# Identifikasi Berita Hoax Terkait Virus Corona Menggunakan Long Short-Term Memory

**Rani Kurnia Putri[1], Muhammad Athoillah[2,*]**

[1]Program Studi Pendidikan Matematika, Universitas PGRI Adi Buana Surabaya, Indonesia

[2]Program Studi Statistika, Universitas PGRI Adi Buana Surabaya, Indonesia

Email: athoillah@unipasby.ac.id

**Abstrak**

Coronavirus merupakan salah satu penyakit menular yang merupakan turunan dari virus SARS-CoV-2. Penyebaran virus yang begitu cepat dan masif menjadikan coronavirus seketika merubah wajah dunia dalam berbagai sektor seperti ekonomi, politik, bahkan pendidikan. Hal ini tentunya menjadikan virus ini sebagai object utama dalam berbagai headline berita. Ironisnya, dengan masifnya berita yang bermunculan tidak semua berita tersebut adalah berita yang benar. Kominfo dalam laman resminya mencatat bahwa sepanjang 2021-2022 telah ditemukan 2.154 berita hoax terkait dengan isu coronavirus. Dalam penelitian ini dibangun sebauh sistem yang mampu mengidentifikasi berita hoax atau buka hoax terkait dengan isu coronavirus dengan menggunakan algoritma *Long Short-Term Memory* (LSTM). Jaringan LSTM adalah jenis jaringan saraf berulang (Recurrent Neural Networks) yang termasuk dalam area kompleks *deep learning*, algoritma ini yang mencoba meniru cara otak manusia beroperasi dan mengungkap hubungan mendasar dalam data sekuensial yang diberikan. Hasil dari penelitian menunjukkan nilai rata-rata yang didapat adalah 51,09% untuk nilai presisi, 51,00% untuk nilai *Recall* sama dengan perhitungan hasil Akurasi dan 50,41% untuk nilai *F-Measure*. hasil ini mengindikasikan walaupun secara nilai hasilnya masih dikatakan kurang baik, namun secara konsistensi hasil identifikasi ini bisa dikatakan sangat baik jika dilihat dari nilai setiap uji coba tidak lebih dari 5 digit dari keseluruhan data yang diuji dengan skema *k-fold cross validation.*

**Kata kunci:** hoax; covid-19; *long short-term memory.*

## PENDAHULUAN

Coronavirus atau juga dikenal sebagai virus Covid-19 merupakan salah satu penyakit menular yang merupakan turunan dari virus SARS-CoV-2 dan ditemukan pertama kali pada akhir tahun 2019 di suatu daerah bernama Wuhan (Cina). Coronavirus dapat menyebar dari hidung maupun mulut orang yang telah terinfeksi dalam partikel cairan kecil ketika si penderita berbicara, batuk, bersin, bahkan saat bernapas (He et al., 2020). Penyebaran virus yang begitu cepat dan masif ini menjadikan virus corona primadona dan seketika merubah wajah dunia. Tidak hanya dalam hal kesehatan, virus ini pada akhirnya menggoncang dunia dalam berbagai sektor baik ekonomi, politik, bahkan pendidikan (Hiscott et al., 2020).

Keberadaan coronavirus yang fenomenal ini tentunya menjadikan virus ini sebagai object utama dalam berbagai headline berita elektronik maupun cetak dalam skala nasional ataupun internasional. Ironisnya, dengan masifnya skala jumlah berita yang bermunculan tidak semua berita tersebut adalah berita yang benar. Kominfo dalam laman resminya mencatat bahwa sepanjang 23 Januari 2020 – 28 Maret 2022, ditemukan 2.154 berita hoax terkait dengan isu virus corona (Kominfo, 2022). Hoax pada dasarnya berarti sebagai kabar atau berita palsu yang memiliki maksud untuk mengelabui bahkan memprovokasi pendengar atau pembacanya agar mempercayai berita palsu tersebut(Simarmata et al., 2019). Sedangkan dampak dari adanya berita hoax yang dipercayai masyarakat dapat mengakibatkan adanya kesalahpahaman, kekhawatiran sehingga dapat memicu kegaduhan dalam masyarakat. (Putri et al., 2020).

