ANALISIS SENTIMEN BERITA KRIMINAL DETIK.COM MENGGUNAKAN WEB SCRAPING DAN NAIVE BAYES CLASSIFIER

PROPOSAL

Disusun sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Komputer
dari Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang

Oleh:
ROFIZAIDAN MAULANA
1810631170087

PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SINGAPERBANGSA KARAWANG
KARAWANG
2023

LEMBAR PENGESAHAN

ANALISIS SENTIMEN BERITA DETIK.COM MENGGUNAKAN WEB SCRAPING DAN NAIVE BAYES CLASSIFIER

PROPOSAL

Disusun sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Komputer
dari Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang
Oleh:

ROFIZAIDAN MAULANA
1810631170087

Disetujui oleh:

Pembimbing I	**Penguji I**
Dadang Yusup, M.Kom.	
0424048202	

Karawang, 07 Agustus 2023
Diketahui dan disahkan
oleh:
Dekan Fakultas Ilmu Komputer

Dr. Mayasari, S.Si., M.Hum.
NIDN. 0426097905

KATA PENGANTAR

Dalam kesempatan ini, dengan rasa syukur dan dedikasi, saya mempersembahkan proposal ini yang berjudul "Analisis Sentimen Berita Detik.com Menggunakan Web Scraping dan Naive Bayes Classifier". proposal ini merupakan salah satu persyaratan untuk memperoleh gelar Sarjana Informatika di Universitas Singaperbangsa Karawang.

Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap berita kriminal pada situs Detik.com dengan menggunakan metode web scraping untuk pengumpulan data komentar dan metode klasifikasi Naive Bayes Classifier untuk analisis sentimen. Dalam penelitian ini, saya berusaha untuk merangkum data dan mengekstrak informasi penting dari komentar pada berita kriminalitas dengan tujuan untuk memberikan pemahaman yang lebih mendalam tentang persepsi dan tanggapan masyarakat terhadap fenomena kriminalitas.

Saya menyadari bahwa penulisan proposal ini tidak terlepas dari berbagai keterbatasan dan tantangan. Namun, semangat untuk berkontribusi dalam pengembangan ilmu pengetahuan dan teknologi, khususnya dalam analisis data berita, telah mendorong saya untuk menyelesaikan penelitian ini. Dalam proses penyusunan proposal ini, terdapat banyak pihak yang telah memberikan dukungan, bimbingan, dan dorongan kepada saya. Saya ingin mengucapkan terima kasih yang tulus kepada:

1. Ibu Dr. Mayasari, M.Hum., yang menjabat sebagai Dekan Fakultas Ilmu Komputer.
2. Bapak Garno, S.Kom., M.Kom., selaku Wakil Dekan Bidang Akademik dan Kemahasiswaan Fakultas Ilmu Komputer.

3. Bapak Muhammad Jajuli, S.Si, M.Si., selaku Wakil Dekan bidang umum dan keuangan Fakultas Ilmu Komputer.
4. Ibu Betha Nurina Sari, M.Kom., selaku Koordinator Program Studi Informatika Fakultas Ilmu Komputer.
5. Bapak Dadang Yusup, M.Kom., selaku dosen pembimbing yang dengan tulus telah memberikan dukungan, arahan, saran, serta motivasi
6. Seluruh Dosen yang telah memberikan saya ilmu yang sangat bermanfaat serta Staf Tata Usaha Fakultas Ilmu Komputer
7. Kedua orang tua yang saya cintai, yang telah membimbing saya dengan sangat baik dengan dukungan moral maupun materi, nasihat dan do'a, sehingga perkuliahan dan penyusunan penelitian ini terlaksana dengan baik
8. Teman-teman mahasiswa Program Studi Informatika angkatan 2018 khususnya kelas D yang telah berjuang bersama selama menempuh pendidikan di Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang
9. Semua pihak yang telah membantu dalam menyelesaikan proposal ini yang tidak dapat disebutkan satu persatu. Akhir kata, saya berharap proposal ini dapat memberikan manfaat dan sumbangsih positif bagi pengembangan ilmu pengetahuan dan teknologi, khususnya di bidang analisis data dan informasi anime. Semoga hasil penelitian ini dapat memberikan inspirasi dan membuka peluang bagi penelitian selanjutnya.

Akhir kata, saya berharap proposal ini dapat memberikan manfaat dan sumbangsih positif bagi pengembangan ilmu pengetahuan dan teknologi,

khususnya di bidang analisis data dan informasi anime. Semoga hasil penelitian ini dapat memberikan inspirasi dan membuka peluang bagi penelitian selanjutnya.

Karawang, 07 Agustus 2023

Rofizaidan Maulana

DAFTAR ISI

BAB 1

PENDAHULUAN

1. Latar Belakang

Tepat di tahun 2021, Indonesia mengalami peningkatan jumlah insiden kekerasan kolektif sebanyak 1.221 kasus. Tim Collective Violence Early Warning (CVEW) dari Centre for Strategic and International Studies (CSIS) mengumpulkan data ini dari 75 media daring di Indonesia (Alexandra et al., 2022). Kekerasan kolektif mencakup berbagai bentuk konflik, seperti konflik etnis, terorisme, separatisme, dan konflik politik lainnya. Frekuensi kekerasan kolektif mengalami peningkatan terutama pada bulan Agustus dan Oktober. Dalam menghadapi tahun politik, penting untuk melakukan peringatan dini dan tindakan dari semua pemangku kepentingan guna mengantisipasi potensi meningkatnya kembali kekerasan kolektif (Panggabean, 2018). Pengumpulan data akan terus berlangsung hingga tahun pemilu 2024. Pada kuartal pertama, terdapat 206 insiden, sedangkan pada kuartal keempat atau menjelang akhir tahun, jumlah insiden meningkat 70 persen menjadi 370. Bulan Agustus dan Oktober menjadi bulan dengan lonjakan tertinggi, dengan masing-masing 151 dan 162 kasus (Dian , 2022).

Berita kriminalitas memiliki peran yang signifikan dalam menyampaikan informasi tentang kejadian-kejadian kejahatan yang terjadi di masyarakat. Berita kriminal mencakup berbagai jenis kejahatan, seperti pembunuhan, penipuan, pemerkosaan, pencurian, perampokan, narkoba, dan penganiayaan, yang melanggar peraturan hukum (Yuniwati et al., 2021). Berita kriminalitas dapat memberikan wawasan tentang tingkat kejahatan, pola kejahatan, dan respons masyarakat terhadap kejadian tersebut. Berita kriminal memiliki daya tarik yang tinggi bagi pembaca karena mencerminkan ancaman

dan menghadirkan elemen drama dan emosi (Hardjo, 2019). Oleh karena itu, menganalisis sentimen terhadap berita kriminalitas dapat memberikan pemahaman yang lebih mendalam tentang persepsi dan tanggapan masyarakat terhadap fenomena kriminalitas yang terjadi (Panuntun et al., 2023).

Selanjutnya, dalam penelitian ini, dilakukan penggunaan metode web scraping untuk mengumpulkan data komentar pada berita kriminalitas dari situs Detik.com. Web scraping adalah proses pengumpulan informasi dari situs web dengan menelusuri dan mengindeks dokumen HTML (Fahrudin et al., 2023). Web scraping dipilih sebagai metode pengumpulan data karena memungkinkan peneliti untuk mengakses dan mengumpulkan data secara otomatis dari situs web yang menjadi sumber informasi. Proses web scraping melibatkan pembuatan template scraping, eksplorasi navigasi situs, otomatisasi navigasi dan ekstraksi informasi, serta ekstraksi dan penyimpanan data (Wibowo et al., 2019). Dengan menggunakan web scraping, peneliti dapat mengumpulkan data komentar secara efisien dan memperoleh dataset yang mencakup berbagai perspektif dan pendapat dari pembaca situs Detik.com terkait berita kriminalitas (Fahrudin et al., 2023).

Setelah melakukan pengumpulan data komentar, langkah selanjutnya adalah melakukan analisis sentimen menggunakan metode klasifikasi Naive Bayes Classifier. Naive Bayes Classifier adalah metode klasifikasi yang berdasarkan pada teorema Bayes dengan asumsi sederhana yaitu independensi antara setiap pasangan fitur (Sari & Wibowo, 2019). Metode ini telah banyak digunakan dalam analisis sentimen karena efisiensinya dalam melakukan klasifikasi teks dan kemampuannya dalam mengatasi data yang memiliki dimensi tinggi. Dengan menerapkan Naive Bayes Classifier, peneliti dapat mengklasifikasikan sentimen komentar pada berita kriminalitas menjadi positif, negatif,

atau netral, sehingga memberikan wawasan tentang pandangan masyarakat terhadap berita tersebut (Normawati & Prayogi, 2021).

