Jurnal TRANSFORMATIKA
Vol.20, No.2, Januari 2023, pp. 23 - 29
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

Perbandingan LSTM dengan *Support Vector Machine* dan Multinomial *Naïve Bayes* pada Klasifikasi Kategori *Hoax*

Puji Winar Cahyo[1*], Ulfi Saidata Aesyi [2]
[1] Program Studi Informatika, FTTI, Universitas Jenderal Achmad Yani Yogyakarta
Jl. Siliwangi Ringroad Barat Banyuraden Gamping, Sleman, Jl. Siliwangi Ringroad Barat Banyuraden Gamping, Sleman, e-mail: pwcahyo@unjaya.ac.id
[2] Program Studi Sistem Informasi, FTTI, Universitas Jenderal Achmad Yani Yogyakarta
Jl. Siliwangi Ringroad Barat Banyuraden Gamping, Sleman, Jl. Siliwangi Ringroad Barat Banyuraden Gamping, Sleman, e-mail: ulfiaesyi@gmail.com

ARTICLE INFO

History of the article :
Received 30 Oktober 2022
Received in revised form 15 Desember 2022
Accepted 3 Januari 2023
Available online 30 Januari 2023

Keywords:

svm; lstm; multinomial naïve bayes; klasifikasi; deep learning

*** Correspondece:**

Telepon:
+62 8562636509

E-mail:
pwcahyo@unjaya.ac.id

ABSTRACT

Hoax is fake news, now massively spread through social media. The impact of hoaxes is that people's misperceptions in understanding of news are very high. With the existence of hoaxes are spreading through social media, it requires the public to think smart when receiving the news. Currently, many ways to prevent hoaxes, right now we have Fact Checker Directory Platform which is a truth platform sourced from several fact check sites. On the truth check platform, every news detected as hoaxes has been categorized into specific type of hoax, manually by the validator. For this reason, this research attempts to automatically categorize the types of hoaxes using comparation of Deep Learning with Machine Learning classifications. Deep Learning uses Long Short Term Memory Network (LSTM), while Machine Learning uses Support Vector Machine (SVM) and Multinomial Naive Bayes. Through the build model process, SVM produces the best accuracy quality of 0.74, Multinomial Naïve Bayes produces an accuracy quality of 0.62 while LSTM displays 0.49. The results of low accuracy in LSTM need to be evaluated on model architecture and data normalization during preprocessing.

1. PENDAHULUAN

Hoax merupakan berita palsu yang kini masif tersebar melalui media sosial [1] melalui *hoax* tersebut masyarakat mudah terkecoh dengan pemberitaan, sehingga kesalahan persepsi dalam memahami berita menjadi sangat tinggi. Dengan adanya *hoax* yang tercipta dan tersebar melalui media sosial menuntut masyarakat untuk berfikir cerdas dalam menerima berita yang tersebar. Masyarakat yang cerdas akan lebih selektif serta berusaha memastikan berita yang diterima dan

disebarkan kembali adalah benar merupakan fakta [2]. Saat ini upaya penangkalan *hoax* sudah banyak dilakukan, diantaranya dengan dibuatnya situs cek kebenaran berita turnbackhoax.id yang merupakan milik komunitas Masyarakat anti fitnah indonesia (Mafindo), halaman laporan isu *hoax* yang diinisiasi oleh Kementerian Komunikasi dan Informatika, halaman cek fakta yang diinisiasi oleh berita *online* liputan6. Situs cek kebenaran yang dibuat rata-rata menyajikan data berita yang telah dikonfirmasi atas kebenarannya. Konfirmasi kebenaran yang dilakukan menggunakan cara yang masih *manual* yaitu dengan cara *validator* kebenaran membaca judul, isi, mencari sumber berita yang ada dan membandingkan berita yang tersebar dengan berita lain yang hampir mirip. Apabila berita tersebut terbukti merupakan *hoax* atau berita palsu maka *validator* akan melakukan konfirmasi pada situs cek kebenaran tersebut.