Dari penjabaran tersebut, maka jika diketahui fakta jumlah berita hoax yang tersebar begitu besar, tentunya hal ini patut untuk diwaspadai dan dihentikan semaksimal mungkin. Memberikan edukasi yang baik kepada masyarakat untuk dapat menyaring berita yang diterima dengan baik merupakan cara terbaik untuk menangkal semua dampak negatif dari berita hoax yang telah disebutkan sebelumnya (Nugraha, 2019). Adapun salah satu bentuk edukasi kepada masyarakat adalah dengan menyediakan platform khusus bagi masyarakat untuk dapat memverifikasi berita yang didapat tersebut merupakan berita hoax atau tidak. Saat ini sudah terdapat beberapa platform media yang dapat dijadikan rujukan untuk memverifikasi berita hoax atau tidak, seperti laman web kominfo milik pemerintah atau laman web *turbackhoax.id* yang dikelola secara swadaya oleh masyarakat yang tergabung dalam MAFINDO (Masyarakat anti hoax indonesia). Namun begitu, identifikasi hoax berita yang dilakukan selama ini pada umumnya dilakukan secara manual. Sebagai contoh, pada laman web *turnbackhoax.id,* identifikasi berita hoax diperoleh dari berbagai laporan masyarakat yang terkumpul dari berbagai media sosial khususnya pada media sosial facebook dalam forum bernama FAFHH (forum anti fitnah hasut dan hoax) (Meinarni & Iswara, 2018).

Pada dasarnya, dalam bidang maching learning, identifikasi berita dapat dikategorikan dalam bidang *text mining*. Ide dasar dari metode pengenalan teks ini adalah dengan

mempelajari (*train*) karakteristik atau pola dari data masukan yang telah diketahui sebelumnya label data tersebut positif (+) atau negatif (-). Dari hasil pengenalan pola tersebut kemudian terbentuklah model sistem yang selanjutnya dapat digunakan untuk mengevaluasi data baru yang masuk sebagai data positif (+) atau Negatif (-) (Žižka et al., 2019) Beberapa algoritma yang dapat digunakan dalam bidang *text mining* adalah *Support Vector Machine*, *K-Means, Naïve Bayes, Artificial Neural Network* dan *Deep Learning* (Jo, 2019; Žižka et al., 2019).

*Long Short-Term Memory (LSTM)* merupakan salah satu algoritma *deep learning* yang sering digunakan dalam berbagai penelitian beberapa tahun terakhir ini. Sebut saja Ghimire dkk (Ghimire et al., 2019) yang menggunakan metode LSTM untuk mengestimasi jumlah gelombang radiasi matahari, atau penelitian yang dilakukan oleh Krishan dkk (Krishan et al., 2019) memodelkan kualitas udara di negara india menggunakan metode yang sama (LSTM). Selain untuk setimasi, algoritma LSTM juga dapat digunakan untuk melakukan klasifikasi, termasuk identifikasi teks seperti yang dilakukan oleh Gauri dkk (Jain et al., 2019) yang mendeteksi adanya Spam atau tidak pada platform media sosial seperti facebook, twitter dll.

Berdasarkan pemaparan yang telah disebutkan, penulis melakukan penelitian untuk membangun sebuah sistem deteksi atau identifikasi berita hoax terkait dengan coronavirus secara otomatis menggunakan algoritma *Long Short-Term Memory (LSTM).* Pada penelitian yang telah dilakukan ini, data yang digunakan adalah data berupa narasi berita hoax atau bukan hoax yang dihimpun dari beberapa artikel yang tersedia dalam laman web *turnbackhoax.id.* Detail dari proses penelitian ini akan dijelaskan pada bab selanjutnya.

## METODE

### Data Penelitian

Data dari penelitian ini merupakan data teks berita yang diambil dari laman web *turnbackhoax.id* dalam rentang waktu sepanjang tahun 2021. Data tersebut merupakan data narasi berita yang dihimpun dan diverifikasi kebenarannya oleh MAFINDO dari berbabagai platform media berita online maupun berita yang beredar dalam media sosial. Selanjutnya, data yang diperoleh dibagi menjadi dua bagian yaitu data training dan data testing dengan proporsi 8:2 artinya 80% dari data tersebut digunakan untuk proses pembelajaran (pengenalan) pola sedangkan 20% sisanya digunakan untuk proses uji coba model yang telah dibentuk dari proses pembelajaran sebelumnya. Sedangkan untuk kebutuhan validasi, dalam penelitian ini digunakan metode validasi *k-fold cross validation* dimana data set yang ada dibagi menjadi 5 bagian yang terdiri dari kombinasi data training dan testing yang kemudian dari kombinasi tersebut dilakukan proses uji coba secara berulang sesuai dengan aturan *k-fold cross validation* (Zhang & Yang, 2015).

