkan bahwa besarnya nilai Accuracy ditentukan oleh variasi learning rate dan jumlah layer nya, dan hasil eksperimen dari penelitian ini sudah membuktikan bahwa arsitektur Deep Learning dan Konvensional telah mendapatkan pembaruan dari baseline pada rujukan utama pada penelitian ini yang hanya mendapatkan nilai Accuracy dengan 0.5212, F1 Score 0.6762 serta AUC 0.6211.

REFERENSI

- [1] A. Wulandari, "Perbandingan Klasifikasi Pergerakan Harga Saham Pt. Astra Internasional Tbk Menggunakan Vector Auto Regressive (var) Stasioner Dan Logistic," *eProceedings Eng.*, vol. 7, no. 1, pp. 2614–2626, 2020.
- [2] B. Beers, "How the News Affects Stock Prices," 2021. <https://www.investopedia.com/ask/answers/155.asp>.
- [3] N. C. C. A. Phitaloka, "Web Content Mining Di Sektor Perbankan Pada Lq45 Untuk Pendukung Keputusan Investasi Saham," *Telematika*, vol. 16, no. 1, p. 18, 2019, doi: 10.31315/telematika.v16i1.2989.
- [4] M. V. Sukriti Jaitly, "IRJET- Forecasting Stock Market Trends using News Headline Analysis," *Irjet*, vol. 8, no. 9, pp. 1141–1144, 2021.
- [5] Aaron7sun, "Daily News for Stock Market Prediction," 2016. <https://www.kaggle.com/aaron7sun/stocknews/data>.
- [6] H. F. Fadli, "Identifikasi Cyberbullying pada Media Sosial Twitter Menggunakan Metode LSTM dan BiLSTM," 2019.
- [7] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," *Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019*, pp. 3285–3292, 2019, doi: 10.1109/BigData47090.2019.9005997.
- [8] H. Zheng, H. Wang, and J. Chen, "Evolutionary Framework with Bidirectional Long Short-Term Memory Network for Stock Price Prediction," *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/8850600.
- [9] A. G. Salman, Y. Heryadi, E. Abdurahman, and W. Suparta, "Single Layer & Multi-layer Long Short-Term Memory (LSTM) Model with Intermediate Variables for Weather Forecasting," *Procedia Comput. Sci.*, vol. 135, pp. 89–98, 2018, doi: 10.1016/j.procs.2018.08.153.

