NAMED ENTITY RECOGNITION (NER) BAHASA INDONESIA BERBASIS *MULTI CLASS CLASSIFICATION*

(Skripsi)

Oleh

VINDO RIZKIYANTO
1817051059

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

ABSTRAK

NAMED ENTITY RECOGNITION (NER) BAHASA INDONESIA BERBASIS *MULTI CLASS CLASSIFICATION*

Oleh

VINDO RIZKIYANTO

Penelitian ini bertujuan untuk menghasilkan sistem yang dapat mengekstraksi informasi berupa data teks Bahasa Indonesia dan mendeskripsikan hasil implementasi *Named Entity Recognition* berbasis *Multi Class Classification*. Metode yang digunakan adalah *extreme programming*. Penelitian ini juga melakukan beberapa tahapan seperti studi literatur, pembuatan model, perancangan sistem, dan pengujian sistem. Jumlah kata yang diperoleh dari hasil pengumpulan data adalah 8017 kata. Dalam tahapan pengumpulan data dilakukan dengan bantuan *library beautifulSoup* untuk proses *scraping website* dengan melakukan *import request* untuk mendapatkan *link* berita yang akan di *scraping* dan html5lib untuk mendapatkan konten dari *website* dengan *parsing* html. Total entitas label yang diperoleh dari hasil pelabelan data adalah 1197 entitas label. Dalam penelitian ini pelabelan data dibagi menjadi 9 kelompok. Pelatihan data dilakukan dengan menerapkan 6 skenario. Hasil skenario yang didapatkan pada pelatihan data menghasilkan *precision, recall,* dan *f1-score*. Dalam tahapan proses pembagian data, penelitian ini menggunakan 2 skenario pembagian data, pertama yaitu 70% *training data* dan 30% *testing data*, kemudian pada *training* data dibagi menjadi 70% *training data* dan 30% *validation data*. Pada skenario kedua data dibagi menjadi 60% *training data* dan 40% *testing data*, kemudian pada *training data* akan dibagi menjadi 60% *training data* dan 40% *validation data*. Berdasarkan hasil dari skenario yang telah dilakukan pada proses evaluasi kinerja model dapat diketahui bahwa skenario 6 dengan pembagian 60% data *training*, dan 40% *data testing* menggunakan *hyperparameter epoch* 100, *batch size* 1000, dan *learn rate* 0.001, skenario ini merupakan konfigurasi *hyperparameter* bawaan *library spaCy* dan terbukti memiliki hasil nilai *precision, recall,* dan *f1-score* terbaik dibandingkan dengan skenario lainnya. Sistem yang telah dibuat diuji menggunakan metode black box testing, dengan hasil yang didapatkan sesuai yang diharapkan berdasarkan skenario uji yang di lakukan.

Kata kunci: *Extreme Programming, Named Entity Recognition, library beautifulSoup dan Multi Class Classificatio*n.

NAMED ENTITY RECOGNITION (NER) BAHASA INDONESIA BERBASIS *MULTI CLASS CLASSIFICATION*

Oleh

VINDO RIZKIYANTO

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA KOMPUTER

Pada

Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

Judul Skripsi	:	***NAMED ENTITY RECOGNITION* (NER) BAHASA INDONESIA BERBASIS *MULTI CLASS CLASSIFICATION***
Nama Mahasiswa	:	**Vindo Rizkiyanto**
Nomor Induk Mahasiswa	:	**1817051059**
Program Studi	:	**Ilmu Komputer**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Ir. Kurnia Muludi, M.S.Sc.
NIP. 19640616 198902 1 001

2. Ketua Jurusan Ilmu Komputer

Didik Kurniawan, S.Si., M.T.
NIP 19800419 200501 1 004

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Ir. Kurnia Muludi, M.S.Sc.** :

Penguji
Bukan Pembimbing : **Prof. Admi Syarif, Ph.D.**

Penguji
Bukan Pembimbing : **Aristoteles, S.Si., M.Si.**

2. Plt. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP 19711001 200501 1 002

Tanggal Lulus Ujian Skripsi : **28 Maret 2023**

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini adalah:

Nama	: Vindo Rizkiyanto
NPM	: 1817051059
Fakultas / Jurusan	: MIPA / Ilmu Komputer
Program Studi	: S1 Ilmu Komputer
Alamat	: Jl. Lapas Raya No. 51 Way Huwi, Jati Agung, Lampung Selatan, Lampung.

Dengan ini menyatakan bahwa, dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebut dalam daftar pustaka.

Bandar Lampung, 30 Maret 2023

Vindo Rizkiyanto
NPM. 1817051059

RIWAYAT HIDUP

Penulis lahir di Bandar Lampung, 30 Mei 2000, dilahirkan sebagai anak kedua dari dua bersaudara. Pendidikan yang telah ditempuh oleh penulis diantaranya, menyelesaikan pendidikan dasar di SDN 1 Way Huwi Lampung Selatan pada tahun 2012. Penulis menyelesaikan pendidikan menengah pertama di SMP Al-azhar 3 Bandar Lampung pada tahun 2015. Penulis melanjutkan pendidikan menengah atas di SMAN 4 Bandar Lampung dan lulus pada tahun 2018. Perjalanan pendidikan penulis dilanjutkan dengan terdaftar sebagai mahasiswa Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung pada tahun 2018 melalui jalur SBMPTN.

Penulis turut aktif mengikuti beberapa kegiatan di kampus selama menjadi mahasiswa, antara lain.

1. Menjadi anggota bidang *External* Himpunan Mahasiswa Jurusan Ilmu Komputer (Himakom) periode 2019/2020 dan 2020/2021.
2. Menjadi anggota bidang Kaderisasi Rohani Islam (Rois) FMIPA Unila periode 2019/2020.
3. Menjadi anggota bidang Kajian dan Keumatan Rohani Islam (Rois) FMIPA Unila periode 2020/2021.
4. Melaksanakan Kerja Praktik pada bulan Februari periode 2019/2020 di PDDI-LIPI.
5. Melaksanakan Kuliah Kerja Nyata (KKN) di Kelurahan Way Huwi, Kecamatan Jati Agung, Kota Bandar Lampung periode 2020/2021.

MOTTO

“Maka sesungguhnya bersama kesulitan ada kemudahan,
sesungguhnya bersama kesulitan ada kemudahan”
(Q.S. Asy-Syarh: 5-6)

“Tak ada penyakit yang tak bisa disembuhkan kecuali kemalasan. Tak ada obat yang tak berguna selain kurangnya pengetahuan"
(Ibnu Sina)

“Kemauan untuk sukses harus lebih besar dari ketakutan dan kegagalan”
(Vindo Rizkiyanto)

PERSEMBAHAN

Puji syukur kehadirat Allah SWT yang selalu memberikan limpahan rahmat dan karunia-Nya, shalawat beriring salam semoga selalu tercurah kepada Nabi Muhammad SAW beserta keluarga dan para sahabat. Dengan segenap kerendahan hati, penulis mempersembahkan karya tulis sederhana ini sebagai rasa tanggung jawab dalam menyelesaikan pendidikan dan tanda bakti kasih tulus kepada:

1. Orang tua penulis tersayang Triyanto dan Dwi Fatonah yang telah sepenuh hati membesarkan, mendidik, dan mengasihi dengan sabar. Terima kasih telah senantiasa mendoakan, menyayangi dan memberikan dukungan dengan penuh ketulusan. Semoga Allah SWT senantiasa memberikan kesehatan dan kesempatan kepada penulis, untuk selalu bisa membahagiakan serta membanggakan kalian di dunia dan akhirat.
2. Kakak penulis Tony Reza Apriyanto dan Monica Dhamayanti yang telah banyak memberikan doa dan kasih serta semangat kepada penulis.
3. Seluruh keluarga besar penulis yang telah senantiasa memberikan do'a dan segala bentuk motivasi serta perhatian yang luar biasa.
4. Para pendidik yang senantiasa memberikan pelajaran dan pendidikan terbaik dalam membimbing penulis.
5. Sahabat tercinta yang selalu ada dalam setiap langkah perjuangan penulis dan senantiasa saling mengingatkan kebaikan dan kesabaran.
6. Almamater tercinta Universitas Lampung.

SANWACANA

Alhamdulillah segala puji bagi Allah SWT, karena atas nikmat dan rahmat-Nya penulis dapat menyelesaikan penyusunan skripsi ini sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer di FMIPA Universitas Lampung.

Dalam kesempatan ini penulis mengucapkan terima kasih kepada:

1. Ibu Prof. Dr. Ir. Lusmeilia Afriani, D.E.A., I.P.M., selaku Rektor Universitas Lampung.
2. Bapak Dr. Eng. Heri Satria, S.Si., M.Si., selaku Plt. Dekan FMIPA Universitas Lampung.
3. Bapak Didik Kurniawan, S.Si., M.T., selaku Ketua Jurusan Ilmu Komputer.
4. Bapak Ardiansyah, S.Kom., M.Kom., selaku Pembimbing Akademik.
5. Bapak Dr. Ir. Kurnia Muludi, M.S.Sc., selaku Pembimbing I atas kesediaan, kesabaran dan keikhlasan dalam memberikan kritik dan saran yang positif, motivasi dan bimbingan kepada penulis selama menyelesaikan skripsi.
6. Bapak Prof. Admi Syarif, Ph.D., selaku Pembahas I yang banyak memberikan masukan dan kritik yang bersifat positif dan membangun untuk perbaikan skripsi ini.
7. Bapak Aristoteles, S.Si., M.Si., selaku Pembahas II yang banyak memberikan masukan dan kritik yang bersifat positif dan membangun untuk perbaikan skripsi ini.
8. Bapak dan Ibu dosen serta staf Ilmu Komputer Universitas Lampung yang telah membimbing penulis dalam pembelajaran di Universitas Lampung.
9. *Support system* penulis Dwi Herliani yang telah menyemangati dan memberikan dukungan penuh tanpa henti kepada penulis dengan tulus.

10. Sahabat seperjuangan M. Azriel Bintang Saputra yang telah membantu dan memberi dukungan dalam menyelesaikan skripsi ini.
11. Kepada semua pihak yang telah membantu terselesaikannya skripsi ini.

Penulis berdoa semoga semua amal dan bantuan yang telah diberikan dapat dijadikan amal sholeh serta mendapat pahala dari Allah SWT dan semoga skripsi bermanfaat.

Bandar Lampung, 30 Maret 2023

Vindo Rizkiyanto

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang Masalah

Manusia melakukan komunikasi menggunakan bahasa, hampir semua kegiatan yang dilakukan manusia menggunakan Bahasa. Bahasa dapat digunakan untuk mendapatkan suatu informasi dan membantu dalam berkomunikasi. Di dalam sebuah informasi ada dua tipe data yang didapatkan, yaitu *structured data* dan *unstructured data*. Data tidak terstruktur merupakan data informasi yang harus diproses terlebih dahulu sebelum digunakan. *File* teks, multimedia, dan jenis data lainnya adalah contoh data tidak terstruktur (Wahyunita, 2019).

Portal berita *online* merupakan sumber informasi atau data *online* yang berbentuk data tidak terstruktur atau *unstructured data.* Terdapat banyak jenis portal berita *online*, salah satunya adalah portal berita *online* pada bidang pertanian. Pertanian merupakan salah satu bidang kegiatan yang mengelola sumber daya alam hayati dengan menggunakan bantuan teknologi, modal, tenaga kerja, dan manajemen agar dapat menghasilkan komoditas pertanian yang mencakup tanaman kebutuhan pangan, hortikultura, perkebunan, dan peternakan dalam suatu agroekosistem. Bidang pertanian akhir-akhir ini mengalami penurunan terhadap kualitas produksi yang dihasilkan terutama pada pertanian Indonesia, yang dimana penurunan tersebut terjadi akibat dampak dari virus/wabah penyakit *Covid-19* yang melanda dan juga perubahan iklim cuaca yang signifikan. Banyak informasi berita yang disampaikan mengenai bidang pertanian baik dari media televisi maupun portal berita *online* yang membahas berbagai macam masalah yang terjadi pada bidang pertanian terutama terkait menurunnya kualitas produksi yang dialami bidang pertanian untuk kebutuhan pangan masyarakat.

Unstructured data dapat menimbulkan beberapa masalah saat mencoba mengurutkannya karena format dan lokasinya dapat sangat bervariasi. Namun, dengan *Natural Language Processing* (NLP), data tidak terstruktur dapat secara otomatis diformat dan dianalisis dengan benar. Dengan memanfaatkan *Machine learning* dan *Natural Language Processing* (NLP) untuk membaca teks tidak terstruktur, lalu mengategorikan dan menganalisisnya seperti yang dilakukan manusia, tetapi dalam waktu singkat dan dengan akurasi total. Manusia harus membaca teks lengkap dari sebuah dokumen untuk mengekstrak informasi secara manual. Ketika sebuah dokumen tertulis panjang, dibutuhkan beberapa saat bagi manusia untuk membacanya dan menemukan informasi yang ada pada dokumen. Oleh sebab itu, dibuatlah *Named Entity Recognition* yang digunakan untuk menemukan informasi berguna, seperti organisasi, lokasi dan nama orang seseorang dalam dokumen teks. Informasi pada dokumen teks dapat diambil lebih cepat dengan menggunakan *Named Entity Recognition*.

Penelitian melakukan berbagai tugas NLP menggunakan bantuan dari *Library spaCy*. *SpaCy* merupakan *library open source* yang berguna menangani Pemrosesan Bahasa Alami (NLP), yang digunakan untuk melakukan tugas pemrosesan POS *Tagging, Named Entity Recognition, Dependency Parsing*, dan lain sebagainya (Yanti, dkk, 2021). NER *spaCy* digunakan pada *unstructured data*, yaitu merubah *unstructured data* menjadi data yang terstruktur dengan memilih dan mengelompokkan kata atau frasa menjadi beberapa kelompok atau *Multi Class Classification*. *Multi Class Classification* adalah mengklasifikasikan elemen ke dalam kelas yang berbeda. Untuk mendapatkan informasi maka dibutuhkan suatu pendekatan alternatif menggunakan *Named Entity Recognition* (NER) untuk mengekstraksi informasi secara otomatis pada data berupa teks portal berita *online* yang tidak terstruktur tersebut agar dapat disimpan dalam bentuk data terstruktur untuk mendapatkan informasi yang diinginkan.

Penelitian yang dilakukan oleh Willyawan pada tahun 2018 yaitu menghasilkan model *Named Entity Recognition* (NER) bahasa Indonesia menggunakan data pada Wikipedia dan DBpedia, yang mana penelitian ini dibatasi dengan informasi berupa pengenalan organisasi, nama orang dan nama tempat, berlandaskan pada

data teks yang diinputkan. Penelitian yang dilakukan oleh Atika pada tahun 2021 adalah mengekstraksi informasi data pada penyakit tropis di Indonesia berupa, entitas yang terkait dengan lokasi kejadian, waktu kejadian, nama penyakit tropis dan jumlah korban yang divisualisasikan menggunakan *Rules-Based*. Penelitian ini juga memanfaatkan *Named Entity Recognition* dan *Library spaCy* untuk membangun ekstraksi informasi dan memproses teks untuk pengklasifikasian. Berdasarkan penelitian terdahulu maka peneliti melakukan penelitian terkait pembuatan model *named entity recognition* yang digunakan untuk mendapatkan informasi terstruktur dari data portal berita *online* bidang pertanian dengan pengklasifikasian ke dalam 9 kategori atau disebut dengan *multiclass classification* dan mengimplementasikan model tersebut ke dalam sistem informasi berbasis website yang dibuat menggunakan pemrograman python dengan bantuan *framework flask*.

1.2 Rumusan Masalah

Rumusan masalah penelitian ini yaitu,

1.2.1 bagaimana membuat sistem yang dapat mengekstraksi informasi berupa data teks Bahasa Indonesia?

1.2.2 bagaimana implementasi *Named Entity Recognition* berbasis *Multi Class Classification*?

1.3 Batasan Masalah

Adapun batasan masalah pada penelitian ini, yaitu:

1.3.1 penelitian ini berfokus pada data bidang pertanian,

1.3.2 teks berita yang digunakan menggunakan bahasa Indonesia,

1.3.3 penelitian ini mengambil lebih dari dua entitas yaitu nama, tanggal, hari, lokasi, penyakit, produk pertanian, musim, organisasi, harga.

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini yaitu,

1.4.1 menghasilkan sistem yang dapat mengekstraksi informasi berupa data teks Bahasa Indonesia.

1.4.2 mendeskripsikan hasil implementasi *Named Entity Recognition* berbasis *Multi Class Classification*.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1.5.1 memperoleh data yang terstruktur,

1.5.2 mengklasifikasikan informasi secara terstruktur berdasarkan pelabelan kata atau frasa,

1.5.3 mengumpulkan data yang saling berhubungan untuk diidentifikasi agar memperoleh informasi yang akurat.