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, penelitian ini bertujuan untuk melakukan analisis sentimen guna memahami sentimen positif dan negatif publik terhadap berita kriminal di Detik.com. Selanjutnya, analisis ini akan diuji menggunakan algoritma Naive Bayes Classifier untuk mengevaluasi tingkat keakuratannya. Tujuan penelitian ini juga mencakup pemahaman lebih mendalam mengenai opini publik setelah munculnya berita-berita kriminal serta visualisasi data hasil analisis. Oleh karena itu, judul penelitian ini adalah "Analisis Sentimen Berita Kriminal Detik.com Menggunakan Web Scraping dan Naive Bayes Classifier".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, terdapat beberapa rumusan masalah yang ingin dijawab dalam penelitian ini, yaitu:

1. Bagaimana melakukan pengambilan data berita dari situs detik.com menggunakan metode web scraping?
2. Sejauh mana metode naïve bayes classifier berhasil dalam menganalisis data ulasan berita kriminalitas detik.com?

1.3 Batasan Masalah

Dalam skop penelitian ini, terdapat beberapa batasan yang perlu ditegaskan, yaitu:

1. Tujuan utama penelitian ini adalah melakukan analisis sentimen pada komentar berita kriminal yang ada di Detik.com.
2. Data yang menjadi subjek penelitian ini terbatas pada berita yang telah diterbitkan dalam kurun waktu tahun 2022.
3. Analisis sentimen akan dilakukan pada komentar dari situs berita kriminal detik.com menggunakan metode Naive Bayes Classifier.
4. Komentar harus jelas dan dalam kata-kata yang sopan.
5. Komentar wajib dapat dikategorikan kedalam 3 kategori yang terdiri dari kategori positif, negatif dan netral.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengimplementasikan metode pengambilan data berita dari situs Detik.com menggunakan teknik web scraping.

2. Mendeskripsikan proposalkan metode Metodologi klasifikasi Naive Bayes dalam penelitian sentimen pada berita Detik.com yang telah diambil.

1.5 Manfaat Penelitian

Diharapkan penelitian ini dapat memberikan kontribusi, baik dalam aspek teoritis maupun praktis, antara lain:

1.5.1 Manfaat Teoritis:

1. Memperluas pemahaman tentang penggunaan web scraping: Penelitian ini memberikan kontribusi dalam memperluas pemahaman tentang metode web scraping dalam pengumpulan data komentar pada berita kriminal di situs Detik.com. Hasil penelitian ini dapat menjadi referensi bagi peneliti lain yang tertarik dengan penggunaan web scraping dalam analisis sentimen.
2. Memperdalam pemahaman tentang analisis sentimen: Penelitian ini berkontribusi dalam memperdalam pemahaman tentang analisis sentimen, terutama dalam konteks pendapat dan tanggapan masyarakat terhadap berita kriminal di situs Detik.com. Dengan menerapkan metode Naive Bayes Classifier, penelitian ini memberikan wawasan tentang klasifikasi sentimen yang dapat digunakan dalam studi sentimen selanjutnya.

1.5.2 Manfaat Praktis:

1. Menyediakan pemahaman tentang respons masyarakat terhadap berita kriminal: Penelitian ini memberikan pemahaman yang lebih baik tentang sentimen masyarakat terhadap berita kriminal di situs Detik.com. Informasi ini dapat digunakan oleh pihak berwenang, media, dan masyarakat untuk mendapatkan

gambaran yang lebih jelas tentang persepsi dan tanggapan terhadap fenomena kriminalitas yang terjadi.

2. Meningkatkan kesadaran tentang isu-isu kejahatan: Melalui analisis sentimen pada komentar berita kriminal di situs Detik.com, penelitian ini dapat memberikan informasi yang lebih akurat tentang pandangan masyarakat terhadap isu-isu kejahatan. Hal ini dapat meningkatkan kesadaran publik tentang isu-isu kejahatan yang ada dan membantu dalam pengembangan strategi pencegahan dan penanggulangan kejahatan yang lebih efektif.
3. Menyediakan informasi yang lebih baik kepada pembaca: Hasil penelitian ini dapat digunakan oleh media, terutama situs Detik.com, untuk memahami dan memperbaiki kualitas berita kriminal yang disajikan kepada pembaca. Dengan memahami sentimen pembaca terhadap berita kriminal, media dapat menghasilkan konten yang lebih relevan dan bermanfaat bagi masyarakat.

1.6 Metodologi Penelitian

Dalam penelitian ini, digunakan metodologi penelitian yang meliputi

1. Data Selection
2. Data Preprocessing
3. Transformasi
4. Penambangan data
5. Penafsiran serta Evaluasi

1.7 Sistematika Penulisan

Struktur penulisan proposal ini mencakup tiga bab dengan urutan sebagaimana berikut ini:

BAB 1 PENDAHULUAN

Pada bagian ini, akan dijelaskan mengenai latar belakang penelitian, perumusan masalah, tujuan penelitian, manfaat penelitian, dan tata cara penulisan selanjutnya.

BAB 2 TINJAUAN PUSTAKA

Pada bagian ini, akan diulas teori-teori dan studi sebelumnya yang memiliki relevansi dengan penelitian ini. Topik yang akan diulas meliputi web scraping, analisis sentimen, dan metode Naive Bayes Classifier.

BAB 3 METODOLOGI PENELITIAN

Dalam bab ini akan dijelaskan metode yang akan diimplementasikan dalam penelitian ini, termasuk rancangan penelitian, proses pengumpulan data, dan langkah-langkah penilaian sentimen dengan penerapan klasifikasi teknik Naive Bayes Classifier.

1.8 Jadwal Penelitian

Rencana kerja penelitian untuk penelitian ini adalah sebagai berikut:

Tabel 1.1 Jadwal Penelitian

No	Kegiatan	Bulan ke -											
		1	2	3	4	5	6	7	8	9	10	11	12
1	Data Selection												
2	Data Preprocessing												
3	Transformation												
4	Data Mining												
5	Interpretation/Evaluation												

BAB 2
LANDASAN TEORI

2.1 Analisis Sentimen (Sentiment Analysis)

Pengertian analisis sentimen merupakan bagian penelitian yang mengkaji pendapat, perasaan, penilaian, sikap dan perasaan orang terhadap suatu entitas dan karakteristiknya yang diungkapkan dalam teks tertulis (Sari & Wibowo, 2019). Entitas yang relevan dapat berupa produk, layanan, organisasi, orang, peristiwa, masalah, atau topik tertentu. Menurut (Ruslim et al., 2019), analisis sentimen digunakan dalam beberapa konteks sebagai berikut: a) Pendukung keputusan seperti memutuskan buku mana yang akan dibeli, hotel mana yang akan dikunjungi atau film yang akan ditonton. b) Analisis pendapat digunakan dalam dunia usaha untuk mengukur pendapat individu tentang suatu produk. Contohnya adalah pencarian produk Google. c) Analisis prakiraan dan tren: Dengan melakukan analisis sentimen, Anda dapat memprediksi perkembangan pasar dengan mengamati opini publik. Teknik yang umum digunakan untuk menerapkan analisis sentimen dan riset opini saat ini adalah klasifikasi (Singgalen, 2023). Adapun (Ruslim et al., 2019) menjelaskan bahwa dua pendekatan umum digunakan untuk mengklasifikasikan emosi, yaitu kosa kata subjektif dan pembelajaran mesin. Kosakata subyektif adalah sekumpulan kata dengan titik-titik yang menunjukkan sifat teks yang positif, negatif, netral dan obyektif. Dalam pendekatan ini, skor subyektif dari kata-kata yang relevan dihitung secara terpisah untuk bagian teks yang dipertimbangkan. Sebaliknya, pendekatan pembelajaran mesin melibatkan klasifikasi otomatis menggunakan fitur tekstual yang diekstraksi dari teks. Metode machine learning ini terdiri dari dua jenis yaitu supervised dan unsupervised (Hikmawan et al., 2020). menjelaskan

bahwa proses klasifikasi analisis sentimen meliputi pengumpulan data, preprocessing data, pemodelan, validasi, dan produksi output. Data dapat dikumpulkan dari berbagai sumber seperti Twitter dengan menggunakan kata kunci tertentu. Pada tahap preprocessing, data disiapkan agar siap digunakan pada tahap pemodelan. Semakin baik langkah preprocessing dilakukan, semakin tinggi tingkat akurasi langkah pemodelan (Ruslim et al., 2019). Dalam analisis sentimen, langkah preprocessing biasanya meliputi penandaan, penghapusan stopword, dan detangling. Pada tahap pemodelan, data digunakan sebagai data pelatihan untuk algoritma klasifikasi tertentu. Pada tahap evaluasi, kinerja model yang dibuat dievaluasi. Estimasi ini biasanya memberikan nilai akurasi model. Jika akurasi belum mencapai target yang diinginkan, model dievaluasi kembali untuk meningkatkan akurasi (Hikmawan et al., 2020). Merujuk pada pendapat para ahli di atas, dapat disimpulkan bahwa analisis emosi melibatkan identifikasi pendapat, perasaan, penilaian, sikap dan perasaan seseorang terhadap suatu satuan dan ciri-cirinya dalam bentuk teks tertulis. Pendekatan umum untuk analisis sentimen adalah dengan menggunakan metode klasifikasi seperti leksikon subyektif dan pendekatan pembelajaran mesin. Proses analisis sentimen meliputi langkah-langkah pengumpulan data, pengolahan data, pemodelan, validasi dan evaluasi hasil (Singgalen, 2023).

