

Geographic Artificial Intelligence GeoAI dan Natural Language Processing dalam Analisis Data Spatial

Endin Fahrudin¹, Samsu Supriyatna², Firman Darmawan³

^{1,2}Program Studi Sistem Informasi Fakultas Ilmu Komputer, Universitas Pamulang, Indonesia

³Badan Pertanahan Nasional, Indonesia

dosen02729@unpam.ac.id, dosen02830@unpam.ac.id, firmmandarmawanbnp@gmail.com

Article Info

Article history:

Received Mar 10, 2025

Revised Sept 09, 2025

Accepted Sept 13, 2025

Keywords:

Geographic
Artificial Intelligence
Natural Language Processing
Fuzzy Similarity
Spatial Data
Textual Data

ABSTRACT

Geographic Artificial Intelligence (GeoAI) and Natural Language Processing (NLP) are two rapidly advancing technologies in spatial data analysis. Geo AI integrates artificial intelligence with geographic data to identify patterns, make predictions, and support decision-making. NLP enables the processing and analysis of textual data related to spatial information, such as documents or reports. This research applies GeoAI, NLP, and Fuzzy Similarity methods to spatial tax and textual land data to identify data similarities and spatial patterns. The study utilizes artificial intelligence to extract spatial information from text and perform machine learning-based analysis, enhancing the efficiency and accuracy of geographic data interpretation. The results demonstrate that combining GeoAI and NLP improves the understanding of spatial patterns in unstructured data and supports location-based decision-making. This research contributes to developing more accurate spatial analysis techniques for various fields, such as urban planning, disaster management, and environmental analysis.

Corresponding Author:

Endin Fahrudin,
Program Studi Sistem Informasi,
Fakultas Ilmu Komputer Universitas Pamulang,
Jl. Puspatek Raya No 10, Serpong - Tangerang Selatan.
Email: dosen02729@unpam.ac.id

1. PENDAHULUAN

Indonesia merupakan negara kepulauan terbesar di dunia dengan total pulau mencapai 13.466, yang terdiri dari luas daratan 1.922.570 km² dan luas perairan 3.257.483 km². Data tersebut sesuai dengan informasi geospasial yang dikeluarkan oleh Badan Informasi Geospasial (BIG). Keragaman wilayah geografis Indonesia, mulai dari dataran rendah, dataran tinggi, hingga pegunungan, menghadirkan tantangan tersendiri dalam proses pemetaan dan analisis lokasi. Selain itu kepemilikan tanah yang bersertifikat perlu dilakukan kesesuaian dengan Surat Pemberitahuan Pajak Terhutang (SPPT) yang harus dibayarkan ke negara. Kemajuan teknologi, khususnya dalam bidang Geographic Artificial Intelligence (GeoAI) dan Natural Language Processing (NLP), memberikan peluang baru untuk mengatasi tantangan ini. GeoAI menggabungkan teknologi kecerdasan buatan dengan data geospasial untuk menganalisis dan memprediksi pola serta tren dalam data spasial. Sementara itu, NLP memungkinkan pemrosesan dan analisis data tekstual yang terkait dengan informasi geografis, seperti laporan cuaca, deskripsi lokasi, dan ulasan publik [1], [2]. Penerapan GeoAI dan NLP dalam analisis data spasial dapat memperbaiki akurasi pemetaan, meningkatkan efektivitas pengambilan keputusan, dan memberikan wawasan yang lebih mendalam mengenai dinamika geografis. Dengan menggunakan metode konvolusi untuk ekstraksi fitur dan teknik NLP untuk analisis teks, kita dapat mengidentifikasi pola, tren, dan hubungan dalam data spasial yang sebelumnya sulit diungkap. Teknologi ini juga berpotensi mengatasi

masalah terkait variasi dalam data dan memungkinkan integrasi informasi dari berbagai sumber. Misalnya, menggabungkan data pertanahan dengan deskripsi tekstual data perpajakan untuk memperoleh gambaran yang lebih komprehensif mengenai kondisi geografis dan pembayaran pajak bumi serta bangunan. Dengan demikian, penerapan GeoAI dan NLP bukan hanya meningkatkan kemampuan pemetaan tetapi juga memberikan alat yang lebih efektif untuk analisis dan perencanaan berbasis data di Indonesia [2].

Dalam pendekatan yang dilakukan antara data spasial atau tekstual perpajakan serta data spasial atau tekstual pertanahan untuk mendapatkan kemiripan data menggunakan metode Fuzzy Similarity. Metode Fuzzy Similarity pada prinsipnya mempelajari pola (fluktuasi) dari data spasial dan tekstual baik yang dimiliki oleh pertanahan [3], [4]. Penelitian ini bertujuan untuk mendapatkan gambar berupa pola data spasial melalui serangkaian proses, mulai dari tahapan identifikasi masalah dan analisis kebutuhan, pengumpulan dan pemrosesan, representasi dokumen, hingga penggunaan Geographic Artificial Intelligence (GeoAI), Natural Language Processing (NLP) dan Metode Fuzzy Similarity [3]. Selain itu, penelitian ini mencakup tahapan validasi model dan evaluasi hasil, sehingga menghasilkan pemahaman yang komprehensif dan relevan tentang data geospasial yang dapat memberikan kontribusi signifikan dalam pengembangan strategi mitigasi dan penanggulangan bencana di masa depan. Dengan pendekatan metodologi yang canggih dan terintegrasi ini, diharapkan penelitian ini dapat memberikan kontribusi berarti dalam pemahaman mendalam tentang data spasial serta membuka jalan menuju upaya-upaya yang lebih efektif dalam melindungi masyarakat dan aset-aset penting di Indonesia [5]. Dunia tengah berada dalam era digital di mana data yang tak terbatas tersedia di ujung jari kita. Penelitian ini diharapkan akan memberikan kontribusi yang berarti dalam memperluas wawasan kita tentang Geographic Artificial Intelligence serta membantu pembuat kebijakan untuk mengambil tindakan preventif yang tepat, serta meningkatkan ketangguhan masyarakat dalam pertanahan dan perpajakan masa depan. Dengan menggabungkan pengetahuan data tekstual atau spasial dari pertanahan dan perpajakan dengan kekuatan analisis Geographic Artificial Intelligence (GeoAI), Natural Language Processing (NLP) dan Metode Fuzzy Similarity, kita dapat membangun suatu khazanah baru mengenai geospasial berbasis teknologi terbaru [5].

