SKRIPSI

PART-OF-SPEECH TAGGING DAN NAMED ENTITY RECOGNITION MENGGUNAKAN RANDOM FOREST DAN RULE-BASED CHUNKING UNTUK GEOCODING ARTIKEL ILMIAH

Disusun dan diajukan oleh:

MUH. YOGA TRIATMOJO H.W.
D121 19 1007

PROGRAM STUDI SARJANA TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2024

LEMBAR PENGESAHAN SKRIPSI

PART-OF-SPEECH TAGGING DAN NAMED ENTITY RECOGNITION MENGGUNAKAN RANDOM FOREST DAN RULE-BASED CHUNKING UNTUK GEOCODING ARTIKEL ILMIAH

Disusun dan diajukan oleh

Muh. Yoga Triatmojo H.W.
D121191007

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika
Fakultas Teknik Universitas Hasanuddin
Pada tanggal .....
dan dinyatakan telah memenuhi syarat kelulusan

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
Mukarramah Yusuf, B.Sc., M.Sc., Ph.D	Elly Warni, S.T., M.T.
NIP 19831008 201212 2 003	NIP 19820216 200812 2 001

Ketua Program Studi,

Prof. Dr. Ir. Indrabayu, S.T., M.T., M.Bus.Sys., IPM, ASEAN. Eng.
NIP 19730716 200212 1 004

ii

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini ;

Nama	:	Muh. Yoga Triatmojo H.W.
NIM	:	D121191007
Program Studi	:	Teknik Informatika
Jenjang	:	S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Part-of-Speech Tagging dan *Named Entity Recognition* Menggunakan *Random Forest* dan *Rule-Based Chunking* Untuk *Geocoding* Artikel Ilmiah

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilalihan tulisan orang lain dan bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Semua informasi yang ditulis dalam skripsi yang berasal dari penulis lain telah diberi penghargaan, yakni dengan mengutip sumber dan tahun penerbitannya. Oleh karena itu semua tulisan dalam skripsi ini sepenuhnya menjadi tanggung jawab penulis. Apabila ada pihak manapun yang merasa ada kesamaan judul dan atau hasil temuan dalam skripsi ini, maka penulis siap untuk diklarifikasi dan mempertanggungjawabkan segala resiko.

Segala data dan informasi yang diperoleh selama proses pembuatan skripsi, yang akan dipublikasi oleh Penulis di masa depan harus mendapat persetujuan dari Dosen Pembimbing.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan isi skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Gowa, 10 Oktober 2024

Yang Menyatakan

Muh. Yoga Triatmojo H.W.

ABSTRAK

MUH. YOGA TRIATMOJO H.W. *Part-of-Speech Tagging dan Named Entity Recognition Menggunakan Random Forest dan Rule-Based Chunking Untuk Geocoding Artikel Ilmiah* (dibimbing oleh Mukarramah Yusuf dan Elly Warni)

Bagian judul dan abstrak pada artikel ilmiah merupakan bagian penting bagi peneliti dalam menemukan informasi mengenai kesenjangan dan perbedaan dari sebuah penelitian terhadap penelitian lainnya. Salah satu informasi kesenjangan yang penting adalah informasi mengenai lokasi penelitian tersebut dilakukan. Dalam mencari informasi lokasi penelitian dari banyak artikel ilmiah, peneliti perlu untuk membaca banyak artikel ilmiah secara manual, selain itu terdapat banyak nama lokasi di Indonesia memiliki nama yang serupa tetapi lokasi dan titik koordinatnya berbeda dikarenakan adanya perbedaan pada wilayah geografis atau administratif dari lokasi tersebut.

Tujuan dari penelitian ini adalah untuk membuat sistem yang mampu mengekstraksi entitas lokasi penelitian dari file artikel ilmiah berbahasa Indonesia kemudian menampilkannya dalam bentuk titik penanda pada peta digital secara akurat meskipun terdapat lokasi lain yang memiliki nama yang serupa. Penelitian ini dibagi menjadi empat tahapan yaitu proses POS *tagging* menggunakan *Random Forest*, pembentukan frasa menggunakan *Rule-Based Chunkin*g, proses NER menggunakan *Random Forest*, dan proses *geocoding* menggunakan Google API.

Hasil pengujian pada setiap proses dengan menggunakan metrik evaluasi *f1-score* adalah sebesar 95,69% pada proses POS *tagging*, sebesar 90,15% pada proses pembentukan frasa, dan sebesar 91,67% pada proses NER. Penelitian ini kemudian diimplementasikan dalam bentuk WebGIS dengan hasil pengujian akurasi dalam mengekstraksi entitas lokasi penelitian dari file artikel ilmiah sebesar 90,41% yang diuji menggunakan 73 file artikel ilmiah berbahasa Indonesia dengan nama lokasi penelitian yang serupa. Sedangkan pada evaluasi keakuratan *geocoding* yang diukur menggunakan metrik *match type* mendapatkan hasil akurasi sebesar 98,48%. Sehingga sistem WebGIS ini berhasil mendapatkan akurasi sebesar 90,41% dalam mengekstraksi artikel ilmiah berbahasa Indonesia yang memiliki nama lokasi penelitian yang serupa dengan tingkat keakuratan proses *geocoding* sebesar 98,48%.

Kata Kunci: *Part-Of-Speech Tagging, Named Entity Recognition, Geocoding, Random Forest, Rule-Based Chunking*

ABSTRACT

MUH. YOGA TRIATMOJO H.W. *Part-of-Speech Tagging and Named Entity Recognition Using Random Forest and Rule-Based Chunking for Geocoding Scientific Articles (supervised by Mukarramah Yusuf and Elly Warni)*

Title and abstract sections in a scientific article are crucial for researchers to find information about the gaps and differences between one study and another. One important piece of gap information is the location where the research was conducted. To search for location information from many scientific articles, researchers must manually read numerous articles. Additionally, many place names in Indonesia are identical, but their locations and coordinates differ due to variations in geographical or administrative regions of those places.

The purpose of this research is to develop a system capable of extracting location entities from Indonesian-language scientific article files and accurately displaying them as markers on a digital map, even when there are other locations with similar names. This research is divided into four stages: POS tagging using Random Forest, phrase formation using Rule-Based Chunking, NER processing using Random Forest, and geocoding using the Google API.

The test results for each process using the f1-score evaluation metric were 95.69% for the POS tagging process, 90.15% for the phrase formation process, and 91.67% for the NER process. This research was then implemented in the form of a WebGIS, with accuracy testing results in extracting research location entities from scientific article files showing 90.41%, tested using 73 Indonesian-language scientific article files containing similar research location names. Meanwhile, for geocoding accuracy evaluation measured using the match type metric, the accuracy result was 98.48%. Therefore, this WebGIS system successfully achieved an accuracy of 90.41% in extracting Indonesian-language scientific articles with similar research location names, with geocoding process accuracy reaching 98.48%.

