Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2024117968
Vol. 11, No. 5, Oktober 2024, hlm. 965-970 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

PENGENALAN ENTITAS BERNAMA MENGGUNAKAN BI-LSTM PADA CHATBOT BAHASA INDONESIA

Anshar Fariz Zulhilmi[1], Rizal Setya Perdana[2], Indriati[3]

[1,2,3]Universitas Brawijaya, Malang
Email: [1]ansharfarizzulhilmi@gmail.com, [2]rizalespe@ub.ac.id, [3]indriati.tif@ub.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 02 November 2023, diterima untuk diterbitkan: 30 Oktober 2024)

Abstrak

Institusi publik perlu mengitegrasikan *e-government* ke dalam struktur pengelolaan mereka. Sistem pelayanan terpadu harus dapat menyelesaikan masalah yang dihadapi oleh pengguna layanan. Apabila sistem pelayanan terpadu hanya mengandalkan manusia, maka sistem pelayanan terpadu dapat terhambat. *Chatbot* adalah salah satu solusi untuk menggantikan peran manusia dalam sistem pelayanan terpadu. Salah satu komponen pada *chatbot* adalah pengenalan entitas bernama. Pada penelitian ini, pengenalan entitas bernama dilakukan dalam beberapa tahap. Tahapan-tahapan tersebut antara lain penghilangan *noise*, pelabelan data, pembuatan kamus kata dan label, *encoding* urutan dan pemisahan data, inisiasi model, dan pelatihan model. Model yang digunakan yakni *bidirectional long-short term memory*. Skor F1 terbaik yang didapat dari pengujian adalah 87,44% dengan *hyperparameter* jumlah *layer* sebanyak 2, *hidden size* sebanyak 100, dan *learning rate* sebesar 0,01. Kemudian, penambahan jumlah *layer* maupun *hidden size* kurang berpengaruh terhadap skor F1 yang dihasilkan oleh model. *Learning rate* memengaruhi seberapa cepat model mencapai solusi optimal.

Kata kunci: *chatbot, pengenalan entitas bernama, NER, Bi-LSTM, skor F1*

NAMED ENTITY RECOGNITION USING BI-LSTM IN INDONESIAN LANGUAGE CHATBOT

Abstract

Public institutions must integrate e-government into their management structures. An integrated service system must be able to solve the problems faced by service users. If the integrated service system only relies on humans, then the integrated service system can be hampered. Chatbot is one of the solutions to replace the human role in an integrated service system. One component of the chatbot is named entity recognition. In this study, the named entity recognition was carried out in several stages. These stages include noise removal, data labeling, word and label dictionary creation, sequence encoding and data separation, model initiation, and model training. The model used is bidirectional long-short term memory. The best F1 score obtained from the test is 87.44% with hyperparameters of the number of layers of 2, hidden size of 100, and learning rate of 0.01. The addition of the number of layers and hidden size has little effect on the F1 score produced by the model. The learning rate affects how fast the model reaches the optimal solution.

Keywords: *chatbot, named entity recognition, NER, Bi-LSTM, f1-score*

1. PENDAHULUAN

Standar pemerintahan pusat dan daerah di Indonesia telah menjadi berbasis *e-government* (Aritonang, 2017). Pemerintahan daerah khususnya institusi publik diharuskan mengintegrasikan e-government dalam struktur pengelolaan. Dengan adanya *e-government* tersebut, institusi publik dapat menyelenggarakan sistem pelayanan terpadu sesuai dengan Peraturan Pemerintah Republik Indonesia Nomor 96 Tahun 2012 pasal 12. E-government dapat tercipta apabila institusi publik menyediakan sebuah sistem pelayanan terpadu.

