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

by John Doe

Submission date: 11-Jun-2021 01:30PM (UTC+0700)
Submission ID: 1604546346
File name: engan_Machine_Learning_dan_Deep_Learning_pada_Ulasan_Wisata.docx (71.38K)
Word count: 2328
Character count: 14936

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

***Abstract*— Negara Indonesia adalah negara yang dipenuhi dengan pariwisata. Wisatawan akan melihat ulasan wisata tersebut, tetapi dengan banyaknya ulasan wisatawan menjadi kesulitan mendapatkan informasi tersebut. Metode *Named Entity Recognition* (NER) adalah sebuah ekstraksi informasi dan pemrosesan dokumen secara terstruktur dan tidak terstruktur. NER dapat digunakan untuk mengidentifikasi informasi misal wisata, nama tempat, fasilitas, maupun suasana. Literatur ini mengkaji sebanyak 12 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 dibidang 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 Indonesia adalah negara yang dipenuhi dengan pariwisata. Sebelum mendatangi salah satu objek pariwisata tersebut biasanya wisatawan akan melihat berbagai macam ulasan yang ada. Dengan adanya *TripAdvisor* yaitu *website* wisata terbesar yang menyajikan berbagai ulasan dari wisatawan yang pernah melakukan perjalanan akan memudahkan calon wisatawan untuk memilih tempat wisata yang akan dikunjungi setelah melihat ulasan tersebut[1]. Tetapi dengan banyaknya ulasan wisatawan menjadi kesulitan dalam mendapatkan informasi.

Named Entity Recognition (NER) adalah sebuah ekstraksi informasi dan pemrosesan dokumen secara terstruktur maupun tidak terstruktur yang mengacu pada orang, tempat, organisasi dan waktu[2]. NER bermanfaat dalam mengidentifikasi misal nama orang, nama tempat, nama organisasi dan nama perusahaan. Entitas tersebut bisa dijadikan informasi dengan ekstraksi fitur, tetapi tidak mudah untuk mengekstraksi fitur dikarenakan entitas tersebut adalah nama orang, nama tempat, nama organisasi, nama orang dan nama lokasi yang ada dalam nama perusahaan[3]. Sehingga entitas nama tersebut akan berawal dengan huruf kapital sehingga dalam mengekstraksi fitur tersebut akan sulit karena mesin tidak dapat mengekstraksi entitas yang berhuruf kapital.

Oleh sebab itu literatur ini dikerjakan yang bertujuan untuk melihat komparasi beberapa metode yang digunakan pada NER, agar mengetahui apa yang telah digunakan pada metode NER penelitian sebelumnya serta melihat hasilnya. Tinjauan literatur ini juga dilakukan agar mengetahui seberapa banyak yang melakukan penelitian dengan NER di bidang Pariwisata, melihat Pariwisata saat ini sangat menonjol dalam sektor ekonomi di Indonesia. Tinjauan literatur ini dapat mempunyai donasi serta menunjang peneliti untuk memastikan metode-metode yang baik untuk NER dan yang paling utama di bidang pariwisata.

II. METODOLOGI PENELITIAN

Pengumpulan data merupakan mencari referensi yang dapat dilakukan menggunakan *Google Scholar*. Pada table 1 dijelaskan bahwa terdapat Kriteria Inklusi yang berarti kriteria yang harus dipenuhi dan Kriteria Eksklusi yang berarti kriteria pendukung dari kriteria inklusi . Kata kunci pencarian literatur yaitu "*NER for Tourism*". Nama peneliti dan judul peneliti bisa dilihat di tabel 2. Referensi yang digunakan menggunakan Bahasa Inggris dan Bahasa Indonesia. Total referensi literatur yang digunakan pada literatur ini berjumlah 12 literatur dengan 3 literatur membahas bukan dibidang wisata, 7 membahas dibidang wisata dan 2 literatur tidak disebutkan label entitasnya, dilihat pada tabel 3.