GeoPandas adalah pustaka Python yang dirancang untuk menyediakan alat dalam pengolahan data geospasial dengan menggunakan struktur data tabular yang menyerupai Pandas. Pustaka ini dikembangkan untuk menjawab kebutuhan analisis spasial yang semakin kompleks dalam pemrograman Python [6]. Dengan memanfaatkan kekuatan Python dalam pengelolaan data geografis, GeoPandas menggabungkan kemampuan analisis data non-spasial dari Pandas dengan fungsi manipulasi data spasial, sehingga memungkinkan analisis yang komprehensif dalam satu kerangka kerja. GeoPandas adalah extension dari pandas, yang menambahkan dukungan untuk tipe data tambahan yang diperlukan untuk bekerja dengan data spasial dimana dapat melakukan operasi spasial dan menghasilkan visualisasi data spasial. Fungsi inti GeoPandas mencakup hal berikut:

- 1) Membaca dan menulis data spasial;
- 2) Struktur data spasial;
- 3) Manajemen proyek;
- 4) Visualisasi data spasial;
- 5) Manipulasi data;
- 6) Geocoding

Dalam beberapa tahun terakhir, GeoPandas telah menjadi salah satu alat penting dalam analisis data geospasial, menggantikan atau melengkapi peran yang sebelumnya didominasi oleh perangkat lunak GIS (*Geographic Information System*) tradisional [4].

GeoPandas menawarkan berbagai fungsi untuk analisis data spasial yang sebagian besar diadaptasi dari fungsi yang disediakan oleh Shapely. Beberapa operasi umum yang dapat dilakukan dengan menggunakan GeoPandas antara lain:

1. Penggabungan (*Union*): Fungsi ini digunakan untuk menggabungkan dua atau lebih geometri menjadi satu kesatuan, yang berguna untuk menyatukan wilayah atau menghapus batasan dalam analisis spasial.
2. Pemotongan (*Intersection*): Operasi ini digunakan untuk memotong dua geometri dan menemukan area tumpang tindih di antara keduanya. Fungsi ini sering diterapkan dalam analisis area yang terkena dampak dari berbagai faktor, seperti bencana alam atau pembangunan infrastruktur.
3. Buffering: Fungsi buffering membuat zona penyangga di sekitar geometri tertentu. Fungsi ini sangat berguna dalam analisis *proximity*, misalnya dalam menentukan jarak aman dari sungai atau jalan raya.
4. Penggabungan Data (*Spatial Join*): Penggabungan data ini mengintegrasikan data dari dua GeoData-Frame berdasarkan lokasi spasial mereka. Operasi ini serupa dengan SQL JOIN, namun dilakukan berdasarkan hubungan spasial antara geometri yang dibandingkan.

2. METODE

2.1. *Similarity Text Natural Language Processing*

Penelitian ini dilakukan dengan pendekatan Geographic Artificial Intelligence (GeoAI), Natural Language Processing (NLP) dan Metode Fuzzy Similarity terhadap data spasial dan textual dari pertanahan dan perpajakan untuk menemukan pola data spasial yang akurat. Penelitian ini melibatkan sejumlah tahapan yang terorganisir dengan baik, dimulai dari pengumpulan data hingga analisis, hingga hasil dan kesimpulan. Berikut adalah uraian rinci mengenai metode yang digunakan dalam penelitian ini:

1. Tahapan identifikasi dan analisis kebutuhan

Dalam tahap ini, langkah yang dilakukan mencakup identifikasi struktur dan karakteristik dari data yang akan diolah, seperti asal data, jenis teks, format, volume, dan metadata yang terkandung di dalamnya. Selain itu, tahap ini juga melibatkan analisis mendalam terhadap kebutuhan yang diinginkan dari hasil akhir proses Geographic Artificial Intelligence (GeoAI), Natural Language Processing (NLP) dan Metode Fuzzy Similarity, termasuk tujuan analisis, kriteria keberhasilan, kecepatan dan skala yang diharapkan, serta pemilihan teknologi dan algoritma yang paling sesuai. Analisis yang teliti pada tahap identifikasi dan analisis kebutuhan membentuk dasar yang kokoh untuk perencanaan selanjutnya dalam proses similarity, memastikan bahwa pendekatan yang diambil sesuai dengan tujuan bisnis atau penelitian yang ingin dicapai [7].

2. Pengumpulan dan Pemrosesan Data atau Dokumen

Setelah identifikasi kebutuhan dan struktur data dilakukan, tahap berikutnya adalah pengumpulan data textual dan spasial dari berbagai sumber. Data utama yang digunakan dalam penelitian ini berasal dari dua instansi penting, yaitu Kementerian Agraria dan Tata Ruang/Badan Pertanahan Nasional (ATR/BPN) yang menyediakan data terkait pertanahan, serta Direktorat Jenderal Pajak untuk data perpajakan. Kedua jenis data ini sangat penting untuk mendapatkan gambaran komprehensif dalam menganalisis keterkaitan antara kondisi lahan dan implikasinya terhadap perpajakan. Data yang dikumpulkan terdiri dari data spasial, seperti peta dan koordinat lahan, serta data tekstual yang berisi deskripsi atau dokumen administratif terkait hak atas tanah, pajak, dan catatan hukum lainnya. Pengolahan data melibatkan beberapa tahap krusial, antara lain:

a. Preprocessing Data Spasial

Pada tahap ini, data spasial diproses untuk memastikan kualitasnya, seperti melakukan validasi koordinat geografis dan merapikan format data sesuai standar yang digunakan dalam GeoAI. Pemetaan dilakukan untuk mengelompokkan wilayah berdasarkan karakteristik fisik dan administratif [8].

b. Preprocessing Data Tekstual

Data teks yang dikumpulkan kemudian diproses melalui Natural Language Processing (NLP) untuk tokenisasi, stemming, dan normalisasi. Tahap ini penting untuk menyederhanakan data dan mempersiapkannya untuk dianalisis menggunakan teknik Fuzzy Similarity [9].

c. Normalisasi dan Integrasi Data

Kedua tipe data (spasial dan textual) perlu dinormalisasi dan diintegrasikan. Ini melibatkan pemetaan antara informasi geografis dan teks administratif sehingga keduanya dapat digunakan bersama dalam analisis lanjutan.

d. Penerapan Geographic Artificial Intelligence (GeoAI) dan NLP

Setelah data diproses, langkah berikutnya adalah menerapkan GeoAI untuk mengidentifikasi pola geografis dari data spasial. GeoAI menggabungkan algoritma kecerdasan buatan dengan analisis spasial untuk menemukan hubungan dan tren yang mungkin tidak terlihat secara langsung melalui analisis manual. Model ini dapat digunakan untuk mendeteksi perubahan penggunaan lahan atau pola yang mempengaruhi perpajakan dan kebijakan lahan. Di sisi lain, NLP digunakan untuk menganalisis data teks dari dokumen pertanahan dan perpajakan. Metode ini membantu dalam mengekstraksi informasi penting dari teks dan mengidentifikasi entitas atau kata kunci yang relevan dengan analisis lahan dan perpajakan.

2.2. Metode Fuzzy Similarity untuk Analisis Data

Setelah data spasial dan tekstual diproses, metode Fuzzy Similarity digunakan untuk menghitung kemiripan antara data lahan dan perpajakan yang ada. Metode ini memungkinkan pendekatan yang lebih fleksibel dan toleran terhadap ketidakpastian dalam data, terutama ketika menangani variasi atau ketidaksesuaian format antara data spasial dan teks. Fuzzy Similarity mengukur sejauh mana dua elemen (baik spasial maupun tekstual) memiliki kemiripan, membantu dalam membuat keputusan yang lebih akurat mengenai pemetaan lahan dan penilaian pajak.