Sistem pelayanan pada Universitas Brawijaya (UB) telah mengintegrasikan *e-government*. Contoh penerapan *e-government* pada Universitas Brawijaya adalah aplikasi-aplikasi yang berhubungan dengan aktivitas-aktivitas akademik seperti SIAM UB, SIADO, dan Gapura UB. Dalam penggunaan aplikasi-aplikasi tersebut, adakalanya *civitas academica* UB menemui kendala. Untuk menyelesaikan kendala tersebut, pengguna aplikasi

tersebut memerlukan bantuan pihak yang memiliki kapabilitas menyelesaikan permasalahan pengguna aplikasi. Pengguna aplikasi UB dapat menghubungi layanan Helpdesk TIK UB untuk menyelesaikan berbagai kendala yang dihadapinya. Admin Helpdesk TIK UB akan memberikan saran saran maupun instruksi-instruksi yang dapat menyelesaikan masalah yang dihadapi pengguna.

Admin Helpdesk TIK UB tidak selamanya dapat melayani pengguna. Hal ini dikarenakan Admin Helpdesk TIK UB adalah manusia yang memiliki kebutuhan-kebutuhan jasmani dan rohani. Selain itu, admin Helpdesk TIK tidak bisa melayani pengguna aplikasi Universitas Brawijaya apabila tidak pada jam kerja. Admin juga tidak dapat melayani banyak pengguna sekaligus secara bersamaan. Admin yang berupa manusia dapat digantikan perannya dalam membantu pengguna sistem oleh sebuah *chatbot*.

Chatbot adalah sebuah *software* komputer yang memfasilitasi sebuah percakapan alamiah dengan penggunanya (Shingte et al., 2021). *Chatbot* menggunakan pembelajaran mesin untuk memahami pertanyaan dari *user* kemudian memberikan respon yang tepat sesuai dengan masukan *user*. Salah satu tahapan ekstraksi informasi dalam *chatbot* adalah pengenalan entitas bernama.

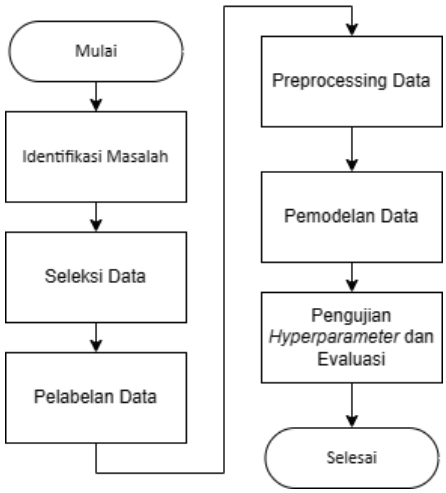
Pada penelitian ini akan dilakukan pengenalan entitas bernama menggunakan algoritma *bidirectional long-short term memory* (Bi-LSTM). Menurut Panchendrarajan dan Amaresan (2018), Bi-LSTM dapat menangkap konteks urutan yang telah dilewati maupun urutan yang akan datang. Hal ini sesuai untuk pengenalan entitas bernama yang membutuhkan konteks kata-kata yang berada didepan maupun belakang. Entitas-entitas yang dikenali meliputi orang, aplikasi, dan organisasi. Kemudian akan dibandingkan pengaruh konfigurasi *hyperparameter* terhadap skor F1. Hasil training dari Bi-LSTM akan diuji dan di evaluasi menggunakan metrik penilaian skor F1 maupun waktu latih.

2. METODE PENELITIAN

Penelitian ini akan dilakukan dalam beberapa tahap. Tahapan-tahapan tersebut antara lain identifikasi masalah, seleksi data, *preprocessing* data, pemodelan data, dan pengujian *hyperparameter* dan evaluasi. Tahapan-tahapan tersebut dilihat pada Gambar 1.

2.1 Identifikasi Masalah

Admin Helpdesk TIK UB tidak selamanya bisa menanggapi permasalahan- permasalahan pengguna secara real time 24 jam dalam seminggu. Admin tersebut adalah seorang manusia yang memiliki kebutuhan lain. Oleh karena itu, *chatbot* dibuat untuk memudahkan pekerjaan admin. *Chatbot* tersebut dapat melayani secara *real-time* bagi pengguna yang membutuhkan bantuan Helpdesk TIK UB.