2.2 Berita

2.2.1 Pengertian Berita

Beberapa ahli telah menjelaskan pengertian teks berita. Menurut (Alfianistiawati, 2021), berita adalah informasi nyata tentang fakta dan opini yang menarik perhatian masyarakat. Berita adalah sumber informasi sosial dalam masyarakat yang terus berkembang. Informasi ini biasanya disampaikan dalam bentuk berita. Namun, tidak semua orang bisa menulis berita sesuai dengan kebutuhannya. Dalam pernyataan yang

sama (Alexander & Firza, 2023) menjelaskan bahwa berita adalah laporan atau pemberitahuan tentang suatu peristiwa atau keadaan yang bersifat umum dan baru saja terjadi. Berita yang dilaporkan oleh wartawan melalui media. Kemampuan menulis berita yang baik merupakan modal utama seorang penulis berita. Oleh karena itu, laporan yang akan disampaikan harus jelas dan dapat dipahami oleh pembaca atau pendengarnya. (Alwasi'a, 2020) juga memberikan pengertian bahwa berita adalah laporan tercepat dari gagasan atau fakta terkini yang benar, menarik dan relevan bagi sebagian besar khalayak di media arus utama seperti surat kabar, radio, televisi atau media online. Oleh karena itu, berita dapat disampaikan melalui berbagai saluran media baik secara lisan maupun tulisan. (Suryadi et al., 2023) menjelaskan bahwa berita adalah hasil dari proses yang kompleks dalam menyusun dan menentukan peristiwa dan topik yang termasuk dalam kategori tertentu. Tidak semua peristiwa dapat dijadikan berita, sehingga perlu dicari tahu peristiwa atau fakta nyata yang dapat dikelompokkan menurut topik tertentu. Pendapat senada disampaikan oleh (Yuniwati et al., 2021)bahwa menulis berita merupakan keterampilan yang memerlukan pemikiran, karena ada unsur 5W 1H yang harus dikembangkan menjadi berita dari beberapa paragraf. Unsur-unsur tersebut menjawab pertanyaan apa (apa yang terjadi), siapa (siapa yang terlibat dalam peristiwa itu), mengapa (mengapa peristiwa itu terjadi), di mana (di mana peristiwa itu terjadi), kapan (kapan terjadinya) dan bagaimana (bagaimana atau bagaimana terjadinya).) telah terjadi). Dalam dunia pendidikan seringkali siswa belum sepenuhnya memahami cara menulis berita dengan memperhatikan unsur 5W 1H tersebut. (Alwasi'a, 2020) menyatakan bahwa headline berita harus berkaitan dengan fakta yang terkandung dalam berita tersebut. Isi berita utama harus logis, masuk akal, dan dapat diterima secara logis. Pemilihan judul yang tepat juga penting saat menulis teks berita, karena judul harus mencerminkan isi

berita secara keseluruhan. Berdasarkan penjelasan di atas dapat disimpulkan bahwa berita adalah informasi tentang peristiwa atau hal baru yang disajikan dalam berbagai bentuk seperti cetak, siaran, di internet atau secara lisan. Jurnalis mengumpulkan berita dengan memperhatikan fakta, opini, dan peristiwa yang menarik perhatian publik. Berita juga dapat mempengaruhi pembacanya terhadap hal-hal yang dikandungnya. Saat menulis berita, penting untuk menjaga keakuratan dan aktualitas fakta dan memilih judul yang sesuai dengan isi berita, yang mencerminkan informasi umum (Hardjo, 2019).

2.2.2 Berita Online

Perkembangan zaman menyebabkan munculnya media baru. Media lama cenderung mengalami perubahan eksistensi, meski di sisi lain juga menyambut media baru sebagai bentuk perkembangan. Media tradisional cetak dan elektronik kini telah bertransformasi menjadi bentuk digital yang sering disebut portal berita online atau media online(Alexander & Firza, 2023). Situs berita online ini menyajikan informasi terkini setiap hari tentang berbagai peristiwa atau kejadian yang berkaitan dengan kehidupan sehari-hari seperti pendidikan, olahraga, teknologi, politik, dan gaya hidup sehat (Suryadi et al., 2023).

2.3 Detik.com

2.3.1 Pengertian Detik.com

Dimulai oleh empat orang, Budiono Darsono, Yayan Sopyan, Abdul Rahman dan Didi Nugrahad, www.detik.com diluncurkan pada 9 Juli 1998. Situs web ini independen dari media cetak yang ada dan beroperasi sebagai media online independen. Berbeda dengan media online generasi pertama, www.detik.com menghadirkan jenis berita baru yang ringkas dan to the point. Untuk menjaga kecepatan, berita detik.com tidak selalu sesuai dengan unsur 5W 1H seperti dalam jurnalisme tradisional (Alwasi'a, 2020). Budiono

menyajikan model “rolling news”, yaitu menggunakan cara penyajian berita secara berurutan, seperti stasiun berita CNN atau kantor berita luar negeri seperti AP, AFP atau Reuters. Konsep tersebut diterima dengan baik oleh masyarakat, dengan pergeseran konsumsi berita dari media cetak ke media online (Alwasi’a, 2020).

2.4 *Data Mining*

Data Mining adalah proses yang menggunakan berbagai teknik dan alat data untuk mencari hubungan dan pola yang tersembunyi, serta meringkas data dan mengekstrak informasi yang sebelumnya tidak diketahui (Hartama et al., 2019). Dalam *data mining*, model data tersembunyi dianalisis dari berbagai perspektif untuk mendapatkan informasi yang bermanfaat, yang kemudian dikumpulkan di ruang umum seperti gudang data untuk dianalisis dan informasi lebih lanjut (Arhami et al., 2020). Proses *data mining* melibatkan penerapan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin untuk mengekstraksi dan mengidentifikasi nilai data penting dari basis data besar (Hartama et al., 2019). Penambangan data membantu mengekstraksi informasi yang bermanfaat pada repositori basis data superior dan bisa dianggap menjadi proses penggalian informasi anyar mengenai data baru untuk mendukung pengambilan keputusan. Penambangan data juga dikenal sebagai penambangan data, yang merupakan bagian penting dari proses penemuan data basis data (KDD). Proses ini mengubah data mentah menjadi model data yang menarik dan memberi pengguna informasi yang diperlukan seperti informasi (Arhami et al., 2020).

2.5 *Web Scraping*

Web Scraping, juga dikenal sebagai pengutipan data web, adalah ekstraksi data dari situs web yang ada atau teknik mengekstraksi data dari situs web. Tangkapan web melibatkan

perayapan dengan mengamati dokumen HTML di situs web tempat Anda ingin mengambil data, lalu menandai bagian HTML yang berisi data yang diinginkan untuk digunakan dalam aplikasi tangkapan web yang dibuat (Fahrudin et al., 2023). Prosedur *web scraping* dilaksanakan dengan mengutip arsip semi-terstruktur sebagaimana HTML maupun XHTML, menganalisisnya dan mengekstraksi informasi dari halaman tersebut untuk tujuan lain. Perlu dicatat bahwa web scraping berbeda dengan data mining lantaran data mining melibatkan penggalian informasi untuk mengenali pola atau arah semantik pervolume informasi yang besar (Suryadi et al., 2023). Aplikasi penangkap web maupun agen cerdas, otomatis, atau otonom berfokus pada cara mendapatkan informasi melalui pengumpulan data (Wibowo et al., 2019). Proses pengikisan web melibatkan beberapa langkah seperti yang dijelaskan oleh (Nurkholis et al., 2023):

1. Mewujudkan model *scraping*: Langkah-langkah ini melacak arsip HTML laman web untuk di *scraping* maupun di *scrap*. Informasi yang Anda cari dapat diidentifikasi dengan kode HTML.
2. Menjelajahi Navigasi Situs Web: Langkah ini melibatkan navigasi situs web untuk mendapatkan informasi yang dapat dimuat atau disalin untuk disalin di pengikis situs web yang diciptakan.
3. Mengotomatiskan pengarah serta ekstraksi data: Dengan mempertimbangan data yang diperoleh dalam langkah pertama dan kedua, *scraping* laman web dikerjakan yang mengotomatiskan ekstraksi data dari situs web tertentu.
4. Penggalian data serta penyimpanan historis: Data yang diperoleh dari prosedur 3 terarsip pada daftar maupun database sebagai hasil scraping.

2.6 Text Mining

2.6.1 Proses Text Mining

Proses Text Mining Data yang diproses pada ruang text mining terdiri dari kelompok besar teks atau berkas yang disebut korpus. Proses penambangan teks dijelaskan oleh (Mahariani & Nurmalasari, 2022) memiliki lima langkah utama:

1. *Text Pre-Processing*

   Dalam langkah ini, kami menggali korpus untuk arti yang benar. Tujuan utama dari langkah ini adalah menyiapkan data teks agar dapat diproses. Langkah ini melibatkan pelipatan huruf untuk membuat semua kata menjadi huruf kecil dan menghapus karakter lain yang dianggap sebagai pemisah dalam korpus.

