Tinjauan Literatur *Named Entity Recognition* dengan Machine Learning dan Deep Learning pada Ulasan Wisata

Muhammad Irfan
Program Studi Teknik Informatika
Universitas Islam Indonesia
Yogyakarta, Indonesia
17523149@students.uii.ac.id

Ahmad Fathan Hidayatullah
Program Studi Teknik Informatika
Universitas Islam Indonesia
Yogyakarta, Indonesia
fathan@uii.ac.id

***Abstract*— Negara Indonesia merupakan salah satu negara tujuan wisata. Banyak para wisatawan mencari informasi tentang tempat wisata dengan mencari ulasan melalui internet. Namun, mencari ulasan melalui internet tidaklah mudah. Perlu adanya tambahan informasi secara rinci mengenai beberapa hal terkait tempat wisata seperti lokasi wisata, fasilitas, suasana, dan sebagainya. Untuk mengidentifikasi informasi berkaitan dengan hal tersebut, pendekatan *named entity recognition* (NER) dapat dilakukan. Metode *Named Entity Recognition* (NER) adalah sebuah ekstraksi informasi dan pemrosesan dokumen secara terstruktur dan tidak terstruktur. Untuk mengetahui perkembangan penelitian mengenai NER di bidang pariwisata, penelitian ini mengkaji sebanyak 8 referensi literatur mengenai NER pada lingkup wisata yang didapat dengan mencari di *Google Scholar* dengan kata kunci pencarian "*NER for Tourism"*. Penelitian NER paling banyak ditemukan di bidang wisata dengan menggunakan model *Bidirectional Encoder Representations from Transformer* (BERT). Model BERT adalah pelatihan dalam representasi kata yang berguna untuk mencegah agar kata tidak menjadi ambigu sehingga tidak terjadi kesalahan dalam pengenalan entitas. Hasil dari penelitian ini diharapkan dapat membantu pengembangan NER pada bidang pariwisata.**

Keywords— Named entity recognition, Wisata, BERT

I. PENDAHULUAN

Negara Indoensia merupakan salah satu negara tujuan wisata. Sebelum mendatangi suatu objek pariwisata, para wisatawan akan melihat berbagai macam ulasan yang ada melalui internet. Salah satunya adalah dengan mengunjungi website seperti TripAdvisor yang menyediakan ulasan mengenai berbagai macam lokasi wisata. Keberadaan website tersebut sangatlah memudahkan calon wisatawan untuk memilih tempat wisata yang akan dikunjungi setelah melihat ulasan tersebut [1]. Akan tetapi, banyaknya ulasan yang tersedia dapat menyulitkan pengunjung website untuk mendapatkan informasi secara lebih rinci.

Named Entity Recognition (NER) merupakan proses ekstraksi informasi dan pemrosesan dokumen secara terstruktur maupun tidak terstruktur yang mengacu nama orang, tempat, organisasi dan waktu [2]. Dengan mengetahui entitas tersebut, maka akan dapat diperoleh penting tentang ulasan wisata seperti nama tempat wisata, nama lokasi, fasilitas, dan sebagainya. Namun demikian, melakukan ekstraksi entitas tersebut ada beberapa permasalahan yang dihadapi seperti adanya penulisan entitas yang beragam dan tidak konsisten. Oleh karena itu, penerapan NER dengan *machine learning* dan *deep learning* sangat penting untuk mempermudah proses ekstraksi entitas tersebut.

Penelitian ini melakukan kajian literatur mengenai Named Entity Recognition dengan *machine learning* dan *deep learning* pada ulasan wisata. Kajian literatur ini dilakukan agar mengetahui seberapa banyak yang melakukan penelitian dengan NER di bidang wisata. Tinjauan literatur ini dapat mempunyai peran serta menunjang peneliti untuk memilih metode-metode yang baik untuk NER dan terutama di bidang wisata.