2.3. Analisis dan Interpretasi Hasil

Setelah menerapkan metode GeoAI, NLP, dan Fuzzy Similarity, data dianalisis untuk menemukan pola dan hubungan yang relevan antara kondisi data pertanahan dan perpajakan. Hasil analisis ini diinterpretasikan untuk memberikan wawasan yang dapat digunakan dalam rekomendasi data sah dan tidak sah dan juga sebagai perencanaan kebijakan public, khususnya yang terkait dengan pengelolaan pertanahan dan perpajakan di Indonesia.

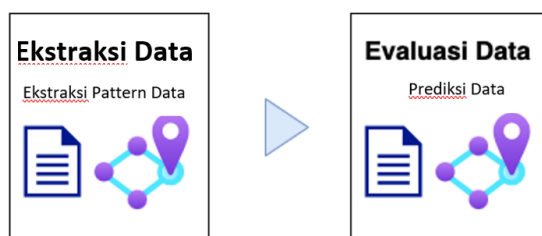
2.4. Validasi dan Evaluasi

Pada tahap akhir, model yang telah dibuat diuji dan divalidasi menggunakan data uji yang terpisah. Validasi dilakukan untuk memastikan bahwa hasil yang diperoleh konsisten dan dapat diandalkan. Evaluasi dilakukan untuk mengukur keakuratan model, apakah dapat memenuhi kriteria keberhasilan yang telah ditetapkan pada tahap identifikasi awal [10].

3. HASIL DAN PEMBAHASAN

3.1. Metode Teknologi *Artificial Intelligence* dalam Analisis Data Teksual

Training model digunakan untuk melatih atau membangun model. *Validation dataset* adalah himpunan data yang digunakan untuk mengoptimasi saat melatih model. Model dilatih menggunakan training dataset, kemudian kinerja saat latihan tersebut diuji menggunakan validation dataset. Hal ini bertujuan untuk melihat kemampuan model pada saat training apakah dapat mengenal pola secara umum.



Gambar 1. Ekstraksi Data PBT dan NOP

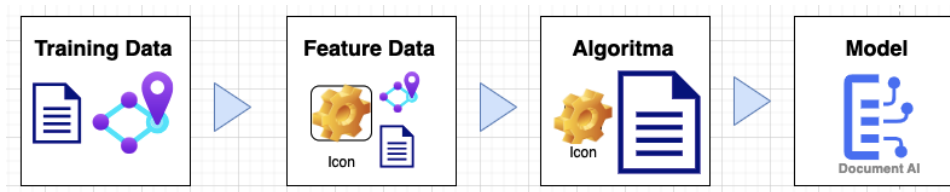
Evaluasi Model Data dilakukan dengan menguji kinerja model pada data yang sebelumnya tidak terlihat. Data tak terlihat yang digunakan adalah set pengujian.

Data informasi yang relatif lengkap dapat mendukung deteksi kesesuaian sehingga dilakukan pengumpulan data untuk mendapatkan informasi atau materi. Adapun metode pengumpulan data sebagai berikut :

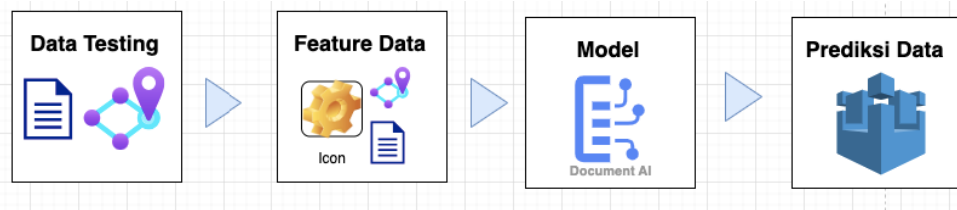
- a. *Data training*
Data training atau data latih adalah data yang digunakan untuk membangun model klasifikator yang sudah diketahui data NIB dan NOP pada spatial dan tekstual. Pada penelitian ini data yang digunakan sebagai *data training*.
- b. *Data Testing*
Data testing atau data uji adalah data yang belum diketahui label untuk kemudian diprediksi kelasnya menggunakan model klasifikasi yang sudah dibangun dari data *training*.

3.1.1. Prediksi Model dan *Testing*

Dalam proses pengembangan model *Machine Learning*, terdapat dua tahapan penting yaitu pelatihan (*training*) dan pengujian (*testing*) model. Berikut adalah langkah-langkah yang dilakukan dalam kedua tahapan ini:



Gambar 2. *Data training*



Gambar 3. Data testing

a. *Data Training dan Model Cleansing*

- Pengumpulan *Data Training*: *Data training* adalah kumpulan data spasial pertanahan dan textual perpajakan yang digunakan untuk melatih modelnya yang akan dijadikan pola dan struktur data.
- Feature Extraction: Data training yang telah didapatkan akan menjadi data yang kemudian diolah datanya. Data inilah menjadi representasi data lebih mudah diolah oleh algoritma.
- Pemilihan Algoritma: Algoritma dipilih sangat penting mempelajari feature data yang didapat dari data training dan membuat model. Model Algoritma ini akan di latih menggunakan data training.
- Model Training: Model berdasarkan pelatihan menggunakan data training yang sudah di gabung antara data spasial pertanahan dan textual perpajakan. Model ini dapat menggeneralisasi pola data tanpa terpengaruh oleh data yang tidak punya relasi antar data spasial pertanahan dan textual perpajakan.
- Model Cleansing: Model yang sudah dilatih dapat akan disesuaikan dengan mengurangi atau menghilangkan data yang tidak punya relasi antar data spasial pertanahan dan textual perpajakan, sehingga model mendapat data yang relevan untuk mendapatkan akurasi prediksi yang baik.

b. *Uji Model dan Prediksi*

- Pengumpulan Data Testing: Data testing adalah kumpulan data spasial pertanahan dan textual perpajakan yang digunakan untuk menguji model yang dibuat, sehingga dapat memberikan penilaian akurasi data yang objektif .
- Feature Extraction pada Data Testing: Data testing spasial pertanahan dan textual perpajakan diolah menjadi feature data, termasuk data training. Feature data testing dimasukkan ke dalam model untuk membuat data prediksi.
- Uji Model (Testing): Model yang sudah dibuat akan terlatih dan diaplikasikan pada data testing untuk mendapatkan hasilnya. Data hasil prediksi akan dibandingkan dengan data yang aslinya untuk mengevaluasi kinerja model yang dibuat.
- Model Refinement: Dari data yang di hasil pada uji model, dapat disesuaikan dengan memperbaiki kinerja algoritma menggunakan data lebih banyak atau merubah parameter guna mendapatkan hasil akurasi prediksi yang meningkat.
- Prediksi Data Baru: Model diperbaiki dan divalidasi, dapat digunakan untuk memprediksi data baru lebih baik dan lebih akurat.