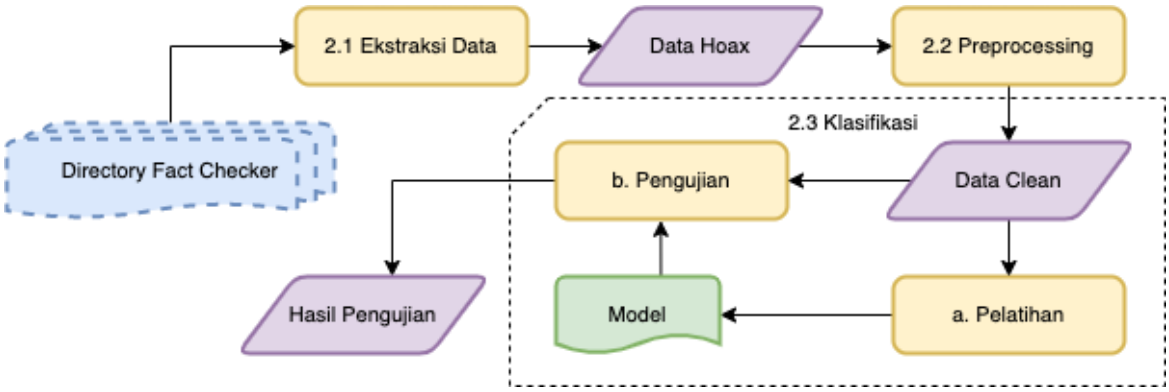
Penelitian terkait *hoax* saat ini sudah banyak dilakukan, diantaranya dengan dibuatnya Platform Directory Fact Checker yang digunakan untuk cek kebenaran yang merujuk pada situs turnbackhoax dan kominfo.go.id, hasil penelitian yang dilakukan pada tahun 2020 menunjukan hasil bahwa dalam tahun tersebut banyak *hoax* yang tersebar dengan peringkat pertama adalah kategori *fabrikasi content* (berita 100 persen berita palsu yang dibuat-buat) dan kategori kedua adalah *misleading content* (pembahasan berita yang dipelintir isinya dengan tujuan untuk mendiskreditkan) [3]. Disamping itu, penelitian terkait analisis dilakukan dengan menggunakan metode kulitatif melalui pengambilan data penyebaran kuesioner pada orang dewasa untuk mengetahui kewaspadaan dan respon terhadap *hoax*. Hasil penelitian yang dilakukan membuktikan bahwa orang dewasa sudah cukup waspada terhadap sebaran *hoax* di media sosial dengan perilaku tidak dengan sembarangan menyebarkan *hoax* di media sosial [4]. Untuk memudahkan deteksi *hoax* yang tersebar melalui media sosial maka dibutuhkan penelitian terkait deteksi berita *hoax*, penelitian menggunakan metode klasifikasi menggunakan metode Rocchio untuk deteksi *hoax* menghasilkan tingkat akurasi 83.5 persen lebih tinggi dibanding dengan metode Multinomial Naïve Bayes yang mencapai akurasi 63.83 persen [5]. Klasifikasi berita *hoax* lainnya dilakukan dengan menggunakan metode Support Vector Machine (SVM) menggunakan *kernel linier* dengan skenario pengujian 80:20 menghasilkan nilai akurasi cukup tinggi sebesar 97.06 persen. Penelitian deteksi *hoax* pada tweet covid-19 menggunakan Long Short Term Memory Network (LSTM) model *deep learning* menggunakan Convolutional Neural Network (CNN) dengan fitur ekstraksi menggunakan Word2Vec menghasilkan tingkat akurasi 79.71 persen yang dilakukan pada 1000 berita *hoax* dan bukan *hoax* [6].

Penelitian sebelumnya telah banyak membahas cara deteksi *hoax* melalui metode klasifikasi data teks, mulai dari penggunaan metode secara kualitatif sampai penelitian dengan menggunakan pembelajaran mesin. Pada klasifikasi *hoax* yang telah dilakukan belum sampai lebih mendetail pada kategorisasi jenis berita *hoax* yang tersebar, seperti yang telah diketahui saat ini bahwa terdapat beberapa kategori jenis hoax yang telah dikenal oleh masyarakat diantaranya adalah satir / parodi : tidak ada niat jahat, namun bisa mengecoh, *false connection*: judul berbeda dengan isi berita, *false context*: konten disajikan dengan narasi konteks yang salah, *misleading content*: konten dipelintir untuk menjelekkan, *imposter content*: tokoh publik dicatut namanya, *manipulated content*: konten yang sudah ada diubah untuk mengecoh, *fabricated content* : 100% konten palsu [3]. Disamping melakukan validasi berita, validator cek kebenaran berita masih menggunakan cara *manual* untuk melakukan pelabelan terhadap kategori jenis berita yang tersebar. Oleh karena itu penelitian ini berusaha untuk membuat model pembelajaran mesin untuk klasifikasi secara otomatis kategori jenis *hoax* dengan menggunakan Long Short Term Memory Network (LSTM) dengan model *deep learning* Recurrent Neural Network (RNN).

