Jurnal Elektronik Ilmu Komputer Udayana
Volume 12, No 3. Februari 2024

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Implementasi Long-Short Term Memory (LSTM) pada Klasifikasi Kategori Berita

Anak Agung Ngurah Andhika Satrya Nugraha[a1], Ida Bagus Made Mahendra[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Bali, Indonesia
[1]andhikangurah@gmail.com
[2]ibm.mahendra@unud.ac.id

Abstrak

Pada era informasi ini, berita dan informasi baru menyebar dengan sangat cepat, terlalu banyak berita baru yang muncul setiap harinya. Oleh karena itu, diperlukan sebuah cara untuk memilah berita yang ingin dilihat. Salah satu cara untuk memilah berita adalah dengan membagi berita ke dalam beberapa kategori. Pembagian kategori pada berita masih dilakukan secara manual dengan memberi label kategori pada berita yang ingin diunggah. Penelitian ini membahas tentang implementasi metode Long-Short Term Memory (LSTM) untuk mengklasifikasikan berita berdasarkan judulnya ke dalam 7 kategori, yaitu finance, food, health, internet, otomotive, sport, dan travel. Terdapat dua model yang diimplementasikan pada penelitian ini, yaitu model LSTM dan model Bidirectional LSTM. Pengujian dilakukan dengan menggunakan nilai akurasi, yaitu perbandingan antara judul berita yang berhasil diklasifikasikan dengan keseluruhan judul berita. Berdasarkan hasil pengujian, didapatkan bahwa model LSTM yang dibuat berhasil mengklasifikasikan berita dengan akurasi sebesar 85.36%, model Bidirectional LSTM juga berhasil mengklasifikasikan berita dengan akurasi sebesar 84.15%.

Kata kunci: *LSTM, Bidirectional LSTM, Klasifikasi, Natural Language Processing, Berita*

1. Pendahuluan

Saat ini, informasi tidak dapat dipisahkan dari kehidupan manusia, sangat banyak berita dan informasi baru yang muncul setiap harinya. Karena terlalu banyaknya berita yang ada, diperlukan sebuah cara untuk memilah berita yang ingin dilihat setiap harinya, salah satu cara yang dapat digunakan adalah dengan membagi berita yang ada ke dalam beberapa kategori, sehingga pembaca dapat memilih kategori berita yang ingin dibaca.

Pembagian kategori berita biasanya dilakukan secara manual oleh pembuat berita saat berita akan diunggah, berita yang sudah diberikan kategori kemudian dapat dikelompokkan pada aplikasi dan pengguna aplikasi tersebut dapat memilih kategori berita yang ingin dibaca. Pemberian kategori berita secara manual adalah cara yang digunakan untuk mengelompokkan berita saat ini, namun terdapat kekurangan pada cara pengelompokan ini. Pengunggah berita perlu mengetahui kategori berita yang akan diunggah, pengunggah berita juga perlu mengetahui kategori berita apa saja yang tersedia pada aplikasi tersebut, ini akan membuat berita tidak dapat langsung diunggah karena kategori berita perlu ditentukan terlebih dahulu secara manual.

Oleh karena itu, diperlukan suatu cara untuk dapat mengelompokkan berita secara otomatis. Salah satu cara yang dapat digunakan untuk mengelompokkan berita secara otomatis adalah menggunakan *natural language processing* (NLP), NLP digunakan untuk memproses judul berita dan menentukan kategori berita berdasarkan judulnya. Terdapat berbagai macam metode pada NLP untuk melakukan klasifikasi teks, namun pada penelitian ini, metode *long-short term memory* (LSTM) digunakan untuk membuat model klasifikasi berita.

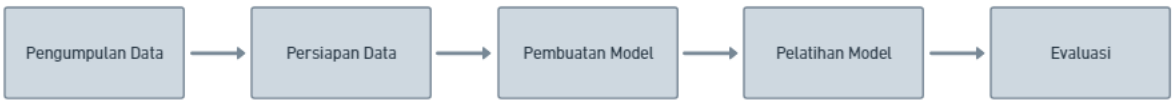
Long-short term memory (LSTM) merupakan salah satu metode yang digunakan pada *recurrent neural network* (RNN) untuk mengatasi masalah *vanishing error* [1]. Metode ini menggunakan blok memori khusus yang dapat mengingat urutan kronologis pada data input [2], dalam konteks NLP, metode ini digunakan untuk mengingat urutan kronologis dari kata-kata pada kalimat. Salah satu kelebihan dari metode ini adalah memungkinkan model *neural network* untuk mengingat urutan kata, sehingga model tidak hanya menentukan makna kalimat berdasarkan banyaknya kata yang muncul, namun juga

berdasarkan kata apa yang muncul terlebih dahulu. Bidirectional LSTM merupakan metode lanjutan dari LSTM, pada metode ini, data dimasukkan dalam urutan maju dan mundur, sehingga model juga dapat mengingat kata-kata setelah kata yang sedang diproses. Model yang sudah dilatih selanjutnya dapat diuji menggunakan *testing data*, di mana nilai akurasi dapat ditentukan dengan membandingkan jumlah berita yang berhasil diklasifikasikan dan jumlah total berita pada data *testing.*

Terdapat banyak penelitian terkait yang sudah dilakukan sebelumnya tentang klasifikasi berita, namun, kebanyakan dari penelitian yang ada membahas tentang klasifikasi berita palsu. Salah satu penelitian yang ada berhasil mengklasifikasikan berita palsu menggunakan metode Random Forest dengan akurasi sebesar 84% [3]. Penelitian ini akan berfokus pada klasifikasi berita berdasarkan kategori.