3.1.2. Data Selisih Textual Pertanahan dan Perpajakan

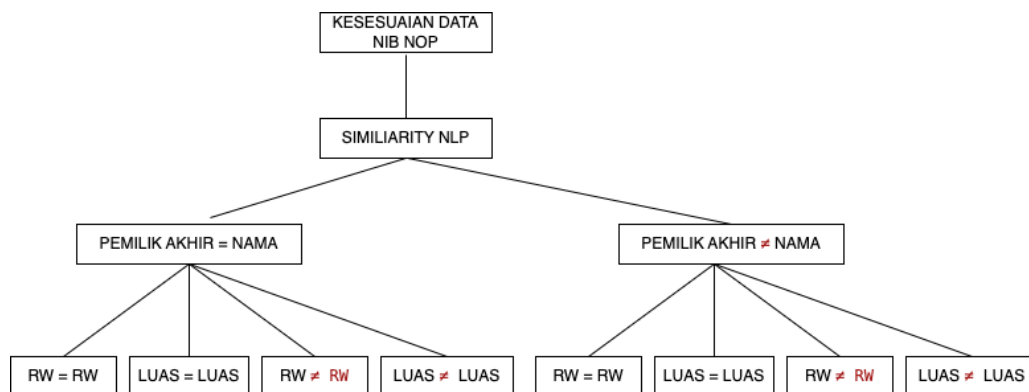
Sinkronisasi data tekstual dan spasial bidang tanah diperlukan untuk menyatukan informasi penting seperti NIB dan atribut pendukung lainnya dengan lokasinya secara geografis. Hal ini penting untuk melakukan pemetaan yang akurat, analisis lanjutan, hingga kelanjutan manajemen data.

Proses sinkronisasi diperoleh dengan melakukan *overlay* peta foto udara dan PBT, serta penyesuaian data tekstual dan spasial untuk menghitung luas dan jumlah bidang, menentukan *gap* dan *overlap* antar bidang, memeriksa ada atau tidak adanya NIB, serta mencocokkan data tekstual dan spasial pertanahan, termasuk identifikasi selisih atau perbedaan data. Berikut adalah hasil sinkronisasi data pertanahan di Kota Bogor pada tabel dibawah ini.

Wilayah		Inventarisasi Data Perpajakan						
No	Kota Bogor	Penyamaan Posisi Blok PBB						
	Kecamatan Bogor Barat	(Spatial Adjusment/Perekat)		Keterangan	Data Spasial	Data Tekstual	Selisih Data	Persentase
	Peta Blok PBB							
1	Cilendek Barat							
			Jumlah Nop	2416	2414	2		
			Jumlah Bidang Tidak ada Nop	43	0	43		
			Jumlah Total Bidang	2459	2414	45		
			Jumlah Bidang NOP dapat diandingkan	509				21%
			Jumlah Bidang NOP Tidak dapat diandingkan	1950				79%
			Matching Data					
			-NOP (V Spasial = V Textual)	1103	1103	0		29%
			-NOP (V Spasial = X Textual)	1356	0	1356		36%
			-NOP (X Spasial = V Textual)	0	1311	-1311		35%
2	Cilendek Timur							
			Jumlah Nop	1525	2015	-490		
			Jumlah Bidang Tidak ada Nop	74		74		
			Jumlah Total Bidang	1599		1599		
			Jumlah Bidang NOP dapat diandingkan	449				28%
			Jumlah Bidang NOP Tidak dapat diandingkan	1150				72%
			Matching Data					
			-NOP (V Spasial = V Textual)	601	601	0		20%
			-NOP (V Spasial = X Textual)	998	0	998		33%
			-NOP (X Spasial = V Textual)	0	1414	-1414		47%
3	Curug Mekar							
			Jumlah Nop	3274	1673	1601		
			Jumlah Bidang Tidak ada Nop	106	0	106		
			Jumlah Total Bidang	3380	1673	1707		
			Jumlah Bidang NOP dapat diandingkan	1839				54%
			Jumlah Bidang NOP Tidak dapat diandingkan	1541				46%
			Matching Data					
			-NOP (V Spasial = V Textual)	1085	1085	0		27%
			-NOP (V Spasial = X Textual)	2295	0	2295		58%
			-NOP (X Spasial = V Textual)	0	588	-588		15%

Gambar 4. Data Gap dan Overlap

3.1.3. Model Natural Language Processing – Geopandas



Gambar 5. Similiarity NLP-GeoPandas

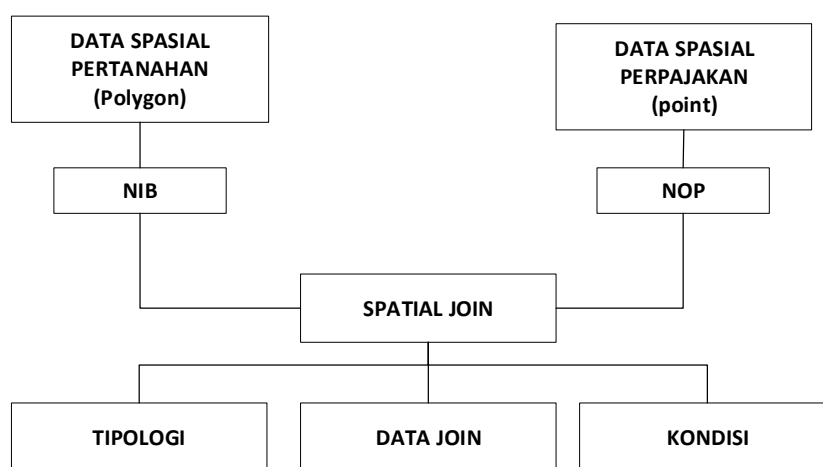
Model NLP-Geopandas dirancang untuk mengevaluasi kesesuaian data antara Nomor Induk Bidang (NIB) dan Nomor Objek Pajak (NOP). Berikut adalah langkah-langkah yang dilakukan untuk membangun model *Natural Language Processing* (NLP) – *Geopandas* untuk klasifikasi kesesuaian data NIB dan NOP:

1. Kesuaian Data NIB dan NOP
Langkah pertama dalam proses ini adalah mengevaluasi kesesuaian data antara NIB dan NOP dengan menggunakan teknik *Natural Language Processing* (NLP) untuk mengukur *similarity* (kemiripan) antara kedua set data tersebut.
2. Pengecekan *Similarity* dengan *Natural Language Processing* (NLP)
Proses ini melibatkan analisis *similarity* menggunakan *Natural Language Processing* untuk memastikan bahwa nama pemilik akhir yang tercatat di data NIB dan NOP adalah konsisten.
3. Klasifikasi Berdasarkan Kesamaan Nama Pemilik Akhir
Setelah *similarity* diukur, data dibagi menjadi dua kategori:
 - 1) Pemilik Akhir = Nama: Dalam kategori ini, nama pemilik akhir di NIB dan NOP adalah sama.
 - 2) Pemilik Akhir ≠ Nama: Dalam kategori ini, nama pemilik akhir di NIB dan NOP tidak sama.
4. Verifikasi RW dan Luas

Untuk setiap kategori di atas, dilakukan verifikasi lebih lanjut untuk dua aspek tambahan: RW (Rukun Warga) dan luas tanah:

