PROPOSAL

PRAKTIK KERJA LAPANGAN

PEMANFAATAN BIG DATA DAN MACHINE LEARNING UNTUK MENDUKUNG PENGUMPULAN DATA SENSUS EKONOMI 2026

Studi Kasus: Provinsi Sumatera Selatan

RISET 6

POLITEKNIK STATISTIKA STIS

JAKARTA

2024

BAB I
PENDAHULUAN

1.1 Latar Belakang

Peningkatan kebutuhan pengguna pada data yang cepat dan terkini mendorong lembaga penyedia data di berbagai negara untuk terus berinovasi demi memenuhi permintaan tersebut (BPS, 2020). Badan Pusat Statistik (BPS) berperan dalam penyediaan data statistik nasional maupun internasional untuk menghasilkan statistik yang mempunyai kebenaran akurat dan menggambarkan keadaan yang sebenarnya dalam rangka mendukung Indonesia maju (Renstra, 2019). Sebagai pemegang peranan tertinggi untuk data *official statistics*, BPS memiliki program penting yaitu menyelenggarakan tiga sensus utama setiap dekade, yaitu : Sensus Penduduk, Sensus Pertanian, dan Sensus Ekonomi (BPS, 2015). Dari ketiganya, Sensus Ekonomi (SE) bertujuan untuk menyediakan data dasar seluruh kegiatan ekonomi sebagai landasan penyusunan kebijakan dan perencanaan pembangunan nasional.

Sesuai Undang-Undang Nomor 16 Tahun 1997 tentang Statistik, Sensus Ekonomi dilaksanakan setiap sepuluh tahun sekali pada tahun yang berakhiran enam. Pelaksanaan SE 2026 menjadi kesempatan penting untuk menyediakan informasi mengenai struktur dan karakteristik usaha, ekonomi digital, dan ekonomi lingkungan (Sekretariat SE, 2026). Selain itu, SE 2026 diharapkan dapat menjawab berbagai isu strategis, seperti daya saing usaha, peta perekonomian wilayah, jumlah dan kontribusi UMKM terhadap perekonomian nasional, serta penerapan ekonomi digital dan keberlanjutan lingkungan.

Usaha Menengah Mikro Kecil (UMKM) memainkan peran krusial dalam perekonomian Indonesia, terbukti pada tahun 2023 kontribusinya mencapai 61% terhadap Produk Domestik Bruto (PDB) nasional menurut data Kementerian Koperasi dan Usaha Kecil Menengah (Kemenkop UKM). Hal ini menunjukkan bahwa UMKM yang ada di Indonesia sangat potensial untuk dikembangkan hingga dapat berkontribusi lebih besar lagi bagi perekonomian. Salah satu provinsi yang menjadikan UMK sebagai tulang punggung perekonomian lokal adalah Sumatera Selatan, dengan kontribusinya mencapai 98,99% dari total usaha non-pertanian, yang mencerminkan peran pentingnya dalam mendukung

pemerataan kesejahteraan melalui jumlahnya yang signifikan, sifatnya yang padat karya, penggunaan bahan baku lokal, dan kemampuannya menyerap tenaga kerja dalam jumlah besar (Sani et al., 2022). Untuk memperoleh gambaran yang akurat terkait potensi dan tantangan UMKM dalam menghadapi isu-isu ekonomi, termasuk digitalisasi dan keberlanjutan lingkungan, diperlukan data yang akurat, aktual dan komprehensif.

Menurut Siregar (2023), BPS harus mampu menyediakan data yang berkualitas dalam arti lengkap, akurat, tepat waktu, benar, dan relevan. Data yang akurat diperoleh melalui berbagai tahapan dalam kerangka *Generic Statistical Business Process Model* (GSBPM), sebuah model yang dikembangkan untuk mengelola proses statistik dari pengumpulan data, hingga analisis, dan publikasi hasil. Menurut UNECE (2014), pengumpulan data merupakan salah satu tahapan dalam GSBPM, yakni proses mengumpulkan segala informasi yang diperlukan dengan menggunakan berbagai metode.

Akan tetapi, tantangan besar muncul dalam penyediaan data yang responsif terhadap perubahan dan sentimen masyarakat yang dinamis. Meskipun SE telah menjadi salah satu instrumen utama pengumpulan data, metode tradisional yang digunakan belum sepenuhnya mampu menjawab kebutuhan akan data yang cepat, akurat, dan efisien. SE 2016, misalnya, menghadapi berbagai kendala seperti tingginya tingkat *under coverage* dan *underestimate*. Beberapa kesalahan yang sering ditemukan antara lain pada isian jumlah tenaga kerja, tahun beroperasi, pendapatan usaha, serta pengeluaran usaha. Bahkan, beberapa isian data ditemukan kosong atau tidak wajar (Rembangkab BPS, 2016).

Big data muncul sebagai solusi potensial untuk mengatasi tantangan tersebut. Dengan karakteristiknya yang mencakup volume data besar, kecepatan tinggi, dan variasi yang luas, *big data* menawarkan kemampuan untuk menyediakan data yang lebih relevan dan aktual. Dengan pengelolaan yang baik dan benar, big data dapat diusulkan sebagai pelengkap statistik resmi (Letouzé & Jütting, 2015). Selain itu, penelitian yang dilakukan oleh Nunes & Ashofteh (2024) menunjukkan bahwa penggunaan *big data* dalam statistik nasional memungkinkan respons yang lebih cepat terhadap perubahan dinamis di masyarakat. Namun, di Indonesia, pemanfaatan *big data* dalam konteks SE masih menghadapi kendala seperti validasi data, regulasi, privasi, keterbatasan sumber daya manusia, serta teknologi pengolahan

data (BPS, 2020). Salah satu teknik yang dapat digunakan untuk mengoptimalkan pemanfaatan *big data* adalah *machine learning*.

Machine Learning adalah ilmu yang memungkinkan komputer untuk belajar secara otomatis dari pengalaman, tanpa bergantung pada aturan yang diprogram secara eksplisit, dan menerapkan pengetahuan yang diperoleh tersebut ke situasi baru (Yung et al., 2018). *Machine learning* menawarkan potensi besar untuk meningkatkan kualitas informasi dan statistik dengan mengatasi tantangan seperti perubahan distribusi data (*concept drift*), yang dapat menyebabkan bias dalam model statistik jika tidak ditangani dengan baik (Meertens et al., 2022).

Pada era saat ini, integrasi antara *big data* dan *machine learning* menjadi alat resmi pada *National Statistics Office* dalam proses statistik (Nunes & Ashofteh, 2024). Dalam buku karya Rachedi, K., & Rachedi, F. (2021), salah satu contoh penerapan terbaik *big data* dalam *official statistics* pernah dilakukan di Belgia dengan cara membandingkan antara data sensus dan data seluler untuk melakukan estimasi kepadatan demografi. Hasilnya menunjukkan bahwa koefisien korelasi dari kepadatan data sensus dengan kepadatan data seluler sebesar 0.85.

Selain digunakan untuk *big data*, *machine learning* juga dapat dimanfaatkan untuk otomatisasi klasifikasi baku. Menurut penelitian Ahilathasan dan Pekarskaya (2020) di Norwegia, proses pengkodean *Standard Industrial Classification* (SIC) yang dilakukan secara manual, seperti pada pendaftaran lebih dari 70.000 perusahaan setiap tahun oleh Statistics Norway (SSB), sangat memakan waktu. Oleh karena itu, SSB memanfaatkan model ML yang mampu mempercepat dan menyederhanakan proses tersebut. Dengan menggunakan teknik seperti *Fasttext*, ML dapat memberikan saran kode terbaik, mempercepat pengambilan keputusan, dan menghasilkan efisiensi biaya yang signifikan dalam jangka panjang. Penelitian tersebut menunjukkan bahwa meskipun model ini hanya mencapai akurasi 22% untuk prediksi dengan tingkat kepercayaan di atas 95%, model tersebut tetap dapat memberikan lima saran kode terbaik yang membantu mempercepat proses klasifikasi. Cara ini dapat diadopsi untuk meningkatkan efisiensi waktu dan pengurangan biaya dalam pengumpulan data *official statistics* di Indonesia.

Oleh karena itu, integrasi antara *big data* dan *machine learning* dapat menjadi salah satu solusi untuk mengatasi permasalahan dalam proses pengumpulan data yang berkaitan dengan proses pemetaan wilayah sensus dan Klasifikasi Baku Lapangan Usaha Indonesia (KBLI). Pendekatan ini dapat membantu BPS dalam meningkatkan metodologi, mengidentifikasi kelemahan, dan mengoptimalkan penggunaan Big Data sebagai sumber informasi, sehingga menghasilkan data yang lebih relevan dan berdampak signifikan (Sensuse dkk., 2021).

1.2 Identifikasi dan Batasan Penelitian

Penelitian ini dibatasi pada dua aspek utama berbasis *Big Data* dan *Machine Learning*, yaitu otomatisasi Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) dan pemetaan sebaran usaha.

1.2.1 Pemetaan Sebaran Usaha Berdasarkan Big Data

Sitorus et al. (2021) menyatakan bahwa pemetaan digital atau pemetaan sebaran usaha merupakan salah satu cara untuk meningkatkan digitalisasi usaha. Pemanfaatan teknologi visualisasi data dalam pemetaan wilayah usaha memberikan keuntungan strategis dalam hal efisiensi waktu dan sumber daya. Teknologi ini memungkinkan identifikasi awal area potensial dengan lebih cepat, mengurangi biaya survei dengan memfokuskan pencacahan pada area yang telah teridentifikasi, serta mengoptimalkan alokasi sumber daya penelitian (Widodo et al., 2024).

Santoso dan Nasir (2021) menyatakan bahwa pemetaan adalah sebuah kegiatan pengelompokan sekumpulan wilayah yang berhubungan dengan posisi geografis yang mencakup daerah dan perkembangan masyarakat. Menurut Saputro (2016), pemetaan adalah sebuah proses pengorganisasian data geografis. Data tersebut akan menjadi peta visualisasi penyebaran kondisi daerah tertentu berdasarkan kategori tertentu serta merepresentasikan keadaan sebenarnya ke dalam peta dasar.

Usaha adalah suatu kegiatan ekonomi yang bertujuan menghasilkan barang/jasa untuk dijual atau ditukar untuk mendapatkan keuntungan tertentu dengan modal yang digunakan dalam usaha (Annisa, 2020). Menurut Zaharudin (2009), usaha dapat diartikan sebagai bentuk pekerjaan yang dilakukan secara konsisten dan berkelanjutan agar mendapatkan

keuntungan yang didirikan dan berkedudukan di suatu tempat. Usaha dapat dilakukan oleh perorangan atau kelompok, baik yang berbentuk badan hukum atau tidak berbentuk badan hukum.

Pemetaan wilayah usaha adalah proses pengorganisasian, identifikasi, dan visualisasi data geografis untuk menggambarkan lokasi dan sebaran usaha dalam suatu wilayah tertentu. Proses ini melibatkan pemetaan titik titik koordinat (latitude dan longitude) dari lokasi usaha untuk menunjukkan lokasi akurat dimana letak usaha tersebut. Pengelompokkan berdasarkan posisi geografis wilayahnya sehingga terbentuk peta visualisasi yang mencerminkan kondisi usaha suatu daerah.

Pada pemetaan usaha, big data dimanfaatkan sebagai sumber data. Pemanfaatan *big data* akan menghasilkan sebaran usaha yang akurat sesuai dengan kondisi lapangan yang ada. Pengumpulan *big data* dilakukan dengan metode *web scraping* pada Google maps. *Website* Google maps merupakan layanan pemetaan yang dikembangkan oleh Google yang mudah dan familiar digunakan oleh pengguna. Pengguna dapat menambahkan tempat usaha yang dimiliki ke Google Maps dengan mudah dan cepat. Kemudahan menambahkan tempat oleh pengguna memungkinkan daftar lokasi usaha yang tersedia banyak dan cakupannya luas. Variabel yang dikumpulkan melalui metode *web scraping* antara lain nama usaha, alamat, kategori usaha, jam buka, titik koordinat lokasi usaha (*latitude* dan *longitude*), nomor telepon.