Keywords: Part-Of-Speech Tagging, Named Entity Recognition, Geocoding, Random Forest, Rule-Based Chunking

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR SINGKATAN DAN ARTI SIMBOL

Lambang/Singkatan	Arti dan Keterangan
$\boldsymbol{h}$	Hipotesa atau klasifikasi
$\boldsymbol{x}$	Input vektor
$\boldsymbol{\theta_k}$	*Independent* dan *identically distributed random vectors*
$\boldsymbol{C_{rf}}$	*Class* hasil klasifikasi *Random Forest*
$\boldsymbol{C_n}$	*Class* prediksi dari *tree* ke-n pada *Random Forest*
$\boldsymbol{x_i}$	Kategori fitur pada *vectorization feature*
$\boldsymbol{v_j}$	Nilai unik dari setiap fitur pada *vectorization feature*
TAPI	*True positive*
TN	*True negative*
FP	*False positive*
FN	*False Negative*
POS	*Part-of-Speech*
NER	*Named Entity Recognition*
NLP	*Natural Language Processing*
CART	*Classification and Regression Tree*
IE	*Information Extraction*
HTML	*Hypertext Markup Language*
CSS	*Cascading Style Sheets*
SIG	Sistem Informasi Geografis
WebGIS	*Web Geographic Information System*

DAFTAR LAMPIRAN

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah SWT, yang telah melimpahkan rahmat, hidayah dan karunia-Nya sehingga penulis dapat menyelesaikan penyusunan tugas akhir dengan judul " *Part-of-Speech Tagging* dan *Named Entity Recognition* Menggunakan *Random Forest* dan *Rule-Based Chunking* Untuk *Geocoding* Artikel Ilmiah". Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana (S1) pada Program Studi Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin

Penulis menyadari bahwa skripsi ini tidak mungkin terselesaikan tanpa adanya dukungan, bantuan, bimbingan dan nasehat dari berbagai pihak selama penyusunan skripsi ini. Pada kesempatan ini penulis menyampaikan terima kasih setulus-tulusnya kepada:

1. Allah SWT atas berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan tugas akhir ini.
2. Kedua orang tua penulis, Marjono Istianto, S.T., M.M. dan Supriatin Supriadi, S.km. yang selalu mendukung penulis dalam menempuh pendidikannya, selalu mendoakan penulis demi kelancaran urusan perkuliahannya, dan selalu memberi semangat penulis saat mengerjakan skripsinya. Penulis tidak akan sampai dititik yang sekarang tanpa doa dan restu kedua orang tua penulis, penulis mengucapkan banyak terima kasih kepada kedua orang tua penulis atas jasa dan kerja kerasnya untuk memfasilitasi penulis dalam menjalankan dunia perkuliahan.
3. Ibu Mukarramah Yusuf, B.Sc., M.Sc., Ph.D. selaku pembimbing I dan Ibu Elly Warni, S.T., M.T. selaku pembimbing II, yang senantiasa menyediakan waktu, tenaga, pikiran, dan perhatian yang luar biasa dalam mengarahkan penulis untuk menyelesaikan tugas akhir.
4. Segenap Dosen dan Staf Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin yang telah banyak membantu penulis selama masa perkuliahan.
5. Kak Hadhi Ichsan Saputra dan Kak Yudha Febri Aribowo selaku kakak kandung penulis dan juga keluarga penulis yang lainnya yang selalu

mendukung dan membantu penulis dalam menjalani kehidupan selama masa perkuliahan.

6. Khalik, Sabda, Juan, Abib, Sayyid, Wira, Zaki, Ali Baba, Hijir, dan teman-teman saya dari grup Albabman yang selalu menemani, menghibur, dan membantu penulis saat penulis membutuhkan bantuan.
7. Rayyan, Nurdita, Rajab, Artia, Dea, Besse, Citra, Halil, Ijlal, Debi dan teman-teman S19NIFIER lainnya yang selalu menemani penulis dalam menjalani masa perkuliahan dan menyelesaikan tugas akhir.
8. Teman-teman KKN Posko Ujung Lare 2019 yang telah memberi pengalaman yang tidak terlupakan kepada penulis selama menjalani masa KKN.
9. Serta pihak-pihak lain yang tidak tersebutkan dan tanpa sadar telah menjadi inspirasi dan membantu penulis dalam menyelesaikan tugas akhir.

Penulis berharap semoga Tuhan membalas segala kebaikan yang telah diterima oleh penulis dari berbagai pihak yang telah membantu mempermudah penulis dalam mengerjakan tugas akhir ini. Penulis menyadari bahwa tugas akhir ini masih jauh dari kata sempurna, oleh karena itu penulis mengharapkan segala bentuk saran serta masukan yang membangun dari berbagai pihak. Semoga tugas akhir ini dapat memberikan pengetahuan dan manfaat bagi penulis dan pembaca.

Gowa, 10 Oktober 2024

Penulis

Muh. Yoga Triatmojo H.W.

BAB I
PENDAHULUAN

1.1 Latar Belakang

Artikel ilmiah merupakan tulisan yang menyajikan hasil penelitian, kajian, atau pemikiran ilmiah yang disusun secara sistematis oleh peneliti dan biasanya dipublikasikan melalui jurnal ilmiah atau seminar. Berdasarkan data yang dikutip dari Dikti Kemdikbud, bahwasanya saat ini jumlah publikasi ilmiah di Indonesia meningkat pesat dalam tujuh tahun terakhir. Pada tahun 2020, Indonesia berada pada peringkat 21 dalam hal jumlah publikasi artikel ilmiah secara global yang mana terjadi peningkatan signifikan dari tahun 2016 dimana Indonesia hanya berada pada peringkat 45 secara global (Asy'ari dkk., 2022).

Bagian judul dan abstrak pada artikel ilmiah merupakan bagian yang paling penting dari artikel ilmiah dikarenakan berisikan informasi inti dan utama terkait penelitian yang telah dilakukan. Bagian judul dan abstrak yang jelas dan informatif dapat membantu peneliti lain agar mampu dengan cepat memahami tujuan, metode, dan temuan utama penelitian yang disajikan. Informasi yang didapatkan pada bagian ini dapat menjadi manfaat bagi peneliti yang ingin memulai penelitian baru dengan topik yang sama dikarenakan dapat menjadi sumber untuk menemukan kesenjangan dan kekurangan dari penelitian tersebut (Husain dkk., 2021).

Lokasi penelitian seperti lokasi studi kasus atau lokasi pengumpulan data merupakan salah satu informasi penting yang dapat ditemukan pada artikel ilmiah. Bagi peneliti lain yang akan melakukan penelitian dan berniat untuk menjadikan artikel ilmiah lainnya sebagai referensi terutama untuk penelitian yang memerlukan ruang lingkup lokasi tertentu, informasi mengenai lokasi penelitian dari sumber referensi menjadi sangat penting. Hal ini dikarenakan informasi lokasi penelitian dapat memberikan konteks spesifik tentang di mana penelitian serupa pernah dilakukan yang mana dapat memengaruhi alasan pemilihan tujuan, metode, hingga relevansi dari penelitian tersebut. Melalui informasi ini, peneliti lain dapat menjadikan artikel ilmiah tersebut sebagai bahan rujukan penelitian dengan melihat apakah tujuan dan metode yang sama dapat diterapkan pada wilayah yang berbeda atau perlu disesuaikan. Informasi lokasi penelitian juga dapat digunakan untuk

melakukan penelitian analisis komparatif antara lokasi yang berbeda dengan membandingkan pemilihan metode dan hasil yang didapatkan apakah konsisten atau dipengaruhi oleh variabel yang berbeda pada masing-masing lokasi. Setelah mendapatkan informasi rujukan yang sesuai, maka artikel ilmiah beserta lokasi penelitiannya tersebut dapat digunakan untuk membantu dalam penulisan laporan penelitian khususnya pada bab pendahuluan (Tullu, 2019).