- 1) $RW = RW$: RW yang tercatat di kedua set data adalah sama.
 - 2) $RW \neq RW$: RW yang tercatat di kedua set data berbeda.
 - 3) $Luas = Luas$: Luas tanah yang tercatat di kedua set data adalah sama.
 - 4) $Luas \neq Luas$: Luas tanah yang tercatat di kedua set data berbeda.
5. Proses Verifikasi
- 1) Jika Pemilik Akhir = Nama:
 - Verifikasi apakah RW di kedua data sesuai ($RW = RW$) atau tidak sesuai ($RW \neq RW$).
 - Verifikasi apakah luas tanah di kedua data sesuai ($Luas = Luas$) atau tidak sesuai ($Luas \neq Luas$).
 - 2) Jika Pemilik Akhir $\neq$ Nama:
 - Verifikasi apakah RW di kedua data sesuai ($RW = RW$) atau tidak sesuai ($RW \neq RW$).
 - Verifikasi apakah luas tanah di kedua data sesuai ($Luas = Luas$) atau tidak sesuai ($Luas \neq Luas$).

Dengan menggunakan model NLP dan teknik analisis data dari Geopandas, proses ini memastikan kesesuaian dan konsistensi data antara NIB dan NOP, serta melakukan verifikasi lebih lanjut terhadap parameter RW dan luas tanah untuk memberikan klasifikasi yang lebih mendetail.



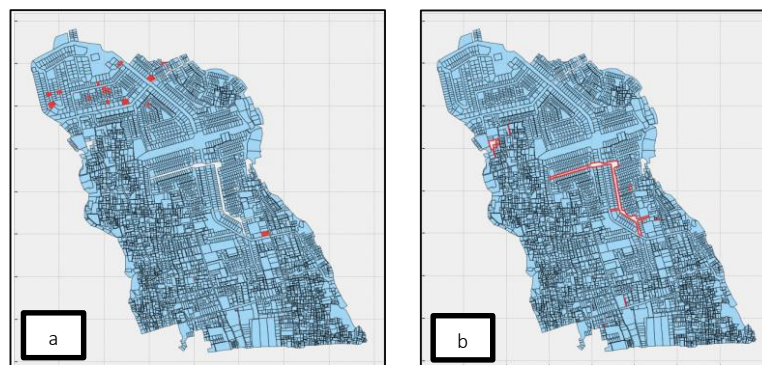
Gambar 6 *Spatial JOIN*

Model *Spatial Join* – *Geopandas* dan *Pandas* dilakukan untuk mendapatkan klasifikasi dari data spasial pertanahan yang berupa data *polygon* yang kemudian dijoinkan dengan data spasial perpajakan yang berupa *point*. Dari analisis tersebut diperoleh klasifikasi data join berdasarkan tipologi dan kondisi data. Tipologi yang dimaksud adalah sebagai berikut :

1. *One To One* (1 NIB -> 1 NOP)
 - 1 (satu) NIB memiliki 1 (satu) NOP. Setiap bidang tanah yang memiliki NIB hanya memiliki 1 (satu) NOP.
2. *One To Many* (1 NIB -> NOP, NOP, NOP)
 - 1 (satu) NIB dengan banyak NOP. Satu Bidang tanah dengan 1 (satu) NIB memiliki banyak NOP
3. *Many To One* (NIB, NIB, NIB -> NOP)
 - Banyak NIB dengan 1 NOP. Banyak bidang tanah dengan banyak NIB memiliki 1 (satu) NOP
4. *One To None* (1 NIB -> no NOP)
 - 1 (satu) NIB belum ada NOP. Satu bidang tanah dengan 1 NIB belum memiliki NOP

3.1.4. Hasil Deteksi Gap dan Overlap Data Polygon

Analisa data spasial adalah untuk mengetahui data spasial yang beririsan, overlap dan gap, dan memisahkan data yang sudah dipisahkan data intersection, overlap dan gap menjadi data cleansing, sehingga data yang didapatkan hasil similarity by nama menjadi bersih



Gambar 7. a. Overlap dan b. Gaps

3.1.5. Nilai Scoring Similarity by Nama

Natural Language Processing (NLP) memungkinkan komputer dan perangkat *digital* mengenali, memahami, dan menghasilkan teks dan ucapan dengan menggabungkan linguistik komputasional—pemodelan bahasa manusia berbasis aturan—bersama dengan pemodelan statistik.

Natural Language Processing (NLP) merupakan sebuah teknologi *machine learning* yang memberi komputer kemampuan untuk menginterpretasikan, memanipulasi, dan memahami bahasa manusia.

a. N-GRAM

Model probabilistik *N-gram*, merupakan model yang digunakan untuk memprediksi kata berikutnya yang mungkin dari kata $N-1$ sebelumnya. Model statistika dari urutan kata ini seringkali disebut juga sebagai model bahasa (*language models / LMs*).

Model estimasi seperti *N-gram* memberikan probabilitas kemungkinan pada kata berikutnya pada data “Pemilik Akhir” yang di sandingkan dengan “Nama” yang mungkin dapat digunakan untuk melakukan kemungkinan penggabungan pada keseluruhan kalimat. Model *N-gram* merupakan model yang paling penting dalam pemrosesan **Similarity by Nama** untuk mendapatkan memperkirakan probabilitas kata berikutnya maupun keseluruhan *sequence*. *N-gram* cukup esensial pada banyak hal dimana kata-kata demi kata yang ada di data dapat lebih tepat disimpulkan.

PEMILIK AKHIR	NAMA
AYUNAH AKO	ayunah ako
SAINURLINA	sainurlina
SAINURLINA	sainurlina
SYARIFULLAH	ir. syarifullah
NURHASAN	nurhasan

Gambar 7. Persandingan data dan probabilitas kemungkinan kata

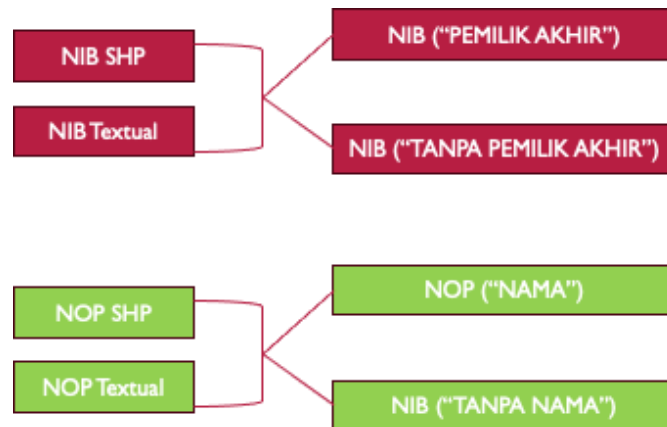
b. TF-IDF

Frekuensi Istilah - Frekuensi Dokumen Terbalik atau *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah metode statistik yang banyak digunakan dalam *Natural Learning Processing* (NLP) dan pengambilan informasi. Metode ini mengukur seberapa penting suatu istilah dalam suatu dokumen relatif terhadap kumpulan dokumen (yaitu, relatif terhadap korpus).