2. METODOLOGI

Penelitian yang dilakukan lebih berfokus pada analisis terhadap proses klasifikasi data teks dari Platform Directory Fact Checker yang merupakan *platform* cek kebenaran dari kumpulan situs cek

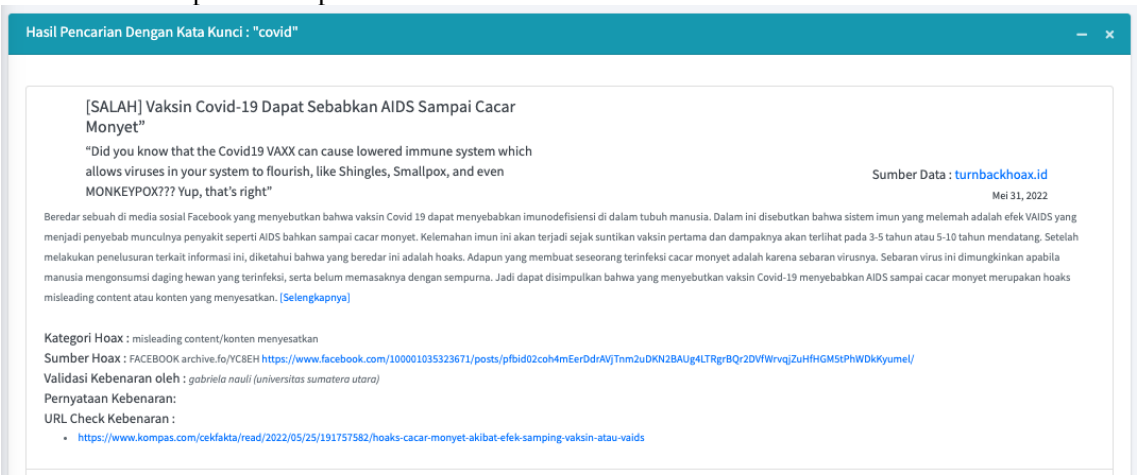
kebenaran. Beberapa tahapan yang dilakukan pada penelitian ini dapat dilihat pada Gambar 1. Dapat dilihat pada Gambar 1, data *hoax* di hasilkan dari beberapa situs cek kebenaran melalui proses ekstraksi data. Setelah didapatkan data *hoax* kemudian di lakukan *preprocessing* dengan tujuan membersihkan data agar siap untuk digunakan pada proses pelatihan dan pengujian [7], secara detail dari tahapan tersebut dapat dijelaskan pada pembahasan berikut ini.



Gambar 1. Tahapan Penelitian

2.1 Ekstraksi Data

Tahap ekstraksi data merupakan proses dimana data diambil dari halaman situs [8]. Ekstraksi data yang digunakan pada penelitian ini adalah data yang diambil dari berbagai situs cek kebenaran melalui Platform Directory Fact Checker [3]. Data yang diambil berjumlah 2000 data *hoax* dengan kategori yang sudah tersedia pada masing-masing data, kemudian terbagi menjadi 500 data kategori *hoax fabricated content*, 500 data kategori *hoax manipulated content*, 500 data kategori *hoax misleading content*, 500 data kategori *hoax false context*. Contoh data *hoax* pada Platform Directory Fact Checker dapat dilihat pada Gambar 2.


