BAB II
TINJAUAN PUSTAKA

Tinjauan pustaka adalah kumpulan dari data dan informasi ilmiah berupa teori-teori, metode, maupun pendekatan yang pernah dikembangkan dan telah didokumentasikan dalam bentuk buku, jurnal, naskah, catatan, rekaman sejarah dan dokumen serupa. Maka dari itu tujuan dari penyusunan kajian pustaka ini adalah untuk mendapatkan gambaran untuk menghindari adanya peniruan atau plagiasi terhadap penelitian milik orang lain.

Penelitian dengan judul Analisis Sentimen Respons Twitter terhadap Persyaratan Badan Penyelenggara Jaminan Sosial (BPJS) di Kantor Pertanahan oleh Ridho Darman (2023) membahas tentang sentimen publik di Twitter mengenai kebijakan pemerintah Indonesia yang mengharuskan BPJS sebagai kartu kesehatan dalam transaksi real estat. Penelitian ini menggunakan Tweet Harvest sebagai alat pengumpul data. Analisis digunakan menggunakan metode kuantitatif, data dianalisisi menggunakan Python dan metode leksikon.

Data yang terkumpul pada tahap pengumpulan data untuk *keyword* "bpjs" dan "syarat tanah" dengan *limit* 500 data adalah 480 baris data *tweet*. Pengolahan sentimen dari data didapatkan hasil 244 *tweet* bernilai positif, 140 *tweet* bernilai negatif dan sisanya 39 *tweet* bernilai netral. Polarisasi sentimen dilakukan menggunakan *library* "matplotlib.pyplot" dan menghasilkan polarisasi 57,7% respon pengguna Twitter positif terhadap persyaratan BPJS di kantor pertanahan, 33,1% lainnya memberikan respon negatif dan 9,2% netral.

Berdasarkan nilai hasil analisis yang telah penelitian oleh Ridho Darman tersebut menunjukkan bahwa cuitan dari akun pemerintah umumnya positif, sedangkan akun pribadi menunjukkan lebih banyak respons negatif, analisa sentimen yang telah dilakukan memberikan wawasan bagi pemerintah untuk menyempurnakan dan meningkatkan kebijakan layanan publik [9].

Penelitian oleh Mantika, A.M., dkk (2024) dengan judul Sentiment Analysis on Twitter Using Naïve Bayes and Logistic Regression for the 2024 Presidential Election melakukan pengujian analisis sentimen publik pada media sosial Twitter mengenai kandidat-kandidat pada pemilihan presiden Indonesia tahun 2024 menggunakan metode Naïve Bayes and Logistic Regression. Studi tersebut bertujuan untuk menentukan model analisis sentimen yang lebih akurat untuk mencerminkan opini publik dengan melakukan klasifikasi cuitan Twitter menjadi sentimen positif, netral, dan negatif.

Penelitian ini menggunakan pendekatan metodis dengan melibatkan alat pengumpul data media sosial X yaitu "Tweet Harvest", prapemrosesan data, dan analisis sistematis menggunakan algoritma. Hasil dari penelitian ini menunjukkan hasil bahwa Naïve Bayes umumnya mengungguli Logistic Regression dalam hal akurasi di berbagai kandidat, serta memberikan informasi tentang sentimen publik yang dapat mempengaruhi strategi kampanye [10].

Penelitian yang dilakukan oleh R. Yunita, dkk (2023) dengan judul Perbandingan Algoritma SVM Dan Naïve Bayes Pada Analisis Sentimen Penghapusan Kewajiban Skripsi menggunakan algoritma Support Vector Machine (SVM) dan Naïve Bayes untuk membahas tentang analisis sentimen terhadap kebijakan penghapusan kewajiban skripsi di perguruan tinggi di Indonesia. Penelitian ini menggunakan "Tweet Harvest" untuk mengumpulkan dataset. Dataset yang telah dikumpulkan dipilih berdasarkan kriteria tertentu dan diberi label sentimen positif dan negatif secara manual untuk analisis sentimen.

Proses preprocessing dilakukan untuk mempersiapkan data teks sebelum dilakukan analisis sentimen. Evaluasi dilakukan dengan membandingkan kinerja kedua algoritma Support Vector Machine (SVM) dan Naïve Bayes menggunakan metrik seperti akurasi, recall, precision, dan F1-score. Hasil penelitian memberikan wawasan mengenai persepsi publik terhadap kebijakan tersebut di media sosial Twitter. Selain itu, penelitian ini juga menentukan metode yang paling baik dalam melakukan klasifikasi teks berisi sentimen publik [11].