1.5.4 menghasilkan sistem informasi yang dapat mendeteksi kelompok kata entitas berdasarkan kata entitas tertentu.

II. TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian ini tidak terlepas dari penelitian yang sudah dilakukan sebelumnya yang tujuannya untuk mendukung penelitian ini. Penelitian terdahulu yang digunakan sebagai referensi dalam penelitian dapat dilihat pada Tabel 1.

Tabel 1. Penelitian Terdahulu

No	Judul	Peneliti (Tahun)
1	Ekstraksi Informasi Berita Online Dengan *Named Entity Recognition* (NER) Dan *Rule-Based* Untuk Visualisasi Penyakit Tropis Di Indonesia	(Atika, 2021)
2	*Named Entity Recognition* (NER) Bahasa Indonesia Menggunakan *Conditional Random Field* dan *Pos-Tagging*	(Willyawan, 2018)
3	Implementasi *Named Entity Recognition* Pada *Factoid Question Answering System* Untuk Cerita Rakyat Indonesia	(Kurniawati, Indriati dan Adikara 2018)
4	*Named-Entity Recognition* pada Teks Berbahasa Indonesia menggunakan Metode Hidden Markov Model dan POS-Tagging	(Yusliani, Sufa, Firdaus, Abdiansah, Sazaki, 2021)
5	Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier	(Wulandari dkk, 2021)
6	Kannada Named Entity Recognition And Classification (Nerc) Based On Multinomial Naïve Bayes (Mnb) Classifier	(Amarappa, 2015)
7	Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study: Power Failure in the Special Region of Yogyakarta)	(Yanti dkk, 2021)
8	Serverless Named Entity Recognition untuk Teks Instruksional Pertanian Kota	(Gelar dkk, 2021)

No	Judul	Peneliti (Tahun)
9	Indonesian Named Entity Recognition Untuk Domain Zakat Menggunakan Conditional Random Fields	(Widiyanti, 2022)
10	Named Entity Recognition Untuk Data Review Tempat Wisata Dengan Metode "Bidirectional Encoder Representations from Transformers"	(Irfan, 2022)
11	Pengembangan Model Named Entity Recognition Untuk Pengenalan Entitas Pada Data Obat Indonesia	(Arianto, 2023)
12	Named Entity Recognition pada Resep Makanan dengan Metode Bidirectional Long Short-Term Memory dan Bidirectional Encoders Representations from Transformers	(Saputro, 2022)
13	Named Entity Recognition For Quranic Text Using Rule Based Approaches	(Tarmizi and Saad, 2022)
14	Hybrid Conditional Random Fields and K-Means for Named Entity Recognition on Indonesian News Documents	(Santoso et al, 2020)
15	Penggunaan Named Entity Recognition dan Artificial Intelligence Markup Language untuk Penerapan Chatbot Berbasis Teks	(Christianto dkk, 2015)
16	Rule-based Named Entity Recognition (NER) to Determine Time Expression for Balinese Text Document	(Wahyuni dan ER, 2021)
17	Named Entity Recognition Menggunakan Hidden Markov Model dan Algoritma Viterbi pada Teks Tanaman Obat	(Setiyoaji dkk, 2017)
18	Penerapan Named Entity Recognition Untuk Mengenali Fitur Produk Pada E-Commerce Menggunakan Rule Template Dan Hidden Markov Model	(Dirgantara dkk, 2018)
19	Ekstraksi Informasi pada Dokumen Teks Menggunakan Metode Named-Entity Recognition untuk Sistem Autoful Formulir Lowongan SIM Magang MyITS StudentConnect	(Hadinata, 2022)
20	Analisis Pembentukan Modul Named Entity Recognition (NER) berbasis Algoritma Conditional Random Field (CRF) pada Sistem Repositori LIPI	(Riyanto, 2017)

2.1.1 Ecstasy Informasi Berita Online Dengan *Named Entity Recognition* (NER) Dan *Rule-Based* untuk Visualisasi Penyakit Tropis di Indonesia (Atika, 2021)

Penelitian yang dilakukan oleh Atika pada tahun 2021 adalah mengekstraksi informasi data pada penyakit tropis di Indonesia berupa, entitas yang terkait dengan lokasi kejadian, waktu kejadian, nama penyakit tropis dan jumlah korban yang divisualisasikan menggunakan *Rules-Based*. Penelitian ini juga memanfaatkan *Named Entity Recognition* dan *Library spaCy* untuk membangun ekstraksi informasi dan memproses teks untuk pengklasifikasian.

2.1.2 *Named Entity Recognition* (NER) Bahasa Indonesia Menggunakan *Conditional Random Field* dan *Pos-Tagging* (Willyawan, 2018)

Penelitian yang dilakukan oleh Willyawan pada tahun 2018 dengan menghasilkan model *Named Entity Recognition* (NER) dengan bahasa Indonesia, yang mana penelitian ini dibatasi dengan informasi berupa pengenalan organisasi, nama orang dan nama tempat, berlandaskan pada data teks yang diinputkan. Penelitian ini menggunakan teknik *conditional random field*, dalam membuat model *Named Entity Recognition* (NER) dan *POS-Tagging* guna memaksimalkan jumlah data yang berlabel dan nilai hasil pengujian pada data model.

2.1.3 Implementasi *Named Entity Recognition* Pada *Factoid Question Answering System* Untuk Cerita Rakyat Indonesia (Kurniawati, dkk. 2018)

Penelitian ini dilakukan untuk membuat anak-anak mudah dalam memahami isi cerita, maka peneliti menyusun *question answering system* dengan memakai *Named Entity Recognition.* Pengklasifikasian pada penelitian ini yaitu, memakai metode *naive bayes*. Penelitian ini menggunakan 4 *named entity* untuk mengetahui istilah kata yang akan dijadikan kunci jawaban diantaranya yaitu *product, location, person* dan *none, none* bukan sebuah *entity*. Selain itu, pertanyaan yang diajukan dalam sistem tanya jawab ini adalah pertanyaan faktual, dengan jawaban singkat dan fakta faktual, bukan deskripsi. Data yang dipakai adalah lima cerita rakyat Indonesia, yang berasal dari internet. Hasil klasifikasi pada *named entity* menghasilkan nilai, *precision* 34,22%, *accuracy* 64,65% dan

recall 13,13% sedangkan pada *question answering accuracy* sistem didapatkan *accuracy* sebesar 16,7%.

2.1.4 *Named-Entity Recognition* pada Teks Berbahasa Indonesia menggunakan *Metode Hidden Markov* Model dan POS-Tagging (Yusliani, dkk. 2021)

Penelitian ini menggunakan metode *Hidden Markov Model* (HMM) yang merupakan cara yang digunakan untuk mengenali entitas bernama pada kata. Metode ini terbagi menjadi dua tahapan yaitu tahap *training* dan tahap *testing*. Pada tahapan *training*, dibutuhkan beberapa data yang memiliki label untuk mendapatkan model pengetahuan dengan nilai probabilitas dari masing-masing kata yang terdapat pada data latih. Nilai probabilitas berguna untuk mengenali kata yang tidak diketahui labelnya. Jika kata yang ingin dicari tidak terdapat pada data latih, maka kata itu akan memiliki nilai probabilitas nol. Nilai probabilitas nol pada kata mengakibatkan kata tersebut tidak dapat diketahui label entitasnya. Maka dari itu, penelitian ini menggunakan *part-of-speech tagging* agar tidak terdapat kata dengan nilai probabilitas nol. Pengujian dilakukan pada teks bahasa Indonesia dengan jumlah kalimat sebanyak 511 kalimat. Pengujian menghasilkan nilai *recall* 83.82%, *precision* 89.31% dan nilai *f-measure* 86.14%.

2.1.5 Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier (Wulandari dkk, 2021)

Penelitian ini akan menggunakan rule based dan NBC untuk NER dalam dokumen biologi sel. Dengan 19 dokumen latih diproses dan dianotasi manual untuk mencari Named Entity (NE) dan didapat 1135 data latih berbentuk kata. Dokumen uji ditokenisasi dan diberi POS Tag oleh tagger site terlebih dulu yang kemudian di cari bigram dan trigram. Selanjutnya proses rule based, jika dalam rule based tidak ditemukan solusi, maka akan masuk pada proses ekstraksi fitur dan NBC. Menggunakan 16 NE class, 18 aturan, dan 7 fitur dilakukan pengujian dengan tiga skenario yaitu pengujian rule based, NBC, dan kombinasi keduanya. Didapatkan rata-rata precision, recall dan f-measure tertinggi pada rule based yaitu 0,85 dengan *micro average.* Dengan *macro average recall* dan *f-measure*

tertinggi didapatkan kombinasi yaitu 0,66 dan 0,45, sedangkan *precision* tertinggi didapatkan rule based yaitu 0,39.

2.1.6 Kannada Named Entity Recognition and Classification (Nerc) Based On Multinomial Naïve Bayes (Mnb) Classifier (Amarappa, 2015)

NERC dalam bahasa Kannada adalah tugas yang penting dan menantang. Tujuan dari penelitian ini adalah untuk mengembangkan model baru untuk NERC, berdasarkan Multinomial Naïve Bayes (MNB) Classifier. Itu Metodologi yang diadopsi dalam makalah ini didasarkan pada ekstraksi fitur korpus pelatihan, dengan menggunakan istilah frekuensi, frekuensi dokumen terbalik dan menyesuaikannya dengan tf-idf-vectorizer. Makalah tersebut membahas tentang berbagai masalah dalam mengembangkan model yang diusulkan. Rincian implementasi dan kinerja evaluasi dibahas. Eksperimen dilakukan pada korpus pelatihan berukuran 95.170 token dan uji korpus 5.000 token. Diamati bahwa model bekerja dengan *Precision*, *Recall* dan *F1-measure* of 83%, 79% dan 81% masing-masing.

2.1.7 Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study: Power Failure in the Special Region of Yogyakarta) (Yanti dkk, 2021)

Model yang dibangun menggunakan data *training* yang diambil dari data keluhan masyarakat di Twitter terkait gangguan listrik di D.I. Yogyakarta. Sebab, berdasarkan laporan statistik PLN tahun 2019, Provinsi D.I. Yogyakarta merupakan provinsi yang sering mengalami gangguan listrik. Adapun penelitian ini dibuat dengan tujuan membangun model NER berbahasa Indonesia terkait gangguan listrik di Provinsi D. I. Yogyakarta dengan SpaCy, mengetahui hasil *evaluation metric* dari model yang telah dibangun, dan memetakan persebaran serta mengetahui perbandingan lokasi yang tersebutkan di *tweet* terkait gangguan listrik di Provinsi D. I. Yogyakarta pada tahun 2020. Penelitian ini menghasilkan performa hasil yang baik dengan hasil perhitungan *precision* 95.52%, *recall* 93.27%, dan *f1-score* 94.38%.

2.1.8 Serverless Named Entity Recognition untuk Teks Instruksional Pertanian Kota (Gelar dkk, 2021)

Model NER dapat diimplementasikan menjadi aplikasi Serverless, dengan menggunakan pendekatan Fungsional as Services (FaaS) dan Backend as Services (BaaS). Pada penelitian ini telah dikembangkan tiga model NER untuk data teks instruksional pertanian sub topik budidaya tanaman buah. Model berbasis Toc2Vec dengan optimasi efisiensi, Model Toc2Vec dengan optimasi akurasi dan Model berbasis IndoBERT. Model berbasis Transformer memiliki nilai f1-score terbaik sebesar 0.71 disusul Model Toc2Vec Efisiensi sebesar 0.60 dan Model Toc2Vec Efektif dan 0.57. Model Toc2Vec tidak dapat memprediksi entitas numerik dengan baik, Prediksi entitas COUNT, PERIOD dan VERIETAS selalu tertukar. Selain itu Ukuran model berbanding lurus dengan kecepatan prediksi kata per detik, dalam hal ini Model Toc2Vec optimasi efisiensi unggul, model tersebut mudah diimplementasikan menjadi *serverless* berbasis FaaS dan BaaS. Fungsionalitas dari *serverless* ML telah berhasil diuji menggunakan metode Blackbox.

2.1.9 Indonesian Named Entity Recognition Untuk Domain Zakat Menggunakan Conditional Random Fields (Widiyanti, 2022)

Pada penelitian ini penulis mendefinisikan *named entity class* untuk domain zakat sebanyak 12 *class*, yaitu rukun islam, zakat, jenis zakat, zakat mal, mustahik, muzzaki, lembaga pengelola zakat, lembaga amil zakat, syariat zakat, nisab zakat, tarif zakat dan person. Penulis menggunakan metode *Conditional Random Fields* dengan 3 skenario pembagian data latih dan data uji. Hasil yang diperoleh dari penelitian ini dengan hasil rata-rata evaluasi kinerja model Indonesian-NER yaitu *precision* sebesar 0.902, *recall* sebesar 0.834 dan *f1-score* sebesar 0.867.

2.1.10 Named Entity Recognition Untuk Data Review Tempat Wisata Dengan Metode "Bidirectional Encoder Representations from Transformers" (Irfan, 2021)

Tujuan dari penelitian ini adalah membangun model NER yang berguna untuk membantu mengidentifikasi informasi pada *review* tempat wisata, dan mengetahui parameter terbaik yang digunakan dalam membangun model, serta mengetahui

performa model yang dibangun. Dalam penelitian ini, terdapat 7723 *review* tempat wisata dari beberapa tempat wisata yang berada di Indonesia. Kumpulan *review* tersebut dipecah menjadi 207.993 token kata. Kemudian kata-kata yang telah dipecah akan diberi label entitas sesuai dengan kategori, beberapa kategori yaitu nama tempat wisata, nama lokasi, fasilitas, dan suasana. Selain dari kategori tersebut akan diberi label O yang berarti *outside*. Dalam penelitian ini menggunakan metode Bidirectional Encoder Representations from Transformer (BERT). BERT dipilih karena metode BERT ini dirancang untuk melatih representasi dua arah dari teks yang tidak berlabel dengan bersama-sama mengondisikan konteks dari kiri dan kanan di semua lapisan. Dari skenario yang dibuat dalam penelitian ini, diperoleh model dengan rata-rata F1-Score sebesar 75%.

2.1.11 Pengembangan Model Named Entity Recognition Untuk Pengenalan Entitas Pada Data Obat Indonesia (Arianto, 2023)

Penelitian ini bertujuan untuk mengekstrak entitas produk obat, kandungan obat, dan komposisi obat yang digunakan sebagai kategori untuk mengklasifikasikan entitas yang diekstraksi. Penelitian ini menggunakan data obat Indonesia sebagai objek penelitian dan metode yang digunakan adalah Convolutional Neural Network–Bidirectional Gated Recurrent Unit (CNN-BiGRU). Berdasarkan hasil pengujian, hasil akurasi terbaik didapatkan pada model dengan pengaturan *hyperparameter* yaitu CNN kernel size adalah 7, CNN filter adalah 50, CNN layer adalah 1, GRU unit adalah 200 dan GRU layer adalah 1. Dengan hasil akurasi yang didapatkan pada F1-Score sebesar 87%.

2.1.12 Named Entity Recognition pada Resep Makanan dengan Metode Bidirectional Long Short-Term Memory dan Bidirectional Encoders Representations from Transformers (Saputro, 2022)

Tujuan penelitian ini adalah membuat model NER untuk membantu mengidentifikasi entitas pada resep masakan dan mengetahui performa model yang dibangun. Terdapat 3999 resep masakan yang didapatkan dari website Kaggle. Resep yang terkumpul akan dipecah menjadi 196.696 token kata. Kemudian token- token tersebut akan dilabeli sesuai dengan kategorinya, yaitu

nama bahan makanan, bumbu masakan dan alat. Selain kategori yang sudah disebutkan, maka token akan diberi label O (*outside*). peneliti menggunakan lima langkah penelitian, yaitu mengumpulkan *dataset*, *preprocessing*, ekstraksi fitur dengan *word embedding* berupa Word2Vec, pemodelan NER, evaluasi dan deteksi entitas serta menggunakan metode Bidirectional Long Short-Term Memory (BiLSTM) dan Bidirectional Encoder Representaionsfrom Transformers (BERT) pada pemodelannya. Dari skenario-skenario model yang telah diuji, diperoleh model BERT yang memiliki rata-rata F1-Score tertinggi sebesar 94,4%, dibandingkan model BiLSTM yang memiliki rata-rata F1-Score tertinggi sebesar 77,9%.

