ANALISIS SENTIMEN TENTANG CHATGPT PADA KOLOM KOMENTAR SITUS REDDIT DENGAN METODE *LEXICON BASED*

SKRIPSI

OLEH:

PAGAS SUKARNO PUTRA

198160044

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS MEDAN AREA

MEDAN

2024

ANALISIS SENTIMEN TENTANG CHATGPT PADA KOLOM KOMENTAR SITUS REDDIT DENGAN METODE *LEXICON BASED*

SKRIPSI

Diajukan Sebagai Salah Satu Syarat untuk Memperoleh

Gelar Sarjana di Fakultas Teknik

Universitas Medan Area

Oleh:

PAGAS SUKARNO PUTRA

198160044

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS MEDAN AREA

MEDAN

2024

HALAMAN PENGESAHAN

Judul Skripsi	: Analisis Sentimen Tentang Chatgpt Pada Kolom Komentar Situs Reddit Dengan Metode Lexicon Based
Nama	: Pagas Sukarno Putra
NPM	: 198160044
Fakultas	: Teknik
Program Studi	: Teknik Informatika

Disetujui Oleh:

Komisi Pembimbing

Andre Hasudungan Lubis, S.Ti, MSe
Pembimbing

Diketahui:

Dr. Ing. Supriatno, ST, MT
Dekan Fakultas Teknik

Rizki Muliono, S.Kom, M.Kom
Ketua Program Studi

Tanggal Lulus: 28 Maret 2024

HALAMAN PERNYATAAN

Saya menyatakan bahwa skripsi yang saya susun sebagai syarat memperoleh gelar sarjana merupakan hasil karya tulis saya sendiri. Adapun bagian-bagian tertentu dalam penulisan skripsi ini yang saya kutip dari hasil karya orang lain telah dituliskan sumbernya dengan jelas sesuai dengan norma, kaidah, dan etika penulisan ilmiah.

Saya bersedia menerima sanksi pencabutan gelar akademik yang saya peroleh dan sanksi-sanksi lainnya dengan peratutan yang berlaku apabila dikemudian hari ditemukan adanya plagiat dalam skripsi ini.

Medan, 28 Maret 2024

Pagas Sukarno Putra

198160044

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR/SKRIPSI/TESIS UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Medan Area, saya yang bertanda tangan dibaawah ini:

Nama	: Pagas Sukarno Putra
NPM	: 198160044
Fakultas	: Teknik
Program Studi	: Teknik Informatika
Jenis Karya	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Medan Area **Hak Bebas Royalti Nonekslusif (Non-exclusive Royalty-Free Right)** atas karya ilmiah saya yang berjudul: **"Analisis Sentimen tentang ChatGPT pada Kolom Komentar Situs Reddit dengan Metode *Lexicon Based*"**. Bersama dengan perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Nonekslusif ini Universitas Medan Area berhak menyimpan, mengalihkan media/format, mengelola dalam bentuk pangkalan *database*, memelihara, dan mempublikasikan skripsi saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Medan
Pada tanggal : 28 Maret 2024
Yang menyatakan

(Pagas Sukarno Putra)

ABSTRAK

ChatGPT merupakan salah satu kecerdasan buatan yang saat ini tengah banyak digunakan oleh pengguna. Aplikasi *chatbot* ini, dirilis pada November 2022 oleh sebuah laboratorium riset kecerdasan buatan yang bernama OpenAI. Setelah peluncurannya, ChatGPT mendapatkan beragam komentar dari para pengguna internet di berbagai platform dengan isi komentar yang positif, netral, ataupun negatif, seperti pada situs Reddit. Banyaknya pengguna Reddit mengakibatkan sulitnya untuk menganalisis apakah komentar yang diutarakan bersentimen positif, netral atau negatif. Oleh karena itu, dibutuhkan suatu teknik serta metode yang sesuai untuk permasalahan tersebut. Penelitian ini menggunakan teknik analisis sentimen dengan metode *Lexicon Based* dan kamus kata VADER untuk mencari polaritas kata yang bersentimen positif, netral, dan negatif dan menggunakan data komentar dari situs Reddit sebagai data penelitian. Dengan tahapan penelitian yang dilakukan yaitu pengumpulan data, *data reduction*, *text preprocessing*, implementasi metode *Lexicon* VADER, hasil analisis sentimen, validasi penelitian dan juga visualisasi data. Hasil dari penelitian ini yaitu data komentar yang memiliki jenis sentimen positif berjumlah 14.250 komentar dengan presentase sebanyak 53,2%, sentimen netral berjumlah 6.767 komentar dengan presentase sebanyak 24.8% dan sentimen negatif berjumlah 6.009 dengan presentase sebanyak 22.0% dari total keseluruhan jumlah data komentar sebanyak 27.296 data. Hasil validasi *clustering* menggunakan *Silhouette Index* mendapatkan nilai 0.532 dengan nilai mendekati angka 1 maka kualitas klasterisasi sudah baik. Penelitian ini menunjukkan bahwa komentar pengguna Reddit terhadap ChatGPT memiliki sentimen yang positif.

Kata Kunci: Analisis Sentimen, Reddit, ChatGPT, Lexicon Based, VADER

ABSTRACT

ChatGPT is an artificial intelligence that is currently used by many users. This chatbot application was released in November 2022 by an artificial intelligence research laboratory called OpenAI. After its release, ChatGPT received various comments from netizens on different platforms with positive, neutral, or negative comments, such as the Reddit website. The large number of Reddit users makes it difficult to analyze whether the comments expressed are positive, neutral, or negative. Therefore, a technique and method that is suitable for this problem is needed. This research used sentiment analysis techniques with the Lexicon Based method and the VADER word dictionary to search for the polarity of words with positive, neutral, and negative sentiments, and used comment data from the Reddit website as research data. The research stages performed were data collection, data reduction, text preprocessing, implementation of the Lexicon-VADER method, sentiment analysis results, research validation, and data visualization. The results of this research were that the comment data that had a positive sentiment type were 14,250 comments with a percentage of 53.2%, neutral sentiment were 6,767 comments with a percentage of 24.8%, and negative sentiment were 6,009 with a percentage of 22.0% of the total number of comment data of 27,296 records. The results of clustering validation using the Silhouette Index obtained a value of 0.532 with a score close to 1, which means the quality of clustering was good. This research showed that Reddit users' comments against ChatGPT had positive sentiments.

***Keywords**: Analysis of Sentiment, Reddit, ChatGPT, Lexicon Based, VADER*

RIWAYAT HIDUP

Penulis dilahirkan di Sampali pada tanggal 06 Juni 2001 dari Bapak Parno dan Ibu Suyatini. Penulis merupakan anak ke-dua dari tiga bersaudara dan memiliki adik laki-laki serta kakak perempuan. Penulis pertama kali mengenyam pendidikan di bangku SD Negeri 104202 Bandar Setia pada tahun 2007 dan lulus pada tahun 2013. Kemudian penulis melanjutkan pendidikan ke jenjang SMP pada tahun 2013 di SMP Negeri 35 Medan dan lulus pada tahun 2016. Pada tahun yang sama, penulis melanjutkan ke jenjang selanjutnya yaitu di SMK Negeri 1 Percut Sei Tuan dan lulus pada tahun 2019. Pada bulan September tahun 2019, penulis melanjutkan pendidikan di bangku kuliah dan terdaftar sebagai mahasiswa Fakultas Teknik dengan mengambil Program Studi Teknik Informatika di Universitas Medan Area.