   Setelah pembungkus dilipat, tokenisasi dan rooting dilakukan. Tokenisasi memecah kata-kata dalam dokumen menjadi kata-kata independen. Kata-kata yang sebenarnya kemudian diubah menjadi kata-kata dasar. Namun, tahapan sebenarnya tidak selalu digunakan karena bisa membingungkan dan tidak cukup spesifik untuk mewakili arti sebenarnya dari kata yang dimodifikasi (Mahariani & Nurmalasari, 2022).

2. *Features Generation*

   Setelah pelabelan dan penghapusan teks pada tahap preprocessing, langkah selanjutnya adalah mereduksi komponen kata-kata pada konteks korpus yang diidentifikasi stopwords. Stopwords adalah frasa yang bukan berarti atau tidak edukatif tetapi sering nampak dalam teks. Stop word didokumentasikan dalam daftar stop yang berbeda untuk setiap bahasa. Sesudah menempuh fase stopwords, fase ini pun bertujuan demi memperoleh representasi korpus yang dikehendaki (Mahariani & Nurmalasari, 2022).

Prosedur yang umum diaplikasikan yaitu bentuk *bag of words*. *bag-of-words model* mewaliki korpus sebagai kumpulan kata, yang kemudian dijumlahkan berdasarkan kemunculannya. Dalam representasi *bag of words*, seluruh kata direpresentasikan sebagai unsur terpisah yang nilai numeriknya dihitung menggunakan bobot.

Representasi korpus diidentifikasi dalam bentuk matriks yang disebut term document matrix. Istilah Document Matrix mengacu pada sekumpulan dokumen yang digunakan dalam proses klasifikasi dokumen. Setiap arsip diwakili sebagai satu set aspek dan dapat digambarkan selaku $D\ i$ = [$wi1wi2\ \ldots\ wit\ \ldots\ wnk$], di mana $D\ i$ adalah dokumen D dan adalah s. muncul di dokumen i. Matriks ini dipenuhi dengan nilai kehadiran kata tersebut. Dalam term matriks dokumen, baris adalah data dari dokumen, sedangkan kolom adalah fungsi yang digunakan. Jika ditulis dalam bentuk matriks (Mahariani & Nurmalasari, 2022).

3. Features Selection

Langkah ini terus mengurangi dimensi. Meskipun pada langkah sebelumnya menghilangkan kata-kata nondeskriptif (stop words), tidak semua kata dalam dokumen memiliki makna yang signifikan. Oleh karena itu, untuk mengurangi dimensi, pemilihan fitur dilakukan hanya untuk kata-kata yang bermakna dan secara signifikan mewakili isi dokumen. Kata-kata yang dianggap penting dipilih berdasarkan intensitas kemunculannya dan keinformatifan seluruh korpus (Asgarnezhad et al., 2021).

4. Analisis teks

Dua pendekatan umum yang digunakan dalam text mining, yaitu pengelompokan dan klasifikasi. Klasifikasi membagi dokumentasi menjadi dua

set data: data pelatihan dan data uji. Tujuan klasifikasi adalah untuk menemukan pola yang tidak dapat dilihat pada korpus. Data training digunakan untuk membangun model, sedangkan data test digunakan untuk memvalidasi model (Asgarnezhad et al., 2021).

2.7 Analisis Sentimen atau Opinion Mining

Analisis sentimen, yang juga dikenal sebagai penambangan pendapat, merupakan bidang penelitian yang memfokuskan pada kajian pendapat, interpretasi, penilaian, sikap, dan perasaan individu terhadap berbagai objek, seperti produk, layanan, organisasi, individu, subjek, masalah, dan karakteristiknya (Suryadi et al., 2023). Penelitian dalam analisis opini dan jajak pendapat mengalami perkembangan pesat, didorong oleh beberapa faktor (Herdhianto, 2020):

1. Kemajuan prosedur pembelajaran mesin pada pemrosesan bahasa bawaan dan informasi.
2. Tersedianya data yang memadai untuk algoritma pembelajaran mesin, terutama melalui pelatihan online dan pengembangan situs agregasi ulasan.
3. Kesadaran tentang tantangan intelektual menarik yang muncul dari aplikasi cerdas dan bisnis.

Ungkapan "analisis sentimen" awalnya digunakan dari (Purnamawati, 2021), sementara ungkapan "opinion mining" diperkenalkan dari (Darmawan, 2022). Dalam penelitian (Herdhianto, 2020)menerapkan teknik pemrosesan bahasa alami untuk melakukan klasifikasi emosi secara online. Sementara itu, (Purnamawati, 2021) memperkenalkan alat jajak pendapat yang mengumpulkan opini tentang topik tertentu

dan mengklasifikasikannya berdasarkan analisis subyektif. (Darmawan, 2022) telah mengidentifikasi masalah khusus serta meningkatkan metode otomatis untuk memilah antara ulasan pro dan kontra.

2.7.1 Model Opinion Mining

Model opinion mining merupakan pendekatan yang digunakan untuk menganalisis opini atau sentimen dalam teks. Opini dapat diungkapkan tentang berbagai hal, seperti produk, layanan, topik, individu, organisasi, atau acara. Entitas yang dikenali dalam analisis ini disebut sebagai "objek". Objek ini memiliki komponen atau bagian dan atribut. Komponen dapat terdiri dari subkomponen dan atributnya sendiri (Purnamawati, 2021). Dalam model opinion mining, istilah "fitur" digunakan untuk merujuk pada komponen dan atribut tersebut. Sebuah kalimat dapat mengandung opini tentang lebih dari satu fitur. Sebagai contoh, pada kalimat "kualitas gambar kamera ini bagus, tetapi masa pakai baterainya pendek," "ketajaman gambar" dan "masa pakai baterai" yaitu fitur atau atribut dari objek yang disampaikan (Darmawan, 2022).

Perlu dicatat bahwa adanya dokumen positif tentang suatu objek tidak selalu berarti bahwa pemilik opini memiliki pandangan positif secara keseluruhan terhadap semua sudut atau karakteristik dari bagian tersebut. Sebaliknya, arsip negatif tidak selalu berpengaruh bahwa pemilik opini merasa tidak menyukai seluruh hal tentang objek tersebut (Herdhianto, 2020).

Berdasarkan model survei opini (Darmawan, 2022), terdapat tiga tugas utama dalam analisis opini:

1. Identifikasi karakteristik objek: Tugas ini mencari karakteristik yang menjadi fokus opini, seperti "kualitas gambar" dalam contoh sebelumnya.
2. Mengatur orientasi pendapat: Tugas ini bertujuan untuk mengatur apakah pendapat tentang suatu elemen bersifat pro, kontra dan tidak memihak. Pada contoh sebelumnya, opini tentang "kejernihan gambar" menyandang pendekatan yang positif.
3. Klasifikasi sinonim: Tugas ini mencari kata atau frase berbeda yang digunakan untuk mengungkapkan karakteristik yang sama.

2.7.2 Klasifikasi Sentimen

Masalah klasifikasi emosi telah menjadi fokus penelitian akademik, di mana pengklasifikasian emosi dan subjektivitas dapat dianggap sebagai dua masalah klasifikasi yang terpisah. Terdapat dua aspek yang relevan dalam hal ini (Purnamawati, 2021):

1. Klasifikasi arsip opini menjadi opini pro dan kontra.
2. Klasifikasi frasa atau kalimat sebagai subjektif atau objektif, serta mengklasifikasikan frasa atau kalimat subjektif sebagai positif, negatif, atau netral.

Meskipun terkadang istilah yang berbeda digunakan untuk menggambarkan dua aspek tersebut, tujuan klasifikasi ini tetap sama, yaitu untuk menentapkan orientasi sebuah kalimat atau arsip. Rincian analisis pendapat (Herdhianto, 2020) bisa diuraikan seperti dibawah ini:

1. Klasifikasi Sentimen / Analisis Sentimen Tingkat Dokumen: Klasifikasi sentimen yaitu langkah untuk mengukur apakah suatu arsip mengungkapkan sentimen pro atau kontra. Tahap demikian pula sering disebut sebagai klasifikasi sentimen tingkat dokumen.
2. Klasifikasi Subjektivitas / Analisis Sentimen Tingkat Kalimat: Klasifikasi subjektivitas, juga dikenal sebagai klasifikasi subjektivitas tingkat kalimat, adalah langkah guna mengidentifikasi jika sebuah kalimat memiliki sudut pandang pribadi atau tidak pribadi, serta mengidentifikasi pendapat yang diungkapkan. Klasifikasi subjektivitas melibatkan dua tugas:

   a. Memastikan apakah kalimat itu memiliki sudut pandang subyektif atau obyektif.

   b. Jika kalimat bersifat subjektif, menentukan apakah kalimat tersebut mengungkapkan pendapat positif atau negatif.