Dengan memanfaatkan Google Maps, didapatkan beragam informasi tentang lokasi usaha yang dapat dimanfaatkan untuk analisis ekonomi dan sosial. Di sisi lain, data resmi dari lembaga statistik, seperti Badan Pusat Statistik (BPS), menyediakan informasi terstandar mengenai usaha yang dapat digunakan untuk mendukung validasi data dari sumber tidak resmi seperti Google Maps. Integrasi kedua sumber data ini berpotensi meningkatkan efisiensi dan akurasi dalam *pre-list* usaha di berbagai wilayah, termasuk Provinsi Sumatera Selatan.

Pre-list usaha adalah daftar awal lokasi usaha yang dapat menjadi dasar survei atau pengumpulan data lebih lanjut. Untuk mengembangkan *pre-list* usaha, data hasil scraping Google Maps yang sudah didapat digunakan sebagai sumber data awal yang mencakup informasi tentang lokasi usaha secara luas dan mendetail. Kemudian, data *Official Statistic*

berperan untuk memastikan validitas dan reliabilitas data yang diperoleh dari Google Maps, sehingga hasil pemetaan lebih akurat.

Tantangan dari pemanfaatan data hasil scraping dari Google Maps adalah data tersebut seringkali mencakup informasi yang tidak relevan atau duplikat. Data tidak relevan yang dimaksud adalah data lokasi non-usaha yang ikut terambil dari proses scraping. Sehingga, diperlukan metodologi khusus untuk membersihkan dan mengklasifikasikan data secara akurat. Proses pembersihan data dilakukan untuk menghilangkan data yang duplikat, sedangkan klasifikasi data dilakukan untuk memisahkan lokasi usaha dari lokasi non-usaha. Dengan dua proses tersebut, diharapkan data yang dihasilkan hanya mencakup data yang relevan yaitu data mengenai lokasi usaha di Provinsi Sumatera Selatan.

Dalam penelitian ini, perlu untuk melakukan pembatasan terkait wilayah penelitian. Penelitian ini berfokus pada pemetaan rumah tangga usaha di Provinsi Sumatera Selatan.

1.2.2 Pemetaan Sebaran Usaha Berdasarkan Hasil Listing

Sensus Ekonomi 2026 yang dilakukan oleh BPS merupakan salah satu instrumen utama untuk memperoleh data yang komprehensif terkait dengan kondisi ekonomi di berbagai sektor. Salah satu elemen vital dalam sensus ini adalah Rumah Tangga Usaha (RTU), yang menggambarkan struktur ekonomi mikro di tingkat rumah tangga. Rumah Tangga Usaha merujuk pada unit rumah tangga yang perekonomiannya ditopang dari hasil keuntungan berusaha secara mandiri (Setiawan & Wibisono, 2021).

Dalam penelitian ini, pre-list usaha yang telah dihasilkan melalui integrasi data big data dari web scraping Google Maps dan data resmi dari lembaga statistik, menjadi dasar bagi kegiatan listing manual yang dilakukan di lapangan. Tahap ini bertujuan untuk memastikan keakuratan data awal dan melengkapi informasi usaha yang mungkin belum tercakup dalam pre-list. Data hasil scraping, meskipun kaya akan cakupan informasi, seringkali memiliki kekurangan seperti informasi yang kurang relevan, usaha yang tidak aktif, atau perubahan kategori usaha. Listing manual diperlukan untuk menyempurnakan dataset yang digunakan.

Listing manual adalah proses pengumpulan data secara langsung melalui kunjungan lapangan untuk mencatat lokasi usaha yang ada pada setiap rukun tetangga unit (RTU). Menurut Suharyanto (2020), listing adalah proses mendata secara rinci dan sistematis unit

observasi tertentu dengan menggunakan daftar pencacahan. Daftar ini berfungsi untuk memverifikasi data awal dan menambahkan informasi baru yang mungkin belum tercakup dalam pre-list. Tahap listing manual memegang peranan penting dalam memastikan data usaha yang terkumpul sesuai dengan kondisi lapangan. Terkadang, data hasil web scraping maupun data resmi masih mengandung kekurangan seperti informasi yang kurang relevan, lokasi usaha yang tidak lagi aktif, atau adanya perubahan kategori usaha. Menurut Nasution dan Wahyuni (2019), pelaksanaan listing memerlukan pendekatan berbasis wilayah untuk memastikan seluruh area tercakup secara merata. Setiap RTU menjadi unit kerja yang didata, sehingga menghasilkan data mikro yang lebih akurat dan dapat dipetakan secara detail.

Setelah kegiatan listing selesai, data hasil listing akan dipetakan kembali menggunakan perangkat lunak pemetaan geografis. Titik-titik koordinat hasil listing akan digabungkan dengan data pre-list yang telah diperbarui. Proses ini memungkinkan visualisasi penyebaran usaha di suatu wilayah yang lebih akurat dan mendetail. Penelitian oleh Handayani et al. (2022) menunjukkan bahwa pemanfaatan hasil listing manual dalam pemetaan sebaran usaha memberikan informasi lebih rinci mengenai pola penyebaran usaha pada tingkat mikro seperti kelurahan atau RT.

1.2.3 Otomatisasi KBLI

Menurut Peraturan BPS Nomor 2 Tahun 2020, KBLI adalah mengklasifikasikan aktivitas/kegiatan ekonomi Indonesia yang menghasilkan produk/output baik berupa barang maupun jasa berdasarkan lapangan usaha yang digunakan sebagai acuan standar dan alat koordinasi, integrasi, serta sinkronisasi penyelenggaraan statistik. KBLI dibangun berdasarkan kerangka konseptual aktivitas produksi yang mengelompokkan unit produksi menjadi industri yang lebih rinci berdasarkan kesamaan aktivitas ekonomi, dengan mempertimbangkan input, proses dan teknologi produksi, karakteristik output yang dihasilkan, dan penggunaan output. Pada prinsipnya, kegiatan yang memiliki proses yang sama dan menggunakan teknologi yang sama dalam memproduksi barang atau jasa, dikelompokkan bersama dalam satu kode KBLI (BPS dalam KBLI , 2020).

Dalam pelaksanaan SE, proses pengklasifikasi lapangan usaha masih dilakukan secara manual berdasarkan jawaban dari pertanyaan terbuka. Setiap rincian pertanyaan terbuka

harus diisi sejelas mungkin oleh petugas pencacah agar pengawas mampu menangkap informasi lapangan usaha dan jenis pekerjaan yang tepat. Proses ini memerlukan waktu dan tenaga yang besar. Selain itu, pemahaman petugas yang sering kali dapat menimbulkan masalah seperti potensi perbedaan persepsi dan interpretasi antar individu, sangat berisiko menghasilkan inkonsistensi dalam hasil klasifikasi (Dwicahayaniawan et al., 2024). Hal ini menyebabkan diperlukan adanya sistem otomatisasi yang mampu mengklasifikasikan usaha secara cepat dan akurat berdasarkan kode KBLI (Suadaa et al., 2023).

1.3 Rumusan Masalah

a. Bagaimana cara membuat *pre-list* usaha menggunakan data Google Maps dan Official Statistik?
b. Bagaimana cara membedakan lokasi usaha dan bukan usaha berdasarkan hasil scraping google maps?
c. Bagaimana cara menyediakan dataset rumah tangga usaha di Provinsi Sumatera Selatan?
d. Bagaimana cara membuat visualisasi dari lokasi rumah tangga usaha di Sumatera Selatan?
e. Bagaimana menyediakan dataset yang berisi deskripsi pekerjaan dan aktivitas ekonomi dari responden di Provinsi Sumatera Selatan yang teranotasi sesuai dengan kategori KBLI secara akurat?
f. Bagaimana membangun model klasifikasi yang akurat untuk memberikan rekomendasi kode Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) menggunakan teknik *machine learning* dan *natural language processing* (NLP)?

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan dari penelitian ini yaitu:

a. Membuat visualisasi peta *pre-list* usaha menggunakan data Google Maps dan Official Statistics.
b. Membuat pemodelan usaha dan bukan usaha berdasarkan hasil scraping google maps.

c. Membuat dataset rumah tangga usaha di provinsi Sumatera Selatan berdasarkan latitude dan longitude yang dikumpulkan hasil listing.
d. Membuat visualisasi peta lokasi rumah tangga usaha di provinsi Sumatera Selatan berdasarkan hasil listing.
e. Menyediakan dataset yang berisi deskripsi pekerjaan dan aktivitas ekonomi dari responden di Provinsi Sumatera Selatan yang telah dianotasi sesuai dengan kategori KBLI.
f. Membangun model klasifikasi yang akurat untuk memberikan rekomendasi kode Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) menggunakan teknik *machine learning* dan *natural language processing* (NLP)

1.5 Manfaat Penelitian

Berdasarkan tujuan tersebut, penelitian ini diharapkan dapat memberikan manfaat kepada beberapa pihak antara lain:

a. Visualisasi peta prelist usaha menggunakan data Google Maps dan Official Statistics dapat menjadi pertimbangan dalam mengalokasikan petugas sebelum melakukan listing sekaligus memandu petugas pencacah untuk melakukan listing.
b. Membantu BPS dalam menyediakan kerangka sampel listing rumah tangga usaha untuk menunjang Sensus Ekonomi 2026 di Provinsi Sumatera Selatan.
c. Dataset yang berisi deskripsi pekerjaan dan aktivitas ekonomi dari responden di Provinsi Sumatera Selatan yang telah dianotasi sesuai dengan kategori KBLI.
d. Model klasifikasi untuk memberikan rekomendasi kode Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) menggunakan teknik *machine learning* dan *natural language processing* (NLP).

BAB II
TINJAUAN PUSTAKA

2.1 Pemetaan Wilayah Sebaran Usaha

2.1.1 Big Data

Semakin berkembangnya zaman, maka semakin banyak pula informasi yang telah terkumpul dari waktu ke waktu. Informasi-informasi yang telah terkumpul cukup kompleks, beragam, dan jumlahnya juga besar. Informasi tersebut dikelola dalam bentuk data besar yang saat ini lebih dikenal dengan istilah *big data*. *Big Data* merupakan sekumpulan data yang yang sangat besar, kompleks, dan cepat tumbuh, yang sulit untuk dikelola, diproses, dan dianalisis menggunakan metode tradisional (Kurniawan et al., 2024). Fenomena *Big Data* muncul karena pertumbuhan data yang sangat pesat dari berbagai sumber, seperti media sosial, sensor, transaksi bisnis, dan perangkat *Internet of Things* (IoT).

Secara umum, big data didefinisikan sebagai sekumpulan data yang memiliki karakteristik 3V, yaitu Volume, Velocity, dan Variety yang tinggi (UNECE, 2013).

a. Volume

   Volume yang tinggi merujuk pada sangat banyaknya baris data dengan atribut/kolo/variabel yang banyak pula, sehingga memerlukan kapasitas yang besar untuk penyimpanan datanya.

b. Velocity

   Velocity yang tinggi merujuk pada cepatnya data dihasilkan, sehingga perubahan yang terjadi membutuhkan kecepatan pula dalam hal pengumpulan, penyimpanan, pengolahan, serta analisisnya. Adanya keterlambatan dalam melakukan analisis akan mengakibatkan hasil analisis kehilangan momentum karena data terbaru tersedia dengan sangat cepat pula.

c. Variety

Variety yang tinggi merujuk pada varians atau ragam data yang tinggi. Keragaman data dapat ditinjau dari keragaman format data yang dihasilkan maupun isi atau *content*-nya yang sangat beragam.