Penelitian yang dilakukan oleh H. Faradian, dkk (2024) dengan judul Analisis Sentimen Terhadap Penutupan Tiktok Shop Menggunakan Algoritma Naïve Bayes Classifier Pada Media Sosial X melibatkan pengumpulan data melalui scraping data pada platform X menggunakan "Tweet Harvest" yang memungkinkan peneliti untuk mengumpulkan sejumlah data tweet yang berkaitan dengan penutupan TikTok Shop, sehingga dapat dilakukan analisis sentimen terhadap opini dan pendapat yang tersebar di platform X terkait topik tersebut.

Data yang terkumpul dari proses tweet harvest ini kemudian menjadi dasar untuk dilakukan analisis lebih lanjut menggunakan Naïve Bayes Classifier guna mengklasifikasikan sentimen positif dan negatif terhadap penutupan TikTok Shop. Data yang dikumpulkan berasal dari tanggal 29 September 2023 hingga 29 November 2023 dengan kata kunci "TikTok Shop dilarang". Data tersebut kemudian dibagi menjadi data training dan data testing, kemudian disimpan dalam file dengan format csv.

Pelabelan data merupakan tahap selanjutnya, data dilabeli sebagai sentimen positif atau negatif menggunakan lexicon based dari data training yang telah melalui proses preprocessing. Label ini digunakan sebagai klasifikasi kelas pada analisis sentimen. Metode klasifikasi yang dipilih dalam penelitian ini adalah Naïve Bayes Classifier karena dikenal sebagai metode yang sederhana, cepat, dan memiliki tingkat akurasi yang tinggi berdasarkan teorema Bayes. Data hasil analisis sentimen dapat dimanfaatkan untuk meningkatkan kinerja dan sebagai bahan acuan bagi pihak terkait, termasuk pemerintah, dalam mengevaluasi respon masyarakat terhadap penutupan TikTok Shop di Indonesia [12].

Penelitan yang berjudul Analisis Sentimen: Pengaruh Jam Kerja Terhadap Kesehatan Mental Generasi Z oleh M. D. A. Fahreza, dkk (2024) menggunakan "Tweet Harvest" untuk melakukan pengumpulan data pada media sosial Twitter dengan kata kunci "jam kerja generasi Z", "kesehatan mental generasi Z", dan "tingkat stres kerja Gen Z". Penelitian ini menggunakan 2 akun Twitter yang berbeda dengan total 2956 data tweet yang dikumpulkan. Data yang telah

dikumpulkan kemudian dilakukan proses pra-pemrosesan data untuk membersihkan, merapikan, dan mengubah data mentah untuk analisis lebih lanjut.

Penelitian ini menggunakan berbagai algoritma stemming, seperti Sastrawi, Nazief Adriani, dan Arifin Setiono dalam mengubah kata-kata menjadi bentuk dasar analisis sentimen. Hasil dari penelitian ini memberikan wawasan tentang respons masyarakat terhadap topik kesehatan mental Generasi Z dan dampak jam kerja, serta menunjukkan pentingnya kesadaran akan kesehatan mental dalam lingkungan kerja [13].

Penelitian yang berjudul Sentimental Analysis of YouTube Videos oleh Baravkar, A., dkk (2021) menggali pemanfaatan analisis sentimen terhadap komentar video Youtube untuk meningkatkan cara menemukan video edukatif yang berkualitas. Penelitian ini menargetkan platform Youtube sebagai sumber utama distribusi konten edukatif. Pengumpulan data pada penelitian dilakukan menggunaan API Youtube dengan metadata jumlah tontonan, suka, dan komentar. Analisis sentimen dilakukan dengan pembelajaran mesin menggunakan regresi logistik untuk menganalisis sentimen dalam komentar, model tersebut dilatih menggunakan dataset ulasan produk Amazon yang memiliki kesamaan dalam hal ulasan pengguna.

Video diurutkan berdasarkan peringkat video untuk memprioritaskan video yang tidak hanya populer tetapi memberikan tanggapan positif dari penonton. Hasil dari penelitian ini menunjukkan bagaimana analisis sentimen dapat diintegrasikan dalam algoritma pemeringkatan untuk meningkatkan relevansi dan kualitas hasil pencarian video edukatif di Youtube serta dengan pendekatan ini berpotensi merubah cara pelajar dan pendidik dalam menemukan konten edukatif di platform digital [14].

