

## Implementasi Long Short Term Memory pada klasifikasi Teks

Hannan Asrawi

Program Studi Informatika Medis Fakultas Ilmu Komputer (Fikom) Universitas Almuslim

[Hannanasrawi05@gmail.com](mailto:Hannanasrawi05@gmail.com)

*Meskipun keduanya terkadang digunakan secara bergantian, istilah “pembelajaran mesin” dan “AI” memiliki arti yang berbeda. Tidak semua kecerdasan buatan adalah pembelajaran mesin adalah perbedaan yang signifikan di antara keduanya. Pada awalnya, machine learning dan natural language processing (NLP) saling berkaitan karena machine learning sering digunakan sebagai alat bantu untuk tugas-tugas NLP. Keuntungan dari NLP adalah dapat melakukan analisis dan memeriksa banyak data, termasuk komentar di akun media sosial dan ratusan evaluasi pelanggan online. Penelitian ini menggunakan Long Short Term Memory sebagai algoritma klasifikasi teks dengan menggunakan 18.000 dokumen newsgroup dari 20 newsgroup dari The UCI KDD Archive.*

**Kata kunci:** Long Short Term Memory, Machine Learning, NLP, Teks Klasifikasi.

### PENDAHULUAN

Pembelajaran mesin, sebuah subbidang dari kecerdasan buatan, didefinisikan sebagai kemampuan robot untuk meniru perilaku manusia yang cerdas. Sistem AI melakukan tugas-tugas kompleks dengan cara yang mirip dengan cara manusia memecahkan masalah, meskipun semua kecerdasan buatan (AI) adalah pembelajaran mesin, perbedaan yang krusial adalah tidak semua AI adalah pembelajaran mesin. Algoritme pembelajaran mesin saat ini digunakan di mana-mana. Algoritma ini membantu kita berinteraksi, membeli barang secara online, dan menggunakan media sosial dengan aman dan efisien. Karena bidang dan teknologi yang mendukung pembelajaran mesin berkembang pesat, kita baru saja mulai mempelajarinya.

Mengklasifikasikan teks ke dalam satu atau beberapa kelompok untuk mengatur, menyusun, dan menyaringnya berdasarkan kriteria apa pun adalah proses mengklasifikasikan teks. Contohnya termasuk makalah penelitian, rekam medis, dokumen hukum, dan bahkan sesuatu yang sederhana seperti evaluasi produk, karena data teks/dokumen jauh lebih umum daripada jenis data lainnya, sangat penting untuk menemukan cara-cara inovatif untuk menggunakannya.

Teknik deep learning menunjukkan keunggulannya dalam klasifikasi teks dengan memberikan hasil yang mutakhir pada sejumlah tantangan tolok ukur akademis yang umum. Klasifikasi menggunakan deep learning Deep learning adalah teknik yang mendorong komputer untuk belajar dengan melakukan, seperti yang dilakukan manusia. Deep learning, sebuah teknologi yang memungkinkan mobil swakemudi untuk mengenali rambu-rambu berhenti dan memperingatkan pejalan kaki dari tiang lampu, adalah elemen kunci dari system.

LSTM memiliki lebih banyak gerbang dan parameter daripada GRU, yang membuatnya lebih kuat dan dapat disesuaikan. Sementara GRU memiliki satu keadaan tersembunyi yang melayani kedua tujuan tersebut, yang dapat membatasi kapasitasnya, LSTM memiliki keadaan dan keluaran sel yang berbeda, yang memungkinkannya untuk menyimpan dan mengeluarkan informasi yang berbeda. Selain itu, sensitivitas LSTM dan GRU dapat berbeda terhadap hiperparameter seperti panjang urutan, laju pembelajaran, atau tingkat putus sekolah.

Jaringan *Long Short-Term Memory* (LSTM) adalah jenis jaringan saraf berulang yang mampu mempelajari ketergantungan urutan dalam masalah prediksi urutan. Ini adalah

perilaku yang diperlukan dalam domain masalah yang kompleks seperti terjemahan mesin, pengenalan suara, dan banyak lagi. LSTM adalah area pembelajaran mendalam yang kompleks. Mungkin sulit untuk memahami apa itu LSTM, dan bagaimana istilah-istilah seperti dua arah dan urutan-ke-urutan berhubungan dengan bidang ini.