II. METODOLOGI PENELITIAN

Pengumpulan data dilakukan dengan mencari referensi melalui *Google Scholar.* Tabel I menjelaskan tentang kriteria inklusi dan kriteria eksklusi untuk mencari literatur NER di bidang pariwisata. Kriteria inklusi merupakan kriteria yang harus dipenuhi dalam penelitian dan kriteria eksklusi merupakan kriteria yang tidak dipenuhi dalam penelitian. Kata kunci pencarian literatur yaitu "*NER for Tourism"* dan "*NER Tourism*". Nama peneliti dan judul peneliti dapat dilihat di tabel II. Referensi yang digunakan menggunakan Bahasa Inggris dan Bahasa Indonesia. Total referensi literatur yang digunakan pada literatur ini berjumlah 8 literatur yang membahas di bidang wisata, dilihat pada tabel II.

TABEL I. KRITERIA INKLUSI DAN EKSKLUSI YANG DITERAPKAN PADA NER

Kriteria Inklusi	Kriteria Eksklusi
Yang berhubungan dengan NER	Tidak termasuk dalam literatur yang berhubungan dengan NER
Literatur membahas tentang NER di bidang wisata	Literatur membahas NER namun bukan di bidang wisata
Literatur NER di bidang wisata memakai Bahasa Indonesi dan Bahasa Inggris	Literatur mengenai NER di bidang wisata yang memakai bahasa selain dengan Bahasa Inggris dan Bahasa Indonesia

TABEL II. NAMA PENELITI DAN JUDUL

Nama Peneliti	Judul penelitian
Saputro, Khurniawan Eko Kusumawardani, Sri Suning Fauziati, Silmi [1]	Development of Semi-Supervised Named Entity Recognition to Discover New Tourism Places
Zahra, Annisa Hidayatullah, Ahmad Fathan Rani, Septia[2]	Kajian Literatur Named Entity Recognition pada Domain Wisata
Hu, Yan Nuo, Minghua Tang, Chao[3]	A Deep Learning Approach for Chinese Tourism Field Attribute Extraction
Chantrapornchai, Chantana Tunsakul, Aphisit [4]	Information Extraction on Tourism Domain using SpaCy and BERT
Xue, Leyi Cao, Han Ye, Fan Qin, Yuehua [5]	A method of Chinese Tourism Named Entity Recognition Based on BBLC Model
Putra, Muhammad Fakhri Despawida Aulia Hidayatullah, Ahmad Fathan[6]	Tinjauan Literatur : Named Entity Recognition pada Ulasan Wisata
Liang, Xuchao Cao, Han Zhang, Weizhen [7]	Knowledge Extraction Experiment Based on Tourism Knowledge Graph Q A Data Set
Zahra, Annisa Hidayatullah, Ahmad Fathan Rani, Septia [8]	Pemodelan Named Entity Recognition pada Artikel Wisata dengan Metode Bidirectional Long Short-Term Memory dan Conditional Random Fields

III. DISKUSI

A. *Proses Named Entity Recognition*

Berdasarkan literatur yang telah didapat, ada beberapa tahapan yang sering dilakukan dalam NER.

1) Dataset

Dari literatur yang yang telah didapatkan, para peneliti memperoleh dari berbagai sumber data, di antaranya dari pencarian top google dengan kata kunci "top tourism place" [1], website dan artikel [2], MSRA dan CTFAE [3], TripAdvisor, Traveloka, dan Hotel.com [4], perjalanan trip Ctrip dan Mafengwo [5], TripAdvisor [6], Shaanxi Tourist Attractions dan Nlpcc universal [7], website dan artikel [8].

2) Preprocessing

Preprocessing merupakan sebuah proses mengolah dan menyiapkan data agar lebih mudah untuk diproses. Pada tahap ini, ada berbagai jenis *preprocessing* yang dapat digunakan sesuai kebutuhannya yaitu normalisasi, tokenisasi, dan menghilangkan kata yang tidak penting [1][2]. Melakukan *preprocessing* dalam bentuk pemecahan kalimat, *POS tagging,* dan chunking. Ada juga yang melakukan chunking dengan memakai library SpaCy noun chunk [2][4].