KATA PENGANTAR

Puji syukur kehadiran Tuhan Yang Maha Esa, atas berkat dan karunia-Nya sehingga penulis dapat menyelesaikan penulisan skripsi yang berjudul "Analisis Sentimen Tentang ChatGPT Pada Kolom Komentar Situs Reddit Dengan Metode *Lexicon Based*". Skripsi ini merupakan salah satu syarat untuk menyelesaikan pendidikan untuk mencapai gelar sarjana di Program Studi Teknik Informatika Fakultas Teknik Universitas Medan Area.

Pada kesempatan ini, penulis mengucapkan banyak terima kasih juga kepada pihak-pihak yang telah memberikan banyak dukungan serta arahan sehingga penulis bisa menyelesaikan penelitian ini dengan baik, untuk itu penulis menyampaikan ucapan terimakasih kepada:

1. Bapak Prof. Dr. Dadan Ramdan, M.Eng, M.Sc. selaku Rektor Universitas Medan Area.
2. Dr.Eng. Supriatno, ST, MT selaku Dekan Fakultas Teknik Universitas Medan Area.
3. Bapak Rizki Muliono, S.Kom, M.Kom selaku Kepala Program Studi Teknik Informatika Universitas Medan Area.
4. Bapak Andre Hasudungan Lubis, S.Ti, M.Sc, selaku Dosen pembimbing yang telah membantu penulis dari segi materi sehingga penulis dapat menyelesaikan skripsi ini.
5. Orang tua penulis yaitu Bapak Parno dan Ibu Suyatini yang telah mendoakan tiada henti dan memberikan semangat serta membantu penulis dalam segi

materi dan moril sehingga penulis dapat menyelesaikan skripsi ini dengan sebaik baiknya.

6. Seluruh Dosen dan Staf Program Studi Teknik Informatika Universitas Medan Area.
7. Seluruh teman-teman yang sudah memberikan dukungannya selama penulisan proposal skripsi ini, khususnya teman-teman Teknik Informatika angkatan 2019.
8. Seluruh pihak yang tidak dapat disebutkan satu persatu yang membantu dalam menyelesaikan skripsi ini.

Penulis menyadari bahwa penelitian ini masih memiliki kekurangan,oleh karena itu kritik dan saran yang bersifat membangun sangat penulis harapkan demi kesempurnaan penelitian ini. Penulis berharap tugas penelitian ini dapat bermanfaat baik kalangan pendidikan maupun masyarakat. Akhir kata penulis ucapkan terima kasih.

Medan, 28 Maret 2024

Penulis,

Pagas Sukarno Putra

NPM 198160044

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pada era perkembangan teknologi sekarang ini, kecerdasaan buatan atau *Artificial Intelligence* (AI) sudah sangat berkembang pesat. Kecerdasan buatan ialah sebuah istilah yang mengacu pada simulasi proses kecerdasan dan pemikiran manusia oleh mesin yang terhubung dengan lautan data dan informasi. Kecerdasan buatan atau AI pada kenyataannya bisa dilihat dari aplikasi *smartphone*, mobil *self-driving*, mesin otomatis dan robot di perusahaan hingga kamera pengawas dengan pengenalan wajah. Kita juga tazhu bahwa ada alat-alat canggih seperti *Cortana, Siri, Alexa, dan Google Assistant,* yang merupakan asisten cerdas untuk mempermudah hidup manusia (Pabubung, 2021).

Salah satu kecerdasan buatan yang sekarang sedang banyak digunakan yaitu ChatGPT. ChatGPT ialah aplikasi *chatbot* yang dirilis pada November 2022 oleh sebuah laboratorium riset kecerdasan buatan asal Amerika Serikat yang bernama OpenAI. Mesin ini menerapkan teknologi *Natural Language Processing* (NLP) atau pemrosesan bahasa alami yang dapat menjawab pertanyaan seseorang dalam bentuk teks (disebut *prompt*) yang diketikkan dalam aplikasi ChatGPT (Setiawan & Luthfiyani, 2023).

Setelah diluncurkannya ChatGPT, *chatbot* AI ini mendapatkan beragam komentar dari para pengguna internet di seluruh dunia. Terdapat komentar yang positif akan hadirnya ChatGPT, dan terdapat pula komentar yang kurang menyenangkan atau bernilai negatif. Komentar mengenai ChatGPT banyak

diutarakan oleh pengguna Internet di berbagai *platform* seperti di media sosial, forum-forum serta situs-situs yang membahas tentang ChatGPT. Salah satu forum yang sering membahas tentang ChatGPT yaitu forum Reddit. Reddit terkenal sebagai wadah untuk berdiskusi berbagai topik dengan bentuk sebuah konten. Reddit didirikan pada tahun 2005 di Amerika Serikat oleh Steve Huffman, Aaron Swartz, dan Alexiz Ohanian. Hingga saat ini, Reddit telah memiliki pengguna sebanyak 300 juta lebih di seluruh dunia (Rachmawaty, 2021).

Dengan banyaknya pengguna pada forum Reddit, komentar mengenai suatu topik akan sangatlah banyak seperti halnya komentar mengenai hadirnya ChatGPT ini. Untuk mengetahui apakah komentar-komentar yang diutarakan oleh pengguna Reddit mengandung sentimen terhadap ChatGPT maka dibutuhkan suatu teknik serta metode untuk menganalisis komentar-komentar tersebut. Sehingga pada penelitian ini akan membahas teknik dan metode dalam menganalisis komentar pengguna Reddit dengan topik ChatGPT. Pentingnya penelitian ini dilakukan ialah apakah menganalisis komentar dengan suatu teknik dan metode dapat mengetahui sentimen pengguna Reddit terhadap ChatGPT dan apakah komentar tersebut memiliki sentimen yang positif, negatif ataupun netral sehingga pihak OpenAI selaku pengembang ChatGPT dapat mengetahui sentimen publik mengenai ChatGPT serta mengetahui kekurangan dan kelebihan ChatGPT dari komentar-komentar tersebut.

Analisis sentimen ialah teknik yang sering digunakan dalam menganalisis komentar. Analisis sentimen adalah teknik untuk mengolah bermacam-macam pendapat yang diungkapkan dalam bentuk teks dengan menggunakan media yang berkaitan erat pada produk, atau lainnya (Habibah dkk., 2023). Dalam melakukan

analisis sentimen untuk tahap mengklasifikasikan data membutuhkan suatu metode yang bertujuan mempermudah pengguna mengetahui pandangan yang termasuk kategori positif, negatif, atau netral serta mendapatkan presentase akurasi pada setiap sentimen serta akurasi pada kinerja metode (Amaliah & Nuryana, 2022). Untuk menganalisis sentimen, ada beberapa metode yang bisa dipakai, salah satunya yaitu menggunakan metode *Lexicon Based*.

Metode *Lexicon Based* ialah pendekatan yang praktis, sederhana serta efektif dalam melakukan analisis sentimen. Facebook, Twitter dapat dimanfaatkan sebagai sumber data serta platform media sosial lainnya yang mencerminkan pendapat tentang suatu produk atau layanan. Keuntungan menggunakan metode *Lexicon Based* adalah tidak memerlukan data yang sudah diberi label atau proses pembelajaran khusus. Kata-kata dalam metode *Lexicon Based* akan dinilai berdasarkan nilai polaritas yang bermaksud untuk memahami pendapat atau tenggapan masyarakat (Hernikawati , 2021).