Gambar 1 Metode Penelitian

Terdapat komponen-komponen pada *chatbot* tersebut yang memiliki fungsi sebagai pemahaman pesan yang diberikan pengguna. Salah satu komponen tersebut yakni pengenalan entitas bernama. Pengenalan entitas bernama akan mengekstraksi informasi-informasi penting yang berkaitan dengan pengguna seperti nama, organisasi, dan aplikasi. Ekstraksi informasi dapat dilakukan dengan mempelajari pola-pola tertentu pada pesan pengguna.

2.2 Seleksi Data

Seleksi data dilakukan pada dataset. Data yang dikumpulkan yakni dokumen pesan dan dokumen entitas. Dokumen entitas berperan sebagai label. Untuk entitas akan dicari kategori tertentu. Kategori-kategori tersebut dapat dilihat pada tabel 1.

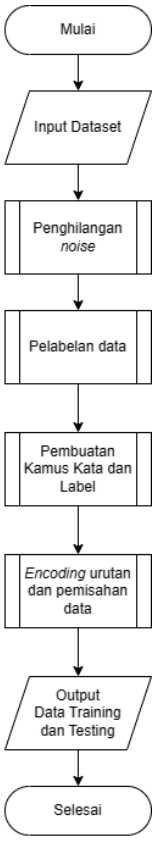
Tabel 1 Contoh Label dan Data

Label	Deskripsi	Contoh Data
B-orang	Kata awal dari entitas orang	Ahmad
I-orang	Kata yang mengikuti entitas orang dan termasuk entitas tersebut	Dhani
B-aplikasi	Kata awal dari entitas aplikasi/layanan yang disediakan oleh Universitas Brawijaya	MS
I-aplikasi	Kata yang mengikuti entitas aplikasi dan termasuk entitas tersebut	Office
B-organisasi	Kata awal dari entitas organisasi yang berada di Universitas Brawijaya.	Fakultas
I-organisasi	Kata yang mengikuti entitas organisasi dan termasuk entitas tersebut	Ilmu Sosial dan Ilmu Politik
O	Kata lain yang tidak termasuk kategori entitas yang diteliti	mohon

2.3 *Preprocessing* data

Data perlu dilakukan *preprocessing* terlebih dahulu agar mendapat data yang bersih sebelum dilakukan proses training. Tahapan-tahapan yang

dilakukan yakni yakni penghilangan *noise*, pelabelan data, pembuatan kamus kata dan label, serta *encoding* label urutan dan pemisahan data. Tahapan penghilangan *noise* adalah tahapan untuk menghilangkan karakter-karakter yang tidak dibutuhkan. Tahapan pembuatan kamus kata dan label merupakan tahapan untuk membuat kamus kata dan label yang akan digunakan untuk *encoding* urutan. Kemudian *encoding* label dan pemisahan data adalah tahapan untuk mengkonversi kata-kata sesuai dengan nilai pada kamus label dan kemudian memisahkan data menjadi data *training* dan data *testing*. Gambar 2 merupakan alur dari *preprocessing* data.



Gambar 2 Diagram Alir *Preprocessing*

2.4 Pemodelan Data

Tahapan pemodelan data untuk pengenalan entitas bernama dilakukan menggunakan arsitektur Bi-LSTM. Input yang diberikan adalah label dan kalimat. Layer embedding akan terlebih dahulu diinisiasi menggunakan bobot *pretrained* dari Fasttext. Input akan melalui *layer embedding*. Kemudian, setelah dari *layer embedding* input akan menuju LSTM. *Layer* LSTM akan menerima dua masukan, yakni satu *layer* digunakan untuk menerima kalimat yang utuh dan satu *layer* lagi menerima kalimat yang dibalik. Selanjutnya, input akan menuju *fully connected layer* atau *linear layer* sebelum akhirnya akan dikenali jenis entitasnya.

