BAB 2

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penelitian ini merupakan penelitian yang membahas mengenai ekstraksi data tidak terstruktur menggunakan *Named Entity Recognition*. Penelitian ini sudah dilakukan oleh peneliti sebelumnya. Adapun tinjauan pustaka penelitian ini sebagai berikut.

Penelitian yang dilakukan oleh Dayinta Warih Wulandari (2018) dengan judul "*Named Entity Recognition* (NER) pada Dokumen Biologi Menggunakan *Rule Based* dan *Naive Bayes Classifier*" menggunakan objek penelitian berupa dokumen biologi. Hasil penelitian adalah pada percobaan *Rule Based* nilai rata-rata menggunakan *micro average* yaitu *precision* 0,855, *recall* 0,855 *dan f-measure* 0,855.

Penelitian yang dilakukan oleh Desi Atika (2021) dengan judul "Ekstraksi Informasi Berita Online dengan *Named Entity Recognition* (NER) dan *Rule-Based* untuk Visualisasi Penyakit Tropis di Indonesia" menggunakan objek penelitian berupa berita online. Hasil penelitian adalah perhitungan *Evaluation Scorer* yaitu *precision* 84%, *recall* 73% dan *f-score* 82%.

Penelitian yang dilakukan oleh Ni Made Sinta Wahyuni dan Ngurah Agus Sanjaya (2021) dengan judul "*Rule-based Named Entity Recognition (NER) to Determine Time Expression for Balinese Text Document*" menggunakan objek

penelitian berupa dokumen teks berbahasa Bali. Hasil penelitian adalah Metode *Rule Based* NER berhasil mengekstraksi entitas dari 112 ayat Surat Al-Anbiya dengan nilai f-score 0,94.

Penelitian yang dilakukan oleh Dwi Adi Bangkit (2022) dengan judul "Implementasi *Named-Entity Recognition* dan *Optical Character Recognition* untuk Aplikasi Pendeteksi Kehalalan Bahan Makan" menggunakan objek penelitian berupa komposisi bahan makanan. Hasil penelitian adalah model *Named Entity Recognition* yang dibangun pada penelitian mampu membaca entitas Halal, Haram, Syuhbat dengan rata-rata *f-score* 0,967.

Penelitian yang dilakukan oleh Shasha Arzila Tarmizi dan Saidah Saad (2022) dengan judul "*Named Entity Recognition for Quranic Text Using Rule Based Approaches*" menggunakan objek penelitian berupa teks Al-Quran. Hasil penelitian adalah NER dengan *Rule Based approach* dapat diimplementasikan pada dokumen teks bahasa Bali dan mampu memberikan hasil sesuai aturan, dengan hasil rata-rata nilai *presisi* 0,85, *recall* 0,87, dan *f-measure* 0,85.

Adapun perbedaan penelitian ini dengan penelitian-penelitian di atas adalah objek yang digunakan pada penelitian, yaitu dokumen daftar riwayat hidup pelamar kerja. Ringkasan perbandingan dengan penelitian sebelumnya dapat dilihat pada **Tabel 2.1**.

Tabel 2.1 Tinjauan Pustaka

No	Penulis	Judul	Metode	Objek Penelitian
1	Wulandari (2018)	*Named Entity Recognition (NER)* pada Dokumen Biologi Menggunakan *Rule Based* dan *Naive Bayes Classifier*	*Named Entity Recognition* dengan pendekatan *Rule Based*	Dokumen biologi
2	Atika (2021)	Ekstraksi Informasi Berita Online	*Named Entity*	Berita *Online*

No	Penulis	Judul	Metode	Objek Penelitian
		dengan *Named Entity Recognition* (NER) dan *Rule-Based* untuk Visualisasi Penyakit Tropis di Indonesia	*Recognition* dengan pendekatan *Rule Based*	
3	Wahyuni dan Sanjaya (2021)	*Rule-based Named Entity Recognition* (NER) *to Determine Time Expression for Balinese Text Document*	*Named Entity Recognition* dengan pendekatan *Rule Based*	Dokumen teks berbahasa Bali
4	Bangkit (2022)	Implementasi *Named-Entity Recognition* dan *Optical Character Recognition* untuk Aplikasi Pendeteksi Kehalalan Bahan Makanan	*Named Entity Recognition* dengan pendekatan *Rule Based*	Komposisi bahan makanan
5	Tarmizi dan Saad (2022)	*Named Entity Recognition for Quranic Text Using Rule Based Approaches*	*Named Entity Recognition* dengan pendekatan *Rule Based*	Teks Al-Quran
6	Risky Eliana Dewi	Sistem Ekstraksi Daftar Riwayat Hidup Menggunakan *Named Entity Recognition*	*Named Entity Recognition* dengan pendekatan *Rule Based*	Dokumen daftar riwayat hidup pelamar kerja