2.1.2 Google Maps

Google maps merupakan sebuah peta digital yang dapat diakses secara gratis oleh penggunanya. Website ini menyediakan peta untuk permukaan seluruh dunia dalam bentuk satelit maupun peta jalan. Selain itu, google maps juga menawarkan pencarian suatu tempat dan rute perjalanan (Mahdia & Noviyanto, 2013). Google juga menyediakan Application Programming Interface (API) untuk memudahkan pengguna dalam memanfaatkan google maps dalam mengembangkan aplikasi. Google maps API merupakan kumpulan dari kelas-kelas pada JavaScript sehingga dapat diakses melalui JavaScript dengan terlebih dahulu melakukan pendaftaran API Key.

2.1.3 Geocoding

Geocoding, disebut juga *forward geocoding,* merupakan suatu proses untuk mendapatkan informasi koordinat geografis (lintang dan bujur) dari suatu alamat atau nama tempat (Cambon et al., 2021). *Geocoding* sudah marak dilakukan sebagai salah satu proses dalam tahapan pemrosesan data untuk melakukan analisis spasial. Pada kasus pemetaan usaha, konversi data ini bertujuan mendapatkan koordinat dari suatu data usaha yang hanya memiliki informasi alamat, sehingga usaha tersebut dapat dimasukkan ke dalam peta yang akan dibuat. Proses *geocoding* dimulai dari penerjemahan suatu entri alamat, lalu mencari alamat yang ada di dalam data referensi, kemudian memberikan kandidat terbaik sebagai fitur titik pada peta.

2.1.4 Web Scraping

Web Scraping adalah teknik untuk mengekstraksi data dari suatu halaman web dan menyimpannya ke dalam suatu basis data untuk dianalisis lebih lanjut. (B Zhao,

2022). Data tersebut berasal dari halaman web yang dibuat menggunakan bahasa *mark-up* seperti HTML atau XHTML. Setelah diambil, dokumen ini dianalisis untuk memperoleh wawasan baru yang bermanfaat. Proses ini biasanya melibatkan perangkat lunak atau robot yang nantinya mengunjungi situs web yang diperlukan untuk mengekstraksi data yang relevan. (Glez-Peña et al., 2013). Selain itu, proses pengambilan data juga bisa menggunakan HTTP yang dapat diproses secara manual oleh pengguna atau otomatis oleh bot atau web crawler.

2.1.5 Algoritma Jaro-Winkler

Yulianingsih (2017) menyatakan bahwa Jaro-Winkler merupakan varian dari metrik Jaro Distance yang umum digunakan untuk mendeteksi duplikasi data. Algoritma ini dioptimalkan untuk string pendek, yang mana skor jarak yang lebih tinggi menunjukkan tingkat kesamaan yang lebih besar antara dua string. Skor 0 menandakan tidak ada kesamaan, sedangkan skor 1 menunjukkan kesamaan sempurna. Pada penelitian sebelumnya, algoritma ini digunakan untuk mendeteksi plagiarisme pada dokumen teks dengan akurasi lebih tinggi, dan proses yang lebih cepat daripada algoritma *Levenshtein*

Alur kerja dari algoritma ini yaitu, (1) menghitung panjang string, (2) menghitung jumlah karakter yang ada pada kedua string, dan (3) menghitung jumlah karakter yang tertukar. Kemudian dikembangkan untuk menghasilkan rumus penghitungan jarak (*dj*) antara string *s1* dan *s2* seperti berikut.

$$dj = \frac{1}{3} \times \left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right)$$

di mana:

dj = jarak Jaro

s1 = panjang string pertama

s2 = panjang string kedua

m = jumlah karakter yang ada pada kedua string

t = jumlah karakter yang tertukar

Selanjutnya, hitung jarak Jaro-Winkler dengan rumus berikut:

$$dw = dj + (l.p - l.p.dj),$$

di mana:
dw = jarak Jaro-Winkler
l = panjang *prefix* (panjang karakter yang sama dari kedua string sampai ditemukan yang tidak sama dengan nilai maksimal 4)
p = konstanta *scaling factor* bernilai 0,1

Pada penelitian ini, algoritma Jaro-Winkler digunakan untuk mendukung integrasi data hasil *scraping* dan direktori BPS dengan mencocokkan string usaha yang sama, menyimpan salah satunya, dan menambahkan string usaha yang hanya ada di salah satu sumber untuk melengkapi data.

2.1.6 Algoritma Jarak Haversine

Menurut Farid dan Yunus (2017), algoritma Haversine digunakan untuk menghitung jarak antara dua titik di permukaan bumi berdasarkan panjang garis lurus antara dua titik tanpa mengabaikan kelengkungan yang dimiliki bumi. Formula Haversine ini digunakan untuk menghitung jarak antara dua titik di permukaan bumi, dengan menghitung jarak berdasarkan perbedaan garis bujur dan garis lintang (Chopde dan Nichat, 2013). Input yang digunakan untuk pengukuran dua titik koordinat yaitu *latitude* dan *longitude*. Formula yang digunakan sebagai berikut.

$$d = 2cos(lat2)\ arcsin\left(\sqrt{sin^2\left(\frac{lat2-lat1}{2}\right)cos(lat1)}\right)sin^2\left(\frac{lon2-lon1}{2}\right)$$

di mana
d = jarak havershine
lat1 = koordinat lintang (latitude) lokasi pertama
lon1 = koordinat bujur (longitude) lokasi pertama

lat2 = koordinat lintang (latitude) lokasi kedua

lon2 = koordinat bujur (longitude) lokasi kedua

Algoritma Haversine dihitung menggunakan pustaka *math* dalam Python. Hasil dari penghitungan ini berupa skor jarak Haversine yang bertipe numerik.

2.1.7 Peta Choropleth

Peta Choropleth adalah teknik visualisasi yang sangat berguna dalam analisis data spasial secara eksploratif (Andy Hermawan et al., 2024). Teknik ini memetakan data dalam bentuk gradasi warna untuk menunjukkan perbedaan nilai di berbagai wilayah, sehingga memudahkan pemahaman tentang distribusi geografis dari berbagai variabel. Dengan visual yang intuitif, peta ini membantu mengidentifikasi pola spasial, tren, dan anomali dalam data. Peta Choropleth sangat relevan di berbagai bidang seperti demografi, ekonomi, lingkungan, dan kesehatan masyarakat untuk menghasilkan wawasan berharga. Sejarah pembentukan peta ini juga dijelaskan oleh Korycka-Skorupa & Pasławski (2017), bahwa peta-peta choropleth awalnya digunakan untuk menggambarkan situasi sosial dan demografis, serta tema 'statistik moral'. Saat ini, peta choropleth dapat dengan mudah dibuat menggunakan software bantuan seperti ArcGIS (Cromley, 2009).

2.1.8 Peta Titik (Point Map)

Point map adalah salah satu metode visualisasi data pada peta yang digunakan untuk memetakan titik-titik koordinat (latitude dan longitude). Teknik ini cocok digunakan ketika peneliti ingin melihat tampilan informasi geografis secara terperinci. Berbeda dengan visualisasi data agregat, point map hanya menyajikan lokasi titik tertentu dari data.

Point map cocok untuk menunjukkan penyebaran dan kepadatan suatu wilayah, seperti persebaran UMKM (Usaha Mikro, Kecil, dan Menengah). Dengan menggunakan poin map, peneliti dapat melihat lokasi akurat UMKM berdasarkan latitude dan longitude, sekaligus memahami konsentrasi UMKM di suatu wilayah, apakah padat atau jarang berdasarkan titik koordinatnya (Huck et al., 2015).

Selain itu, point map juga sangat berguna dalam memvisualisasikan data spasial yang membutuhkan presisi tinggi, karena setiap titik pada peta mencerminkan posisi geografi yang tepat sesuai dengan latitude dan longitude (Xiao et al., 2020).

2.2 Otomatisasi KBLI

2.2.1 Klasifikasi Baku Lapangan Usaha Indonesia (KBLI)

Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) adalah standar yang digunakan untuk mengelompokkan semua kegiatan ekonomi di Indonesia berdasarkan jenis aktivitas usaha. KBLI disusun oleh Badan Pusat Statistik (BPS) dengan tujuan untuk menyediakan kerangka klasifikasi aktivitas ekonomi yang terstruktur dan terstandarisasi secara nasional dan internasional. KBLI digunakan dalam penyusunan statistik ekonomi, dasar perencanaan kebijakan, serta evaluasi dan perizinan usaha.

KBLI telah mengalami beberapa kali revisi untuk mengikuti perkembangan ekonomi dan munculnya bidang usaha baru yang tidak tercakup oleh versi sebelumnya. Klasifikasi lapangan usaha di Indonesia dimulai dengan *Klasifikasi Lapangan Usaha Indonesia* (KLUI) tahun 1977 yang mengacu pada ISIC Rev. 2 (1968), kemudian mengalami revisi menjadi KLUI 1983, KLUI 1997, KBLI 2000, dan seterusnya. KBLI terbaru yang digunakan saat ini adalah KBLI 2020, yang merupakan penyempurnaan dari KBLI 2015. KBLI 2020 tetap mengacu pada *International Standard Industrial Classification of All Economic Activities* (ISIC) Rev. 4 hingga level empat digit, yang merupakan standar internasional yang digunakan untuk mengelompokkan kegiatan ekonomi di seluruh dunia. Pada level lima digit, KBLI 2020 menyesuaikan dengan kondisi ekonomi di Indonesia, mengakomodasi kebutuhan klasifikasi yang lebih rinci dan spesifik untuk kegiatan ekonomi yang ada di Indonesia.

KBLI memiliki struktur hierarkis yang terdiri dari beberapa tingkat klasifikasi, yaitu sebagai berikut.

1. Kategori : Level tertinggi dalam hierarki KBLI, terdiri dari 21 kategori yang diberi kode satu huruf alfabet (A sampai U). Kategori ini mengelompokkan

aktivitas ekonomi ke dalam sektor-sektor utama seperti pertanian, manufaktur, perdagangan, dan jasa.

2. Golongan Pokok : Setiap kategori diuraikan menjadi satu atau beberapa golongan pokok yang memiliki kode dua digit angka. Golongan pokok adalah pengelompokan yang lebih rinci dari sektor ekonomi.
3. Golongan : Golongan pokok selanjutnya dipecah menjadi golongan yang memiliki tiga digit angka, di mana dua digit pertama mewakili golongan pokok yang terkait dan satu digit terakhir mewakili aktivitas ekonomi dalam golongan tersebut.
4. Sub Golongan : Subgolongan terdiri dari empat digit angka yang menggambarkan aktivitas ekonomi spesifik yang tercakup dalam golongan tertentu. Ini adalah rincian lebih lanjut dari golongan.
5. Kelompok : Ini adalah level terendah dari struktur KBLI, yang memiliki lima digit angka. Kelompok digunakan untuk mengelompokkan kegiatan ekonomi yang lebih homogen sesuai dengan jenis aktivitas yang dilakukan.

KBLI digunakan untuk memastikan bahwa semua data statistik ekonomi yang dikumpulkan oleh berbagai instansi di Indonesia dapat dibandingkan secara nasional dan internasional. Selain itu, KBLI juga digunakan sebagai acuan dalam penyusunan kebijakan dan perizinan usaha. Misalnya, melalui *Online Single Submission* (OSS), sistem perizinan usaha di Indonesia, setiap bidang usaha diwajibkan untuk merujuk pada kode KBLI yang relevan. Pengklasifikasian aktivitas ekonomi dalam KBLI mengikuti prinsip-prinsip sebagai berikut.

a. Keseragaman Proses Produksi: Aktivitas ekonomi yang menggunakan proses dan teknologi produksi yang sama dikelompokkan dalam satu kode KBLI.
b. Fungsi Ekonomi: Klasifikasi didasarkan pada fungsi pelaku ekonomi dalam menggunakan input seperti tenaga kerja, modal, dan bahan baku untuk menghasilkan output berupa barang atau jasa.