**Pra-Proses**

Untuk mendapatkan hasil yang optimal, sebelum data teks diklasifikasikan dengan model LSTM, perlu dilakukan pra-proses pada data teks yang digunakan. Dalam proses ini beberapa pra-poses yang dilakukan adalah:

(1) Tokenisasi teks yaitu proses pengubahan teks atau kalimat menjadi bagian kata per kata sehingga dapat di identifikasi data per kata tersebut.

(2) *Lowercase* yaitu proses pengubahan huruf pada teks menjadi huruf kecil semua, sehingga dapat diproses secara rata.

(3) *Erase Punctuation* yaitu proses penghapusan tanda baca pada teks sehingga tidak ada huruf tanpa makna yang terproses.

**Klasifikasi dengan Long Short-Term Memory**

Struktur data teks secara alami adalah data yang berurutan. Sepotong teks atau kalimat merupakan hasil dari urutan kata-kata, yang mungkin memiliki hubungan dan ketergantungan di antara satu kata dengan kata lainnya. Oleh karena itu, jaringan *Long Short-Term Memory* (LSTM) adalah salah satu algoritma yang tepat untuk mempelajari serta menggunakan dependensi jangka panjang untuk mengklasifikasikan data urutan. Jaringan LSTM adalah jenis jaringan saraf berulang (*Recurrent Neural Networks*) yang dapat mempelajari ketergantungan jangka panjang antara langkah-langkah waktu dari data urutan (Hochreiter & Schmidhuber, 1997).

Secara garis besar, untuk dapat mengklasifikasikan data teks dengan LSTM maka hal pertama yang dilakukan adalah dengan mengubah data tersebut menjadi urutan data. Proses transformasi ini dapat dilakukan dengan menggunakan pengkodean kata yang memetakan dokumen ke dalam urutan indeks numerik. Setelah data teks ditransformasikan, maka data ini dapat diproses dengan model LSTM. Untuk hasil yang lebih baik, sertakan juga lapisan penyisipan kata di dalam jaringan LSTM. Penyisipan ini dapat memetakan kata-kata dalam data menjadi vektor numerik daripada indeks skalar. Penyisipan ini juga menangkap detail semantik dari kata-kata, sehingga data dari kata-kata dengan arti yang sama mempunyai vektor yang serupa (Brownlee, 2017; Johnson & Zhang, 2016).

*Long Short-Term Memory* (LSTM) memproses urutan panjang variabel $x = (x_1, x_2, \cdots, x_n)$ dengan menambahkan data baru secara bertahap ke dalam satu slot memori dengan *gate* yang mengontrol urutan variabel dan sejauh mana data baru tersebut harus "dihafal", data mana yang harus "dilupakan" ataupun data mana yang harus "diekspos". Pada setiap waktu $t$, maka memori $c_t$ dan $h_t$ diperbaharui dengan persamaan berikut ini:

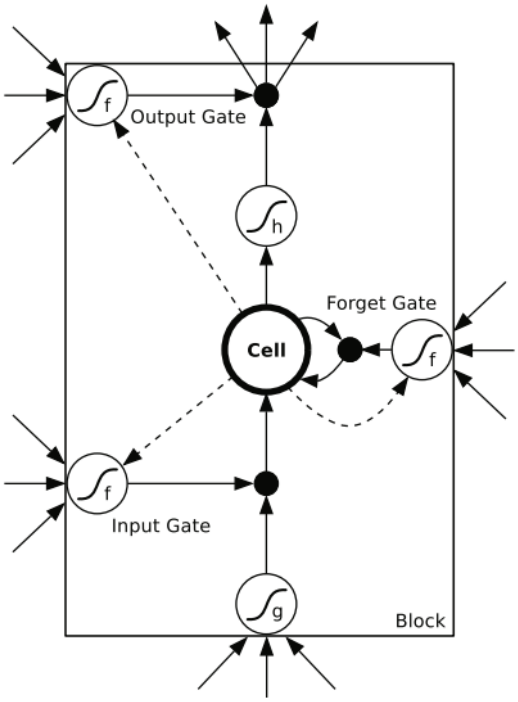
$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ c_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} W.[h_{t-1}, x_1]$$

$$c_t = f_t \odot c_{t-1} + i_t \odot c_t$$

$$h_t = o_t \odot \tanh(c_t)$$

Dimana $i$, $f$ dan $o$ merupakan *gate activation*. Dibandingkan dengan *Recurrent Neural Networks* (RNN) standar, LSTM menggunakan pembaruan memori aditif dan memisahkan memori $c$ dari keadaan tersembunyi $h$, yang berinteraksi dengan lingkungan saat membuat prediksi (Cheng et al., 2016).