Penelitian yang dilakukan oleh Himawan, A., dkk (2020) dengan judul "Implementation of Web Scraping to Build a WebBased Instagram Account Data Downloader Application" bertujuan memfasilitasi akases yang lebih mudah ke data Instagram untuk berbagai kegunaan, mulai dari penelitan akademis hingga analisis

bisnis, tanpa memerlukan layanan API yang mahal. Penelitian dilakukan menggunakan bahasa pemrograman Python dengan library Beautiful Soup untuk memberikan pengumpulan data Instagram secara otomatis melalui web scraping sebagai alternatif gratis API yang memiliki *limit*. Sistem di uji menggunakan 15 akun instagram dan mampu mengunduh 2412 entri data perakun, termasuk postingan, suka, dan komentar yang kemudian dikelola dalam antarmuka pengumpulan data.

Arsitektur aplikasi meliputi Antarmuka Pengguna Web untuk interaksi pengguna. Proses ini dijelaskan dari analisis data awal hingga fase desain dan implementasi, menenekankan utilitas web scraping tanpa akses API. Hasil dari aplikasi memberikan ekstraksi data dengan kemampuan mengelola data melalui fungsionalitas seperti menghapus, ekspor ke format csv, excel, atau Json. Kinerja aplikasi membuktikan efikasi dalam pengumpulan data besar tanpa biaya untuk pengguna [15].

Penelitian yang dilakukan oleh Djufri, Mohammad (2020) dengan judul PENERAPAN TEKNIK WEB SCRAPING UNTUK PENGGALIAN POTENSI PAJAK (Studi Kasus pada Online Market Place Tokopedia, Shopee dan Bukalapak) bertujuan untuk mengeksplorasi dan menenilai efektivitas teknik web scraping dalam mengidentifiakasi potensi pajak dari transaksi e-commerce di Indonesia. Sistem yang dibuat bekerja dengan mengumpulkan data transaksi e-commerce menggunakan bahasa pemrograman pyton dan php serta ekstensi Google Chrome.

Data yang terkumpul disimpan dalam database NoSql dan dilakukan analisis dengan mengintegrasikan kedalam sistem Business Intelligence. Objek penelitian ini adalah transaksi yang umum terjadi di marketplace, mencakup semua penjual yang aktif di platform tersebut, dengan protokol keamanan dan kepatuhan data yang ketat untuk melindungi informasi pengguna.

Hasil dari penelitian ini berhasil menunjukkan bagaimana teknik web scraping dapat digunakan secara efektif untuk mengumpulkan data transaksi dari marketplace online di Indonesia serta menunjukkan data yang relevan dapat

dikumpulkan secara sistematis dan analitis. Hal tersebut memberikan informasi data tentang potensi pajak dari ekonomi digital yang berkembang pesat di Indonesia [16]. Penelitian yang berjudul Development of web crawler to build Indonesian text corpus oleh J. Hendryli, dkk (2020) membahas tentang pembuatan web crawler yang bertujuan mengekstrak konten berita Indonesia dari situs web detikNews dan membangun dataset dari teks yang telah dikumpulkan tersebut.

Penelitian ini menggunakan model pengembangan perangkat lunak waterfall, yang meliputi tahapan analisis kebutuhan, desain sistem, implementasi, pengujian, dan pemeliharaan. Crawler ini dikembangkan menggunakan bahasa pemrograman Python, dengan bantuan pustaka Scrapy dan BeautifulSoup4 untuk scraping dan pembersihan data. Crawler ini akan mengumpulkan data dari halaman indeks detikNews dan mengumpukan semua halaman berita dari tanggal yang ditentukan sampai crawler dihentikan. Data yang diekstrak meliputi judul, nama penulis, tanggal, isi berita, url, dan waktu publikasi.

Crawler ini dirancang untuk mengabaikan konten gambar dan video dan hanya fokus pada teks. Crawler ini berhasil mengumpulkan lebih dari 790.000 item berita dari Mei 2011 hingga April 2020, dengan total kata mencapai 190 juta kata dari 14 juta kalimat. Kesimpulan dari penelitan ini adalah untuk menekankan bahwa crawler yang dikembang berhasil membangun sebuah korpus besar teks berita Indonesia yang dapat menjadi sumber daya penting untuk penelitan NLU bahasa Indonesia [17].

Penelitian yang berjudul CLICK-ID: A novel dataset for Indonesian clickbait headlines oleh William A., dkk (2020) membahas tentang pembuatan web crawler untuk melatih model klasifikasi clickbait menggunakan teknik pemrosesan bahasa alami (NLP). Penelitian ini bertujuan untuk membantu dalam pengembangan solusi yang lebih efektif mengatasi masalah clickbait di Indonesia. Dataset dikumpulkan melalui scraping dari 12 penerbit berita online di Indonesia. Data yang terkumpul mencakup informasi detail dari setiap artikel seperti judul, penerbit, tanggal publikasi, kategori, sub-kategori, dan konten lengkap artikel dalam format CSV dan XLSX. Pengumpulan dataset dilakukan menggunakan web

scraper yang dibangun dengan Python dan library Scrapy untuk mengumpulkan 46.517 judul berita.