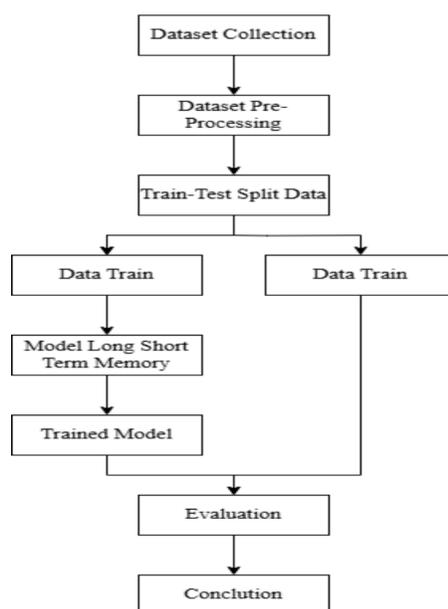
Penelitian ini menerapkan model deep learning LSTM atau *Long Short Term Memory*, untuk menentukan model mana yang lebih unggul dalam kategorisasi teks dari kedua model tersebut dan untuk menilai kinerja. LSTM dan GRU dapat dianggap sebagai variasi RNN yang efektif terlepas dari perbedaannya karena memiliki beberapa karakteristik yang sama. Keduanya menggunakan gerbang untuk mengontrol aliran data dan mengatasi masalah gradien yang hilang atau meledak. Keduanya dapat mengenali pola berurutan dalam data dan menemukan ketergantungan yang bertahan lama.

Data dari sejumlah sumber, termasuk 20 data Newsgroup, akan digunakan dalam penelitian ini. Data ini berkaitan dengan 18.000 dokumen newsgroup yang didistribusikan (hampir) seragam di antara 20 newsgroup yang berbeda. kutipan dari Arsip KDD UCI. GRU dua arah digunakan untuk mengkategorikannya lebih lanjut. Untuk menentukan mana dari dua model algoritma deep learning yang paling efektif dalam situasi ini, perbandingan dianggap penting.

Bagian selanjutnya dari esai ini disusun sebagai berikut. Studi terkait dibahas di Bagian 2, dataset dan pendekatan yang disarankan dibahas di Bagian 3, temuan eksperimental ditampilkan di Bagian 4, dan kesimpulan dan langkah selanjutnya dibahas di Bagian 5.

## METODE PENELITIAN

Dalam penelitian kami, kami menggunakan model GRU dengan mekanisme perhatian dan menyajikan model analisis teks untuk mengklasifikasikan teks berdasarkan model fundamental yang disebutkan di atas. Gambar 1 berikut ini menggambarkan struktur model tersebut:



Gambar 1. Struktur model

Dataset yang umum digunakan untuk studi pembelajaran mesin dalam aplikasi teks, seperti klasifikasi teks dan pengelompokan teks, adalah kumpulan 20 newsgroup. Dokumen ini ditulis dalam Bahasa Inggris, dan sebuah file bernama list.csv berisi referensi ke nomor document\_id dan newsgroup yang menyertainya. Data tersebut pada awalnya diproses terlebih dahulu sebelum melanjutkan ke langkah berikutnya.

Untuk mempersiapkan data teks untuk dimasukkan ke dalam model, diperlukan sebuah teknik prapemrosesan teks. Data teks terdiri dari noise dalam berbagai bentuk, termasuk emosi, tanda baca, dan teks dalam berbagai keadaan. Untuk itu, teknik tokenisasi, huruf kecil, penghilangan stopword, dan pembersihan data digunakan. Selain itu, data juga akan dikategorikan menggunakan LSTM.