Dalam metode *Lexicon* Based memerlukan kamus kata sebagai acuan nilai polaritas dari kata yang akan dianalisis. Kamus kata yang sering dipakai dibeberapa penilitian terdahulu yaitu VADER. VADER akan menganalisis teks mengikuti *lexicon* (*a library*) yang menghasilkan kelas sentimen berupa positif, negatif dan netral dengan tambahan *compound score* atau skor total (Sumitro dkk., 2021).

Beberapa penelitian terdahulu yang menggunakan metode *Lexicon Based* seperti penelitian yang dilakukan oleh (Sumitro dkk., 2021) yang berjudul "Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode *Lexicon Based*" dengan menggunakan metode *Lexicon based* dan juga

kamus kata VADER yang menghasilkan presentase sentimen positif 20,25%, agak positif 23,9%, negatif 1,88%, agak negatif 9,6% dan netral 44,36%.

Pada penelititan (Amaliah & Nuryana, 2022) yang berjudul "Perbandingan Akurasi Metode *Lexicon Based* Dan *Naïve Bayes Classifier* Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter" dengan menggunakan metode *Lexicon Based* dengan kamus kata VADER serta metode *Naïve Bayes Classifier*, dengan jumlah 147 data *tweets* mempunyai hasi akurasi masing-masing sebesar 67% dengan polaritas tertinggi yaitu positif sebesar 64,63% dan akurasi 78% dengan polaritas tertinggi yaitu positif sebesar 53,74%.

Dan juga penelitian (Habibah dkk., 2023) dengan judul "Analisis Sentimen Mengenai Penggunaan E-Wallet Pada Google Play Menggunakan *Lexicon Based* dan *K-Nearest Neighbor*" dengan menggunakan *Lexicon Based* dan KNN pada aplikasi Dana menghasilkan akurasi tertinggi pada k=6 sebesar 78% serta sentimen positif 4.921 ulasan, negatif 2.550 ulasan, dan netral 1.529. Pada aplikasi Ovo menghasilkan akurasi tertinggi pada k=9 sebesar 75,33% dengan sentimen positif 5.289 ulasan, negatif 1.789 ulasan dan netral 1.992 ulasan. Pada aplikasi LinkAja menhasilkan akurasi tertinggi pada k=8 sebesar 73,5% dengan sentimen positif 6.037 ulasan, negatif 1.705 ulasan dan netral 1.285 ulasan.

Berdasarkan pemaparan yang telah disampaikan diatas, penelitian tugas akhir ini akan membahas mengenai sentimen para pengguna Reddit berdasarkan komentar yang telah diutarakan pada kolom komentar *subreddit* ChatGPT dengan menggunakan metode *Lexicon Based* dengan kamus kata VADER sehingga dapat

mengetahui sentimen pengguna Reddit terhadap ChatGPT apakah memiliki sentimen yang positif, negatif, ataupun netral.

Untuk data yang akan digunakan pada penelitian tugas akhir ini yaitu data sekunder berupa *dataset* dari situs penyedia *dataset* yaitu Kaggle.com yang berjudul *"ChatGPT Reddit*. Pada *dataset* ini akan dilakukan *pre-processing* terlebih dahulu dengan bantuan bahasa pemrograman *phyton* dan Google Colaboratory sehingga data dapat diolah untuk mengetahui jenis sentimen dengan metode *Lexicon Based* berdasarkan kamus kata VADER.

1.2 Rumusan Masalah

Dari penjelasan pada latar belakang diketahui rumusan masalah pada penelitian tugas akhir ini adalah bagaimana mengetahui sentimen dari beragam jenis komentar tentang ChatGPT di situs Reddit dengan menerapkan metode *Lexicon Based* berdasarkan kamus kata VADER dan mengelompokkan komentar tersebut ke dalam 3 jenis sentimen sehingga diperoleh apakah komentar tersebut positif, negatif, atau netral.

1.3 Batasan Masalah

Berdasarkan rumusan masalah diatas, maka diketahui batasan masalah pada penelitian tugas akhir ini yaitu:

1. Menggunakan metode *Lexicon Based* sebagai metode pada penelititan.
2. Menggunakan kamus kata VADER yang berbasis bahasa Inggris .
3. Jenis sentimen yang digunakan pada penelitian akan dibagi menjadi 3 jenis sentimen, yaitu sentimen positif, netral, dan negatif.

1.4 Tujuan Penelitian

Tujuan yang akan dicapai pada penelitian tugas akhir ini yaitu :

1. Mengimplementasikan teknik analisis sentimen dengan menggunakan metode *Lexicon Based* berdasarkan kamus kata VADER terhadap komentar mengenai ChatGPT di situs Reddit.
2. Dapat mengetahui jenis sentimen dari tiap data komentar menggunakan metode *Lexicon Based* berdasarkan kamus kata VADER.

1.5 Manfaat Penelitian

Manfaat dari penelitian tugas akhir ini yaitu dapat memahami serta menerapkan metode *Lexicon Based* dalam menganalisis sentimen sehingga mengetahui sentimen pada setiap data komentar para pengguna Reddit terhadap kehadiran ChatGPT serta kata apa saja yang sering dipakai pengguna Reddit dalam berkomentar dan pihak OpenAI dapat mengetahui kelebihan serta kekurangan yang ada pada ChatGPT.

1.6 Sistematika Penulisan

Sistematika penulisan pada penelitian tugas akhir ini terdiri dari lima bab dengan setiap bab memiliki pembahasan sebagai berikut :

BAB I PENDAHULUAN

Bab pertama pada penelitian tugas akhir ini membahas mengenai latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan

BAB II TINJAUAN PUSTAKA

Bab kedua pada penelitian tugas akhir ini membahas teori-teori yang berhubungan dengan penelitian tugas akhir yaitu komentar, *text mining,* analisis sentimen, *Natural Language Processing*, ChatGPT, Reddit, *text pre-processing*, metode *Lexicon Based*, VADER, *phyton*, Google Colaboratory, visualisasi data dan juga membahas penelitian terdahulu yang berkaitan dengan penelitian tugas akhir ini.

BAB III METODOLOGI PENELITIAN

Bab ketiga pada penelitian tugas akhir ini membahas mengenai tahapan-tahapan penelitian seperti *datasets, data reduction, text pre-processing, Lexicon Based* VADER, hasil klasifikasi sentimen, dan visualisasi hasil dari analisis sentimen.

BAB IV HASIL & PEMBAHASAN

Bab keempat pada penelitian tugas akhir ini membahas tentang hasil penelitian yang telah dilakukan dan pembahasan hasil yang telah dicapai.

BAB V PENUTUP

Bab kelima pada penulisan penelitian tugas akhir ini yaitu kesimpulan pada penelitian tugas akhir ini serta saran kepada penelitian selanjutnya.

BAB II
TINJAUAN PUSTAKA

2.1 Komentar

Menurut KBBI (Kamus Besar Bahasa Indonesia), komentar merujuk pada pendapat atau tanggapan terhadap berita, pidato, dan hal-hal sejenisnya. Diharapkan bahwa komentar-komentar tersebut mengandung pandangan yang cerdas, informatif, dan relevan dengan artikel yang dibahas. Namun realitanya tidak selalu demikian. Konten yang muncul dalam kolom komentar seringkali mengandung kata kasar, termasuk intimidasi, kata-kata kotor, dan ujaran kebencian (Rendragraha dkk., 2021).

Di dalam kolom komentar, setiap individu memiliki kebebasan untuk menuliskan komentarnya dengan beragam gaya tulisan, bahasa, dan struktur kalimat. Komentar-komentar yang ada di kolom tersebut tidak selalu positif, terkadang ada yang mengandung humor, kontroversial, dan bahkan ada yang kasar serta menghina pemilik akun (Putri dkk., 2021).