2. Metode Penelitian

Metode yang digunakan pada penelitian ini terdiri dari 5 bagian, yaitu pengumpulan data, persiapan data, pembuatan model, pelatihan model, dan evaluasi. Model yang dibuat mengimplementasikan *layer Word Embedding* untuk merepresentasikan judul berita sebagai vektor dan *layer* LSTM untuk mengklasifikasikan judul berita.



Gambar 1. Diagram Metode Penelitian

2.1. Pengumpulan Data

Dataset yang digunakan pada penelitian ini adalah dataset judul dan kategori berita yang didapatkan melalui platform Github [4]. Dataset ini berisi 91017 data yang terdiri dari tanggal *posting* dari berita. URL berita, judul berita dalam Bahasa Indonesia, dan kategori berita.

2.2. Persiapan Data

Adapun prosedur-prosedur yang dilakukan setelah mengumpulkan data yaitu:

a. Menghapus kategori “news” dan “hot”, penghapusan ini dilakukan karena kategori tersebut merupakan kategori umum yang berisi berita dari berbagai kategori
b. Mengubah semua huruf pada judul berita menjadi *lowercase*
c. Menghapus *stop words* pada judul berita
d. Mengubah data kategori menjadi vektor dengan metode *one-hot encoding*
e. Memisahkan dataset menjadi 80% *training data* dan 20% *testing data*

2.3. Tokenisasi

Setelah dataset dibagi menjadi *training data* dan *testing data*, dilakukan tokenisasi untuk mengubah setiap kata pada judul berita ke dalam bentuk numerik. 10.000 kata terbanyak pada dataset akan diberi representasi angka, sedangkan kata-kata lainnya diberikan representasi khusus (*out-of-value token*). Kemudian, judul berita dapat direpresentasikan sebagai *sequence* numerik. Panjang *sequence* yang digunakan adalah 20, untuk *sequence* dengan jumlah token kurang dari 20, nilai 0 akan ditambahkan sebagai *padding* di awal *sequence*, sedangkan untuk *sequence* dengan jumlah token lebih dari 20, dilakukan pemotongan di awal *sequence* sehingga hanya 20 token terakhir yang digunakan.

2.4. Word Embedding

Word Embedding adalah sebuah metode yang digunakan untuk mengubah *sequence* kata menjadi vektor [5]. Pada penelitian ini, *Word Embedding* digunakan sebagai layer awal pada model, setiap judul berita yang berupa sequence 20 token diubah menjadi vektor berukuran 64 nilai [6]. Vektor ini kemudian akan digunakan untuk klasifikasi di *layer* selanjutnya.

2.5. Long-Short Term Memory (LSTM)

Long-Short Term Memory merupakan metode berbasis gradient yang diciptakan untuk mengatasi masalah *vanishing error* pada model *recurrent neural network* (RNN) [1]. Pada penelitian ini, LSTM digunakan sebagai layer pada model setelah dilakukan *Word Embedding*. banyaknya unit LSTM yang digunakan adalah sebanyak 64 [6]. Setiap *unit* pada *layer* LSTM

terdiri dari 3 bagian, yaitu *forget gate*, *input gate*, dan *output gate*. *Forget gate* digunakan untuk me-*reset state internal* dari *unit* sehingga *vanishing gradient* dapat dihindari. *Input gate* digunakan untuk memasukkan data baru ke dalam *state internal* dari *unit*. *Output gate* menggunakan *state internal* dari *unit* dan nilai *input* untuk menghasilkan nilai *output* dari *unit* [1]. Ketiga bagian ini memungkinkan *unit* LSTM untuk menghasilkan prediksi dan mengingat data yang diberikan sebagai *state*.

2.6. Bidirectional LSTM

Bidirectional LSTM merupakan metode yang digunakan untuk mengatasi kelemahan dari LSTM, yaitu ketidakmampuan LSTM dalam memproses data dari masa depan [7]. Pada *Bidirectional LSTM*, dilakukan proses input secara maju dan mundur (*forward* dan *backward*) menggunakan dua *layer* LSTM terpisah [1]. Pada penelitian ini, kedua *layer* LSTM yang digunakan memiliki unit yang sama, yaitu sebanyak 64 [6]. Model *Bidirectional LSTM* pada penelitian ini digunakan sebagai perbandingan dengan model LSTM untuk menentukan metode yang lebih baik dalam klasifikasi kategori berita.

