ANALISIS SENTIMEN MEMBANDINGKAN PENGGUNA APLIKASI *E-COMMERCE* TOKOPEDIA DAN SHOPEE MENGGUNAKAN ALGORITMA *NAIVE BAYES*

SKRIPSI

STEVEN

20201000012

TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2025

ANALISIS SENTIMEN MEMBANDINGKAN PENGGUNA APLIKASI *E-COMMERCE* TOKOPEDIA DAN SHOPEE MENGGUNAKAN ALGORITMA *NAIVE BAYES*

SKRIPSI

Diajukan sebagai salah satu syarat untuk kelengkapan gelar kesarjanaan pada

Program Studi Teknik Informatika

Jenjang Pendidikan Strata 1

STEVEN

20201000012

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2025

LEMBAR PERSEMBAHAN

Dengan penuh rasa syukur kepada Tuhan Yang Maha Esa, saya mempersembahkan karya SKRIPSI ini kepada:

1. Kedua orang tua tercinta, yang selalu memberikan dukungan, doa, dan kasih sayang yang tiada henti.
2. Kakak dan adikku yang telah memberikan dukungan semangat serta dorongan yang senantiasa diberikan.
3. Teman-teman pada grup "dunia terbalik" yang selalu memberikan dukungan dan berjuang bersama.
4. Para senior di tempat ibadah yang selalu memberikan motivasi dan dukungan yang diberikan.
5. Dosen pembimbing yang telah dengan sabar membimbing dan memberikan ilmu serta arahan sehingga skripsi ini dapat diselesaikan.

UNIVERSITAS BUDDHI DHARMA

LEMBAR PERNYATAAN KEASLIAN SKRIPSI

Yang bertanda tangan di bawah ini.

NIM : 20201000012

Nama : STEVEN

Jenjang Studi : Strata 1

Program Studi : Teknik Informatika

Peminatan : Database Development

Dengan ini saya menyatakan bahwa:

1. Skripsi ini adalah asli dan belum pernah diajukan untuk mendapat gelar akademik Sarjana atau kelengkapan studi, baik di Universitas Buddhi Dharma maupun di Pereguruan Tinggi lainnya.
2. Skripsi ini saya buat sendiri tanpa bantuan dari pihak lain, kecuali arahan dosen pembimbing.
3. Dalam Skripsi ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dan dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan daftar pustaka.
4. Dalam Skripsi ini tidak terdapat pemalsuan (kebohongan), seperti buku, artikel, jurnal, data sekunder, pengolahan data, dan pemalsuan tanda tangan dosen atau Ketua Program Studi Universitas Buddhi Dharma yang dibuktikan dengan keasliannya.
5. Lembar pernyataan ini saya buat dengan sesungguhnya, tanpa paksaan dan apabila dikemudian hari atau pada waktu lainnya terdapat penyimpangan dan ketidak benaran dalam pernyataan ini, saya bersedia menerima sanksi akademik berupa pencabutan gelar akademik yang telah saya peroleh karena Skripsi ini serta sanksi lainnya sesuai dengan peraturan dan norma yang berlaku.

Tangerang, 05 Februari 2025

Steven

20201000012

UNIVERSITAS BUDDHI DHARMA

LEMBAR PERSETUJUAN PUBLIKASI KARYA ILMIAH

Yang bertanda tangan di bawah ini.

NIM : 20201000012

Nama : STEVEN

Jenjang Studi : Strata 1

Program Studi : Teknik Informatika

Peminatan : Database Development

Dengan ini menyetujui untuk memberikan ijin kepada pihak Universitas Buddhi Dharma, Hak Bebas Royalti Non – Eksklusif (Non-exclusive Royalty Fee Right) atas karya ilmiah kami yang berjudul: "ANALISIS SENTIMEN MEMBANDINGKAN PENGGUNA APLIKASI *E-COMMERCE* TOKOPEDIA DAN SHOPEE MENGGUNAKAN ALGORITMA *NAIVE BAYES"*.

Dengan Hak Bebas Royalti Non – Eksklusif ini pihak Universitas Buddhi Dharma berhak menyimpan, mengalih-media atau format-kan, mengelolanya dalam pangkalan data (database), mendistribusikannya, dan menampilkan atau mempublikasikannya di internet atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis atau pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan Universitas Buddhi Dharma, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Tangerang, 05 Februari 2025

Steven

20201000012

UNIVERSITAS BUDDHI DHARMA

LEMBAR PENGESAHAN PEMBIMBING

ANALISIS SENTIMEN MEMBANDINGKAN PENGGUNA APLIKASI *E-COMMERCE* TOKOPEDIA DAN SHOPEE MENGGUNAKAN ALGORITMA *NAIVE BAYES*

Dibuat Oleh:

NIM : 20201000012

NAMA : Steven

Telah disetujui untuk dipertahankan dihadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

Peminatan Database Development

Tahun Akademik 2024/2025

Disahkan oleh,

Tangerang, 08 Januari 2025

Pembimbing,

Indah Fenriana, S.Kom., M.Kom

NIDN : 0406028801

UNIVERSITAS BUDDHI DHARMA

LEMBAR PENGESAHAN SKRIPSI

ANALISIS SENTIMEN MEMBANDINGKAN PENGGUNA APLIKASI *E-COMMERCE* TOKOPEDIA DAN SHOPEE MENGGUNAKAN ALGORITMA *NAIVE BAYES*

Dibuat Oleh:

NIM : 20201000012

NAMA : Steven

Telah disetujui untuk dipertahankan dihadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

Peminatan Database Development

Tahun Akademik 2024/2025

Disahkan oleh,

Tangerang, 05 Februari 2025

Dekan,	**Ketua Program Studi,**
Dr. Yakub, M.Kom., M.M.	Hartana Wijaya, S.Kom, M.Kom
NIDN : 0304056901	NIDN : 0412058102

UNIVERSITAS BUDDHI DHARMA

LEMBAR PENGESAHAN TIM PENGUJI

Nama : Steven

NIM : 20201000012

Fakultas : Sains dan Teknologi

Judul Skripsi : ANALISIS SENTIMEN MEMBANDINGKAN PENGGUNA APLIKASI *E-COMMERCE* TOKOPEDIA DAN SHOPEE MENGGUNAKAN ALGORITMA *NAÏVE BAYES*

Dinyatakan LULUS setelah mempertahankan di depan Tim Penguji Komprehensif pada hari Rabu, 05 Februari 2025

Nama Penguji :	Tanda Tangan :
Ketua Sidang : **Riki, M.Kom** 0431128204	……..
Penguji I : **Junaidi Akbar, S.Pd., M.Pd.T** 0431079403	……..
Penguji II : **Indah Fenriana, S.Kom., M.Kom** 0406028801	……..

Mengetahui,

Dekan Fakultas Sains dan Teknologi

Dr. Yakub, M.Kom., M.M.

NIDN : 0304056901

KATA PENGANTAR

Dengan mengucapkan Puji Syukur kepada Tuhan Yang Maha Esa, yang telah memberikan Rahmat dan karunia-Nya kepada penulis sehingga dapat menyusun dan menyelesaikan Skripsi ini dengan judul **ANALISIS SENTIMEN MEMBANDINGKAN PENGGUNA APLIKASI *E-COMMERCE* TOKOPEDIA DAN SHOPEE MENGGUNAKAN ALGORITMA *NAIVE BAYES*.**Tujuan utama dari pembuatan Skripsi ini adalah sebagai salah satu syarat kelengkapan dalam menyelesaikan program pendidikan Strata 1 Program Studi Teknik Informatika di Universitas Buddhi Dharma. Dalam penyusunan Skripsi ini penulis banyak menerima bantuan dan dorongan baik moril maupun materiil dari berbagai pihak, maka pada kesempatan ini penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Ibu Dr. Limajatini, S.E., M.M., B.K.P. sebagai Rektor Universitas Buddhi Dharma.
2. Bapak Dr. Yakub, S.Kom., M.Kom., M.M. Dekan Fakultas Sains dan Teknologi Universitas Buddhi Dharma.
3. Bapak Hartana Wijaya, S.Kom., M.Kom sebagai Ketua Program Studi Teknik Informatika Universitas Buddhi Dharma.
4. Ibu Indah Fenriana, S.Kom., M.Kom., sebagai pembimbing yang telah membantu dan memberikan dukungan serta harapan untuk menyelesaikan penulisan Skripsi ini.
5. Orang tua dan keluarga yang selalu memberikan dukungan baik moril dan materil.
6. Teman-teman yang selalu membantu dan memberikan semangat.

Serta semua pihak yang terlalu banyak untuk disebutkan satu-persatu sehingga terwujudnya penulisan ini. Penulis menyadari bahwa penulisan Skripsi ini masih belum sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan di masa yang akan datang.

Akhir kata semoga ini dapat berguna bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Tangerang, 05 Februari 2025

Penulis

ANALISIS SENTIMEN MEMBANDINGKAN PENGGUNA APLIKASI *E-COMMERCE* TOKOPEDIA DAN SHOPEE MENGGUNAKAN ALGORITMA *NAÏVE BAYES*
98 Halaman + xvii / 31 Tabel / 34 Gambar / 9 Lampiran

ABSTRAK

Perkembangan pengguna internet di Indonesia telah mendorong peningkatan penggunaan aplikasi *e-commerce* seperti Tokopedia dan Shopee. Namun, menganalisis sentimen pengguna dari banyaknya ulasan menjadi tantangan tersendiri karena kompleksitas bahasa, variasi ekspresi, dan volume data yang besar. Penelitian ini bertujuan membandingkan sentimen pengguna terhadap Tokopedia dan Shopee menggunakan algoritma Naïve Bayes dengan dua teknik pembobotan kata, yaitu *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Count Vectorizer*. Data diperoleh melalui scraping ulasan *Naive Bayes*, dengan total 28.259 ulasan (14.406 untuk Shopee dan 13.853 untuk Tokopedia). Evaluasi model menggunakan akurasi dan *Confusion Matrix* menunjukkan bahwa TF-IDF memberikan performa lebih baik dengan akurasi 80% dibandingkan *Count Vectorizer*. Hasil analisis sentimen mengungkapkan bahwa Shopee memiliki sentimen positif lebih tinggi daripada Tokopedia, terutama dalam aspek kepuasan pengguna terhadap fitur dan layanan pelanggan. Temuan ini menegaskan pentingnya fokus pada pengalaman pengguna dan layanan pelanggan untuk meningkatkan daya saing platform *e-commerce*. Penelitian ini memberikan kontribusi praktis bagi pengembang aplikasi dalam mengidentifikasi area perbaikan, seperti peningkatan fitur dan layanan pelanggan. Bagi peneliti lain, studi ini dapat menjadi referensi dalam analisis sentimen menggunakan algoritma *Naïve Bayes* dan teknik pembobotan kata. Selain itu, bagi konsumen, hasil penelitian ini dapat menjadi pertimbangan dalam memilih *platform e-commerce* berdasarkan ulasan pengguna. Dengan demikian, penelitian ini tidak hanya memperkaya literatur analisis sentimen di Indonesia tetapi juga memberikan rekomendasi strategis bagi pengembangan *e-commerce* di masa depan.

Kata Kunci : Perbandingan *E-commerce*, *Naïve Bayes*, TF-IDF, *Count Vectorizer*

SENTIMEN ANALYSIS COMPARING USERS OF THE TOKOPEDIA AND SHOPEE E-COMMERCE APPLICATIONS USING THE NAÏVE BAYES ALGORITHM
98 *Pages* + xvii / 31 *Tables* / 34 *Figures* / 9 *Attachment*

ABSTRACT

The growth of internet users in Indonesia has driven the increased use of e-commerce applications such as Tokopedia and Shopee. However, analyzing user sentiment from a large number of reviews presents its own challenges due to language complexity, variations in expression, and the large volume of data. This study aims to compare user sentiment toward Tokopedia and Shopee using the Naïve Bayes algorithm with two word-weighting techniques, namely Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorizer. Data was obtained through scraping reviews from the Google Play Store, with a total of 28,259 reviews (14,406 for Shopee and 13,853 for Tokopedia). Model evaluation using accuracy and Confusion Matrix showed that TF-IDF performs better with an accuracy of 80% compared to Count Vectorizer. Sentiment analysis results revealed that Shopee has a higher positive sentiment than Tokopedia, particularly in terms of user satisfaction with features and customer service. These findings emphasize the importance of focusing on user experience and customer service to enhance the competitiveness of e-commerce platforms. This study provides practical contributions for app developers in identifying areas for improvement, such as enhancing features and customer service. For other researchers, this study can serve as a reference for sentiment analysis using the Naïve Bayes algorithm and word-weighting techniques. Additionally, for consumers, the results of this study can be a consideration in choosing e-commerce platforms based on user reviews. Thus, this research not only enriches the literature on sentiment analysis in Indonesia but also provides strategic recommendations for the future development of e-commerce.

Keywords : Comparasion E-commerce , Naïve Bayes, TF-IDF, Count Vectorizer

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi dan komunikasi, khususnya internet, telah mengubah cara masyarakat dalam hal melakukan transaksi jual beli. *E-commerce*, atau aplikasi jual beli, telah menjadi solusi dalam memenuhi kebutuhan konsumen yang menginginkan kemudahan berbelanja tanpa harus keluar rumah. Di Indonesia, berbagai platform *E-commerce* seperti Tokopedia dan Shopee telah muncul sebagai pemain utama dalam industri ini, menciptakan persaingan yang kompetitif dalam menawarkan beragam produk dan layanan.(Bintang Zulfikar Ramadhan et al., 2023) pasar *E-commerce* di Indonesia terus berkembang pesat. Menurut laporan dari Statista pada penelitian(Ulhaq et al., 2024) jumlah pengguna *E-commerce* di Indonesia diprediksi akan terus meningkat, mencerminkan adopsi yang sangat luas dari belanja online di kalangan konsumen penduduk di Indonesia. Pemilihan Tokopedia dan Shopee sebagai objek penelitian didasarkan pada beberapa pertimbangan strategis. Kedua platform ini merupakan pemain dominan dalam industri *e-commerce* Indonesia, dimana berdasarkan data dari *Naive Bayes*, Tokopedia telah mencapai lebih dari 100 juta unduhan dengan 7 juta ulasan(data *Naive Bayes*), sementara Shopee mencatatkan lebih dari 100 juta unduhan dengan 14 juta ulasan per tahun 2024 (data *Naive Bayes*).

Seiring dengan meningkatnya penggunaan platform aplikasi *E-commerce* juga memunculkan berbagai tantangan masalah. Masalah utama meliputi persaingan ketat antar-platform,keterbatasan akses,keamanan data,pengalaman pengguna,kepuasan pelanggan,dsb .Dengan masalah tersebut platform aplikasi *e-commerce* perlu berinovasi untuk mengatasi tantangan ini dan memenuhi ekspektasi konsumen(Nur'aeni et al., 2024). dengan adanya *Naive Bayes* konsumen kini dapat dengan mudah memberikan *feedback* melalui ulasan dan

penilaian di berbagai platform aplikasi pada *Naive Bayes*. Ulasan ini membuat sumber informasi yang berharga bagi perusahaan dalam memahami pandangan dan tingkat kepuasan pengguna terhadap layanan yang mereka tawarkan. (Fransiska & Irham Gufroni, 2020). Untuk mengolah ulasan konsumen yang tersedia di *Naive Bayes*, diperlukan metode yang mampu menangani data teks dalam jumlah besar secara sistematis.

Analisis sentimen merupakan sebuah alat untuk mengidentifikasi dan mengklasifikasikan emosi atau opini yang terkandung dalam teks, seperti positif, negatif, atau netral. Teknik ini banyak diterapkan dalam berbagai bidang, termasuk pemasaran, dan penelitian perilaku pengguna.(Tabinda Kokab et al., 2022) melalui implementasi sistem analisis sentimen menggunakan *machine learning* algoritma *Naïve Bayes* pada ulasan pandangan pengguna terhadap aplikasi *E-commerce* tokopedia Maupun shopee di *Naive Bayes*, penelitian ini dapat menjadi sumbangan pemikiran untuk pengembangan teori terkait dan juga dapat membantu bagi *E-commerce* Tokopedia maupun *E-commerce* Shopee dalam meningkatkan kualitas layanan aplikasi pengguna(Pratmanto *et al*., 2020).

Dampak sentimen pengguna terhadap keberlangsungan bisnis *e-commerce* sangat signifikan, karena sentimen positif dapat meningkatkan kepuasan pelanggan dan loyalitas, sementara sentimen negatif dapat mengakibatkan penurunan minat belanja dan reputasi perusahaan(Kusuma & Cahyono, 2023) Selain itu, integrasi media sosial dengan *e-commerce* juga berperan penting dalam memperkuat posisi perusahaan di pasar digital, yang menunjukkan bahwa sentimen pengguna dapat mempengaruhi pertumbuhan dan keberlangsungan bisnis *e-commerce*(Michael, 2024) Oleh karena itu, pemantauan dan analisis sentimen pengguna menjadi kunci untuk menjaga daya saing dan keberlangsungan bisnis di era digital ini.