2.5 Pengujian Hyperparameter dan Evaluasi

Pengujian *hyperparameter* akan dilakukan pada jumlah *layer*, *hidden size*, dan *learning rate*. Tolak ukur perbandingan *hyperparameter* jumlah layer dan *hidden size* adalah skor F1 dan waktu yang dibutuhkan untuk training. Kemudian, *learning rate* menggunakan tolak ukur skor F1 dan grafik *loss* untuk pembanding. Kemudian *hyperparameter-hyperparameter* yang terbaik akan dipilih untuk digunakan pada analisis *confusion matrix*.

3. LANDASAN KEPUSTAKAAN

3.1 Penelitian Terdahulu

Penelitian pertama terkait pengenalan entitas bernama adalah penelitian yang dilakukan oleh Rachman dkk. (2017). Penelitian tersebut membahas penggunaan bidirectional long short-term memory (Bi-LSTM) pada Twitter berbahasa Indonesia. Data yang digunakan diberikan label *begin-inside-outside* (BIO). Label *begin* menandakan awalan dari sebuah entitas sedangkan *inside* menandakan bagian dari sebuah entitas. Tanda *outside* merupakan bagian teks yang bukan merupakan entitas. Fitur-fitur yang digunakan yakni *word embedding*, *neighboring word embedding*, dan *pos-tagging*. Entitas yang dijadikan label yakni organisasi, orang, dan lokasi. Pada penelitian ini fitur *word embedding* dan pos-tagging mendapatkan skor terbaik dengan *precision* sebesar 78,33%, *recall* sebesar 76,56%, dan skor F1 sebesar 77,08%.

Penelitian selanjutnya yang terkait pengenalan entitas bernama adalah penelitian yang dilakukan oleh Azarine, Arif Bijaksana, dan Asror (2019). Penelitian tersebut melakukan pengenalan entitas bernama menggunakan *hidden markov model*. Penandaan entitas pada data menggunakan label BIO. Entitas-entitas yang menjadi objek penelitian tersebut adalah orang, lokasi dan organisasi. Penambahan fitur berupa *pos-tagging* juga dilakukan untuk meningkatkan akurasi. Skor F1 yang didapatkan pada penelitian ini adalah 64,06%. Skema penandaan BIO terbukti dapat menghasilkan akurasi yang baik.

Penelitian selanjutnya terkait pengenalan entitas bernama pada *chatbot* adalah penelitian yang dilakukan oleh Li dkk. (2019). Pada penelitian tersebut, *chatbot* dibangun dengan tujuan untuk memberikan hotel yang sesuai dengan pengguna. Kemudian, salah satu model yang digunakan oleh penelitian tersebut model Bi-LSTM dengan pre-trained embedding. Hasil skor F1 dari pengenalan entitas bernama adalah 65% untuk entitas hotel, 87% untuk entitas lokasi serta 0,96 untuk gabungan entitas hotel dan lokasi.

Penelitian selanjutnya yang terkait yakni penelitian yang dilakukan oleh Hien dkk. (2018) membahas mengenai arsitektur chatbot untuk menjawab pertanyaan mahasiswa universitas yang terkait dengan layanan satuan pendidikan pada

Faculty of Information Technology of Ho Chi Minh City University of Science. Penelitian menggunakan data sebanyak 1560 pesan pengguna. *Chatbot* tersebut dipercaya dapat lebih membantu mahasiswa dibandingkan chatbot pada umumnya.

3.2 Chatbot

Chatbot adalah program komputer yang memungkinkan interaksi alami dengan pengguna (Shingte dkk., 2021). Chatbot menggunakan pembelajaran mesin ataupun aturan-aturan tertentu untuk memahami permintaan pengguna dan memberikan respons yang sesuai berdasarkan input pengguna. Jika chatbot memahami apa yang diinginkan pengguna, maka respon yang tepat dapat diberikan kepada pengguna.

Jenis *chatbot* terbagi menjadi dua, yakni *linguistic chatbot* dan *artificial intelligence chatbot* (Nirala, Singh, dan Purani, 2022). Jenis *linguistic chatbot* memiliki input yang terbatas sehingga jawaban yang dihasilkan merupakan jawaban yang sudah didefinisikan diawal pembuatan chatbot. Kemudian jenis *artificial intelligence chatbot* menggunakan banyak logika dan konteks kalimat untuk memahami keinginan *user*. Jenis chatbot ini juga dapat mempelajari input dari *user* untuk membuat balasan yang lebih baik kepada *user*.