2.2 Text Mining

Penambangan teks atau *text mining* ialah teknik eksplorasi serta analisis data berbentuk teks yang tidak tertata dengan menggunakan bantuan *software* untuk menidentifikasi topik, konsep, kata kunci, pola, serta atribut yang berkaitan dalam jumlah data berbentuk teks yang besar (Fathonah & Herliana, 2021). *Text mining* dapat didefinisikan juga sebagai proses komputer dalam menemukan informasi baru yang sebelumnya tidak diketahui, dan secara otomatis mengekstrak informasi dari berbagai sumber yang berbeda (Fitriani dkk., 2021).

Dalam mengolah teks yang diambil dari media sosial yang menggunakan bahasa tidak formal atau baku, penting untuk melakukan *text mining*. Proses *text mining* umumnya melibatkan pengelompokan teks, ekstraksi konsep/entitas, kategorisasi teks, produksi taksonomi granular, penarikan kesimpulan dari dokumen, analisis sentimen, dan pemodelan hubungan antar entitas (Utama dkk., 2019).

2.3 Analisis Sentimen

Analisis sentimen ialah suatu sub-bidang dalam NLP yang bertujuan untuk membentuk sistem yang dapat mengidentifikasi dan menghasilkan opini dari teks. Saat ini, data dalam bentuk teks sangat melimpah di Internet, baik dengan format media sosial, blog, forum, maupun situs yang memuat ulasan (Fathonah & Herliana, 2021). Analisis sentimen diidentifikasi sebagai penggalian opini, yaitu bidang yang luas dari NLP, linguistik komputasi, dan *text mining* yang memiliki tujuan untuk menganalisis pendapat, sentimen, evaluasi, sikap, penilaian, dan emosi seseorang tentang suatu topik, produk, layanan, organisasi, individu, atau kegiatan tertentu (Asri dkk., 2022).

Pengertian lain dari analisis sentimen adalah teknik yang digunakan untuk menggali informasi berbentuk opini (sentimen) individu mengenai suatu peristiwa atau isu. Analisis sentimen dapat digunakan untuk mengetahui pendapat umum mengenai kepuasan terhadap layanan, analisis pesaing berdasarkan data tekstual kebijakan, isu, *cyber bullying*, dan prediksi harga dari saham (Sumitro dkk., 2021). Mengelompokkan teks ke dalam kalimat atau dokumen kemudian ditentukan

apakah pendapat yang dinyatakan dalam kalimat atau dokumen tersebut positif atau negatif ialah tugas dasar dari analisis sentimen (Aripiyanto dkk., 2022).

2.4 Natural Language Processing

NLP atau *Natural Language Processing* ialah suatu sistem dan algoritma yang memungkinkan komputer mampu mengerti serta mengerjakan tugas yang berkaitan dengan bahasa manusia. NLP dapat menjalankan analisis bahasa manusia yang berbentuk tulisan ataupun lisan untuk mendapatan informasi yang bermanfaat (Aripiyanto dkk., 2022). NLP ialah sebuah disiplin ilmu komputer yang merupakan bagian dari AI atau kecerdasan buatan, yang berkaitan dengan bahasa dan interaksi antara komputer dengan bahasa alami manusia, seperti bahasa Indonesia atau bahasa Inggris. Fokus utama dari penelitian NLP adalah untuk mengembangkan mesin yang dapat memahami dan mengerti makna dari bahasa manusia serta memberikan respons yang tepat sesuai dengan konteksnya (Yunefri dkk., 2021).

NLP dasar melibatkan beberapa tugas, termasuk deteksi bahasa, tokenisasi dan parsing, part-of-speech tagging, lemmatization/stemming, dan identifikasi hubungan semantik. Pada dasarnya, NLP berfungsi untuk memecah bahasa menjadi unit-unit yang lebih kecil, memahami hubungan di antara unit-unit tersebut, dan menganalisis bagaimana unit-unit tersebut saling berinteraksi untuk membentuk makna (Rosyadi dkk., 2020).

2.5 ChatGPT

Teknologi *chatbot* adalah salah satu bentuk implementasi dari *Natural Language Processing,* yang merupakan cabang ilmu AI (*Artificial Intelligence*)

yang menekuni interaksi manusia dan komputer menggunakan bahasa alami. Komputer memakai bahasa mesin yang setiap orang belum tentu mengerti, namun NLP membolehkan komputer untuk mendengarkan ucapan, membaca teks, menginterpretasikannya, menganalisis sentimen, dan mengidentifikasi bagian yang penting (Rosyadi dkk., 2020).

ChatGPT ialah aplikasi *chatbot* yang dirilis pada November 2022 oleh sebuah laboratorium riset kecerdasan buatan asal Amerika Serikat yang bernama OpenAI. Mesin ini menerapkan teknologi *Natural Language Processing* (NLP) atau pemrosesan bahasa alami yang dapat menjawab pertanyaan seseorang dalam bentuk teks (disebut *prompt*) yang diketikkan dalam aplikasi ChatGPT (Setiawan & Luthfiyani, 2023).

Bagi pengguna, ChatGPT memiliki kemampuan yang menguntungkan, seperti memnghasilkan teks yang serupa dengan manusia, mempercepat proses penulisan, dan dalam pemecahan masalah tertentu memberikan soluis yang cepat serta akurat. Walaupun memiliki banyak keuntungan, ChatGPT juga memiliki resiko yang besar seperti kemampuan ChatGPT dalam menghasilkan teks yang mungkin mengandung bias atau ketidakakuratan informasi (Misnawati, 2023).

2.6 Reddit

Reddit terkenal sebagai wadah untuk berdiskusi berbagai topik dengan bentuk sebuah konten. Reddit didirikan pada tahun 2005 di Amerika Serikat oleh Steve Huffman, Aaron Swartz, dan Alexiz Ohanian. Hingga saat ini, Reddit telah memiliki pengguna sebanyak 300 juta lebih di seluruh dunia (Rachmawaty, 2021).

Di Reddit, isi konten diatur dalam berbagai komunitas yang disebut *subreddit*. Setiap *subreddit* berisi berbagai postingan yang memiliki komentar terkait. Dalam postingan dan komentar tersebut, pengguana dapat menautkan ke konten di tempat lain di web, termasuk koten dari subreddit lain di Reddit. Salah satu jeni tautan *subreddit*, dimana pengguna dapat langsung menautkan ke *subreddit* tertentu, bukan kiriman, komentar, atau situs web eksternal (Krohn & Weninger, 2022).

Komentar Reddit diatur ke dalam *threads* (utas) dengan judul dan beberapa *threads* tersebut menjadi lebih populer dan menerima lebih banyak suara positif dan jumlah komentar yang lebih tinggi dari peserta forum. Dengan demikian, jumlah komentar dalam sebuah *threads* dapat digunakan sebagai penentu popularitas diskusi ini di kalangan *Redditor*, sedangkan skor komentar dapat menunjukkan popularitas komentar tertentu (Long dkk., 2023).