Dalam mencari lokasi penelitian dari banyak artikel ilmiah, peneliti diharuskan untuk membaca banyak file artikel ilmiah terkait. Padahal rata-rata waktu yang diperlukan oleh orang dewasa untuk membaca satu artikel ilmiah yang memuat ±500 kata (1-2 halaman) adalah ± 2 menit. Artinya hanya untuk membaca bagian judul dan abstrak dari satu artikel ilmiah saja membutuhkan waktu ± 2 menit yang menyebabkan untuk membaca bagian judul dan abstrak pada artikel ilmiah dalam jumlah yang lebih banyak lagi pastinya akan memerlukan waktu yang lebih lama lagi. Mengacu pada hal tersebut, sistem ekstraksi lokasi penelitian secara otomatis dapat diterapkan untuk membantu menyelesaikan permasalahan ini (Brysbaert, 2019).

Ekstraksi lokasi penelitian dari artikel ilmiah secara otomatis dapat dilakukan dengan menggunakan teknik *Natural Language Processing* (NLP*)*. NLP dapat diterapkan dalam beberapa jenis pekerjaan salah satunya adalah *Information Extraction* (IE). IE memiliki *sub-task* yang dapat membantu proses pengidentifikasian dan ekstraksi informasi yang berupa entitas atau biasa disebut dengan *Named Entity Recognition* (NER). Sehingga dengan menggunakan bantuan NER, entitas di dalam sebuah teks dapat di ekstraksi yang mana salah satu dari entitas yang dapat diekstraksi adalah entitas lokasi. Salah satu cara meningkatkan performa dan akurasi dari model NER diperlukan untuk melakukan metode *Part-of-Speech tagging* (POS *tagging*) terlebih dahulu. POS *tagging* adalah proses yang berfungsi untuk mengklasifikasikan kelas kata dari setiap kata dalam suatu kalimat (Yusliani dkk., 2021).

Terdapat beberapa metode yang dapat digunakan untuk melakukan proses POS *tagging* dan NER, salah satu metode singular yang memiliki keakuratan tinggi dan kompleksitas waktu inferensi yang baik adalah dengan menggunakan *machine learning classifier*. Algoritma *Random Forest* merupakan salah satu algoritma yang

sering digunakan dalam proses klasifikasi teks karena kinerja dan akurasinya yang baik terhadap dataset yang bervariasi dengan fitur yang beragam (Parmar dkk., 2019). Terdapat beberapa penelitian yang telah menggunakan *Random Forest* pada proses POS *tagging* diantaranya adalah Newman dkk. (2022) yang menganalisis mengenai penerapan POS *tagging* pada identifikasi anotasi *source code* dengan hasil algoritma *Random Forest* memiliki nilai akurasi tertinggi dengan nilai 86% dibandingkan dengan beberapa algoritma lainnya seperti *Decision Tree, Support Vector Classifier* dan *Multinomial Naïve Bayes*. Pada proses NER juga telah terdapat beberapa penelitian yang menggunakan *Random Forest* dengan hasil evaluasi yang baik diantaranya adalah Gultiaev dan Domashova (2022) yang dalam penelitiannya membahas mengenai penerapan NER dalam mengekstraksi entitas dokumen teks berbahasa Rusia, algoritma *Random Forest* memiliki nilai *f1-score* tertinggi dengan nilai 0,916 dibandingkan dengan beberapa algoritma lainnya seperti *Naïve Bayes, Logistic Regression*, dan *Support Vector Classifi*er.

Permasalahan lainnya yang terjadi adalah apabila peneliti tidak mengetahui lokasi penelitian secara akurat atau ditemukannya lokasi penelitian yang memiliki nama yang serupa dengan lokasi lainnya. Contohnya seperti lokasi “Kota Depok” dan “Kecamatan Depok Kabupaten Sleman” yang terdapat di dua provinsi yang berbeda yaitu di Jawa Barat dan Daerah Istimewa Yogyakarta. Permasalahan seperti ini memerlukan usaha yang lebih besar bagi peneliti untuk mengetahui ketepatan lokasi yang tercantum pada artikel ilmiah tersebut. Sehingga pemetaan lokasi menggunakan peta digital diperlukan agar dapat menemukan lokasi yang tepat. Proses tersebut dapat dilakukan dengan mengimplementasikan metode *geocoding*. *Geocoding* merupakan proses konversi alamat dalam bentuk teks ke dalam koordinat geografis. Koordinat geografis adalah sistem referensi yang digunakan untuk menentukan lokasi suatu titik di permukaan bumi berdasarkan garis lintang (*latitude*) dan garis bujur (*longitude*). Agar proses *geocoding* mendapatkan hasil yang lebih akurat maka alamat yang di input juga harus dalam bentuk yang lengkap dan rinci (Sulistyo dkk., 2023).

Pada dasarnya, untuk melakukan proses *geocoding* diperlukan akses terhadap *Application Programming Interface* (API) tertentu yang berfungsi untuk menghubungkan aplikasi terhadap data titik koordinat secara global. Telah terdapat

beberapa penelitian yang membandingkan antara berbagai API dari berbagai platform dalam melakukan proses *geocoding* yang diantaranya adalah penelitian yang dilakukan oleh Monir dkk., (2021) yang membandingkan antara QGIS, Google Maps (API), dan ArcGIS dengan hasil penelitian bahwa Google Maps (API) memiliki tingkat akurasi alamat yang sangat baik yaitu sebesar 97% secara global. *Geocoding* API dari Google Maps telah digunakan pada berbagi aplikasi geospasial dikarenakan kemudahan dalam penggunaannya dan juga terdapat kredit gratis yang diberikan kepada pengguna dengan tetap mempertahankan keakuratan penandaan lokasinya (Monir dkk., 2021).

Agar mampu mendapatkan hasil *geocoding* yang akurat meskipun terdapat lokasi lain dengan nama yang serupa maka inputan pada proses *geocoding* juga harus lengkap. Namun, salah satu permasalahan yang sering dialami dalam proses NER adalah kesulitan melakukan proses *chunking* terhadap entitas dengan panjang kata yang bervariasi (Htay dkk., 2019). *Chunking* adalah proses pembentukan frasa atau *chunk* yang dibentuk dari kata-kata (token) yang ada dalam suatu teks kalimat. Metode *chunking* seperti n-gram memiliki kelemahan terhadap entitas-entitas dengan jumlah kata yang bervariasi. Metode *Rule-Based Chunking* adalah metode yang mampu digunakan untuk mengatasi masalah pembentukan *chunk* dengan jumlah kata yang bervariasi, dikarenakan proses pembentukannya didasarkan oleh aturan-aturan sintaksis dari hasil pelabelan POS *tagging*. Proses ini membuat metode *Rule-Based Chunking* menjadi lebih fleksibel dalam pembentukan *chunk* yang spesifik dengan jumlah kata yang bervariasi karena *chunk* yang dibentuk dapat disesuaikan dengan kebutuhan penelitian (Hatta Fudholi dkk., 2022).