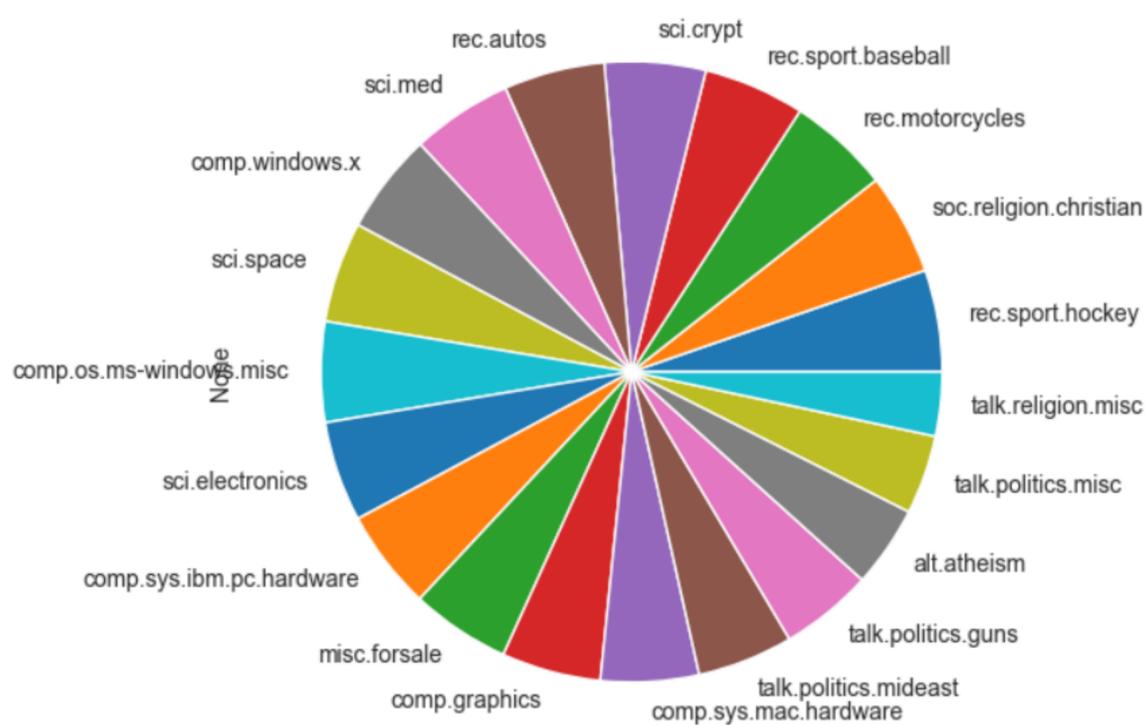
*Long Short-Term Memory (LSTM)* adalah singkatan dari jaringan memori jangka pendek yang digunakan dalam bidang *Deep Learning*. Ini adalah berbagai jaringan saraf berulang *Recurrent neural networks (RNN)* yang mampu mempelajari ketergantungan jangka panjang, terutama dalam masalah prediksi urutan.

LSTM terdiri dari empat jaringan saraf dan banyak blok memori yang dikenal sebagai sel dalam struktur rantai. Unit LSTM konvensional terdiri dari sel, gerbang input, gerbang keluaran, dan gerbang lupa. Aliran informasi ke dalam dan keluar sel dikendalikan oleh tiga gerbang, dan sel mengingat nilai dalam interval waktu yang sewenang-wenang. Algoritma LSTM diadaptasi dengan baik untuk mengkategorikan, menganalisis, dan memprediksi rangkaian waktu dengan durasi yang tidak pasti.

Epoch adalah seluruh siklus pelatihan model pembelajaran mesin menggunakan seluruh dataset pelatihan; setiap sampel pelatihan dalam dataset diproses oleh model selama satu epoch, dan bobot serta pantulan dataset dihitung oleh algoritme. Epoch juga merupakan jumlah total iterasi dari semua data pelatihan dalam satu siklus untuk melatih model pembelajaran mesin. Akurasi dan hasil kerugian dari setiap model dibandingkan, dan perbandingan tersebut kemudian digunakan sebagai evaluasi untuk menarik kesimpulan.

## HASIL DAN PEMBAHASAN

Hasil dari uji coba kategorisasi data yang telah dilakukan akan dijelaskan pada bagian ini. Aplikasi Jupyter Notebook digunakan untuk melakukan penelitian ini dengan menggunakan bahasa pemrograman Python. Dari The UCI KDD Archive, 1999an, sebuah koleksi dataset yang terdiri dari 20 newsgroup diambil. Tetapi fungsi `sklearn.datasets.fetch_20_news_groups` digunakan sebagai sumber pemrograman.