2.1.13 Named Entity Recognition For Quranic Text Using Rule Based Approaches (Tarmizi and Saad, 2022)

Makalah ini menjelaskan pembuatan metode Pengenalan Entitas Bernama berbasis aturan untuk mengekstrak entitas yang ada dalam terjemahan bahasa Inggris terhadap makna teks Alquran dan evaluasi kinerjanya. Bernama penandaan entitas, tugas umum dalam anotasi teks, di mana entitas (kata benda) dalam teks tidak terstruktur diidentifikasi dan diberi kelas. Beberapa aturan dibangun untuk mengekstraksi beberapa jenis entitas seperti nama nabi dan orang, penciptaan, lokasi, waktu, dan berbagai nama Tuhan. Aturan dibangun terutama menggunakan ekspresi *regular* dan gazetteer. Rules yang telah dibangun menghasilkan presisi dan *recall* yang tinggi serta *F-score* yang memuaskan lebih dari 90%. Hasil dari percobaan ini dapat digunakan sebagai anotasi dalam membangun model pembelajaran mesin mengekstrak entitas dari jenis domain yang sama khususnya pada teks Alquran atau umumnya dalam domain Islam teks.

2.1.14 Hybrid Conditional Random Fields and K-Means for Named Entity Recognition on Indonesian News Documents (Santoso dkk, 2020)

Penelitian ini mengusulkan sistem NER untuk Dokumen Berita Indonesia dengan menggunakan Hybrid Conditional Bidang Acak (CRF) dan K-Means. Pendekatan hybrid adalah mencoba memasukkan penyematan kata sebagai kluster dari K-Means dan ambil sebagai fitur di CRF. Hasil pengelompokan dari K-Means menunjukkan bahwa kata yang memiliki kesamaan makna dikelompokkan *cluster*. Penyematan kata dalam penelitian ini menggunakan Word2Vec bahasa Indonesia.

Dataset terdiri dari 51.241 entitas dari Berita Online Indonesia. Peneliti melakukan beberapa percobaan dengan membagi korpus menjadi *dataset* pelatihan dan pengujian menggunakan pemisahan persentase. Penelitian ini menggunakan 4 skenario untuk percobaan, yaitu 60-40, 70-30, 80-20, dan 90-10. Performa terbaik untuk model yang dicapai dalam skenario 60-40 dengan F1-Score sekitar 87,18% dan juga meningkat sekitar 5,01% dibandingkan dengan model baseline.

2.1.15 Penggunaan Named Entity Recognition dan Artificial Intelligence Markup Language untuk Penerapan Chatbot Berbasis Teks (Christianto dkk, 2015)

Pada penelitian ini, *chatbot* dibuat untuk memenuhi kebutuhan informasi di ITHB dengan menggunakan Named Entity Recognition (NER) dan Artificial Intelligence Markup Language (AIML). NER digunakan untuk membantu mengenali pola (kata kunci) kalimat dari bahasa sehari-hari manusia (*Natural Language Processing*). AIML digunakan untuk memberikan jawaban yang relevan dan sesuai dengan pola (kata kunci) kalimat yang telah ditemukan di dalam bahasa manusia. Selain itu, pada penelitian ini juga dilakukan beberapa optimasi seperti optimasi pada proses perhitungan Naïve Bayes pada NER, proses *spelling correction*, dan proses *pattern matching* yang terbukti dapat mempercepat dan meningkatkan akurasi sistem *chatbot* dalam proses pencarian jawaban. Berdasarkan hasil pengujian, sistem *chatbot* ini dapat mengenali pola kalimat bahasa manusia dengan akurasi (NER) hingga 97% dan sistem dapat memberikan jawaban yang tepat dengan akurasi hingga 90% berdasarkan pola yang telah ditemukan tersebut.

2.1.16 Rule-based Named Entity Recognition (NER) to Determine Time Expression for Balinese Text Document (Wahyuni dan ER, 2021)

Dalam penelitian mengembangkan APM yang mampu mengidentifikasi entitas ekspresi waktu dalam dokumen teks berbahasa Bali. Entitas ekspresi waktu menjadi komponen penting dalam teks karena biasanya diikuti oleh fakta dan informasi penting. NER dibangun dengan menggunakan pendekatan berbasis aturan. Itu aturan dibangun berdasarkan pengamatan langsung terhadap dokumen

dan memperhatikan morfologi dan struktur kontekstual. Berdasarkan percobaan yang dilakukan, didapatkan hasil rata-rata dari nilai presisi, *recall*, dan *f-measure* adalah 0,85, 0,87, dan 0,85.

2.1.17 Named Entity Recognition Menggunakan Hidden Markov Model dan Algoritma Viterbi pada Teks Tanaman Obat (Setiyoaji dkk, 2017)

Penelitian ini mengimplementasikan algoritma hidden markov model dan algoritma viterbi yang dibutuhkan pada data latih. Data latih yang digunakan di beri pos-*tagging* pada setiap kata yang termasuk dalam entitas dalam penelitian dan dilakukan *preprosesing* (tokenisasi, *filtering*, dan *stemming*). Data latih yang sudah dilakukan pemberian pos-*tagging* di proses dan di jadikan model HMM. Model HMM digunakan untuk penghitungan pada algoritma viterbi. Algoritma viterbi menghitung probabilitas setiap kata pada data uji, probabilitas tertinggi akan dijadikan sebagai *output* dari sistem NER pada penelitian ini. Pada penelitian ini nilai tertinggi untuk rata-rata nilai *precision* 0.5447 yang didapat dari skenario pertama dengan data latih sebagai data uji. Nilai rata-rata *recall* tertinggi di dapat pada skenario pertama dengan nilai 0.7402. Nilai *f-measure* tertinggi pada skenario pertama dengan nilai 0.5606. Secara keseluruhan pengenalan entitas nilai yang paling rendah pada skenario ke tiga pada entitas nama dengan *f-measure* 0.4100, dan yang paling tinggi pada skenario pertama pada entitas guna dengan *f-measure* 0.7255.

2.1.18 Penerapan Named Entity Recognition Untuk Mengenali Fitur Produk Pada E-Commerce Menggunakan Rule Template Dan Hidden Markov Model (Dirgantara dkk, 2018)

Pada penelitian ini NER yang dilakukan menggunakan metode Hidden Markov Model dan Rule Template dengan 6 entitas yaitu MEREK, TIPE, HARGA, SPEK, N_SPEK dan N_TAG. Secara keseluruhan pengenalan entitas yang dilakukan pada penelitian ini menghasilkan nilai akurasi pada Rule Template sebesar 97.20% dan nilai akurasi pada Hidden Markov Model sebesar 92.23%.

2.1.19 Recognition untuk Sistem Autoful Formulir Lowongan SIM Magang MyITS StudentConnect (Hadinata, 2022)

Penelitian ini membuat sistem menggunakan NER yang bertujuan untuk meningkatkan kinerja dan efisiensi SIM Magang dalam melakukan pendistribusian informasi terkait magang yang tersedia untuk kalangan mahasiswa-mahasiswa ITS. Dilakukan pengamatan performa terhadap ketepatan analisa NER dengan menggunakan data latih berupa poster lowongan magang sebanyak 24 buah. Setelah itu, didapatkan hasil berupa optimizer Adam dengan *epochs* sebanyak 1000 yang dapat bekerja dengan performa paling baik dengan nilai *precision* 0.53023, *recall* 0.56755, dan *f1-score* 0.54565.

2.1.20 Analisis Pembentukan Modul Named Entity Recognition (NER) berbasis Algoritma Conditional Random Field (CRF) pada Sistem Repositori LIPI (Riyanto, 2017)

Penelitian ini mengimplementasikan algoritma Conditional Random Field (CRF) berbasis Stanford NER dan pendekatan berbasis aturan (rule-based) dalam pembuatan modul Named Entity Recognition (NER) untuk pengelolaan data bibliografi dari publikasi jurnal ilmiah pada sistem *repository* PDII-LIPI. NER yang digunakan adalah NER yang dapat mengenali entitas nama penulis, tahun terbit, nama jurnal, volume terbitan, nomor terbitan, dan nomor halaman dalam bahasa Indonesia dan bahasa Inggris. Hasil eksperimen terbaik untuk kinerja model yang telah dibangun menghasilkan nilai *precision* 98.10%, *recall* 98.31%, dan 98.20% pada *F-measure*.

2.2 Uraian Tinjauan Pustaka

2.2.1 Bahasa Indonesia

Bahasa yang digunakan pada media *online* dalam menunjang kemajuan bidang ilmu pengetahuan dan teknologi yaitu Bahasa Indonesia. Bahasa Indonesia digunakan sebagai sarana pengembangan kebudayaan nasional, ilmu pengetahuan dan teknologi. Bahasa yang baik merupakan bahasa yang memiliki nilai rasa yang tepat dan sesuai pada situasi penggunaannya, sedangkan bahasa Indonesia yang benar merupakan bahasa yang konsisten menerapkan kaidah-kaidah. Penggunaan bahasa Indonesia yang baik dan benar juga akan memberikan pemikiran yang baik

dan benar nyatanya, bahasa Indonesia merupakan wujud identitas bagi bangsa Indonesia dan menjadi sarana komunikasi bagi masyarakat (Khair, 2018).

2.2.2 Sistem Informasi

Sistem informasi merupakan kumpulan komponen berbasis komputer yang saling berkolaborasi untuk mengolah data, mengumpulkan dan menyimpan untuk menghasilkan data yang berguna serta bermakna bagi proses pengambilan keputusan di tingkat manajemen. (Kristiawan dan Sukadi, 2016). Data yang dikumpulkan, dikelompokkan, dan diolah sedemikian rupa, sehingga menjadi satu kesatuan informasi yang saling mendukung dan menjadi sebuah informasi yang berharga bagi penerimanya dikenal dengan sistem informasi (Heriyanto, 2018). Pengertian sistem informasi menurut Mamed Rofendy Manalu (2015) sistem informasi merupakan, suatu sistem pada sebuah organisasi yang berdasarkan kombinasi antara teknologi, media prosedur, fasilitas, orang-orang dan pengendalian yang ditujukan agar menghasilkan jalur komunikasi yang penting, memberi sinyal kepada manajemen, memproses tipe transaksi rutin tertentu dan sebagainya. Terhadap kendala internal maupun eksternal yang penting dan menyediakan suatu dasar dari informasi yang digunakan dalam pengambilan keputusan.

Berdasarkan paparan paragraf sebelumnya sistem informasi yang dimaksud dalam penelitian ini merupakan suatu sistem yang dapat dilakukan secara manual maupun berbasis komputer yang diperlukan untuk kebutuhan sehari-hari dalam penyimpanan, pengolahan data, dan mendukung kegiatan operasional sebuah organisasi untuk menghasilkan informasi valid yang digunakan dalam pengambilan keputusan.

2.2.3 *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* sering diterjemahkan ke dalam *bytecode*, format perantara, dan kemudian dieksekusi (Ratna, 2020).

2.2.4 *Framework Flask Python*

Flask merupakan *framework* web yang ditulis menggunakan *python*. Ini dapat diklasifikasikan ke dalam *microframework*, karena tidak membutuhkan alat atau pustaka tertentu dan memiliki basis data bawaan. Tidak memiliki lapisan abstraksi database. Namun ekstensi dukungan dari *flask* yang dapat menambahkan fitur aplikasi. Ekstensi ada yang digunakan pada pemetaan relasional objek, penanganan unggahan, validasi formulir, berbagai teknologi autentikasi terbuka, dan beberapa *tools framework* umum. Ekstensi lebih sering diperbarui daripada program inti *Flask* (Singh, dkk, 2019).

Flask merupakan sebuah aplikasi *microframework* yang digunakan pada Bahasa pemrograman *Python* dan dibuat menggunakan *toolkit Werkzeug* serta *template Jinja2*. *Flask* diciptakan oleh Armin Ronacher pertama kali dan kemudian dirilis pada April 2010. Kelebihan dari *framework flask* yaitu *flask* menyederhanakan inti dari frameworknya seminimal mungkin agar menjadi lebih cepat dan ringan, oleh karena itu *flask* disebut sebagai *microframework*. Maka dari itu *framework flask* lebih banyak disukai karena dapat dengan mudah dipahami dan dimengerti dibandingkan *framework* lainnya. *Framework flask* yang sederhana tetapi memiliki fungsi yang dapat ditambahkan dan dikembangkan sesuai dengan kebutuhan (Samudera, 2015). *Flask* dibagi ke dalam 2 bagian kategori yaitu, *Static File* yang mempunyai semua kode yang digunakan untuk pembuatan *website* seperti *Java Script*, *file* gambar, kode CSS dan *file* yang berisi semua template Jinja yang menyertakan halaman HTML. *Library* yang akan digunakan harus diakomodasi dalam lingkungan virtual saat mengembangkan aplikasi dengan *framework flask* (Ningtyas dan Setiyawati, 2021).

Berdasarkan paparan diatas, *flask* pada penelitian ini adalah *flask* salah satu *microframework* yang paling diminati karena mudah dipahami dan dimengerti serta memiliki fitur yang banyak sehingga dapat dikembangkan sesuai kebutuhan

pengguna *framework* ini. *Flask* memiliki berkembang menggunakan ekstensi terbaik, dan dengan API yang bagus. *Flask* memiliki semua manfaat *template* yang cepat dan banyak, fitur *Web Server Gateway Interface* yang kuat, serta kemampuan dalam pengujian yang menyeluruh pada aplikasi web dan tingkat *library* yang ekstensif.

2.2.5 HTML *(Hypertext Markup Language)*

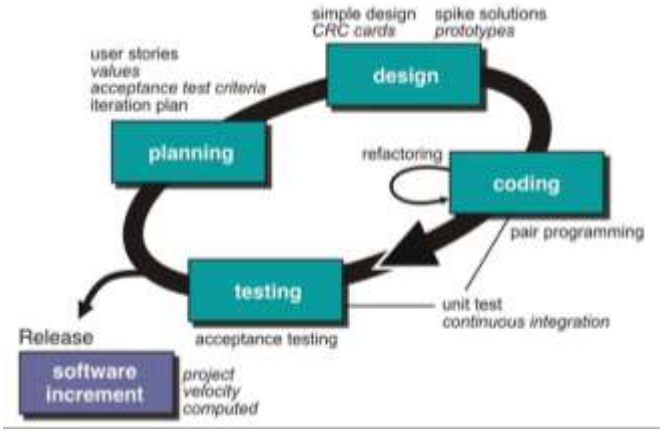
Hypertext Markup Language merupakan bahasa pemrograman yang dipakai untuk membuat sebuah situs *website* atau *homepage*. Setiap dokumen pada *website* ditulis dengan menggunakan format kode HTML. Semua format gambar, dokumen, *hyperlink* yang dapat diklik, dokumen multimedia *form* yang dapat diisi dan sebagainya didasarkan atas HTML (Fauzi, dkk, 2015).

2.2.6 *Extreme Programming*

Menurut Ramadhani dan Riyadi (2019), *Extreme Programming* biasa dikenal sebagai XP, merupakan bagian dari berbagai macam metode yang ada pada rekayasa perangkat lunak dan dalam metodologi pengembangan perangkat lunak *agile*. Pada umumnya, *extreme programming* digambarkan sebagai pendekatan yang digunakan dalam pengembangan perangkat lunak untuk meningkatkan efisiensi dan fleksibilitas integrasi pengembangan perangkat lunak, dengan pemikiran sederhana tanpa mengurangi kualitas perangkat lunak yang dibuat. *Extreme Programming* di kembangkan oleh Jeffries, Cunningham dan beck yang dikenal sebagai *lightweight discipline* pada pengembangan perangkat lunak berdasarkan 4 *core value*. Empat nilai utama yang sangat mendasar dan menjadi ciri utama dalam metodologi *extreme programming*, adalah: *simplicity, communication, feedback, dan courage.*

1. *Simplicity* (Kesederhanaan), yaitu mencoba menemukan solusi paling praktis dan sederhana.
2. *Communication* (Komunikasi), nilai ini sangat diutamakan dalam setiap tahap *extreme programming* untuk menentukan hal yang diinginkan *customer*. Komunikasi dilakukan oleh *customer* dan *developer* pada tahap *coding* hingga penyelesaian tahap akhir program, sehingga pada saat terjadi kesalahan dapat langsung diperbaiki.

3. *Feedback* (Umpan balik), *extreme programming* memungkinkan *project* mendapat masukan atau umpan balik lebih awal dan sesering mungkin baik dari *customer* atau dari *stakeholder* lainnya yang berwenang pada projek atau pun dari pihak lain.
4. *Courage* (Keberanian), tahap ini berani mencoba sesuatu yang baru, berani mengerjakan kembali dan setiap ada kekeliruan langsung diperbaiki. Dalam pengembangan sistem yang menerapkan metode *extreme programming* terdapat 4 tahapan yang dilakukan, yaitu: *Planning, Design, Coding,* dan *Testing*.