Hasil Pencarian Dengan Kata Kunci : "covid"

[SALAH] Vaksin Covid-19 Dapat Sebabkan AIDS Sampai Cacar Monyet"

"Did you know that the Covid19 VAXX can cause lowered immune system which allows viruses in your system to flourish, like Shingles, Smallpox, and even MONKEYPOX??? Yup, that's right"

Sumber Data : turnbackhoax.id

Mei 31, 2022

Beredar sebuah di media sosial Facebook yang menyebutkan bahwa vaksin Covid 19 dapat menyebabkan imunodefisiensi di dalam tubuh manusia. Dalam ini disebutkan bahwa sistem imun yang melemah adalah efek VAIDS yang menjadi penyebab munculnya penyakit seperti AIDS bahkan sampai cacar monyet. Kelemahan imun ini akan terjadi sejak suntikan vaksin pertama dan dampaknya akan terlihat pada 3-5 tahun atau 5-10 tahun mendatang. Setelah melakukan penelusuran terkait informasi ini, diketahui bahwa yang beredar ini adalah hoaks. Adapun yang membuat seseorang terinfeksi cacar monyet adalah karena sebaran virusnya. Sebaran virus ini dimungkinkan apabila manusia mengonsumsi daging hewan yang terinfeksi, serta belum memasaknya dengan sempurna. Jadi dapat disimpulkan bahwa yang menyebutkan vaksin Covid-19 menyebabkan AIDS sampai cacar monyet merupakan hoaks misleading content atau konten yang menyesatkan. [Selengkapnya]

Kategori Hoax : misleading content/konten menyesatkan
Sumber Hoax : FACEBOOK archive.fo/YC8EH https://www.facebook.com/100001035323671/posts/pfbid02coh4mEerDdrAVjTnm2uDKN2BAUg4LTRgrBQr2DVfWrvqjZuHfHGMStPhWDkKyumel/
Validasi Kebenaran oleh : *gabriela nauli (universitas sumatera utara)*
Pernyataan Kebenaran:
URL Check Kebenaran :
- https://www.kompas.com/cekfakta/read/2022/05/25/191757582/hoaks-cacar-monyet-akibat-efek-samping-vaksin-atau-vaids

Gambar 2. Tahapan Penelitian

2.2 Preprocessing

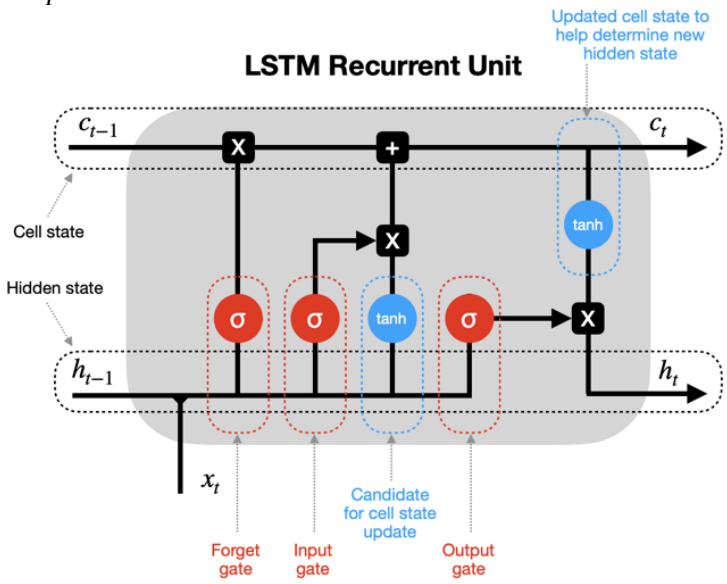
Tahap *preprocessing* merupakan proses dimana data hoax hasil dari tahap 2.1 Ekstraksi Data akan dibersihkan menjadi data bersih sehingga data siap untuk diproses pada tahap klasifikasi [9]. Langkah pada tahap *preprocessing* sangat bervariasi, tergantung kebutuhan dan kecocokan pada

proses analisis yang akan dilakukan [10]. Beberapa langkah *preprocessing* yang diterapkan pada penelitian ini, diantaranya [11].

a. Case Folding: Format huruf seluruh teks yang semula bervariasi huruf besar dan huruf kecil akan dikonversi menjadi keseluruhan huruf kecil
b. URL Removal dan Karakter Khusus: Uniform Resource Locator (URL) dan karakter khusus seperti tanda petik, *hashtag*, karakter khusus lainnya yang ada didalam berita *hoax* akan dihilangkan.
c. Tokenisasi: Pemisahan teks kedalam bagian-bagian tertentu.
d. Stopword Removal: Menghilangkan kata yang tidak dipelukan untuk proses lanjutan.
e. Slangword Conversion: Normalisasi kata apabila terdapat kata slang didalam berita *hoax*, seperti contoh: bejibun menjadi bertumpuk, bingit menjadi banget.
f. Emoticon Conversion: Normalisasi *emoticon* untuk dapat dikonversi kedalam kata, seperti contoh: ☺ menjadi senang, ☹ menjadi sedih