TABLE I. KRITERIA INKLUSI DAN EKSKLUSI YANG DITERAPKAN PADA NER

Kriteria Inklusi	Kriteria Eksklusi
Yang berhubungan dengan akademi	Tidak termasuk dalam literatur yang berhubungan dengan akademi
Literatur membahas tentang NER dibidang wisata	Literatur membahas NER namun bukan dibidang wisata
Literatur NER dibidang wisata memakai bahasa indonesia atau bahasa inggris	Literatur mengenai NER dibidang wisata yang memakai bahsa selain dengan bahasa inggris atau bahasa indonesia

TABLE II. NAMA PENELITI DAN JUDUL

Nama Peneliti	Judul penelitian
Cambria, Erik Valdivia, Ana Luzón, M Victoria Herrera, Francisco [1]	AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS Sentiment Analysis in TripAdvisor
Štravs, Miha Zupančič, Jernej [2]	Named entity recognition using gazetteer of hierarchical entities
Alfred, Rayner Leong, Leow Chin On, Chin Kim Anthony, Patricia [3]	Malay Named Entity Recognition Based on Rule-Based Approach
Hu, Yan Nuo, Minghua Tang, Chao[4]	A Deep Learning Approach for Chinese Tourism Field Attribute Extraction
Saputro, Khurniawan Eko Kusumawardani, Sri Suning Fauziati, Silmi [5]	Development of semi-supervised named entity recognition to discover new tourism places
Zahra, Annisa Fathan Hidayatullah, Ahmad Rani, Septia [6]	Kajian Literatur Named Entity Recognition pada Domain Wisata
Chantrapornchai, Chantana Tunsakul, Aphisit [7]	Information extraction on tourism domain using SpaCy and BERT
Hakala, Kai Pyysalo, Sampo [8]	Biomedical Named Entity Recognition with Multilingual BERT
Xue, Leyi Cao, Han Ye, Fan Qin, Yuehua [9]	A method of chinese tourism named entity recognition based on bblc model
Fakhiri Despawida Aulia Putra, Muhammad Fathan Hidayatullah, Ahmad [10]	Tinjauan Literatur : Named Entity Recognition pada Ulasan Wisata
Sun, Cong Yang, Zhihao [11]	Transfer Learning in Biomedical Named Entity Recognition: An Evaluation of BERT in the PharmaCoNER task
Taher, Ehsan Hoseini, Seyed Abbas Shamsfard, Mehrnoush [12]	Beheshti-NER: Persian named entity recognition Using BERT
Liang, Xuchao Cao, Han Zhang, Weizhen [13]	Knowledge Extraction Experiment Based on Tourism Knowledge Graph Q A Data Set

III. PEMBAHASAN

Pembahasan ini membahas mengenai hasil kajian literatur serta pembahasan literatur yang akan dianalisis. Bersumber pada riset yang dilakukan, implementasi dari NER membutuhkan sebagian tahapan untuk menghasilkan nilai keluaran yang diinginkan, dibawah ini merupakan tahap-tahap yang dilakukan pada NER.

1) Pencarian Data

Data yang diolah dapat menggunakan kumpulan data yang telah ada sebelumnya dan berlabel, atau membuat kumpulan data sendiri. Menurut tabel 2 tersebut hingga bisa disimpulkan bahwa sebagian besar literatur menggunakan kumpulan data yang dibuat sendiri oleh para peneliti, ada contoh literatur yang memakai dataset yang sudah ada sebelumnya ialah Microsoft Research Asia(MSRA)[4].

2) Pembersihan Data

Pembersihan data atau preprocessing adalah 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[5][6]. Melakukan preprocessing dalam bentuk pemecahan kalimat, *POS tagging* dan chunking. Ada juga yang melakukan chunking dengan memakai metode SpaCy noun chunk[6][7].