Terdapat beberapa metode yang digunakan untuk mengklasifikasikan emosi. Secara umum, metode-metode ini dapat dibagi menjadi metode supervised dan unsupervised. Penelitian ini menggunakan metode supervised yang memerlukan data pelatihan dan uji. Dua label kelas, yaitu positif atau negatif, telah ditentukan sebelumnya, dan klasifikasi emosi dianggap sebagai masalah pembelajaran yang diawasi. Menurut (Suryadi et al., 2023), beberapa fitur yang digunakan dalam pembelajaran mesin antara lain:

1. *Terms* dan Frekuensi: Pendekatan ini mirip dengan pencarian informasi, di mana teks direpresentasikan sebagai vektor dengan setiap entri yang sesuai dengan istilah atau n-gram.
2. *Part-of-Speech Tagging: Part-of-Speech* (POS) tagging adalah sistem yang secara otomatis memberikan label kata-kata dalam sebuah kalimat. Misalnya, kalimat "Saya minum jamu" akan memiliki tag PRP (kata ganti orang) untuk "Saya", VBT (kata kerja transitif) untuk "minum", dan NN (kata benda umum) untuk "jamu". Sistem ini menerima kalimat sebagai input dan mengeluarkan keluaran dengan kata-kata yang telah ditandai.
3. Opinion Words and Expressions: Opinion Words adalah kata-kata yang umumnya digunakan untuk mengungkapkan perasaan positif atau negatif. Dalam implementasinya, POS tagging sering digunakan untuk memilah kata-kata opini.
4. *Syntactic* / Sintaks: Sintaksis memainkan peran penting dalam opinion mining, terutama dalam mendeteksi subjektivitas. Studi sebelumnya telah menganalisis informasi sentimental, termasuk hubungan sintaksis dan struktur dokumen, dengan membuat kumpulan besar kalimat dari sumber online Jepang dengan tujuan menciptakan leksikon terowongan(Suryadi et al., 2023).

2.8 Algoritma Naïve Bayes

Naive Bayes merupakan metode klasifikasi yang menggunakan probabilitas dan statistik, diperkenalkan oleh seorang ilmuwan Inggris bernama Thomas Bayes. Metode ini digunakan untuk melakukan prediksi masa depan berdasarkan data dari masa lalu (Darmawan, 2022). Dalam Naïve Bayes, probabilitas untuk setiap kelas dihitung dengan

syarat bahwa kelas tersebut adalah benar, berdasarkan vektor informasi objek yang diamati (Herdhianto, 2020). Metode ini mengasumsikan bahwa atribut objek yang digunakan adalah independen. Proses perhitungan probabilitas untuk memperoleh perkiraan akhir melibatkan jumlah frekuensi dari tabel keputusan (Normawati & Prayogi, 2021).

Dalam teorema Bayes, probabilitas dapat dinyatakan sebagai berikut: P (H | X) = (P(X | H) * P(H)) / P(X). Di mana X adalah pembuktian, H adalah hipotesis, P (H | X) adalah probabilitas bahwa hipotesis H benar dengan syarat X, atau dengan kata lain, P (H | X) merupakan probabilitas posterior dari H dengan syarat X. P(X | H) adalah probabilitas posterior dari X dengan syarat H. P (H) adalah probabilitas prior hipotesis H, dan P(X) adalah probabilitas prior dari bukti X (Sari & Wibowo, 2019).

Alur kerja Naive Bayes, sebagaimana dijelaskan oleh (Mahariani & Nurmalasari, 2022), meliputi langkah-langkah berikut:

1. Membaca data training.
2. Menghitung jumlah peluang untuk setiap variabel, dan mencari nilai probabilitasnya.
3. Memperoleh nilai dalam tabel mean, standar deviasi, dan probabilitas.

2.9 Knowledge Discovery in Database (KDD)

Proses penemuan informasi penting berdasarkan kumpulan data yang jumlahnya sangat besar. Seringkali dalam proses data mining menerapkan berbagai metode probabilitas dan statistik, database systems, hingga memanfaatkan artificial intelligence,

dan machine learning (Suyanto, 2017). Banyak sekali permasalahan data mining yang dapat ditemukan dalam kehidupan sehari-hari, seperti transaksi bank, permainan saham, penjualan barang, dan sebagainya. Hal tersebut disebabkan karena data-data yang terus bertambah banyak dan semakin kompleks setiap harinya, sehingga data-data tersebut tidak dapat ditangani dengan baik, Untuk menyelesaikan permasalahan tersebut, diperlukan proses data mining secara bertahap, mulai dari selection, preprocessing, transformation, hingga knowledge.

2.10 Penelitian Sebelumnya

Di bawah ini disajikan tabel yang berisi penelitian terdahulu yang dapat dijadikan referensi sebagai acuan dalam penelitian ini:

Penelitian ke-1	
Judul	ANALISIS SENTIMEN PADA REVIEW APLIKASI BERITA ONLINE MENGGUNAKAN METODE *MAXIMUM ENTROPY*
Tahun	2020
Penulis	Annisa Alwasi'a
Ringkasan	Perkembangan teknologi telah berdampak signifikan pada masyarakat, terutama melalui internet sebagai pendorong kemajuan teknologi informasi dan komunikasi. Detikcom, sebagai portal berita online terkemuka di Indonesia, berupaya mempertahankan dan meningkatkan performanya dengan memperhatikan evaluasi khalayak terhadap layanan dan berita yang

	disuguhkan. Dalam pengkajian ini, analisis sentimen digunakan untuk mengkelompokan ulasan user Detikcom menjadi sentimen positif atau negatif.
Kesimpulan	Hasil analisis menggunakan algoritma Maximum Entropy menunjukkan akurasi tinggi, 95,69%, dengan kinerja sistem mengklasifikasikan kelas positif 97,45% dan kelas negatif 86,17%. Hasil analisis memberikan wawasan tentang topik yang sering dibahas oleh pengguna, dengan kata-kata positif seperti "bagus," "baik," "oke," "update," dan "mantap," serta kata-kata negatif seperti "tidak," "buruk," "iklan," "Detik," dan "oprasi." Informasi ini berguna untuk menilai kelebihan dan kesenjangan aplikasi Detikcom, serta ulasan negatif dapat dijadikan pertimbangan dalam pemecahan masalah dengan diagram fishbone.
Penelitian ke-2	
Judul	SENTIMENT ANALYSIS MENGGUNAKAN NAÏVE BAYES CLASSIFIER (NBC) PADA TWEET TENTANG ZAKAT
Tahun	2020
Penulis	Adhyaksa Herdhianto
Ringkasan	Twitter menjadi basis data gagasan dan sentimen khalayak yang dapat dimanfaatkan untuk studi kemasyarakatan. Salah satu isu kemasyarakatan yang

	menjadi perhatian utama di negara-negara bertumbuh, terlingkup Indonesia, adalah kemelaratan. Pengentasan kemelaratan dapat dilakukan melalui zakat, akan tetapi potensi zakat di Indonesia belum optimal. Oleh karena itu, pengkajian ini bertujuan untuk menelaah sentimen masyarakat mengenai zakat melalui Twitter memanfaatkan Algoritma Naïve Bayes Classifier (NBC) lewat pemilahan fitur Term-Frequency
Kesimpulan	Hasil kategorisasi sentimen dari 50 tweet pada informasi uji menunjukkan dominasi sentimen positif dibandingkan sentimen negatif maupun netral. Hal ini terjadi karena metode Lexicon Based memiliki lebih banyak kata positif dalam kamusnya dibandingkan kata negatif. Data tweet tentang zakat diambil dari query zakat, zakat fitra, dan zakat fitrah pada tanggal 04 Juni 2019. Hasil penelitian menunjukkan bahwa akurasi Algoritma Naïve Bayes Classifier dengan seleksi fitur Term-Frequency mencapai 74%, dan presisi sentiment positif sebesar 79,3% serta sentiment negatif sebesar 66,7%.
Penelitian ke-3	
Judul	ANALISIS SENTIMEN PENGGUNA APLIKASI WHATSAPP DENGAN ALGORITMA MACHINE LEARNING CLASSIFIER BERBASIS SMOTE

Tahun	2021
Penulis	Annida Purnamawati
Ringkasan	Maksud pengkajian ini ialah untuk mengamati serta mengkaji bagaimana orang mengkritik suatu situasi dengan label positif atau negatif serta mengekspresikan agresi mereka terhadap aplikasi WhatsApp. Penelitian ini juga membahas permasalahan data imbalance pada beberapa ulasan tentang aplikasi WhatsApp dan mencari hasil ketepatan dari analisis sentimen menggunakan algoritma *Decision Tree, K-Nearest Neighbor,* beserta *Support Vector Machine.*
Kesimpulan	Algoritma SVM dengan 2-Gram (bigram) dan SMOTE menghasilkan akurasi sebesar 89.09%, f1 score 89.09%, precision 89.13%, recall 89.09%, serta auc score 0.89, menunjukkan klasifikasi yang baik. Penggunaan teknik partitioning + SMOTE pada algoritma kategorisasi teruji mengoptimalkan akurasi algoritma SVM, meskipun tidak secara konkret. Dengan demikian, boleh dirangkum bahwa implementasi optimasi terunggul untuk model ini merupakan algoritma SVM beserta N-gram + SMOTE, yang dapat menjadi solusi untuk permasalahan klasifikasi sentimen pada aplikasi WhatsApp.
Penelitian ke-4	