Natural Learning Processing (NLP) digunakan untuk mengevaluasi pentingnya kata-kata berbeda dalam yang ada pada data “Pemilik Akhir” Pertanahan dan “Nama” Perpajakan. Hal ini berguna dalam klasifikasi teks dan untuk membantu model pembelajaran mesin membaca kata-kata. Kata-kata dalam dokumen teks diubah menjadi angka penting melalui proses vektorisasi teks. Ada banyak skema penilaian vektorisasi teks yang berbeda, dengan TF-IDF menjadi salah satu yang paling umum.

Secara keseluruhan, TF-IDF adalah alat yang sangat penting dalam *toolkit* NLP, membantu dalam memprioritaskan kata-kata yang paling relevan dan bermakna dalam teks untuk berbagai aplikasi dan analisis.

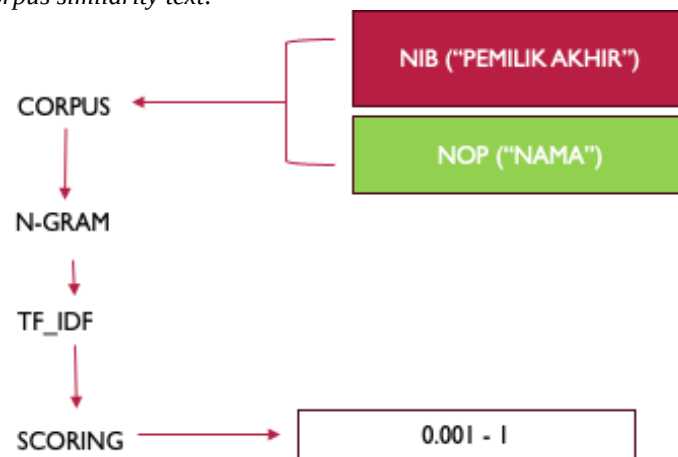
Pada kajian ini, data yang digunakan adalah data dari *join* NIB antara data SHP dengan tekstual pertanahan dengan data *join* NOP antara data SHP perpajakan dan tekstual perpajakan.



Gambar 8. Join data tekstual NIB dan NOP

Data pertanahan yang di *join* akan dipisahkan menjadi dua, yaitu data NIB yang mempunyai “Pemilik Akhir” dan NIB tanpa “Pemilik Akhir”. Sedangkan data perpajakan yang di *join* akan dipisahkan menjadi dua, yaitu data NOP yang mempunyai “NAMA” dan NOP tanpa “NAMA”.

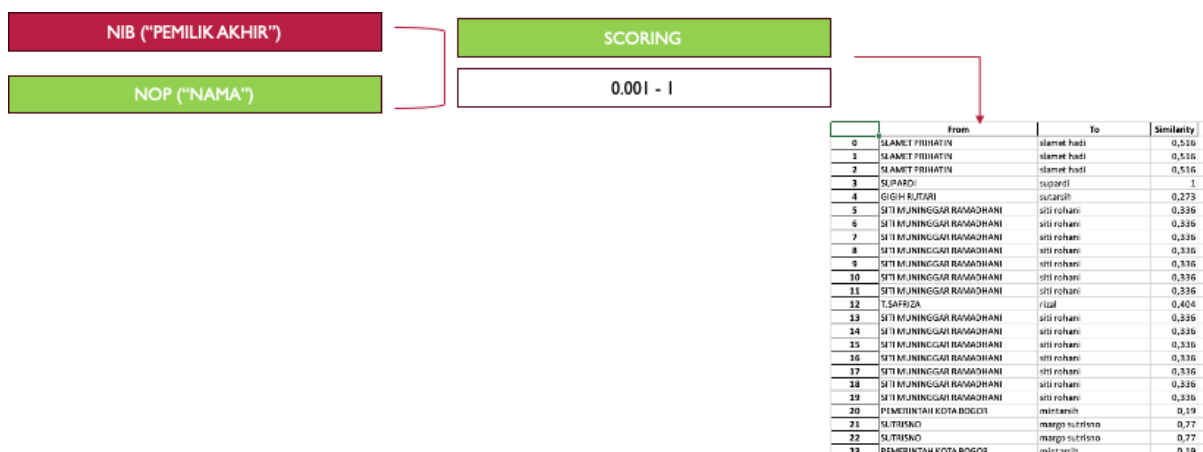
Data join NIB pertanahan yang mempunyai “Pemilik Akhir” dan data join NOP perpajakan yang mempunyai “NAMA” dijadikan *corpus similarity text*.



Gambar 9. Proses Analisis menggunakan TF-IDF

c. Cosine Similarity Text

Kesamaan kosinus adalah cara matematis untuk mengukur seberapa mirip dua kumpulan informasi. Dalam istilah yang paling sederhana membantu dalam memahami hubungan antara data “**Pemilik Akhir**” Pertanahan dengan “**Nama**” Perpajakan, bukan hanya membandingkannya berdasarkan nilai masing-masing. *Cosine Similarity* merupakan diantara metode pengukuran kesamaan terhadap mekanisme sistem temu kembali data, mengukur kesamaan antara dua dokumen atau teks. Dalam proses pengukuran kesamaan *Cosine Similarity* menghitung nilai dari sudut yang dihasilkan. Nilai yang dihasilkan dalam bentuk sudut vector antara 0 – 1, yang mana semakin mendekati 1 maka *query* dan kata memiliki kemiripan yang besar dan sedangkan lebih mendekati 0 maka memiliki kemiripan yang rendah. *Cosine similarity* dapat dihitung menggunakan persamaan.



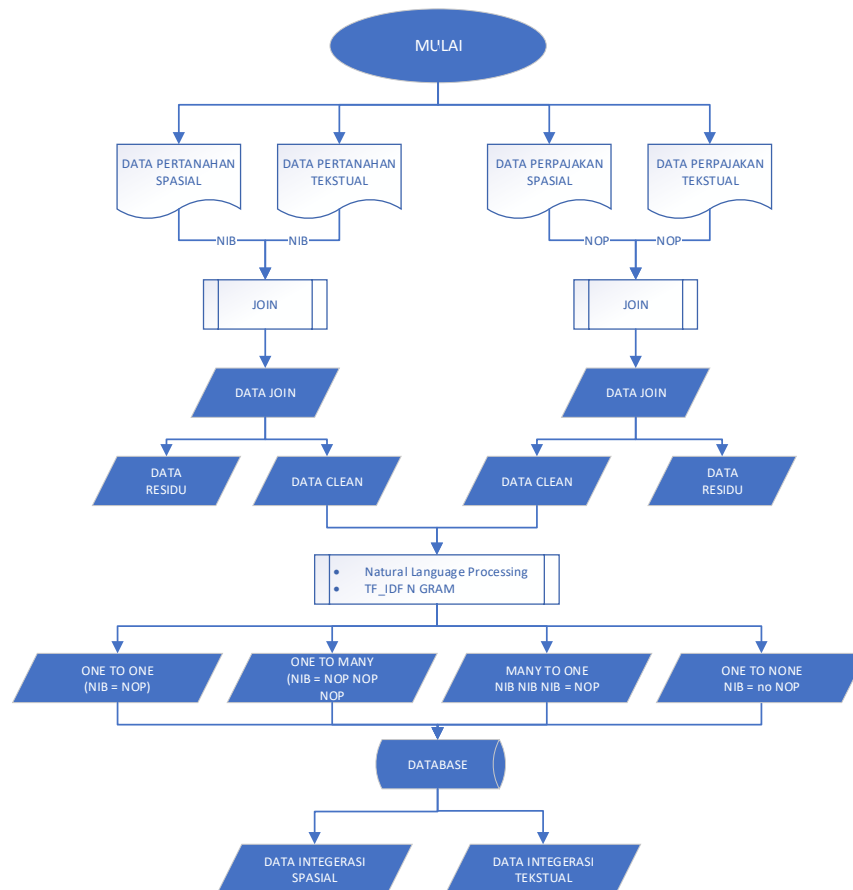
Gambar 10 Implementasi cosine similarity dalam scoring TF-IDF