3.3 Pengenalan Entitas Bernama

Pengenalan entitas bernama atau yang biasa disebut *named entity recognition* (NER) dalam bahasa Inggris adalah tahapan mengklasifikasi suatu bagian dari teks ke dalam salah satu kategori entitas (Taher, Hoseini dan Shamsfard, 2020). Contoh dari kategori entitas tersebut adalah orang, lokasi, dan organisasi. Kategori entitas sendiri dapat ditentukan sendiri sesuai dengan kebutuhan. Untuk mengenali bagian teks yang termasuk kategori entitas tertentu, pengenalan entitas bernama membutuhkan *resource* teks yang berlabel untuk pembuatan aturan maupun pelatihan algoritma *machine learning* atau *deep learning*.

3.4 LSTM

Arsitektur long short-term memory (LSTM) terdiri dari sekumpulan koneksi jaringan-jaringan secara rekurens yang disebut blok memori (Graves, 2012). Tiga unit multiplikatif dan juga hubungan dengan dirinya sendiri digabung menjadi sebuah blok memori. Tiga unit multiplikatif, yaitu input, *output*, dan *forget gate*, memudahkan untuk proses *write*, *read*, dan *reset* pada sel blok memori. Unit-unit multiplikatif ini memudahkan arsitektur LSTM untuk mengatasi *vanishing gradient problem* dengan menyimpan informasi dalam jangka waktu yang lama.

4. PENGUJIAN DAN HASIL

4.1 Pengujian Jumlah Layer

Setiap layer Bi-LSTM terdiri dari satu *forward layer* dan satu *backward layer*. Jumlah layer Bi-LSTM yang akan diuji yakni 1,2, dan 3. Setiap layer akan dibandingkan nilai skor F1 dan waktu yang dibutuhkan untuk *training*. Tabel 2 berikut menampilkan hasil pengujian dari jumlah *layer*.

Tabel 2 Hasil Pengujian Jumlah Layer

No.	Jumlah *Layer*	Skor F1 (%)	Waktu *training* yang dibutuhkan (dalam detik)
1	1	86,84	253
2	2	**86,97**	378
3	3	85,79	490

Pengujian layer Bi-LSTM menghasilkan beberapa kesimpulan. Dari Tabel 2, nilai skor F1 terbaik diperoleh pada pengujian jumlah *layer* Bi-LSTM sebanyak dua dengan nilai 86,97%. Hasil pengujian tersebut mengindikasikan bahwa jumlah *layer* tidak serta merta menambah kemampuan model untuk memiliki performa terbaik.

4.2 Pengujian *Hidden Size*

Hidden size adalah hyperparameter yang mempengaruhi jumlah fitur *hidden state* yang dihasilkan. *Hidden state* pada layer LSTM berfungsi untuk menyimpan nilai input yang sebelumnya sehingga model dapat mempelajari hubungan antar input. Tabel 3 menampilkan hasil pengujian *hidden size*.

Tabel 3 Hasil Pengujian *Hidden Size*

No.	*Hidden size*	Skor F1 (%)	Waktu *training* yang dibutuhkan (dalam detik)
1	50	86,58	274
2	100	**86,77**	378
3	150	86,35	657
4	200	86,75	590

Pengujian *hidden size* menghasilkan skor F1 terbaik pada hidden size sebesar 100. Hasil pada tabel 3 menandakan bahwa penambahan *hidden size* tidak berkorelasi dengan performa yang ditunjukkan oleh model.