Dalam penelitian yang dilakukan oleh Hatta Fudholi dkk. (2022) yang membahas mengenai penerapan NER dalam mengekstraksi entitas obat mendapatkan kesimpulan bahwa dengan menggunakan metode POS *tagging* dan *Rule-Based Chunking* dengan aturan berdasarkan kategori kelas kata dapat meningkatkan evaluasi NER dari nilai *f1-score* sebesar 0,870 menjadi sebesar 0,892 pada teks domain kesehatan. Selain itu, dalam penelitian yang dilakukan oleh Syahroni dan Harsono (2019) yang membahas mengenai penggunaan metode *Rule-Based Chunking* dalam pembentukan frasa Bahasa Indonesia mendapatkan nilai akurasi sebesar 93,32% pada proses pengubahan level kata menjadi frasa. Aturan

untuk pembentukan frasa Bahasa Indonesia cukup mudah untuk diterapkan dikarenakan struktur dasarnya yang konsisten (Syahroni & Harsono, 2019).

Berdasarkan uraian diatas, penelitian ini akan melakukan teknik POS *tagging* dan NER menggunakan algoritma *Random Forest* dan juga menggunakan metode *Rule-Based* pada proses pembentukan *chunk* berdasarkan hasil dari POS *tagging*. Hasil dari penelitian ini akan diimplementasikan ke dalam bentuk WebGIS (*Web Geographic Information System*) dengan tujuan untuk membantu para peneliti dalam menemukan kesenjangan dan perbedaan yang terdapat pada artikel ilmiah dalam jumlah yang banyak berdasarkan entitas lokasi penelitiannya secara tepat menggunakan peta digital.

1.2 Rumusan Masalah

1. Bagaimana mengimplementasikan POS *tagging* dan NER menggunakan *Random Forest* dan *Rule-Based Chunking* pada dataset artikel ilmiah berbahasa Indonesia?
2. Bagaimana hasil kinerja POS *tagging* dan NER menggunakan *Random Forest* dan *Rule-Based Chunking* pada dataset artikel ilmiah berbahasa Indonesia?
3. Bagaimana membuat sistem WebGIS yang mampu memvisualisasikan lokasi penelitian dari artikel ilmiah berbahasa Indonesia dan mampu membedakan lokasi tersebut jika terdapat lokasi lain dengan nama yang serupa?

1.3 Tujuan Penelitian

1. Untuk mengimplementasikan POS *tagging* dan NER menggunakan *Random Forest* dan *Rule-Based Chunking* pada dataset artikel ilmiah berbahasa Indonesia.
2. Untuk menganalisis hasil kinerja POS *tagging* dan NER menggunakan *Random Forest* dan *Rule-Based Chunking* pada dataset artikel ilmiah berbahasa Indonesia.
3. Untuk membuat sistem WebGIS yang mampu memvisualisasikan lokasi penelitian dari artikel ilmiah berbahasa Indonesia dan mampu membedakan lokasi tersebut jika terdapat lokasi lain dengan nama yang serupa.

1.4 Manfaat Penelitian

1. Bagi peneliti agar dapat memudahkan dalam mengekstraksi lokasi penelitian dari artikel ilmiah berbahasa Indonesia kemudian menampilkan hasilnya ke dalam peta digital agar mampu mendapatkan kesenjangan penelitian berdasarkan lokasi penelitiannya.
2. Bagi penulis agar menambah ilmu pengetahuan dan pengalaman sekaligus menerapkan teori yang didapat dari perkuliahan dalam perusahaan maupun dunia kerja.
3. Bagi Universitas Hasanuddin bisa menjadi referensi tambahan bagi akademisi untuk mengembangkan lebih lanjut sistem yang ada.

1.5 Ruang Lingkup

1. Penelitian ini mencakup lokasi penelitian yaitu lokasi nama desa, kelurahan, kecamatan, kota, kabupaten, provinsi, pulau, dan organisasi.
2. Artikel ilmiah yang digunakan adalah artikel ilmiah berbahasa Indonesia dan memuat informasi lokasi penelitian yang terletak di wilayah Indonesia.
3. Data yang digunakan berasal dari hasil ekstraksi teks artikel ilmiah berbahasa Indonesia pada bagian judul dan abstrak.
4. Sistem dibuat menggunakan bahasa pemrograman *Python* dan *framework flask*.
5. Penelitian ini menggunakan algoritma *Random Forest* pada proses POS *tagging* dan NER.
6. Penelitian ini menggunakan metode *Rule-Based Chunking* pada proses pembentukan frasa (*chunk*).

BAB II
TINJAUAN PUSTAKA

2.1 Artikel Ilmiah

Artikel ilmiah dapat diartikan sebagai karya tulis berisi ulasan maupun penelitian dengan mengutamakan objektivitas dari penulisnya dan disusun secara sistematis. Karya tulis tersebut bersifat ilmiah jadi tidak bisa ditulis secara sembarangan. Bukan fakta yang disembunyikan sehingga dengan mempublikasikan hasil penelitian, masyarakat akan mendapatkan informasi yang dibutuhkan. Artikel ilmiah umumnya dikumpulkan dan dipublikasikan ke dalam satu jurnal ilmiah. Struktur dalam artikel ilmiah secara umum biasanya terdiri dari judul, abstrak, pendahuluan, metode penelitian, hasil penelitian, kesimpulan, dan referensi yang digunakan. Artikel ilmiah memiliki peran yang sangat penting dalam dunia akademik dan ilmiah dikarenakan memungkinkan peneliti untuk membagikan temuan dan pengetahuan baru mereka dengan komunitas akademik dan masyarakat luas. Ini membantu memperluas basis pengetahuan di berbagai bidang studi sehingga mampu mendorong kemajuan dalam ilmu pengetahuan dan teknologi (Asy'ari dkk., 2022). Gambar 1 berikut ini merupakan contoh tampilan dari artikel ilmiah.

ISTISMAR : Jurnal Kajian, Penelitian Ekonomi dan Bisnis Islam | **Vol. 3 No.2 2021**

STRATEGI PROGRAM GERAKAN KALENG INFAQ NAHDHATUL ULAMA (KOIN NU) DI UPZISNU DESAPACARPELUK KECAMATAN MEGALUH

Pipit Widya Tutik[1], Kholis Firmansyah[2], Naily El Muna[3]

[1]Universitas KH. A. Wahab Hasbullah, Jombang [2]Universitas Islam Negeri Raden Mas Said, Surakarta [3] Universitas KH. A. Wahab Hasbullah, Jombang

tutikturi@gmail.com[1] Kholisfirmansyah@unwaha.ac.id[2], elmunanaily09@gmail.com[3]