2.7. Pelatihan Model

Pada tahap ini, model yang sudah dibuat akan dilatih menggunakan *training data*, *training data* terdiri dari 80% dataset total, sedangkan 20% sisanya akan digunakan pada proses evaluasi. Proses *training* dilakukan sebanyak 15 *epoch*.

2.8. Metode Evaluasi

Nilai-nilai yang digunakan untuk menguji dan mengevaluasi model adalah nilai akurasi, nilai *precision*, nilai *recall*, dan nilai F1. Nilai akurasi adalah perbandingan antara jumlah berita yang berhasil diklasifikasikan dengan jumlah total berita. Metrik evaluasi selain akurasi dihitung per kategori (kelas) berita, di mana nilai yang digunakan adalah nilai rata-rata dari evaluasi masing-masing kategori. Cara menentukan nilai metrik evaluasi dapat dilihat pada persamaan berikut.

$$precision = \frac{TP}{TP+FP} \quad (1)$$

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1\ score = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

Di mana TP (*true positive*) adalah banyaknya berita yang berhasil diklasifikasikan sebagai kategori yang dievaluasi, FP (*false positive*) adalah banyaknya berita yang tidak berhasil diklasifikasikan sebagai selain kategori yang dievaluasi, TN (*true negative*) adalah banyaknya berita yang berhasil diklasifikasikan sebagai kategori selain kategori yang dievaluasi, dan FN

(*false negative*) adalah banyaknya berita yang tidak berhasil diklasifikasikan sebagai kategori yang dievaluasi.

3. Hasil dan Pembahasan

Penelitian ini menggunakan 2 model untuk klasifikasi kategori berita, yaitu model LSTM dan Bidirectional LSTM. Model dibuat menggunakan Jupyter Notebook dengan library Tensorflow dan Keras. Layer yang digunakan pada masing-masing model dapat dilihat pada tabel 1.

Tabel 1. Layer yang Digunakan pada Model

LSTM	Bidirectional LSTM
Embedding Layer (Ukuran vektor output = 64)	Embedding Layer (Ukuran vektor output = 64)
LSTM Layer (64 unit, *forward*)	LSTM Layer (64 unit, *backward*)
-	LSTM Layer (64 unit, *forward*)
Hidden Layer (64 Layer)	Hidden Layer (64 unit)
Output Layer (7 unit/kelas)	Output Layer (7 unit/kelas)

Kedua model yang sudah dibuat kemudian dilatih menggunakan data *training* sebanyak 15 *epoch* dengan *optimizer* Adam. Hasil evaluasi model LSTM dan Bidirectional LSTM dapat dilihat pada tabel 2.

Tabel 2. Hasil Evaluasi Model LSTM

Metrics	LSTM	Bidirectional LSTM
Accuracy	0.8536	0.8415
Average Precision	0.85	0.84
Average Recall	0.85	0.84
Average F1-Score	0.85	0.84

Berdasarkan tabel 2, model LSTM sedikit lebih unggul dibandingkan dengan model Bidirectional LSTM dengan akurasi sebesar 85.36%. Kedua model ini memiliki akurasi yang hampir sama.

4. Kesimpulan

Model LSTM yang dibuat dapat mengklasifikasikan kategori berita berdasarkan judulnya dengan akurasi yang cukup tinggi, yaitu sebesar 85.36%. Selanjutnya, akurasi dari model dapat ditingkatkan

menggunakan metode lain atau dengan *hyperparameter tuning*. Model yang sudah dilatih juga dapat diimplementasikan pada aplikasi lain seperti aplikasi web dan android.

Referensi

[1] R. C. Staudemeyer and E. R. Morris, “Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks,” pp. 1–42, 2019, [Online]. Available: http://arxiv.org/abs/1909.09586.
[2] Trivusi, “Mengenal Algoritma Long Short Term Memory (LSTM),” 2022. https://www.trivusi.web.id/2022/07/algoritma-lstm.html (accessed Nov. 20, 2022).
[3] N. Ghaniaviyanto Ramadhan, F. Dharma Adhinata, A. Jala, T. Segara, P. Rakhmadani, and F. Informatika, “Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression,” *J. Ris. Komputer)*, vol. 9, no. 2, pp. 2407–389, 2022, doi: 10.30865/jurikom.v9i2.3979.
[4] Ibamibrahim, “Dataset Judul Berita Indonesia.” https://github.com/ibamibrahim/dataset-judul-berita-indonesia.
[5] K. S. Witanto, N. A. S. ER, and A. E. Karyawati, “Implementasi LSTM pada Analisis Sentimen Review Film Menggunakan Adam dan RMSprop Optimizer,” vol. 10, no. 4, pp. 351–362, 2022.
[6] Keras, “Keras layers API.” https://keras.io/api/layers/.
[7] M. Ilmiah, *Implementasi Metode Bidirectional Long Short-Term Memory (Bi-LSTM) untuk Prediksi Kasus Positif Covid-19 di Indonesia*. 2022.

Halaman ini sengaja dibiarkan kosong