2.7 Text Preprocessing

Text preprocessing merupakan proses membersihkan data yang tidak lengkap atau tidak sempurna, sehingga data yang dimanfaatkan sudah terbebas dari beragam jenis emoji dan atribut yang tidak relevan (Habibah dkk., 2023). *Preprocessing* ialah metode dalam *data mining* yang menyertakan transformasi data tidak terstruktur menjadi terstruktur dan dapat dipahami. Data yang tidak terstruktur acapkali tidak lengkap, mungkin mengandung kesalahan, dan tidak konsisten. Teknik *pre-processing* terbukti efektif dalam menangani masalah tersebut. Teknik *pre-processing* memiliki langkah-langkah termasuk *tokenizing, case folding*, *filtering*, dan *stemming* (Cindo dkk., 2019). *Text preprocessing* digunakan untuk mempersiapkan data sebelum dilakukan analisis sentimen yang bertujuan untuk

mengektraksi informasi terkait sentimen penulis dalam data tersebut, apakah sentimen tersebut negatif atau positif (Yunefri dkk., 2021).

2.7.1 Cleaning

Data *cleaning* merupakan proses evaluasi kualitas data dengan melakukan modifikasi,perubahan, atau penghapusan data yang dianggap tidak diperlukan, tidak lengkap, tidak akurat, atau memiliki format atau struktur yang tidak sesuai dalam basis data, dengan tujuan menghasilkan data yang berkualitas tinggi. Proses ini juga dikenal sebagai data *cleaning* atau data *scrubbing* (Darwis dkk., 2021). Tahapan *cleaning,* digunakan untuk membersihkan kata-kata dengan tujuan mengurangi *noise* seperti hastag, html, tautan, nama pengguna, dan konten yang tidak relevan. Selain itu, proses ini juga melibatkan penghilangan tanda baca seperti titik (.), koma (,), dan karakter aksen lainnya (Vindua & Zailani, 2023).

2.7.2 Text Normalization

Normalisasi teks (*Text Normalization*) merupakan proses mengubah tulisan yang disingkat atau dimodifikasi menjadi bentuk yang sesuai dengan cara berbicara. Tulisan yang disingkat atau dimodifikasi adalah tulisan yang maknanya tetap sama, tetapi kata-katanya tidak ditulis secara lengkap atau ditambahi huruf-huruf tertentu (Alrajak dkk., 2020). Pengertian lain *text normalization* ialah proses memperbaiki teks dalam dokumen yang mengandung kata-kata tidak baku atau singkatan bahasa sehari-hari menjadi kata-kata yang bermakna. Misalnya, “yg” diganti dengan “yang”, “tdk” diganti dengan “tidak” dan sebagainya (Ashari dkk., 2020).

2.7.3 Case folding

Pada setiap data, seringkali terdapat variasi penulisan huruf yang tidak konsisten, seperti adanya huruf kecil dan juga huruf besar atau kapital. Maka dari itu, diperlukan langkah-langkah untuk mengubah semua huruf menjadi huruf kecil agar konsisten dalam bentuk standar. Proses ini disebut *case folding*, dimana *case folding* digunakan untuk menkonversi huruf-huruf alfabet menjadi huruf kecil atau *lower case* (Hidayati dkk., 2023).

2.7.4 Tokenizing

Tokenizing merupakan tahapan untuk membagi deretan atau rangkaian kata pada sebuah kaliamat, paragraf, atau halaman ke dalam token atau potongan kata tunggal yang berdiri sendiri dan dapat dikenali secara individual (Darwis dkk., 2021). Pada prinsipnya, *tokenizing* ialah tahapan memisahkan kata per kata yang membentuk sebuah kalimat dalam dokumen. Biasanya, setiap kata diidentifikasi serta dipisahkan dengan spasi, maka dari itu tahapan *tokenizing* menggunakan karakter berupa spasi dalam dokumen untuk menguraikan kata-kata (Fitriyah dkk., 2020).

2.7.5 Stopwords Removal

Stopwords merujuk pada kata-kata umum yang sering muncul dan dianggap tidak memiliki maknya yang signifika. Menggunakan *stopwords removal* terbukti meningkatkan akurasi sistem klasifikasi sentimen jika dibandingkan dnegan tidak menggunakan *stopwords removal* (Fitriyah dkk., 2020). Adapun daftar lengkap dari *stopwords* yang berdasarkan gambar berikut ini (Birks dkk., 2020).

Stop words								
A	**Between**	**Few**	**How**	**My**	**Re**	**Theirs**	**Wasnt**	**Your**
About	Both	For	I	Myself	S	Them	We	Youre
Above	But	From	If	Needn	Same	Themselves	Were	Yours
After	By	Ft	In	Neednt	Shan	Then	Weren	Yourself
Again	Can	Further	Into	No	Shant	There	Werent	Yourselves
Against	Comp	Had	Is	Nor	She	These	What	Youve
Ain	Complainant	Hadn	Isn	Not	Shes	They	When	
All	Couldn	Hadnt	Isnt	Now	Should	This	Where	
Am	Couldnt	Has	It	O	Shouldn	Those	Which	
An	D	Hasn	Its	Of	Shouldnt	Through	While	
And	Did	Hasnt	Its	Off	Shouldve	Time	Who	
Any	Didn	Have	Itself	On	So	To	Whom	
Are	Didnt	Haven	Just	Once	Some	Too	Why	
Aren	Do	Havent	Ll	Only	Stated	Under	Will	
Arent	Does	Having	M	Or	Such	Unknown	With	
As	Doesn	He	Ma	Other	Suspect	Until	Won	
At	Doesnt	Her	Me	Our	Suspects	Up	Wont	
Be	Doing	Here	Mightn	Ours	T	Ve	Wouldn	
Because	Don	Hers	Mightnt	Ourselves	Than	Very	Wouldnt	
Been	Dont	Herself	More	Out	That	Victim	Y	
Before	Down	Him	Most	Over	Thatll	Victims	You	
Being	During	Himself	Mustn	Own	The	Was	Youd	
Below	Each	His	Mustnt	Property	Their	Wasn	Youll	

Gambar 2.1 Daftar Stopwords Bahasa Inggris

2.7.6 Stemming

Stemming yaitu proses pada sistem IR (*Information Retrieval*) yang mengubah kata-kata hasil *filtering* menjadi kata akar dengan menggunakan aturan-aturan tertentu. Proses *stemming* pada teks berbahasa Indonesia berbeda dengan *stemming* pada teks berbahasa Inggris. Pada teks berbahasa Inggris, prosesnya hanya melibatkan penghilangan surfiks, sedangkan pada teks berbahasa Indonesia, selain surfiks, juga dilakukan penghilangan prefix dan konfiks (Fitriani dkk., 2021). *Stemming* juga dapat dikatakan tahapan untuk menghapus imbuhan seperti awalan, sisipan, akhiran, dan kombinasi awalan dan akhiran dengan tujuan mencari kata dasar dari kata tersebut. Tujuan dari tahap *stemming* adalah untuk mengurangi jumlah indeks yang berbeda dari suatu data (Hidayati dkk., 2023).

2.8 Lexicon Based

Metode *Lexicon Based* ialah pendekatan yang praktis, sederhana serta efektif dalam melakukan analisis sentimen. Facebook, Twitter dapat dimanfaatkan sebagai sumber data serta platform media sosial lainnya yang mencerminkan pendapat tentang suatu produk atau layanan. Keuntungan menggunakan metode *Lexicon Based* adalah tidak memerlukan data yang sudah diberi label atau proses pembelajaran khusus. Kata-kata dalam metode *Lexicon Based* akan dinilai berdasarkan nilai polaritas yang bermaksud untuk memahami pendapat atau tenggapan masyarakat (Hernikawati , 2021).