Beberapa penelitian terdahulu telah membuktikan keefektifan algoritma *Naïve Bayes* untuk membuat model sentimen analisis dengan pendekatan *machine learning* yaitu(Huang

et al., 2023), (Pratmanto et al., 2020), (Ernianti Hasibuan & Elmo Allistair Heriyanto, 2022) juga pada beberapa studi yang dilakukan oleh (Huang et al., 2023) ini bertujuan untuk mengeksplorasi sejauh mana metode klasifikasi ulasan pada *platform E-commerce* menggunakan algoritma *machine learning* dapat diandalkan. Pada penelitian ini, fitur seleksi yang digunakan untuk pembobotan kata yaitu *Term Frequency-Inverse Document Frequency* (*TF-IDF*), sebagai salah satu teknik dalam prosesnya. Hasil penelitian dengan menunjukan akurasi bervariasi pada beberapa algoritma seperti *Decision Tree, Random Forest,Support Vector Machine* dan *Naïve Bayes* yang salah satunya algoritma mendapatkan akurasi tertinggi sebesar 98.17%. Pada penelitian yang dilakukan oleh (Pratmanto et al., 2020) mengklasifikasi sentimen pada aplikasi *E-commerce* shopee sebanyak 140 ulasan sebagai data latih dan 60 ulasan sebagai data uji, algoritma *Naïve bayes* berhasil mencapai akurasi sebesar 96,67% menujukan bahwa klasifikasi yang dilakukan termasuk dalam kategori sangat baik mengidentifikasi sentimen pada ulasan aplikasi Shopee, pada penelitian tersebut juga menggunakan sebuah fitur yaitu *TF-IDF* untuk pembobotan kata. Penelitian yang dilaksanakan (Ernianti Hasibuan & Elmo Allistair Heriyanto, 2022) dengan mengklasifikasikan ulasan aplikasi *E-commerce* Amazon Shopping menggunakan algoritma *naïve bayes* menghasilkan akurasi rata-rata sebesar 82.15%.Penelitian menunjukkan bahwa metode ini efektif dalam mengidentifikasi sentimen dari ulasan pengguna, dan menggunakan fitur *TF-IDF*, dalam proses pembobotan kata. Pemilihan algoritma *Naive Bayes* dalam penelitian ini didasarkan pada beberapa keunggulan yang sesuai untuk analisis sentimen dalam konteks *e-commerce*. Dengan algoritma *Naive Bayes* memiliki kemampuan yang baik dalam menangani dataset teks yang besar dengan komputasi yang relatif ringan dan cepat, yang sangat penting mengingat volume ulasan yang besar dari kedua platform *e-commerce*. Algoritma ini juga efektif dalam menangani fitur kategorik dan menunjukkan performa yang baik dalam klasifikasi teks pendek seperti ulasan pengguna.

Meskipun beberapa penelitian sebelumnya telah menunjukkan keefektifan dalam menganalisis sentimen *e-commerce*, masih terdapat beberapa keterbatasan yang perlu diatasi. Penelitian (Pratmanto et al., 2020) hanya menggunakan dataset terbatas sebanyak 140 ulasan sebagai data latih dan 60 ulasan sebagai data uji pada platform Shopee, yang mungkin tidak cukup representatif untuk menggambarkan keseluruhan sentimen pengguna. Sementara itu, studi yang dilakukan (Huang et al., 2023) berfokus pada perbandingan berbagai algoritma *machine learning* namun tidak mempertimbangkan pendekatan hybrid dengan lexicon-based yang dapat meningkatkan akurasi klasifikasi sentimen. Penelitian (Ernianti Hasibuan & Elmo Allistair Heriyanto, 2022) yang mencapai akurasi 82.15% hanya menggunakan TF-IDF sebagai metode pembobotan kata, tanpa membandingkan dengan teknik pembobotan lain seperti *Count Vectorizer* yang mungkin lebih sesuai untuk karakteristik tertentu dari ulasan *e-commerce*. Selain itu, mayoritas penelitian terdahulu cenderung menganalisis platform *e-commerce* secara terpisah, tanpa melakukan perbandingan langsung antara dua platform besar seperti Tokopedia dan Shopee yang dapat memberikan wawasan berharga tentang preferensi dan kepuasan pengguna di Indonesia. Keterbatasan ini menunjukkan perlunya penelitian yang lebih komprehensif dengan menggunakan kombinasi pendekatan lexicon-based dan *machine learning*, serta membandingkan efektivitas berbagai teknik pembobotan kata dalam konteks analisis sentimen *e-commerce*.

Hasil pengelompokan sentimen berdasarkan peringkat score berbasis *lexicon, dan Machine learning*, dengan *Lexicon-based* dapat menangkap nuansa sentimen dalam konteks bahasa Indonesia dan bahasa informal yang umum digunakan dalam ulasan, sementara *machine learning* dengan algoritma *Naive Bayes* dapat mengenali pola-pola kompleks dalam data. Pada setiap *lexicon* memiliki kamus kata yang telah diklasifikasikan dengan intensitas sentimen tertentu, yang membantu dalam mengidentifikasi perasaan atau opini publik

terhadap topik tertentu, dimana yang hasilnya minus akan dilabelkan menjadi sentimen negatif, dan untuk nilai plus sebagai positif, dan dengan nilai kosong sebagai sentimen netral.(Sham & Mohamed, 2022) Salah satu pendekatan yang dapat digunakan dalam sentimen analisis dengan cara menggunakan pendekatan *machine learning*. Hasil dari penelitian ini diharapkan dapat menyumbang kontribusi yang signifikan baik secara teoritis maupun praktis. Dengan melakukan perbandingan sentimen antara pengguna aplikasi Tokopedia dan Shopee menggunakan algoritma *Naive Bayes* dan pendekatan *lexicon-based.* Berdasarkan latar belakang yang sudah dijabarkan, penelitian yang akan dibuat dengan judul "**ANALISIS SENTIMEN MEMBANDINGKAN PENGGUNA APLIKASI *E-COMMERCE* TOKOPEDIA DAN SHOPEE MENGGUNAKAN ALGORITMA *NAIVE BAYES***". Penelitian ini akan membandingkan hasil dari jumlah ulasan sentimen dari aplikasi *E-commerce* tokopedia dan shopee. Karena kedua aplikasi tersebut sering sekali diminati oleh pengguna dikarenakan terdapat berbagai promo,harga yang kompetitif, kebijakan pengembalian yang layak, dsb. Karena itu penelitian ini memilih kedua data aplikasi tersebut. Dengan penggunaan fitur untuk pembobotan kata yang berbeda, yaitu *TF-IDF* dan *Count Vectorizer*. Penelitian ini akan diimplementasikan ke dalam aplikasi website. Pemilihan metode TF-IDF dan *Count Vectorizer* dalam penelitian ini bertujuan untuk menentukan teknik pembobotan kata terbaik yang dapat menangkap pola sentimen secara optimal pada platform *E-commerce* untuk Tokopedia dan Shopee. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam beberapa aspek penting. Dari sisi pengembangan aplikasi *e-commerce*, hasil analisis sentimen yang dapat membantu Tokopedia dan Shopee mengidentifikasi fitur-fitur yang perlu ditingkatkan atau diperbaiki berdasarkan umpan balik pengguna. Dalam konteks layanan pelanggan, pemahaman yang lebih mendalam tentang sentimen pengguna memungkinkan kedua platform untuk mengembangkan strategi layanan yang lebih responsif

dan personal, seperti memperbaiki sistem penanganan keluhan atau meningkatkan kecepatan respons terhadap masalah pelanggan. Dengan hasil perbandingan sentimen antara kedua platform dapat memberikan wawasan berharga tentang faktor-faktor yang mempengaruhi kepuasan pengguna dalam berbelanja online, termasuk preferensi terhadap antarmuka aplikasi, sistem pembayaran, atau layanan pengiriman. Selain itu, penggunaan kombinasi pendekatan lexicon-based dan *machine learning* dalam penelitian ini dapat berkontribusi pada pengembangan model analisis sentimen yang lebih akurat, khususnya untuk konteks *e-commerce* Indonesia yang memiliki karakteristik unik dalam penggunaan bahasa dan pola interaksi pengguna.

Untuk memaksimalkan manfaat dan aksesibilitas hasil penelitian, analisis sentimen ini akan diimplementasikan dalam bentuk aplikasi berbasis website. Aplikasi ini akan dilengkapi dengan beberapa fitur utama seperti dashboard visualisasi yang menampilkan perbandingan sentimen antara Tokopedia dan Shopee secara real-time, grafik tren sentimen berdasarkan periode waktu, dan jumlah kata yang paling banyak di bahas pada setiap aplikasi tersebut. Implementasi dalam bentuk website ini memungkinkan stakeholder terkait, baik dari pihak *e-commerce* maupun peneliti, untuk dengan mudah mengakses dan menganalisis hasil perbandingan sentimen pada kedua platform. Selain itu, antarmuka website yang interaktif akan memudahkan pengguna dalam memahami pola sentimen dan mengidentifikasi area-area yang memerlukan peningkatan layanan pada masing-masing *platform*.

1.2 Identifikasi Masalah

Berdasarkan uraian di atas, permasalahan yang dapat disusun adalah sebagai berikut:

1) Keterbatasan waktu dalam menganalisis sejumlah besar ulasan pengguna pada *Google Play Store* menyebabkan sulitnya mengetahui perbandingan kualitas layanan antara aplikasi Tokopedia dan Shopee berdasarkan sentimen pengguna.
2) Belum mengetahui tingkat kepuasan antara rating dan komentar sesuai yang diberikan oleh pengguna antara perbandingan *platform* Tokopedia dan Shopee berdasarkan komentar atau sentimen yang diberikan oleh konsumen.

1.3 Ruang Lingkup

Ruang lingkup penelitian ini adalah sebagai berikut:

1) *Program* dibuat berbasis *website*.
2) *Dataset* yang digunakan merupakan *dataset* sekunder yang didapatkan melalui *scraping Naive Bayes*.
3) *Dataset* terdiri dari 14.406 dari Shopee dan 13.853 data Tokopedia.
4) Informasi yang dihasilkan berupa presentase dari klasifikasi positif, negatif, maupun netral. Yang dinilai melalui kamus sentimen-bahasa/leksikon/inset at master · onpilot/sentimen-bahasa · GitHub
5) Informasi atas perbedaan signifikan dalam sentimen pengguna antara aplikasi *E-commerce* tokopedia dan shopee.
6) Evaluasi dilakukan menggunakan akurasi dan *Confusion Matrix*.

1.4 Tujuan dan Manfaat

1.4.1 Tujuan

Tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut:

1) Mengurangi keterbatasan waktu dalam menganalisis ulasan pengguna dengan mengambil data ulasan pada *Google Play Store* secara manual dengan menerapkan metode analisis sentimen berbasis *Naïve Bayes* serta pembobotan kata menggunakan TF-IDF dan *Count Vectorizer*.
2) Meningkatkan kepuasan terhadap perbandingan *platform* sesuai yang diberikan oleh pengguna antara Tokopedia dan Shopee berdasarkan komentar dan sentimen.

1.4.2 Manfaat

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1) Melakukan analisis sentimen ulasan aplikasi tokopedia dan shopee secara otomatis tanpa perlu campur tangan manusia.
2) Memberikan kontribusi positif terhadap peningkatan kualitas layanan aplikasi *E-commerce* Tokopedia dan Shopee maupun layanan aplikasi lainnya terkait sentimen pengguna untuk perbandingan aplikasi untuk *platform Google Play Store*.

1.5 Sistematika Penulisan

Sistematika penulisan disusun untuk mempermudah pemahaman dan analisis penelitian. Dalam laporan ini, materi dipaparkan dalam beberapa sub-bab yang secara umum dapat dijelaskan sebagai berikut:

BAB I PENDAHULUAN

Isi dari bab ini meliputi latar belakang masalah, identifikasi masalah, perumusan masalah, lingkup penelitian, tujuan dan manfaat penelitian, metode penelitian, dan sistematika penulisan.

BAB II LANDASAN PEMIKIRAN TEORITIS

Isi dari bab ini meliputi uraian terkait teori-teori yang mendasari penelitian ini yang dibagi menjadi tiga bagian yaitu teori umum, teori khusus, dan teori rancangan, tinjauan studi, dan kerangka pemikiran.

BAB III ANALISA MASALAH DAN PERANCANGAN APLIKASI

Pada bab ini berisi uraian terkait pengumpulan data, metode yang digunakan, *Flowchart* dan perancangan *interface*.

BAB IV PENGUJIAN DAN IMPLEMENTASI

Pada bab ini berisi uraian terkait evaluasi model yang telah dikembangkan, tampilan program, dan pengujian aplikasi.

BAB V KESIMPULAN DAN SARAN

Isi dari Bab ini meliputi uraian terkait simpulan penelitian berdasarkan hasil pengujian dan implementasi yang telah dilaksanakan dan saran-saran yang dapat dikembangkan untuk penelitian selanjutnya.

BAB II

LANDASAN TEORI

2.1 Teori Umum

2.1.1 *E-commerce*

E-commerce atau biasa disebut dengan Toko online merupakan wadah digital yang memudahkan jual-beli barang melalui internet. Di platform ini, penjual dapat memamerkan produknya dan pembeli bisa langsung membeli dengan mudah. Keuntungan utamanya adalah penjual tidak perlu mengeluarkan biaya sewa tempat atau membayar biaya tambahan untuk membuka toko, karena *platform e-commerce* menyediakan ruang jualan secara gratis.(Dikananda et al., 2024)

2.1.2 Analisis Sentimen

Proses mengurai teks untuk mengetahui apakah pesan tersebut bersifat negatif, positif, atau netral. Pada konteks ini, analisis sentimen dapat membantu pengguna memahami pendapat, emosi, dan sikap yang terkandung dalam sebuah teks. (Ranjan & Mishra, 2020), teknik ini dapat digunakan untuk menganalisis opini atau sentimen orang terhadap topik, acara, individu, isu, layanan, produk, organisasi, dan atribut-atributnya di media sosial. Dengan kata lain, pengguna dapat mengetahui bagaimana orang merespons atau merasa terhadap berbagai hal berdasarkan teks yang ada. Ada tiga metode teknik dalam melakukan klasifikasi sentimen yaitu *hybrid approach, lexicon based*, dan *machine learning* (V. Kevin Sitanayah Que, Ade Iriani, 2020)

2.1.3 *Text Mining*

Text Mining merupakan cara memproses ekstraksi informasi yang berguna dari teks. Dengan melibatkan identifikasi pola, hubungan, dan informasi penting lainnya dari dokumen teks. seperti dokumen *Word*, *PDF*, kutipan teks, atau sebagainya.

Dalam konteks analisis sentimen, *text mining* digunakan untuk melakukan pengekstrakan dan menganalisis opini, sentimen, dan sikap dari teks yang dihasilkan oleh pengguna(Birjali *et al*., 2021). Dan menurut (Alhaq *et al*., 2021) *text mining* adalah suatu proses ekstraksi informasi teks di mana seorang pengguna menggunakan alat analisis, yang merupakan bagian dari *data mining*,yang dimaksud dengan kategorisasi.

Text mining dapat dianggap sebagai topik penelitian yang relatif baru. Penggunaan *text mining* dapat menawarkan solusi untuk masalah seperti pemrosesan, pengorganisasian, dan analisis volume besar teks tak terstruktur. Banyak bidang penelitian menggunakan *text mining*, termasuk pengembangan perangkat lunak, media online, pemasaran, pendidikan, dan politik. Pre-proses teks diperlukan untuk analisis text mining melalui beberapa tahap, seperti normalisasi, tokenisasi, *filtering*, *stemming*, *tagging*, dan analisis. Tahap-tahap ini sama dengan proses data mining.(Samsir *et al*., 2021)

2.1.4 *Text Preprocesing*

Text Preprocesing merupakan langkah-langkah yang diperlukan untuk mempersiapkan data teks sebelum dilakukan analisis sentimen. Proses ini biasanya meliputi pembersihan teks, penghapusan tanda baca, normalisasi kata, stemming serta penghilangan stop words. Dengan tujuan untuk mengurangi noise dalam data dan meningkatkan kualitas input yang akan digunakan dalam model analisis sentimen (Huang et al., 2023)

2.1.5 Data

Data merupakan suatu informasi yang terstruktur dari suatu fenomena atau realitas yang dapat diidentifikasikan secara spesifik. Yang memungkinan untuk dihitung, dibandingkan, disagregasi, dibedakan, dan dapat diulang atau direkonstruksi(Marchionini, 2023)

Menurut (Rukhmana, 2021) data dapat diperoleh dari 2 cara yaitu :

a. Data Primer

   Data primer dapat diperoleh dengan cara langsung melalui sumber asli dengan cara pengamatan, pengukuran, atau interaksi sistematis dengan subjek penelitian. Perolehan data dilakukan oleh peneliti sendiri menggunakan instrumen penelitian yang tervalidasi.

b. Data Sekunder

   Data dapat diperoleh dari repositori atau dokumentasi yang telah ada sebelumnya, dimana pengumpulan datanya dilakukan oleh pihak lain. Data ini umumnya telah mengalami pemrosesan dan tersedia dalam format yang terstruktur melalui berbagai media.