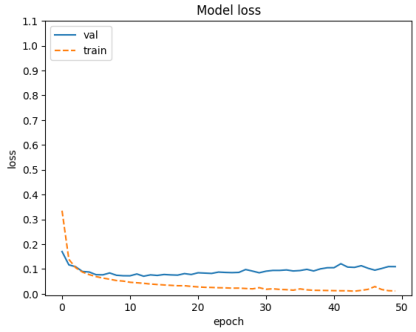
4.3 Pengujian *Learning Rate*

Pengujian *learning rate* akan membandingkan nilai *learning rate* sehingga dapat diketahui pengaruhnya pada model. *Learning rate* adalah *hyperparameter* yang mengatur besaran perubahan bobot yang dilakukan oleh model. Tabel 4 menampilkan hasil pengujian *learning rate*.

Tabel 4 Hasil Pengujian Learning Rate

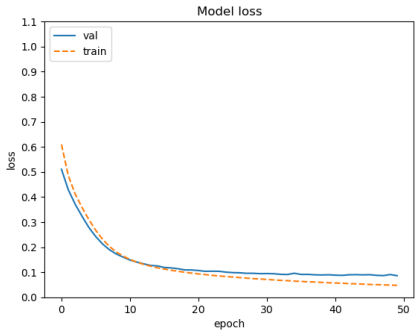
No.	Nilai *learning rate*	Skor F1 (%)
1	0,1	86,81
2	0,01	**87,44**
3	0,001	41,19

Gambar 3 adalah hasil grafik nilai *loss* untuk *learning rate* 0,1.



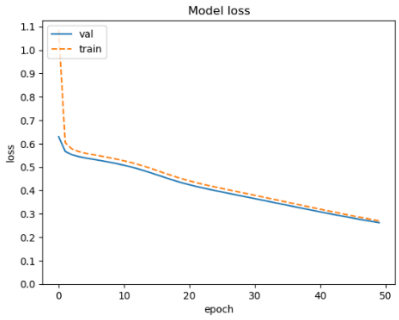
Gambar 3 Hasil Grafik Nilai Loss untuk *Learning Rate* 0,1

Kemudian Gambar 4 adalah grafik nilai *loss* untuk pengujian *learning rate* sebesar 0,01.



Gambar 4 Hasil Grafik Nilai Loss untuk *Learning Rate* 0,01

Gambar 5 adalah grafik nilai *loss* untuk pengujian *learning rate* sebesar 0,001.

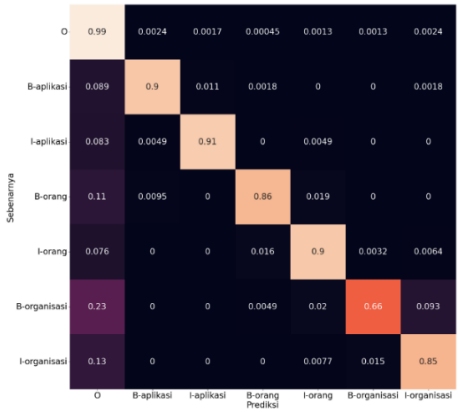


Gambar 5 Hasil Grafik Nilai *Loss* untuk *Learning Rate* 0,01

Pada Gambar 3 model dengan *learning rate* 0,1 memiliki nilai loss yang cenderung meningkat sehingga tidak dapat mencapai solusi optimal. Pada Gambar 4, model dengan *learning rate* 0,01 memiliki grafik nilai *loss* berjalan beriringan antara nilai *loss* pada *training* dan validasi. Sedangkan pada gambar 5, model dengan *learning rate* 0,001 terlalu lambat untuk memperbarui bobot sehingga model belum mencapai performa optimal.

4.4 Hasil Penggunaan Hyperparameter

Confusion matrix hasil pengujian kombinasi *hyperparameter* terbaik dapat dilihat pada Gambar 6 dibawah.



Gambar 6 Confusion Matrix Hyperparameter Pengujian

Terlihat bahwa untuk label O memiliki rasio true positive terbaik dengan nilai 0,99. Tingginya skor F1 dikarenakan label O adalah label dengan jumlah label terbanyak. Label O dapat memiliki jumlah ratusan kali lipat lebih banyak daripada jumlah label lain. Kemudian model juga masih sering salah mengidentifikasi yang bukan dari label O. Kemudian rasio true positive terendah didapatkan oleh label B-organisasi. Untuk label B-Organisasi, Model masih sering salah membedakan B-organisasi dengan I-Organisasi. Hal ini dapat terlihat dari nilai rasio false negative untuk label B-organisasi terhadap I-Organisasi. Fenomena ini salah satunya disebabkan oleh data pengguna Helpdesk TIK yang menuliskan organisasi secara variatif seperti nama fakultas diletakkan sebelum nama prodi atau nama prodi diletakkan sebelum nama fakultas.