3) Pelabelan Data

Proses yang akan dilakukan sesudah melakukan preprocessing yaitu melakukan pelabelan data yang sudah di preprocessing atau diolah, tetapi semua literatur yang membahas di bidang wisata dalam pemberian pelabelan pada data dientitasnya jelas berbeda, seperti yang ditujukkan pada tabel 3. Pelabelan data yang menggunakan format BIO, 'B' yang berarti awal entitas, 'I' yang berarti entitas perantara dan 'O' yang berarti kata non-entitas lainnya[8]. Pelabelan data pada riset sebelumnya hanya dibidang wisata saja seperti nama orang, nama lokasi, nama organisasi, waktu dan lain-lain[9].

4) Ekstraksi fitur

Tahap ektraksi fitur ini dapat dipakai untuk mengubah data asli yang berawal dari teks menjadi berupa vektor, disebabkan dalam penggunaan *deep learning* tidak bisa mengolah data dengan data asli, maka harus diubah terlebih dahulu dari teks menjadi vektor. Menggunakan ekstraksi fitur dapat mengetahui informasi karakter dari sebuah data dan untuk menggali informasi yang berguna untuk ke tahap selanjutnya[9][6]. Word embedding yaitu system yang dapat melakukan pengubahan dari sebuah teks menjadi angka, berbagai jenis word embbeding yaitu Word2vec, GloVe dan Skip-gram[10]. Dalam pengenalan entitas ini, sebagian besar peneliti menggunakan Word2Vec untuk mendapatkan penyisipan kata dalam sebuah kata, namun word embedding tersebut tidak dapat membedakan kata yang sama karena konteks saat menyandikan kata yang menghasilkan banyak kesalahan dalam pengenalan.

5) Pemodelan NER

Langkah selanjutnya setelah ekstraksi fitur yaitu merancang sebuah model Named Entity Recognition yang terdiri menggunakan satu metode atau lebih dari satu metode didalamnya, model Named Entity Recognition berada pada tabel 3. Dalam tabel 3 tersebut bisa dilihat penggunaan model paling banyak dipakai yaitu model BERT. BERT singkatan dari *Bidirectional Encoder Representations from Transformer*. Model BERT adalah pelatihan dalam representasi kata yang berguna untuk mencegah kata agar tidak terjadi ambigu sehingga tidak terjadi kesalahan dalam pengenalan entitas[6][9]. BERT dapat disesuaikan untuk membuat model kompetitif untuk berbagai tugas hilir seperti pengenalan nama, ekstraksi relasi dan penjawab pertanyaan[11]. BERT juga menjadi solusi dari transfer learning yang baik untuk memecahkan masalah sumber daya yang terendah[12].

6) Evaluasi

Evaluasi adalah pengukuran kinerja model yang telah dibangun yang digunakan untuk pengujian data, untuk setiap kategori entitas, hasil ekstraksi biasanya diukur dengan nilai *Precission*, *Recall* dan F1-*Score* atau F-*Measure*[13].

IV. DISKUSI

NER literatur, metode yang digunakan dan hasil *F1-Score* dari penelitian sebelumnya bisa dilihat di tabel 3.

TABLE III. TABEL OVERVIEW NER LITERATUR, METODE YANG DIGUNAKAN DAN HASIL F1-*SCORE* DARI PENELITIAN SEBELUMNYA

Referensi	Dataset	Ukuran Dataset	Bahasa Dataset	Model	Label Entitas	F1-Score
[2]	Data diperoleh dari hubungan hierarkis	8.915 Kalimat	Slovenian dan English	Gazetteer	Person, location, organization, time, thing, other.	Tidak disebutkan
[3]	Data diperoleh dari website, artikel	8.700 kata	Malay	Rule Based	Person, location, organization	Mendapatkan *F1-Score :* 89.47%
[4]	MSRA	55.280 Kalimat	Cina	BERT – ResCNNs – BLSTM – CRF	Location, organization, person	95.41%
	CTFAE	15.845 Kalimat	Cina	BERT – ResCNNs – BLSTM – CRF	Include area, construction time, location, nickname, internal attraction	92.17%
[5]	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.	Mendapatkan *F1-Score* : 69.1%
[6]	Tidak disebutkan	Tidak disebutkan	Inggris	BiLSTM dan CRF	Tidak disebutkan	Tidak disebutkan
[7]	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%