2.1.6 Website

Website merupakan satu kumpulan halaman atau *page* pada suatu domain yang berisi berbagai konten yang dapat di visualisasikan oleh pengguna. Website biasanya mengandung gambar, ilustrasi, video, dan teks untuk berbagai tujuan. Setiap website memiliki desain dan struktur unik yang dirancang untuk menarik minat pengunjung dan menyampaikan informasi secara efektif. (Fitriani et al., 2022)

2.2 Teori Khusus

2.2.1 *Naïve Bayes*

Naïve Bayes adalah Algoritma klasifikasi *Bayesian* bersumber pada teorema *Bayes* yang terdapat kemampuan klasifikasi yang sama dengan algoritma *Decission Tree,* dan *Neural Network*. Algoritma *Naïve Bayes* digunakan untuk mengklasifikasikan bahwa sampel data masuk ke dalam kategori yang sesuai berdasarkan distribusi probabilitas dan memilih kategori dengan probabilitas tertinggi sebagai kategori yang diprediksi. Karena algoritma *Naive Bayes* stabil dan mudah dioperasikan dan banyak digunakan dalam klasifikasi sentimen(Li *et al*., 2020)

Salah satu keandalan *Naïve Bayes* adalah kemampuannya untuk bekerja dengan baik meskipun dalam kasus dimensi fitur yang besar. Meskipun sederhana, metode ini mampu memberikan hasil dengan akurasi tinggi dalam banyak kasus, terutama pada data teks.(Putri et al., 2020)

Rumus dasar *Naïve Bayes* untuk analisis sentimen dapat diilustrasikan dengan menggunakan teorema Bayes. Rumus umumnya adalah

$$P(Sentimen|Dokumen) = \frac{P(Dokumen|Sentimen).P(Sentimen)}{P(Dokumen)}$$

- P(*Sentimen|Dokumen)* adalah probabilitas sentimen bersasarkan dokumen yang diamati.
- P*(Dokumen|Sentimen)* adalah probabilitas dokumen muncul jika sentimen tertentu sudah diketahui.

- P*(Sentimen)* variabel acak yang mewakili sentimen (positif, negatif, atau netral) dari pengguna terhadap aplikasi *E-COMMERCE* Tokopedia dan Shopee.
- P*(Dokumen)* adalah variabel acak yang mewakili data ulasan atau feedback pengguna terhadap kedua aplikasi tersebut.

Metode "*Naive*" dalam *Naive Bayes* mengasumsikan bahwa setiap fitur (kata dalam konteks analisis sentimen) bersifat independen, meskipun kenyataannya ini mungkin tidak sepenuhnya benar. Dengan asumsi ini, rumus tersebut dapat disederhanakan menjadi:

$$P(Sentimen|Dokumen) \propto P(Sentimen) \cdot \prod_{i=1}^{n} P(Kata_i|Sentimen)$$

- $P(Kata_i|Sentimen)$ adalah probabilitas kemunculan kata ke-i jika sentimen sudah diketahui.

Proses pelatihan *Naive Bayes* melibatkan menghitung probabilitas ini dari data pelatihan yang telah diberi label, dan selanjutnya, model dapat digunakan untuk mengklasifikasikan sentimen dokumen baru berdasarkan probabilitas yang diestimasi. Dalam konteks analisis sentimen untuk aplikasi *E-commerce* Tokopedia dan Shopee, kata-kata akan diambil dari ulasan pengguna sebagai fitur-fitur yang digunakan dalam rumus *Naive Bayes*.

2.2.2 *TF-IDF*

Term Frequency - Inverse Document Frequency proses menghitung atau mengekstrak kata menjadi angka vektor yang digunakan untuk mengetahui seberapa berat kata dalam dokumen atau korpus. Bobot ini dapat digunakan untuk mengetahui seberapa penting kata dalam dokumen.(Tri Putra et al., 2023) Metode Term Frequency-Inverse Document Frequency *(TF-IDF)* adalah alat statistik yang berguna

untuk menentukan seberapa penting sebuah kata atau kelompok kata tertentu dalam sebuah dokumen.(Suryani & Edy, 2020) berikut merupakan rumus dari TF-IDF :

$$TF = \frac{\text{Jumlah kemunculan kata}}{Total\ kata\ dalam\ dokumen}$$

$$IDF = log\ \frac{\text{Total Dokumen}}{Dokumen\ yang\ mengandung\ kata}$$

$$TF - IDF = TF * IDF$$

2.2.3 *Count Vectorizer*

Count Vectorizer merupakan teknik dasar *Natural Language Processing* (NLP) yang digunakan untuk ekstraksi fitur dari data teks. teknik ini memainkan peran penting dalam mengubah teks tak terstruktur dari media sosial, menjadi format terstruktur yang sesuai untuk algoritma *machine learning*.(Turki & Roy, 2022) *Count Vectorizer* begitu penting untuk algoritma *machine learning* karena dapat mengubah data tekstual menjadi format numerik yang dapat dianalisis,kemudian menjadikannya alat dasar dalam tugas pemrosesan bahasa alami (*natural language*) seperti klasifikasi emosi.(Ahmed et al., 2022)

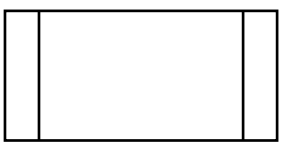
2.3 Teori Rancangan

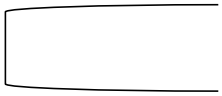
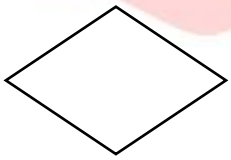
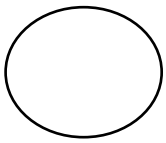
2.3.1 *Flowchart*

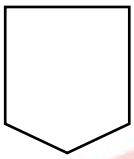
Flowchart merupakan representasi grafis dari langkah-langkah dalam suatu proses, menggunakan simbol-simbol standar untuk menunjukkan berbagai jenis tindakan atau keputusan. Ini membantu dalam memvisualisasikan alur kerja dan hubungan antar langkah.(Rusydy et al., 2024)

Secara umum, *Flowchart* terbagi menjadi dua jenis: *system Flowchart* dan *program Flowchart*. *Sistem Flowchart* mencakup penyelesaian masalah yang saling berhubugan dan berinteraksi untuk mencapai tujuan tertentu. Sementara itu, *program Flowchart* berfokus pada langkah-langkah penyelesaian masalah unit untuk mencapai hasil tertentu dan berfungsi sebagai pedoman untuk program komputer. (Chaudhuri, 2020) Simbol umum seperti terminal, proses, panah, dan lain-lain digunakan dalam *Flowchart*:

Tabel 2.1 Tabel *Flowchart*

Simbol	Nama	Deskripsi
	*Terminal*	Berfungsi sebagai pernyataan awal atau akhir suatu program
	*Input/Output*	Berfungsi sebagai pernyataan proses input atau output
	*Computer Processing*	Berfungsi sebagai pernyataan yang di proses oleh computer
	*Predefined Processing*	Berfungsi sebagai menggambarkan proses yang tidak dijelaskan secara spesifik dalam *Flowchart*. Biasanya,

		simbol ini mewakili subprogram.
	*Comment*	Berfungsi sebagai memberi penjelasan yang lebih jelas terhadap suatu elemen
	Flow	Berfungsi sebagai penghubung antara simbol yang lain biasa disebut Connecting line
	Document Input/Output	Berfungsi sebagai memberi pernyataan bahwa inputan bentuk dokumen fisik atau output yang perlu dicetak
	Decission	Berfungsi sebagai menandakan pada kondisi tertentu yang akan memungkinkan antara jawaban ya dan tidak
	On-page Connector	Berfungsi sebagai keluar – masuk atau meneruskan

		proses dalam lembar kerja yang sama
	Off-page Connector	Berfungsi sebagai penyambungan proses dalam lembar kerja yang berbeda

Sumber : (Setiawan, 2021)

2.3.2 HTML

HTML (*Hyper Text Markup Language)* sebuah bahasa *markup* yang berfungsi sebagai pengorganisir struktur dan konten halaman web. Yang mencakup elemen-elemen seperti teks, gambar, dan interaksi pengguna yang diwakili dalam format yang dapat dipahami oleh browser.(Gur et al., 2023)

Cara kerja HTML cukup sederhana. Proses dimulai ketika klien mengakses URL (*Uniform Resource Locator*) melalui peramban (*browser*). Lalu, peramban menghubungi server web yang mengirimkan semua informasi yang diperlukan. Setelah informasi diterima, peramban web langsung menerjemahkan kode HTML dan menampilkannya di layar pengguna.(Wiyanto et al., 2022)

2.3.3 *Jupyter Notebook*

Jupyter Notebook digunakan sebagai lingkungan untuk menguji kemampuan model dalam menyelesaikan masalah pemrograman dan *data science*. Selain itu, terdapat penjelasan tentang struktur *Jupyter Notebook*, termasuk sel-sel yang terdiri dari kode dan dokumentasi, serta bagaimana model dapat memanfaatkan konteks dari sel-sel sebelumnya untuk menghasilkan solusi yang tepat.(Chandel et al., 2022)

2.3.4 *Python*

Python merupakan bahasa pemrograman serbaguna yang bersifat terinterpretasi dan tingkat tinggi. Bahasa ini diciptakan oleh Guido van Rossum pada tahun 1991. Bahasa ini memiliki filosofi desain unik yang menekankan pada kejelasan dan keterbacaan kode, salah satunya melalui penggunaan spasi putih yang memiliki makna struktural. Filosofi ini membuat Python mudah dipahami dan ditulis, baik untuk proyek kecil maupun besar. (Wikipedia, 2023)

Python Bahasa pemrograman yang sangat populer dan sering digunakan karena fleksibilitasnya yang luar biasa Bahasa ini digunakan hampir di seluruh bidang teknologi, mulai dari pengembangan perangkat lunak, aplikasi web, analisis data, kecerdasan buatan, hingga pengembangan game. dan lainnya. Penggunaan Python tidak terbatas pada satu domain saja. Di dunia pengembangan perangkat lunak, Python juga digunakan untuk membuat aplikasi *desktop* dan *web*. Di bidang *data science*, *Python* menjadi bahasa yang dominan untuk analisis data, pemrosesan statistik, dan pembuatan model *machine learning*. (Biznet, 2023)

2.3.5 *Google Colab Research*

Google Colab adalah Platform berbasis *cloud* ini memungkinkan pengguna membuat dan menjalankan dokumen *notebook* secara online langsung melalui browser, tanpa perlu repot menginstal perangkat lunak atau mengatur konfigurasi komputer sendiri. Untuk melakukan pemrograman dan analisis data menggunakan *Python*(Guntara, 2023)

Google Colab merupakan dokumen digital yang memungkinkan pengguna menulis, menyimpan, dan berbagi program komputer melalui Google Drive. Platform ini sangat berguna untuk menjalankan kode Python dan proyek *machine*

learning dengan mudah dan cepat.(Handika, 2024)*Google Colab* disini digunakan untuk mengscrape data dari masukkan dan ulasan pada platform *Naive Bayes* menjadi *file .CSV.*

2.3.6 *Scikit-Learn*

Scikit-Learn sebuah pustaka Python yang sangat populer untuk *machine learning* dan analisis data yang mencakup berbagai algoritma untuk klasifikasi, regresi, klastering, dan pengurangan dimensi, serta alat untuk pra-pemrosesan data, pemilihan fitur, dan evaluasi model.(Pölsterl, 2020)

2.3.7 *Flask*

Flask adalah salah satu framework web yang paling populer dalam bahasa pemrograman Python. Keunggulan utama *Flask* terletak pada kemudahan penggunaannya dan dukungan ekstensif. Dengan dukungan penuh untuk *routing, templating*, dan integrasi *database, Flask* memungkinkan pengembang untuk membuat aplikasi web kompleks dengan kode yang bersih dan terstruktur. Yang dikenal sebagai *microframework, Flask* dirancang untuk memberikan kemudahan dan fleksibilitas dalam pengembangan aplikasi web.(Ghimire, 2020)

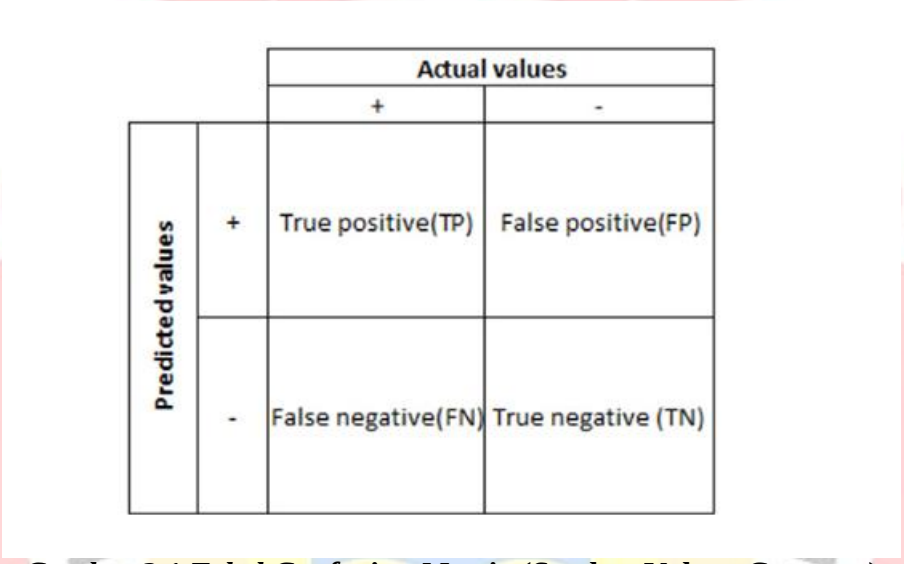
2.4 Teori Pengujian

2.4.1 *Black Box Testing*

Black Box Testing merupakan metode pengujian perangkat lunak di mana peneliti menguji fungsionalitas sistem tanpa melihat kode internal. Pengujian dilakukan dengan memasukkan data uji dan memeriksa hasilnya, fokus pada kesesuaian output dengan kebutuhan atau spesifikasi yang telah ditentukan sebelumnya.(Fahrezi et al., 2022)

2.4.2 *Confusion Matrix*

Confusion Matrix difungsikan untuk menilai performa model klasifikasi, terutama dalam hal klasifikasi biner, *Confusion Matrix* adalah tabel yang menggambarkan jumlah perkiraan yang tepat dan salah yang dibuat oleh model dibandingkan dengan data sebenarnya. Berikut adalah beberapa komponen utama dari *Confusion Matrix*(Valero-Carreras et al., 2023). Tabel *Confusion Matrix* terlihat pada Gambar 2.1



***Gambar 2.1 Tabel Confusion Matrix* (Sumber Valero-Carreras)**

True Positive (TP) adalah Jumlah data yang benar-benar positif dan berhasil dikenali sebagai positif, False Positive (FP) adalah Jumlah data yang sebenarnya negatif, tetapi salah dikenali sebagai positif, True Negative (TN) adalah Jumlah data yang benar-benar negatif dan berhasil dikenali sebagai negative, dan False Negative (FN) adalah Jumlah data yang sebenarnya positif, tetapi salah dikenali sebagai negative.

Tabel *Confusion Matrix*, yang ditunjukkan pada Gambar 2.1, digunakan untuk menilai kinerja pengklasifikasi dengan mengukur akurasi, presisi, recall, dan ukuran f-score ini dapat diperoleh dengan menggunakan rumus berikut:

$$Akurasi = \frac{TP + TN}{TP + TN + TP + FN}$$

$$Akurasi\ Keseluruhan = \frac{TPpositif + TPNetral + TPNegatif}{Total\ Data}$$

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

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

$$F - Measure = \frac{2 * Recall * Presisi}{Recall + Presisi}$$

2.4.3 Kurva ROC-AUC

Kurva ROC-AUC merupakan sebuah diagram yang menunjukkan hubungan antara tingkat true positif (TP) dan tingkat false positif (FP) pada nilai di ambang atas(threshold) yang digunakan oleh model klasifikasi. Luas di bawah kurva ROC-AUC menilai kinerja model untuk memisahkan antara kelas positif dan negatif. Tabel 2.2 berisi pedoman interpretasi AUC.

Tabel 2.2 **Kategori *AUC***

Nilai AUC	Kategori
0.90-1.00	***Excellent Classification***
0.80-0.90	***Good Classification***
0.70-0.80	***Fair Classification***
0.60-0.70	***Poor Classification***
<0.60	***Failure***

2.5 Tinjauan Studi

2.5.1 Penelitian oleh Staphord Bengensi, Timoty Oladunni, Ruth Olusegun, dan Halima Audu

Penelitian berjudul "*A MACHINE LEARNING-SENTIMEN ANALYSIS ON MONKEYPOX OUTBREAK : AN EXTENSIVE DATASET TO SHOW THE POLARITY OF PUBLIC OPINION FROM TWITTER TWEET"* yang dilakukan oleh Staphord Bengensi, Timoty Oladunni, Ruth Olusegun, dan Halima Audu pada tahun 2023 menyoroti pentingnya memahami persepsi publik terhadap wabah Monkeypox, peneliti menggumpulkan data yang berasal dari twitter API, dan dibagi menjadi 3 kelas positif, negatif dan netral dengan membandingkan metode pembobotan kata TF-IDF dan *Count Vectorizer*. Dan mendapatkan akurasi sebesar 93.48% pada SVM dengan pembobotan kata *Count Vectorizer*

2.5.2 Penelitian oleh Nurul Zalza Bilal Jannah, Kusnawi

Penelitian berjudul "*COMPARISON OF NAÏVE BAYES AND SVM IN SENTIMEN ANALYSIS OF PRODUCT REVIEWS ON MARKETPLACES*" yang dilakukan oleh Nurul Zalza Bilal Jannah, Kusnawi pada tahun 2024 bertujuan untuk analisis sentimen terhadap ulasan produk di marketplace Shopee dengan memanfaatkan algoritma *Naïve Bayes* dan *Support Vector Machine* (SVM) sebagai model penelitian. Dalam penelitian , peneliti mengumpulkan data dengan cara scraping dari marketplace shopee dengan menggunakan metode lexicon untuk pelabelan dan menggunakan *feature extraction* TF-IDF dan menggunakan model *k-fold cross-validation*. Dan penelitian yang dilakukan menunjukan bahwa SVM mendapatkan akurasi sebesar 88.58%.