Judul	IMPLEMENTASI WEB SCRAPING DAN SENTIMENT ANALYSIS TERHADAP BERITA MENGGUNAKAN MACHINE LEARNING
Tahun	2023
Penulis	Ary Suryadi, Wahid Andika Syb'an, Nazzala Alfa'inna, Eni Heni Hermaliani
Ringkasan	Penelitian ini bertujuan mengembangkan aplikasi buat mengutip berita pada instrumen online dan mengekstrak sentimen tentang teks warta tersebut. Aplikasi ini juga menciptakan acuan machine learning yang dapat mengklasifikasikan berita selaku otomatis dengan akurasi tinggi. Metode web scraping digunakan untuk mengumpulkan data berita dari berbagai sumber online, beserta versi *machine learning* digunakan bagi analisis sentimen. Bukti penelitian berupa 100 berita terkait "Universitas Nusa Mandiri" menurut bulan Desember 2022.
Kesimpulan	Pada penelitian ini, web scraping diimplementasikan dengan memanfaatkan instrumen pelacak Google News dan menggunakan kata vital "Universitas Nusa Mandiri". Penilaian analysis yang dipakai adalah *Aspect-Based* Penilaian *Analysis*. Data diolah lewat mencari berita terkait "Universitas Nusa Mandiri" mengenai agenda 01 Desember 2022 hingga 31

	Desember 2022, dan hasilnya ditemukan 100 warta. Dari analisis tersebut, sekitar 87% warta menunjukkan penilaian positif, sedangkan 13% warta menunjukkan penilaian negatif.
Penelitian ke-5	
Judul	ANALISIS SENTIMEN REVIEW PELANGGAN E-COMMERCE DI INDONESIA MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER
Tahun	2022
Penulis	Tito Dwiki Darmawan
Ringkasan	Penelitian ini berfokus pada analisis sentimen review pelanggan di lima e-commerce tertinggi di Indonesia: Lazada, Bukalapak, Blibli.com, Tokopedia, dan Shopee. Tujuannya adalah membantu *e-commerce* dalam memahami sentimen pelanggan untuk meningkatkan layanan. Metode yang digunakan adalah *Naïve Bayes Classifier* untuk mengelompokkan review menjadi sentimen positif dan negatif berdasarkan frekuensi pengalaman sebelumnya.
Kesimpulan	Hasil penelitian menunjukkan Lazada memiliki sentimen positif tertinggi (97.0%), diikuti Bukalapak (94.6%), Shopee (88.5%), Blibli.com (76.1%), dan Tokopedia (34.4%). Akurasi metode *Naïve Bayes* pada masing-masing *e-commerce* adalah 56.23% untuk

	Lazada, 93.05% untuk Bukalapak, 87.82% untuk Shopee, 55.31% untuk Blibli.com, dan 94.94% untuk Tokopedia. Pengkajian ini divalidasi memakai *confusion matrix* beserta 10-*fold cross-validation.*

2.11 Penelitian Sekarang

Penelitian ini bertujuan untuk melakukan analisis sentimen pada berita yang terdapat di situs Detik.com dengan menggunakan metode web scraping dan Naive Bayes Classifier. Langkah pertama penelitian ini adalah mengumpulkan data berita dari situs Detik.com menggunakan metode web scraping, yang memungkinkan pengambilan data secara otomatis dan spesifik dari bagian relevan berupa berita-berita yang menjadi objek analisis sentimen. Proses web scraping dilakukan pada halaman-halaman web yang menggunakan bahasa markup XHTML atau HTML dengan menggunakan bahasa pemrograman Python sebagai alat utama.

Setelah data berita berhasil dikumpulkan, penelitian selanjutnya melibatkan penerapan metode Naive Bayes Classifier digunakan untuk mengkategorikan sentimen berita menjadi baik atau buruk. Hasil analisis sentimen diharapkan dapat memberikan wawasan mengenai pandangan dan opini publik terhadap ulasan pers yang ada pada situs Detik.com. Penelitian ini menjadi relevan mengingat pentingnya pemahaman terhadap sentimen masyarakat terhadap berita dan pemberitaan online.

BAB 3

OBJEK DAN METODOLOGI PENELITIAN

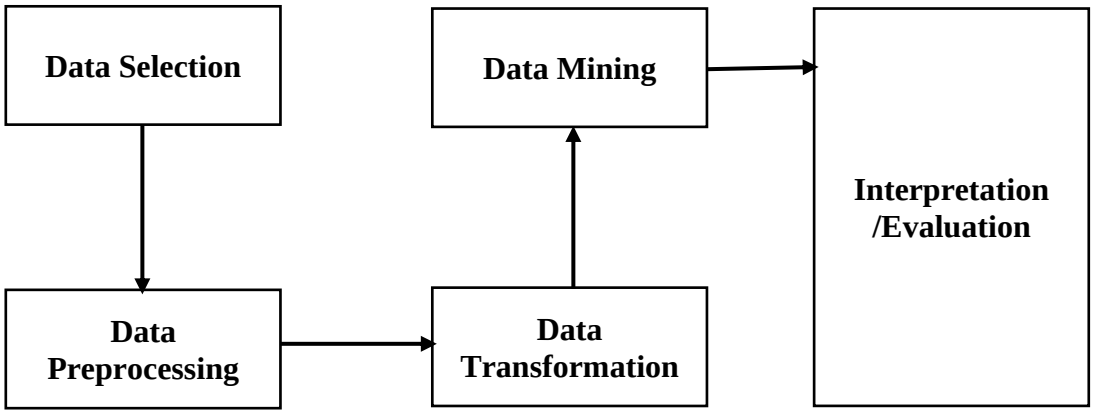
3.1 Objek Penelitian

Materi penelitian ini akan melibatkan pengumpulan data komentar berita kriminal dari situs Detik.com menggunakan web scraping. Data yang diambil akan melibatkan berita-berita terkait kriminal yang dipublikasikan pada tahun 2022.

Berikut adalah sampel berita kriminal dan komentarnya pada situs Detik.com

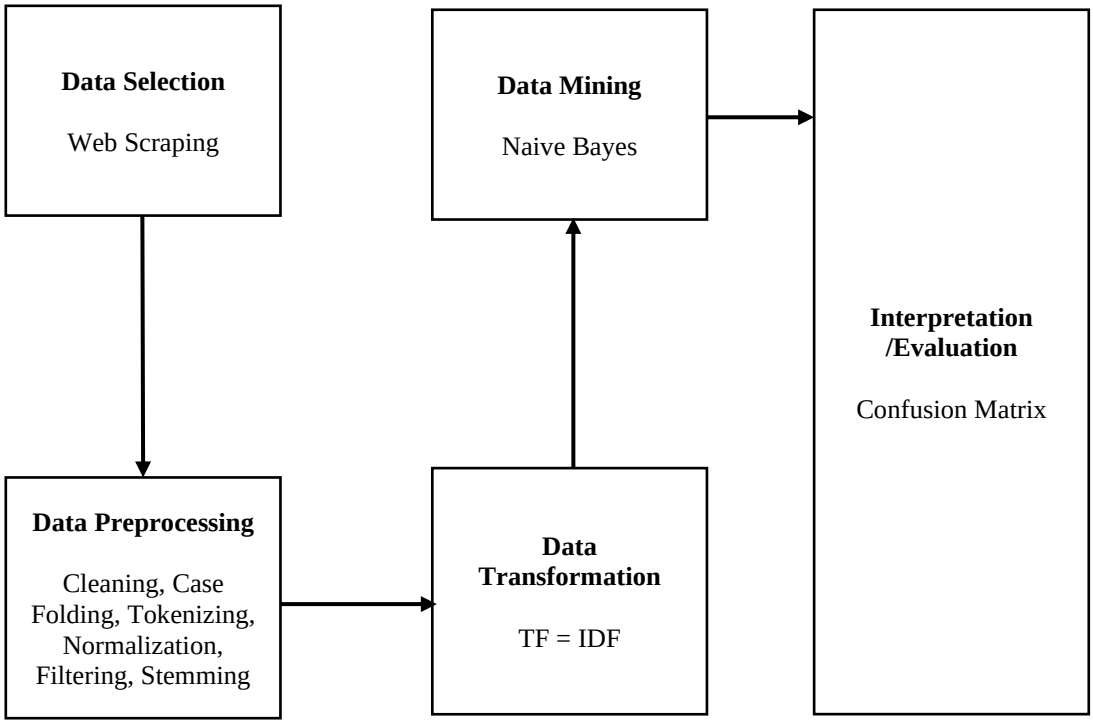
3.2 Metodologi Penelitian

Pendekatan studi yang diterapkan dalam studi ini adalah Knowledge Discovery and Data Mining (KDD), yang melibatkan lima tahap proses, yaitu proses Data Selection, Data Preprocessing, Transformation, Data Mining dan Interpretation/Evaluation (Arhami et al., 2020) .