Gambar 1. *Extreme Programming Process*

(Supriyatna, 2018).

2.2.7 *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 saling berdekatan. Fokus yang dilakukan *spaCy* yaitu menyelesaikan sesuatu hal daripada menyelesaikan pendekatan akademis. *SpaCy* diikuti dengan algoritma *Parts of Speech* dan *Named Entity Recognition*. Fitur yang tersedia dalam *library spaCy* yaitu *Tokenization,*

POS-Tagging, Text Classification, dan *Named Entity Recognition* (Desikan, 2018).

Gavrilov dkk, (2020) menyatakan bahwa *spaCy* tidak melakukan yang paling akurat dalam evaluasi mereka, ia melakukan akurasi pemeliharaan tercepat yang sebanding. Model *spaCy* bersifat statistik dan setiap "keputusan" yang mereka buat adalah prediksi. Pustaka *spaCy* tidak menawarkan model pra-pelatihan untuk bahasa Indonesia, tetapi memberikan kesempatan untuk melakukan pelatihan dan mendapatkan model sendiri. Pada penelitian ini, arsitektur *library* yang digunakan pada *spaCy* adalah *nlp.pipeline.*

Berdasarkan paparan diatas *spaCy* yang dimaksud dalam penelitian ini merupakan *library* yang bersifat *open-source* yang dapat digunakan dan diakses oleh semua orang dan memiliki berbagai macam fitur di dalamnya yang dapat digunakan dalam pembuatan model. *SpaCy* biasa digunakan dalam melakukan pemrosesan dan pengolahan data untuk mengenali sebuah entitas menggunakan *Named Entity Recognition. SpaCy* dapat mengenali sebuah entitas dari kata dengan sangat cepat, sehingga dapat melakukan pelabelan pada *token* yang saling berdekatan. Model *Architecture spaCy* ditunjukkan pada Gambar 2.

Gambar 2. Model *Architecture spaCy.*

(*Library Architecture spaCy*, 2022*)*.

2.2.8 *Text Mining*

Text mining adalah proses penggalian informasi dari data berbasis teks, biasanya dari dokumen, dengan tujuan untuk mengidentifikasi istilah-istilah yang dapat menggambarkan isi dokumen secara akurat untuk melakukan analisis hubungan antar dokumen. Untuk mendapatkan informasi yang bermakna dari sumber data adalah tujuan dari *text mining*. *Text mining* menggunakan rangkaian dokumen dengan format tertentu sebagai sumber datanya (Putri dan Setiadi, 2014). *Text mining*, disebut juga sebagai Teks Data Mining (TDM) alias *knowledge discovery in text* (KDT), secara khusus dikembangkan guna pemrosesan ekstraksi informasi dari dokumen-dokumen teks tidak terstruktur *(unstructured)* (Yulian, 2018).

Text mining dapat juga disebut proses mengeksplorasi serta menganalisis data tidak terstruktur dengan jumlah besar dan dibantu oleh perangkat lunak yang dapat mengenali pola, konsep, kata kunci, topik, serta atribut lain pada data. *Text mining* bekerja dalam mengatur dan menyusun data dengan cara tertentu, sehingga dapat dilakukan analisis secara kualitatif dan kuantitatif. Ini adalah sub bidang *Data Mining* (DM), yang juga dikenal sebagai *Knowledge Discovery in Databases* (KDD). KDD digunakan untuk menemukan pengetahuan dari berbagai sumber data, termasuk data teks, *database* relasional, data web, *user* log data, dll (Hotho, dkk, 2005).

Text mining merupakan sebuah langkah dalam analisis teks yang dilakukan secara otomatis oleh komputer untuk mencari informasi yang berkualitas dari suatu rangkaian teks yang terangkum pada dokumen. Gagasan awal dalam pembuatan *text mining* yaitu, untuk menemukan pola-pola informasi yang dapat digali dari data teks yang tidak terstruktur (Hamzah, 2012). *Text mining* yaitu sebuah proses menambang data yang berupa data teks, dimana sumber data yang digunakan umumnya berasal dari dokumen teks dan tujuannya untuk menemukan kata yang dapat mewakili isi dokumen. Sehingga dapat dilakukannya analisis keterkaitan antar dokumen tersebut.

Text mining mengubah informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi tidak dalam bentuk data (Aditya, 2015).

Menurut Fauziah dkk (2019), *text mining* didefinisikan sebagai sebuah proses penggalian informasi, yang dimana seorang pengguna berinteraksi dengan serangkaian dokumen menggunakan *tool* analisis yang merupakan komponen-komponen pada data mining. Tujuan dari *text mining* adalah untuk menemukan kata-kata yang dapat mewakili isi di dalam sebuah dokumen, sehingga dapat dilakukan analisis hubungan antar dokumen. (Ghiffarie, dkk, 2019). Berdasarkan paparan diatas yang dimaksud *text mining* pada penelitian ini yaitu proses menambang data yang dilakukan untuk mendapatkan sebuah informasi dari dokumen teks menggunakan *tools* analisis yang ada. *Text mining* dapat digunakan untuk mendapatkan informasi pada data tidak terstruktur atau semi terstruktur. *Text mining* bekerja dalam mengatur dan menyusun data dengan cara tertentu, sehingga dapat dilakukan analisis secara kualitatif dan kuantitatif.

2.2.9 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) yaitu perangkat lunak yang digunakan untuk mengolah bahasa manusia. Istilah "pemrosesan bahasa alami" (NLP) mengacu pada metode yang digunakan dalam menganalisis dan menyajikan teks berdasarkan analisis linguistik, dengan tujuan memungkinkan untuk menerjemahkan suatu bahasa ke dalam berbagai aplikasi atau permainan (Wulandari, dkk, 2018). Dengan memberikan komputer berupa pengetahuan bahasa manusia, NLP bertujuan untuk membuat komputer memahami bahasa manusia.

Natural Language Processing (NLP) atau pemrosesan bahasa alami adalah subbidang kecerdasan buatan yang mengkaji tentang komunikasi manusia dan komputer menggunakan bahasa alami. (Suciadi, 2001). Dua fungsi utama *Natural Language Processing* (NLP) adalah *Natural Language Understanding* (NLU) dan *Natural Language Generation* (NLG). NLP juga bertanggung jawab atas peringkasan otomatis (*automatic summarization*), *Information Extraction* (IE), *Information Retrieval* (IR), *Named Entity Recognition* (NER), dan proses terkait lainnya selain NLG dan NLU (Amarappa and Sathyanarayana, 2015).

Natural Language Processing (NLP) atau pengolahan bahasa alami adalah, bagian dari bidang ilmu *Artificial Intelligence* (kecerdasan buatan) yang mempelajari tentang komunikasi manusia dan komputer menggunakan bahasa alami. Bahasa alami yaitu bahasa yang telah berkembang secara alami dan digunakan oleh manusia untuk berkomunikasi. Contoh bahasa alami termasuk bahasa Hindi, Inggris, Prancis, dan Jerman. Pada kajian ilmiah, bahasa berdasarkan perspektif komputasi dikenal sebagai "bahasa alami" atau "*Computational Linguistic*". *Natural Language Processing* (NLP), merupakan bidang komputer dan linguistik manusia yang berkaitan dengan interaksi antara komputer dan bahasa manusia (Chaurasia and Kumar, 2010).

Berdasarkan paparan diatas yang dimaksud *Natural Language Processing* (NLP) dalam penelitian ini adalah pemrosesan bahasa alami yang dikembangkan agar komputer dapat berkomunikasi dengan bahasa manusia sehingga menghasilkan respon pengelolahan bahasa yang sesuai. pemrosesan bahasa alami mengacu pada metode yang digunakan dalam menganalisis dan menyajikan teks berdasarkan analisis linguistik, dengan tujuan memungkinkan untuk menerjemahkan suatu bahasa ke dalam berbagai aplikasi atau permainan.

2.2.10 *Named Entity Recognition* (NER)

Named Entity Recognition (NER) yaitu jenis frase kata benda. Tujuan dari NER adalah untuk mencari dan mengidentifikasi jenis entitas bernama dalam teks. Hubungan antara *named entity* dan *question answering system* dapat ditentukan dengan menggunakan NER. Fungsi utama NER adalah mencari entitas bernama dan mengidentifikasi tipenya (Wulandari, dkk, 2018). *Named entitiy* (NE) adalah kata benda yang menjelaskan jenis individu tertentu, seperti organisasi, lokasi, nama orang dan sebagainya. *Natural Language Processing* (NLP) sekarang banyak menggunakan *named entity* secara ekstensif. (Bird, 2009).

Dalam *Natural Language Processing* (NLP), *Named Entity Recognition* (NER) adalah komponen *Information Extraction* (IE). Langkah pertama dalam ekstraksi informasi disebut pengenalan entitas, yang mengubah teks menjadi data dengan menyusunnya. Kata dan frasa dapat diidentifikasi dan dikategorikan menurut jenis

entitasnya menggunakan *named entity recognition*. *Named Entity Recognition* dapat digunakan di *semantic web, question answering* dan *machine translation* (Leonandya, dkk, 2015).

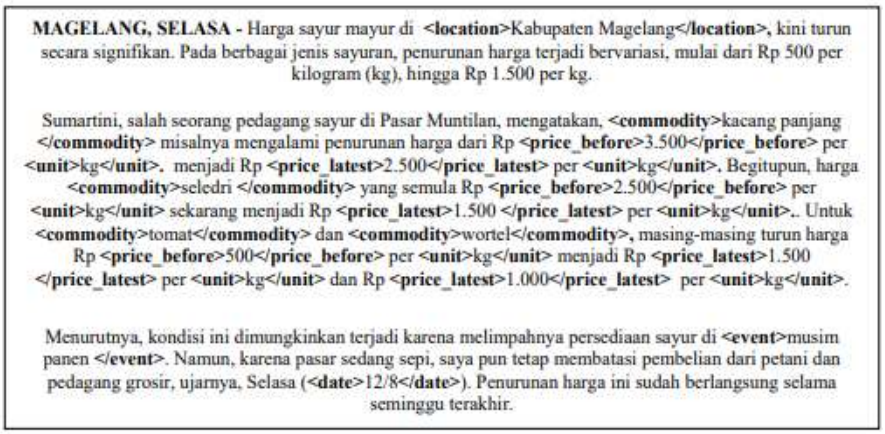
Named Entity Recognition (NER) adalah salah satu tugas penting dari sistem *Information Extractions* (IE) yang digunakan untuk mengekstrak entitas deskriptif. Ini membantu mengidentifikasi entitas generik atau domain-independen seperti lokasi, orang dan organisasi, dan entitas domain-spesifik seperti penyakit, obat-obatan, bahan kimia, protein, dll (Marrero, dkk, 2013). *Named entity recognition* (NER) juga disebut sub tugas penting dalam pemrosesan bahasa alami (NLP). Itu hasil pengakuan dan klasifikasi layak kata benda dalam dokumen teks banyak digunakan di pengambilan informasi, ekstraksi informasi, terjemahan mesin, penjawab pertanyaan dan peringkasan otomatis (Nadeau and Sekine, 2007). Tujuan utama NER adalah untuk menemukan dan mengklasifikasikan entitas bernama ke dalam kategori yang telah ditentukan seperti, lokasi, peristiwa, nama orang, waktu, organisasi, nilai moneter, jumlah dan persentase.

NER berkaitan dengan mengekstraksi entitas dunia nyata dari teks seperti orang, organisasi, atau peristiwa. *Named Entity Recognition* (NER), digunakan untuk mengekstrak informasi yang berguna dan penting dari dokumen teks mentah yang tidak terstruktur. Proses dokumen terstruktur dan tidak terstruktur merupakan bagian dari *Named Entity Recognition* (NER), yang merupakan inti dari ekstraksi informasi. Istilah "identifikasi" mengacu pada orang, lokasi, organisasi, bisnis, waktu, dan tanggal (Kurniawati, dkk, 2018).

Named Entity Recognition (NER) sangat membantu dalam proses ekstraksi informasi dan merupakan komponen dari proses *text mining*. Tujuan utama NER adalah untuk mencari dan mengatur nama teks ke dalam kelas yang telah ditentukan (Zhang, dkk, 2004). Pada penelitian ini, penulis menerapkan metode NER pada dokumen berita pertanian menggunakan bahasa Indonesia. *Named Entity Recognition* (NER) adalah jenis frase kata benda. Tujuan dari NER adalah untuk mencari dan mengidentifikasi jenis entitas bernama dalam teks.

Hubungan antara *named entity* dan *question answering system* dapat ditentukan dengan menggunakan NER. Fungsi utama NER adalah mencari entitas bernama dan mengidentifikasi tipenya. Menggunakan kamus untuk mencari jenis setiap kata dalam teks adalah cara paling dasar untuk mengidentifikasi entitas bernama. Namun, ada beberapa masalah dalam menggunakan kamus untuk mengidentifikasi entitas bernama, salah satunya adalah ambiguitas (Wulandari, dkk, 2018).

Berdasarkan paparan di atas yang dimaksud *Named Entity Recognition* (NER) dalam penelitian ini yaitu untuk untuk menemukan dan mengklasifikasikan entitas bernama pada kategori yang telah ditentukan seperti, lokasi, peristiwa, nama orang, waktu, organisasi, nilai moneter, jumlah dan persentase dan lain sebagainya.



MAGELANG, SELASA - Harga sayur mayur di **<location>**Kabupaten Magelang**</location>**, kini turun secara signifikan. Pada berbagai jenis sayuran, penurunan harga terjadi bervariasi, mulai dari Rp 500 per kilogram (kg), hingga Rp 1.500 per kg.

Sumartini, salah seorang pedagang sayur di Pasar Muntilan, mengatakan, **<commodity>**kacang panjang **</commodity>** misalnya mengalami penurunan harga dari Rp **<price_before>**3.500**</price_before>** per **<unit>**kg**</unit>**. menjadi Rp **<price_latest>**2.500**</price_latest>** per **<unit>**kg**</unit>**. Begitupun, harga **<commodity>**seledri **</commodity>** yang semula Rp **<price_before>**2.500**</price_before>** per **<unit>**kg**</unit>** sekarang menjadi Rp **<price_latest>**1.500 **</price_latest>** per **<unit>**kg**</unit>**.. Untuk **<commodity>**tomat**</commodity>** dan **<commodity>**wortel**</commodity>**, masing-masing turun harga Rp **<price_before>**500**</price_before>** per **<unit>**kg**</unit>** menjadi Rp **<price_latest>**1.500 **</price_latest>** per **<unit>**kg**</unit>** dan Rp **<price_latest>**1.000**</price_latest>** per **<unit>**kg**</unit>**.

Menurutnya, kondisi ini dimungkinkan terjadi karena melimpahnya persediaan sayur di **<event>**musim panen **</event>**. Namun, karena pasar sedang sepi, saya pun tetap membatasi pembelian dari petani dan pedagang grosir, ujarnya, Selasa (**<date>**12/8**</date>**). Penurunan harga ini sudah berlangsung selama seminggu terakhir.

Gambar 3. Anotasi Berita Web
(Kuspriyanto, dkk, 2010)

2.2.11 Ekstraksi Informasi

Proses penggalian informasi inti dari dokumen tidak terstruktur menjadi dokumen terstruktur yang mudah dikenali dikenal sebagai proses ekstraksi informasi (*Information Extraction*). *Unstructured data* adalah data yang belum terorganisasi, sedangkan *structured data* adalah data yang telah terorganisasi untuk memudahkan pencarian data (Ismaya, 2014). Ekstraksi informasi adalah bagian bidang ilmu pada pengolahan bahasa alami, yang mengubah teks tidak terstruktur menjadi informasi terstruktur. Situs web mengirimkan berbagai jenis

informasi tidak terstruktur, sehingga memerlukan pengembangan teknologi yang dapat menganalisis teks dan menemukan pengetahuan yang relevan dalam bentuk informasi terstruktur.

Informasi utama yang terkandung dalam konten halaman web adalah salah satu contoh informasi yang tidak terstruktur. Meskipun berbagai metode ekstraksi informasi manual ataupun otomatis telah dikembangkan oleh berbagai peneliti, akurasi dan kecepatan ekstraksi masih memerlukan perbaikan (Susanti dan Mustofa, 2015). Proses ekstraksi informasi mampu menggunakan teknik pendekatan berbasis aturan ataupun berbasis statistik. Teknik pendekatan berbasis aturan dibangun dan dirancang menggunakan keahlian seorang pakar, yang dimana para ahli dengan keahliannya mencocokkan himpunan teks yang digunakan. Metode pendekatan berbasis statistik dirancang dan dibangun menggunakan data *train*, dimana data latih sesuai dengan teks yang digunakan sebagai objek ekstraksi informasi (Ilyas dan Khodra, 2015). Berdasarkan uraian diatas yang dimaksud dengan ekstraksi informasi adalah proses untuk mendapatkan informasi dari sebuah dokumen ataupun berita *online* yang biasanya merupakan data atau informasi yang tidak terstruktur sehingga sulit dimengerti. Data teks yang tidak terstruktur akan diubah menjadi sebuah teks informasi yang terstruktur sehingga mudah dikenali.