2.5.3 Penelitian oleh Saiful Ulya, Achmad Ridwan, Widya Cholid Wahyudin, Fida Maisa Hana

Penelitian berjudul "*TEXT MINING* SENTIMEN ANALISIS PENGGUNA APLIKASI *MARKETPLACE* TOKOPEDIA BERDASAR RATING DAN KOMENTAR PADA *NAIVE BAYES*" yang dilakukan oleh Achmad Ridwan, Widya Cholid Wahyudin, Fida Maisa Hana pada tahun 2022 bertujuan untuk analisis sentimen terhadap ulasan komentar aplikasi marketplace tokopedia menggunakan Naïve bayes, Decisison tree, Deep Learning. Di dalam penelitian, peneliti menggumpulkan data dengan cara scraping mengunakan google scrape dengan menggunakan metode lexicon untuk pelabelan dan menggunakan *feature extraction TF-IDF* dan penelitian yang dilakukan menunjukan bahwa, *Naïve Bayes* mendapatkan akurasi sebesar 89,13%.

2.5.4 Penelitian oleh Safitri Juanita, Krisna Adiyarta, M. Syafrullah

Penelitian berjudul "*SENTIMEN ANALYSIS ON E-MARKETPLACE USER OPINIONS USING LEXICON-BASED AND NAÏVE BAYES MODEL*" yang dilakukan oleh Safitri Juanita, Krisna Adiyarta, M. Syafrullah pada tahun 2022 bertujuan untuk analisis sentimen terhadap ulasan marketplace pada twitter dan Instagram menggunakan *Naïve bayes* klasifikasi yaitu *Gausian* dan *Multinomial Naïve bayes*. dengan menggunakan metode lexicon untuk pelabelan dan feature extraction *TF-IDF* dan *Multinomial Naïve Bayes* mendapatkan akurasi sebesar 90%.

2.5.5 Penelitian oleh Dinar Ajeng Kristiyanti, Dwi Andini Putri, Elly Indrayuni, Acmad Nurhadi, Akhmad Hairul Umam

Penelitian berjudul "*E-WALLET SENTIMEN ANALYSIS USING NAÏVE BAYES AND SUPPORT VECTOR MACHINE ALGORITHM*" yang dilakukan oleh Dinar Ajeng Kristiyanti, Dwi Andini Putri, Elly Indrayuni, Acmad Nurhadi, Akhmad

Hairul Umam pada tahun 2020 bertujuan untuk analisis sentimen terhadap ulasan E-Wallet. Dalam penelitian ini data dikumpulkan melalui Scrape pada *Naive Bayes* dan dikelompokan kedalam dua bagian yaitu positif dan negatif, dengan menggunakan *machine learning Naïve Bayes dan Support Vector Machine* sebagai klasifikasi, dan menggunakan *feature extraction N-Gram* (*Natural Language Processing*) dan model *Naïve Bayes* mendapatkan akurasi sebesar 94%.

2.5.6 Penelitian oleh Dany Pratmanto, Rousyati Rousyati, Fanny Fatma Wati, Andrian Eko Widodo, Suleman Suleman, Ragil Wijianto

Penelitian berjudul "*APP REVIEW SENTIMEN ANALYSIS SHOPEE APPLICATION IN NAIVE BAYES USING NAIVE BAYES ALGORITHM*" yang dilakukan oleh Dany Pratmanto, Rousyati Rousyati, Fanny Fatma Wati, Andrian Eko Widodo, Suleman Suleman, Ragil Wijianto pada tahun 2020 bertujuan untuk analisis sentimen terhadap ulasan pengguna aplikasi Shopee di *Google play*. Data didapatkan melalui scrape ulasan dari *Google play*, data terbagi menjadi 3 bagian yaitu positif,negative, dan netral, dan menggunakan model *machine learning Naïve bayes* menggunakan *feature extraction Count Vectorizer* dan *TF-IDF*, dan model *Naïve Bayes* mendapatkan akurasi sebesar 96,67%.

2.5.7 Penelitian oleh Ghulam Musa Raza, Zainab Saeed Butt, Seemab Latif, Abdul Wahid

Penelitian berjudul "*SENTIMEN ANALYSIS ON COVID TWEETS : AN EXPERIMENTAL ANALYSIS ON THE IMPACT OF COUNT VECTORIZER AND TF-IDF ON SENTIMEN PREDICTIONS USING DEEP LEARNING MODELS*" yang dilakukan oleh Ghulam Musa Raza, Zainab Saeed Butt, Seemab Latif, Abdul Wahid pada tahun 2021 Ditujukan untuk analisis sentimen terhadap cuitan di Twitter terkait COVID-19. Data didapatkan melalui Twitter API data terbagi menjadi dua

kelas yaitu positf dan negatif, dengan memanfaatkan model klasifikasi *Support Vector Machine (SVM),Logistic Regression,Bernoulli Naïve Bayes,Single Layer Perceptron,Multi-Layer Perceptron,* dan membandingkan *feature extraction* TF-IDF dan *Count Vectorizer*, dan hasil yang didapatkan sebesar 93.15% dengan metode SVM dan TF-IDF

2.5.8 Penelitian oleh Bintang Zulfikar Ramadhan, Ibnu Riza, Iqbal Maulana

Penelitian berjudul "ANALISIS SENTIMEN ULASAN PADA APLIKASI *E-COMMERCE* DENGAN MENGGUNAKAN ALGORITMA *NAÏVE BAYES*" yang dilakukan oleh Bintang Zulfikar Ramadhan, Ibnu Riza, Iqbal Maulana pada tahun 2022 Berfokus pada analisis sentimen terhadap ulasan pengguna pada *Naive Bayes*, dengan menggunakan *machine learning Naïve Bayes,* data terbagi menjadi dua kelas yaitu positif dan negative, dan menggunakan *feature extraction TF-IDF*, hasil penelitian menunjukan *Naïve Bayes* mendapatkan akurasi 92%

2.5.9 Penelitian oleh Mohammed Shenify

Penelitian berjudul "*SENTIMEN ANALYSIS OF SAUDI E-COMMERCE USING NAÏVE BAYES ALGORITHM AND SUPPORT VECTOR MACHINE*" yang dilakukan oleh Mohammed Shenify pada tahun 2024 bertujuan untuk mengklasifikasi opini publik mengenai *E-commerce* di Twitter. Data didapatkan melalui scraping komentar pada twitter, dan data terbagi menjadi dua kelas yaitu positif dan negative. Dengan menggunakan *machine learning Naïve Bayes* dan *Support Vector Machine,* dan menggunakan *feature extraction TF-IDF*, hasil penelitian menunjukan *SVM* mendapatkan akurasi 89%.

2.5.10 Penelitian oleh V Oktaviani, B Warsito, H Yasin, R Santoso, Suparti

Penelitian berjudul "*SENTIMEN ANALYSIS OF E-COMMERCE APPLICATION IN TRAVELOKA DATA REVIEW ON GOOGLE PLAY SITE USING NAÏVE BAYES CLASSIFIER AND ASSOCIATION METHOD*" pada tahun 2021 bertujuan untuk analisis sentimen aplikasi *e-commerce* Traveloka menggunakan klasifikasi *Naive Bayes*dan *Association*, data didapatkan melalui scraping melalui web *Google play*, data terbagi menjadi dua kelas yaitu positif, dan negative, dengan menggunakan *feature extraction TF-IDF*, dan hasil penelitian menunjukan *Naïve Bayes* mendapatkan akurasi 91,20%

2.5.11 Penelitian oleh Artanti Inez Tanggraeni, Melkior N. N. Sitokdana

Penelitian berjudul "Analisis Sentimen Aplikasi *E-Government* Pada *Google play* Menggunakan Algoritma *Naïve Bayes"* pada tahun 2022 bertujuan untuk analisis sentimen aplikasi *E-Government* menggunakan algoritma Naïve Bayes dan didapatkan melalui *web scraping* melalui web *Google play*, data dilabelkan secara manual dan menjadi dua kelas yaitu positif dan negative, dengan menggunakan *feature extraction TF-IDF*, dan hasil penelitian menunjukan Naïve Bayes mendapatkan akurasi 89%

2.5.12 Penelitian oleh Nurhaliza Agustina. C.A, Desy Herlina Citra, Wido Purnama,Chairun Nisa, Amanda Rozi Kurnia

Penelitian berjudul "IMPLEMENTASI ALGORITMA *NAIVE BAYES* UNTUK ANALISIS SENTIMEN ULASAN SHOPEE PADA *NAIVE BAYES*" pada tahun 2022 bertujuan untuk analisis sentimen aplikasi ulasan Shopee pada *Google play* menggunakan algoritma *Naïve Bayes*, dan data didapatkan melalui *Web Scraping* Data, dengan menggunakan *feature extraction TF-IDF*, dan hasil peneitian menujukan Naïve Bayes mendapatkan akurasi 83%

2.5.13 Penelitian oleh Sendi Alpin Rizaldi, Syariful Alam, Imay Kurniawan

Penelitian berjudul "ANALISIS SENTIMEN PENGGUNA APLIKASI JMO (JAMSOSTEKMOBILE) PADA *NAIVE BAYES* MENGGUNAKAN METODE *NAIVE BAYES*" pada tahun 2023 bertujuan untuk analisis sentimen aplikasi JMO pada ulasan *Google play* menggunakan metode Naïve Bayes, data didapatkan dengan Web Scraping, melabelkan dengan cara manual, menggunakan feature extraction *TF-IDF*, dan hasil penelitian menunjukan *Naïve Bayes* mendapatkan akurasi 95%

2.5.14 Penelitian oleh Gilbert Darmanawan, Syaiful Alam, M. Imam Sulistyo

Penelitian berjudul "ANALISIS SENTIMEN BERDASARKAN ULASAN PENGGUNA APLIKASI MYPERTAMINA PADA *GOOGLE PLAYSTORE* MENGGUNAKAN METODE *NAÏVE BAYES*" pada tahun 2023 bertujuan untuk analisis sentimen aplikasi Mypertamina pada ulasan *Google play*, data didapatkan dengan cara Web Scraping, dan menggunakan feature extraction *TF-IDF*, dan hasil penelitian menunjukan *Naïve bayes* mendapatkan akurasi 91,6%

2.5.15 Penelitian oleh Anggista OktaviaPraneswara, Nuri Cahyono

Penelitian berjudul "ANALISIS SENTIMEN ULASAN APLIKASI TIKTOK *SHOP SELLER CENTER* DI *GOOGLE PLAYSTORE* MENGGUNAKAN ALGORITMA *NAIVE BAYES*" pada tahun 2023 bertujuan untuk analisis sentimen aplikasi Tiktok Shop Seller Center pada *Google play*, data didapatkan melalui Web scraping, data yang didapatkan 2 yaitu positif dan negatif, dengan menggunakan feature extraction *TF-IDF*, dan hasil penelitian menunjukan *Naïve Bayes* mendapatkan akurasi 87,6%

2.5.16 Rangkuman Model Penelitian

Penelitian sejenis ini menampilkan referensi atau studi terdahulu yang memiliki keterkaitan tema dengan penelitian yang sedang dilakukan. Dalam konteks analisis sentimen, berbagai sumber penelitian sebelumnya yang telah menggunakan algoritma *Naïve Bayes* untuk menganalisis dataset, yang bertujuan memberikan landasan teoritis dan perbandingan metodologi penelitian.

Tabel 2.3 Rangkuman Tinjauan Studi

Peneliti	Metode	Dataset	Akurasi	Kelebihan	Kekurangan
(Bengesi et al., 2023)	*SVM vs Random Forest vs MLP vs Naïve Bayes vs KNN vs XGBoost + TF-IDF + Count Vectorizer*	Twitter API	93.48%	Model SVM dengan *Count Vectorizer* terbukti lebih baik	Penelitian ini mencakup 103 bahasa terhadap tweet dalam Bahasa non – inggris dan memungkinkan beberapa kata bisa terabaikan atau tidak sama dengan arti aslinya
(Jannah & Kusnawi, 2024)	*Naïve Bayes + SVM*	Shopee	88.58%	SVM terbukti lebih baik	Tidak menggunakan d ataset yang lebih luas
(Ulya et al., 2022)	*Naïve Bayes vs Decision Tree vs Deep Learning*	*Naive Bayes*	89.13%	*Naïve Bayes* terbukti lebih baik	Belum memiliki label polaritas emosi yang spesifik, seperti senang, sedih, amarah, dan lain-lain.
(Juanita et al., 2022)	*Gaussian Naïve Bayes vs Multinomial Naïve Bayes*	Twitter API	90%	Mengeksplorisasi algoritma *Naïve bayes*	Tidak dicantumkan akurasi yang didapat, sejumlah data tidak seimbang saat didapatkan
(Kristiyanti et al., 2020)	*Naïve Bayes vs SVM vs*	*Naive Bayes*	94.00%	*Naïve Bayes* terbukti lebih baik	Pelabelan tidak akurat karena komentar

	Random Forest				dengan rating 3 atau 4 dianggap positif
(Pratmanto et al., 2020)	*Naïve Bayes* + TF-IDF	Shopee	96.66%	*Naive Bayes* + TF-IDF mampu meningkatkan akurasi diatas 90%	Hanya mengindentifikasi sentimen positif dan negatif saja, data yang dikumpulkan hanya 200 data
(Raza et al., 2021)	*SVM vs Bernoulli Naïve Bayes vs Single Layer Perceptron vs Multi Layer Perceptron Logistic Regression*	Twitter API	93.15%	SVM + TF-IDF terbukti lebih baik	Tidak menjelaskan tentang teknik yang digunakan untuk mengatasi noise dalam data
(Bintang Zulfikar Ramadhan et al., 2023)	*Naïve Bayes* + TF-IDF	*Naive Bayes*	92%	*Naïve Bayes* + TF-IDF terbukti lebih baik menggunakan Skenario 80:20	Data yang dikumpulkan hanya selang 1 bulan saja, sehingga memungkinkan sebagian pandangan luas bisa terabaikan, pelabelan masih dilakukan secara manual dan hanya mencakup sentimen positif dan negatif saja
(Shenify, 2024)	*Naïve Bayes* vs *Support Vector Machine (SVM)*	Twitter API	89%	SVM + TF-IDF terbukti lebih baik menggunakan Skenario 70:20	Tidak dijelaskan metode yang dilakukan untuk mengtranslate bahasa arab menjadi bahasa inggris
(Oktaviani et al., 2021)	*Naïve Bayes +vs Association*	*Naive Bayes* Traveloka	91.20%	*Naïve Bayes* terbukti lebih baik menggunakan skenario 80:20	Hanya mencakup sentimen positif dan negatif saja

(Tanggraeni & Sitokdana, 2022)	*Naïve Bayes* + TF-IDF	*Naive Bayes*	87%	*Naïve Bayes* terbukti mampu meningkatkan akurasi	Hanya mencakup sentimen positif dan negatif saja, sedikitnya jumlah data yang dilatih dan diuji
(Agustina et al., 2022)	*Naïve Bayes* + *10-fold cross*	*Naive Bayes*	83%	*Naïve Bayes* terbukti lebih baik menggunakan skenario hold-out / 80:20	Hanya mencakup sentimen positif dan negatif saja, tidak dijelaskan metode yang digunakan untuk melabelkan data
(Sendi Alpin Rizaldi, Syariful Alam, 2023)	*Naïve Bayes* + *10-fold cross*	*Naive Bayes*	95%	*Naïve Bayes* dapat meningkatkan nilai akurasi menggunakan metode 10-fold cross	Hanya mencakup sentimen positif dan negatif saja.
(Darmawan et al., 2023)	*Naïve Bayes* + *TF-IDF*	*Naive Bayes*	91.6 %	*Naïve Bayes* terbukti bisa meningkatkan akurasi menggunakan skenario 80:20	Hanya mencakup sentimen positif dan negatif saja
(Cahyono & Anggista Oktavia Praneswara, 2023)	*Naive Bayes* + TF-IDF	*Naive Bayes*	87.6%	*Naïve Bayes* terbukti bisa meningkatkan akurasi menggunakan skenario 90:10	Hanya mencakup sentimen positif dan negatif saja, tidak dijelaskan metode yang digunakan untuk labeling data

Berbagai penelitian telah membandingkan algoritma dalam analisis sentimen, menunjukkan bahwa SVM dengan *Count Vectorizer* memiliki akurasi tertinggi sebesar 93.48% (Bengesi et al., 2023), sementara *Naïve Bayes* dengan TF-IDF juga terbukti sangat efektif dengan akurasi 96.66% pada dataset terbatas (Pratmanto et al., 2020). SVM sering kali unggul dibandingkan Naïve Bayes, terutama saat dikombinasikan dengan TF-IDF

(Jannah & Kusnawi, 2024; Shenify, 2024; Raza et al., 2021), sedangkan penelitian lain menunjukkan keunggulan *Naïve Bayes* dibandingkan *Random Forest* dan SVM, dengan akurasi mencapai 94% (Kristiyanti et al., 2020). Sebagian besar penelitian menggunakan Twitter API (Bengesi et al., 2023; Juanita et al., 2022; Raza et al., 2021; Shenify, 2024), sementara yang lain menggunakan Shopee (Jannah & Kusnawi, 2024; Pratmanto et al., 2020) dan Traveloka (Oktaviani et al., 2021). Teknik TF-IDF dan validasi silang (*10-fold cross-validation*) terbukti meningkatkan akurasi model secara signifikan (Sendi Alpin Rizaldi & Syariful Alam, 2023; Agustina et al., 2022). Namun, sebagian besar penelitian hanya menganalisis sentimen positif dan negatif, tanpa klasifikasi emosi yang lebih spesifik, serta menghadapi keterbatasan dalam jumlah dataset dan metode pelabelan (Bintang Zulfikar Ramadhan et al., 2023; Tanggraeni & Sitokdana, 2022; Cahyono & Anggista Oktavia Praneswara, 2023).