Abstrak: *Desa Pacarpeluk merupakan salah satu desa yang berada di Kecamatan Megaluh, Kabupaten Jombang yang memiliki mayoritas penduduk bermata pencaharian dibidang pertanian, selain di bidang pertanian, Desa Pacarpeluk memiliki potensi sebagai sentra pembuatan keripik dan buah semangka saat musim tanam sebagai produk unggul. Peneliti tertarik untuk menganalisis manajemen syariah dalam program koin infaq untuk memenuhi kebutuhan pokok kaum dhuafa, yang dimana masyarakat Desa Pacarpeluk sangat antusiasme dalam melakukan kegiatan program koin infaq ini. Peneliti ini merupakan jenis penelitian kualitatif deskriptif karena peneliti ikut berpatisipasi dilapangan. Pendekatan yang digunakan adalah kualitatif, data-data yang dikumpulkan dengan cara survey secara langsung. Serta wawancara langsung dengan narasumber. Dalam program koin infaq ini sudah menerapkan fungsi manajemen syariah dengan melakuakan fungsi seperti Perencanaan ini dengan menentukan perumusan sasaran yang akan menerima bantuan dan penetapan program seperti santunan duka, jaminan pengobatan, santunan persalinan, jenguk warga sakit, peduli bencana dan pemberian sembako, pengorganisasian ini dengan membentuk strukturorganisasi yang terdiri pembina, ketua, sekertaris,bendahara dan tim fundraising, Penggerakan ini dengan menerapkan pembimbingan terhadap anggota-anggota yang mengikuti kegiatan program koin infaq, Pengawasan ini dilakukan dengan menerapkan kegiatan rapat akhir tahun.*

Kata Kunci: *Manajemen Syariah dan Program Koin Infaq*

Gambar 1 Artikel ilmiah
Sumber: (Widya Tutik dkk., 2021)

Gambar 1 merupakan contoh dari halaman pertama artikel ilmiah yang memuat bagian judul dan abstrak. Bagian judul dan abstrak pada artikel ilmiah

adalah bagian yang paling penting dan paling pertama dibaca oleh seorang peneliti dikarenakan pada bagian ini berisikan banyak informasi inti dan utama mengenai penelitian yang dilakukan dalam artikel ilmiah tersebut. Bagian judul pada artikel ilmiah adalah komponen yang sangat penting dikarenakan fungsinya sebagai gambaran singkat namun jelas mengenai isi dan fokus utama dari penelitian yang dilakukan. Judul bukan hanya sekadar rangkaian kata, tetapi merupakan pintu masuk pertama bagi pembaca ke dalam dunia penelitian yang disajikan di dalam artikel ilmiah tersebut. Sedangkan bagian abstrak pada artikel ilmiah adalah sebuah ringkasan singkat dari keseluruhan isi artikel yang mencakup tujuan penelitian, metode yang digunakan, hasil yang diperoleh, dan kesimpulan yang diambil. Abstrak bertujuan untuk memberikan gambaran umum tentang penelitian kepada pembaca, sehingga mereka dapat dengan cepat memahami inti dari studi tanpa harus membaca keseluruhan artikel ilmiah (Tullu, 2019).

2.2 *Geocoding*

Debu *Geocoding* merupakan proses konversi data tekstual yang menjelaskan lokasi geografis, seperti alamat jalan, nama tempat, atau kode pos menjadi koordinat geografis ang dapat digunakan dalam sistem informasi geografis (SIG) seperti WebGIS atau aplikasi peta digital. Koordinat geografis adalah sistem referensi yang digunakan untuk menentukan lokasi suatu titik di permukaan bumi berdasarkan garis lintang (*latitude*) dan garis bujur (*longitude*). *Geocoding* memungkinkan pengguna untuk mencari, memetakan, dan menganalisis lokasi berdasarkan informasi deskriptif. Proses *geocoding* secara umum akan meningkatkan kemudahan dalam pencarian alamat dan juga memainkan peran penting dalam berbagai aplikasi, termasuk analisis spasial, perencanaan kota, logistik, dan pelayanan (Sulistyo dkk., 2023).

Proses *geocoding* biasanya melibatkan beberapa langkah penting, yang meliputi *parsing*, pencocokan, dan interpolasi. *Parsing* adalah tahap pertama di mana alamat atau deskripsi lokasi diuraikan menjadi komponen-komponen yang lebih kecil seperti nomor rumah, nama jalan, dan nama desa atau kota secara lengkap. Setelah *parsing*, tahap pencocokan dilakukan dengan mencari komponen-komponen ini dalam basis data referensi geografis. Basis data ini dapat diakses

dengan menggunakan *key* API khusus yang yang telah menyimpan data referensi geografis secara lengkap. Pada saat ini telah terdapat berbagai platform digital umum yang menyediakan layanan *geocoding* API yang mana layanan-layanan ini sering digunakan dalam pengembangan perangkat lunak atau penelitian, beberapa contohnya adalah *Bing Maps* API, *ArcGIS online*, dan *Geocoding* API dari Google (Monir dkk., 2021)

2.3 Google Maps Platform

Google Maps Platform adalah serangkaian layanan API yang disediakan oleh Google untuk pengembang aplikasi. Platform ini memungkinkan pengembang untuk mengintegrasikan fitur-fitur dari Google Maps ke dalam aplikasi atau situs web mereka. Terdapat berbagai macam fitur yang disediakan yang mana diantaranya adalah fitur *maps embed* yaitu fitur untuk mempilkan peta digital Google Maps ke dalam situs web, fitur *geocoding* dan *reverse geocoding* yaitu fitur untuk mengubah alamat menjadi koordinat geografis atau sebaliknya, fitur *address validation* untuk melakukan validasi alamat dan berbagai fitur lainnya (McQuire, 2019).

Agar dapat mengakses berbagai fitur ini, pengembang perlu menggunakan *key* API *khusus* yang didapatkan dengan cara membuat akun Google Cloud terlebih dahulu. Sebagian besar fitur yang disediakan pada Google Maps Platform adalah fitur berbayar sesuai pemakaian, yang artinya yang artinya pengguna hanya perlu membayar untuk jumlah permintaan dan fitur yang digunakan. Meskipun berbayar layanan Google Cloud memberikan kredit gratis kepada penggunanya setiap bulannya sehingga selama tidak melewati jumlah kredit gratis yang diberikan pengguna tidak perlu membayar untuk setiap layanan yang telah digunakan (McQuire, 2019).

2.4 *Natural Language Processing* (NLP)

NLP adalah sub-bidang kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. Tujuan dari teknologi ini adalah untuk membuat mesin mampu membaca dan menalar sesuai bahasa manusia dan secara otomatis memprosesnya. Kajian NLP mencakup segmentasi tuturan *(speech*

segmentation), penandaan kelas kata (POS *tagging*), segmentasi teks *(text segmentation),* penentuan makna *(word sense disambiguation),* dan ekstraksi informasi pada teks seperti NER. Perkembangan teknologi NLP sangat dipengaruhi oleh kemajuan dalam pembelajaran mesin dan ketersediaan data teks dalam jumlah besar. Penggunaan model berbasis statistik dan pembelajaran mendalam telah membawa NLP ke tingkat baru dalam hal akurasi dan kemampuan (Rumapea, 2021).