2.2.12 *Multi Class Classification*

Multi Class Classification yang dimaksud pada penelitian ini yaitu pengklasifikasian data dan membentuk data menjadi beberapa kelompok berdasarkan kesamaan. Dalam kumpulan data, variabel atau fitur memiliki peran penting dalam mengklasifikasikan data. *Multi Class Classification* berarti data yang akan diklasifikasi memiliki lebih dari dua kelas dalam variabel target, seperti pada Tabel 2.

Tabel 2. Contoh *Multi Class Classification.*

Nama	Tanggal	Lokasi	Penyakit	Produk Pertanian
Vindo	29-Januari-2020	Lampung	Covid-19	Minyak kelapa
Rizki	01/03/2021	Pasar	Tikus	Lada
Yanto	Januari-2021	Desa Umbul	Wereng	Cabai

Berdasarkan tabel di atas diketahui bahwa variabel target memiliki enam kategori yaitu Nama, Tanggal, Lokasi, Penyakit, Produk_pertanian, Musim, merupakan data yang akan dikelompokkan berdasarkan teks yang diambil dari portal berita *online* pada bidang pertanian, sebagai contoh pemahaman konseptual *Multi Class Classification*.

2.2.13 *UML (Unified Modelling Language)*

Bahasa standar yang digunakan untuk menjelaskan dengan jelas *requirement*, desain, analisis, dan arsitektur pemrograman berorientasi objek adalah *Unified Modeling Language* (UML*)* (Nugroho, 2010). Sistem pengembangan perangkat lunak berorientasi objek ditentukan oleh *Unified Modeling Language* (UML), bahasa yang didasarkan pada grafik dan gambar untuk visualisasi. (Manalu, 2015). Pada kenyataannya, pemodelan digunakan untuk menyederhanakan masalah yang sulit agar lebih mudah dipelajari dan dipahami. Dari perspektif sistem perangkat lunak, UML diperlukan untuk pemodelan visual untuk menentukan, menggambar, membangun, dan mendokumentasikan. *Unified Modeling Language* (UML) yang biasa digunakan pada rancangan pengembangan sistem diantaranya *use case diagram, class diagram, activity diagram*, dan *sequence diagram*.

2.2.14 Balsamiq *Mockup*

Balsamiq adalah *tools* perancangan *mockup* yang bersifat *cloud* yang dapat jalankan pada sistem operasi linux, mac os dan windows (Krisnayani, dkk, 2016). *Balsamiq mockup* atau *wireframing* merupakan sketsa dalam pembuatan antarmuka web dalam bentuk kerangka dengan penggambaran menggunakan perangkat lunak. Dengan *tools* yang sederhana, balsamiq dapat dipahami dengan mudah sehingga pengguna dapat dengan mudah menggunakan. Kerangka kerja balsamiq dapat dibagikan dengan pengguna lain melalui email atau online (Faranello, 2012).

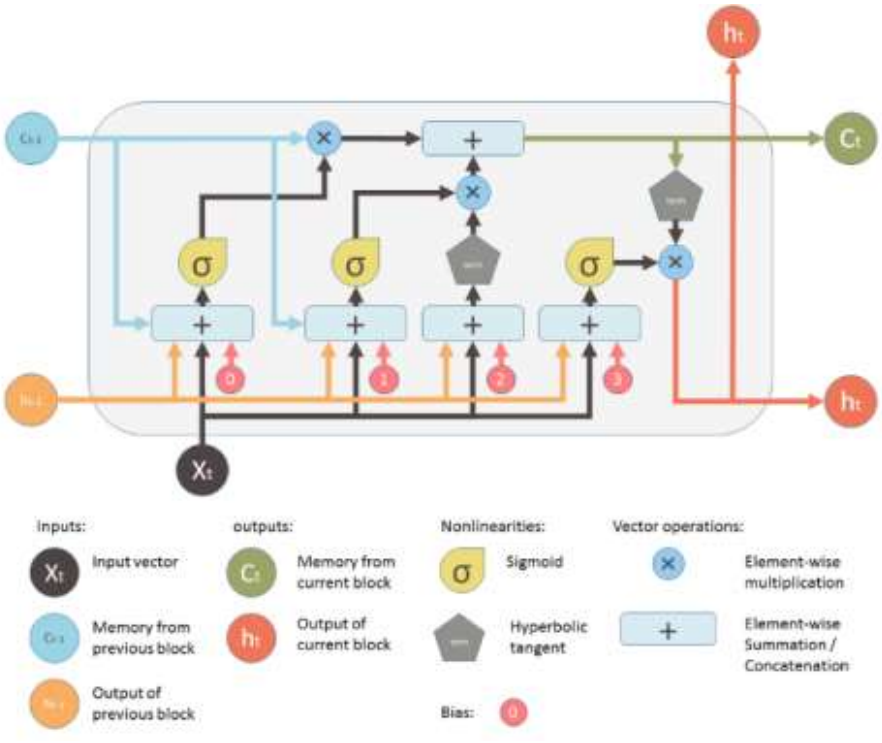
2.2.15 *Black Box Testing*

Fungsi perangkat lunak adalah fokus pengujian yang dilakukan *black-box*, yang berfokus pada kualitas perangkat lunak. Kesalahan dalam fungsi, antarmuka, struktur data, kinerja, inisialisasi, dan terminasi adalah tujuan dari pengujian *black*

box (Mustaqbal, dkk, 2015). Metode pengujian *black box* mudah digunakan, karena hanya membutuhkan batas atas dan bawah dari setiap data yang diharapkan. Jumlah bidang entri data yang akan diuji, aturan entri yang harus dipenuhi, dan kasus atas dan batas semuanya dapat digunakan untuk memperkirakan jumlah data uji (Cholifah, dkk, 2018).

2.2.16 *Long Short-Term Memory*

Salah satu perkembangan jaringan saraf tiruan yang dapat digunakan untuk memodelkan data deret waktu adalah *Long Short-Term Memory* (LSTM). LSTM dikenal juga sebagai jaringan saraf dengan arsitektur yang dapat disesuaikan, memungkinkannya dibentuk dengan cara apa pun yang diperlukan untuk aplikasi apa pun. Dalam LSTM, sel menyimpan nilai atau status (*cel state*) untuk waktu yang singkat atau lama. Ada *memory block* di LSTM yang memilih nilai keluaran yang relevan dengan masukan yang diberikan. LSTM memiliki keunggulan dalam hal ini (Wiranda dan Sadikin, 2019). Berikut Arsitektur LSTM dapat dilihat pada Gambar 4.



Gambar 4. Ilustrasi *Long Short-Term Memory* (LSTM) (Rizki, dkk, 2020).

Penjelasan *gate* yang terdapat dalam satu sel memori *Long Short-Term Memory* (LSTM) (Wiranda dan Sadikin, 2019) dapat dilihat sebagai berikut:

1. *Input Gate* (it)

Lapisan sigmoid dilalui oleh keluaran sebelumnya dan masukan baru melalui gerbang masukkan. Gerbang ini mengembalikan 0 atau dimana rumus dari it adalah:

$it = \sigma(WiSt{-}1 + WiXt)$

Dengan,

Wi = Bobot dari *Input Gate*.

$St{-}1$ = *State* sebelumnya atau *state* pada waktu t-1.

Xt = *Input* pada waktu t.

σ = Fungsi aktivasi *sigmoid*.

Nilai *gate input* dikalikan dengan *output* dari lapisan kandidat ($\tilde{C}$). Rumus dari ($\tilde{C}$) adalah:

$\tilde{C} = tanh\ (WcSt{-}1 + WcXt)$

Dengan:

C = *Intermediate cell state*.

Wc = Bobot dari *cell state*.

St-1 = *State* sebelumnya atau *state* pada waktu t-1.

Xt = *Input* pada waktu t.

2. *Forget Gate* (ft)

Forget gate merupakan *layer sigmoid* yang mendapatkan hasil pada waktu t – 1 dan masukan pada waktu-t dengan mengkombinasikannya serta mengimplementasikan fungsi aktivasi *sigmoid*, dikarenakan *sigmoid* hasil *t* dari gerbang ini yaitu 0 atau 1. Jika ft = 0, maka keadaan (*state*) sebelumnya akan dilupakan sementara, apabila ft = 1 *state* sebelumnya tidak akan berubah. Rumus dari ft yaitu:

$ft = \sigma\ (WfSt{-}1 + WfXt)$ (5)

Keterangan:

Wf = Bobot dari *forget gate*.

$St-1$ = *State* sebelumnya atau *state* pada waktu t - 1.

Xt = Masukkan pada waktu t.

σ = Fungsi aktivasi *sigmoid*.

3. *Output Gate* (ot)

Output gate atau gerbang hasil mengatur banyaknya *state* yang lewat ke dalam *output* dan bekerja menggunakan cara yang sama dengan gerbang lainnya, tahap terakhir akan menghasilkan *cell state* yang baru (ht). Rumus dari ot dan ht yaitu:

$ot = \sigma(WoSt-1 + WoXt)$

$ht = ot * tanh(ct)$

Keterangan:

Wo = Bobot dari *output gate*.

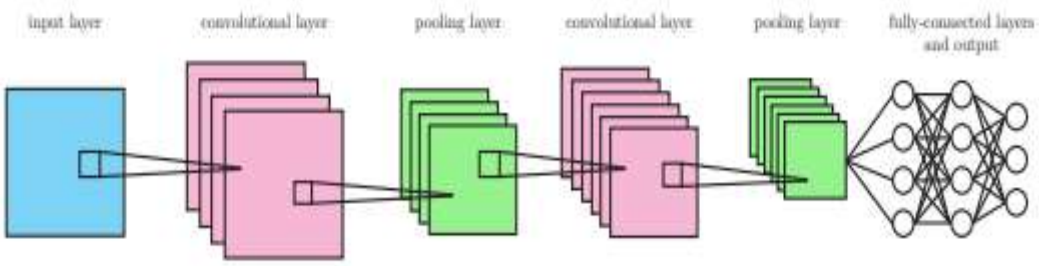
$St-1$ = *State* sebelumnya atau *state* pada waktu t - 1.

Xt = *Input* / masukkan pada waktu t.

σ = Fungsi aktivasi *sigmoid*.

2.2.17 *Convolutional Neural Network* (CNN)

Convolutional Neural Network (CNN) yaitu jaringan saraf dengan multilayer jenis *feed forward network*, yang terdiri dari dua atau lebih lapisan dalam yang dihubungkan oleh *fully connected layer*. Secara umum, CNN sangat mirip dengan jaringan saraf konvensional, yang terdiri dari neuron dengan fungsi bobot, bias, dan aktivasi. Kekuatan utama CNN terdapat pada arsitekturnya, yaitu dapat mengekstrak informasi dari objek dengan prediktif seperti gambar, teks dan klip audio (Sartini, 2020). Berikut ini merupakan arsitektur lapisan-lapisan pada *convolutional neural network* dapat dilihat pada Gambar 5.



Gambar 5. Lapisan-lapisan pada CNN (Stenroos, 2017)

Convolutional Neural Network (CNN) dibagi menjadi dua bagian yaitu, *feature extraction layer* dan *fully connected layer*. *Feature extraction layer* terdiri dari 2 komponen yaitu, *convolutional layer* dan *pooling layer* (Sartini, 2020).

1. *Convolution Layer*

Lapisan konvolusi di CNN adalah lapisan terkenal yang menjadi karakteristik utama dari algoritma ini. Dengan menggeser filter, lapisan konvolusi akan menggabungkan data masukan atau keluaran dari lapisan sebelumnya untuk menghasilkan keluaran yang dikenal sebagai *feature map* atau *activation map*. Bobot pada setiap lapisan konvolusi akan diperbarui selama proses pelatihan atau pelatihan pada data yang dilatih untuk meningkatkan hasil klasifikasi (Maulana dan Rochmawati, 2020).

2. *Pooling layer*

Setelah *convolutional layer*, *pooling layer* digunakan untuk meringkas informasi yang dihasilkan oleh *convolutional layer*. Vektor baru dibuat dengan menggabungkan (*pooled*) vektor yang dihasilkan. *Pooling* yang umum digunakan yaitu *max pooling* dan *average pooling*. Dengan *downsampling* (mengurangi ukuran *feature map*), *pooling layer* berupaya mempercepat komputasi dengan mencegah *overfitting* dan memperbarui lebih sedikit parameter (Sartini, 2020).

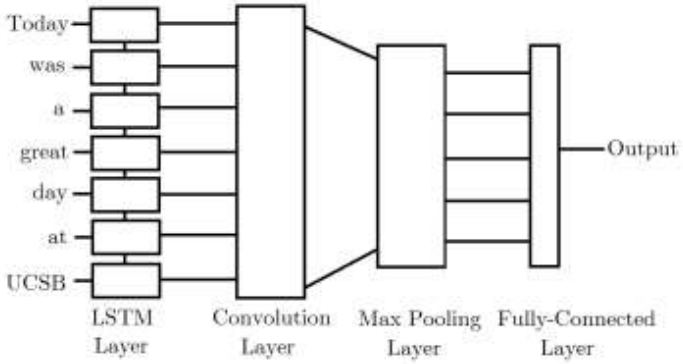
3. *Fully Connected Layer*

Feature map perlu diratakan atau dibentuk kembali menjadi vektor untuk digunakan sebagai *input* dari *fully connected layer*, karena hasil dari *convolutional layer* dan *pooling layer* masih berupa *array* multidimensi. *Multilayer perceptron* memiliki *activation function, hidden layer, output layer,* dan *loss function* yang sama dengan *fully connected layer*. (Sartini, 2020).

2.2.18 LSTM-CNN

Model LSTM-CNN, terdiri dari lapisan LSTM awal yang akan menerima penyematan kata pada setiap token sebagai masukan. Asumsinya adalah token keluaran akan menyimpan informasi dari semua token sebelumnya selain yang sekarang. Dengan kata lain, lapisan LSTM mengubah pengkodean input. Keluaran lapisan LSTM kemudian diumpamakan menjadi *convolutional layer* yang di harapkan dapat mengekstraksi fitur lokal. Keluaran lapisan konvolusi pada

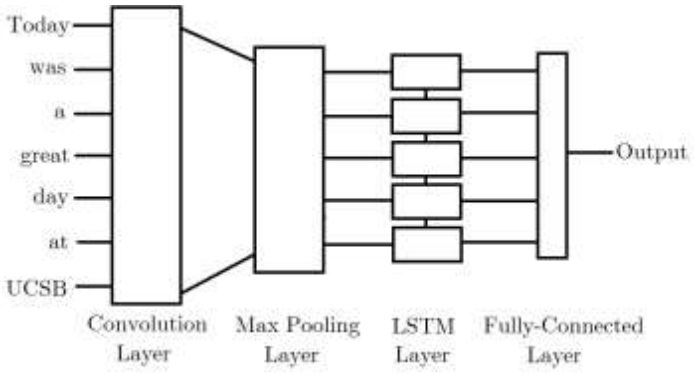
akhirnya akan digabungkan ke dimensi yang lebih kecil dan pada akhirnya menjadi keluaran sebagai label positif atau negatif (Sosa, 2017).



Gambar 6. LSTM-CNN Model (Sosa, 2017)

2.2.19 CNN-LSTM

Kombinasi model pada CNN-LSTM terdiri dari lapisan konvolusi awal yang akan menerima penyisipan kata sebagai masukan. *Output* nya kemudian akan digabungkan ke dimensi yang lebih kecil, yang kemudian dimasukkan ke dalam lapisan LSTM. Intuisi di balik model ini adalah lapisan konvolusi akan mengekstraksi fitur lokal dan lapisan LSTM kemudian akan dapat menggunakan pengurutan fitur tersebut untuk dipelajari tentang pengurutan teks masukan. Dalam praktiknya, model ini tidak sekuat model LSTM-CNN kami yang lain diusulkan (Sosa, 2017).



Gambar 7. CNN-LSTM Model.