Kesamaan data teks “Pemilik Akhir” akan disandingkan dengan “Nama” akan menemukan yang paling mendekati dengan metode N-GRAM dan *scoring* TF-IDF.

PEMILIK AKHIR	NAMA	Similarity
AYUNAH AKO	ayunah ako	1
SAINURLINA	sainurlina	1
SAINURLINA	sainurlina	1
SYARIFULLAH	ir. syarifullah	1
NURHASAN	nurhasan	1
NURHASANAH	nurhasan	0,904
SRI NILLA WATI	sri nilla wati	1
R. LINA YULIANA, SE	r lina yuliana	1
IZHARRY AGUSJAYA MOENZIR	iz harry agusjaya moenzir	0,905
HARRY ZULFIKAR	harry zulfikar	1
DEWI DAMAYANTI	dewi damayanti	1
SRI YUNI HARTATI	sri yuni hartati	1
ACEP ISKANDAR	acep iskandar	1
THIO CHRISTONI	thio christoni	1
PHOW HOK GWAN	phow hok gwah	0,914
LILY ISKANDAR	lily iskandar	1
NURLAILA M. ALI	nurlaila m ali	1
HARTATI SURYANINGSIH	hartati suryaningsih	1
FEBRI IRWANDY	febri irwandy	1

Gambar 11. Output similarity yang dihasilkan

Match Join Data NIB dengan NOP Metode Geopandas NLP (TF-IDF, N Gram, Consine Similarity) Berdasarkan proses *join* data spasial dan tekstual yang dilakukan sebelumnya, didapatkan data *clear* berupa *database* bidang tanah dan *database* central perpajakan. *Database* tersebut digunakan untuk mengkaitkan (*join*) atribut NOP pada NIB yang sesuai dengan menggunakan kesamaan atribut “nama” di data pertanahan dan data perpajakan. Berikut adalah alur atau proses deteksi kesesuaian PBT dan NOP secara tekstual dengan metode *Geopandas NLP* (Gambar 10).



Gambar 12. Deteksi Kesesuaian PBT dan NOP secara Tekstual dengan Metode Geopandas NLP

Pengukuran kesamaan dokumen dilakukan dengan membandingkan satu kolom atau beberapa kolom dengan data yang lainnya tentu sulit dan memakan banyak waktu jika dilakukan secara manual. Oleh sebab itu, dibangunlah sebuah aplikasi atau sistem yang dapat membandingkan data dan mengukur tingkat kesamaan dokumen secara otomatis yang dapat mempermudah memperoleh data yang diinginkan. Pengukuran tingkat kesamaan pada aplikasi dengan menggunakan algoritma TF-IDF. Algoritma TF-IDF adalah algoritma untuk menentukan bobot dari suatu term (kata) pada suatu data.

Dengan menggunakan algoritma TF-IDF untuk membandingkan dokumen-dokumen pertanahan PBT dengan NOP Perpajakan, AI yang dapat mengukur kesamaan antar dokumen. Algoritma yang dipakai dalam pengukuran kesamaan teks yaitu: tekstual *fuzzy similarity*, *TF-IDF* dan *ontology based query*. Berikut penyandingan data tet pertanahan pada kolom “Pemilik Akhir” dengan data text perpajakan pada kolom “Nama”.

	From	To	Similarity
0	nan	nan	1.000
1	PT. BANGUN PURIMULIA	yulia	0.283
2	ESIH SUKAESIH	sukarsih	0.404
3	nan	nan	1.000
4	MOH. NASIR	nasir naisin	0.513
...	...	...	...
6721	SUTISNA	sutisna	1.000
6722	SUTISNA	sutisna	1.000
6723	AGUS WAHYUDI	agus	0.514



penyamaan teks dengan *scoring* 0.2 sampai dengan 1 (0.2 – 1), kesamaan data yang paling mendekati 0.9 - 1

	From	To	Similarity
0	nan	nan	1.000
1	PT. BANGUN PURIMULIA	yulia	0.283
2	ESIH SUKAESIH	sukarsih	0.404
3	nan	nan	1.000
4	MOH. NASIR	nasir naisin	0.513
6720	SUTISNA	sutisna	1.0
6721	SUTISNA	sutisna	1.0
6722	SUTISNA	sutisna	1.0



penyamaan teks dengan *scoring* 0.2 sampai dengan 0.8 (0.2 – 0.8)

Gambar 13. Penyandingan “Pemilik Akhir” dan “Nama” dengan *scoring*

Berikut adalah hasil penyandingan data NIB dengan NOP dengan kesamaan “Pemilik Akhir” dan “Nama”:

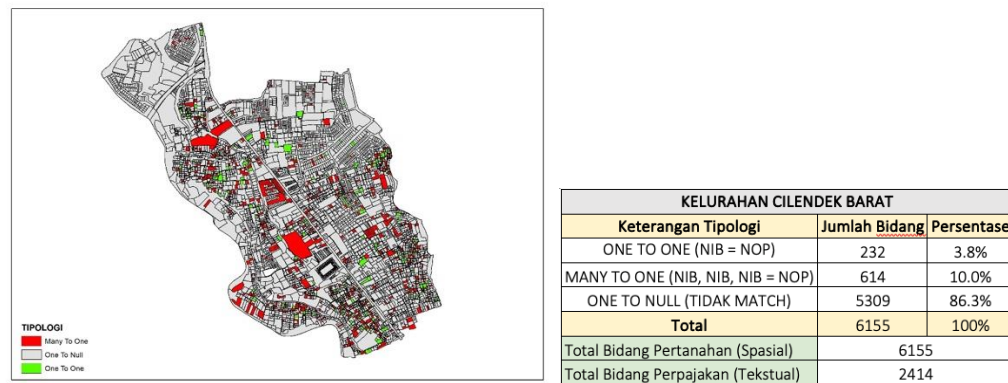
PEMILIK PERTAMA	PEMILIK AKHIR	TIPE HAK	From	To	Similarity	keyname_nop	d_nop	d_luas	geometry_y	NO	NAMA
EVI HERAWATI	EVI HERAWATI	Hak Milik	EVI HERAWATI	evi herawati	1.0	1323	3.271040e+17	50.62	POLYGON ((694890.94 9272662.56, 694886.55 9272...	1751.0	evi herawati
SUMANTA AR.	SUMANTA AR.	Hak Milik	SUMANTA AR.	sumanta. ar	1.0	2241	3.271040e+17	255.24	POLYGON ((694856.71 9272355.24, 694864.61 9272...	1921.0	sumanta. ar
HANDAYANI	HANDAYANI	Hak Milik	HANDAYANI	handayani	1.0	2215	3.271040e+17	132.84	POLYGON ((694346.04 9272468.98, 694342.07 9272...	981.0	handayani
ANDA	ANDA	Hak Milik	ANDA	anda	1.0	1270	3.271040e+17	264.01	POLYGON ((694807.99 9272731.8, 694808.48	1675.0	anda