2.2 Dasar Teori

2.2.1 Bahasa Pemograman *Python*

Python merupakan bahasa pemrograman tingkat tinggi yang diinterpretasikan, berorientasi objek, interaktif dan dapat berjalan hampir di semua platform, seperti pada Windows, Linux, Mac dan lainnya. Python merupakan bahasa pemrograman tingkat tinggi yang mudah dipelajari, karena sintaksnya yang elegan dan jelas, yang dikombinasikan dengan penggunaan modul struktur data yang canggih, efisien dan siap digunakan. Kode sumber aplikasi bahasa pemrograman python biasanya diterjemahkan ke dalam *bytecode* terlebih dahulu, selanjutnya menjadi format perantara, dan kemudian dieksekusi (Ratna, 2020).

2.2.2 Ekstraksi Informasi

Ekstraksi informasi adalah proses menemukan informasi terstruktur dari dokumen yang tidak terstruktur atau terstruktur sebagian. Ekstraksi informasi dilakukan dengan mengambil semua data sesuai dengan aturan yang telah ditetapkan, kemudian data yang diambil dipisahkan untuk mendapatkan informasi yang diinginkan (Kurnia, 2020).

2.2.3 Evaluasi Skor

Evaluasi pada sistem *information Extraction,* matriks yang biasa digunakan adalah *precision* dan *recall*. *Precision* merupakan jumlah item yang telah diidentifikasi dengan benar atau relevan dari jumlah total item yang di identifikasi. Sedangkan *recall* merupakan rasio jumlah item relevan yang ditemukan dengan total jumlah item dalam kumpulan item yang dianggap relevan. *F-score* merupakan pengukuran kinerja yang diperoleh dari kombinasi nilai *precision* dan *recall*. Perhitungan untuk mengukur kekuatan dari sistem sebagai berikut (Drovo dkk., 2019):

a) Rumus *Recall*

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (1)$$

b) Rumus *Precision*

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (2)$$

c) Rumus *F-score*

$$\mathrm{F}_1 = 2\,\frac{precision\,.\,recall}{precision\,+\,recall} \quad (3)$$

Keterangan

True Positive (TP) : Jumlah prediksi dimana pengklasifikasian dilakukan dengan benar memprediksi kelas positif sebagai positif.

False Positive (FP) : Jumlah prediksi dimana pengklasifikasian salah memprediksi kelas negatif sebagai positif.

False Negative (FN) : Jumlah prediksi dimana pengklasifikasian salah memprediksi kelas positif sebagai negatif.

2.2.4 *Framework Flask Python*

Framework Flask Python adalah *framework web* bahasa *Python* dan termasuk dalam kategori *microframework*. *Flask* bertindak sebagai kerangka aplikasi dan antarmuka *web*. Pengembang dapat membuat halaman web yang terstruktur dan mengelola interaksi dengan mudah menggunakan *flask* dan *python* (Irsyad, 2018).

Fitur yang termasuk dalam *framework flask* adalah *debugger* cepat, server pengembangan terintegrasi, dukungan pengujian unit terintegrasi, kompatibilitas dengan mesin aplikasi *RESTful request dispatching*, *Google*, *Jinja 2 templating* berbasis *Unicode* dan mendukung untuk *secure cookie*. *Flask* juga memiliki berbagai kelebihan seperti (Yunius, 2017):

1. Ringan karena memiliki *core* sederhana dan desain modular.
2. Mampu untuk menangani fungsi HTTP *request* dengan lebih mudah.
3. API yang koheren dan baik.
4. Mempunyai banyak dokumentasi dan terstruktur dengan baik, penuh dengan contoh yang dapat digunakan langsung.