Total data yang terkumpul sebanyak 15.000 judul dianotasi dengan label clickbait atau non-clickbait. Proses Anotasi dilakukan dengan pendekatan konsensus dimana setiap judul dianotasi oleh tiga annotator, dan mayoritas menentukan label akhir. Menggunakan dataset yang telah dianotasi, peneliti mengembangkan model klasifikasi clickbait berbahasa Indonesia. Model ini menggunakan arsitektur Bi-LSTM dan CNN yang telah terbukti efektif dalam tugas-tugas NLP serupa.

Eksperimen dilakukan dengan menggunakan metode validasi silang 5-fold untuk menilai kinerja model. Hasil dari penelitian ini mengisi celah keterbatasan dataset untuk sumber daya NLP berbahasa Indonesia dan membantu pengembangan solusi untuk mengatasi masalah clickbait di Indonesia. Diharapkan pula dapat dimanfaatkan bagi berbagai aplikasi NLP dalam bahasa Indonesia di masa depan [18].

Penelitian oleh Mansur A., dkk. (2021) dengan judul The Performance of Indonesia's President: A Sentiment Analysis in social media dengan tujuan mengukur persepsi publik dan mendapatkan umpan balik yang dapat digunakan oleh pembuat kebijakan untuk meningkatkan kinrja kepemimpinan. Penelitian ini menggunakan web scraping dengan bahasa pemrograman Python dan Selenium Library untuk mengumpulkan data komentar dari Youtube. Penelitian ini menggunakan TextBlob, sebuah library Python untuk pemrosesan Natural Language Processing (NLP).

Analisa sentimen digunakan untuk mengklasifikasikan komentar sebagai positif, negatif, atau netral. Hasil dari penelitian ini adalah wawasan mengenai persepsi publik terhadap kinerja kepemimpinan Jokowi, mengidentifikasi faktor-faktor yang berkontribusi terhadap sentimen yang positif dan negatif [19].

Tabel 2.1. Tabel Perbandingan Penelitian

Peneliti	**A. Himawan, A. Priadana, dan A. W. Murdiyanto** [15]	**Djufri, Mohammad** [16]	**Janson Hendryli, Viny C Mawardi** [17]	**Andika William, Yunita Sari** [18]	**Agus Mansur, Zuhdi Allamsyah, Putri Amalia** [19]	**Johanes Daulat Tamba**
Judul	Implementation of Web Scraping to Build a WebBased Instagram Account Data Downloader Application	PENERAPAN TEKNIK WEB SCRAPING UNTUK PENGGALIAN POTENSI PAJAK (Studi Kasus pada Online Market Place Tokopedia, Shopee dan Bukalapak)	Development of web crawler to build Indonesian text corpus	CLICK-ID: A novel dataset for Indonesian clickbait headlines	The Performance of Indonesia's President: A Sentiment Analysis in Social Media	Pembangunan Sistem Otomasi Pengumpul Data dan Pelabelan Sentimen pada Media Sosial X dan Youtube.
Platform	*Website*	*Website*	*Website*	Terminal Lokal	Terminal Lokal	*Website*
Bahasa Pemrograman	Python	Python dan Php	Python	Python	Python	Python
Framework	Flask	Tidak Disebutkan	*Scrapy*	*Scrapy*	Tidak Disebutkan	Django

Database / Penyimpanan	NoSql	NoSql	Tidak Disebutkan	CSV dan XLSX	CSV	Sqllite
Metode Pengembangan	Tidak Disebutkan	Tidak Disebutkan	Waterfall	Tidak Disebutkan	Tidak Disebutkan	*Waterfall*
Pengelolaan Data Pengguna	Tidak	Tidak	Tidak Disebutkan	Tidak	Tidak	Ya
Objek Penelitian	Umum	Direktorat Jenderal Pajak Indonesia	Umum	Umum	Umum	Umum
Entitas Online	Instagram	Tokopedia, Bukalapak, dan Shopee	detikNews	detikNews, Kompas, Liputan6, Okezone, Tribunnews, Republika, SINDOnews, Tempo, Fimela, KapanLagi	Youtube	X dan Youtube
Menampilkan Hasil Analisis?	Ya	Ya	Ya	Ya	Ya	Ya
Menyediakan fitur pelabelan sentimen?	Tidak	Tidak	Tidak	Tidak	Ya	Ya
Menyediakan Fitur Unduh Laporan?	Ya	Ya	Tidak	Tidak	Tidak	Ya