3) Pelabelan Data

Proses yang akan dilakukan sesudah melakukan *preprocessing* yaitu melakukan pelabelan data. Dari beberapa literatur yang ada, format pelabelan menggunakan format BIO. Pada format tersebut, "B" merupakan awal entitas, "I" merupakan entitas perantara, dan "O" merupakan kata non-entitas [5]. Pelabelan data dari riset sebelumnya hanya di bidang wisata seperti nama orang, nama lokasi, nama organisasi, waktu dan lain-lain [5].

4) Ekstraksi fitur

Proses ektraksi fitur digunakan untuk mengubah data asli yang berawal dari teks menjadi vektor, karena dalam penggunaan *deep learning* tidak dapat mengolah data dengan data asli, maka harus diubah terlebih dahulu dari teks menjadi vektor. Dari proses ekstraksi fitur, akan dapat diketahui informasi karakter dari sebuah data dan menggali informasi yang berguna untuk menuju ke tahap selanjutnya [5][2]. Salah satu contoh proses ekstraksi fitur dengan menerapkan *word embedding.*

Word embedding merupakan sistem yang dapat melakukan pengubahan dari sebuah teks menjadi angka, berbagai jenis *word embbeding* yaitu Word2vec, GloVe, dan Skip-gram [6]. Dalam pengenalan entitas ini, sebagian besar peneliti menggunakan Word2Vec. Namun penggunaan Word2vec pada NER menimbulkan masalah yaitu dalam membedakan kata yang sama pada konteks yang berbeda.

5) Pemodelan NER

Proses yang akan dilakukan setelah ekstraksi fitur yaitu merancang sebuah model Named Entity Recognition yang terdiri dari satu metode atau lebih metode didalamnya, model Named Entity Recognition berada pada tabel III. Dalam tabel III tersebut, dapat diketahui bahwa model yang paling banyak digunakan yaitu model BERT. BERT (*Bidirectional Encoder Representations from Transformer)* merupakan bentuk representasi kata untuk mencegah kata agar tidak terjadi ambigu dan kesalahan dalam pengenalan entitas [2][5].

6) Evaluasi

Untuk mengukur kinerja model NER yang telah dibangun, peneliti menggunakan precision, recall dan f1-score [7]. Hasil evaluasi dari penelitian sebelumnya dapat dilihat pada tabel III.