3.3 Rancangan penelitian

Dalam skripsi ini, Algoritma Naïve Bayes akan diterapkan untuk melakukan analisis sentimen berdasarkan tahapan-tahapan dari metode Knowledge Discovery and Data Mining (KDD).



Selanjutnya, rancangan penelitian akan diuraikan sebagai berikut:

3.3.1 Data Selection

Pada tahap permulaan, eksekusi penggalian data dan pencantuman label data. Data yang akan diambil berasal dari komentar-komentar berita Detik.com yang berfokus pada periode 1 Januari 2022 hingga 31 Desember 2022. Eksekusi pengambilan data dengan teknik web scraping serta bahasa pemrograman Python. Selanjutnya, data ulasan akan dilabeli secara manual dengan dua label, yaitu positif dan negatif, dan akan divalidasi oleh ahli bahasa.

3.3.2 Preprocessing

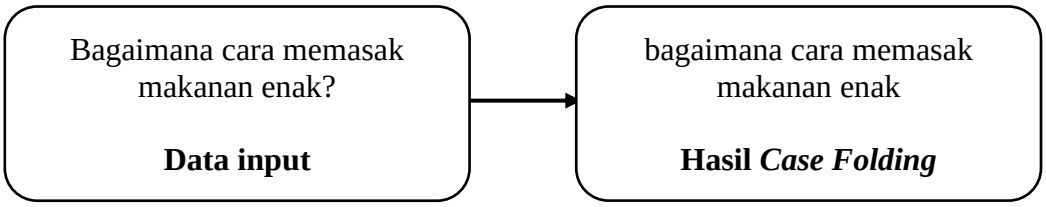
Bab persiapan data pada penelitian ini Proses pemeriksaan arti yang akurat dan analisis tata bahasa untuk memastikan kebenaran susunan teks. Bertujuan untuk menyusun teks agar menjadi data bermutu yang siap untuk proses berikutnya. Berikut ini adalah langkah-langkah tahapan preprocessing yang akan diimplementasikan dalam penelitian ini.

1. **Cleaning**

   Pada tahap cleaning, akan dilakukan penghapusan atribut yang dianggap tidak berpengaruh pada proses klasifikasi, seperti karakter tanda baca, angka, dan emoji. Selanjutnya,

2. **Case Folding**

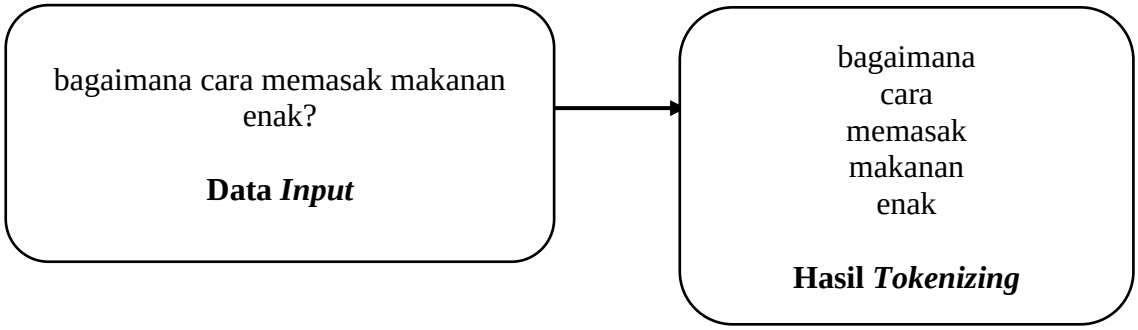
   pada tahap case folding, seluruh bentuk huruf dalam dokumen akan diubah menjadi huruf kecil. Hanya huruf dari 'a' sampai 'z' yang akan diterima sebagai data dokumen. Gambar 3.2 menggambarkan hasil dari penerapan case folding tersebut.



Gambar 3.2 Contoh Dari Penggunaan Case Folding

3. Tokenizing

Pada fase pemisahan kata atau analisis sintaksis, dilakukan proses pemisahan data masukan menjadi kata-kata berdasarkan setiap kata yang membentuknya. Selain itu,



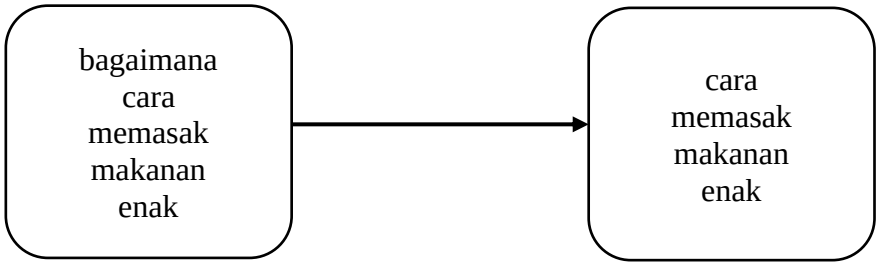
spasi digunakan untuk memisahkan kata-kata. Gambar 3.3 menjelaskan hasil dari penerapan tokenizing tersebut.

Input	Output
Data input setelah dari langkah case folding	Himpunan kata

4. Filtering

Fase filtering merupakan langkah untuk memilih dan menggunakan kata-kata yang penting dari hasil tahap sebelumnya, yaitu tahap tokenizing. Proses filtering dapat dilakukan dengan menerapkan algoritma stoplist (untuk menghilangkan kata-kata yang dianggap tidak penting) atau menggunakan wordlist (untuk menyimpan kata-kata yang penting). Stoplist atau stopword

adalah kata-kata non-deskriptif yang dihapus dalam pendekatan bag-of-words (pendekatan kata demi kata). Beberapa contoh stopword adalah "yang", "dan", "di", "dari", dan lain-lain.



Data Input **Hasil Filtering**

Penjelasan:

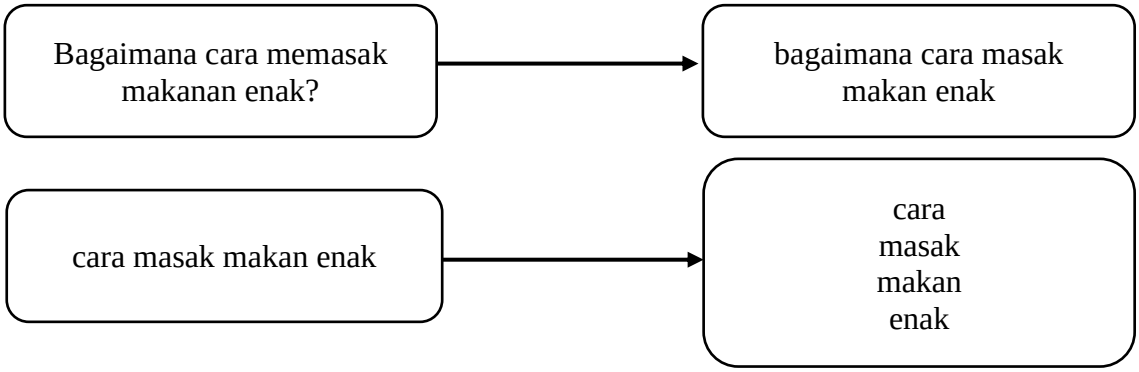
Input	Output
Gabungan kata yang dihasilkan setelah tokenizing	Gabungan term yang akan dianalisis menggunakan teknik svd

Gambar 3.4 Hasil Dari Penggunaan Filtering

5. Stemming

Proses stemming merupakan suatu langkah yang terdapat dalam sistem Information Retrieval (IR) yang bertujuan untuk mengubah struktur kata-kata dalam dokumen tertentu, istilah-istilah dasarnya (root word) berdasarkan peraturan yang berlaku. Dalam konteks bahasa Inggris, proses stemming umumnya dilakukan karena struktur imbuhan bahasa ini bersifat stabil dan mudah dijalankan. Namun, pada proses stemming dalam bahasa Indonesia memiliki perbedaan dengan bahasa Inggris karena bahasa Indonesia memiliki struktur imbuhan yang lebih kompleks, sehingga mengakibatkan kesulitan

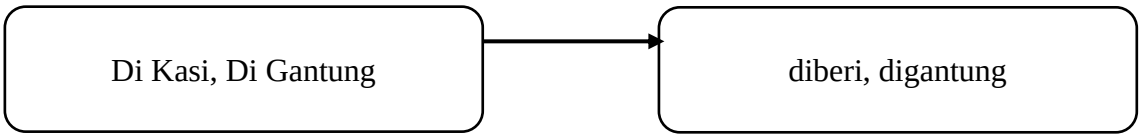
dalam proses pengolahan. Hasil dari penerapan stemming ini ditunjukkan dalam gambar 3.5.