Metode *Lexicon Based* ialah metode yang bersifat *unsupervised learning* dimana prosesnya tidak memerlukan data latih dan metode ini membutukan kamus kata yang berisi kata-kata positif maupun negatif (Aripiyanto dkk., 2022). Pendekatan *lexicon* adalah suatu metode yang melibatkan penggunaan kamus sentimen yang memuat kata-kata opini yang kemudian dibandingkan dengan data yang akan mementukan nilai pada kata. Dalam kamus *lexicon,* setiap kata dinilai berdasarkan polaritas kata. Pertama-tama dalam pendekatan *lexicon*, langkah awalnya ialah memutuskan kata-kata untuk dianalisis dari kumpulan teks. Proses pemilihan setiap kata tersebut dapat dilakukan dengan menggunakan teknik *part-of-speech* untuk mencari kata-kata berdasarkan jenis tertentu seperti *noun, adjective, dan adverb* (Cindo dkk., 2019) .

Dalam kamus *lexicon*, kata-kata yang teridentifikasi akan dihitung skornya berdasarkan dengan jumlah kata pada setiap teks atau kalimat (Ismail & Hakim, 2023).

$$S_{positive} \sum_{i \in t}^{n} positive\ score_i \quad (2.1)$$

$$S_{negative} \sum_{i \in t}^{n} negative\ score_i \quad (2.2)$$

Nilai ($S_{positive}$) merujuk pada bobot kalimat yang diperoleh dengan menjumlahkan skor polaritas n kata opini positif, sedangkan nilai ($S_{negative}$) mengacu pada bobot nilai yang diperoleh dengan menjumlahkan skor dari polaritas n kata bernilai opini negatif. Berdasarkan persamaan nilai sentimen pada satu kalimat, maka dapat ditemukan persamaan 3 yang memungkinkan kita memutuskan jenis sentimen dengan menbandingkan jumlah nilai dari positif, negatif, serta netral (Ismail & Hakim, 2023).

$$Sentence_{sentiment} \begin{Bmatrix} positive\ if\ S_{positive} > S_{negative} \\ neutral\ if\ S_{positive} = S_{negative} \\ negative\ if\ S_{positive} < S_{negative} \end{Bmatrix} \quad (2.3)$$

Apabila terdapat lebih banyak kata positif daripada kata negatif dalam suatu teks, maka teks tersebut akan dikategorikan sebagai sentimen positif. Sebaliknya, apabila terdapat lebih banyak kata bernilai negatif daripada kata bernilai positif, maka teks tersebut akan dikelompokkan kedalam sentimen negatif. Dan jika jumlah kata bernilai positif serta negatif sama banyak dalam suatu teks, maka teks tersebut akan dikategorikan sebagai sentimen netral (Ismail & Hakim, 2023).

2.9 VADER

C.J Hutto dan Eric Gilbert di tahun 2014, memperkenalkan VADER (*Valence Aware Dictionary and Sentiment Reasoner*) berdasarkan *human-centric*

yang menyatukan analisis kualitatif dengan validasi empiris serta didasarkan pada kebijaksanaan serta penilaian oleh manusia. VADER ialah metode yang dimanfaatkan untuk model dalam menganalisis sentimen dan mengukur variasi data berdasarkan intensitas kekuatan emosi yang terdapat di dalamnya dengan menggunakan kamus data *lexicon* yang tersedia (Abimanyu, 2022).

Hutto dan Gilbert mengemukakan bahwa setiap fitur leksikal memiliki skor rata-rata nol dan standar deviasi kurang dari 2,5. Terdapat lebih dari 7500 fitur leksikal dengan skor valensi teruji yang menunjukkan polaritas sensorik dan intensitas perasaan pada skala -4 hingga 4 yang mencakup polaritas positif dan negatif. Contohnya, 'oke' memiliki skor 0,9, 'untuk' memiliki skor 3,1, 'jelek' memiliki skor -2,5, dan 'sakit' memiliki skor -1,5. Vader akan mencetak setiap teks, dan skor positif, negatif, atau netral akan dihasilkan. Selanjutnya, semua titik akan dijumlahkan untuk membentuk *compound*, yaitu matriks yang menghitung semua skor yang dinormalisasi dari -1 hingga +1. Jika skor komposit lebih besar dari 0,05, maka dianggap positif. Sebaliknya, jika skor komposit kurang dari -0,05, dianggap negatif, dan jika skor komposit antara -0,05 dan 0,05, dianggap netral (Asri dkk., 2022). Adapun jenis sentimen berdasarkan total dari compound score atau skor komposit yang dapat dilihat pada Tabel 2.1 dibawah ini.

Tabel 2.1 Jenis Sentimen Berdasarkan Total Skor Komposit

Compound Score	Jenis Sentimen
S > 0,05	Positif
-0,05 < S < 0,05	Netral
S < -0,05	Negatif

Hutto menggunakan rumus normalisasi *score* seperti dibawah ini (Abimanyu, 2022).

$$\frac{x}{\sqrt{x^2 + \alpha}} \tag{2.4}$$

Dimana hal ini, x mewakili jumlah nilai sentimen dari setiap kata yang membentuk kalimat sementara alpha ialah sebuah konstanta normalisasi yang secara default sitetapkan sebagai 15. Oleh karena itu, metode analisis sentimen VADER paling efektif digunakan untuk dokumen berukuran kecil atau pendek bukan dari dokumen yang lebih besar, seperti Reddit, *tweet*, dan kalimat, (Abimanyu, 2022).

2.10 Silhouette Index

Salah satu metode validasi yang menggunakan kriteria internal adalah *Silhouette Index*. Metode ini akan mengukur seberapa baik setiap objek ditempatkan di klaster masing-masing dengan melihat jarak rata-rata antara objek dengan anggota klaster yang sama dan jarak rata-rata antara objek dengan anggota klaster lain (Nahdliyah dkk., 2019).

Untuk menghitung nilai SI, langkah pertama adalah menghitung rata-rata jarak antara objek yang disebut i dengan semua objek lain di klaster yang sama menggunakan persamaan berikut :

$$a(i) = \frac{1}{|A| - 1} \sum j \ \in_{A, j \neq i} \ d(i, j) \tag{2.5}$$

Berikutnya, kita harus menghitung rata-rata jarak antara objek i dengan semua objek lain di klaster yang berbeda, kemudian pilih nilai terbesar dan terkecil. Setelah mendapatkan nilai rata-rata jarak, langkah berikutnya adalah menghitung nilai SI dengan rumus berikut :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2.6)$$

Langkah terakhir ialah membandingkan nilai *Silhouette Index* yang didapat. Nilai *Silhouette Index* memiliki rentang antara -1 hingga 1. Nilai yang lebih mendekati dengan angka 1 menunjukkan data semakin baik (Reinaldi dkk., 2021).

2.11 Python

Guido Van Rossum menciptakan Python sebagai bahasa pemrograman tingkat tinggi pada tahub 1991. Saat ini. Python merupakan pemrograman yang multifungsi dan dapat digunakan untuk *machine learning* dan *deep learning.* Python dipilh sebagai bahasa pemrograman penelititan karena memiliki sintaksis yang mudah ditulis. Selain itu, Python juga memiliki *library* yang lengkap dan didukung oleh komunitas yang kuat karena bersifat *opne source.* Untuk menulis *source code* Python, tersedia berbagai pilihan IDE seperti VS Code, Sublime Text, PyCharm, atau dapat juga menggunakan IDE *online* seperti Jupyter notebook dan Google Colab (Alfarizi dkk., 2023).