TABEL III. TABEL OVERVIEW MODEL, LABEL ENTITAS, DAN HASIL EVALUASI DARI PENELITIAN NER SEBELUMNYA

Referensi	Dataset	Ukuran Dataset	Bahasa Dataset	Model	Label Entitas	Hasil Evaluasi
[1]	Data dalam penelitian ini dikumpulkan dari top google dengan kata kunci "top tourism place in%s". Sumber data umum didapat dengan seratus halaman yang menggambarkan wisata di lima benua.	2.686 Entitas unik	Inggris	Yet Another Two Stage Idea dikombinasi Naïve Bayes Classifier dan K-Nearest Neighbor	Nature, city, region, negative.	*F1-score* : 69.1%
[2]	Data diperoleh dari website, artikel	Tidak disebutkan	Inggris	BiLSTM dan CRF	Nama wisata, tempat penginapan, fasilitas, dan lokasi	Tidak disebutkan
[3]	MSRA	55.280 Kalimat	Cina	BERT – ResCNNs – BLSTM – CRF	Location, organization, person	F1-score : 95.41%
	CTFAE	15.845 Kalimat	Cina	BERT – ResCNNs – BLSTM – CRF	Include area, construction time, location, nickname, internal attraction	F1-score : 92.17%
[4]	Menggunakan data dari komentar ulasan Tripadvisor, Traveloka, Hotels.com yang meliputi ulasan hotel, ulasan restoran dan ulasan wisata	Data restoran 18.700 Data hotel 11.859 Data wisata 12.523	Tidak disebutkan	Library SpaCy, BERT	Location, facility, organization	Dari dataset menggunakan model BERT dan SpaCy mendapatkan akurasi 95% - 98%
[5]	Data teks diperoleh dari perjalanan yang dilakukan Ctrip dan Mafengwo	13. 464 Kalimat	Cina	BERT – BLSTM – CRF (BBLC)	Person, location, organization, time, thing.	F1-score metode BLSTM-CRF : 85.52 % F1-score metode CRF : 88 % F1-score metode BBLC : 84.79 %
[6]	Website TripAdvisor	Tidak disebutkan	Indonesia	BiLSTM,CNN dan CRF	Nama wisata, nama lokasi, dan fasilitas	Tidak disebutkan
[7]	Menggunakan data set dari Shaanxi Tourist Attractions Menggunakan data set dari Nlpcc universal	700 pasangan tanya jawab	Tidak disebutkan	BERT-BiLSTM-CRF	Tidak disebutkan	Shaanxi tourist mendapatkan akurasi : 89.77% Nlpcc universal mendapatkan akurasi : 96.93%
[8]	Data diperoleh dari website, artikel	183.507 data ulasan	Inggris	BiLSTM dan CRF	Nama wisata, tempat penginapan, fasilitas, dan lokasi	F1-score : 75.25%