2.2.2 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan (*artificial intelligence*) yang berfokus pada interaksi antara komputer dan bahasa

manusia alami. Tujuan utama dari NLP adalah memungkinkan komputer untuk memahami, memproses, dan menghasilkan bahasa yang digunakan manusia, baik dalam bentuk teks maupun suara (Jurafsky & Martin, 2009). NLP menggabungkan berbagai teknik dari linguistik, ilmu komputer, dan pembelajaran mesin untuk mengekstraksi makna, struktur, dan konteks bahasa manusia, sehingga memungkinkan aplikasi seperti pengenalan suara, analisis teks, dan terjemahan otomatis.

Secara umum, NLP terbagi dalam dua cabang besar, yaitu *natural language understanding* (NLU) dan *natural language generation* (NLG). NLU berfokus pada bagaimana komputer memahami makna dari bahasa manusia, sementara NLG berkaitan dengan bagaimana komputer dapat menghasilkan teks atau ucapan yang menyerupai bahasa alami manusia. Beberapa komponen penting dalam NLP antara lain sebagai berikut.

1. *Tokenization*: Proses memecah teks menjadi unit-unit yang lebih kecil seperti kata atau frasa.
2. *Part-of-Speech Tagging* (POS): Identifikasi dan penandaan kelas kata (seperti kata benda, kata kerja, dsb.) dalam kalimat.
3. *Named Entity Recognition* (NER): Identifikasi entitas khusus dalam teks, seperti nama orang, tempat, atau organisasi.
4. *Sentiment Analysis*: Proses mengidentifikasi emosi atau sentimen dalam teks, apakah positif, negatif, atau netral.
5. *Machine Translation*: Proses penerjemahan otomatis teks dari satu bahasa ke bahasa lainnya.

Teknik-teknik NLP banyak menggunakan model statistik dan pembelajaran mesin, termasuk penggunaan algoritma seperti *naive bayes*, *support vector machines* (SVM), dan *deep learning* dengan arsitektur seperti *recurrent neural networks* (RNN) dan *transformers*. *Transformer*, dengan model seperti *Bidirectional Encoder Representations from Transformers* (BERT) dan GPT (*Generative Pretrained Transformer*), telah meningkatkan kemampuan NLP secara signifikan dalam beberapa tahun terakhir.

2.2.3 *Machine Learning*

Machine Learning (ML) untuk klasifikasi teks adalah bidang yang menggabungkan teknik pembelajaran mesin dengan pemrosesan bahasa alami (*Natural Language Processing*) untuk mengklasifikasikan teks berdasarkan label atau kategori tertentu. Tujuan dari klasifikasi teks adalah mengidentifikasi dan mengelompokkan teks ke dalam kelas-kelas yang telah ditentukan berdasarkan pola yang dipelajari dari data (Manning, Raghavan, & Schütze, 2008). Setelah teks direpresentasikan, model *machine learning* dapat dilatih untuk melakukan klasifikasi. Beberapa algoritma populer untuk klasifikasi teks adalah sebagai berikut.

1. *Naive Bayes* : Algoritma probabilistik sederhana yang sangat cocok untuk klasifikasi teks (Manning et al., 2008).
2. *Support Vector Machines* (SVM): Algoritma yang menemukan *hyperplane* terbaik untuk memisahkan kelas dalam ruang fitur.
3. *Decision Trees* dan *Random Forests* : Algoritma berbasis pohon keputusan yang membagi data ke dalam subset untuk membuat keputusan.
4. *Deep Learning*: Dengan perkembangan jaringan saraf tiruan, pendekatan seperti *Recurrent Neural Networks* (RNN) dan *Transformers* telah digunakan untuk klasifikasi teks dengan hasil yang sangat baik.

Penggunaan *machine learning* untuk klasifikasi teks memiliki banyak aplikasi praktis, seperti pengelompokan dokumen berita, klasifikasi email spam, analisis sentimen, hingga klasifikasi topik dalam data media sosial. Model berbasis *machine learning* secara bertahap meningkatkan akurasi dan efisiensi dalam tugas-tugas ini dengan belajar dari data yang lebih besar dan teknik yang lebih canggih.

Menurut Manning, Raghavan, dan Schütze (2008), klasifikasi teks adalah salah satu tugas penting dalam *information retrieval* dan memiliki relevansi tinggi di berbagai domain, mulai dari sistem pencarian hingga analisis konten.

2.2.4 *Supervised Learning*

Supervised learning merupakan metode pembelajaran mesin yang menggunakan dataset berlabel sebagai dasar pelatihan model, di mana setiap input

dalam dataset memiliki pasangan label yang diketahui (Pan & Yang, 2010). Dengan memanfaatkan data yang telah dianotasi, algoritma supervised learning dilatih untuk mengenali pola dan hubungan antara input dan output agar dapat membuat prediksi yang akurat pada data baru yang belum pernah dilihat. Teknik ini telah menjadi landasan bagi berbagai aplikasi di bidang klasifikasi teks, prediksi harga, dan pengenalan wajah (Alzubaidi et al., 2021).

Beberapa algoritma populer dalam supervised learning antara lain adalah regresi linier, *Support Vector Machine* (SVM), dan *Random Forest*. Algoritma ini masing-masing memiliki cara kerja yang berbeda dalam mempelajari hubungan antara fitur dan label, tetapi tujuan akhirnya adalah menghasilkan model yang mampu memprediksi label dengan akurasi yang tinggi (Goodfellow et al., 2016).

2.2.5 Transfer Learning

Transfer learning merupakan metode dalam pembelajaran mesin yang memungkinkan pemanfaatan model yang telah dilatih pada satu tugas untuk diterapkan pada tugas lain yang serupa, terutama saat data yang tersedia pada tugas baru terbatas (Pan & Yang, 2010). Konsep ini sangat berguna di bidang pemrosesan bahasa alami (NLP) dan visi komputer, di mana model yang telah dilatih pada dataset skala besar, seperti BERT untuk NLP atau ResNet untuk visi komputer, dapat digunakan sebagai dasar untuk menyelesaikan tugas yang lebih spesifik (Goodfellow et al., 2016).

Dalam transfer learning, model awal dilatih pada domain yang lebih umum dan luas, kemudian di-*fine-tune* atau disesuaikan dengan data baru yang lebih spesifik terhadap tugas yang ingin diselesaikan. Pendekatan ini telah terbukti mampu meningkatkan akurasi model sekaligus mengurangi waktu pelatihan, karena model telah mengakumulasi pengetahuan dasar dari dataset awal yang digunakan (Alzubaidi et al., 2021).

Transfer learning semakin berkembang dan diterapkan di berbagai bidang, terutama dalam konteks aplikasi yang membutuhkan akurasi tinggi namun data pelatihan yang terbatas. Sebagai contoh, dalam NLP, model seperti BERT atau GPT-3 yang dilatih pada corpus bahasa yang besar dapat dengan mudah disesuaikan untuk tugas-tugas spesifik seperti klasifikasi teks atau analisis sentimen (Pan & Yang, 2010).

2.2.6 *Text Preprocessing*

Text preprocessing adalah tahap awal yang sangat penting dalam pemrosesan teks, terutama dalam aplikasi *machine learning* dan *natural language processing* (NLP). Tahap ini bertujuan untuk mempersiapkan teks mentah agar dapat digunakan dalam algoritma pembelajaran mesin dengan merapikan data, mengurangi kompleksitas, dan mengubahnya menjadi representasi yang dapat dipahami oleh model. *Text preprocessing* sangat mempengaruhi performa dan akurasi model yang akan dibangun (Manning, Raghavan, & Schütze, 2008). Beberapa teknik utama dalam *text preprocessing* adalah sebagai berikut.

1. *Tokenization* : Proses memecah teks menjadi unit-unit yang lebih kecil seperti kata, frasa, atau simbol. Tokenisasi membantu memisahkan teks ke dalam unit yang lebih dapat diolah, seperti memecah kalimat menjadi daftar kata-kata.
2. *Stopword Removal* : *Stopword* adalah kata-kata umum yang sering tidak memiliki makna kontekstual signifikan dalam proses pemodelan. Penghapusan *stopword* bertujuan untuk mengurangi dimensi teks dan fokus pada kata-kata yang lebih bermakna.
3. *Stemming* dan *Lemmatization* : *Stemming* adalah proses mengurangi kata ke bentuk dasarnya dengan menghapus akhiran (*suffix*) kata. *Lemmatization* lebih canggih daripada *stemming* karena mempertimbangkan bentuk kata yang benar secara linguistik, bukan sekadar menghapus akhiran (Jurafsky & Martin, 2009).
4. *Lowercasing* : Mengubah semua karakter dalam teks menjadi huruf kecil bertujuan untuk menghindari perbedaan dalam representasi kata yang sama.
5. *Punctuation Removal* : Tanda baca sering kali dihapus dari teks karena tidak memiliki makna khusus dalam analisis teks.

6. TF-IDF (*Term Frequency-Inverse Document Frequency*): Teknik ini digunakan untuk menghitung pentingnya suatu kata dalam dokumen tertentu relatif terhadap seluruh dokumen dalam korpus. Kata yang sering muncul dalam satu dokumen tetapi jarang muncul di dokumen lain dianggap lebih penting. TF-IDF sangat berguna dalam merepresentasikan kata-kata dalam model numerik untuk pembelajaran mesin.
7. *Vectorization*: Setelah *preprocessing*, teks harus direpresentasikan dalam bentuk numerik agar dapat diproses oleh algoritma pembelajaran mesin. Beberapa teknik populer untuk merepresentasikan teks adalah *Bag of Words* (BoW) dan *Word Embeddings* seperti Word2Vec atau GloVe.

2.2.7 *Text Classification*

Text classification adalah proses pengelompokan teks ke dalam satu atau lebih kategori yang telah ditentukan sebelumnya. Kategori ini bisa berupa label sederhana seperti "positif" atau "negatif" dalam analisis sentimen, atau lebih kompleks seperti pengelompokan dokumen berdasarkan topik. Tujuan dari *text classification* adalah untuk secara otomatis mengidentifikasi dan menetapkan label yang sesuai berdasarkan pola dalam teks (Manning, Raghavan, & Schütze, 2008). *Text classification* memiliki beberapa jenis berdasarkan tugas yang dihadapi, di antaranya adalah sebagai berikut.

1. *Binary Classification*: Mengelompokkan teks ke dalam dua kelas. Misalnya, klasifikasi email menjadi *spam* dan *non-spam*.
2. *Multiclass Classification*: Mengklasifikasikan teks ke dalam lebih dari dua kelas. Contohnya adalah pengelompokan berita ke dalam berbagai kategori seperti olahraga, politik, dan hiburan.
3. *Multilabel Classification*: Setiap teks dapat memiliki lebih dari satu label. Contohnya adalah mengklasifikasikan artikel ilmiah yang bisa masuk dalam lebih dari satu bidang keilmuan.

Text classification memiliki berbagai aplikasi penting di dunia nyata, berikut adalah contohnya.

1. *Spam Detection* : Sistem email dapat secara otomatis mengidentifikasi dan menyaring *email spam* menggunakan model klasifikasi teks.
2. *Sentiment Analysis* : Teks seperti ulasan produk, posting media sosial, atau komentar online dapat diklasifikasikan berdasarkan sentimen, seperti positif, negatif, atau netral.
3. *Topic Categorization* : Dokumen teks dapat dikelompokkan ke dalam kategori topik tertentu, seperti pengelompokan artikel berita berdasarkan topik seperti olahraga, politik, atau teknologi.
4. *Document Filtering* : Klasifikasi dokumen berdasarkan tipe, misalnya untuk menyaring konten yang tidak pantas di media sosial.