Implementasi NLP telah meluas ke berbagai aplikasi praktis, mulai dari chatbot dan asisten virtual hingga analisis media sosial dan pengenalan suara. Dalam industri, NLP digunakan untuk meningkatkan layanan pelanggan melalui otomatisasi respons dan analisis sentimen, serta dalam pencarian informasi dan pemrosesan dokumen. Tantangan utama yang masih dihadapi dalam NLP termasuk menangani ambiguitas bahasa, variasi dialek, dan masalah bias dalam data pelatihan. Penelitian terus berlanjut untuk mengatasi tantangan ini dan meningkatkan keandalan serta etika penggunaan teknologi NLP. Dengan terus berkembangnya metode dan algoritma, NLP diharapkan akan semakin canggih dan mampu memberikan kontribusi yang signifikan dalam berbagai aspek kehidupan sehari-hari dan industri (Rumapea, 2021).

2.5 *Part-of-Speech Tagging* (POS *Tagging*)

POS *tagging* adalah proses menandai setiap kata dalam sebuah teks berdasarkan struktur atau kelas kata yang sesuai, seperti contohnya kelas kata benda (*noun*), kata kerja (*verb*), kata sifat (*adjective*), dan lain-lain. Kelas kata adalah kategori yang digunakan untuk mengelompokkan kata-kata berdasarkan fungsi, makna, atau ciri-ciri gramatikalnya dalam suatu bahasa. Sejarah POS *tagging* berawal dari analisis linguistik manual yang dilakukan oleh para ahli bahasa, namun mulai berkembang pesat pada era 1960-an dan 1970-an dengan diperkenalkannya metode statistik dan komputasi. Salah satu tonggak penting dalam perkembangan POS *tagging* adalah proyek Brown Corpus yang dimulai pada 1960-an, yang memberikan dasar data untuk pengembangan model POS *tagging* otomatis (Juwantara dkk., 2021).

POS *tagging* memiliki berbagai kegunaan dalam bidang NLP. Salah satu kegunaan utamanya adalah membantu dalam *parsing* dan pembentukan *chunk* frasa, di mana struktur kelas kata dari sebuah kalimat dianalisis untuk memahami hubungan antara kata-kata. Selain itu, POS *tagging* digunakan dalam NER untuk membantu mengidentifikasi dan mengklasifikasikan entitas dalam teks. Ini juga bermanfaat dalam aplikasi seperti *machine translation*, *text-to-speech synthesis*, dan *information retrieval*, di mana pemahaman yang lebih dalam tentang struktur kalimat dapat meningkatkan akurasi dan relevansi hasil. Dengan memberikan konteks kelas kata, POS *tagging* memungkinkan sistem NLP untuk menangani ambiguitas kata dan meningkatkan kualitas analisis (Nguyen dkk., 2019).

Seiring perkembangan teknologi dan metode komputasi, POS *tagging* telah mengalami transformasi signifikan. Awalnya, pendekatan berbasis aturan (*Rule-Based*) dan model probabilistik menjadi metode utama. Namun, dengan kemajuan dalam teknologi NLP membuat teknik POS *tagging* juga terus berkembang. Penggunaan teknologi *Artificial Intelligence* (AI) seperti *machine learning* dan *deep learning* menjadi semakin sering digunakan dalam teknik POS *tagging* dikarenakan memiliki konsistensi nilai keakuratan pengenalan kelas kata dalam kalimat dan efisiensi waktu pembuatan yang lebih baik dari sebelumnya (Newman dkk., 2022).

2.6 *Rule-Based Chunking*

Rule-Based Chunking adalah teknik dalam NLP yang menggunakan aturan-aturan eksplisit yang ditentukan sebelumnya untuk mengidentifikasi dan menandai unit-unit bermakna dalam teks, seperti frasa nominal atau frasa verbal. Aturan-aturan ini biasanya didasarkan pada pola-pola linguistik dan kelas kata yang umum ditemukan dalam bahasa yang diproses atau biasanya disebut sebagai "*grammar*". Pengertian *chunking* sendiri adalah proses penggabungan kata-kata menjadi unit-unit yang bermakna yang mana bentuk dari unit ini adalah gabungan dari kata yang dapat juga disebut sebagai frasa. Salah satu tonggak penting dalam perkembangan *chunking* berbasis aturan adalah pengenalan konsep *shallow parsing*, yang berfokus pada identifikasi struktur permukaan dari kalimat tanpa perlu melakukan analisis sintaksis penuh. Pada tahun 1990-an, muncul berbagai sistem *chunking* berbasis

aturan yang sukses diterapkan dalam berbagai aplikasi NLP, seperti sistem penandaan frasa kata benda yang dikembangkan oleh Steven Abney, yang menjadi landasan untuk banyak penelitian dan aplikasi selanjutnya (Hatta Fudholi dkk., 2022).

Proses *Rule-Based Chunking* dimulai dengan melakukan POS *tagging* pada teks untuk menentukan kelas kata setiap kata dalam kalimat. Setelah POS tagging, aturan-aturan *chunking* diterapkan untuk mengelompokkan kata-kata menjadi *chunks*. Contohnya, sebuah aturan dapat menyatakan bahwa setiap rangkaian kata yang terdiri dari kata sifat diikuti oleh kata benda harus ditandai sebagai frasa kata benda. Aturan-aturan ini sering ditulis dalam bentuk pola reguler atau ekspresi reguler yang mencerminkan struktur gramatikal tertentu. Salah satu keunggulan utama dari *Rule-Based Chunking* adalah kemampuannya yang fleksibel dalam pembentukan chunk yang spesifik dengan panjang kata yang bervariasi dikarenakan cara kerjanya yang berdasarkan aturan sintaksis sehingga hasilnya dapat disesuaikan dengan kebutuhan peneliti (Htay dkk., 2019).

2.7 *Named Entity Recognition* (NER)

NER adalah sebuah teknik yang secara otomatis dapat mengidentifikasi dan mengekstraksi kata pada teks atau korpus menurut kategori atau kelas yang telah ditentukan, seperti berdasarkan kategori lokasi, nama, organisasi, dan sebagainya. NER memungkinkan untuk melakukan pengekstraksian informasi penting dari sekumpulan data teks yang tidak terstruktur. Teknik NER dapat digunakan dalam semua karya tulis digital yang memiliki teks di dalamnya seperti artikel ilmiah, dokumen, berita, sosial media, *e-book*, dll. Penerapan NER dapat mencakup berbagai bidang, seperti pada mesin penerjemah, pengenalan suara, *chat bot*, dan mesin pencari (Hatta Fudholi dkk., 2022).

Konsep NER pertama kali muncul pada akhir 1980-an dan awal 1990-an, didorong oleh kebutuhan untuk mengekstraksi informasi dari teks secara otomatis. Salah satu penggerak utama dalam perkembangan NER adalah *Message Understanding Conference* (MUC) yang diselenggarakan oleh *Defense Advanced Research Projects Agency* (DARPA). Pada saat itu, tujuan awal MUC mengembangkan NER adalah untuk mendorong kemajuan dalam otomatisasi

ekstraksi informasi dari teks yang tidak terstruktur pada kategori entitas nama orang, organisasi, lokasi, tanggal, dan waktu. Seiring berkembangnya waktu NER tidak hanya digunakan pada kelima entitas tersebut tetapi dapat digunakan pada entitas lainnya seperti nama obat dan penyakit pada domain biologi, istilah hukum pada domain hukum, nama produk dan transaksi pada domain finansial, dan sebagainya (Widiyanti dkk., 2023).