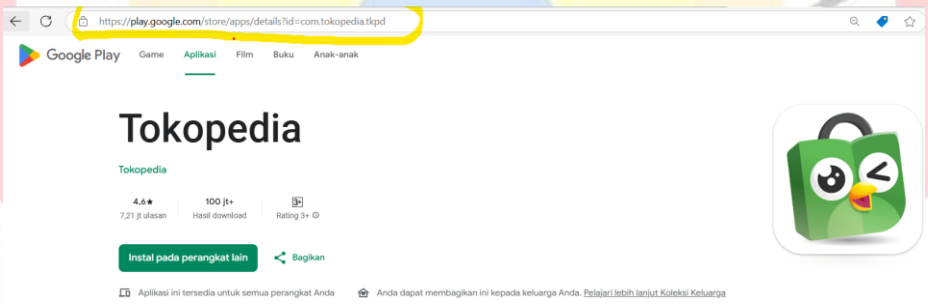
Oleh karena itu, fokus utama penelitian ini adalah mengeksplorasi dua metode pembobotan kata, yakni *TF-IDF* dan *Count Vectorizer*, telah dibuktikan oleh (Bengesi et al., 2023) bahwa *Count Vectorizer* mendapaktan akurasi yang baik itu didapatkan menggunakan SVM, pada penelitian ini akan menggunakan skenario 80:20 karena sudah terbukti berhasil dalam meningkatkan akurasi dengan menggunakan algoritma *machine learning Naïve Bayes*, dan dalam penelitian ini mencakup sentimen positif,negatif dan netral dalam mengindentifikasi sentimen pengguna pada aplikasi platform *Naive Bayes* salah satunya Tokopedia dan Shopee.

BAB III

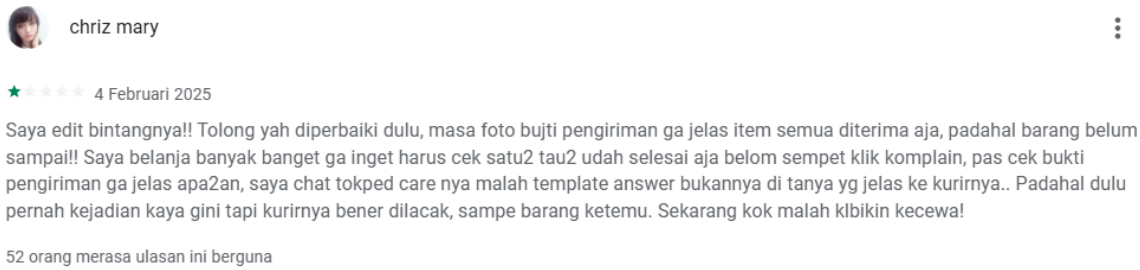
METODOLOGI PENELITIAN

3.1 Pengumpulan Data :

Pengumpulan data dikerjakan dengan cara mengambil *feedback*/ulasan pengguna pada kolom ulasan aplikasi Tokopedia dan Shopee pada *platform Naive Bayes*.Dimana data yang terkumpul di dapatkan dari Januari 2023 hingga Desember 2024 dengan total 20.594 untuk data shopee dan 19.539 untuk data tokopedia. Melalui pemakaian *google-play-scraper* yang dikumpulan data berupa ulasan peringkat, dan komentar pengguna terkait pengalaman pengguna menggunakan kedua aplikasi. Penggambilan data dilakukan berdasarkan Sort.Newest atau ulasan terbaru. Berikut Contoh Proses scraping yang dilakukan :



Gambar 3.1 Halaman *Website Naive Bayes* (contoh tokopedia)



Gambar 3.2 Contoh Halaman Ulasan Dari Tokopedia

Pada halaman tersebut perlu menggunakan fungsi google-play-scraper untuk mengambil data ulasan dengan secara otomatis menggunakan parameter num_reviews. Lalu Menyimpan data seperti userName, score, content, dan at ke dalam *DataFrame*.

Hal ini berfokus untuk mendapatkan detail seperti username ,score ,detail waktu ,ulasan. Selanjutnya, Data yang berhasil dikumpulkan dapat diproses dan disusun dalam format yang tepat untuk keperluan analisis mendatang. Contoh data dari proses pengumpulan informasi dapat ditemukan dalam Tabel 3.1.

Tabel 3.1 Tabel Sample data scraping tokopedia

userName	Score	at	content
Pengguna Google	5	2024-11-08 01:53:54	sangat membantu
Pengguna Google	1	2024-11-08 01:27:30	Kurang inovasi nya
Pengguna Google	5	2024-11-07 18:19:14	Gimana nih tokped banyak kasus di ID Express?
Pengguna Google	3	2024-11-07 02:18:39	diskonnya jarang

Deksirpsi : data ini merupakan contoh hasil pengumpulan data hasil scraping menggunakan *Google-Play-Scrapper* data Tokopedia

Tabel 3.2 Tabel Sample data scraping shopee

userName	Score	at	content
Pengguna Google	2	2024-11-08 02:56:27	Sekarang Shoope pelit banget yaa Ama koin live

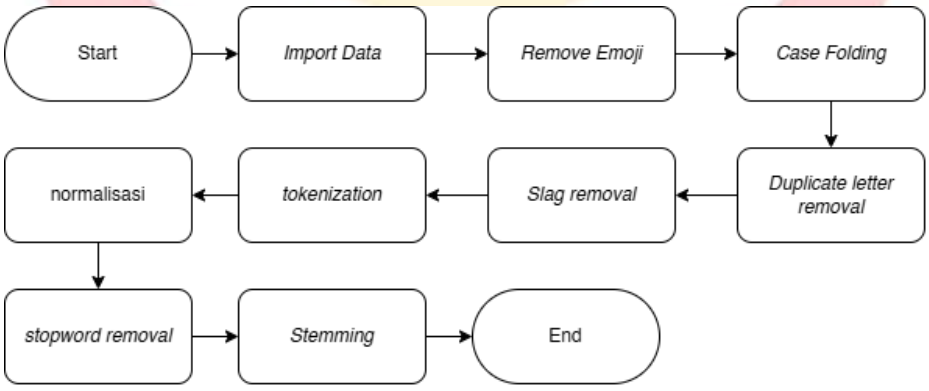
Pengguna Google	5	2024-11-08 02:50:57	be the best menthol jiwa
Pengguna Google	3	2024-11-08 02:34:21	ini shopee kenapa g bisa cod lagi
Pengguna Google	5	2024-11-08 01:43:48	Sangat bagus dan harga terjangkau

Deksirpsi : data ini merupakan contoh hasil pengumpulan data hasil scraping menggunakan *Google-Play-Scrapper* data Shopee

3.2 Pengolahan Data

3.2.1 Text Preprocessing

Pada tahap ini, peneliti melakukan *remove emoji*, *case folding*, *duplicate letter removal*, *slang removal*, *tokenization*, normalisasi, *stemming*, dan *stopwords removal* pada *preprocessing* teks. Dari total keseluruhan yang didapat dari cleaning sampai stemming terdapat 1.929.026 untuk Shopee dan 1.008.844 untuk Tokopedia



Gambar 3.3 *Flowchart Text Preproessing*

Proses pemrosesan teks menggunakan library NLTK dan beberapa corpus yang dilakukan melalui beberapa tahapan sistematis:

1. Import Data

Mengambil data dari scrapping melalui *Google-Play-Scrapper*.

2. Penghapusan Emotikon

Membersihkan teks dari simbol-simbol emotikon dan emoji untuk memastikan penulisan kata yang konsisten dengan Menggunakan regular expression (regex) untuk mendeteksi dan menghapus karakter non-alfanumerik.

3. *Case Folding*

Case folding adalah proses mengubah seluruh teks menjadi huruf kecil untuk memastikan konsistensi dalam pemrosesan kata. Misalnya, kata *"Shopee"*, *"shopee"*, dan *"SHOPEE"* akan diubah menjadi *"shopee"*. Proses ini penting agar perbedaan huruf besar dan kecil tidak mempengaruhi analisis teks.

4. *Duplicate Letter removal*

Teknik ini digunakan untuk menghapus huruf-huruf yang berulang secara tidak wajar dalam sebuah kata. Contohnya, kata *"hebaaat"* akan diubah menjadi *"hebat"*. Teknik ini penting untuk mengurangi variasi kata yang tidak diperlukan akibat penggunaan berlebihan huruf dalam penulisan informal.

5. *Slag Removal*

Slang atau kata-kata tidak baku dikeluarkan dari teks, dengan Menggunakan kamus slang Indonesia yang telah dipersiapkan. Misalnya, *"gak"* akan diubah menjadi *"tidak"*, *"gw"* menjadi *"saya"*, dan *"btw"*

menjadi *"ngomong-ngomong"*. Hal ini membantu meningkatkan akurasi dalam analisis sentimen dengan menghindari variasi kata yang tidak standar.

6. *Tokenizing*,

Proses memecah teks menjadi unit-unit yang lebih kecil (token) seperti kata, frasa, atau symbol. Misalnya, kalimat *"Saya suka belanja di Shopee"* akan diubah menjadi daftar token: ["Saya", "suka", "belanja", "di", "Shopee"]. Teknik ini mempermudah analisis lebih lanjut, seperti pencocokan kata dengan kamus atau teknik pembobotan teks.

7. Normalisasi

Mengubah kata-kata tidak baku menjadi bentuk standar dengan menggunakan kamus mapping kata baku dan tidak baku. Contohnya, *"dpt"* diubah menjadi *"dapat"*, *"krn"* menjadi *"karena"*, dan *"udh"* menjadi *"sudah"*. Langkah ini sangat penting agar kata-kata memiliki bentuk yang seragam sebelum diproses lebih lanjut.

8. *Stopwords removal*,

Proses membuang kata-kata umum yang tidak ada maknanya. Dengan Menggunakan corpus *stopwords* bahasa Indonesia dari NLTK. seperti *"yang"*, *"di"*, *"dan"*, *"ke"*, *"ini"*, dan lain-lain. Proses ini dilakukan dengan menggunakan corpus *stopwords* bahasa Indonesia, menghapus stopwords membantu meningkatkan efisiensi model dengan hanya menyisakan kata-kata yang memiliki makna signifikan.

9. *Stemming*

proses menyederhanakan teks dengan mengurangi variasi kata yang memiliki makna dasar yang sama. Misalnya, *"membeli"* menjadi *"beli"*, *"terbaik"* menjadi *"baik"*, dan *"pelangganan"* menjadi *"pelanggan"*. Dalam bahasa Indonesia.

Tabel 3.3 Proses *Text Preprocessing*

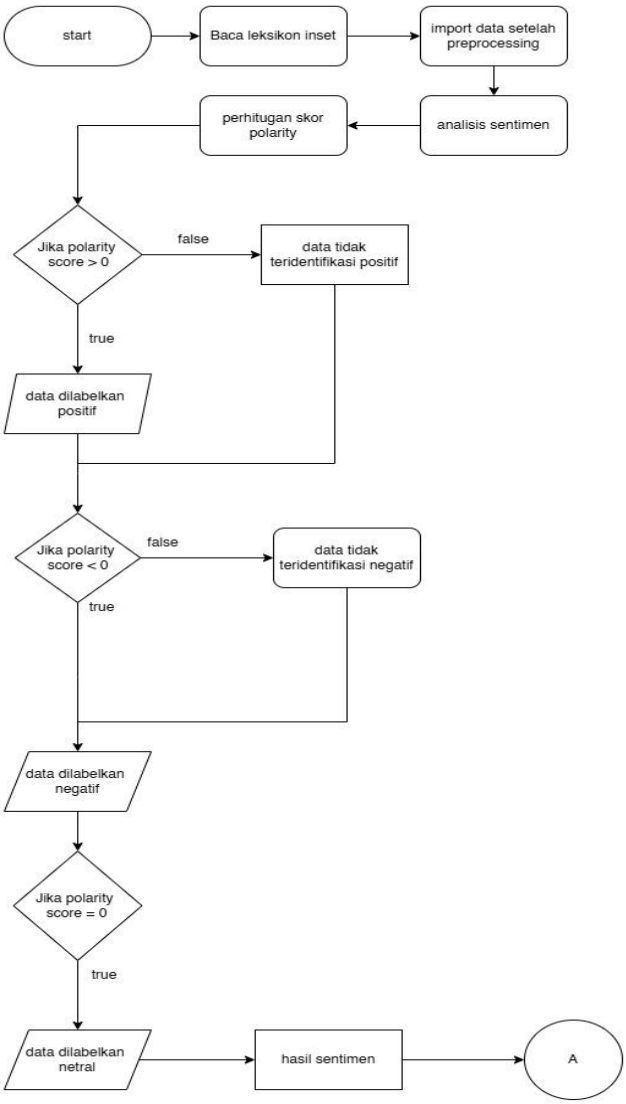
Content	Cleaning	Normalize	Tokenize	Stopword Removal	Stemming Data
Shopee sekarang kenapa pengiriman selalu merekomendasikan Spx nggak kek dulu, saya sangat kecewa karena pengiriman selalu lemot meskipun ada vc senilai 10k tapi kalo barang emang lagi urgent juga bakalan kecewa, kenapa nggak kek dulu J&T yang lebih cepat, bahkan sampe 2 minggu barang saya belum nyampe dan bukan sekali dua kali, kalo memang nggak mau tersaingi oleh jasa pengiriman lain harusnya pihak shopee memperbaiki keluhan yang ada khususnya jasa pengiriman.	shopee sekarang kenapa pengiriman selalu merekomendasikan spx nggak kek dulu saya sangat kecewa karena pengiriman selalu lemot meskipun ada vc senilaik tapi kalo barang emang lagi urgent juga bakalan kecewa kenapa nggak kek dulu j t yang lebih cepat bahkan sampe minggu barang saya belum nyampe dan bukan sekali dua kali kalo memang nggak mau tersaingi oleh jasa pengiriman lain harusnya pihak shopee memperbaiki keluhan yang ada khususnya	shopee sekarang kenapa pengiriman selalu merekomendasikan spx tidak kayak dulu saya sangat kecewa karena pengiriman selalu lemot meskipun ada vc senilaike tapi kalau barang memang lagi urgent juga bakalan kecewa kenapa tidak kayak dulu juta yang lebih cepat bahkan sampai minggu barang saya belum sampai dan bukan sekali dua kali kalau memang tidak mau tersaingi oleh jasa pengiriman lain harusnya pihak shopee memperbaiki keluhan yang ada khususnya	shopee, sekarang, kenapa, pengiriman, selalu, merekomendasikan, spx, tidak, kayak, dulu, saya, sangat, kecewa, karena, pengiriman, selalu, lemot, meskipun, ada, vc, senilai, ke, tapi, kalau, barang, memang, lagi, urgent, juga, bakalan, kecewa, kenapa, tidak, kayak, dulu, juta, yang, lebih, cepat, bahkan, sampai, minggu, barang, saya, belum, sampai, dan, bukan, sekali, dua, kali, kalau, memang, tidak, mau, tersaingi, oleh, jasa, pengiriman, lain, harusnya, pihak, shopee, memperbaiki,	shopee, pengiriman, merekomendasikan, spx, kayak, kecewa, pengiriman, lemot, vc, senilai, barang, urgent, kecewa, kayak, juta, cepat, barang, tersaingi, jasa, pengiriman, shopee, memperbaiki, keluhan, jasa, pengiriman	shopee kirim rekomendasi spx kayak kecewa kirim lemot vc nila barang urgent kecewa kayak juta cepat barang saing jasa kirim shopee baik keluh jasa kirim