2.2.8 Evaluasi Model

Evaluasi model *text classification* sangat penting untuk menilai kinerja dan efektivitas model dalam mengelompokkan teks ke dalam kategori yang tepat. Evaluasi model dilakukan dengan menggunakan berbagai metrik untuk mengukur seberapa baik model dalam melakukan klasifikasi pada data yang tidak terlihat (*testing data*). Proses evaluasi ini membantu dalam memilih model terbaik, memahami kesalahan yang terjadi, dan mengoptimalkan kinerja model secara keseluruhan (Manning, Raghavan, & Schütze, 2008).

1. Confusion Matrix

Confusion matrix adalah alat evaluasi yang memberikan gambaran tentang performa model dengan menampilkan jumlah prediksi yang benar dan salah untuk setiap kelas. Dalam kasus klasifikasi biner, *confusion matrix* terdiri dari sebagai berikut.

- *True Positive* (TP): Kasus di mana model memprediksi kelas positif dengan benar.
- *True Negative* (TN): Kasus di mana model memprediksi kelas negatif dengan benar.
- *False Positive* (FP): Kasus di mana model salah memprediksi kelas positif.
- *False Negative* (FN): Kasus di mana model salah memprediksi kelas negatif.

2. Metrik Evaluasi

Berikut adalah beberapa metrik utama yang digunakan dalam mengevaluasi model *text classification*.

- *Accuracy*: Mengukur persentase prediksi yang benar (baik positif maupun negatif) dari keseluruhan data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Metrik ini memberikan gambaran umum tentang kinerja model, tetapi kurang informatif ketika terdapat ketidakseimbangan kelas (*imbalanced data*).

- *Precision*: Mengukur seberapa banyak prediksi positif yang benar dari semua prediksi positif yang dibuat oleh model.

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

Precision sangat penting ketika biaya kesalahan positif tinggi, seperti dalam deteksi *spam* atau penyakit.

- *Recall* (atau *Sensitivity*): Mengukur seberapa banyak prediksi positif yang benar dari semua data yang seharusnya positif.

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

Recall relevan ketika penting untuk menangkap sebanyak mungkin kelas positif, misalnya dalam deteksi penipuan atau diagnosis medis.

- *F1*-Score: Merupakan rata-rata harmonis dari *precision* dan *recall*. *F1-score* berguna ketika ada kebutuhan untuk menyeimbangkan *precision* dan *recall*.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

3. *Cross-Validation*

Cross-validation adalah teknik untuk memastikan bahwa model tidak terlalu tergantung pada satu set data tertentu. Salah satu metode yang populer adalah *k-fold cross-validation*, di mana data dibagi menjadi *k* bagian, dan model dilatih *k* kali, dengan setiap kali menggunakan satu bagian sebagai data uji dan sisanya sebagai data latih. Ini membantu memberikan estimasi kinerja yang lebih akurat dan mengurangi risiko *overfitting*.

4. *Overfitting* dan *Underfitting*

   Overfitting terjadi ketika model sangat cocok dengan data latih sehingga kinerjanya menurun pada data uji, karena model menangkap *noise* atau pola yang hanya ada pada data latih. Sedangkan, *underfitting* terjadi ketika model tidak mampu menangkap pola yang cukup dari data, sehingga kinerjanya buruk baik pada data latih maupun data uji.

2.3 Penelitian Terkait

2.3.1 Studi Kasus UMKM

UMKM adalah salah satu sektor yang memegang peranan penting dalam perekonomian Indonesia. Menurut data Kementerian Koperasi dan UKM, kontribusi UMKM pada tahun 2023 terhadap PDB sebesar lebih dari 60%. Munculnya berbagai *e-commerce* pada saat masa pandemi, membuat pendapatan UMKM meningkat, salah satunya di daerah Sumatera Selatan. Menurut Rianty & Rahayu (2021), omzet perusahaan di Sumatera Selatan meningkat karena adanya pengaruh dari *e-commerce* dengan melakukan penjualan dan promosi produk melalui internet. Hal tersebut juga memudahkan konsumen untuk mengenal produk-produk yang ditawarkan. Salah satu program yang pernah dilaksanakan untuk pengembangan UMKM di Sumatera Selatan, khususnya Kota Palembang, yaitu program Palembang *Go-Digital*. Program ini berupa pemberian bantuan website dan pelatihan penggunaan aplikasi digital (Mirani et al., 2022). Namun, pada pelaksanaannya program ini masih terdapat banyak kendala dan faktor penghalang.

Di sisi lain, setelah pandemi Covid-19, pemerintah Kota Palembang semakin menggencarkan berbagai pemanfaatan teknologi digital untuk mengembalikan perekonomian di daerahnya. Berdasarkan penelitian Asrol et al. (2022), program-program digitalisasi dari pemerintah pusat, daerah, dan lembaga lainnya sangat membantu para pelaku UMKM di daerah Kota Palembang.

Saat ini, data UMKM yang terdaftar di Kemenkop UKM mencakup UMKM formal. Namun, banyak UMKM yang beroperasi tanpa terdaftar secara formal. Oleh

karena itu, pemanfaatan sumber Big Data seperti Google Maps dapat menjadi solusi untuk melengkapi kekurangan tersebut.

2.3.2 Studi Kasus Pemetaan Potensi Ekonomi

Penelitian mengenai pemetaan potensi ekonomi telah beberapa kali dilakukan oleh peneliti terdahulu, contohnya adalah pemetaan potensi ekonomi sektoral dan estimasi pertumbuhan ekonomi di Pamekasan, Jawa Timur yang dilakukan oleh Wahed pada 2018. penelitian ini bertujuan untuk melakukan klasifikasi pertumbuhan sektor ekonomi dan mengetahui sektor basis serta unggulan di wilayah Pamekasan. Penelitian ini menggunakan metode Tipologi Klassen, Location Quotient dan Shift Share. Hasil dari penelitian ini menunjukkan sektor yang masuk dalam kategori unggulan dengan kriteria tergolong sektor maju dan tumbuh dengan pesat, sektor basis dan kompetitif, serta kontribusinya yang besar terhadap pembentukan PDRB Kabupaten Pamekasan, yaitu sektor perdagangan besar dan eceran, reparasi mobil, dan sepeda motor. Sedangkan sub sektor perdagangan besar dan eceran, reparasi mobil, dan sepeda motor yang potensial untuk dikembangkan sebagai sub sektor unggulan, yaitu sub perdagangan besar dan eceran (Wahed, 2018).

Selain pada Kabupaten Pamekasan, penelitian mengenai pemetaan potensi ekonomi juga pernah dilakukan pada Kabupaten Ngawi oleh Widyaningrum pada 2020. peneliti menggunakan data PDRB Kabupaten Ngawi tahun 2015 hingga 2019. Data PDRB digunakan karena dapat mengukur kesejahteraan masyarakat dan peningkatan kualitas standar hidup. Alat analisis yang digunakan yaitu Location Quotient (LQ), Analisis Shift-Share, Model Rasio Pertumbuhan (MRP), Analisis Overlay. Sektor perekonomian diklasifikasikan menjadi sektor basis dan sektor unggulan. Hasil penelitian menunjukkan bahwa sektor basis di Kabupaten Ngawi adalah pertanian; pengangkutan dan komunikasi; jasa-jasa. Dari hasil overlay menunjukkan sektor potensial di Kabupaten Ngawi terdiri dari perdagangan, hotel dan restoran; industri pengolahan; dan konstruksi. Sektor-sektor tersebut memiliki peluang untuk dikembangkan menjadi sektor unggulan. Sedangkan sektor unggulan di

Kabupaten Ngawi terdiri dari angkutan dan komunikasi; jasa-jasa (Widyaningrum & Cahyono, 2020).

2.3.3 Big Data untuk Pemetaan Sebaran Usaha

Penelitian yang dipublikasi oleh Cholifa Fitri Annisa dan Setia Pramana (2021) membahas pemanfaatan data Google Maps dalam statistik resmi, khususnya dalam memetakan usaha di sektor makanan dan minuman di Jawa dan Bali. Dengan metode web scraping, penelitian ini mengumpulkan data spasial terkait lokasi usaha dan membandingkannya dengan data BPS, menemukan kecocokan data sebesar 68,22%. Ini relevan karena menggunakan metode big data dan membandingkan hasilnya dengan data tradisional untuk memastikan akurasi representasi lokasi. Untuk meningkatkan akurasi, disarankan melakukan verifikasi lokasi melalui integrasi dengan sumber data tambahan seperti data dari aplikasi peta lain atau verifikasi lapangan.

Penelitian yang dipublikasikan oleh Madjida dan Takdir (2020) mengembangkan pendekatan *Text Processing* untuk mencocokkan alamat dari Big Data (Google Maps) dengan data BPS (Master File Desa). Pendekatan yang diusulkan mampu mencapai akurasi sebesar 73,58% pada tingkat desa menggunakan algoritma pemrosesan teks yang didasarkan pada Master File Desa (MFD) dari BPS. Tingkat akurasi meningkat pada level administratif yang lebih tinggi, dengan 82,39% kecocokan di tingkat kabupaten dan 71,78% di tingkat provinsi, terutama untuk data yang tidak cocok pada tingkat desa dan kecamatan. Penelitian ini relevan dengan topik pemetaan sebaran usaha dalam mengatasi tantangan integrasi antara data Big Data dengan sumber data tradisional seperti data dari direktori BPS.

2.3.4 Pemetaan Wilayah Berbasis Web Menggunakan Google Maps API

Menurut Sugita, D. O., & Witriyono, H (tanpa tahun), Google Maps dapat diintegrasikan ke dalam web menggunakan Google Maps API. Namun, penggunaannya masih terdapat pembatasan *request* karena fungsi yang digunakan masih bersifat *free*. Di sisi lain, penelitian yang dilakukan oleh Sholikhan, M.,

Prasetyo, S. Y. J., & Hartomo, K. D. (2019) di Kabupaten Kudus dalam pemetaan lokasi UMKM kaligrafi berbasis webGIS menyimpulkan bahwa pemetaan tersebut terbukti "cukup efektif" dan "cukup efisien" sesuai tujuan yang diharapkan. Dengan bantuan Google Maps API, pemetaan wilayah dapat dioptimalisasikan untuk mempermudah akses bagi yang membutuhkan.

2.3.5 Pengembangan *Machine Learning*

Penelitian yang dipublikasi oleh Lya Hulliyyatus Suadaa, Farid Ridho, Anugerah Karta Monika, dan Nucke Widowati K. Projo (2023) membahas pengembangan model klasifikasi teks menggunakan *machine learning* dan *transfer learning* untuk mengkategorisasikan KBLI berdasarkan deskripsi pekerjaan responden dan kategori KBLI yang berasal dari pariwisata. Metodologi *text processing* yang digunakan adalah *case folding*, *cleaning*, *stemming*, *tokenization* dan teknik *extraction* yang digunakan *Term Frequency-Inverse Document Frequency* (TF.IDF). Sedangkan, *text classification* yang digunakan adalah Random Forest, SVM, dan IndoBERT. Hasil dari penelitian ini menggunakan *machine learning* Random Forest memberikan hasil yang lebih baik dari SVM berdasarkan *precision, recall, F1-score,* maupun *accuracy*. *Transfer learning* menggunakan *fine-tuned* IndoBERT didapat *Recall, F1-Score*, *Accuracy* mendapatkan hasil yang lebih baik dibandingkan kedua *machine learning* berturut-turut 0.67, 0.68, dan 0.87. Namun, penelitian ini belum dapat mengeksploitasi kemampuan untuk memahami teks deskripsi pekerjaan yang pendek dan disarankan mengembangkan model yang *robust* dan efektif untuk teks yang pendek.