[8]	Wikipedia	50.000 kalimat	Inggirs dan Spanish	CRF-Multilingual BERT	proteins, genes, and related entities, chemicals that can be normalized to external re-sources,miscellaneous related entities	BERT : 87% CRF-Multilingual BERT : 88%R
[9]	Data teks diperoleh dari perjalanan yang dilakukan Ctrip dan Mafengwo	13. 464 Kalimat	Cina	BERT – BLSTM – CRF (BBLC	Person, location, organization, time, thing, other.	*F1-Score* Entitas person mendapatkan : 84,79 *F1-Score* Entitas location mendapatkan : 91,46 *F1-Score* Entitas organization mendapatkan : 71,13 *F1-Score* Entitas time mendapatkan : 85,22 *F1-Score* Entitas thing mendapatkan : 91,39
[10]	Website TripAdvisor	Tidak disebutkan	Indonesia	LSTM,BiLSTM ,CNN dan CRF	Tidak disebutkan	Tidak disebutkan
[11]	PharmaCoNER	5000 clinical cases	Inggris	Multilingual BERT, BioBERT	Chemical, protein	Multilingual BERT : 89.24% BioBERT : 89.02%
[12]	Arman Peyma	7.682 Kalimat 7.145 Kalimat	Persian	BERT	Person, Organization, Location, Facility, Product, Event, Date, Time, Percent, Money	Arman dataset : 83.5% Peyma dataset: 88.4%
[13]	Menggunakan data set dari Shaanxi Tourist Attractions Menggunakan data set dari Nlpcc universal	700 pasangan tanya jawab	Tidak disebutkan	BERT-BiLSTM-CRF	Tidak disebutkan	Dari data set Shaanxi tourist mendapatkan : 89.77% Dari data set Nlpcc universal mendapatkan : 96.93%

V. ANALISIS

Setelah melakukan analisis semua literatur mendapatkan analisis yang berhubungan dengan metode yang digunakan dalam penelitian sebelumnya, pada riset [2] metode yang akan digunakan adalah metode Gazetteer, karena proses Gazetteer mendapatkan representasi yang dinormalisasi dari setiap entitas yang ada, ketika entitas diberikan dalam bentuk hierarki setiap nama simpul di pohon hierarki dianggap sebagai entitas. Pada riset [3] metode yang akan digunakan adalah metode Gazetteer, karena proses Gazetteer mendapatkan representasi yang dinormalisasi dari setiap entitas yang ada, ketika entitas diberikan dalam bentuk hierarki setiap nama simpul di pohon hierarki dianggap sebagai entitas. Pada penelitian [4] metode yang digunakan BERT yang dikombinasi dengan ResCNNs-BLSTM-CRF. Kombinasi beberapa metode tersebut berhasil mendapatkan hasil F1-*Score* yang cukup tinggi yaitu 95.41%. Pada penelitian [5] penerapan 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 di implementasikan. Pada tahap kedua algoritma ini mencoba untuk memprediksi data tidak berlabel dengan menggunakan pendekatan K-Nearrest Neighbor. Pada penelitian [6] metode yang diterapkan adalah BiLSTM dan CRF, karena tingkat kepopuleran metode CRF sangat tinggi dalam *Named Entity Recognition*, sehingga akan kombinasikan dengan metode