Gambar 14. Hasil kesamaan NIB dan NOP

3.1.6. Hasil Deteksi Kesesuaian data Similarity by Nama

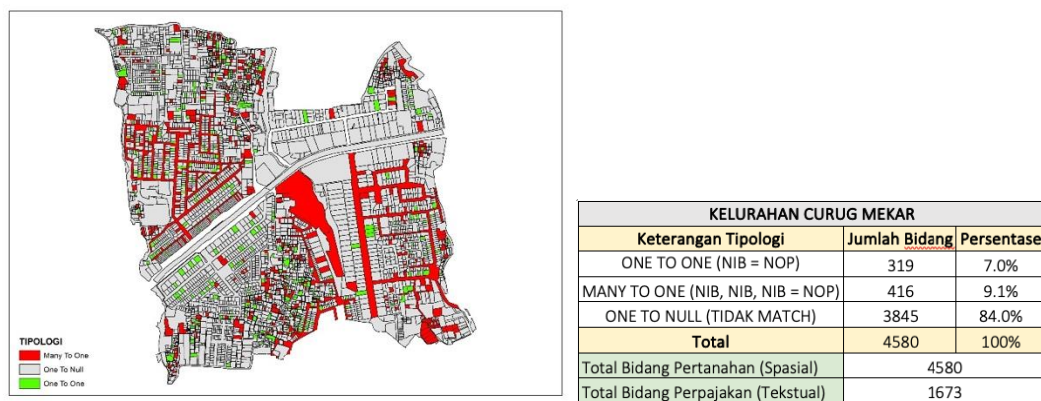
Proses TF-IDF menghasilkan database tekstual NIB dan NOP yang telah terikat (*join*) melalui persamaan nama pemilik akhir pada data pertanahan dan nama pada data perpajakan. Berdasarkan hasil tersebut, dapat diklasifikasikan data *One to One* dan *Many To One*. Berikut adalah bidang *One to One* dan *Many to One* di Kota Bogor yang dapat dilandungkan ke data spasial.

1. Kelurahan Cilendek Barat



Gambar 15. Peta Deteksi Kesesuaian PBT dan NOP di Kelurahan Cilendek Barat

2. Kelurahan Curug Mekar



Gambar 16. Peta Deteksi Kesesuaian PBT dan NOP di Kelurahan Curug Mekar

4. PENUTUP

Penerapan Natural Language Processing Similarity by Nama sangat membantu dalam pembersihan data, memberikan rekomendasi mengenai data yang perlu diperbarui dan data yang perlu dihapus. Hasil olahan data ini sangat bermanfaat untuk verifikasi lapangan guna menyelaraskan data dengan hasil verifikasi lapangan. Selain itu, hasil penelitian menunjukkan bahwa kombinasi GeoAI dan NLP dapat meningkatkan pemahaman terhadap pola spasial dalam data yang tidak terstruktur serta memberikan dukungan untuk pengambilan keputusan berbasis lokasi. Penelitian ini berkontribusi pada pengembangan teknik analisis data spasial yang lebih akurat dan dapat diterapkan di berbagai bidang, seperti perencanaan perkotaan, penanggulangan bencana, dan analisis lingkungan. Dengan demikian, integrasi GeoAI dan NLP tidak hanya meningkatkan efisiensi dalam pengolahan data, tetapi juga memperkuat kemampuan analisis spasial untuk mendukung keputusan yang lebih informatif dan tepat sasaran.

DAFTAR PUSTAKA

- [1] S. Scheider and K. F. Richter, "GeoAI," *Springer Sci. Bus. Media Deutschl. GmbH*, 2023, doi: 10.1007/s13218-022-00797-z.
- [2] L. H. A. Amri and R. A. Wijayanti, "Pemanfaatan Sistem Informasi Geospasial Online untuk Mendukung Pengambilan Keputusan Pemanfaatan Tata Ruang," *J. Teknol. Inf. dan Terap.*, vol. 6, no. 2, pp. 62–66, 2019, doi: 10.25047/jtit.v6i2.111.
- [3] H. Rumapea, "Methomika: Deteksi Kemiripan Artikel Melalui Keywords dengan Metode Fuzzy String Matching dalam Natural Language Processing," *J. Manaj. Inform. Komputerisasi Akunt.*, vol. 5, no. 1, 2021, doi: 10.46880/jmika.vol5no1.pp60-66.
- [4] Z. Chen, X. Ma, L. Wu, and Z. Xie, "An Intuitionistic Fuzzy Similarity Approach for Clustering Analysis of Polygons," *ISPRS Int. J. Geo-Information*, vol. 8, no. 2, 2019, doi: 10.3390/ijgi8020098.

- [5] Y. Heryadi and E. Irwansyah, *Deep Learning: Aplikasinya di Bidang Geospasial*. AWI Technology Press, 2020.
- [6] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Chapter 5 - Diagnosing of disease using machine learning," K. K. Singh, M. Elhoseny, A. Singh, and A. A. B. T.-M. L. and the I. of M. T. in H. Elngar, Eds., Academic Press, 2021, pp. 89–111. doi: <https://doi.org/10.1016/B978-0-12-821229-5.00003-3>.
- [7] I. Saputra and D. Rosiyadi, "Perbandingan Kinerja Algoritma K-Nearest Neighbor, Naïve Bayes Classifier dan Support Vector Machine dalam Klasifikasi Tingkah Laku Bully pada Aplikasi WhatsApp," *Fakt. Exacta*, vol. 12, no. 2, p. 101, 2019, doi: 10.30998/faktorexacta.v12i2.4181.
- [8] I. Mutia, "KAJIAN PENERAPAN E-LEARNING DALAM PROSES," *Fakt. Exacta*, vol. 6, no. 4, pp. 278–289, 2013.
- [9] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Inf. Fusion*, vol. 57, pp. 115–129, May 2020, doi: 10.1016/j.inffus.2019.12.001.
- [10] A. Maftukhah, A. Fadlil, and S. Sunardi, "Butterfly Image Classification using Convolution Neural Network with AlexNet Architecture," *J. Infotel*, vol. 16, no. 1, pp. 82–95, 2024, doi: 10.20895/infotel.v16i1.1004.