Gambar 3.5 Hasil Dari Penggunaan Stemming

6. Normalization

Tahap normalization merupakan proses di mana kata-kata yang tidak baku akan diubah atau diperbaiki menjadi baku (Aditiya et al., 2022).



Gambar 3.6 Hasil Dari Penggunaan Normalization

3.3.3 Text Transformation

Teks Pada bagian ini, penelitian berfokus pada transformasi teks atau pembentukan atribut guna memperoleh representasi dokumen yang diperlukan. Metode ekstraksi fitur yang digunakan dalam tahap ini adalah TF-IDF. Tahap awal dari proses transformasi melibatkan pembagian data menjadi tiga skenario, yakni skenario 1 dengan presentasi rasio 80% data untuk latihan dan 20% rasio data untuk pengujian, skenario 2 dengan presentasi 70% data untuk latihan dan 30% data untuk pengujian, serta skenario 3 dengan presentasi 60% data untuk latihan dan 40% data untuk pengujian.

3.3.4 Data Mining

Pada tahap ini, dilakukan proses klasifikasi sentimen pada data ulasan setelah melewati proses transformasi menggunakan algoritma Naïve Bayes. Terdapat tiga skenario yang digunakan dalam proses klasifikasi ini, yaitu skenario 1 dengan 80% data untuk latihan dan 20% data untuk pengujian, skenario 2 dengan 70% data untuk latihan dan 30% data untuk pengujian, serta skenario 3 dengan 60% data untuk latihan dan 40% data untuk pengujian.

3.3.5 Evaluation

Di tahap terakhir dari penelitian ini, Melaksanakan tinjauan evaluasi terhadap model yang ada sebelumnya dengan maksud menghitung tingkat ketepatan saat menjalankan klasifikasi menggunakan algoritma Naïve Bayes, dan hasil keakuratan pada komentar berita kriminal. Dalam penilaian, Confusion Matrix diterapkan sebagai metode yang umum digunakan untuk mengukur performa model klasifikasi. Peran utama Confusion Matrix dalam analisis prediktif adalah membandingkan nilai sebenarnya dengan hasil prediksi yang dihasilkan oleh model yang telah diterapkan.

Dalam konteks evaluasi ini mencakup berbagai metrik evaluasi yang digunakan, termasuk akurasi, ketepatan, sensitivitas, dan F1-score (f-measure), yang akan memberikan pemahaman tentang sejauh mana performa model Naïve Bayes Classifier dalam analisis sentimen komentar berita Detik.com melalui proses web scraping. Hasil evaluasi ini akan memberikan wawasan yang berharga untuk memahami seberapa baik model dapat mengklasifikasikan sentimen dari komentar-komentar tersebut dan memberikan kontribusi pada analisis keseluruhan terhadap respons pembaca terhadap berita kriminal yang dipublikasikan oleh Detik.com.

DAFTAR PUSTAKA

Aditiya, P., Enri, U., & Maulana, I. (2022). Analisis Sentimen Ulasan Pengguna Aplikasi Myim3 Pada Situs Google Play Menggunakan Support Vector Machine. *JURIKOM (Jurnal Riset Komputer)*, *9*(4), 1020–1028.

Alexander, A., & Firza, M. H. H. (2023). Analisis Kesalahan Ejaan dan Tanda Baca Pada Salah Satu Surat Kabar. *Jurnal Ilmiah Dan Karya Mahasiswa*, *1*(1), 24–28.

Alexandra, L., Fitriani, F., & Satria, A. (2022). *DATASET COLLECTIVE VIOLENCE EARLY WARNING: CUPLIKAN KEKERASAN DAN INTERVENSI DI INDONESIA PADA 2021*.

Alfianistiawati, R. (2021). Konstruksi media massa dalam pembentukan stigma masyarakat mengenai covid-19. *Jurnal Ilmu Komunikasi Acta Diurna*, *17*(2).

Alwasi'a, A. (2020). *Analisis Sentimen pada review Aplikasi Berita Online Menggunakan Metode Maximum Entropy (Studi Kasus: Review Detikcom pada Google Play 2019)*.

Arhami, M., Kom, M., & Muhammad Nasir, S. T. (2020). *Data Mining-Algoritma dan Implementasi*. Penerbit Andi.

Asgarnezhad, R., Monadjemi, S. A., & Soltanaghaei, M. (2021). An application of MOGW optimization for feature selection in text classification. *The Journal of Supercomputing*, *77*, 5806–5839.

Darmawan, T. D. (2022). Analisis Sentimen Review Pelanggan E-Commerce di Indonesia Menggunakan Algoritma Naive Bayes Classifier. *Universitas Dinamika*.

Fahrudin, T. M., Riyantoko, P. A., & Hindrayani, K. M. (2023). Implementation of Web Scraping on Google Search Engine for Text Collection Into Structured 2D List. *Telematika: Jurnal Informatika Dan Teknologi Informasi*, *20*(2), 139–152.

Hardjo, S. (2019). Hubungan Antara Persepsi Terhadap Berita Kriminal di Televisi dengan Kecemasan Ibu Rumah Tangga Akan Tindak kejahatan. *Universitas Medan Area*.

Hartama, D., Windarto, A. P., & Wanto, A. (2019). The application of data mining in determining patterns of interest of high school graduates. *Journal of Physics: Conference Series*, *1339*(1), 012042.

Herdhianto, A. (2020). Sentiment analysis menggunakan Naïve Bayes Classifier (NBC) PADA tweet tentang zakat. *Fakultas Sains Dan Teknologi Universitas Islam Negeri Syarif* .

Hikmawan, S., Pardamean, A., & Khasanah, S. N. (2020). Sentimen Analisis Publik Terhadap Joko Widodo terhadap wabah Covid-19 menggunakan Metode Machine Learning. *Jurnal Kajian Ilmiah*, *20*(2).

Mahariani, A., & Nurmalasari, D. (2022). Distribution and Classification of Community Service Topics Based on the Results of Extracting Information from Proposal Documents Using Text Mining With Naive Bayes Algorithm (Case Study: Polytechnic Caltex Riau). *Jurnal Aksara Komputer Terapan*, *11*.

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *5*(2), 697–711.

Nurkholis, A., Fernando, Y., & Ans, F. A. (2023). METODE VECTOR SPACE MODEL UNTUK WEB SCRAPING PADA WEBSITE FREELANCE. *INTI Nusa Mandiri*, *18*(1), 52–58.

Panggabean, S. R. (2018). *Konflik dan perdamaian etnis di Indonesia*. Pustaka Alvabet.

Panuntun, S. B., Krismawati, D., Pramana, S., & Astuti, E. T. (2023). Analisis Teks Pemberitaan Telemedicine di Indonesia: Pendekatan Sentimen, NER, Topic Modeling, dan Social Network dalam Memahami Isu dan Persepsi. *Indonesian of Health Information Management Journal (INOHIM)*, *11*(1), 56–67.

Purnamawati, A. (2021). Analisis Sentimen Pengguna Aplikasi Whatsapp dengan Algoritma Machine Learning Classifier Berbasis Smote. *Fakultas Teknologi Informasi Universitas Nusa Mandiri*.

Ruslim, K. I., Adikara, P. P., & Indriati, I. (2019). Analisis Sentimen Pada Ulasan Aplikasi Mobile Banking Menggunakan Metode Support Vector Machine dan Lexicon Based Features. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(7), 6694–6702.

Sari, F. V., & Wibowo, A. (2019). Analisis sentimen pelanggan toko online Jd. Id menggunakan metode Naïve Bayes Classifier berbasis konversi ikon emosi. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *10*(2), 681–686.

Singgalen, Y. A. (2023). Analisis Sentimen Wisatawan terhadap Kualitas Layanan Hotel dan Resort di Lombok Menggunakan SERVQUAL dan CRISP-DM. *Building of Informatics, Technology and Science (BITS)*, *4*(4), 1870–1882.

Suryadi, A., Syb'an, W. A., Alfa'inna, N., & Hermaliani, E. H. (2023). Implementasi Web Scraping dan Sentiment Analysis Terhadap Berita Menggunakan Machine Learning. *Swabumi (Suara Wawasan Sukabumi): Ilmu Komputer, Manajemen, Dan Sosial*, *11*(1), 28–34.

Wibowo, F. R., Rusdianto, D. S., & Arwan, A. (2019). Pengembangan Sistem Pengumpulan Promo E-Commerce Berbasis Website Dengan Menerapkan Teknik Web Scraping Dalam Proses Pengambilan Data Promo. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(3), 2887–2893.

Yuniwati, E. S., Wungubelen, B. L., & Saharudin, H. (2021). Pengaruh Tayangan Berita Kriminal Terhadap Kecemasan Ibu Rumah Tangga akan Tindak Kejahatan pada Anak. *Jurnal Penelitian & Pengkajian Ilmiah Mahasiswa (JPPIM)*, *2*(4), 58–64.