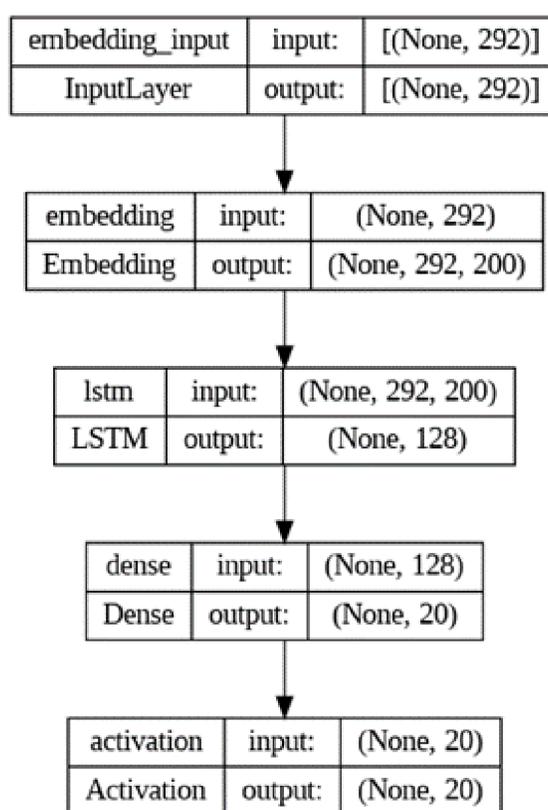
Gambar 2. Target Class Distribution

Langkah pertama dalam proses ini adalah pembersihan data, yang meliputi menghilangkan kata-kata yang berlebihan, mencari data yang hilang, melipat huruf (menghapus tanda baca), mengubah semua huruf menjadi huruf kecil, dan mengubah semua angka pada halaman menjadi huruf kecil. Pengkodean memisahkan teks menjadi unit-unit kecil yang disebut token untuk analisis tambahan, sementara penghapusan stopword menghilangkan kata-kata yang sering digunakan yang dianggap tidak memiliki nilai.

Metode sklearn. datasets digunakan dalam penelitian ini untuk mengakses data dari 20 newsgroup. Ada sekitar 18.000 posting tentang 20 subjek dalam 20 dataset newsgroup. Satu bagian dari posting ini ditujukan untuk pengujian (atau evaluasi kinerja), sedangkan bagian lainnya ditujukan untuk pelatihan atau pengembangan. Berdasarkan pesan yang diposting sebelum dan sesudah tanggal tertentu, set pelatihan dan pengujian dibagi.

Model LSTM pada Gambar menggunakan paradigma berurutan yang mirip dengan tumpukan lapisan standar dengan setiap lapisan memiliki satu tensor input dan satu tensor output, mirip dengan model GRU dua arah. Penanaman model GRU dua arah juga sama. Lapisan LSTM yang terdapat pada jaringan ini terdiri dari lapisan tersembunyi yang memiliki empat blok LSTM atau neuron, lapisan yang terlihat dengan satu input, dan lapisan output yang memprediksi satu nilai. Blok-blok LSTM menggunakan fungsi aktivasi sigmoid standar. LSTM selalu memiliki larik 3D sebagai lapisan inputnya. Tergantung pada argumennya, LSTM dapat menghasilkan larik 2D atau larik 3D, namun dalam hal ini, keluarannya adalah larik 2D.

Keluaran dari lapisan padat berikutnya adalah hasil kali titik dari matriks bobot, yang juga dikenal sebagai kernel, dan tensor input. Fungsi softmax juga digunakan oleh lapisan aktivasi untuk mengubah vektor nilai menjadi distribusi probabilitas. Komponen vektor output, yang bervariasi dari 0 hingga 1, dijumlahkan hingga 1. Setiap vektor ditangani dengan cara yang berbeda. Fungsi ini memberi tahu sumbu input mana yang akan digunakan dengan argumen sumbu. Softmax sering digunakan sebagai aktivasi untuk lapisan terakhir dari jaringan klasifikasi karena hasilnya dapat dianggap sebagai distribusi probabilitas. Model LSTM diilustrasikan pada Gambar 3 di bawah ini.



