Tinjauan Literatur : *Named Entity Recognition* pada Resep Makanan Indonesia

Febrianto Eko Saputro
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM 14.5,Sleman,
Yogyakarta, Indonesia
17523165@students.uii.ac.id

Ahmad Fathan Hidayatullah
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM 14.5, Sleman,
Yogyakarta, Indonesia
fathan@uii.ac.id

***Abstract*— Makalah ini bertujuan untuk mengkaji literatur mengenai penerapan *Named Entity Recognition* pada data resep makanan. Metode *Named Entity Recognition* adalah salah satu subtugas ekstraksi informasi yang berfokus pada pengenalan dan identifikasi unit informasi berupa entitas bernama. Literatur ini mengkaji 5 referensi literatur mengenai NER pada domain masakan yang didapat melalui *Google Scholar* dengan kata kunci pencarian "*Named Entity Recognition for Food Ingredients*". Pada penelitian ini, akan dilakukan analisis mengenai metode, jenis entitas, dan bahasa yang digunakan pada NER untuk data resep masakan. Model yang sering digunakan adalah model rule-based FoodIE. Jenis entitas yang ditemukan meliputi *name, state, unit, quantity, size, temp, dry/fresh, titles, categories, ingredients, steps, appetizers/snacks, breakfast/lunch, desserts, dinner* dan *drinks.* Mengenai bahasa yang digunakan penelitian, bahasa yang digunakan adalah bahasa Inggris.Tinjauan literatur NER terbanyak ditemukan di bidang kuliner dengan menggunakan model. Hasil dari penelitian ini diharapkan dapat membantu pengembangan NER pada bidang kuliner.**

Keywords— NER, kuliner, FoodIE

I. PENDAHULUAN

Named Entity (Entitas bernama) adalah kata atau suatu ekspresi yang secara unik menggambarkan suatu elemen di antara sekumpulan elemen lain dengan atribut yang serupa. Organisasi, nama orang, nama tempat dan penyakit yang ada di bidang biomedis termasuk dari entitas bernama [1]. *Named Entity Recognition* adalah salah satu subtugas ekstraksi informasi yang berfokus pada mengenali dan mengidentifikasi unit informasi seperti nama orang dan nama makanan. Kategoriasi entitias bernama dapat berbeda berdasarkan tujuan tugas NER. Sebagai contoh, diperlukan *preprocessing* seperti *Information Retrival*(Zhang) [2]. Dalam masalah ini, informasi seperti bahan, bumbu, dan alat untuk memasak dapat membantu pengguna dalam mendapatkan informasi itu. .

Oleh karena itu, tinjauan literatur ini bertujuan untuk melakukan perbandingan beberapa metode pada NER, agar mengetahui apa yang sudah digunakan pada metode NER penelitian sebelumnya dan melihat hasilnya. Tinjauan literatur ini dilakukan untuk mengetahui perkembangan penlitian di bidang NER, terutama NER yang diterapkan pada data kuliner dan resep. Selain itu, di penelitian ini juga akan mengungkap tentang metode apa saja yang digunakan untuk melakukan NER. Tinjauan literatur ini dapat mempunyai kontribusi serta menunjang peneliti untuk memastikan metode-metode yang baik untuk NER dan yang paling utama di bidang kuliner.

II. METODOLOGI PENELITIAN

Dalam proses pencarian literatur, ada beberapa tahapan yang dilakukan. Pertama, mencari literatur melalui *Google Scholar* dengan batasan tahun dari tahun 2017 hingga 2021. Kedua, kata kunci yang digunakan untuk mencari literatur adalah "*Named Entity Recognition* for food ingredients". Setelah dilakukan pencarian untuk penelitian terkait, didapatkan beberapa hasil penelitian yang berhubungan dengan NER. Namun dari hasil-hasil tersebut, yang akan diambil sebagai kajian dalam penelitian ini merupakan penelitian yang hanya melakukan NER dalam bidang kuliner. Pada Tabel 1, dapat dijelaskan untuk literatur yang telah di dapatkan menggunakan Bahasa Inggris dan metode yang digunakan dalam literatur tersebut menggunakan deep learning dan rule-based. Dalam mengambil informasi-informasi penting dari penelitian yang telah didapatkan, cara yang digunakan adalah mengkaji dan mengamati langkah-langkah dan hasil dari penelitian tersebut.