IV. ANALISIS

Pada bagian ini, akan dianalisis semua literatur yang telah diperoleh di bab sebelumnya. Analisis dilakukan dengan menyebutkan metode, bahasa *dataset*, entitas, dan hasil. Penelitian yang dilakukan oleh [1] penerapan *Machine Learning* dengan menggunakan NER berbasis semi-supervised dan unsupervised. Metode Yet Another Two Stage Idea yaitu algoritma SSL yang memiliki dua tahap proses. YATSI ini dilatih menggunakan data berlabel, sehingga Naïve Bayes Classifier dipilih karena probabilistik dan mudah untuk diimplementasikan. Pada tahap kedua algoritma ini mencoba untuk memprediksi data tidak berlabel dengan menggunakan pendekatan K-Nearrest Neighbor [1]. Bahasa dataset yang digunakan dalam penelitian ini menggunakan Bahasa Inggris, entitas yang diterapkan nature, city, region, negative. Hasil peforma mendapatkan 69.1%. Penelitian yang dilakukan oleh [2] metode yang diterapkan adalah BiLSTM dan CRF, BiLSTM dipilih karena tingkat kepopuleran metode CRF sangat tinggi dalam *Named Entity Recognition*, sehingga mengombinasikan dengan metode BiLSTM. Menggunakan metode BiLSTM karena dapat digunakan untuk pemberian kode teks input dua arah [2]. Bahasa dataset yang digunakan Bahasa Inggris, entitas yang diterapkan nama wisata, tempat penginapan, fasilitas, dan lokasi. Hasil performa tidak disebutkan. Penelitian yang dilakukan oleh [3] metode yang digunakan BERT yang dikombinasi dengan ResCNNs-BLSTM-CRF. Mekanisme model yang dilakukan yang pertama melakukan pelabelan urutan. Kedua, melakukan penyematan karakter sebagai masukan [3]. Bahasa *dataset* yang digunakan Bahasa Cina, entitas yang diterapkan location, organization, person untuk *dataset* MSRA dan Include area, construction time, location, nickname, internal attraction untuk *dataset* CTFAE. Kombinasi beberapa metode tersebut berhasil mendapatkan hasil F1-*Score* yang cukup tinggi yaitu 95.41% untuk *dataset* MSRA dan 92.17% untuk *dataset* CTFAE. Penelitian yang dilakukan oleh [4] penerapan metode menggunakan Library SpaCy dan BERT. Penerapan Library SpaCy, karena telah dibuat model untuk mengidentifikasi nama organisasi, nama tempat dan fasilitas. Penerapan metode BERT, karena menggunakan format BIO Tag yang dapat mengidentifikasi nama organisasi, nama tempat, dan fasilitas. Mekanisme dari library SpaCy untuk NER merupakan pemrosesan fitur seperti tokenisasi, *part of speech*. Modelnya skokastik dan proses pelatihan dilakukan menggunakan gradien. Mekanisme dari BERT untuk NER tag untuk entitas bernama yang harus dilabeli, dan juga harus melatih model untuk mengenalinya sehingga dapat mengekstrak relasi yang diinginkan [4]. Bahasa dataset yang digunakan tidak disebutkan dalam penelitian, entitas yang diterapkan Location, facility, organization. Hasil performa mendapatkan 95% - 98%. Penelitian yang dilakukan oleh [5] penerapan metode yang digunakan yaitu mengombinasikan BERT-BiLSTM-CRF yang bisa disebut dengan BBLC Model**,** Metode tersebut diterapkan karena mendapat F1 score yang tinggi dalam entitas person, location, organization, time and things. Mekanisme model BBLC untuk NER, BERT bertujuan menghasilkan model bahasa dan encoder transformer untuk membaca seluruh urutan teks, BLSTM memproses seluruh urutan teks, dan CRF menghasilkan pelabelan data dengan format BIO tag [5]. Bahasa *dataset* yang digunakan Bahasa Cina, entitas yang diterapkan person, location, organization, time, thing. Hasil performa BBLC mendapatkan 84.79%. Penelitian [6] metode yang diterapkan adalah BiLSTM, CNN dan CRF pada domain wisata, karena metode BiLSTM dapat memberikan kode teks input dari dua arah dan metode CRF digunakan untuk mengeluarkan label di setiap karakter. Mekanisme seluruh model diatas membandingkan hasil yang dihasilkan setiap medel tersebut [6]. Bahasa *dataset* yang digunakan Bahasa Indonesia, entitas yang diterapkan nama wisata, nama lokasi, dan fasilitas. Hasil performa dari penelitian tersebut tidak disebutkan. Penelitian [7] metode yang digunakan dalam NER yaitu BERT-BiLSTM-CRF yang mendapatkan hasil performa yang cukup baik dengan dataset Shaanxi Tourist yaitu 89.77% dan mendapatkan hasil yang cukup baik dengan dataset Nlpcc Universal yaitu 96.93%. Bahasa *dataset* yang digunakan tidak disebutkan dan entitas yang diterapkan tidak disebutkan. Mekanisme terdiri dari BERT digunakan untuk mendapatkan vector kata, BLSTM digunakan untuk pemberian kode teks input dari dua arah, dan CRF digunakan untuk mengeluarkan urutan label dengan kategori setiap karakter [7]. Penelitian yang dilakukan oleh [8] metode yang diterapkan adalah BiLSTM dan CRF, BiLSTM dipilih karena tingkat kepopuleran metode CRF sangat tinggi dalam *Named Entity Recognition*, sehingga mengombinasikan dengan metode BiLSTM. Menggunakan metode BiLSTM karena dapat digunakan untuk pemberian kode teks input dua arah [8]. Bahasa dataset yang digunakan Bahasa Inggris, entitas yang diterapkan nama wisata, tempat penginapan, fasilitas, dan lokasi. Hasil performa mendapatkan 75.25%.

Dari analisis yang diperoleh, ada dua pendekatan yang berbeda yaitu *machine learning* dan *deep learning*. Pendekatan *machine learning* yang digunakan YATSI, NBC, dan KKN. Pendekatan *deep learning* yang digunakan BERT, ReCNNs, BLSTM dan CRF. Hasil performa menggunakan pendekatan *deep learning* diantaranya [2], [3], [4], [5], [6], [7], [8] lebih baik dari pada menggunakan pendekatan *machine learning* [1].