2.2.18 *Hyperparameter*

Teknik pemodelan pada masa sekarang, banyak melibatkan sejumlah besar parameter pada saat melakukan proses *training* data. Parameter yang biasa disebut dengan *hyperparameter* ini tidak dapat diubah, berbeda dengan parameter lain

yang dapat berubah seiring berjalannya proses pelatihan data. Karena memiliki potensi dalam mempengaruhi situasi dan variasi teknik pemodelan, *hyperparameter* harus ditentukan terlebih dahulu sebelum digunakan (Rijn and Hutter, 2018). *Hyperparameter* adalah parameter yang dapat diatur dan dikontrol untuk meningkatkan kinerja prediksi metode pemodelan (Septiani dkk, 2021). *Hyperparameter* yang biasanya digunakan yaitu, *epoch, batch size*, *learning rate* dan *optimizer.*

1). *Epoch*

Epoch merupakan *hyperparameter* yang menentukan berapa kali proses akan dilakukan dalam masa *training* dalam *neural network* (Wibawa, 2016).

2). *Batch Size*

Batch size yaitu jumlah *training sample* yang digunakan dalam satu *iteration*. *Batch size* digunakan dalam proses *training* untuk menentukan jumlah contoh data *training* dan merupakan salah satu *hyperparameter* terpenting (Rochmawati, 2021).

3). *Learning Rate*

Hyperparameter yang digunakan untuk menghitung nilai koreksi bobot selama proses *training* (Rochmawati, 2021).

4). *Optimizer*

Optimizer Adam yaitu algoritma stokastik berdasarkan perkiraan adaptif dari order rendah. Adam cocok diterapkan pada permasalahan data dengan *gradient* yang menyebar (Wibawa, 2016).

III. METODE PENELITIAN

3.1 Waktu dan Tempat

Tempat dan waktu penilitian ini akan dijelaskan sebagai berikut.

3.1.1 Waktu Penelitian

Penelitian dilaksanakan di bulan Januari 2022 pada semester delapan genap hingga bulan Desember 2022 di semester Sembilan ganjil.

3.1.2 Tempat Penelitian

Penelitian dilaksanakan di Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Universitas Lampung dan Rumah.

3.2 Data dan Alat

3.2.1 Data

Pada Penelitian ini data yang digunakan merupakan data yang didapatkan dari portal berita Bahasa Indonesia bidang pertanian yang dimana data tersebut berupa data yang tidak terstruktur, kata atau frasa pada data nantinya akan dikelompokkan menjadi beberapa entitas. Data yang berupa konten halaman *website* merupakan data tidak terstruktur yang akan diubah menjadi data terstruktur dengan melakukan proses *web scraping* pada halaman portal berita *online.*

3.2.2 Alat

Perangkat keras yang digunakan untuk melakukan penelitian ini yaitu laptop dengan spesifikasi sebagai berikut.

1. *System Manufacturer: Asus.*
2. *System Model: X550JX.*
3. *Processor: Intel Core i7-4720HQ.*
4. RAM*: 8 GB.*

Perangkat lunak atau *software* yang digunakan pada penelitian ini adalah sebagai berikut.

1. Sistem Operasi *Windows 10.*
2. *Visual Studio Code.*
3. *Star*UML.
4. *Jupyter Notebook.*
5. *Anaconda3.*
6. *Browser Chrome.*
7. *Packages Library Python:*

   a. *Library BeautifulSoup*

   Library BeautifulSoup merupakan *library python* yang digunakan untuk melakukan proses web *scraping* agar dapat mendapatkan data dari sebuah *website*. *BeautifulSoup* memiliki metode-metode yang sederhana sehingga dapat memudahkan proses navigasi, pencarian, dan juga modifikasi struktur data pada *website* yang akan di *scraping*.

   b. *Library Request*

   Library Request digunakan untuk mengirim berbagai *request* HTTP. Dalam HTTP memiliki berbagai macam metode yang digunakan dalam melakukan perintah *request* seperti, membuat *request* GET, POST, PUT, DELETE, HEAD, PATCH, dan OPTION. Perintah *request* pada HTML digunakan untuk meminta data dari sebuah *website* dan mengirimkan *request* tersebut ke server, sehingga data tersebut nantinya akan dilakukan *parsing* data dan kemudian disimpan.

   c. *Library* HTML5Lib

   Library HTML5Lib merupakan *Library* yang digunakan untuk melakukan *parsing* data dan sebagai pengurai yang akan mengurai konten teks yang terdapat pada HTML yang telah di *request*. HTML5Lib menormalkan berbagai elemen dan struktur elemen yang sulit untuk dipahami menjadi format umum.

   d. *Library* tqdm

   Library / modul tqdm digunakan untuk menampilkan sebuah progres bar untuk mengetahui persentase proses yang dilakukan dalam bentuk yang sederhana.

e. *Library* JSON

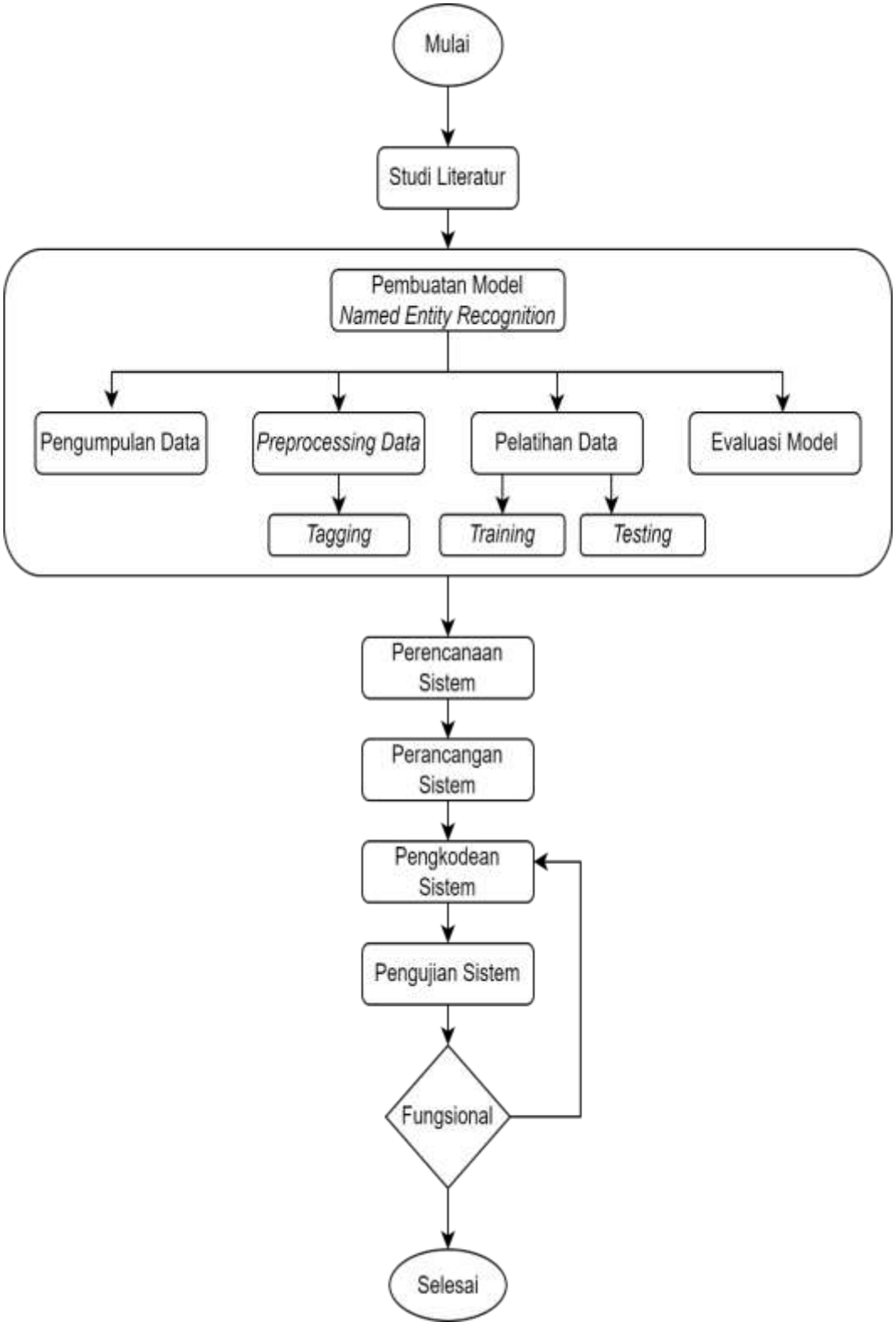
Library JSON merupakan ektensi yang berguna dalam pengelolahan data.

f. *Library spaCy*

Library spaCy digunakan untuk pemrosesan *data language* dan men-*training* NER. *Library spaCy* memiliki fungsi pada model *Text Classification* yang dapat mendeteksi kalimat berdasarkan sentimen negatif, positif dan netral.

3.3 Alur Kerja Penelitian

Penelitian ini akan melakukan beberapa tahapan seperti studi literatur, pembuatan model, perancangan sistem, dan pengujian sistem. Alur penelitian tersebut dapat dilihat pada Gambar 8.



Gambar 8. Alur Kerja Penelitian.

3.3.1 Studi Literatur

Pada tahap proses studi literatur akan dilakukan pengumpulan data yang akan digunakan dalam penelitian. Data yang diperoleh berasal dari berbagai macam sumber portal berita *online*, kemudian akan dipilih dan disesuaikan dengan berita yang dibutuhkan untuk diolah.

3.3.2 Pembuatan Model

Dalam pembuatan model pada penelitian ini diperlukan beberapa tahapan yaitu:

3.3.2.1 Pengumpulan Data

Proses pengumpulan data, dilakukan dengan menggunakan metode *web scraping* pada portal berita *online*. *Web Scraping* biasa juga disebut dengan Ekstraksi Data Web, yaitu teknik yang digunakan dalam mengekstrak sejumlah data yang besar dari sebuah situs *website* yang dimana data akan diekstraksi dan kemudian disimpan ke dalam *file* lokal komputer atau ke dalam *database* dalam bentuk table *spreadsheet*. Adapun langkah untuk mendapatkan data yaitu diperlukan metode *web scraping* pada portal berita. Teknik *web scraping* digunakan untuk mendapatkan informasi dari sebuah *website* berupa data yang besar, nantinya akan diekstrak dan disimpan ke dalam file lokal komputer.

Web scraping berfokus untuk mendapatkan data dengan cara pengambilan dan ekstraksi. Adapun proses *scrapping* yaitu:

1) Melakukan impor fungsi *get()* dari *request module.*
2) Menetapkan alamat halaman *website* menjadi variabel *url*.
3) *Request the server* yaitu melakukan permintaan HTTP isi yang terdapat pada halaman *website* dengan menggunakan fungsi *get(),* kemudian menyimpannya *respon server* kedalam sebuah *variable response*.
4) Menyimpan data konten halaman *website* yang telah diambil dan diubah menjadi data yang terstruktur dalam *data frame* berbentuk .txt.

3.3.2.2 *Preprocessing Data*

Pada tahapan *preprocessing* ini akan melakukan tokenisasi kata kemudian dilakukan pemberian fitur pada kata untuk mendapatkan vektor kata. Beberapa tahapan *preprocessing* akan dilakukan pada *dataset* yang diperoleh yaitu:

1. *Tagging*

   Tagging digunakan untuk memperoleh data *training* yang lebih baik dan meningkatkan kualitas pada *training data*. *Tagging* digunakan untuk memberikan pelabelan pada setiap kata, yang dimana setiap kata kemudian akan digunakan untuk menentukan kata-kata yang memiliki entitas nama, tanggal, hari, lokasi, penyakit, produk_pertanian, musim, organisasi, harga. Proses *tagging* dilakukan menggunakan bantuan dari *website* https://tecoholic.github.io/ner-annotator/ untuk mempermudah proses pelabelan pada data berita. Setiap token yang sesuai pada dokumen bidang pertanian dilakukan secara manual dianotasi menjadi sembilan kelas menggunakan label yang disajikan pada Tabel 3.

Tabel 3. Anotasi Pelabelan Dokumen.

Entity	**Deskripsi**
Nama_orang	Menunjukkan Nama Yang Tertera Atau Tertulis Di Dalam Halaman Web Berita.
Tanggal	Menunjukkan Tanggal Dipublikasikannya Sebuah Berita Pada Halaman Web Serta Tanggal Yang Tertera Pada Isi Berita.
Hari	Menunjukkan Hari Dipublikasikannya Sebuah Berita Pada Halaman Web Serta Hari Yang Tertera Pada Isi Berita.
Lokasi	Menunjukkan Lokasi Dipublikasikannya Sebuah Berita Pada Halaman Web Serta Hari Yang Tertera Pada Isi Berita.
Penyakit	Menunjukkan Jenis Penyakit,Wabah, Dan Hama Yang Tertera Pada Halaman Web.
Produk Pertanian	Menunjukkan hasil dari pertanian yang tertera pada halaman web.
Musim	Menunjukkan Kondisi Cuaca Ataupun Musim Yang Ada Pada Isi Berita
Organisasi	Menunjukkan Nama Sebuah Organisasi Yang Tertera Pada Isi Berita.
Harga	Menunjukkan nominal harga yang tertera pada text berita.

2. Pembagian data menjadi *training set* dan *testing set*

   Proses *splitting* yang dilakukan pada data yang telah di *preprocessing* akan dibagi ke dalam dua bagian yaitu t*raining set* 60%, *testing set* 40% dan *training set* 70%, *testing set* 30%.

3.3.2.3 Pelatihan Data

Pada tahapan ini, menjelaskan tahapan pada saat melakukan *training* pada data model *Named Entity Recognition* Bahasa Indonesia yang telah dibuat. Tahapan

training melakukan kompilasi model yang telah dibuat dengan cara menentukan *loss function* dan *optimizer* menggunakan *hyperparameter* yang akan digunakan dalam tahapan *training*. Dalam melakukan pelatihan data menggunakan dua skenario yang diterapkan pada data untuk menentukan skenario yang menghasilkan nilai lebih besar berdasarkan *hyperparameter* yang ditentukan dari *batchsize*, *epoch*, dan *learning rate*. Skenario yang dilakukan pada penilitian ini dapat dilihat pada Tabel 4.

Tabel 4. Skenario pengujian data.

Skenario	**Pembagian Data**	***hyperparameter***		
		epoch	***batchsize***	***Learning rate***
1	*Training set* 70% *Testing set* 30%	100	1000	0.001
2		50	256	0.0001
3		50	128	0.001
4	*Training set* 60% *Testing set* 40%	100	1000	0.001
5		50	256	0.0001
6		50	128	0.001

3.3.2.4 Evaluasi Model

Untuk mengukur performa dari Model *Named Entity Recognition* menggunakan fungsi yang telah disediakan oleh *spaCy* yaitu *evaluate* yang dapat menghitung skor evaluasi dari model, yang memiliki tiga kriteria umum, untuk mengetahui skor evaluasi digunakan perhitungan nilai *ents_p(precision)* yaitu menghitung *persentase* tingkat akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model, *ents_r(recall)* yaitu menghitung persentase keberhasilan dari model, *ents_f*(F-*score*) merupakan fungsi dari tingkat keseimbangan antara *precision* dan *recall*. Berikut persamaan yang digunakan untuk menghitung hasil evaluasi dari *precision, recall, F-score*.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$f1\ score = 2 \times \frac{\text{recall} \times \text{presisi}}{\text{recall} + \text{presisi}}$$

TP (*True Positif*) = Jumlah prediksi dimana pengklasifikasian dilakukan dengan benar memprediksi kelas positif sebagai positif.

FP (*False Positif*) = Jumlah prediksi dimana pengklasifikasian salah memprediksi kelas negatif sebagai positif.

FN (*False Negative*) = Jumlah Prediksi dimana pengklasifikasian salah memprediksi kelas positif sebagai negatif.