### Gambar 3. Ringkasan Model Long Short Term Memory

## PENUTUP

### Simpulan

Dalam klasifikasi teks yang menggunakan data dari The UCI KDD Archive, yang merupakan kumpulan data 20 newsgroup yang terdiri dari sekitar 18.000 postingan newsgroup pada 20 subjek, kami menerapkan *Long Short-Term Memory*. Hasil dari model *Long Short-Term Memory*, hasil dari 16 epoch memiliki nilai akurasi sebesar 0.9956 dan 0.7395 untuk validasi. Kehilangan 0.0022 dan kehilangan validasi 0.1099 menghasilkan nilai epoch loss.

Disarankan untuk penelitian selanjutnya menggunakan data yang lain, dan diharapkan menggunakan pengukuran evaluasi model machine learning, hal ini sangat penting untuk memastikan bahwa model machine learning tersebut memiliki performa yang baik.

### Saran

Hasil dari model LSTM yang digunakan, yang memiliki struktur sebagai berikut:

Model sekuensial menggunakan lapisan LSTM yang menyematkan setiap bilangan bulat ke dalam sebuah vektor dan kemudian memproses urutan vektor tersebut. Lapisan LSTM yang berukuran 128 unit. Setiap lapisan berulang di Keras menyertakan dua argumen yang berkaitan dengan dropout dengan dropout berulang sebesar 0.1: dropout, sebuah float yang mendefinisikan tingkat dropout untuk unit input lapisan, dan recurrent\_dropout, yang mendefinisikan tingkat dropout dari unit-unit yang berulang. Klasifikasi dilakukan dengan menggunakan lapisan padat dengan 20 unit dan fungsi aktivasi softmax sebagai tambahan.

Hasil percobaan memberikan penjelasan sebagai berikut:

Model memiliki 16 iterasi pelatihan, yang disebut epoch. Percobaan dilakukan sama dengan model sebelumnya dimana informasi mengenai loss function, akurasi pada data training, loss function, dan akurasi pada data validasi (dalam hal ini, val\_loss dan val\_accuracy) ditampilkan pada setiap epoch. Informasi ini dijelaskan pada bagian hasil pada tabel 3 dan 4. Evaluasi model pada data uji dilakukan pada akhir epoch ke-16, dan hasilnya menunjukkan akurasi sebesar 0.9956 dan loss function sebesar 0.0022.

Batasan dari penelitian ini hanya mempertimbangkan hasil pada epoch tersebut, yang kemudian dibandingkan antara kedua model untuk mendapatkan kesimpulan. Evaluasi kinerja dalam masalah klasifikasi biner dan masalah klasifikasi multikelas, bagaimanapun, dapat ditambahkan sebagai saran untuk penelitian tambahan.

## DAFTAR PUSTAKA

- Adriono, E., Somantri, M., & Suryono, C. A. (2022). Model Prediksi Jumlah Pakan menggunakan Algoritma Evolusi Pikiran - Jaringan Syaraf Tiruan Rambat Balik untuk Budidaya Udang. *Jurnal Kelautan Tropis*, 25(2), 266–278. <https://doi.org/10.14710/jkt.v25i2.14256>
- Agustiningsih, K. K., Utami, E., Muhammad, O., & Alsyabani, A. (2022). Sentiment Analysis of COVID-19 Vaccines in Indonesia on Twitter Using Pre-Trained and Self-Training Word Embeddings. 1, 39–46.
- Amalia, J. (2022). Membangun Slang Dictionary Untuk Normalisasi Teks