5. Mudah untuk diimplementasi saat produksi.
6. Mudah untuk dikontrol secara menyeluruh.

2.2.5 *Named Entity Recognition* (NER)

Named Entity Recognition (NER) adalah bagian dari ekstraksi informasi dalam pemrosesan bahasa alami. NER adalah langkah pertama dalam ekstraksi informasi dimana NER digunakan untuk mengidentifikasi dan mengklasifikasikan kata atau frasa ke dalam jenis entitasnya (Dirgantara dkk., 2020).

2.2.6 *PostgreSQL*

PostgreSQL (Post-gress-SQL) adalah sistem manajemen basis data relasional, fungsi utamanya adalah menyimpan data dengan aman dan mengembalikan data sesuai permintaan dari aplikasi perangkat lunak lainnya. Itu dapat menangani beban kerja mulai dari aplikasi mesin tunggal kecil hingga aplikasi Internet besar dengan banyak pengguna bersamaan. Di *macOS Server*, *PostgreSQL* adalah *database default*, *PostgreSQL* juga tersedia untuk *Microsoft Windows* dan *Linux* (disertakan di sebagian besar distribusi) (Setyawan dan Pratiwi, 2020).

2.2.7 REST API (*Representational State Transfer Application Programming Interface*)

REST API mendukung beragam sistem untuk berinteraksi mengirim atau menerima data dengan mudah. Setiap penggunaan REST API didukung oleh URL dan HTTP. REST API menentukan request dan response menggunakan HTTP. Data dalam database pada suatu aplikasi dihubungkan dengan endpoint API pada REST API. Agar informasi dapat lebih mudah dibaca dan di analisa pada sisi

client maka keluaran yang dihasilkan oleh API server berupa JSON. Beberapa method tersedia untuk melakukan komunikasi data, antara lain (Afrianto dan Cahyono, 2022):

1. *Method* GET digunakan untuk mendapatkan data dari *database*.
2. *Method* POST digunakan untuk menyimpan data ke dalam *database*.
3. *Method* DELETE digunakan untuk menghapus data di dalam *database*.

2.2.8 *Rule Based*

Rule-based NER didasarkan pada formulasi aturan yang dibuat sendiri oleh para *engineer* dan *mapping* kamus bahasa untuk mengidentifikasi dan melakukan ekstraksi *named entities*. Keluaran dari pendekatan ini dilakukan dengan memeriksa *rule* dan *mapping* kamus. Untuk setiap jenis klasifikasi yang berbeda, maka aturan yang berlaku juga berbeda. Setiap sistem mendapatkan teks, pertama yang akan dilakukan adalah sistem akan mencari *named entities* dan kemudian melakukan perbandingan dengan *rule* yang sudah ada. Ketika ada *rule* yang cocok, sistem akan memberikan *output* yang diklasifikasikan. Kinerja dari sistem ini akan sangat bergantung pada *coverage* dari *rule* itu sendiri. *Rule-based* adalah sistem yang cocok untuk *often* domain *spesific*, sistem ini tidak bisa digunakan pada *new* domain. Sistem ini juga akan menghasilkan output yang lebih baik jika menggunakan *restricted* domain. Pendekatan ini akan dapat mendeteksi *named entities* dengan kesalahan ejaan (Sharma dan R, 2018).

2.2.9 *SpaCy*

SpaCy merupakan sumber pustaka terbuka (*open-source library*) yang biasa digunakan untuk *Natural Language Processing* tingkat lanjut yang ditulis

dengan Bahasa *Python*. *SpaCy* bisa digunakan untuk melakukan proses ekstraksi informasi atau biasa disebut *Natural Language Processing* dalam memproses data berupa teks untuk digunakan dalam pembelajaran yang lebih mendalam. Sistem pengenalan entitas statistik yang ditampilkan *spaCy* sangat cepat, sehingga dapat memberikan label ke dalam rentan token yang berdekatan. *SpaCy* berfokus untuk menyelesaikan sesuatu hal dengan pendekatan akademis. Fitur yang tersedia dalam *library spaCy* yaitu *Tokenization*, *POS-Tagging, Text Classification*, dan *Named Entity Recognition* (Desikan, 2018).