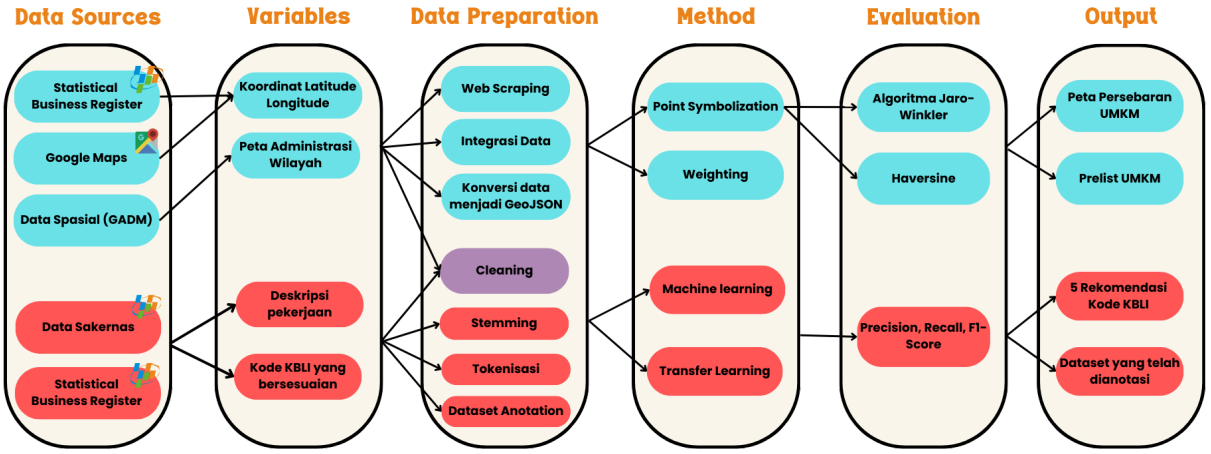
Penelitian yang dipublikasikan oleh Mahendri Dwi Cahyo dan Budi Yuniarto (2020) membahas klasifikasi data Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) untuk industri berdasarkan aktivitas yang dilakukan dengan menggunakan model deep learning, yaitu *Gated Recurrent Unit* (GRU), vektor kata dari *fastText*, dan teknik *Label Smoothing Regularization* (LSR) *loss* yang berfungsi untuk mengurangi dampak kesalahan klasifikasi dalam data. Penelitian ini juga mencakup berbagai teknik *preprocessing*, seperti pembersihan teks dan penghapusan data

duplikat, yang diterapkan pada data sampel Sensus Ekonomi 2016 dan pedoman resmi KBLI. Model ini mencapai akurasi 48,24% dalam merekomendasikan kode KBLI yang sesuai berdasarkan deskripsi aktivitas. Penerapan LSR membantu model mengatasi label yang tidak konsisten, sementara *fine-tuning* dengan data pedoman KBLI memperbaiki akurasi dengan mengarahkan model pada pola umum tiap kategori. Namun, kinerja model tetap terbatas oleh kualitas data latih yang digunakan, seperti adanya kesalahan label dan ambiguitas deskripsi

Penelitian yang dipublikasi oleh Agnes Septi Dwicahayaniawan dan Terry Devara Tri Saadi (2024) berfokus pada penerapan *machine learning* dan *transfer learning* untuk klasifikasi baku pekerjaan berdasarkan data Survei Angkatan Kerja Nasional (Sakernas) di Nusa Tenggara Barat. Studi ini mengeksplorasi penggunaan model IndoBERT, yang berbasis *transformer* dan dilatih secara khusus untuk bahasa Indonesia, dalam membedakan pekerjaan dengan tingkat akurasi yang lebih tinggi dibandingkan metode konvensional seperti *Support Vector Machine* (SVM). Hasil penelitian menunjukkan bahwa IndoBERT memiliki keunggulan dari segi akurasi dan F1-score pada pemodelan KBLI (0,76) dan KBJI (0,65). Namun, keunggulan tersebut tidak terlalu signifikan dengan beban komputasi yang lebih besar dan proses *training* yang menghabiskan waktu lebih lama. Selain itu, tantangan dalam akurasi juga muncul dari inkonsistensi dan *noise* dalam data, yang menghambat model dalam menghasilkan klasifikasi yang andal.

2.4 Kerangka Pikir

PANITIA PELAKSANA PRAKTIK KERJA LAPANGAN
POLITEKNIK STATISTIKA STIS
TAHUN AJARAN 2024/2025
Jalan Otto Iskandardinata No. 64C, Jakarta Timur 13330
Email: pkl@stis.ac.id



BAB III
METODOLOGI

Dalam melakukan penelitian diperlukan sebuah metode yang tepat agar bisa mencapai tujuan penelitian. Salah satu metode yang digunakan dalam penelitian ini yaitu Cross-Industry Standard Process for Data Mining (CRISP-DM). Metode ini menyediakan standar proses baku untuk data mining yang dapat diterapkan ke dalam strategi pemecahan masalah umum pada bisnis atau pada unit penelitian.

3.1. Business Understanding

Langkah pertama dalam metodologi CRISP-DM adalah memahami kebutuhan bisnis penelitian. Hal ini bertujuan untuk memastikan solusi yang dihasilkan relevan dan tepat sasaran (Hayat Suhendar & Widyani, 2023). Pada tahapan pemahaman bisnis ada beberapa hal yang dilakukan antara lain, menentukan tujuan bisnis, menilai situasi, dan menentukan tujuan data mining. Hal ini bertujuan untuk menyusun daftar awal (*pre-list*) lokasi usaha di kabupaten atau kota yang menjadi lokus di Provinsi Sumatera Selatan dengan memanfaatkan teknologi *big data.* Dengan memanfaatkan teknologi tersebut berbagai informasi lokasi usaha dapat diintegrasikan dan dianalisis secara lebih komprehensif agar memberikan panduan dalam proses *pre-listing* dan alokasi petugas untuk persiapan survei PKL.

Selain itu, penelitian ini berfokus pada penyediaan dataset yang memuat deskripsi pekerjaan dan aktivitas ekonomi dari responden di Provinsi Sumatera Selatan. Dataset ini telah dianotasi sesuai dengan kategori dalam KBLI.

Penelitian ini juga bertujuan untuk mengembangan model klasifikasi yang akurat dalam memberikan rekomendasi kode KBLI secara otomatis. Model tersebut diharapkan memiliki tingkat akurasi tinggi, sehingga dapat meningkatkan efisiensi dan konsistensi dalam pengelompokan data usaha.

Dengan memahami kebutuhan tersebut, penelitian ini diarahkan untuk menciptakan solusi berbasis teknologi yang tidak hanya praktis dalam mengimplementasikan, tetapi juga memberikan nilai tambah dalam pengelolaan data usaha, khususnya usaha mikro dan kecil (UMK).

3.2. Data Understanding

3.2.1. Pemetaan Wilayah Sebaran Usaha

Tahapan ini berfokus pada pemahaman terhadap data yang akan digunakan. Sumber data yang digunakan meliputi data spasial yang diperoleh dari Google Maps melalui proses *web scraping* dan direktori usaha BPS. Proses *web scraping* ini dilakukan untuk melengkapi data dari Direktori BPS, sehingga membentuk satu dataset baru yang merupakan gabungan dari hasil scraping dan data BPS. Data yang akan diperoleh dari hasil *web scraping* Google Maps berupa seluruh informasi yang disediakan oleh Google Maps tentang tempat tersebut, seperti nama usaha, kategori usaha, nomor telepon, dan lain-lain. Data-data tersebut akan diidentifikasi mengenai jumlah data yang diperoleh, missing value, dan lain-lain.

3.2.2. Otomatisasi KBLI

Untuk membangun model klasifikasi otomatis yang mengelompokkan kegiatan usaha ke dalam kategori Klasifikasi Baku Lapangan Usaha Indonesia (KBLI), diperlukan pemahaman mendalam terhadap data deskripsi kegiatan usaha. Sumber data utama yang digunakan adalah *dataset* deskripsi kegiatan usaha dan kode KBLI yang bersesuaian, yang diperoleh dari beberapa survei dan *register*. Fokus utama adalah pada data yang berasal dari Provinsi Sumatera Selatan, yang meliputi Survei Angkatan Kerja Nasional (Sakernas) untuk tahun 2021, 2022, dan 2023. Selain itu, digunakan juga data dari beberapa sumber lain, seperti Statistical Business Register (SBR) tahun

2024, Survei Industri Mikro dan Kecil Tahunan (IMK) tahun 2024, serta Survei Karakteristik Usaha (SKU) untuk tahun 2022 dan 2024. Data yang digunakan berupa teks mentah deskripsi kegiatan usaha yang dilengkapi dengan kode KBLI yang relevan, yang akan menjadi *input* utama dalam pengembangan model klasifikasi KBLI. Sebelum data dapat digunakan untuk pemodelan, dilakukan pengolahan lanjutan, termasuk proses tokenisasi, penghilangan *stopwords*, dan pelabelan manual, guna memastikan bahwa data siap untuk tahap pemodelan yang efektif.

3.3. Data Preparation

3.3.1. Pemetaan Wilayah Sebaran Usaha

Pada tahap preparation, data akan diperoleh dengan memanfaatkan Google Maps melalui proses *web scraping* dan data direktori BPS. Data Direktori BPS berupa alamat yang nantinya akan diubah menjadi koordinat menggunakan metode *geocoding*. Proses *web scraping* ini dilakukan untuk melengkapi data dari Direktori BPS, sehingga membentuk satu dataset baru yang merupakan gabungan dari hasil scraping dan data BPS. Dari kedua sumber data tersebut, akan dilakukan proses pencocokan data usaha menggunakan algoritma Jaro-Winkler untuk mencocokkan nama atau atribut teks dan algoritma jarak Haversine untuk mencocokkan lokasi geografis (Cholifa & Pramana, 2020). Proses ini bertujuan untuk menghilangkan data yang redundan sehingga dataset yang dihasilkan menjadi lebih akurat dan konsisten. Dengan kombinasi ini, diharapkan data yang dikumpulkan lebih kaya, akurat, dan relevan untuk analisis lebih lanjut.

Proses pengintegrasian juga termasuk menggunakan metode pemrosesan teks. Menurut Madjida dan Takdir (2020), agar big data dapat diintegrasikan dan dibandingkan dengan data dari sumber data tradisional yaitu survei dan sensus, diperlukan pemrosesan teks untuk

mengubah teks alamat dari big data yang masih berbentuk teks bebas menjadi menjadi format yang terstandarisasi agar sesuai dengan area administratif yang digunakan dalam statistik resmi.

Output dari *Data Preparation* ini yaitu sebuah dataset yang berisi informasi usaha dalam format geojson agar dapat dimanfaatkan pada tahap selanjutnya yaitu *modelling*

3.3.2. Otomatisasi KBLI

Langkah pertama dalam data *preparation* adalah *preprocessing text*, yang mencakup pembersihan data teks mentah untuk menghilangkan elemen-elemen yang tidak relevan, seperti tanda baca, angka, dan karakter khusus (Koshy et al., 2019). Proses ini juga melibatkan tokenisasi, yaitu pemecahan teks menjadi unit kata yang lebih kecil, serta normalisasi kata untuk menyamakan berbagai bentuk kata yang mungkin ada dalam teks. Selanjutnya, dilakukan penghapusan *stopwords*, yaitu kata-kata umum yang tidak memberikan kontribusi signifikan dalam analisis. Proses lainnya yang tak kalah penting adalah *stemming* atau l*emmatization*, yang berfungsi untuk mengubah kata-kata ke dalam bentuk dasarnya, sehingga mempermudah analisis dan klasifikasi (Yuyun et al., 2023).