Dalam arsitektur LSTM dikenal adanya blok memori yang merupakan satu set subnet yang terhubung secara berulang. Blok-blok ini dapat dianggap sebagai versi *chip* memori yang dapat dibedakan dalam komputer digital. Setiap dari blok tersebut berisi satu atau lebih sel memori yang terhubung dengan tiga unit pengganda — input serta *gate* yang menyediakan analog terus menerus dari operasi tulis, baca, dan reset untuk sel.




**Gambar 1.** Blok memori LSTM dengan satu sel (Graves, 2012)

Gambar 1 memberikan ilustrasi blok memori pada LSTM dengan satu sel. Jaringan LSTM dapat dikatakan mirip sama dengan RNN standar, hanya saja unit penjumlahan di lapisan tersembunyi digantikan oleh blok memori, Blok LSTM juga dapat dicampur dengan unit penjumlahan biasa, meskipun ini biasanya tidak diperlukan. Lapisan keluaran yang sama dapat digunakan untuk jaringan LSTM seperti untuk RNN standar. Gerbang (*gate*) perkalian memungkinkan sel memori LSTM untuk menyimpan dan mengakses informasi dalam jangka waktu yang lama, sehingga

mengurangi masalah gradien yang hilang. Misalnya, selama *input gate* tetap tertutup (yaitu memiliki aktivasi mendekati 0), aktivasi sel tidak akan ditimpa oleh input baru yang tiba di jaringan, dan oleh karena itu jaringan tersebut masih akan tetap tersedia untuk data selanjutnya(Graves, 2012).

## HASIL DAN PEMBAHASAN

Hasil dari penelitian ini disajikan dengan menunjukkan hasil dari perhitungan Presisi, *Recall*, Akurasi dan *F-Measure* yang didapatkan dari hasil uji coba sistem dengan menerapkan *k-fold cross validation.* Semua nilai tersebut didapatkan dari perbandingan hasil dari prediksi dengan dataset sebenarnya yang disebut dengan *confusion matrix* seperti yang ditunjukkan tabel 1 berikut ini:

**Tabel 1.** Confusion Matrix

| | | **Sebenarnya** | |
|---|---|---|---|
| | | **Positif (+)** | **Negatif (-)** |
| **Prediksi** | **Positif (+)** | True Positive (TP) | False Negative (FN) |
| | **Negatif (-)** | False Positive (FP) | True Negative (TN) |

Sedangkan untuk mendapatkan nilai Akurasi, Presisi, *Recall* dan *F-Measure* dapat dihitung dengan persamaan berikut ini:

$$\text{Akurasi} = \frac{\text{True Positive (TP)} + \text{True Negative (TN)}}{\text{Jumlah Keseluruhan Data}}$$

$$Recall = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

$$\text{Presisi} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

$$F - Measure = 2\,\frac{\text{Recall} \times \text{Presis}i}{\text{Presisi} + Recall}$$

Hasil dari penelitian menujukkan bahwa hasil klasifikasi berita hoax dengan menggunakan LSTM berjalan cukup baik dengan nilai rata-rata yang didapat adalah 51,09% untuk nilai Presisi, 51,00% untuk nilai *Recall* sama dengan perhitungan hasil Akurasi dan 50,41% untuk

nilai *F-Measure*. Sedangkan untuk hasil keseluruhan dari uji coba yang dilakukan sesuai skema *k-fold cross validation* yang disebutkan sebelumnya, dapat dilihat dari tabel 2 berikut ini:

**Tabel 2.** Hasil Uji Coba Keseluruhan

| No | Presisi | Recall | F-Measure | Akurasi |
|---|---|---|---|---|
| 1 | 41,43% | 41,67% | 41,26% | 41,67% |
| 2 | 56,79% | 56,67% | 56,47% | 56,67% |
| 3 | 53,75% | 53,33% | 52,00% | 53,33% |
| 4 | 51,76% | 51,67% | 51,00% | 51,67% |
| 5 | 51,71% | 51,67% | 51,33% | 51,67% |
| **Rata-rata** | **51,09%** | **51,00%** | **50,41%** | **51,00%** |