BiLSTM. Pada riset [7] penerapan metode menggunakan *Library SpaCy* dan BERT. Penerapan metode *Library SpaCy* karena telah dibuat model yang mengidentifikasi nama organisasi, nama tempat dan fasilitas. Penerapan metode BERT karena menggunakan format BIO Tag yang mengidentifikasi nama organisasi, nama tempat dan fasilitas. Pada riset [8] metode yang akan digunakan adalah metode CRF dan Multilingual BERT, metode CRF adalah metode yang populer dan efektif untuk pelabelan urutan dan dengan demikian merupakan dasar relevan dalam mencari entitas NER. Penerapan metode BERT dengan menggunakan data berformat CoNLL yang beridentik dengan metode CRF.Pada penelitian [9] penerapan metode yang digunakan yaitu mengkombinasikan 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. Pada penelitian [10] metode yang diterapkan adalah BiLSTM pada domain wisata, karena dalam domain wisata masih sedikit yang menggunakan metode BiLSTM sehingga peneliti ini memilih metode BiLSTM untuk diterapkan di NER dibidang wisata. Pada penelitian [11], metode yang digunakan Multilingual BERT dan BioBERT karena mendapatkan F1-score yang cukup baik dengan hasil 89.24% dan 89.02%, NER yang berhasil dikombinasikan dengan metode Multilingual BERT dan BioBERT. Pada penelitian [12] penerapan metode BERT dalam NER ini dilakukan karena BERT memiliki model bahasa dua arah yang telah dilatih sebelumnya yang telah disesuaikan sehingga dapat membangun model pengenalan entitas bernama dalam bahasa Persia. Pada penelitian [13] metode yang digunakan dalam NER yaitu BERT-BiLSTM-CRF yang mendapatkan hasil yang cukup baik dengan dataset Shaanxi Tourist yaitu 89.77% dan mendapatkan hasil yang cukup baik dengan dataset Nlpcc Universal yaitu 96.93%.

VI. KESIMPULAN

Dapat dilihat dari analisis bahwa literatur NER yang membahas bidang wisata masih belum populer. Masih banyak pilihan untuk entitas karena bidang pariwisata sangat luas, maka dari itu yang dapat diidentifikasi dibidang pariwisata termasuk nama tempat, nama organisasi, fasilitas dan masih banyak lagi. Pengenalan dalam entitas yang sangat diperlukan,sehingga dalam melakukan penelitian peneliti harus menyesuaikan dengan kebutuhan, tetapi setiap peneliti memiliki kebutuhan yang berbeda maka peneliti harus membuat atau menyiapkan *dataset* untuk memenuhi kebutuhan dalam melakukan penelitiannya.

Dalam penelitian ini mengkaji 12 literatur yang berhubungan dengan NER pada bidang wisata maupun domain diluar wisata yang diperoleh melalui *Google Scholar*. Dari hasil riset, informasi yang telah didapatkan bahwa saat ini penelitian NER pada domain wisata masih tergolong belum banyak dan model yang populer digunakan adalah BERT(*Bidirectional Encoder Representation from Transformer*), Model BERT adalah pelatihan dalam representasi kata yang berguna untuk mencegah kata agar tidak menjadi ambigu sehingga tidak terjadi kesalahan dalam pengenalan entitas. Penerapan NER pada domain wisata harus dikembangkan lagi sehingga dapat mengidentifikasi kelas entitas yang lebih banyak dan lebih luas.

REFERENCES

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

ORIGINALITY REPORT

11%	10%	3%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journal.uii.ac.id Internet Source	8%
2	Xuchao Liang, Han Cao, Weizhen Zhang. "Knowledge Extraction Experiment Based on Tourism Knowledge Graph Q & A Data Set", 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2020 Publication	1%
3	dblp.dagstuhl.de Internet Source	1%
4	arxiv.org Internet Source	<1%
5	repository.tudelft.nl Internet Source	<1%
6	Chantana Chantrapornchai, Aphisit Tunsakul. "Information Extraction Tasks based on BERT and SpaCy on Tourism Domain", ECTI	<1%

Transactions on Computer and Information Technology (ECTI-CIT), 2021
Publication

7	repository.usu.ac.id Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		