- Menggunakan Pre-Trained Fasttext Model. *JSR: Jaringan Sistem Informasi Robotik*, 6(2), 250–256. <https://doi.org/10.58486/jsr.v6i2.184>
- Dicky Wahyu Hariyanto, W. M. (2020). *Analisis Sentimen Pada Media Sosial Twitter*. 12(1), 79–86.
- Fransiska, S., & Irham Gufroni, A. (2020). Sentiment Analysis Provider by. U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method. *Scientific Journal of Informatics*, 7(2), 2407–7658. <http://journal.unnes.ac.id/nju/index.php/sji>
- Fudholi, D. H., Zahra, A., & Nayoan, R. A. N. (2022). A Study on Visual Understanding Image Captioning using Different Word Embeddings and CNN-Based Feature Extractions. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 4(1), 91–98. <https://doi.org/10.22219/kinetik.v7i1.1394>
- Ghaleini, E. N., Koopialipoor, M., Momenzadeh, M., Sarafraz, M. E., Mohamad, E. T., & Gordan, B. (2019). A combination of artificial bee colony and neural network for approximating the safety factor of retaining walls. *Engineering with Computers*, 35(2), 647–658. <https://doi.org/10.1007/s00366-018-0625-3>
- Hadiprakoso, R. B., Qomariasih, N., & Yasa, R. N. (2021). Identifikasi Malware Android Menggunakan Pendekatan Analisis Hibrid Dengan Deep Learning. *Jurnal Teknologi Informasi Universitas Lambung Mangkurat (JTIULM)*, 6(2), 77–84. <https://doi.org/10.20527/jtiulm.v6i2.82>
- Handayani, S. F., Pratiwi, R. W., Dairoh, D., & Af'idah, D. I. (2022). Analisis Sentimen pada Data Ulasan Twitter dengan Long-Short Term Memory. *JTERA (Jurnal Teknologi Rekayasa)*, 7(1), 39. <https://doi.org/10.31544/jtera.v7.i1.2022.39-46>
- Hastomo, W., Bayangkari Karno, A. S., Kalbuana, N., Meiriki, A., & Sutarno. (2021). Characteristic Parameters of Epoch Deep Learning to Predict Covid-19 Data in Indonesia. *Journal of Physics: Conference Series*, 1933(1). <https://doi.org/10.1088/1742-6596/1933/1/012050>
- Hermanto, D. T., Setyanto, A., & Luthfi, E. T. (2021). Algoritma LSTM-CNN untuk Binary Klasifikasi dengan Word2vec pada Media Online. *Creative Information Technology Journal*, 8(1), 64. <https://doi.org/10.24076/citec.2021v8i1.264>
- Khasanah, I. N. (2021). Sentiment Classification Using fastText Embedding and Deep Learning Model. *Procedia CIRP*, 189, 343–350. <https://doi.org/10.1016/j.procs.2021.05.103>
- Qaisar, S. M. (2020). Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory. *2020 2nd International Conference on Computer and Information Sciences, ICCIS 2020*, 12–15. <https://doi.org/10.1109/ICCIS49240.2020.9257657>
- Salur, M. U., & Aydin, I. (2020). A Novel Hybrid Deep Learning Model for Sentiment Classification. *IEEE Access*, 8, 58080–58093. <https://doi.org/10.1109/ACCESS.2020.2982538>
- Susanty, M., & Sukardi, S. (2021). Perbandingan Pre-trained Word Embedding dan Embedding Layer untuk Named-Entity Recognition Bahasa Indonesia. *Petir*, 14(2), 247–257. <https://doi.org/10.33322/petir.v14i2.1164>

- Tulu, C. N. (2022). Experimental Comparison of Pre-Trained Word Embedding Vectors of Word2Vec, Glove, FastText for Word Level Semantic Text Similarity Measurement in Turkish. *Advances in Science and Technology Research Journal*, 16(4), 147–156. <https://doi.org/10.12913/22998624/152453>
- Umer, M., Ashraf, I., Mehmood, A., Kumari, S., Ullah, S., & Sang Choi, G. (2021). Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model. *Computational Intelligence*, 37(1), 409–434. <https://doi.org/10.1111/coin.12415>
- Zulqarnain, M., Ghazali, R., Ghouse, M. G., & Mushtaq, M. F. (2019). Efficient processing of GRU based on word embedding for text classification. *International Journal on Informatics Visualization*, 3(4), 377–383. <https://doi.org/10.30630/joiv.3.4.289>