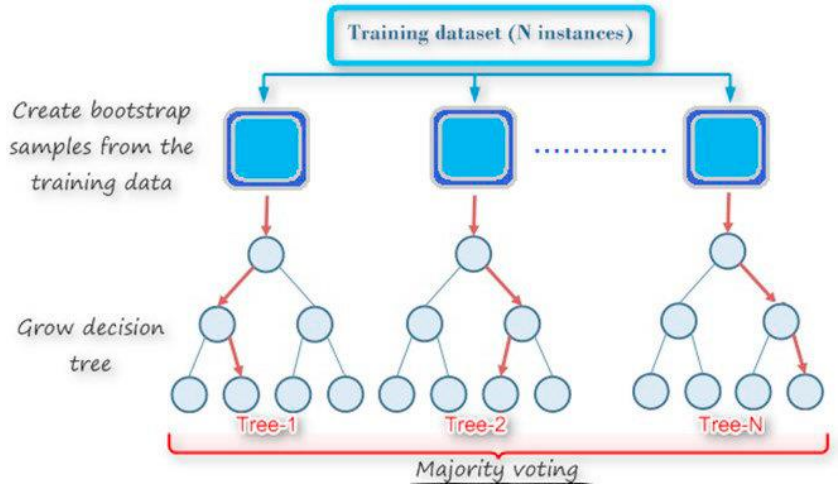
Telah dikembangkan berbagai macam cara untuk meningkatkan akurasi dari NER seperti dengan menggunakan algoritma *machine learning* atau memperbaiki proses *processing* dan ekstraksi fitur pada data yang digunakan. Salah satu cara yang paling sering digunakan dan memberikan peningkatan akurasi yang cukup tinggi pada proses NER adalah dengan melakukan teknik POS *tagging*. Teknik ini berfungsi untuk memberikan kelas pada kata atau token berdasarkan fungsi sintaksisnya pada struktur kalimat. Melalui teknik ini akurasi NER dapat ditingkatkan karena POS *tagging* membantu NER dalam memahami konteks sintaksis kata dalam kalimat dan membantu dalam pembentukan *chunk* yang lebih baik (Nguyen dkk., 2019).

2.8 *Random Forest*

Random Forest merupakan salah satu metode *machine learning* yang dapat digunakan untuk tugas regresi dan klasifikasi dan umumnya digunakan pada data tabular atau data teks dalam jumlah yang banyak. Algoritma *Random Forest* pertama kali diperkenalkan pada tahun 2001 oleh Leo Breiman, seorang profesor di Departemen Statistik *University of California*, Berkeley. Breiman dan timnya memperkenalkan konsep *ensemble learning*, yang menggabungkan prediksi dari beberapa model *machine learning* yang berbeda untuk meningkatkan akurasi prediksi. *Random Forest* adalah salah satu jenis *ensemble learning* yang menggabungkan banyak pohon keputusan (*Decision Tree*) dengan menggunakan metode CART (*Classification and Regression Trees*) untuk untuk membangun model dalam bentuk pohon keputusan yang berfungsi untuk memperbaiki akurasi prediksi. Konsep *Random Forest* sendiri terinspirasi oleh algoritma Breiman sebelumnya, yaitu *bagging*, yang juga merupakan teknik *ensemble learning* yang

menggabungkan banyak model untuk meningkatkan akurasi (Fadiyah Basar dkk., 2022).

Random Forest adalah algoritma yang termasuk ke dalam teknik *supervised learning* yang pertama kali diperkenalkan oleh Leo Breiman pada tahun 2001 dengan menggabungkan teknik *bootstrap aggregating* dengan *resampling*. *Supervised learning* adalah jenis *machine learning* di mana model dilatih menggunakan dataset yang telah diberikan label. Ada beberapa kelebihan dari metode *Random Forest* yaitu hasil akurasi yang bagus. Dan juga mampu mengatasi *missing value*, *data imbalance*, dan *noise* yang ada pada data, serta algoritma ini cocok untuk mengklasifikasikan data dalam jumlah yang besar (Fadiyah Basar dkk., 2022). Adapun ilustrasi alur kerja pada algoritma *Random Forest* dapat dilihat pada Gambar 2 berikut ini.



Gambar 2 Algoritma *Random Forest*
Sumber: (Willy dkk., 2021)

Gambar 2 adalah ilustrasi dari algoritma sederhana *Random Forest* yang menggambarkan bagaimana *Random Forest* dapat menghasilkan prediksi menggunakan *majority voting*. Dari gambar tersebut, terlihat bahwa algoritma *Random Forest* terdiri dari kombinasi beberapa pohon keputusan (*Decision Trees*), di mana setiap pohon bergantung pada nilai vektor acak yang diambil sebagai sampel secara merata pada semua pohon dalam hutan tersebut. Metode *Random Forest* terdiri dari dua tahap yaitu pembentukan hutan dan *voting* hasil klasifikasi (Willy dkk., 2021). Adapun rumus dari dari algoritma *Random Forest* dapat dilihat pada Persamaan (1) dan Persamaan (2) berikut ini.

$$forest = \{h(x, \theta_k), k = 1, \dots\} \quad (1)$$

dimana:

h : Hipotesa atau klasifikasi

x : Input *vector*

θ_k : *Independent and identically distributed random vectors*

Persamaan (1) menyatakan bahwa *forest* terbentuk dari sekumpulan klasifikasi atau hipotesa yang berjumlah *k*. Input tiap hipotesis adalah *x* yang kemudian dilakukan *resampling* dengan *random vector* dari *x* itu sendiri (Willy dkk., 2021).

$$C_{rf} = majority\ vote\ \{C_n(x)\} = 1 \quad (2)$$

dimana:

C_{rf} : *Class* hasil klasifikasi *Random Forest*

x : Input *vector*

C_n : *Class* prediksi dari *tree* ke-n pada *Random Forest*

Setelah dilakukan pembuatan forest, pada Persamaan (2) merupakan langkah selanjutnya yaitu melakukan *voting* untuk klasifikasi dan mengukur performa dari *Random Forest* (Willy dkk., 2021).

2.9 *DictVectorizer*

Dalam NLP, representasi data yang efisien dan efektif adalah kunci untuk membangun model pembelajaran mesin yang kuat. Salah satu teknik yang digunakan untuk mengubah data teks ke dalam bentuk yang dapat diproses oleh algoritma pembelajaran mesin adalah vektorisasi fitur. *DictVectorizer* adalah salah satu alat yang disediakan oleh *library scikit-learn* untuk melakukan transformasi ini. *DictVectorizer* mengubah fitur kategori menjadi fitur numerik menggunakan teknik *one-hot encoding*. Setiap kategori unik dalam fitur akan diubah menjadi kolom baru dalam matriks keluaran, dengan nilai 1 menunjukkan keberadaan kategori tersebut dan 0 menunjukkan ketiadaannya (Azalia dkk., 2019). Adapun rumus *one-hot encoding* yang digunakan dapat dilihat pada Persamaan (3) berikut ini.

$$\begin{cases} 1\ jika\ x_i = v_j \\ 0\ lainnya \end{cases} \quad (3)$$

DictVectorizer menggunakan rumus one-hot encoding yang dapat dilihat pada Persamaan (3) di atas. Di mana x_i adalah fitur kategori dan v_j adalah nilai unik dari fitur tersebut. Misalkan kita memiliki fitur kategori x_i yang dapat mengambil nilai $v_1, v_2, \dots \dots, v_k$. Untuk setiap nilai fitur $x_i = v_j$, *DictVectorizer* akan membuat kolom baru x_{ij} untuk setiap fitur tersebut lalu diberikan nilai 1 jika $x_i = v_j$ (*true*) dan 0 untuk lainnya (*false*). Kolom baru ini akan disimpan dalam format matriks array (Azalia dkk., 2019).