Setelah proses pembersihan teks selesai, langkah berikutnya adalah pelabelan data (*labeling*), yaitu pemberian anotasi pada dataset untuk menentukan kategori KBLI yang sesuai berdasarkan informasi yang tersedia. Proses ini memadukan keahlian domain dengan referensi resmi KBLI, guna memastikan setiap entri data usaha mendapatkan label yang akurat. *Dataset* yang telah dilabeli ini kemudian akan digunakan untuk melatih model klasifikasi berbasis pembelajaran terawasi (*supervised learning*), yang bertujuan untuk mengotomatisasi proses klasifikasi usaha di masa depan (Jie, G et al., 2024).

3.4. Modeling

3.4.1. Pemetaan Wilayah Sebaran Usaha

Pada tahap *modelling*, fokus utamanya adalah pembuatan peta distribusi spasial usaha berupa peta titik (*point map*) dan peta *choropleth*. *Point map* digunakan untuk memvisualisasikan memetakan lokasi usaha berdasarkan dengan data spasial yang telah disiapkan, yaitu koordinat titik lokasi usaha yang nantinya akan digabungkan dengan peta wilayah administratif yang diperoleh dari *Global Administrative Area* (GADM). Di sisi lain, sedangkan peta choropleth digunakan untuk memvisualisasikan kepadatan atau klaster usaha secara agregat di suatu wilayah. Titik-titik koordinat lokasi usaha dikelompokkan berdasarkan kelurahan/desa dan hasil . Hasil pengelompokan tersebut digunakan sebagai menjadi pembobot untuk menghitung kepadatan usaha. Peta ini dibuat untuk empat lokus penelitian: Kota Palembang, Kabupaten Prabumulih, Kabupaten Ogan Ilir, dan Kabupaten Ogan Komering Ilir. Dan peta choropleth digunakan untuk memvisualisasikan kepadatan atau klaster usaha secara agregat di suatu wilayah, titik-titik koordinat lokasi usaha akan dikelompokkan berdasarkan kelurahan/desa dan menjadi pembobot dari kepadatan usaha di empat lokus penelitian yaitu Kota Palembang, Kabupaten Prabumulih, Kabupaten Ogan Ilir, dan Kabupaten Ogan Komering Ilir. Kepadatan usaha diperoleh dengan membagi jumlah usaha di suatu wilayah dengan luas wilayah tersebut yang dinyatakan dalam satuan usaha jiwa per km persegi. Peta ini dibuat untuk empat lokus penelitian: Kota Palembang, Kabupaten Prabumulih, Kabupaten Ogan Ilir, dan Kabupaten Ogan Komering Ilir. Peta-peta ini diharapkan dapat membantu untuk memahami bagaimana pola persebaran usaha usaha tersebar di suatu wilayah penelitian, baik saat sebelum survei

pendahuluan dilakukan maupun dan setelah hasil listing rumah tangga usaha didapatkan.

3.4.2. Otomatisasi KBLI

Pada tahap pemodelan, fokus utama adalah membangun model yang dapat memberikan rekomendasi kode KBLI berdasarkan deskripsi kegiatan usaha. Pendekatan yang digunakan menggabungkan metode *machine learning* dan *deep learning*. Langkah pertama adalah memilih model pembelajaran mesin yang tepat, dengan mempertimbangkan beberapa opsi seperti Support Vector Machine (SVM), Random Forest, dan IndoBERT. Ketiga model ini dipilih karena keunggulannya masing-masing: SVM memiliki kemampuan yang baik dalam menangani data berdimensi tinggi (Wahyuni et al., 2024), Random Forest efisien dalam mengurangi risiko overfitting (Efendi et al., 2024), dan IndoBERT, sebagai model *pretrained* berbasis *deep learning* yang dirancang khusus untuk bahasa Indonesia, menawarkan performa yang kuat dalam pemrosesan teks berbahasa Indonesia (Hidayat et al., 2024).

Setelah pemilihan model, langkah selanjutnya adalah melatih (*training*) dan menguji (*testing*) model. *Dataset* yang telah dipreproses akan dibagi menjadi dua bagian: data pelatihan untuk melatih model agar dapat mengenali pola, dan data pengujian untuk mengevaluasi performa model. Untuk memaksimalkan potensi model berbasis *deep learning,* digunakan teknik *transfer learning* dengan model IndoBERT. Dalam pendekatan ini, model pretrained IndoBERT akan di-fine-tune menggunakan *dataset* yang telah dilabeli sesuai kategori KBLI. Proses fine-tuning ini bertujuan untuk meningkatkan kemampuan model dalam memahami konteks spesifik bahasa Indonesia (Chandradev et

al., 2023), sehingga dapat mencapai akurasi yang lebih tinggi dalam mengklasifikasikan kode KBLI.

3.5. Evaluation

3.5.1. Pemetaan Wilayah Sebaran Usaha

Evaluasi model merupakan bagian penting untuk mengukur seberapa baik model dalam mengklasifikasikan kategori usaha dan non-usaha. Metode yang akan digunakan untuk evaluasi model adalah confusion matrix yang memberikan gambaran rinci tentang kinerja model dengan menunjukkan jumlah prediksi yang benar (True Positive dan True Negative) serta jumlah prediksi yang salah (False Positive dan False Negative). Berdasarkan confusion matrix tersebut, selanjutnya akan dihitung metrik evaluasi seperti *akurasi, presisi, recall*, dan *f1-score* untuk menilai performa model secara menyeluruh.

Salah satu yang menjadi pertimbangan dalam evaluasi adalah masalah keseimbangan data. Jika terdapat indikasi bahwa data tidak seimbang, metode evaluasi yang dapat digunakan adalah K-Fold Cross Validation. Secara umum, metode ini menghasilkan model dengan bias yang rendah. Nilai K pada metode ini dapat disesuaikan dengan kebutuhan. Semakin tinggi nilai K yang ditetapkan, model yang dihasilkan akan memiliki bias rendah, tetapi variansnya cenderung meningkat (Widyaningsih et al., 2021).

3.5.2. Otomatisasi KBLI

Pada tahap evaluasi, model yang telah dikembangkan pada tahap pemodelan akan diuji untuk memastikan ketepatan dan relevansi rekomendasi kode KBLI berdasarkan deskripsi kegiatan usaha. Evaluasi dilakukan menggunakan data pengujian yang terpisah dari data pelatihan, dengan tujuan untuk menguji performa model dalam kondisi yang lebih mendekati situasi dunia nyata. Proses evaluasi

model dibagi menjadi dua langkah utama: menggunakan metrik evaluasi klasifikasi dan teknik *cross-validation*.

Metrik evaluasi yang digunakan meliputi *precision*, *recall*, dan *F-1 score*, yang dihitung dari *confusion matrix* (Azhari et al., 2021). Metrik ini akan menilai akurasi klasifikasi yang dihasilkan oleh model. Selanjutnya, pada tahap *cross-validation*, hasil kode KBLI yang diprediksi oleh model akan dibandingkan dengan kode KBLI yang benar (*ground truth*) untuk memastikan bahwa rekomendasi yang dihasilkan oleh model akurat dan relevan. Dengan pendekatan ini, model dapat dievaluasi secara menyeluruh, memastikan kualitas prediksi dan kemampuannya dalam memberikan rekomendasi yang tepat berdasarkan konteks kegiatan usaha.

3.6. Data Presentation

3.6.1. Pemetaan Wilayah Sebaran Usaha

Output dari pemetaan wilayah sebaran usaha berupa menyajikan hasil pemetaan distribusi lokasi usaha yang divisualisasikan dalam bentuk peta sebaran lokasi usaha (point map) di Provinsi Sumatera Selatan dan peta kepadatan usaha (choropleth map) sebelum survei pendahuluan dan hasil listing di tingkat kelurahan/desa. Selain itu, output yang dihasilkan juga berupa dataset rumah tangga usaha yang didapatkan dari hasil listing.

3.6.2. Otomatisasi KBLI

Output dari otomatisasi KBLI adalah menyajikan model klasifikasi dalam bentuk API yang dapat digunakan untuk merekomendasikan kode KBLI berdasarkan kegiatan usaha secara otomatis dengan akurasi yang tinggi.

PANITIA PELAKSANA PRAKTIK KERJA LAPANGAN
POLITEKNIK STATISTIKA STIS
TAHUN AJARAN 2024/2025
Jalan Otto Iskandardinata No. 64C, Jakarta Timur 13330
Email: pkl@stis.ac.id

DAFTAR PUSTAKA

Ahilathasan, T., & Pekarskaya, T. (2020). Standard Industrial Code Classification by Using Machine Learning. Statistics Norway. Diakses pada 27 November 2024: https://statswiki.unece.org/download/attachments/285216428/ML_WP1_CC_Norway.pdf?version=1&modificationDate=1605171509316&api=v2

Alzubaidi, L., et al. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions." *Journal of Big Data* 8.1 (2021): 53. DOI: 10.1186/s40537-021-00444-8

Andrienko, G., Andrienko, N., & Savinov, A. (n.d.). *Choropleth Maps : Classification Revisited*. Retrieved October 20, 2024, from https://geoanalytics.net/and/papers/ica01.pdf

Andy Hermawan, Fatika Rahma Sanjaya, Gregorius Aldo Primantono, & Muhammad Syahirul Alim. (2024). Pengaruh Penggunaan Keywords Pada Penamaan Listing Airbnb Terhadap Tingkat Popularitas Di Kota Bangkok. Neptunus: *Jurnal Ilmu Komputer Dan Teknologi Informasi, 2(3)*, 32–45. https://doi.org/10.61132/neptunus.v2i3.179

Annisa, C. F., & Pramana, S. (2021). *KAJIAN PEMANFAATAN DATA GOOGLE MAPS DALAM Studi Kasus : Usaha Sektor Penyedia Makan Minum di Pulau Jawa dan Bali Utilization of Google Maps Data for Official Statistics Case Study : Food and Beverage Supply Sector Business in Java and Bali Islands. January.* https://doi.org/10.34123/semnasoffstat.v2020i1.614

Aritomatika, E. J., Sanjaya, A., & Widodo, D. W. (2021). Implementasi algoritma Jaro Winkler distance untuk pendeteksi kesamaan kata dalam pengembangan aplikasi English conversation. *Seminar Nasional Inovasi Teknologi*, UN PGRI Kediri, 24 Juli 2021, 104. e-ISSN: 2549-7952, p-ISSN: 2580-3336.

Asrol, S., Lidyah, R., Hartini, T., & Muhammadinah. (2022). Peran Percepatan Transformasi Digital untuk Keberlanjutan Bisnis dan Pemulihan Ekonomi Pelaku UMKM di Kota Palembang Pascapandemi. *Jurnal Intelektualita: Keislaman, Sosial, Dan Sains, 11(2)*, 242–246. https://doi.org/10.19109/intelektualita.v11i2.14685

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 640. https://doi.org/10.30865/mib.v5i2.2937

Badan Pusat Statistik (2020). Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) 2020. Peraturan BPS No. 2 Tahun 2020.

Badan Pusat Statistik (2022). *Teknik Pengumpulan dan Preprocessing Citra Satelit.*

Cambon, J., Hernagómez, D., Belanger, D., & Possenriede, D. (2021). Tidygeocoder: An R Package for Geocoding. *Journal of Open Source Software, 6*(65), 3544. https://doi.org/10.21105/ joss.03544

Chandradev, V., Suarjaya, I. M. A. D., & Bayupati, I. P. A. (2023). Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT. *Jurnal Buana Informatika, 14*(2), 107–116. https://doi.org/10.24002/jbi.v14i02.7244.

Cholissodin, I., & Riyandani, E. (2018). *ANALISIS BIG DATA*.