5. KESIMPULAN DAN SARAN

Ada beberapa kesimpulan yang dapat ditarik pada penelitian ini. Penambahan *hyperparameter* jumlah layer tidak selalu meningkatkan performa model. Penambahan *layer* akan memperlambat proses *training*. Kemudian, *hidden size* yang semakin banyak tidak mengindikasikan kenaikan performa pada pemodelan entitas bernama. *Learning rate* yang terlalu kecil dapat memperlambat model untuk mencapai konvergen dan sebaliknya *learning rate* yang terlalu besar mengakibatkan model tidak dapat mencapai solusi optimal. Kemudian saran yang perlu diberikan untuk peneliti selanjutnya yakni menggunakan data yang lebih banyak, menggunakan *word embedding* yang lain seperti BERT ataupun *word embedding* lain yang berhubungan dengan domain permasalahan, dan menggunakan arsitektur-arsitektur lain seperti *convolutional neural network*.

6. DAFTAR PUSTAKA

ARITONANG, D.M., 2017. The Impact of E-Government System on Public Service Quality in Indonesia. *European Scientific Journal, ESJ*, 13(35), p.99. https://doi.org/10.19044/esj.2017.v13n35p99.

AZARINE, I.S., ARIF BIJASKSANA, M. and ASROR, I., 2019. Named Entity Recognition on Indonesian Tweets using Hidden Markov Model. In: *2019 7th International Conference on Information and Communication Technology (ICoICT)*. IEEE. pp.1–5. https://doi.org/10.1109/ICoICT.2019.8835277.

GRAVES, A., 2012. Long Short-Term Memory. pp.37–45. https://doi.org/10.1007/978-3-642-24797-2_4.

HIEN, H.T., CUONG, P.-N., NAM, L.N.H., NHUNG, H.L.T.K. and THANG, L.D., 2018. Intelligent Assistants in Higher-Education Environments. In: *Proceedings of the Ninth International Symposium on Information and Communication Technology - SoICT 2018*. New York, New York, USA: ACM Press. pp.69–76. https://doi.org/10.1145/3287921.3287937.

LI, B., JIANG, N., SHAM, J., SHI, H. and FAZAL, H., 2019. Real-World Conversational AI for Hotel Bookings. In: *2019 Second International Conference on Artificial Intelligence for Industries (AI4I)*. IEEE. pp.58–62. https://doi.org/10.1109/AI4I46381.2019.00022.

NIRALA, K.K., SINGH, N.K. and PURANI, V.S., 2022. A survey on providing customer and public administration based services using AI: chatbot. *Multimedia Tools and Applications*, 81(16), pp.22215–22246. https://doi.org/10.1007/s11042-021-11458-y.

Panchendrarajan, R. and Amaresan, A., 2018. Bidirectional LSTM-CRF for named entity recognition. In: *Proceedings of the 32nd Pacific Asia conference on language, information and computation*.

RACHMAN, V., SAVITRI, S., AUGUSTIANTI, F. and Mahendra, R., 2017. Named entity recognition on Indonesian Twitter posts using long short-term memory networks. In: *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE. pp.228–232. https://doi.org/10.1109/ICACSIS.2017.8355038.

SHINGTE, K., CHAUDHARI, A., Patil, A., CHAUDHARI, A. and DESAI, S., 2021. Chatbot Development for Educational Institute. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3861241.

TAHER, E., HOSEINI, S.A. and SHAMSFARD, M., 2020. Beheshti-NER: Persian Named Entity Recognition Using BERT. [online] Available at: <http://arxiv.org/abs/2003.08875>.