3.3.3 Perencanaan Sistem

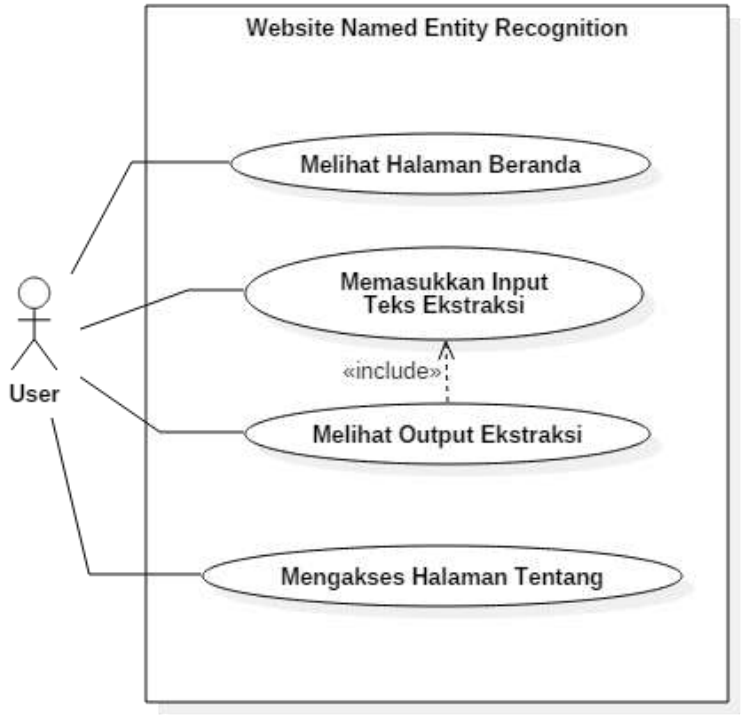
Tahapan perencanaan sistem dilakukan untuk menganalisis masalah pada sistem yang ingin dibuat dan mengetahui kebutuhan yang diperlukan oleh pengguna. Perencanaan sistem dilakukan dengan cara menganalisis penelitian terdahulu, untuk mengetahui kebutuhan apa saja yang diperlukan pada sistem.

3.3.4 Perancangan Sistem

Perancangan sistem dilakukan sesuai dengan kesepakatan yang telah disetujui terkait kebutuhan sistem. Perancangan sistem akan menggambarkan tahapan kerja sistem dan siapa saja yang akan menggunakan sistem. *Diagram Use case* digunakan untuk memberikan gambaran secara ringkas terkait pengguna sistem ini nantinya dan proses apa saja yang dapat dilakukan di dalam sistem. Berikut ini merupakan desain diagram sistem dan desain antarmuka sistem.

1. ***Use Case Diagram***

   Use Case Diagram dari system *Named Entity Recognition* dapat dilihat pada gambar berikut.



Gambar 9. *Use Case Diagram* Sistem.

Berikut merupakan penjelasan tentang *Use Case Diagram* yang dapat dilihat pada Tabel 5.

Tabel 5. Deskripsi *Use Case Diagram* Sistem

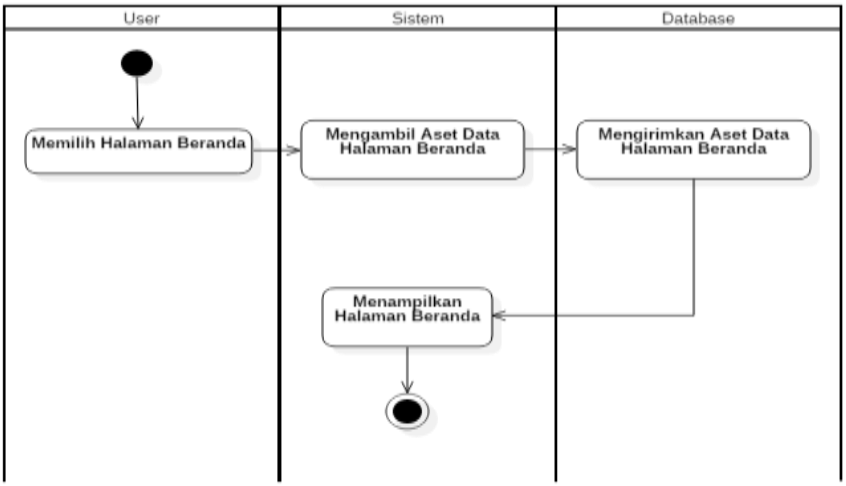
Aktor	Use Case	Deskripsi
User	Melihat Halaman Beranda	*Use Case* ini *user* dapat melihat halaman utama atau beranda pada tampilan *website* yang berisikan informasi singkat tentang *website*.
	Memasukkan *Input* Teks Ektraksi	Pada *Use Case* ini sistem menampilkan *form input* yang akan diisi oleh *user* untuk melakukan proses ekstraksi berita.
	Melihat Output Ekstraksi	Pada *Use Case* ini akan menampilkan halaman hasil dari ektraksi yang dilakukan oleh *user*.
	Melihat Halaman Tentang	Pada *Use Case* ini akan menampilkan halaman berisikan informasi singkat tentang pengembang *website*.

2. *Activity Diagram*

Activity Diagram digunakan untuk menjelaskan aktivitas proses dari setiap *Use Case*. Berikut merupakan *Activity Diagram* yang telah dibuat.

a. *Activity Diagram* Melihat Halaman Beranda

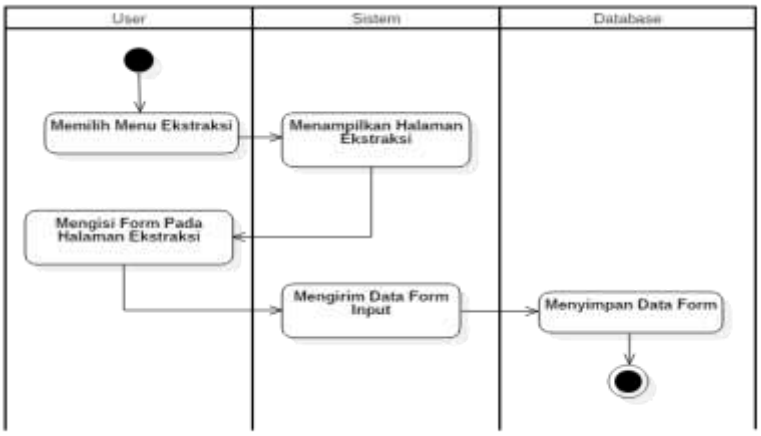
Pada *Activity Diagram* melihat halaman beranda, *user* harus mengakses menu beranda yang ada pada bar sistem, kemudian sistem akan mengambil aset data tampilan, lalu sistem menampilkan tampilan halaman beranda. Berikut merupakan *Activity Diagram* melihat halaman beranda dapat dilihat pada Gambar 10.



Gambar 10. *Activity Diagram* Melihat Halaman Beranda.

b. *Activity Diagram* Memasukkan *Input* Teks Ekstraksi

Pada *Activity* ini *user* harus memilih menu ektraksi terlebih dahulu untuk dapat mengakses halaman ekstraksi, kemudian pada halaman ekstraksi akan menampilkan form yang wajib diisi oleh *user*, agar dapat melakukan proses ektraksi *user* harus mengisi *form* yang ada pada sistem berupa teks yang ingin diekstraksi, setelah itu sistem akan menyimpan teks tersebut ke *database*. Berikut merupakan *Activity Diagram* memasukkan *input* teks ekstraksi dapat dilihat pada Gambar 11.

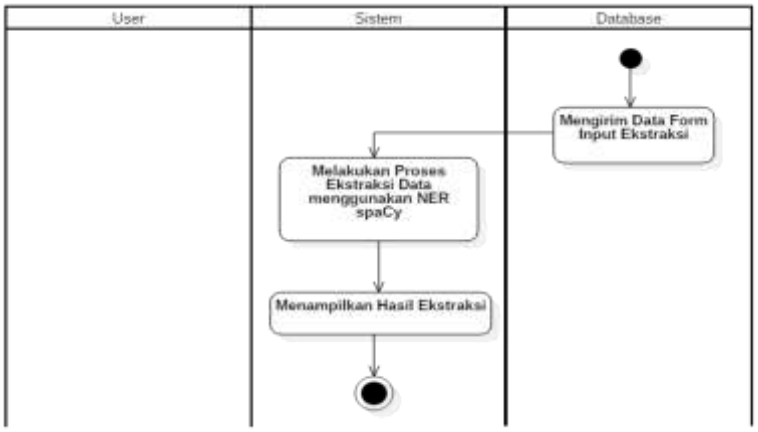


Gambar 11. *Activity Diagram* Memasukkan *Input* Teks Ekstraksi.

c. *Activity Diagram* Melihat *Output* Ekstraksi

Data teks yang sudah diisi oleh *user* akan diproses oleh sistem untuk melakukan ektraksi informasi menggunakan model NER yang sudah

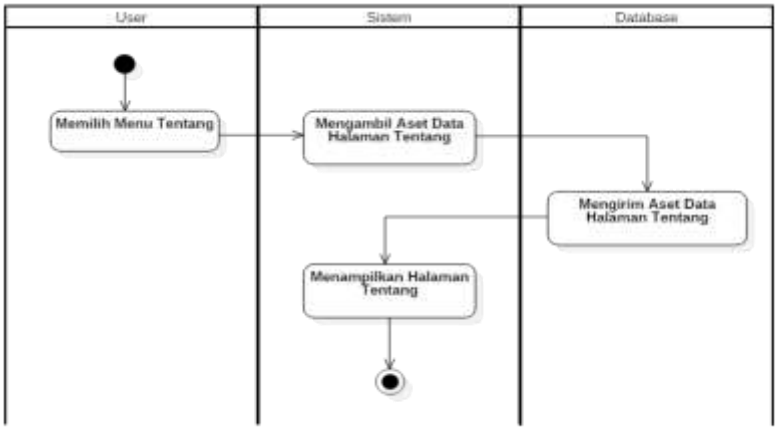
diterapkan pada sistem, pada saat proses ektraksi selesai dilakukan, sistem akan menampilkan hasil dari ekstraksi teks yang sebelumnya dimasukkan. Berikut merupakan *Activity Diagram* melihat *output* ekstraksi yang dapat dilihat pada Gambar 12.



Gambar 12. *Activity Diagram* Melihat *Output* Hasil Ekstraksi.

d. *Activity Diagram* Melihat Halaman Tentang

Dalam mengakses halaman tentang, *user* harus memilih menu tentang yang terdapat pada bar sistem. Pada halaman tentang *user* dapat melihat informasi tentang pengembang *website*. Berikut merupakan *Activity Diagram* melihat halaman tentang dapat dilihat pada Gambar 13.

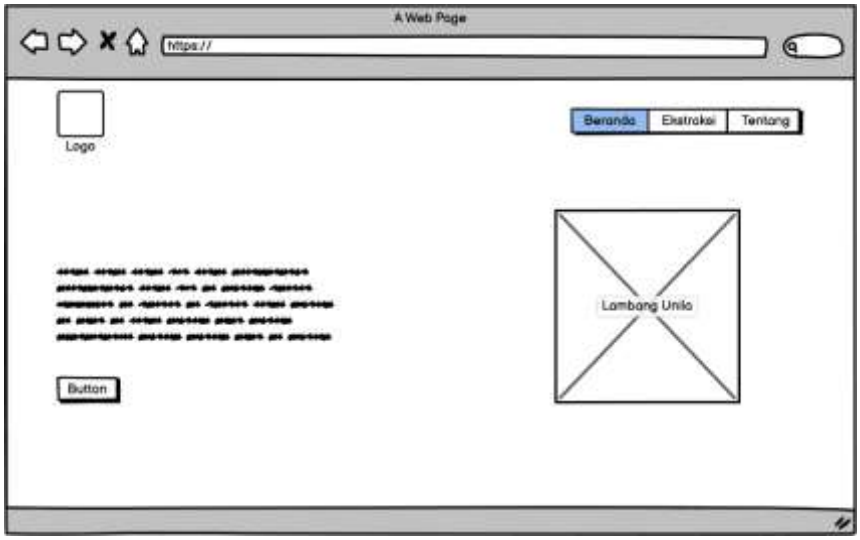


Gambar 13. *Activity Diagram* Melihat Halaman Tentang.

3. Rancangan Tampilan Sistem Menggunakan *Wireframe*

a. Rancangan Tampilan Halaman Beranda

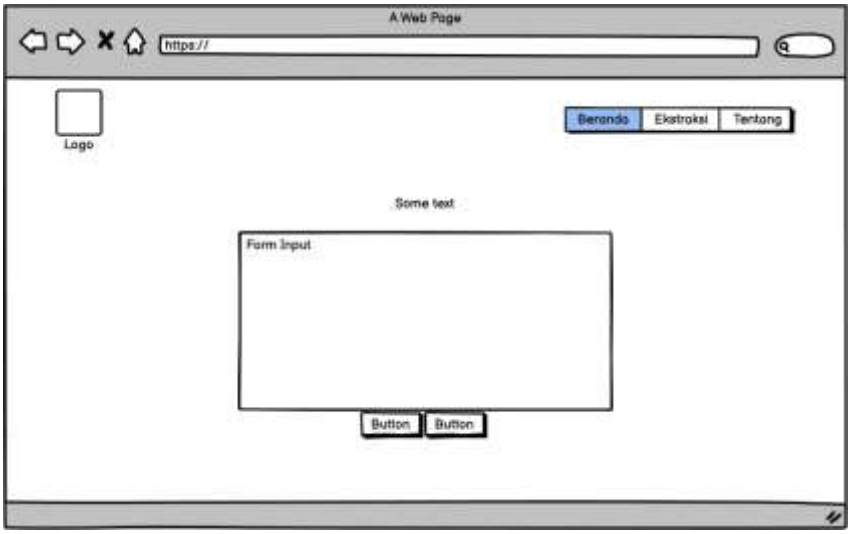
Pada rancangan tampilan antarmuka halaman beranda menampilkan logo universitas lampung dengan keterangan judul sistem. Berikut merupakan rancangan tampilan halaman beranda dapat dilihat pada Gambar 14.



Gambar 14. Rancangan Tampilan Antarmuka Beranda.

b. Rancangan Tampilan Halaman *Input* Teks Ekstraksi

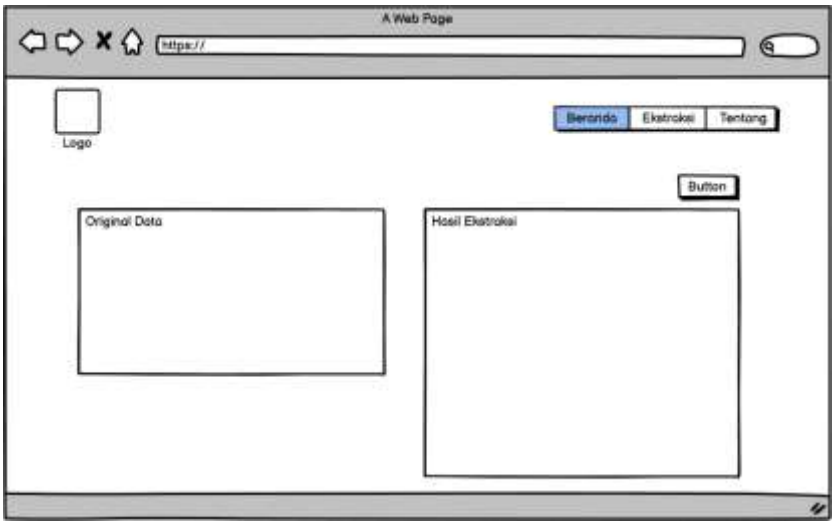
Pada halaman ini *user* harus mengisi *form* yang ada pada halaman ekstraksi berupa kalimat berita yang ingin diekstraksi. Berikut merupakan rancangan tampilan halaman *Input* Teks Ekstraksi dapat dilihat pada Gambar 15.



Gambar 15. Rancangan Tampilan Halaman *Input* Teks Ekstraksi.

c. Rancangan Tampilan Halaman *Output* Ekstraksi

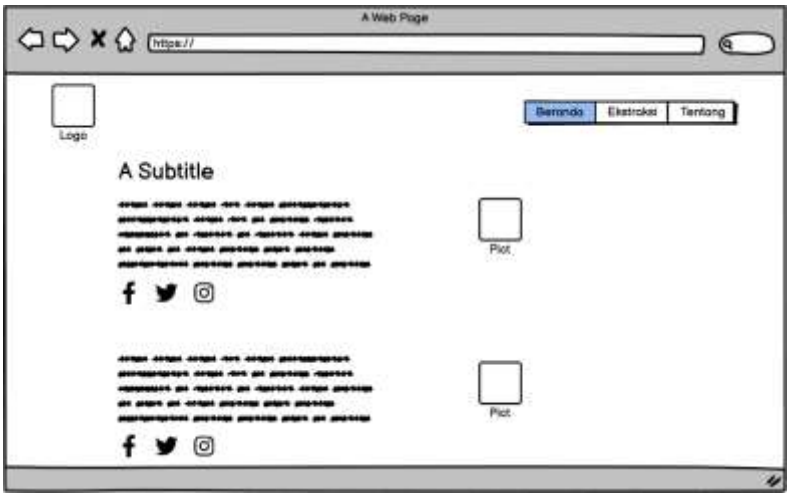
Sistem akan menampilkan hasil proses ektraksi yang telah dilakukan sebelumnya. Halaman *output* ekstraksi akan menampilkan dua *form*, yaitu *form original* dan *form preview* yang dimana *form original* berisi data teks yang sebelumnya dinput oleh *user*, *form preview* berisi data teks hasil ektraksi yang diinput sebelumnya. Berikut merupakan rancangan pada tampilan Halaman *Output* Ekstraksi yang dapat dilihat pada Gambar 16.