2.10 WebGIS

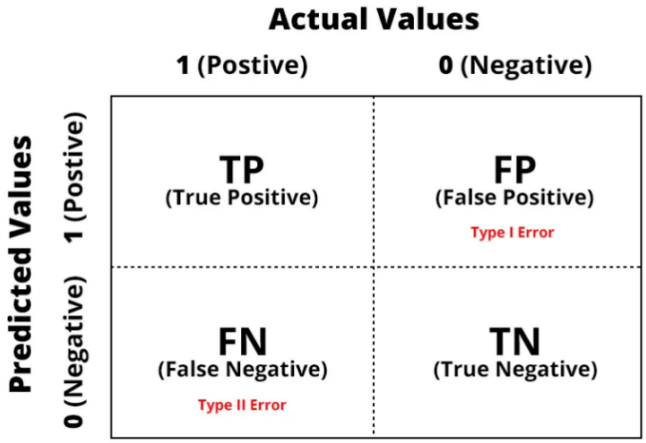
WebGIS adalah aplikasi *Geografis Information System* (GIS) berbasis *website* atau pemetaan digital yang memanfaatkan jaringan internet sebagai media komunikasi yang berfungsi mendistribusikan, mempublikasikan, mengintegrasikan, mengkomunikasikan, dan menyediakan informasi dalam bentuk teks, peta digital serta menjalankan fungsi-fungsi analisis dan *query* yang terkait dengan GIS melalui jaringan internet. WebGIS adalah setiap GIS yang menggunakan teknologi web untuk berkomunikasi antar sistem. WebGIS merupakan jenis sistem informasi yang terdistribusi. Bentuk paling sederhana dari WebGIS yaitu harus terdiri setidaknya dari server dan *client* (Sutanta dkk., 2021).

WebGIS menyediakan mekanisme dan metode baru yang efektif dalam pengembangan sistem informasi geografis (SIG). Adapun arsitektur WebGIS terdiri dari server yang kemudian server tersebut dapat diakses oleh *client* melalui internet. Terdapat beberapa platform aplikasi yang menyediakan layanan ini seperti Google Maps, Google Earth, Open Street Maps, Yahoo Maps dan banyak aplikasi komersial dan non-komersial lainnya menyediakan berbagai jenis informasi terkait geografis. Informasi yang dimaksud dapat berupa peta yang terperinci, citra satelit, dan peta daerah yang mencakup seluruh dunia secara global (Sutanta dkk., 2021).

2.11 *Confusion Matrix*

Confusion matrix juga sering disebut *error matrix*. *Confusion matrix* berfungsi untuk memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil klasifikasi sebenarnya. *Confusion matrix* berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi

pada serangkaian data uji yang nilai sebenarnya diketahui. Gambar 3 berikut ini merupakan *confusion matrix* dengan 4 kombinasi nilai prediksi dan nilai aktual yang berbeda (Krstinić dkk., 2020).



Gambar 3 *Confusion matrix*
Sumber: (Jefiza dkk., 2023)

Gambar 3 merupakan gambar *confusion matrix* yang memiliki 4 istilah yang mengartikan presentasi hasil dari proses klasifikasi. Keempat istilah tersebut adalah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN) sebagai berikut.

1. *True Positive* (TP) merupakan data positif yang diprediksi benar.
2. *True Negative* (TN) merupakan data negatif yang diprediksi benar.
3. *False Positive* (FP) atau tipe 1 *error* merupakan data negatif namun diprediksi sebagai data positif.
4. *False Negative* (FN) atau tipe 2 *error* merupakan data positif namun diprediksi sebagai data negatif.

Confusion matrix dapat digunakan untuk menghitung berbagai *performance metrics* yang berfungsi untuk mengukur kinerja model klasifikasi yang telah dibuat yaitu *accuracy*, *precision*, *recall*, dan *f1-score* (Jefiza dkk., 2023).

1. *Accuracy*

   Accuracy adalah metrik yang digunakan untuk mengukur seberapa sering model prediksi memberikan hasil yang benar, baik itu untuk kelas positif maupun kelas negatif. *Accuracy* dihitung dengan membandingkan jumlah prediksi yang benar yaitu *True Positive* (TP) maupun *True Negative* (TN) dengan keseluruhan jumlah prediksi yang dibuat oleh model. Nilai *accuracy*

dalam konteks pembelajaran mesin dapat diperoleh dengan Persamaan (4) berikut ini:

$$accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) x\ 100 \quad (4)$$

Atau rumus umumnya dapat diperoleh dengan Persamaan (5) berikut ini:

$$accuracy = \left(\frac{Total\ Data\ Benar}{Total\ Data}\right)\ x\ 100 \quad (5)$$

2. *Precision*

Precision adalah metrik evaluasi yang berfungsi untuk mengukur seberapa baik model dalam membuat prediksi yang benar untuk kelas positif dari total prediksi positif yang dilakukan. Lengkapnya, *precision* merupakan rasio prediksi benar positif yang dibandingkan dengan keseluruhan hasil yang diprediksi positif oleh model sehingga dapat diketahui seberapa akurat prediksi positif dari model. *Precision* sering digunakan ketika kesalahan *False Positive* (FP) perlu diminimalkan. Nilai *precision* dapat diperoleh dengan Persamaan (6) berikut ini:

$$precision = \left(\frac{TP}{TP+FP}\right) x\ 100 \quad (6)$$

3. *Recall*

Recall adalah metrik evaluasi yang berfungsi untuk mengukur seberapa baik model mampu menemukan semua *instance* positif yang sebenarnya. *Recall* mengukur sensitivitas model, semakin kecil jumlah *False Negative* (FN) berarti model mampu menemukan hampir semua data positif yang ada. *Recall* sangat berguna dalam situasi di mana *False Negative* (FN) perlu diminimalkan. Nilai *recall* dapat diperoleh dengan Persamaan (7) berikut ini:

$$recall = \left(\frac{TP}{TP+FN}\right)\ x\ 100 \quad (7)$$

4. *F1-score*

F1-score adalah metrik yang digunakan untuk menyeimbangkan *precision* dan *recall* dalam satu nilai, sehingga memberikan gambaran seberapa baik model dalam memprediksi data positif secara akurat dan juga menangkap semua data positif yang sebenarnya. Nilai *f1-score* dapat diperoleh dengan Persamaan (8) berikut ini:

$$f1 - score = 2\ \left(\frac{precision\ x\ recall}{precision + recall}\right)\ x\ 100 \quad (8)$$