2.12 Google Colaboratory

Google Colab ialah sebuah IDE (*Integrated Development Environment*) untuk pemrograman Python di mana pemrosesan dilakukan oleh server Google

yang memiliki perangkat keras berkualitas tinggi. Dari segi *software*, Google Colab sudah menyediakan mayoritas pustaka yang diperlukan pengguna. Dengan jaminan stabilitas servernya, hampir semua pemrosesan dapat dilakukan dengan lancar menggunakan Google Colab, selama koneksi internet tetap lancer (Guntara, 2023a).

"Colaboratory", atau disebut juga "Colab" merupakan produk yang dikembangkan oleh Google Research. Colab mengizinkan pengguna untuk meng*coding* serta menjalankan kode berbasis Python secara *online* melewati peramban web, dan sangat cocok digunakan dalam keperluan analisis data, *machine learning*, dan pendidikan. Secara lebih teknis, Colab ialah layanan notebook Jupyter yang di-*hosting,* yang dapat digunakan tanpa perlu instalasi tambahan, dan menyediakan akses gratis ke sumber daya komputasi termasuk GPU. Meskipun gratis sumber daya yang tersedia dalam Colab tidak dijamin dan terbatas, serta penggunaan terkadang memiliki batasan yang bervariasi (Soen dkk., 2022).

2.13 Visualisasi Data

Visualisasi data ialah teknik untuk menggambarkan informasi yang terkandung dalam data dengan menggunakan grafik atau gambar. Tujuan utama dari visualisasi data adalah untuk memudahkan pemahaman data, mengidentifikasi pola dan tren yang ada, serta mendukung pengambilan keputusan berdasarkan data (Guntara, 2023b). Dalam memvisualisasikan data, ada beberapa tipe visualisasi seperti dengan menggunakan tabel, grafik, diagram, ataupun dengan *word cloud.*

Word Cloud ialah representasi visual yang menggambarkan frekuensi kata-kata dalam sebuah teks. Ukuran huruf pada *Word Cloud* menunjukkan seberapa sering kata tersebut muncul, di mana ukuran huruf yang lebih besar menandakan frekuensi

yang lebih tinggi, sedangkan ukuran huruf yang lebih kecil menandakan frekuensi yang lebih rendah (Julianto, 2022).

Gambar 2. 2 Contoh Word Cloud

2.14 Penelitian Terdahulu

Penelitian tugas akhir yang akan disusun oleh peneliti ini merujuk kepada beberapa penelititan terdahulu yang berkaitan dengan metode *Lexicon Based* dan juga VADER. Adapun penelitian terdahulu tersebut yang dapat dilihat pada Tabel 2.2 berikut ini.

Tabel 2.2 Penelitian Terdahulu

No	Penulis	Judul	Hasil
1.	Putri Amira Sumitro, Rasiban, Dadang Iskandar Mulyana dan	Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter	Dengan memakai metode *lexicon based* dan memanfaatkan *library* vader *sentiment* menggunakan 1.941

	Wahyu Saputro (2021)	Menggunakan Metode *Lexicon Based*	data *tweet* menghasilkan sentimen positif 20,25%, agak positif 23,9%, negatif 1,88%, agak negatif 9,6% dan netral 44,36%.
2.	Fitrah Amaliah dan I Kadek Dwi Nuryana (2022)	Perbandingan Akurasi Metode *Lexicon Based* Dan *Naïve Bayes Classifier* Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter	Menggunakan data dari Twitter sebanyak 600 data *tweets* setelah tahap *preprocessing* menjadi 147 data dengan hasil akurasi metode *lexicon based* dengan kamus kata VADER sebesar 67% dengan hasil polatitas tertinggi yaitu positif sebesar 64,63% dan akurasi pada *naïve bayes classifier* sebesar 78% dengan polaritas tertinggi yaitu positif sebesar 53,74%.
3.	Nurul Habibah, Elvia Budianita, Muhammad Fikry dan Iwan Iskandar (2023)	Analisis Sentimen Mengenai Penggunaan *E-Wallet* Pada Google Play Menggunakan *Lexicon Based* dan *K-Nearest Neighbor*	Dengan menggunakan *Lexicon Based* dengan kamus kata VADER dan KNN pada aplikasi Dana menghasilkan akurasi tertinggi pada k=6 sebesar 78% serta sentimen positif 4.921 ulasan, negatif 2.550, ulasan dan netral 1.529. Pada aplikasi Ovo menghasilkan akurasi tertinggi pada k=9 sebesar 75,33% dengan sentimen positif 5.289 ulasan, negatif 1.789 ulasan dan netral 1.992 ulasan.

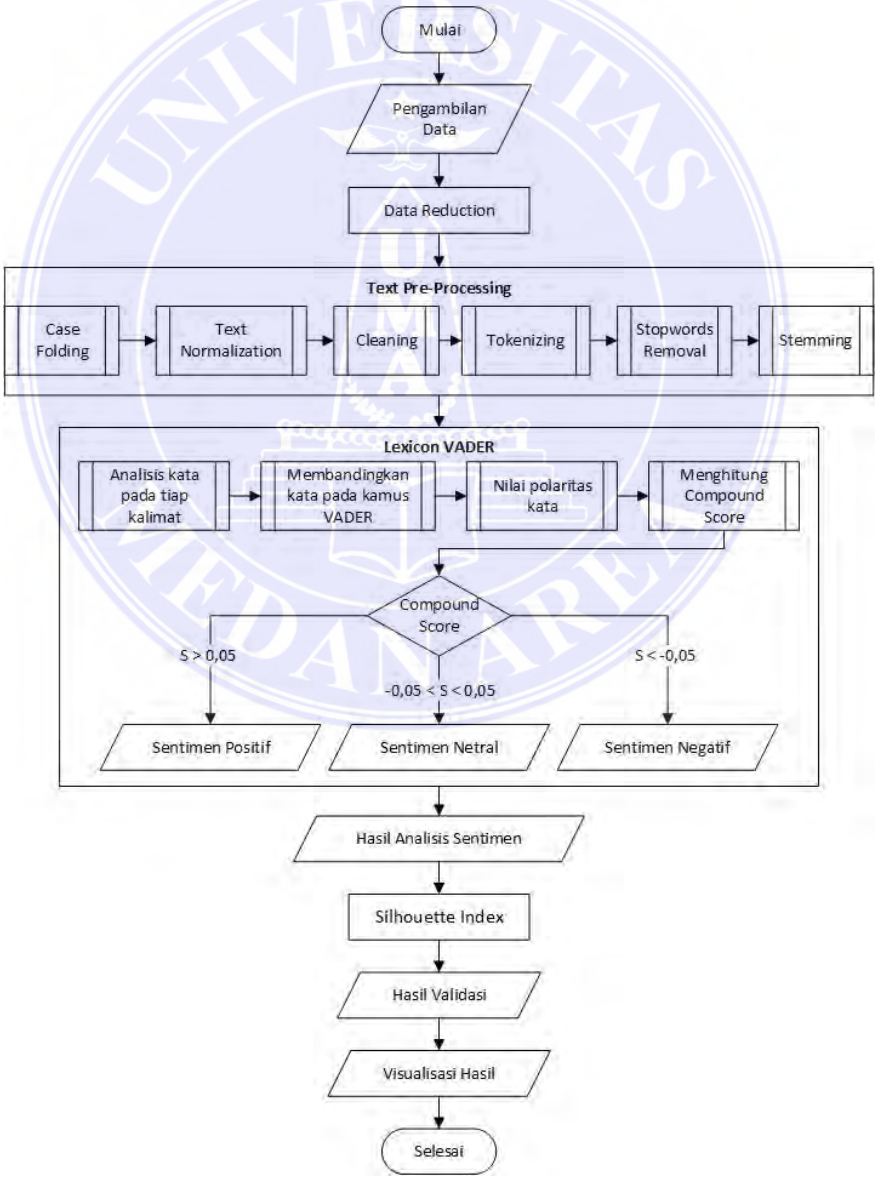
			Pada aplikasi LinkAja menhasilkan akurasi tertinggi pada k=8 sebesar 73,5% dengan sentimen positif 6.037 ulasan, negatif 1.705 ulasan dan netral 1.285 ulasan.