TABLE I. TABEL NAMA PENELITI DAN JUDUL PENELITIAN

Nama Peneliti	**Judul**
(Willis et al., 2017) [3]	Forage: Optimizing Food Use With Machine Learning Generated Recipes
(Popovski et al., 2019) [4]	Foodie: A rule-based named-entity recognition method for food information extraction
(Mahalakshmi et al., 2020) [5]	Exploiting Bi-LSTMs for Named Entity Recognition in Indian Culinary Science
(Diwan et al., 2020) [6]	A named entity based approach to model recipes
(Popovski et al., 2020) [7]	A Survey of Named-Entity Recognition Methods for Food Information Extraction

III. HASIL

Bab ini membahas mengenai hasil kajian literatur serta ulasan terhadap literatur-literatur yang akan dikaji. Bersumber pada penelitian yang dilakukan, implementasi NER membutuhkan beberapa tahapan untuk menciptakan hasil yang diinginkan. Berikut adalah tahapan yang dilakukan pada proses NER.

1) Pengumpulan Data

Data yang akan diolah menggunakan *dataset* yang baru atau *dataset* yang sudah ada sebelumnya dan diberi label. Dapat dilihat rincian *dataset* yang digunakan pada setiap literatur pada Tabel 2 yang menunjukkan *dataset* apa saja yang digunakan pada penelitian sebelum-sebelumnya dan jumlah datanya. *Dataset* pada Tabel 2 berbentuk dokumen. *Dataset* yang mereka gunakan kebanyakan berasal dari website seperti Allrecipes dan MyRecipes. *Dataset* dikategorikan pada tujuh bagian yaitu resep untuk *breakfast, dessert, appetizers, drinks, snacks, dinner* dan *lunch*.

TABLE II. TABEL DATASET DAN JUMLAH DATA

Judul	Dataset	Jumlah
Forage: Optimizing Food Use With Machine Learning Generated Recipes	Meal-Master[3]	60.000 recipes
Foodie: A rule-based named-entity recognition method for food information extraction	Allrecipes[4]	1000 recipes
Exploiting Bi-LSTMs for Named Entity Recognition in Indian Culinary Science	Hebbars kitchen (https://hebbarskitchen.com/).[5]	3550 recipes
A named entity based approach to model recipes	RecipeDB[6]	118.000 recipes
A Survey of Named-Entity Recognition Methods for Food Information Extraction	FoodBase[7]	1000 recipes

2) Preprocessing

Tahapan *preprocessing* digunakan untuk mengolah data yang mentah menjadi data yang siap diolah atau siap digunakan pada proses selanjutnya. Proses *preprocessing* ini bermacam-macam dan proses *preprocessing* data tergantung pada kebutuhan penelitian. Metode *preprocessing* yang digunakan untuk tahapan ini ada *coreNLP tagger, URCEL Semantic Analysis System (USAS)* dan *WordNet Lemmatizer*. Metode tersebut digunakan karena dapat menghilangkan tanda kutip, karakter spesial seperti [&, %, $] dan mengkonversi text nya ke *lower case*. *Preprocessing* memiliki keuntungan seperti mengklasifikasi istilah bahan dengan benar yang berbeda dalam pluralitas, kapitalisasi, kehadiran tanda hubung sebagai entitas yang identitik. Misalnya, baik "tomat" dan "Tomat" adalah output sebagai "tomat", sehingga membuat data yang dihasilkan lebih berguna untuk tujuan yang berbeda [6]. Penggunaan metode *preprocessing coreNLP, USAS* dan *WordNet Lemmatizer* digunakan pada penelitian [3] dan [6].

3) Pelabelan Data

Setelah melakukan *preprocessing,* tahapan selanjutnya adalah pelabelan data. Pelabelan data sangat berguna karena pelabelan data sendiri adalah proses untuk mengidentifikasi data mentah dan menambahkan informasi yang berarti dan informatif untuk menyediakan konteks kepada data sehingga model bisa mempelajari data tersebut. *Stanford POS Twitter* digunakan pada penelitian [6] untuk melakukan *POS Tagging* ke semua data yang ada di dataset mereka. Tujuan penggunan *Stanford POS* ini karena sebagian data bukan kalimat lengkap secara gramatikal.

4) Ekstraksi Fitur

Tahapan ekstraksi fitur sangat membantu dalam pengolahan data karena *deep learning* tidak bisa mengolah semua data terutama jika data tersebut berupa teks maka data mentahan diubah menjadi vector sehingga model dapat bekerja dengan lebih efektif. Penelitian [3] menggunakan *Word2Vec* untuk mengonversi token kata dalam data pelatihan ke fitur yang direpresentasikan vektor untuk model LSTM. Ini sangat membantu untuk model LSTM untuk mempelajari arti kata dan menghasilkan hasil yang memuaskan.Di dalam model *Word2Vec,* mereka menggunakan pendekatan *Skip—ram* karena skip-gram bisa memprediksi sumber konteks kata-kata dari kata target[3].