2.3 Klasifikasi

Tahap klasifikasi yang diterapkan pada penelitian ini menggunakan metode klasifikasi teks. Proses pembelajaran klasifikasi teks yang digunakan adalah Deep Learning (DL) dengan arsitektur Recurrent Neural Network (RNN) dan algoritma Long Short-Term Memory (LSTM). Secara umum LSTM terdiri dari *memory cell*, *input gate*, *output gate* dan *forget gate*. LSTM sangat cocok digunakan untuk klasifikasi dan melakukan prediksi pada *data time series* [12]. LSTM *cell* akan mengambil masukan dan menyimpan dalam kurun waktu tertentu kemudian *input gate* akan melakukan sampai sejauh mana nilai baru berjalan kedalam *cell*, *forget gate* akan mengontrol sejauh mana nilai akan disimpan didalam *cell*, dan *output gate* akan mengontrol sejauh mana nilai dalam *cell* digunakan untuk menghitung nilai aktivasi keluaran dari unit LSTM [13], berikut Gambar 1 merupakan arsitektur dari LSTM [14] *ht* merupakan *hidden state*, *ct* merupakan *cell state*, *xt* merupakan *input vector*.



Gambar 1. Grafik Arsitektur secara umum LSTM

Pada penelitian ini proses klasifikasi dibagi menjadi dua tahap diantaranya:

a. Pelatihan dengan menciptakan model LSTM menggunakan jumlah *hidden state* 64, penentuan *batch_size* yaitu 64, *optimizer* menggunakan *'adam'*. Proses dalam membentuk

model *weight* dan bias akan terus diperbarui hingga mendapat model yang sesuai. Pada iterasi pelatihan dilakukan proses *validation* yang untuk melihat seberapa baik model dari hasil pelatihan;

b. Pengujian dengan mengambil kembali model pembelajaran yang telah dihasilkan untuk mengetahui tingkat efektifitas LSTM yang diterapakan.

3. PEMBAHASAN

Proses klasifikasi pada penelitian ini berfokus pada 4 kategori *hoax* diantaranya adalah *fabricated content*, *manipulated content*, *misleading content* dan *false context*. Total data yang diproses pada klasifikasi ini menggunakan 2000 data *hoax* digunakan untuk proses pelatihan dan pengujian berdasarkan masing-masing *validation_split* yang digunakan. Menggunakan 500 *epoch* untuk membentuk model LSTM menghasilkan 0.34 *validation accuracy*, lebih detail proses *epoch* pembentukan model LSTM dapat dilihat pada Gambar 3.

Gambar 3. LSTM Epoch

Dapat dilihat pada Gambar 3 nilai akurasi dari awal *epoch* meningkat akan tetapi tidak signifikan, pada proses pelatihan akurasi hanya mencapai 0.49 di *epoch* 480 kemudian menurun. Pada proses *testing validation* akurasi hanya mencapai 0.37. Sehingga dapat dikatakan penggunaan LSTM untuk proses pembentukan model masih belum cukup optimal, untuk itu dibandingkan dengan klasifikasi menggunakan pembelajaran mesin Support Vector Machine (SVM) dengan *kernel 'linear'* dan Naïve Bayes Classifier sehingga menghasilkan hasil akurasi yang dapat dilihat pada Tabel 1.

Tabel 1. Hasil pengujian SVM dan Multinomial Naïve Bayes Classifier

	precision	*recall*	*f1-score*	*support*
SVM	0.74	0.74	0.74	600
Multinomial Naïve Bayes Classifier	0.64	0.63	0.62	600

DOI : http://dx.doi.org/10.26623/transformatika.v20i2.5880

Tabel 1 menunjukan bahwa tingkat akurasi *(f1-score)* untuk SVM mencapai 0.74 sedangkan tingkat akurasi *(f1-score)* untuk Multinomial Naïve Bayes Classifier mencapai 0.62. Pembentukan model dengan menggunakan pembelajaran mesin menghasilkan tingkat akurasi yang lebih tinggi dibandingkan dengan menggunakan Deep Learning yang hanya mencapai 0.37.