	jasa pengiriman	jasa pengiriman	keluhan, yang, ada, khususnya, jasa, pengiriman		
Kesel banget kalo mau nilai pesanan. Gak bisa multi tasking. kalo mau balik lagi mau ngetik malah kembali ke awal kan kesel sudah ngetik panjang lebar malah ngulang.. padahal HP Ram Gede 12+8GB!!!!! dan sudah di kunci aplikasi Shopee nya.. TOLONG DI PERBAIKI!!!	kesel banget kalo mau nilai pesanan gak bisa multi tasking kalo mau balik lagi mau ngetik malah kembali ke awal kan kesel sudah ngetik panjang lebar malah ngulang padahal hp ram gede gb dan sudah di kunci aplikasi shopee nya tolong di perbaiki	kesel banget kalau mau nilai pesanan tidak bisa multi tasking kalau mau balik lagi mau mengetik malah kembali ke awal kan kesel sudah mengetik panjang lebar malah ngulang padahal hp ram gede gb dan sudah di kunci aplikasi shopee ya tolong di perbaiki	kesel, banget, kalau, mau, nilai, pesanan, tidak, bisa, multi, tasking, kalau, mau, balik, lagi, mau, mengetik, malah, kembali, ke, awal, kan, kesel, sudah, mengetik, panjang, lebar, malah, ngulang, padahal, hp, ram, gede, gb, dan, sudah, di, kunci, aplikasi, shopee, ya, tolong, di, perbaiki	kesel, banget, nilai, pesanan, multi, tasking, mengetik, kesel, mengetik, lebar, ngulang, hp, ram, gede, gb, kunci, aplikasi, shopee, perbaiki	kesel banget nilai pesan multi tasking etik kesel etik lebar ngulang hp ram gede gb kunci aplikasi shopee perbaiki
Awalnya kaya ragu ...tpi setelah barang sampai dan sesuai...dan dikemas rapih ....berfungsi dengan baik...jdi belanja yakin dishopee semoga shopee dan toko2 yg ada amanah dan lebih baik kedepannya	awalnya kaya ragu tpi setelah barang sampai dan sesuai dan dikemas rapih berfungsi dengan baik jdi belanja yakin dishopee semoga shopee dan toko yg ada amanah dan	awalnya kayak ragu tapi setelah barang sampai dan sesuai dan dikemas rapi berfungsi dengan baik jadi belanja yakin dishopee semoga shopee dan toko yang ada amanah dan lebih	awalnya, kayak, ragu, tapi, setelah, barang, sampai, dan, sesuai, dan, dikemas, rapi, berfungsi, dengan, baik, jadi, belanja, yakin, dishopee, semoga, shopee, dan, toko, yang, ada, amanah,	kayak, ragu, barang, dikemas, rapi, berfungsi, belanja, dishopee, semoga, shopee, toko, amanah, kedepannya, meningkatkn, kenyamanan, kepercayaan, pelnggan	kayak ragu barang kemas rapi fungsi belanja dishopee moga shopee toko amanah depan meningk atkn nyaman percaya

meningkatkn kenyamanan dan kepercayaan pelnggan	lebih baik kedepannya meningkatkn kenyamanan dan kepercayaan pelnggan	baik kedepannya meningkatkn kenyamanan dan kepercayaan pelnggan	dan, lebih, baik, kedepannya, meningkatkn, kenyamanan, dan, kepercayaan, pelnggan		pelnggan
Ini kenapa dah kalo gua searching selalu loading Mulu Ampe 1 jam padahal jaringan aja normal.udah di coba belanja di tiktok aman aman aja...tapi kalo ke shope loading nya gak ngotak bjir.tolong di per baiki untuk developernya. kurang lebihnya mohon maaf sekian trimagaji	ini kenapa dah kalo gua searching selalu loading mulu ampe jam padahal jaringan aja normal udah di coba belanja di tiktok aman aman aja tapi kalo ke shope loading nya gak ngotak bjir tolong di per baiki untuk developerny a kurang lebihnya mohon maaf sekian trimagaji	ini kenapa sudah kalau gua searching selalu loading mulu sampai jam padahal jaringan saja normal sudah di coba belanja di tiktok aman aman saja tapi kalau ke shope loading ya tidak ngotak bjir tolong di per baiki untuk developerny a kurang lebihnya mohon maaf sekian trimagaji	ini, kenapa, sudah, kalau, gua, searching, selalu, loading, mulu, sampai, jam, padahal, jaringan, saja, normal, sudah, di, coba, belanja, di, tiktok, aman, aman, saja, tapi, kalau, ke, shope, loading, ya, tidak, ngotak, bjir, tolong, di, per, baiki, untuk, developernya , kurang, lebihnya, mohon, maaf, sekian, trimagaji	gua, searching, loading, mulu, jam, jaringan, normal, coba, belanja, tiktok, aman, aman, shope, loading, ngotak, bjir, baiki, developernya , lebihnya, mohon, maaf, trimagaji	gua searchin g loading mulu jam jaring normal coba belanja tiktok aman aman shope loading ngotak bjir baik develop er lebih mohon maaf trimagaj i

3.2.2 Pelabelan Data

Pada langkah ini,data akan diklasifikasikan menjadi label sentimen positif,negatif dan netral dengan dingukur tingkat sentimen dari perbandingan *polarity score* dan rating skor yang diberikan, peneliti melakukan pelabelan berdasarkan lexicon karena lexicon dapat meningkatkan efektivitas dan akurasi dalam penggunaan kata sesuai dengan makna aslinya.(Rahman, 2024) dengan menggunakan kumpulan kata dari lexicon InSet bahasa Indonesia. Berikut alur *Flowchart* pelabelan melalui kamus lexicon. Pada Gambar 3.4.



Gambar 3.4 *Flowchart* Pelabelan Data Melalui Kamus *Lexicon*

Kamus lexicon berbentuk _json_inset-pos.txt, dan _json_inset-neg.txt yang didapatkan melalui sentimen-bahasa/leksikon/inset at master · onpilot/sentimen-bahasa · GitHub. yang digunakan Pada penelitian sebelumnya oleh, (Musfiroh et al., 2021) Inset Lexicon dibuat oleh Fajri Koto dan Gemal Y. Rahmaningtyas dengan menggunakan kata-kata yang dikumpulkan dari Twitter, media sosial yang populer di Indonesia. Dengan mengumpulkan pendapat tertulis dan membaginya menjadi pendapat berisi kata yang menggandung positif atau negatif, yang bisa digunakan untuk menganalisis sentimen publik terhadap topik, acara, atau produk tertentu. yang dimana melalui proses *manual scoring* yang didalamnya itu terdapat 3.609 kata positif dan 6.609 kata negatif dan *polarity score* dimana bila kata tersebut positif maka *polarity score* digolong positif +, bila kata tersebut negatif maka *polarity score* digolongkan negatif -, berikut contoh tabel kamus inset positif ditunjukan pada Tabel.3.4 untuk positif dan Tabel 3.5 untuk negatif.

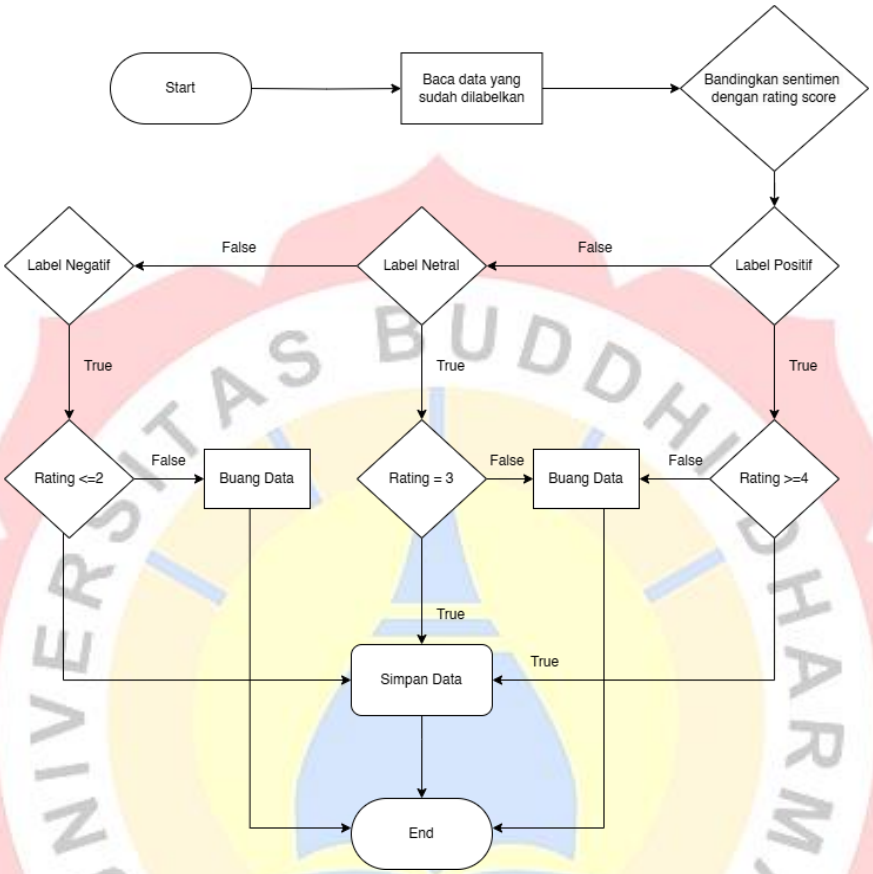
Tabel 3.4 Sample Kamus _Json_Inset-pos.txt

Kata	***Polarity score***	**Interpretasi**
"mantab"	5	Sangat Positif
"welas"	4	Positif
"merespon"	3	Positif Lemah / Netral Positif
"detail"	2	Positif Lemah / Netral Positif
"hendaknya"	1	Positif Sangat Lemah

Tabel 3.5 Sample Kamus _Json_Inset_neg.txt

Kata	***Polarity score***	**Interpretasi**
"lambat"	-1	Negatif Ringan
"mahal"	-4	Negatif Ekstrim
"anjir"	-3	Negatif Kuat

" komplain "	-5	Negatif Ekstrim
"gelo"	-2	Negatif Ringan



Gambar 3.5 *Flowchart* Pelabelan Data Melalui Perbandingan Score

Proses pelabelan sentimen dalam penelitian ini menggunakan pendekatan dua tahap yang mengombinasikan analisis lexicon dan validasi berdasarkan rating pengguna. Seperti yang ditunjukkan pada Gambar 3.5, hasil analisis lexicon dibandingkan dengan nilai rating untuk memastikan konsistensi dan validitas label sentimen. Mengacu pada (Demircan et al., 2021), tiga kategori dengan kriteria sebagai berikut.

1. Bernilai positif (> 0) → teks dianggap memiliki sentimen positif**.**
2. Bernilai negatif (< 0) → teks dianggap memiliki sentimen negatif**.**
3. Bernilai nol (0) → teks dianggap netral**,** karena jumlah kata positif dan negatif seimbang atau tidak mengandung kata yang terdaftar dalam kamus sentimen.

Untuk mempermudah dalam memahami cara perhitungan nilai *polarity score* dan perbandingan dengan rating, berikut ini adalah contoh daftar token kata yang akan diproses dan hasil prosesnya, dapat dilihat pada Tabel 3.4.

***Tabel 3.6* Tabel Perhitungan *Polarity score* dan dibandingkan dengan rating**

Stemming	Total *Polarity Score*	Label dari *polarity score*	Rating	Label Akhir
Mantap, banyak, promo	(5 + 3 + 3) = 11	Positif	5	Positif
anjir, mahal, gelo,	(-3 + -4 + -2) = -9	Negatif	2	Negatif
Udah, mahal, lambat, jelek, lagi	(-4 + -1 + -2) = -7	Negatif	2	Negatif
tanggapi, terimakasih	5	Positif	5	Positif

Data yang telah diberi label menggunakan inset lexicon kemudian dibandingkan dengan rating didapat sebanyak 14.406 buah Untuk shopee dan sebanyak 13.853 untuk tokopedia Informasi tentang jumlah data yang berhasil diberi label dengan benar dan yang mendapat label dengan kesalahan akan dijelaskan dalam Tabel 3.5 dan Table 3.6.

Tabel 3.7 Hasil Labeling Data Shopee

	Benar Dilabelkan	Salah Dilabelkan
Positif	5.382	4.118
Negatif	5.483	1.517
Netral	3.542	553
Total	14.406	6.188

Tabel 3.8 Hasil Labeling Data Tokopedia

	Benar Dilabelkan	Salah Dilabelkan
Positif	5.031	3.436
Negatif	5.077	1.734
Netral	3.745	516
Total	13.853	5.686

3.2.1 *Data Reduction*

Data reduction merupakan bagian dari preprocessing cara menghilangkan data dari atribut yang tidak valid dengan tujuan meminimalkan data dengan tetap menjaga keakuratannya. pada tahap ini dilakukan validasi data dengan mengurangi jumlahnya dari 20.594 menjadi 14.406 untuk data shopee,dan 19.539 menjadi 13.853 untuk data tokopedia pengurangan ini dilakukan dikarenakan banyak data yang terlabelkan tidak sesuai dengan hasil sentimen maupun hasil dari rating/score yang diberikan pengguna. Dengan mengurangi jumlah data, penelitian ini dapat menciptakan sebagian kecil data yang lebih seimbang dan mewakili, memungkinkan proses analisis yang lebih efektif dan akurat.

3.3 Pemodelan Data

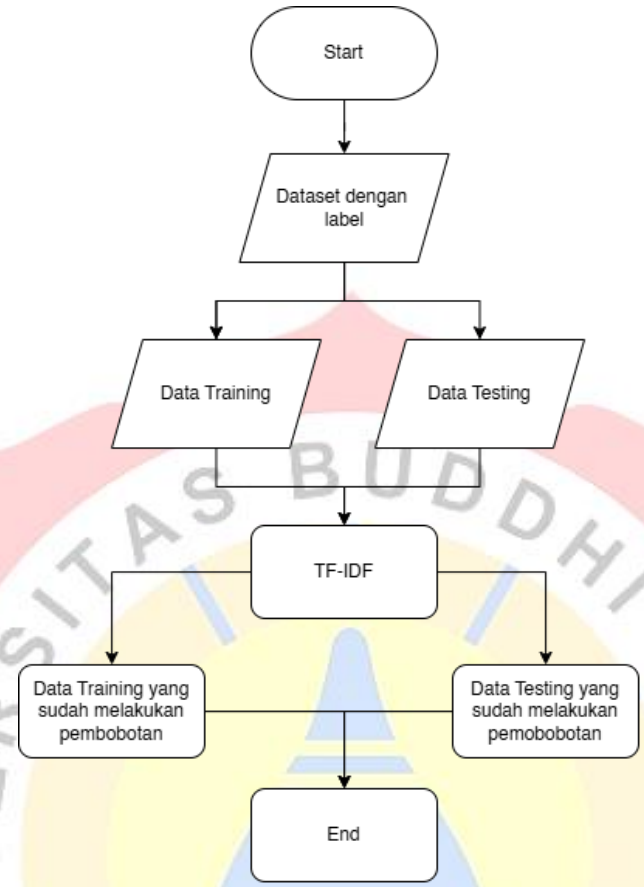
3.3.1 Pembagian Dataset

Pada penelitian ini dataset yang digunakan berjumlah gabungan data shopee dan data tokopedia, yang dibagi menjadi dua bagian yaitu data training dan data testing dengan perbandingan komposisi 80:20, sehingga kedua data bila digabungkan dimana 80% data (22.607) digunakan sebagai data training dan 20% (5.652 ulasan) sebagai data testing.dengan distribusi data pada masing-masing kelas dapat dilihat pada Tabel 3.7.

Tabel 3.9 Distribusi Gabungan Data Shopee dan Tokopedia

Kelas	Training	Testing	Total
Positif	8.329	2.083	10.412
Negatif	8.448	2.112	10.560
Netral	5.820	1.457	7.277
Total	22.607	5.652	28.249

3.3.2 *TF-IDF*

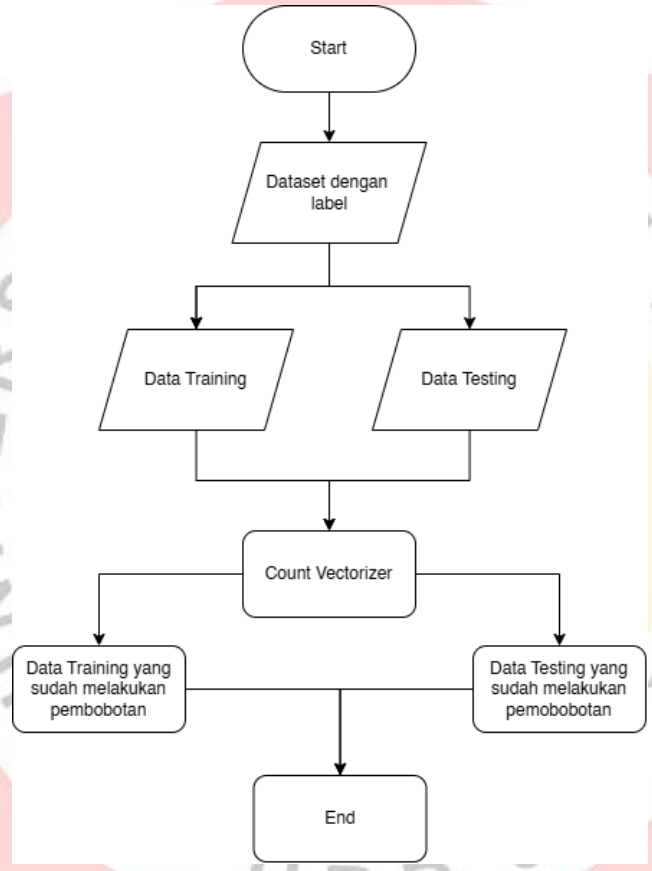


Gambar 3.6 *Flowchart* Pembobotan dengan *TF-IDF*

dalam tahapan ini, dataset yang telah diberi label telah terbagi menjadi dua bagian, yaitu data pelatihan dan data pengujian, dan peneliti melakukan pembobotan kata menggunakan metode *TF-IDF*. Penelitian ini menggunakan model *TF-IDF* dengan parameter yang disesuaikan untuk meningkatkan kualitas representasi teks. Tujuan pertama dari penggunaan max_features=10.000 adalah untuk membatasi jumlah kata yang ada dalam representasi *TF-IDF*. Ini dilakukan untuk meningkatkan efisiensi komputasi, berkonsentrasi pada kata-kata yang dianggap paling penting, dan mengurangi dimensi ruang fitur. Selain itu, parameter min_df=2 digunakan untuk menghindari pemodelan kata-kata acak yang tidak memberikan informasi penting, dan parameter max_df=0.8 digunakan untuk menghindari kata-kata yang muncul di lebih dari 80% dokumen. Ini meningkatkan kemampuan model untuk menemukan pola atau topik tertentu dengan menghindari kata-kata yang muncul di

bawah tujuh dokumen. Penting untuk dicatat bahwa stopword dalam bahasa Indonesia diabaikan karena pemahaman bahwa kata-kata tersebut mungkin tidak memberikan kontribusi yang signifikan dalam analisis. Akibatnya, representasi *TF-IDF* dapat berkonsentrasi pada kata-kata yang sangat relevan. Code untuk menghitung jumlah keseluruhan berada pada Lampiran 3 Listing Program Model Pelatihan.