Gambar 16. Rancangan Tampilan Halaman *Output* Ekstraksi.

d. Rancangan Tampilan Halaman Tentang

Pada halaman tentang akan menampilkan profil singkat mengenai pengembang sistem. Berikut merupakan rancangan tampilan pada Halaman Tentang dapat dilihat pada Gambar 17.



Gambar 17. Rancangan Tampilan Halaman Tentang.

3.3.5 Pengkodean Sistem

Tahap ini merupakan eksekusi hasil dari perencanaan dan perancangan sistem. Pengkodean sistem menggunakan bahasa pemrograman python dengan menggunakan *framework flask.*

3.3.6 Pengujian Sistem

Pengujian sistem akan dilakukan dengan menggunakan metode *black box testing* yaitu dengan melakukan pengujian terhadap fungsional fitur yang ada pada sistem. Dalam melakukan pengujian sistem penguji hanya dapat menguji fitur sistem dan tidak dapat menguji kode program pada sistem, kemudian penguji memastikan fitur yang tersedia dalam sistem dapat bekerja dengan baik atau tidak.

V. SIMPULAN DAN SARAN

5.1 Simpulan

Berdasarkan hasil penelitian dan pembahasan dapat disimpulkan sebagai berikut:

1. Sistem ekstraksi informasi telah dibuat menggunakan bahasa pemrograman python dengan *framework flask* yang dapat melakukan ekstraksi informasi dari sebuah berita dengan menggunakan bahasa indonesia.
2. Sistem menerapkan model *named entity recognition*, dengan hasil evaluasi pada kinerja model menggunakan skenario yang dilakukan mendapat hasil terbaik yaitu, skenario dengan pembagian data 60% data *training* dan 40% data *testing* menggunakan *hyperparameter epoch* 100, *batch size* 1000, dan *learn rate* 0.001 yang memiliki hasil terbaik dengan nilai *precision* 100, *recall* 100, dan *f1-score* 100. Sistem yang telah dibuat diuji menggunakan metode *Black Box Testing*, dengan hasil yang didapatkan sesuai yang diharapkan berdasarkan skenario uji yang di lakukan.

5.2 Saran

Berdasarkan kesimpulan, maka peneliti menyarankan beberapa hal berikut.

1. Data yang digunakan tidak hanya mencakup berita bidang pertanian, melainkan dapat mencakup bidang yang lainnya.
2. Sistem yang dikembangkan dapat dilakukan penambahan fitur seperti, fitur *update* model, atau dapat menggunakan dua model pada sistem.

DAFTAR PUSTAKA

Aditya, B. R. 2015. Penggunaan Web Crawler Untuk Menghimpun Tweets dengan Metode Pre-Processing Text Mining. *Jurnal Infotel*, 7(2): 93-100.

Amarappa, S., and Sathyanarayana, S. V. 2015. Kannada Named Entity Recognition and Classification (Nerc) Based On Multinomial Naïve Bayes (Mnb) Classifier. *International Journal on Natural Language Computing (IJNLC)*, 4(4): 39-52.

Arianto, D. B. 2023. *Pengembangan Model Named Entity Recognition Untuk Pengenalan Entitas Pada Data Obat Indonesiai*. Skripsi. Universitas Islam Indonesia, Yogyakarta, 89.

Atika, D. 2021. *Ekstraksi Informasi Berita Online Dengan Named Entity Recognition (Ner) Dan Rule-Based Untuk Visualisasi Penyakit Tropis Di Indonesia*. Skripsi. Universitas Sumatera Utara, Medan, 74.

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media: Sebastopol.

Chaurasia, M., and Kumar, S. 2010. Natural Language Processing Based Information Retrieval For The Purpose Of Author Identification. *International Journal of Information Technology & Management Information System (IJITMIS)*, 1(1): 45-54.

Cholifah, W. N., Yulianingsih., dan Sagita, S. M. 2018. Pengujian Black Box Testing Pada Aplikasi Action & Strategy Berbasis Android Dengan Teknologi Phonegap. *Jurnal String*, 3(2): 206-210.

Christianto, D., Siswanto, E., dan Chaniago, R. 2015. Penggunaan Named Entity Recognition dan Artificial Intelligence Markup Language untuk Penerapan Chatbot Berbasis Teks. *Jurnal Tematika*, 10(2): 61-68.

Desikan, B. S. 2018. *Natural Language Processing and Computational Linguistics*. Packt: Birmingham.

Dirgantara, M. Y. S., Fauzi, M. A., dan Perdana, R. S. 2018. Penerapan Named Entity Recognition Untuk Mengenali Fitur Produk Pada E-Commerce Menggunakan Rule Template Dan Hidden Markov Model. *Jurnal*

Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(10): 3912-3920.

Faranello, S. 2012. *Balsamiq Wireframes Quickstart Guide*. Packt Publishing: Birmingham.

Fauzi., Wulandari., dan Aprilia, S. 2015. Sistem Informasi Penjualan Produk Berbasis Web Pada Chanel Distro Pringsewu. *Jurnal TAM (Technology Acceptance Model)*, 4(1): 41-47.

Fauziah, S., Sulistyowati, D. N., dan Asra. T. 2019. Optimasi Algoritma Vector Space Model Dengan Algoritma K-Nearest Neighbour Pada Pencarian Judul Artikel Jurnal. *Jurnal PILAR Nusa Mandiri*, 15(1): 21-26.

Gavrilov, D., Gusev, A., Korsakov, I., Novitsky, R., and Serova L. 2020. *Feature Extraction Method From Electronic Health Records In Russia.* Proceeding Of The 26th Conference Of Fruct Association.

Gelar, T., Nanda, A., dan Bakhrun, A. 2022. Serverless Named Entity Recognition untuk Teks Instruksional Pertanian Kota. *Jurnal Teknik Informatika dan Sistem Informasi*, 8(3): 597-606.

Ghiffarie, A., Salsabila, K. D. A., Baistama, R. P., Variadi, M. I., dan Rhajendra, M. D. 2019. Analisis Sentimen Terhadap Produk The Body Shop Tea Tree Oil. *Jurnal Teknologi & Manajemen Informatika*, 5(1): 1-8.

Hadinata, K. C., Navastara, D. A., dan Fabroyir, H. 2022. Ekstraksi Informasi pada Dokumen Teks Menggunakan Metode Named-Entity Recognition untuk Sistem Autoful Formulir Lowongan SIM Magang MyITS StudentConnect. *Jurnal Teknik ITS*, 11(1): 35-41.

Hamzah, A. 2012. *Klasifikasi Teks Dengan Naïve Bayes Classifier (Nbc) Untuk Pengelompokan Teks Berita Dan Abstract Akademis*. Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III.

Heriyanto, Y. 2018. Perancangan Sistem Informasi Rental Mobil Berbasis Web Pada PT.APM Rent Car. *Jurnal Intra-Tech*, 2(2): 64-77.

Hotho, A., Nurnberger, A., and Paab, G. 2005. A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20(1): 1-37.

Ilyas, R., dan Khodra, M. L. 2015. Ekstraksi Informasi 5W1H pada Berita Online Bahasa Indonesia. *Jurnal Cybermatika*, 3(1): 35-41.

Irfan, M. 2022. *Named Entity Recognition Untuk Data Review Tempat Wisata Dengan Metode "Bidirectional Encoder Representations from Transformers"*. Skripsi. Universitas Islam Indonesia, Yogyakarta, 71.

Ismaya, A. 2014. Algoritma Ekstraksi Informasi Berbasis Aturan. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi* (JNTETI), 3(4): 242-247

Khair, U. 2018. Pembelajaran Bahasa Indonesia dan Sastra (BASASTRA) di SD dan MI. *Jurnal Pendidikan Dasar*, 2(1): 81-98.

Krisnayani, P., Arthana, I. K. R., dan Darmawiguna, I.G. M. 2016. Analisa Usability Pada Website UNDIKSHA Dengan Menggunakan Metode Heuristic Evaluation. *Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI) ISSN 2252-9063*, 5(2).

Kristiawan, B., dan Sukadi. 2013. Pembuatan Sistem Informasi Persewaan Mobil Pada Rental Mobil Akur Pacitan. *Indonesian Jurnal on Computer Science – Speed* (IJCSS), 10(4): 15-19.

Kurniawati, Y., Indriati., dan Adikara, P. P. 2018. Implementasi Named Entity Recognition Pada Factoid Question Answering System Untuk Cerita Rakyat Indonesia. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(9): 3142-3149.

Kuspriyanto., Santoso, O. S., Widyantoro, D. H., Sastramihardja, H. S., Muludi, K., and Maimunah. S. 2010. Performance Evaluation of SVM-Based Information Extraction using τ Margin Values. *International Journal on Electrical Engineering and Informatics*, 2(4): 256-265.

Leonandya, R. A., Distiawan, B., and Praptono, N. H. 2015. *A Semi-Supervised Algorithm for Indonesian Named Entity Recognition*. Proceedings International Symposium on Computational and Business Intelligence (ISCBI).

Manalu, M. R. 2015. Implementasi Sistem Informasi Penyewaan Mobil Pada Cv. Btn Padang Bulan Dengan Metode Waterfall. *Jurnal Mantik Penusa*, 18(2): 34-43.

Marerro, M., Urbano, J., Sanchez-Cuadrado, S., Morato, J., and Gomez-Berbis, J. M. 2013. Named Entity Recognition: Fallacies, Challenges and Opportunities. *Journal of Computer Standards and Interfaces, 35(5):* 482-489.

Maulana, A. R., dan Rochmawati, N. 2020. Opinion Mining Terhadap Pemberitaan Corona di Instagram menggunakan Convolutional Neural Network. *Journal of Informatics and Computer Science*, 2(1): 53-59.

Mustaqbal, M. S., Firdaus, R. F., dan Rahmadi, H. 2015. Pengujian Aplikasi Menggunakan Black Box Testing Boundary Values Analysis. *Jurnal Ilmiah Teknologi Informasi Terapan*, 1(3): 30-38.

Nadeau, D., and Sekine, S. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1): 1-20.

Ningtyas, D. F., dan Setiyawati, N. 2021. Implementasi Flask Framework pada Pembangunan Aplikasi Purchasing Approval Request. *Jurnal Janitra Informatika dan Sistem Informasi*, 1(1): 19-34.

Nugroho, A. 2010. *Rekayasa Perangkat Lunak Berorientasi Objek dengan metode USDP*. Andi offset: Yogyakarta.

Putri, E. K., dan Setiadi, T. 2014. Penerapan Text Mining Pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes. *Jurnal Sarjana Teknik Informatika*, 2(3): 73-83.

Ramadhani, E. S., dan Riyadi, S. 2019. Pengembangan E-Budgeting Perusahaan Kelapa Sawit Dengan Metode Extreme Programming. *Jurnal Penelitian Dosen Fikom* (UNDA), 10(1): 1-7.

Ratna, S. 2020. Pengolahan Citra Digital Dan Histogram Dengan Phyton Dan Text Editor Phycharm.*Technologia,* 11(3): 181-186.

Rijn, J. N. V., and Hutter F. 2018. *Hyperparameter Importance Across Datasets*. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Riyanto, S. 2017. *Analisis Pembentukan Modul Named Entity Recognition (NER) berbasis Algoritma Conditional Random Field (CRF) pada Sistem Repositori LIPI*. Skripsi. Universitas Gunadarma, Jakarta, 43.

Rizki, M., Basuki, S., dan Azhar, Y.2020. Implementasi Deep Learning Menggunakan Arsitektur Long Short Term Memory Untuk Prediksi Curah Hujan Kota Malang. *Repositor,* 2(3): 331-338.

Rochmawati, N., Hidayati, H. B., dan Yamasari, Y. 2021. Analisa Learning rate dan Batch size Pada Klasifikasi Covid Menggunakan Deep learning dengan Optimizer Adam. *Journal Information Engineering and Educational Technology*, 5(2): 44-48.

Samudera, N. A. 2015. *Perancangan Sistem Keamanan Ruangan Menggunakan Raspberry Pi*. e-Proceeding of Engineering.

Santoso, J., Setiawan, E. I., Yuniarno, E. M., Hariadi, M., and Purnomo, M. H. 2020. Hybrid Conditional Random Fields and K-Means for Named Entity Recognition on Indonesian News Documents. *International Journal of Intelligent Engineering and Systems*, 13(3): 233-245.

Sartini. 2020. *Analisis Sentimen Twitter Bahasa Indonesia Menggunakan Algoritma Convolutional Neural Network*. Skripsi. Universitas Negeri Semarang, Semarang, 45.

Septiani, A., Jondri., dan Rizal, A. 2021. *Klasifikasi Suara Paru Normal dan Abnormal dengan Menggunakan Discrete Wavelet Transform dan Support Vector Machine*. e-Proceeding of Engineering.

Setiyoaji, A., Muflikhah, L., dan Fauzi, M. A. 2017. Named Entity Recognition Menggunakan Hidden Markov Model dan Algoritma Viterbi pada Teks Tanaman Obat. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 1(12): 1858-1864.

Singh, M., Verma, A., Parasher, A., Chauhan, N., and Budhiraja, G., 2019. Implementation of Database Using Python Flask Framework. *International Journal Of Engineering And Computer Science* (IJECS), 8(12): 24894-24899.

Sosa, P. M. 2017. Twitter Sentiment Analysis using combined LSTM-CNN Models. *ACADEMIA (Accelerating the world's research)*. Diunduh dari: https://www.academia.edu/download/55829451/sosa_sentiment_analysis.pdf. Diakses pada 8 Desember 2022 pukul 19.00 WIB.

SpaCy. 2022. SpaCy. Retrieved from SpaCy.io: SpaCy.io, diakses pada tanggal 21 Oktober 2022 pukul 11.20 wib.

Stenroos, O. 2017. *Object Detection from Images Using CNN*. Aalto University.

Suciadi, J. 2001. Studi Analisis Metode-Metode Parsing Dan Interpretasi Semantik Pada Natural Language Processing. *Jurnal Informatika*, 2(1): 13-22.

Supriyatna, A. 2018. Metode Extreme Programming Pada Pembangunan Web Aplikasi Seleksi Peserta Pelatihan Kerja. *Jurnal Teknik Informatika*, 11(1): 1-18.

Susanti, E., dan Mustofa, K. 2015. Ekstraksi Informasi Halaman Web Menggunakan Pendekatan Bootstrapping pada Ontology-Based Information Extraction. *Indonesian Journal of Computing and Cybernetics Systems (*IJCCS*), 9(2): 111-120.*

Tarmizi, S. A., and Saad, S. 2022. Named Entity Recognition For Quranic Text Using Rule Based Approaches. *Asia-Pacific Journal of Information Technology and Multimedia*, 11(2): 112-122.

Wahyunita, L. 2019. Rekayasa Web Klasifikasi pada Data Tidak Terstruktur. *Jurnal Komunikasi, Media dan Informatika*, 8(2): 88-95.

Wibawa, M. S. 2016. Pengaruh Fungsi Aktivasi, Optimisasi dan Jumlah Epoch Terhadap Performa Jaringan Saraf Tiruan. *Jurnal Sistem Dan Informatika*, 11(1): 1-8.

Willyawan, A. 2018. *Named Entity Recognition (Ner) Bahasa Indonesia Menggunakan Conditional Random Field Dan Pos-Tagging.* Skripsi. Universitas Sumatera Utara, Medan, 54.

Wiranda, L., dan Sadikin, M. 2019. Penerapan Long Short Term Memory Pada Data Time Series Untuk Memprediksi Penjualan Produk Pt. Metiska Farma. *Jurnal Nasional Pendidikan Teknik Informatika*, 8(3): 184-196.

Wulandari, D. W., Adikara, P. P., dan Adinugroho, S. 2018. Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(11): 4555-4563.

Yanti, R. M., Santoso, I., dan Suadaa, L. H. 2021. Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study: Power Failure in the Special Region of Yogyakarta). *Indonesian Journal of Information Systems* (IJIS), 4(1): 76-86.

Yulian, E. 2018. Text Mining dengan K-Means Clustering pada Tema LGBT dalam Arsip Tweet Masyarakat Kota Bandung. *Jurnal Matematika* (Mantik), 4(1): 53-58.

Yusliani, N., Sufa, M. R. P., Firdaus, A., dan Sazaki, Y. 2021. Named-Entity Recognition pada Teks Berbahasa Indonesia menggunakan Metode Hidden Markov Model dan POS-Tagging. *Jurnal Linguistik Komputasional* (JLK), 4(1): 13-20.

Zhang, J., Shen, D., Zhou, G., Su, J., and Tan, C. L. 2004. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37(6): 411–422.