4. KESIMPULAN

Melalui proses penelitian yang telah dilakukan didapatkan hasil bahwa Algoritma SVM merupakan algoritma dengan tingkat akurasi paling baik untuk dapat diterapkan pada pembentukan model kategorisasi jenis *hoax* dengan capaian akurasi 0.74. Sedangkan penggunakan LSTM untuk membentuk model kategorisasi *hoax* menghasilkan akurasi yang cukup rendah hanya mencapai 0.47. Dari hasil tersebut dimungkinkan perlu evaluasi pada model arsitektur LSTM atau perlu peningkatan normalisasi data pada saat *preprocessing*.

5. ACKNOWLEDGEMENT

Peneliti mengucapkan banyak terimakasih kepada Universitas Jenderal Achmad Yani Yogyakarta atas pendanaan penelitian yang telah diberikan. Terutama Program Studi Informatika yang telah memberikan kesempatan pada peneliti untuk bisa mengikuti Skema Penelitian Internal Perguruan Tinggi. Tentunya hasil penelitian ini masih memiliki banyak kekurangan, saran dan masukan yang membangun untuk lebih baiknya penelitian ini tentunya sangat kami butuhkan.

DAFTAR PUSTAKA

[1] M. Iqbal, "Efektifitas Hukum Dan Upaya Menangkal Hoax sebagai Konsekuensi Negatif Perkembangan Interkasi Manusia," *Literasi Huk.*, vol. 3, no. 2, pp. 1–9, 2019.

[2] N. Histimuna Aisyah, "Mahasiswa Cerdas Tangkal Berita Hoaks di Era Disrupsi Melalui Literasi Digital," *J. Keislam. dan Ilmu Pendidik.*, vol. 1, no. 1, pp. 67–82, 2021.

[3] P. W. Cahyo and A. Ulfi Saidata, "Analisis Eksploratif Berita Hoax pada Situs Cek Kebenaran," *J. Inform. Univ. Pamulang*, vol. 7, no. 2, pp. 313–320, 2022.

[4] I. M. Solichin, B. N. Jati, F. Ghalib, N. A. Rakhmawati, and A. M. P. Kualitatif, "Analisis Kewaspadaan dan Respon Orang Dewasa terhadap Hoax," *J. Inf. Eng. Educ. Technol.*, vol. 06, no. 2, pp. 33–36, 2022.

[5] A. Afriza and J. Adisantoso, "Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis," *J. Ilmu Komput. dan Agri-informatika*, vol. 5, no. 1, pp. 1–10, 2018.

[6] P. N. Anggreyani and W. Maharani, "Hoax Detection Tweets of the COVID-19 on Twitter Using LSTM- CNN with Word2Vec," vol. 6, pp. 2432–2437, 2022.

[7] P. W. Cahyo, K. Kusumaningtyas, and U. S. Aesyi, "A User Recommendation Model for Answering Questions on Brainly Platform," *J. INFOTEL*, vol. 13, no. 1, pp. 7–12, 2021.

[8] P. W. Cahyo, M. Habibi, A. Priadana, and A. B. Saputra, "Analysis of Popular Hashtags on Instagram Account The Ministry of Health," in *Proceedings of the International Conference on Health and Medical Sciences (AHMS 2020)*, 2021, vol. 34, no. Ahms 2020, pp. 270–273.

[9] M. Habibi and P. W. Cahyo, "Journal Classification Based on Abstract Using Cosine Similarity and Support Vector Machine," *JISKa*, vol. 4, no. 3, pp. 185–192, 2020.

[10] Y. Hacohen-kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, pp. 1–22, 2020.

[11] P. W. Cahyo and M. Habibi, "Entity Profiling to Identify Actor Involvement in Topics of

Social Media Content," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 14, no. 4, pp. 417–428, 2020.

[12] A. Yadav, C. K. Jha, and A. Sharan, "Optimizing LSTM for time series prediction in Indian stock market," *Procedia Comput. Sci.*, vol. 167, pp. 2091–2100, 2020.

[13] A. Khumaidi, R. Raafi, and I. P. Solihin, "Pengujian Algoritma Long Short Term Memory untuk Prediksi Kualitas Udara dan Suhu Kota Bandung," *J. Telemat.*, vol. 15, no. 1, pp. 13–18, 2020.

[14] S. Dobilas, "LSTM Recurrent Neural Networks — How to Teach a Network to Remember the Past," *towardsdatascience.com*, 2022.