3.3.3 *Count Vectorizer*



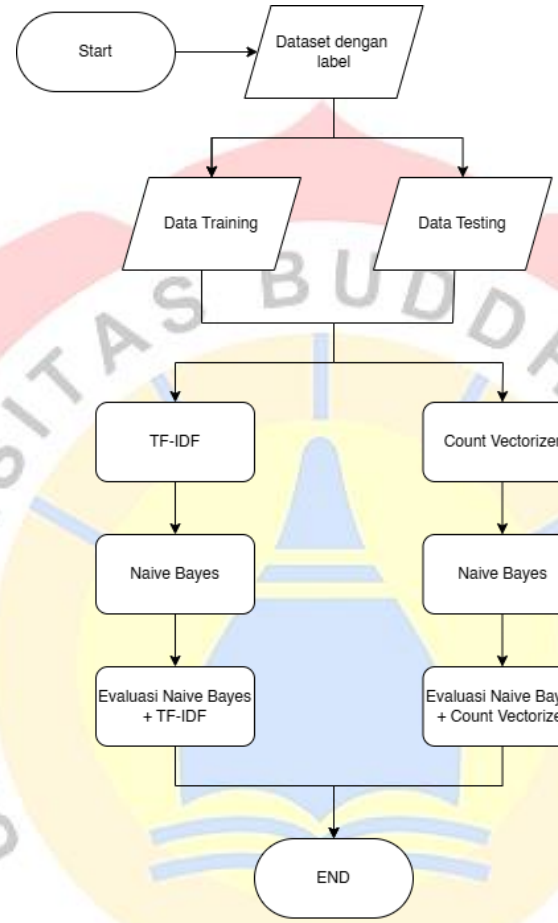
Gambar 3.7 *Flowchart* Pembobotan dengan *Count Vectorizer*

Pada tahap ini, dataset yang telah diberi label telah terbagi menjadi dua bagian, yaitu data pelatihan dan data pengujian, dan peneliti melakukan pembobotan kata menggunakan metode *Count Vectorizer*. Penelitian ini menggunakan model *Count Vectorizer* dengan parameter yang disesuaikan untuk meningkatkan kualitas representasi teks. Tujuan pertama dari penggunaan max_features=10.000 adalah untuk membatasi jumlah kata yang ada dalam representasi *Count Vectorizer*. Ini dilakukan untuk meningkatkan efisiensi komputasi,

berkonsentrasi pada kata-kata yang dianggap paling penting, dan mengurangi dimensi ruang fitur. Selain itu, parameter min_df=2 digunakan untuk menghindari pemodelan kata-kata acak yang tidak memberikan informasi penting, dan parameter max_df=0.8 digunakan untuk menghindari kata-kata yang muncul di lebih dari 80% dokumen. Ini meningkatkan kemampuan model untuk menemukan pola atau topik tertentu dengan menghindari kata-kata yang muncul di bawah tujuh dokumen. Penting untuk dicatat bahwa stopword dalam bahasa Indonesia diabaikan karena pemahaman bahwa kata-kata tersebut mungkin tidak memberikan kontribusi yang signifikan dalam analisis. Akibatnya, representasi *Count Vectorizer* dapat berkonsentrasi pada kata-kata yang sangat relevan. Code untuk menghitung jumlah keseluruhan berada pada Lampiran 3 Listing Program Model Pelatihan.

3.3.4 Naïve Bayes

Pada tahap ini, peneliti membangun model menggunakan algoritma *Naïve Bayes* pada data yang telah di-bobotkan sebelumnya menggunakan dua metode yaitu, TF IDF dan *Count Vectorizer*.



Gambar 3.8 *Flowchart* Model *Naive Bayes*

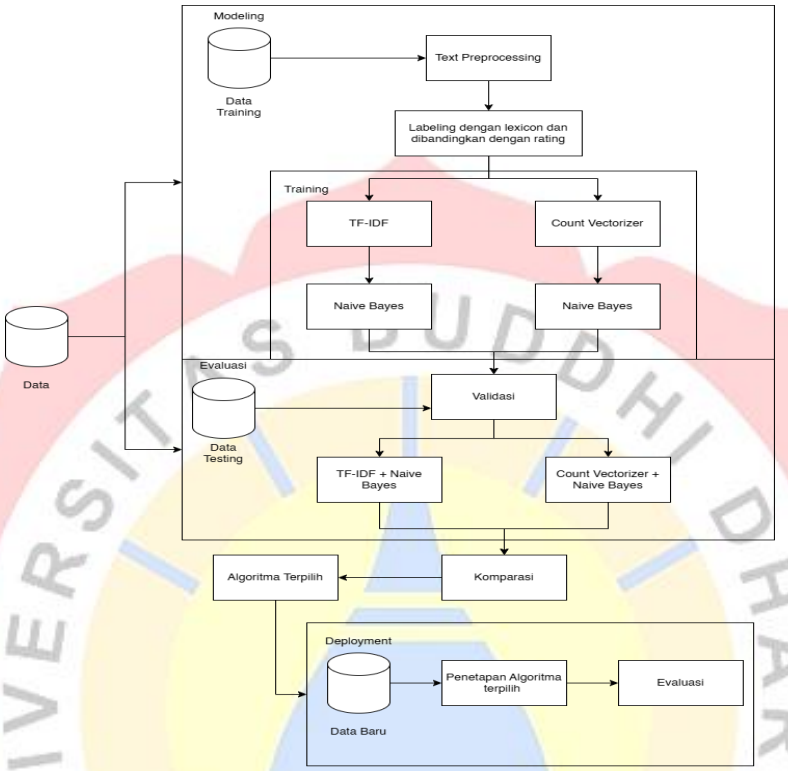
Dalam proses pemodelan. Peneliti menggunakan objek *NB library scikit-learn* dengan pengaturan parameter tertentu yang disesuaikan untuk kebutuhan analisis. Model yang digunakan adalah *MultinomialNB*, yang cocok untuk data dengan fitur bernilai frekuensi atau probabilitas, seperti hasil pembobotan dari *TF-IDF* dan *Count Vectorizer*.

3.3.5 *Paired T-test*

Untuk menguji statistik untuk memastikan perbedaan tersebut signifikan atau tidak digunakannya *Paired T-test* untuk menguji pembobotan kata TF-IDF dan *Count Vectorizer*

apakah terdapat perbedaan yang signifikan, terdapat pada Lampiran 4 Listing Program *Paired T-Test* untuk mengetahui tentang perbedaan yang didapatkan atau tidak.

3.4 Kerangka Pemikiran



Gambar 3.9 Kerangka Pemikiran

3.5 Metode Evaluasi

Dataset dalam penelitian ini dibagi menjadi dua bagian: 20% data pengujian dan 80% data pelatihan karena sudah terbukti berhasil meningkatkan hasil akurasi dari beberapa penelitian sebelumnya seperti (Bintang Zulfikar Ramadhan ,Oktaviani ,Agus tina ,Darmawan). Setelah itu, model dilatih dan diuji dengan data ini. Proses ini menghasilkan nilai. False positive (FP), false negative (FN), true positive (TP), dan true negative (TN) terdiri dari *Confusion Matrix*. Selain itu, nilai-nilai ini digunakan untuk menghitung metrik evaluasi seperti akurasi, presisi, recall, dan skor f-1.

3.6 Perancangan Prototype

3.6.1 Analisa Kebutuhan

Dalam tahapan ini, penelitian dilakukan dengan proses mengambil data komentar dari *Google Play Store* data Tokopedia dan Shopee sebesar 14.406 dari Shopee dan 13.853 data Tokopedia. Setelah itu dilakukan tahap *preprocessing* data, kemudian data melakukan pelabelan menggunakan Label inset yang didapatkan melalui Github pada , pelabelan dilakukan dengan memberi label (positif, negatif, dan netral) pada setiap data Tokopedia dan Shopee yang diperoleh dari web scrapping. Lalu data tersebut dilatih menjadi data training dan data testing dengan skenario 80:20 metode evaluasi halaman 51. Yang lalu dilakukan klasifikasi menggunakan algoritma *Naïve bayes*, dan menghitung akurasi menggunakan *Confusion Matrix*.

3.6.2 Perancangan Model

Tabel 3.10 Contoh data sudah di label

Id	Review	Sentimen
1.	"aplikasi ini bagus dan cepat"	Positif
2.	"fitur lengkap dan bermanfaat"	Positif
3.	"aplikasi biasa saja"	Netral
4.	"layanan cukup baik"	Netral
5.	"aplikasi ini lambat dan mahal"	Negatif
6.	"fitur terbatas dan kurang bagus"	Negatif

Data pada tabel 3.10 merupakan contoh data yang sudah diambil dan di preprocessing lalu dilabeling menggunakan kamus inset dengan kemudian akan dilakukan pembobotan kata TF-IDF maupun *Count Vectorizer*.

a) **Menghitung Pembobotan Kata *Count Vectorizer***

Tabel 3.11 Contoh Pembobotan Kata *Count Vectorizer*

Kata	Positif	Netral	Negatif
aplikasi	1	1	1
ini	1	0	1
bagus	1	0	1
cepat	1	0	0
fitur	1	0	1
lengkap	1	0	0
bermanfaat	1	0	0
biasa	0	1	0
saja	0	1	0
layanan	0	1	0
cukup	0	1	0
baik	0	1	0
lambat	0	0	1
mahal	0	0	1
terbatas	0	0	1
kurang	0	0	1

Deskripsi : angka 1 menunjukan bahwa ada kata tersebut dalam label, bila 0 tidak menunjukan ada kata didalam label tersebut.

Total Kata dalam Setiap Kategori

1. Total kata dalam review Positif = 6
2. Total kata dalam review Netral = 5
3. Total kata dalam review Negatif = 6
4. Total semua kata = 6 + 5 + 6 = 17

Dilanjuti dengan menghitung probabilitas priority

$$P(Positif) = \frac{Jumlah\ dokumen\ Positif}{Total\ dokumen} = \frac{2}{6} = 0.33$$

$$P(Netral) = \frac{Jumlah\ dokumen\ Netral}{Total\ dokumen} = \frac{2}{6} = 0.33$$

$$P(Negatif) = \frac{Jumlah\ dokumen\ Negatif}{Total\ dokumen} = \frac{2}{6} = 0.33$$

Dilanjuti dengan menghitung Probabilitas Kata

$$P(w_i|C) = \frac{(Frekuensi\ kata + 1)}{(Total\ kata\ dalam\ kategori + Jumlah\ kata\ unik)}$$

Jumlah kata unik = 16 **(**total kata berbeda dalam semua kategori**).**

Contoh perhitungan probabilitas kata **"bagus":**

$$P('bagus'|Positif) = \frac{(1\ +1)}{(6+16)} = \frac{2}{22} = 0.0909$$

$$P('bagus'|Netral) = \frac{(0\ +1)}{(5+16)} = \frac{1}{21} = 0.0476$$

$$P('bagus'|Negatif) = \frac{(1\ +1)}{(6+16)} = \frac{2}{22} = 0.0909$$

Semua kata dihitung dengan cara yang sama.

a) Klasifikasi

Misal ingin mengklasifikasikan **"fitur bagus dan cepat"**.

P(Positif | X) ∝ P(Positif) × P(fitur | Positif) × P(bagus | Positif) × P(dan |Positif)
× P(cepat | Positif)

P(Netral | X) ∝ P(Netral) × P(fitur |Netral) × P(bagus |Netral) × P(dan |Netral)
× P(cepat | Netral)

$P(\text{Negatif} \mid X) \propto P(\text{Negatif}) \times P(\text{fitur} \mid \text{Negatif}) \times P(\text{bagus} \mid \text{Negatif}) \times P(\text{dan} \mid \text{Negatif}) \times P(\text{cepat} \mid \text{Negatif})$

Tabel 3.12 Probabilitas Yang Sudah Dihitung

Kata	**Positif(P)**	**Netral (N)**	**Negatif (G)**
Fitur	0.0909	0.0476	0.0909
Bagus	0.0909	0.0476	0.0909
Dan	0.0476	0.0476	0.0476
Cepat	0.0909	0.0476	0.0476

Hitung probabilitas untuk tiap kelas menggunakan *Naïve bayes* :

$$P(\text{Positif} \mid X) = 0.33 \times 0.0909 \times 0.0909 \times 0.0476 \times 0.0909 = 0.33 \times 0.0000035 = 0.0000012$$

$$P(\text{Netral} \mid X) = 0.33 \times 0.0476 \times 0.0476 \times 0.0476 \times 0.0476 = 0.33 \times 0.0000051 = 0.0000017$$

$$P(\text{Negatif} \mid X) = 0.33 \times 0.0909 \times 0.0909 \times 0.0476 \times 0.0476 = 0.33 \times 0.0000018 = 0.0000006$$

Kesimpulan :

P(Positif | X) = 0.0000012

P(Netral | X) = 0.0000017

P(Negatif | X) = 0.0000006

Karena P(Netral) paling tinggi, maka review diklasifikasikan sebagai Netral.

b) Menghitung Pembobotan Kata TF-IDF

Tabel 3.13 Contoh Pembobotan Kata TF-IDF

Kata	D1	D2	D3	D4	D5	D6	DF	IDF
aplikasi	1	0	1	0	1	0	3	Log(6/3) = 0.301
ini	1	0	0	0	1	0	2	Log(6/2) = 0.477
bagus	1	0	0	0	0	1	2	log(6/2) = 0.477
dan	1	1	0	0	1	1	4	log(6/4) = 0.176
cepat	1	0	0	0	0	0	1	log(6/1) = 0.778
fitur	0	1	0	0	0	1	2	log(6/2) = 0.477
lengkap	0	1	0	0	0	0	1	log(6/1) = 0.778
bermanfaat	0	1	0	0	0	0	1	log(6/1) = 0.778
biasa	0	0	1	0	0	0	1	log(6/1) = 0.778
saja	0	0	1	0	0	0	1	log(6/1) = 0.778
layanan	0	0	0	1	0	0	1	log(6/1) = 0.778
cukup	0	0	0	1	0	0	1	log(6/1) = 0.778
baik	0	0	0	1	0	0	1	log(6/1) = 0.778
lambat	0	0	0	0	1	0	1	log(6/1) = 0.778
mahal	0	0	0	0	1	0	1	log(6/1) = 0.778
terbatas	0	0	0	0	0	1	1	log(6/1) = 0.778
kurang	0	0	0	0	0	1	1	log(6/1) = 0.778

Deskripsi : angka 1 menunjukan bahwa ada kata tersebut dalam label, bila 0 tidak menunjukan ada kata didalam label tersebut. DF didapatkan dari hasil keseluruhan pada kata tersebut contoh kata “aplikasi” berada di D1,D3,D5 dan muncul 1 kali maka 1 x 3 = 3

Contoh menghitung IDF (*aplikasi)* = $\log\left(\frac{5}{3}\right) = \log(1.67) = 0.22$

Contoh menghitung TF-IDF (*aplikasi*) = 1 × 0.301 = 0.301

(ini) (D1) = 1 × 0.477 = 0.477

(bagus) (D1) = 1 × 0.477 = 0.477

(dan) (D1) = 1 × 0.176 = 0.176

(cepat) (D1) = 1 × 0.778 = 0.778

Menghitung Probabilitas Prior P(C) Naïve Bayes :

$$P(Positif) = \frac{2}{5} = 0.4$$

$$P(netral) = \frac{1}{5} = 0.2$$

$$P(Negatif) = \frac{2}{5} = 0.4$$

Hitung Probabilitas Kata dalam Setiap Kelas Menggunakan TF-IDF Misalnya, ingin klasifikasikan "fitur bagus":

$$\mathbf{P(Positif) \times P(\ fitur \mid Positif\ ) \times P(bagus \mid Positif) 0.4 \times 0.130 \times 0.078}$$

$$\mathbf{= 0.004}$$

$$\mathbf{P(Netral) \times P(\ fitur \mid Netral\ ) \times P(bagus \mid Netral) 0.2 \times 0 \times 0 = 0}$$

$$\mathbf{P(Negatif) \times P(\ fitur \mid Negatif\ ) \times P(bagus \mid Negatif\ )}$$

$$\mathbf{= 0.0040.4 \times 0.078 \times 0.078 = 0.002}$$

Karena P(Positif) lebih besar, maka review diklasifikasikan sebagai Positif.

c) Menghitung Akurasi pada *Confusion Matrix*

Tabel 3.14 Contoh *Confusion Matrix*

Label Actual	**Positif**	**Netral**	**Negatif**
Positif(2)	2(TP)	0(TN)	2(FN)
Netral(1)	1(FP)	0(TP)	0(FN)
Negatif(2)	0(FP)	0(FN)	2(TP)

True Positive (TP): Model benar dalam memprediksi dokumen sebagai positif berdasarkan bobot kata yang relevan. Disitu ada label actual positif(2) lalu menunjukan bahwa 2 data tersebut sebagai positif

False Positive (FP): Model salah memprediksi dokumen sebagai positif, padahal seharusnya negatif. Ada label actual Netral(1) namun hasil positif 1 artinya salah memprediksi

False Negative (FN): Model salah memprediksi dokumen sebagai negatif, padahal seharusnya positif. Positif(2) namun di prediksi 2 FN seharusnya itu positif

True Negative (TN): Model benar dalam memprediksi dokumen sebagai negatif berdasarkan bobot kata yang relevan.