Chopde, N. R., & Nichat, M. K. (2013). Landmark based shortest path detection by using A* and Haversine formula. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2), 298.

Dwicahayaniawan, A. S., & Saadi, T. D. T. (2024). Potensi pemanfaatan machine learning dan transfer learning untuk klasifikasi baku pekerjaan. *Seminar Nasional Official Statistics, 2024*(1), 681–690. https://doi.org/10.34123/semnasoffstat.v2024i1

Dwi Cahyo, M., & Yuniarto, B. (2023). Deep learning for Indonesia standard industrial classification. In *Proceedings of the 2023 IEEE International Conference on Electrical Engineering and Informatics (ICEEI), Jakarta, Indonesia.* IEEE. https://doi.org/10.1109/ICELTICs50595.2020.9315361.

Efendi, Muhammad Syahrul., Sarwido., & Zyen Akhmad Khanif. (2024). Penerapan Algoritma Random Forest Untuk Prediksi Penjualan Dan Sistem Persediaan Produk. *Resolusi: Rekayasa Teknik Informatika dan Informasi,* 5(1), 12–20.

Elgendy, N., & Elragal, A. (2016). Big Data Analytics in Support of the Decision Making Process. Procedia Computer Science, 100, 1071–1084. https://doi.org/10.1016/J.PROCS.2016.09.251

Farid, & Yunus, Y. (2017). Analisa algoritma Haversine formula untuk pencarian lokasi terdekat rumah sakit dan puskesmas Provinsi Gorontalo. *ILKOM Jurnal Ilmiah*, 9(3), 353.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., Fdez-Riverola, F. (2013). Web scraping technologies in an API world, *Briefings in Bioinformatics, 15*(5), 788–797, https://doi.org/10.1093/bib/bbt026

Goodfellow, I., Bengio, Y., and Courville, A. "Deep learning." MIT Press, 2016.

Hayat Suhendar, M. T., & Widyani, Y. (2023). Machine Learning Application Development Guidelines Using CRISP-DM and Scrum Concept. *2023 IEEE International Conference on Data and Software Engineering (ICoDSE)*, 168–173. https://doi.org/10.1109/ICoDSE59534.2023.10291438

Hidayat, Wildan Amru., & Nastiti, Vinna Rahmayanti Setyaning (2024). PERBANDINGAN KINERJA PRE-TRAINED INDOBERT-BASE DAN INDOBERT-LITE PADA KLASIFIKASI SENTIMEN ULASAN TIKTOK TOKOPEDIA SELLER CENTER DENGAN MODEL INDOBERT. *Jurnal Sistem Informasi*, 11(2), 13–20.

Jie, G., Chen. T., Zhang. T., Cao, Q., Sun, Z., Luo, H. (2024). A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Transaction on pattern Analysis and Machine Intelligence*, 46(12), 9052–9071. https://doi.org/10.1109/TPAMI.2024.3415112

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Education.

Kurniawan, S. D., Widiastuti, R. Y., Hermanto, D. M. C., Mukhlis, I. R., Pipin, S. J., Surianto, D. F., Priyatno, A. M., Pasaribu, A. A., & Judijanto, L. (2024). BIG DATA (*Mengenal Big Data & Implementasinya di Berbagai Bidang*) (Sepriano & Y. Agusdi, Eds.; 1st ed.). PT. Sonpedia Publishing Indonesia.

Koshy, A., Balakumar, N.M.J., Shyna, A., John, A. (2019). Preprocessing Techniques for High Quality Text Extraction from Text Images. *1st International Conference on Innovations in Information and Communication Technology (ICIICT)*. https://doi.org/10.1109/ICIICT1.2019.8741488

Letouzé, E., & Jütting, J. (2015). Official statistics, Big Data and human development. *Harvard Humanitarian Iniciative, MIT Media Lab & Overseas Development Institute*, *June*, 1–26.

Madjida, W. O. Z., & Takdir. (2020). Big Data for Official Statistics: Administrative Area Identification from Plain Text Address. *2020 Asia-Pacific Statistics Week*, *June*. https://doi.org/10.13140/RG.2.2.25667.22565

Mahdia, F., & Noviyanto, F. (2013). Pemanfaatan Google Maps API untuk pembangunan sistem informasi manajemen bantuan logistik pasca bencana alam berbasis mobile web (Studi kasus: Badan Penanggulangan Bencana Daerah Kota Yogyakarta). *Jurnal Sarjana Teknik Informatika*, 1(1), 162. e-ISSN: 2338-5197.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. Retrieved from https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Meertens, Q. A., Diks, C. G. H., Van Den Herik, H. J., & Takes, F. W. (2022). Improving the Output Quality of Official Statistics Based on Machine Learning Algorithms. *J. Off. Stat., 38*(2), pp. 485–50, doi: 10.2478/jos-2022-0023.

Mirani, D., Aryansah, J. E., & Musdalifah, F. S. (2022). Efektifitas Program Palembang Go-Digital Bagi Pelaku UMKM di Masa Pandemi Covid-19. *Kolaborasi : Jurnal Administrasi Publik, 8(1)*, 108–121. https://doi.org/https://doi.org/10.26618/kjap.v8i1.6514

Nag, P. (1984). *Census Mapping Survey*. Concept. https://books.google.co.id/books?id=BZfY6O8k3PsC

Nunes, C. E. R., & Ashofteh, A. (2024). A Review of Big Data and Machine Learning Operations in Official Statistics: MLOps and Feature Store Adoption. *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), 1,* 711-718. DOI:10.1109/COMPSAC61105.2024.00101

Pan, S.J., and Yang, Q. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2010): 1345-1359. DOI: 10.1109/TKDE.2009.191

Rachedi, K., & Rachedi, F. (2021). Using Big Data in Official Statistics for Sustainable Development. In *Big Data Analytics* (pp. 261-270). Apple Academic Press.

Rianty, M., & Rahayu, P. F. (2021). Pengaruh E-Commerce Terhadap Pendapatan UMKM Yang Bermitra Gojek Dalam Masa Pandemi Covid-19. *Akuntansi Dan Manajemen*, 16(2), 153–167.

Sani, K., Akbar, D. A., & Sumantri, R. (2022). Social Safety Net dalam Pemberdayaan Usaha Mikro Kecil dan Menengah pada Masa Pandemi Covid-19 di Kota Palembang. Jurnal Intelektualita: Keislaman, Sosial, dan Sains, 11(2), 316-322. https://doi.org/10.19109/intelektualita.v11i2.10332

Santoso, A., & Nasir, M. (2021). Pemetaan Lahan dan Komoditas Pertanian Berbasis Webgis di Kabupaten OKU Timur. Jurnal Ilmiah Betrik: Besemah Teknologi Informasi dan Komputer, 12(2), 129-138.

Saputro, B. (2016). Klasifikasi dan pemetaan posdaya tematik berbasis masjid menggunakan metode naive bayes classifier [Tesis, Universitas Islam Negeri Maulana Malik Ibrahim].

Seminar Nasional Teknologi dan Sistem Informasi (SITASI) 2023. Analisis Pengaruh Penggunaan Shopee terhadap Kinerja UMKM di Surabaya. SITASI 2023 https://sitasi.upnjatim.ac.id/index.php/sitasi/article/download/558/83

Sensuse, D. I., Suwiyanto, V., Lusa, S., Gandhi, A., Mishbah, M., & Elisabeth, D. (2021). Designing knowledge sharing system for statistical activities in bps-statistics indonesia. Data, 6(5).

Setiawan, F., & Wibisono, S. (2021). *Algoritma bray&curtis berbobot pada cbr penentuan keluarga terdampak covid-19*.

Sharma, G. (2024). Web Crawling and Scraping: A Survey. *International Conference on Healthcare Innovations, Software and Engineering Technologies (HISET)*, 190-192, doi: 10.1109/HISET61796.2024.00063.

Sholikhan, M., Prasetyo, S. Y. J., & Hartomo, K. D. (2019). Pemetaan Lokasi UMKM Kaligrafi Kabupaten Kudus dengan Metode Location Based Service sebagai Media Promosi Berbasis WebGIS. *Indonesian Journal of Computing and Modeling*, 2(1), 8-16.

Siregar, R. (2023). Transformasi Digital Badan Pusat Statistik Indonesia Pada Era Big Data. Madani: Jurnal Ilmiah Multidisiplin, 1(11).

Sitorus, J. H., Achmad, F., Faisal, A., & Suhartono. (2021). Pemetaan Sebaran Umkm Berbasis Webgis Di Desa Pasuruan, Lampung Selatan. *Jurnal Dinamika Pengabdian, 7(1)*, 135–144.

Suadaa, L. H., Ridho, F., Monika, A. K., & Projo, N. W. K. (2023). Automatic text categorization to standard classification of Indonesian business fields (KBLI) 2020. In *Proceedings of the 2023 IEEE International Conference on Electrical Engineering and Informatics (ICEEI), Bandung, Indonesia*. IEEE. https://doi.org/10.1109/ICEEI59426.2023.10346866

Sugita, D. O., & Witriyono, H. APPLICATION OF TOURISM MAPPING IN SOUTH BENGKULU USING WEB-BASED GOOGLE MAPS API TECHNOLOGY.

Sun, Z., Strang, K., & Li, R. (2018). *10 Bigs : Big Data and Its Ten Big Characteristics 10 Bigs : Big Data and Its Ten Big Characteristics*.

Universitas Multimedia Nusantara. (2023). *Kenali karakteristik Big Data: 10V*. Diakses pada 12 Desember 2024, dari https://www.umn.ac.id/en/get-to-know-the-characteristics-of-big-data-10v/.

Wahed, M. (2018). Pemetaan Potensi Ekonomi Sektoral Dan Estimasi Pertumbuhan Ekonomi Kabupaten Pamekasan. *Ekonomi Dan Bisnis, 5(1)*, 1–16 https://doi.org/10.35590/jeb.v5i1.685

Wahyuni, Sonya Dian., & Kusumodestoni, R. H. (2024). Optimalisasi Algoritma Support Vector Machine (SVM) Dalam Klasifikasi Kejadian Data Stunting. *Bulletin of Information Technology (BIT), 5(2)*, 56–64. https://doi.org/10.47065/bit.v5i2.1247

Widyaningrum, E. D. A., & Cahyono, H. (2020). Pemetaan Potensi Wilayah Guna Mendorong Pembangunan Ekonomi Kabupaten Ngawi. *Jurnal Dinamika Pembangunan, 3(2)*, 117–139.

Widodo, W., Adly, E., Cahyati, M. D., Chamim, A. N. N., Satriawan, B. J., Saputra, I. R., Pangestu, M. I. M., & Kusuma, T. Y. T. (2024). Peta Digital Kependudukan Berbasis

QGIS : Inovasi Teknologi untuk Visualisasi Data kependudukan dan Populasi di Dusun Mrisi , Bantul. *Jurnal Surya Masyarakat, 6(2)*, 140–149. https://doi.org/https://doi.org/10.26714/jsm.6.2.2024.140-149

Yung, J. & Saltzman, B. (2018). A Machine Learning Approach to Identifying Different Types of Uncertainty. *Economic Letter, 171*, 58-62. https://doi.org/10.1016/j.econlet.2018.07.003

Yuyun, A. D. Latief, T. Sampurno, H. Hazriani, A. O. Arisha, and M. Mushaf, "Next Sentence Prediction: The Impact of Preprocessing Techniques in Deep Learning," 2023 *International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2023, pp. 1-6. doi: 10.1109/IC3INA60834.2023.10285805.

Zaharuddin, H. (2009). Menangkap Peluang Usaha. Jakarta: Dian Anugrah Perkasa.

Zhao, B. (2020). Encyclopedia of Big Data. Encyclopedia of Big Data, December. https://doi.org/10.1007/978-3-319-32001-4