5) Pemodelan Named Entity Recognition

Langkah selanjutnya adalah membangun model NER yang terdiri dari banyak metode di dalamnya. Model yang digunakan pada literature sebelumnya bisa lihat di Tabel 3. Beberapa ada yang mengkombinasikan meotde lain seperti BiLSTM, LSTM, CRF dan RNN. Model *rule-based* juga di gunakan seperti FoodIE dan NCBO. NER pada domain masakan. yang menggunakan *deep learning* telah dilakukan oleh penelitian [3] dan [5]. Model NER pada domain masakan yang menggunakan *rule-based* telah dilakukan oleh penelitian [6], [4] dan [7].

6) Evaluasi

Pada tahapan ini, proses evaluasi dilakukan kepada data pengujian untuk mengukur kinerja model NER yang sudah dibuat dan dapat diukur dengan nilai *Precision, Recall dan F1 Score* [7]. Dari Tabel 3, dapat diketahui penelitian [3] mendapatkan hasil F1 score 0.885. Penelitian [4] memperoleh hasil 0.9605. Kemudian penelitian [5] medapatkan hasil F1 score sebesar 94.66%. Penelitian [6]mendapatkan hasil sebesar 0.9611 dan pada penelitian [7] dengen menggunakan metode FoodIE mendapatkan hasil sebesar 96.05%.

IV. PEMBAHASAN

Bagian ini akan membahas semua literatur yang telah dikaji. Penelitan [6] menggunakan model Stanford NER dengan *K-Means Clustering* yang digunakan untuk mengelompokkan vector dari bahan-bahan. Kelompok-kelompok tersebut di bentuk atas dasar frekuensi tag melalui approach *Bag-of-Words.* Keputusan pemilihan jumlah cluster yang terbentuk didasarkan pada dua factor: kelemahan terbentuk nya kelompok dan interpretasi kelompok. Bahasa *dataset* yang digunakan dalam penelitian ini menggunakan bahasa Inggris. Entitas yang diterapkan adalah *name, state, unit, quantity, size, temp* dan *dry/fresh*. Penelitian [3] menggunakan model LSTM dan RNN karena kemampuannya untuk mengingat urutan data, menangkap dependensi jarak jauh hingga menghasilkan data baru. *K-Means Clustering* digunakan untuk memisahkan *dataset* resep menjadi delapan-puluh-delapan kelompok. Bahasa *dataset* yang digunakan dalam penelitian ini menggunakan bahasa Inggris. Entitas yang diterapkan adalah *title, categories, ingredients* dan *steps*. Penelitian [4] menggunakan model FoodIE. FoodIE bekerja dengan data tidak terstruktur (lebih khusus, dengan resep yang menyertakan data tekstual berupa instruksi tentang cara menyiapkan hidangan) dan menggunakan *coreNLP* dan *URCEL Semantic Analysis System* (*USAS*). Metode tersebut digunakan karena dapat menghilangkan tanda kutip, dan mengkonversi text nya ke *lower case*. Bahasa *dataset* yang digunakan dalam penelitian ini menggunakan bahasa Inggris. Entitas yang diterapkan adalah *appetizers/snacks, breakfast/lunch, desserts, dinner* dan *drinks.* Penelitian [5] menggunakan model Bi-LSTM, CNN dan CRF. CNN kuat dan sangat akurat dalam mengekstraksi informasi morfologis tertentu dari kata-kata dan mengubahnya menjadi representasi *neural*. Bi-LSTM menggunakan dua *hidden-states* untuk mengasimilasi masa lalu dan menyajikan informasi dan mereka digabungkan untuk mengeluarkan hasil akhir. CRF menggunakan *Conditional Probabilities* ke semua label yang memungkinkan sebelum memilih yang terbaik. Bahasa *dataset* yang digunakan dalam penelitian ini menggunakan bahasa Inggris. Entitas yang diterapkan adalah *Ingredients.* Penelitian [7] menggunakan model FoodIE dan NCBO. *The National Center for Biomedical Ontology* (*NCBO*) adalah ontology-based layanan web yang digunakan untuk menganotasi data tekstual tidak terstruktur dengan entitas ontologi biomedis. Ini terdiri dari dua langkah. Langkah pertama terdiri dari pilihan *dictionary* dengan domain yang sesuai. *Dictionary* dibangun dengan menggabungkan semua nama entitas yang termasuk dalam domain yang diminati. Langkah kedua adalah anotasi yang menggunakan metode Mgrep untuk mengenali entitas dengan menggunakan pencocokan string pada *dictionary.* Bahasa *dataset* yang digunakan dalam penelitian ini menggunakan bahasa Inggris. Entitas yang diterapkan adalah *appetizers/snacks, breakfast/lunch, desserts, dinner* dan *drinks.*