Menghitung Akurasi :

$$Akurasi = \frac{TP}{Total\ Sampel} = \frac{2+0+3}{6} = 0.8(80\%)$$

2 TP positif. 0 Tp netral, 3 TP Negatif dengan jumlah 6 data dan hasil yang didapatkan adalah 80%

Precision setiap kelas

$$Precision_{positif} = \frac{TP_{Positif}}{TP_{Positif} + FP_{Positif}} = \frac{2}{2+1} = 0.67$$

$$Precision_{netral} = \frac{TP_{Netral}}{TP_{Netral} + FP_{Netral}} = \frac{0}{0+0} = 0.67$$

$$Precision_{negatif} = \frac{TP_{negatif}}{TP_{negatif} + FP_{negatif}} = \frac{2}{2+0} = 1$$

Recal Netral :

$$Recal_{positif} = \frac{TP_{Positif}}{TP_{Positif} + FP_{Positif}} = \frac{2}{2+0} = 1$$

$$Recal_{netral} = \frac{TP_{Netral}}{TP_{Netral} + FP_{Netral}} = \frac{0}{0+1} = 0$$

$$Recal_{negatif} = \frac{TP_{negatif}}{TP_{negatif} + FP_{negatif}} = \frac{2}{2+0} = 1$$

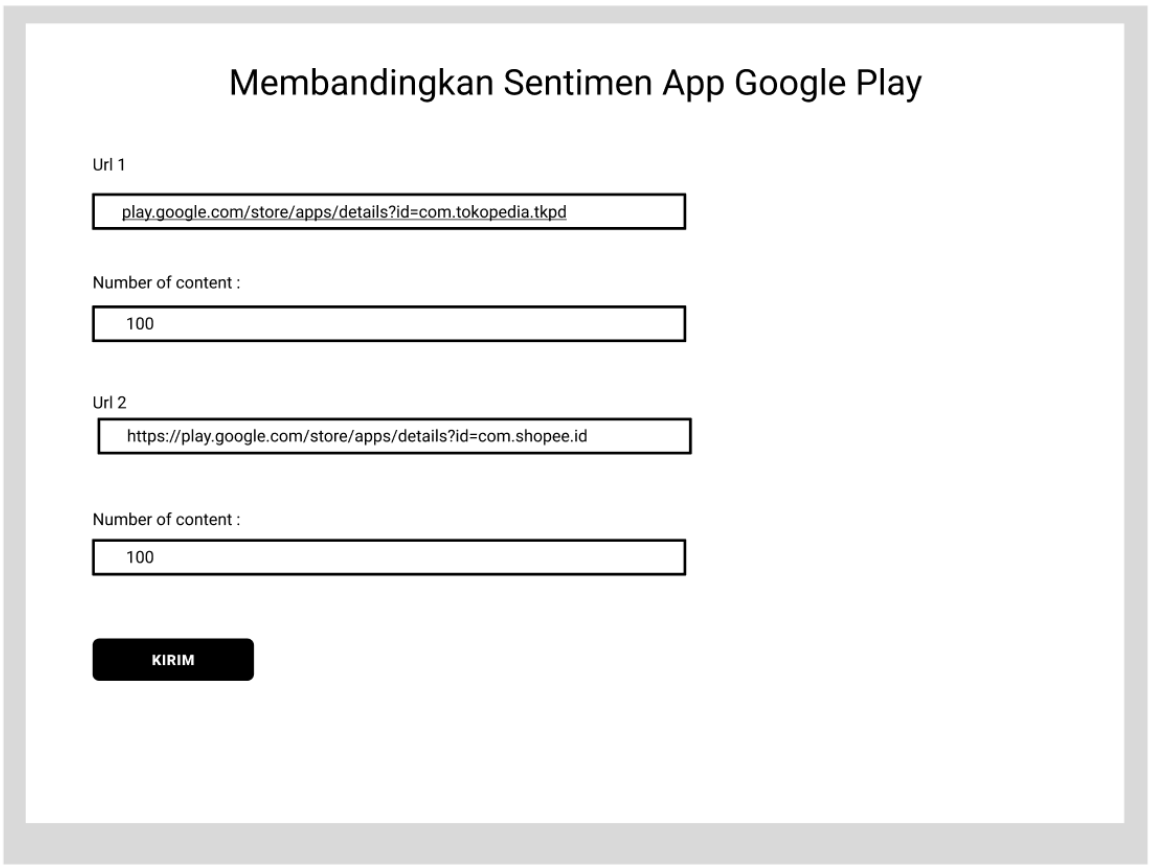
F1-Score Negatif :

$$f1_{Positif} = 2 * \frac{Precision * Recall}{Precisiion + Recall} = 2 * \frac{0.67 * 1}{0.67 + 1} = 0.8$$

$$f1_{negaitf} = 2 * \frac{1 * 1}{1 + 1} = 1$$

3.6.3 Perancangan Aplikasi

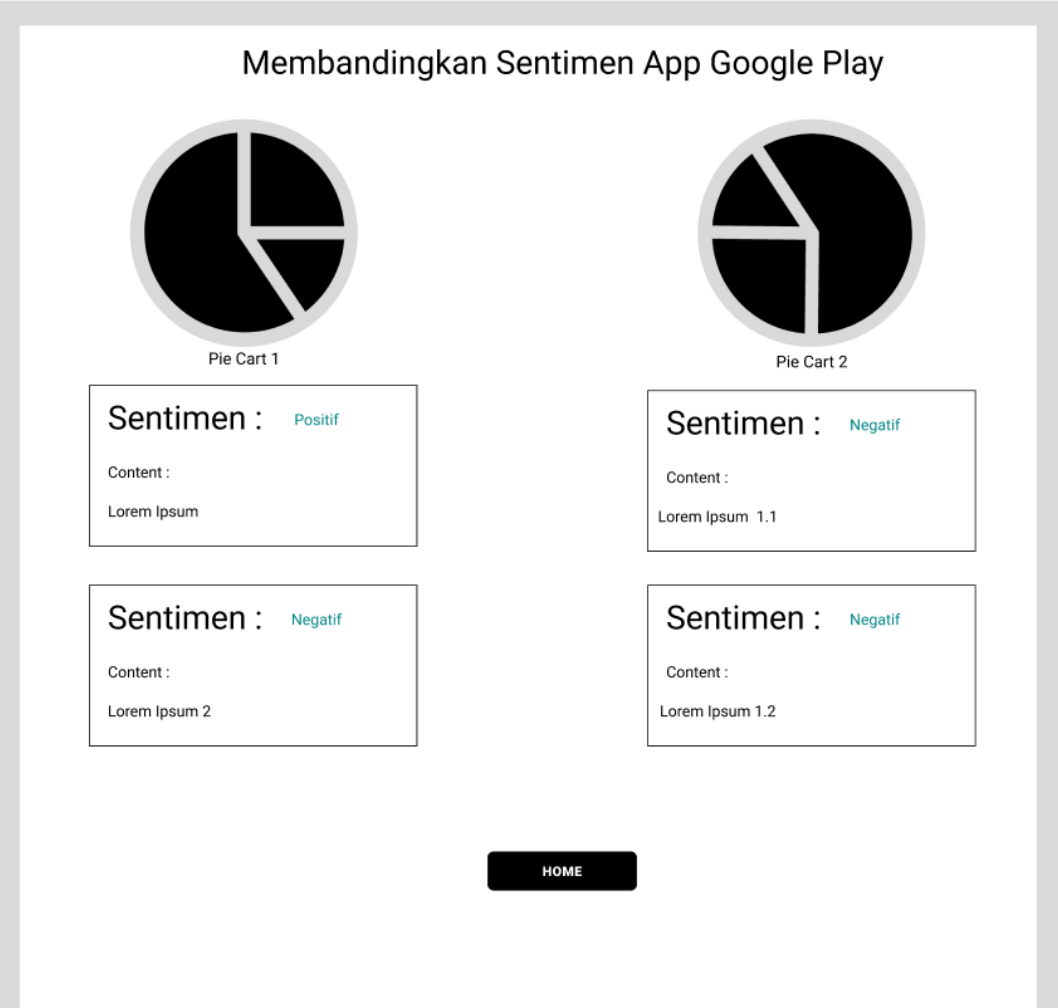
Hasil analisa ulasan pada platform *Google Play Store* ditampilkan dengan pie chart, jumlah sentimen, komentar, dan hasil sentimen pada komentar. Berdasarkan jumlah data yang diambil pada menu utama. Berikut adalah tampilan *prototype* dari aplikasi website membandingkan sentimen app play store.



Gambar 3.10 *Prototype* Halaman Utama

Pada gambar 3.10 ini merupakan contoh hasil yang akan dibuat peneliti dengan terdapat 4 kolom, kolom url 1 atau 2 untuk menginput link url dari *platform*

Google Play Store yang terdapat di gambar itu merupakan link pada aplikasi tokopedia (Url1) dan shopee (Url2). Dan dibawah kolom url terdapat Number of content dimana nantinya user bisa mengisi jumlah angka lebih dari 1, contoh di gambar adalah 100 angka.



Gambar 3.11 *Prototype* Halaman Hasil

Pada Gambar 3.11 ini merupakan contoh hasil setelah url 1-2 dan number of content 1-2 terisi dan terkirim lalu hasil terdapat 2 *Pie chart* untuk menampilkan total keseluruhan jumlah sentimen (positif,negatif,netral) dan hasil komentar pada kedua aplikasi, untuk sebelah bagian kiri itu merupakan hasil inputan Url 1 dan number of content 1 dan sebelah kanan Url 2 dan number of content 2.

3.6.4 *Construction of Prototype*

Dalam pengembangan sistem atau aplikasi, tahap pembuatan *prototype* merupakan langkah krusial yang menghubungkan antara perancangan konseptual dan implementasi akhir. Pada fase ini, ide-ide yang telah dirumuskan dalam perancangan sebelumnya.Konstruksi *prototype* tidak hanya melibatkan pembuatan antarmuka pengguna, tetapi juga mencakup pengembangan *script Function* yang akan mengatur logika dan interaksi dalam aplikasi. Berikut merupakan beberapa *scipt Function* yang digunakan.

```
from google_play_scraper import reviews, Sort
from google_play_scraper.exceptions import NotFoundError
result, continuation_token = reviews(
‘play.google.com/store/apps/details?id=com.tokopedia.tkpd’
lang='id',country='id',sort=Sort.Newest,count = 100, filter_score_with =
None )
```

Function diatas Untuk mengambil ulasan dari Google Play menggunakan play scraper, fungsi yang disediakan untuk mengambil data ulasan berdasarkan ID aplikasi contoh tokopedia. Proses ini melibatkan pengaturan parameter seperti jumlah ulasan yang ingin diambil dan jenis data yang diinginkan.

```
df['cleaning']= df['content'].apply(remove_emoji)
    logger.info('Removed emojis')
     # 2. Basic cleaning
    df['cleaning'] = df['cleaning'].apply(clean_text)
    logger.info('Applied basic text cleaning')
    # 3. Case folding
    df['cleaning']= df['cleaning'].apply(case_folding)
    # 4. Normalisasi (dengan tambahan root words)
    df['normalisasi'], _, _, _ = zip(
        *df['cleaning'].apply(lambda x: replace_taboo_words(x,
kamus_tidak_baku, root_words))
    )
    logger.info('Applied word normalization')
    # 5. Tokenization
    df['tokenize'] = df['normalisasi'].apply(tokenize)
    logger.info('Applied tokenization')
```

```
    # 6. Stopword removal
    df['stopword_removal'] = df['tokenize'].apply(lambda x:
remove_stopwords(x, stop_words))
    logger.info('Removed stopwords')
    # 7. Stemming
    df['stemming_data'] = df['stopword_removal'].apply(lambda x: '
'.join(stem_text(x)))
    logger.info('Applied stemming')
    logger.info("Finished processing data")
```

Function untuk melakukan tahapan preprocessing data dari remove emoji, case folding,normaliasi, tokenizing, stop word removal, dan stemming (proses bisa dilihat pada Bab 3 Pengolahan Data Text preprocessing halaman 35-38)

```
def _initialize_lexicons(self) -> None:
        """Load positive and negative lexicons from files"""
        try:
            pos_path = self.lexicon_dir / '_json_inset-pos.txt'
            neg_path = self.lexicon_dir / '_json_inset-neg.txt'
            with open(pos_path, 'r', encoding='utf-8') as f:
                self.pos_lexicon = json.load(f)
            with open(neg_path, 'r', encoding='utf-8') as f:
                self.neg_lexicon = json.load(f)
```

Function untuk membaca kamus labeling positif dan negatif melalui github yang berbentuk format txt

Dalam tabel diatas dimulai dari *Function* mengambil data menggunakan *google-play-scraper*, lalu dilanjuti *preprocessing data*, *Labeling* data dan Modeling data untuk dilatih menjadi aplikasi. Selanjutnya tabel 3.16. Yang merupakan tahapan untuk merancang aplikasi website tersebut, berikut adalah *script Front-End Function* untuk membuat halaman website aplikasi.

```
    <div class="data-card mt-4">
        <form id="scrapeForm" class="scrapeForm" action="/compare"
method="POST">
            <h4>Input Data 1</h4>
            <div class="mb-3">
                <label for="url1" class="form-label">Google Play Store URL
1:</label>
                <input type="url" class="form-control" id="url1" name="url1"
required>
```

```
            </div>
            <div class="mb-3">
                <label for="num_reviews1" class="form-label">Number of
Reviews 1:</label>
                <input type="number" class="form-control" id="num_reviews1"
name="num_reviews1" min="1" required>
                <small class="form-text text-muted">

                </small>
            </div>
```

Function tersebut untuk menampilkan judul halaman menggunakan form dengan id "scrapeForm" yang akan mengirimkan data ke endpoint "/compare" menggunakan metode POST, Input URL Google Play Store (tipe url) Input jumlah review yang ingin dianalisis (tipe number) dirancang untuk mengumpulkan URL aplikasi dan menentukan berapa banyak review yang ingin dianalisis sentimennya.

```
<h2>Prediction Result Pie Charts</h2>

        <div class="chart-container">
            <div class="chart-wrapper">
                <canvas id="myPieChart1"></canvas>
            </div>
            <div class="chart-wrapper">
                <canvas id="myPieChart2"></canvas>
            </div>
        </div>
```

Function tersebut untuk menampilkan pie chart untuk mengetahui jumlah keseluruhan sentimen yang didapatkan.

```
        <div class="word-stats">
            <div class="word-stats-column">
                <h2>Kata Paling Banyak Pada - App 1</h2>
                <ul>
                    {% for word, count in most_common_words1 %}
                    <li>
                        <span>{{ word }}</span>
                        <span class="word-count">{{ count }}</span>
                    </li>
                    {% endfor %}
                </ul>
```

Function ini akan menampilkan daftar kata-kata yang paling sering muncul dalam review aplikasi pertama, memberikan insight tentang tema atau topik yang sering dibahas oleh pengguna dalam review.

3.6.5 *Deployment Prototype*

Tahapan ini merupakan hasil akhir dari pada *prototype* itu sendiri yang artinya dari analisa kebutuhan sampai pembuatan *prototype* di implementasikan menjadi aplikasi website, yang nantinya akan dapat menjalankan analisis sentimen membandingkan aplikasi *Google play store* Tokopedia dan Shopee. hasil dari tampilan website yang telah dibuat dapat dilihat pada BAB IV Implementasi Website pada Gambar 4.16 – Gambar 4.21 halaman 85 – 87.