## KESIMPULAN

Penelitian ini dilakukan untuk membantu memberikan referensi tentang bagaimana membuat sistem identifikasi berita hoax atau tidak hoax terkait dengan coronavirus (covid-19) dengan menggunakan algoritma *deep learning* atau lebih tepatnya adalah *Long Short-Term Memory* (LSTM). Data yang digunakan dalam penelitian ini adalah data narasi berita terkait dengan coronavirus yang dihimpun dari laman *web turnbackhoax.id* yang selanjutnya diklasifikasn dengan menggunakan jaringan LSTM dan dilakukan pengujian secara berulang dengan mengikuti skema validasi *k-fold cross validation.* Hasil dari penelitian menunjukkan bahwa identifikasi berita hoax terkait coronavirus dengan LSTM mendapatkan hasil yang cukup baik dengan nilai rata-rata yang didapat adalah 51,09% untuk nilai Presisi, 51,00% untuk nilai *Recall* sama dengan perhitungan hasil Akurasi dan 50,41% untuk nilai *F-Measure*, hasil ini mengindikasikan walaupun secara nilai hasilnya masih dikatakan kurang baik, namun secara konsistensi maka hasil identifikasi ini bisa dikatakan sangat baik jika dilihat dari hasil keseluruhan uji coba yang mana margin antara nilai setiap uji coba tidak lebih dari 5 digit.

## UCAPAN TERIMAKASIH

Segenap tim peneliti dengan ini menyampaikan ucapakan terima kasih kepada Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) di Univ PGRI Adi Buana Surabaya atas dukungan penuhnya melalui dana yang diberikan melalui program Hibah Penelitian Penerapan Universitas PGRI Adi Buana Surabaya Tahun 2021/2022.

## DAFTAR PUSTAKA

Brownlee, J. (2017). *Long Short-Term Memory Networks With Python: Develop Sequence Prediction Models with Deep Learning*. Machine Learning Mastery. https://books.google.co.id/books?id=m7SoDwAAQBAJ

Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. *ArXiv Preprint ArXiv:1601.06733*.

Ghimire, S., Deo, R. C., Raj, N., & Mi, J. (2019). Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. *Applied Energy*, *253*, 113541.

Graves, A. (2012). Long Short-Term Memory. In A. Graves (Ed.), *Supervised Sequence Labelling with Recurrent Neural Networks* (pp. 37–45). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24797-2_4

He, F., Deng, Y., & Li, W. (2020). Coronavirus disease 2019: What we know? *Journal of Medical Virology*, *92*(7), 719–725.

Hiscott, J., Alexandridi, M., Muscolini, M., Tassone, E., Palermo, E., Soultsioti, M., & Zevini, A. (2020). The global impact of the coronavirus pandemic. *Cytokine & Growth Factor Reviews*, *53*, 1–9. https://doi.org/https://doi.org/10.1016/j.cytogfr.2020.05.010

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jain, G., Sharma, M., & Agarwal, B. (2019). Spam detection in social media using convolutional and long short term memory neural network. *Annals of Mathematics and Artificial Intelligence*, *85*(1), 21–44.

Jo, T. (2019). Text mining. *Studies in Big Data. Cham: Springer International Publishing*.

Johnson, R., & Zhang, T. (2016). Supervised and semi-supervised text categorization using LSTM for region embeddings. *International Conference on Machine Learning*, 526–534.

Kominfo. (2022). *Penanganan Sebaran Konten Hoaks Covid-19*. https://kominfo.go.id/content/detail/40830/penanganan-sebaran-konten-hoaks-covid-19-senin-28032022/0/infografis

Krishan, M., Jha, S., Das, J., Singh, A., Goyal, M. K., & Sekar, C. (2019). Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India. *Air Quality, Atmosphere & Health*, *12*(8), 899–908.

Meinarni, N. P. S., & Iswara, I. B. A. I. (2018). Hoax and its Mechanism in Indonesia. *International Conference of Communication Science Research (ICCSR 2018)*, 183–186.

Nugraha, M. T. (2019). Hoax di Media Sosial Facebook: Antara Edukasi dan Propaganda Kepentingan. *JSW: Jurnal Sosiologi Walisongo*, *3*(1), 97–108.

Putri, N. F., Vionia, E., & Michael, T. (2020). Pentingnya Kesadaran Hukum Dan Peran Masyarakat Indonesia Dalam Menghadapi Penyebaran Berita Hoax Covid-19. *Media Keadilan: Jurnal Ilmu Hukum*, *11*(1), 98–111.

Simarmata, J., Iqbal, M., Hasibuan, M. S., Limbong, T., & Albra, W. (2019). *Hoaks dan media sosial: saring sebelum sharing*. Yayasan Kita Menulis.

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112.

Žižka, J., Dařena, F., & Svoboda, A. (2019). *Text Mining with Machine Learning: Principles and Techniques*. CRC Press. https://books.google.co.id/books?id=avm7DwAAQBAJ