Setelah dianalisis dapat diketahui bahwa, ada dua pendekatan yang berbeda yaitu *deep learning* dan *rule-based*. Pendekatan *deep learning* yang digunakan LSTM, Bi-LSTM, RNN, CNN dan CRF. Pendekatan *rule-based* yang digunakan FoodIE, Stanford NER dan NCBO. Hasil performa menggunakan pendekatan *rule-based* diantaranya [4], [6] dan [7] lebih baik daripada menggunakan pendekatan *deep learning* diantaranya [3] dan [5].

TABLE III. TABEL MODEL DAN HASIL

Judul	Model	Hasil
Forage: Optimizing Food Use With Machine Learning Generated Recipes	LSTM RNN[3]	LSTM dengan RNN memperoleh hasil sebesar 0.885
Foodie: A rule-based named-entity recognition method for food information extraction	FoodIE [4]	Mendapatkan hasil F1 score 0.9605
Exploiting Bi-LSTMs for Named Entity Recognition in Indian Culinary Science	BiLSTM CNN CRF[5]	Mendapatkan hasil F1 score 94.66%
A named entity based approach to model recipes	Stanford NER *K-means Clustering*[6]	Mendapatkan hasil F1 score 0.9611
A Survey of Named-Entity Recognition Methods for Food Information Extraction	FoodIE NCBO(FoodOn, OntoFood, SNOMED CT) [7]	F1 score FoodIE : 96.05% F1 score SNOMED CT: 63.75% F1 score OntoFood : 32.62% F1 score FoodOn : 63.90%

V. KESIMPULAN

Penelitian ini mengkaji 5 literatur mengenai NER pada domain masakan. Yang didaptkan melalui *Google Scholar* dalam rentang waktu tahun 2017 hingga 2021. Berdasarkan literatur yang dikaji, ada dua pendekatan yang digunakan yaitu

deep learning dan *rule-based.* Dari hasil performa, performa *rule-based* lebih baik dari *deep learning.* Dari beberapa literatur yang diperoleh, model yang sering digunakan yaitu FoodIE. Literatur yang telah ditemukan menggunakan *dataset* bahasa Inggris. Dalam literatur yang dikaji, entitas yang telah ditemukan yaitu *name, state, unit, quantity, size, temp, dry/fresh, titles, categories, ingredients, steps, appetizers/snacks, breakfast/lunch, desserts, dinner* dan *drinks.* Penerapan NER pada domain masakan harus dikembangkan lagi sehingga dapat mengidentifikasi kelas entitas yang lebih banyak.

REFERENCES

[1] B. A. Ben Ali, S. Mihi, I. El Bazi, and N. Laachfoubi, "A recent survey of Arabic named entity recognition on social media," *Rev. d'Intelligence Artif.*, vol. 34, no. 2, pp. 125–135, 2020, doi: 10.18280/ria.340202.

[2] Y. Zhang, "Named Entity Recognition for Social Media Text," p. 32, 2019, [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1366031&dswid=-8297.

[3] A. Willis, E. Lin, and B. Zhang, "Forage: Optimizing Food Use With Machine Learning Generated Recipes," 2017.

[4] G. Popovski, S. Kochev, B. K. Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction," *ICPRAM 2019 - Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods*, no. Icpram, pp. 915–922, 2019, doi: 10.5220/0007686309150922.

[5] G. S. Mahalakshmi, M. N. Sreedhar, R. K. Selvam, and S. Sendhilkumar, "Exploiting Bi-LSTMs for Named Entity Recognition in Indian Culinary Science," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3545088.

[6] N. Diwan, D. Batra, and G. Bagler, "A named entity based approach to model recipes," *Proc. - 2020 IEEE 36th Int. Conf. Data Eng. Work. ICDEW 2020*, pp. 88–93, 2020, doi: 10.1109/ICDEW49219.2020.000-2.

[7] G. Popovski, B. K. Seljak, and T. Eftimov, "A Survey of Named-Entity Recognition Methods for Food Information Extraction," *IEEE Access*, vol. 8, pp. 31586–31594, 2020, doi: 10.1109/ACCESS.2020.2973502.