V. KESIMPULAN

Penelitian ini mengkaji 8 literatur yang berhubungan dengan NER pada bidang wisata yang diperoleh melalui *Google Scholar*. Berdasarkan literatur yang telah dikaji, ada dua pendekatan yang digunakan yaitu *machine learning* dan *deep learning*. Dari hasil temuan, performa *deep learning* lebih bagus dari *machine learning*. Dari beberapa literatur NER yang diperoleh, model yang populer yaitu BERT (*Bidirectional Encoder Representation from Transformer*). Model BERT merupakan pelatihan dalam representasi kata yang berguna untuk mencegah kata agar tidak menjadi ambigu sehingga tidak terjadi kesalahan dalam pengenalan entitas. Literatur yang telah ditemukan menggunakan *dataset* Bahasa Cina, Bahasa Inggris dan Bahasa Indonesia. Dalam literatur yang telah dikaji, entitas yang telah di temukan yaitu nature, city, region, negative, person, location, organization, time, thing, nama wisata, fasilitas. Penerapan NER pada domain wisata harus dikembangkan lagi, sehingga dapat mengidentifikasi kelas entitas yang lebih banyak dan lebih luas.

REFERENCES

[1] K. E. Saputro, S. S. Kusumawardani, and S. Fauziati, "Development of semi-supervised named entity recognition to discover new tourism places," *Proc. - 2016 2nd Int. Conf. Sci. Technol. ICST 2016*, pp. 124–128, 2017, doi: 10.1109/ICSTC.2016.7877360.

[2] A. Zahra, A. F. Hidayatullah, and S. Rani, "Kajian Literatur Named Entity Recognition pada Domain Wisata," *Automata*, vol. 2, no. 1, pp. 0–4, 2021.

[3] Y. Hu, M. Nuo, and C. Tang, "A Deep Learning Approach for Chinese Tourism Field Attribute Extraction," *Proc. - 2019 15th Int. Conf. Comput. Intell. Secur. CIS 2019*, pp. 108–112, 2019, doi: 10.1109/CIS.2019.00031.

[4] C. Chantrapornchai and A. Tunsakul, "Information extraction on tourism domain using SpaCy and BERT," *ECTI Trans. Comput. Inf. Technol.*, vol. 15, no. 1, pp. 108–122, 2021, doi: 10.37936/ecti-cit.2021151.228621.

[5] L. Xue, H. Cao, F. Ye, and Y. Qin, "A method of chinese tourism named entity recognition based on bblc model," *Proc. - 2019 IEEE SmartWorld, Ubiquitous Intell. Comput. Adv. Trust. Comput. Scalable Comput. Commun. Internet People Smart City Innov. SmartWorld/UIC/ATC/SCALCOM/IOP/SCI 2019*, no. September 1995, pp. 1722–1727, 2019, doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00307.

[6] M. F. D. A. Putra and A. F. Hidayatullah, "Tinjauan Literatur : Named Entity Recognition pada Ulasan Wisata," *Automata*, 2021, [Online]. Available: https://journal.uii.ac.id/AUTOMATA/article/view/17391.

[7] X. Liang, H. Cao, and W. Zhang, "Knowledge Extraction Experiment Based on Tourism Knowledge Graph Q A Data Set," *Proc. 2020 IEEE Int. Conf. Power, Intell. Comput. Syst. ICPICS 2020*, pp. 828–832, 2020, doi: 10.1109/ICPICS50287.2020.9202197.

[8] A. Zahra, A. F. Hidayatullah, and S. Rani, "PEMODELAN NAMED ENTITY RECOGNITION PADA ARTIKEL WISATA DENGAN METODE BIDIRECTIONAL LONG SHORT-TERM MEMORY DAN CONDITIONAL RANDOM FIELDS ARTIKEL WISATA DENGAN METODE BIDIRECTIONAL LONG SHORT-TERM MEMORY DAN CONDITIONAL RANDOM FIELDS," 2021.