Dari Tabel penelitian terdahulu diatas, terdapat beberapa penelitian yang memakai metode *Lexicon Based* saja untuk menganalisis sentimen, dan terdapat pula yang mengkombinasikan 2 metode seperti metode *Lexicon Based* dengan *K-Nearest Neigbor.* Serta, terdapat juga penelitian yang melakukan perbandingan dua metode yaitu metode *Lexicon Based* dengan *Naïve Bayes Classifier* untuk mendapatkan akurasi dari kedua metode tersebut. Penelitian ini menggunakan metode *Lexicon Based* dengan menggunakan *datasets* yang bersumber dari Reddit yang lebih menekankan pada analisis sentimen terhadap komentar yang berhubungan dengan ChatGPT

BAB III
METODOLOGI PENELITIAN

3.1 Tahapan Penelitian

Metodologi penelitian ialah suatu tahapan dalam melakukan penelitian yang disusun dengan sistematis serta logis yang bertujuan untuk mendapatkan solusi dari permasalahan yang ada. Adapun tahapan-tahapan dalam menyusun penelitian tugas akhir ini yang dapat dilihat dibawah ini.



Gambar 3.1 Flowchart Tahapan Penelitian

3.2 Pengambilan Data

Data yang akan digunakan pada penelitian tugas akhir ini merupakan data sekunder yang berasal dari situs penyedia *dataset* untuk publik yaitu Kaggle.com. *Dataset* yang dipakai memiliki judul "ChatGPT Reddit: Reddit comments about ChatGPT". *Dataset* ini memuat data komentar pengguana reddit tentang ChatGPT dengan data berjumlah 52.416 data dengan atribut *comment_id*, *comment_parent*, *comment_body* dan dengan *class feature subreddit*. Berikut ini adalah *link* sumber dari *dataset* komentar ChatGPT yang digunakan oleh peneliti https://www.kaggle.com/datasets/armitaraz/chatgpt-reddit.

3.3 Data Reduction

Untuk mendapatkan efesiensi dan kemudahan dalam pengolahan data, akan dilakukan *data reduction* pada *dataset*. Pada tahapan ini bertujuan agar tidak adanya bias dalam menganalisis data. Pada *class feature dataset* yaitu kolom *subreddit* memiliki 4 jenis *class* yaitu r/ChatGPT dengan total 29.710 data, r/dataisbeautiful dengan total 1.091 data, r/Futurology dengan total 5.508 data dan r/technology dengan total 16.671 data. Sehingga data yang akan digunakan hanya class r/ChatGPT dengan 29.710 data. Kemudian dilakukan pemilahan fitur dengan hanya menggunakan kolom *comment_body* sebagai kolom yang akan dilakukan *text preprocessing* kemudian dianalisis menggunakan metode *Lexicon Based* VADER.

3.4 Text Preprocessing

1. *Case Folding*

Pada tahap ini, data yang mempunyai kata dengan huruf besar atau kapital akan diubah kedalam huruf kecil semua. Seperti kata "What" diubah menjadi

"what", "PROBLEM" diubah menjadi "problem", "PRO" diubah menjadi "pro". Seperti "ChatGPT is very GOOD Does this have a negative effect" diubah menjadi "chatgpt is very good does this have a negative effect".

2. *Text Normlaization*

Pada tahapan in, data yang berbentuk singkatan akan diubah ke bentuk semulanya yaitu bentuk kata tanpa singkatan. Seperti "what's" diubah menjadi "what is". "ngl" diubah menjadi "not gonna lie", dan juga kata-kata singkatan lainnya.

3. *Cleaning*

Pada tahapan *cleaning* akan dilakukan pembersihan data yang dianggap sebagai *noise* atau gangguan seperti data duplikat, data tidak memiliki isi atau kosong, serta pembersihan data yang berisi tanda baca yang tidak mengandung sentimen. Tanda baca yang akan dibersihkan seperti tanda titik (.), tanda seru (!), tanda tanya (?), tanda koma(,),tanda titik koma (;), tanda kurung, tanda petik dua ("") dan tanda baca lainnya dan juga akan dilakukan penghapusan karakter yang tidak penting seperti, @, #, % dan sebagainya, serta penghapusan url yang tidak berkaitan dengan analisis sentimen pada penelitian ini. Contohnya seperti: "ChatGPT is very GOOD,.! !!! Does this have a negative effect?" diubah menjadi "ChatGPT is very GOOD Does this have a negative effect".

4. *Tokenizing*

Pada tahapan *tokenizing*, setiap kalimat yang terdiri dari beberapa kata akan dipisahkan menjadi kata per kata sehingga keakuratan data akan meningkat. Sepert kalimat "chatgpt is very good does this have a negative effect" dipecah menjadi "[chatgpt, is, very, good, does, this, have, a, negative, effect]".

5. *Stopwords Removal*

Pada tahapan *stopwords removal,* merupakan proses untuk menghilangkan atau menghapuskan kata yang tidak mengandung sentimen atau makna tertentu. Seperti kata “the”, “and”, “hers”, dll.

6. *Stemming*

Pada tahapan *stemming,* tiap kata yang memiliki kata imbuhan diawal maupun diakhir akan diubah kedalam bentuk kata dasar sehingga kata yang akan dianalisis sentimen hanya hata-kata dasar saja seperti “faster” menjadi “fast”, “using” menjadi “use”, “reading” menjadi “read”, dan lainnya.

3.5 Lexicon VADER

Setelah melalui tahapan *pre-processing*, pada tahapan ini setiap kata pada kalimat dalam data akan dicari nilai polaritas sentimen setiap kata berdasarkan dari kamus kata VADER. Setelah mendapatkan skor polaritas, kemudian dilakukan perhitungan *compound score* atau skor komposit berdasarkan persamaan (2.4) sehingga kalimat dapat dikelompokkan ke dalam jenis sentimen.

3.6 Hasil Analisis Sentimen

Setelah melalui perhitungan skor polaritas dan skor komposit maka akan didapatkan hasil jenis sentimens dari tiap data dengan tiga jenis sentimen yaitu sentimen positif, negatif dan netral. Sehingga dapat dikelompokkan berapa banyak komentar ChatGPT pada situs Reddit yang memiliki sentimen positif, negatif, atau netral.

3.7 Silhouette Index

Pada tahapan ini, data yang telah berhasil dikelompokkan menggunakan metode *Lexicon VADER,* selanjutnya akan dilakukan proses validasi hasil menggunakan metode validasi *Silhouette Index.* Dengan membagi data kedalam 3 klaster berdasarkan 3 jenis sentimen yang telah ditentukan. Nilai yang diambil untuk validasi hasil yaitu data nilai dari *compound score* tiap kalimat pada data komentar.

3.8 Hasil Validasi

Pada tahapan ini merupakan hasil keluaran dari tahapan validasi hasil menggunakan metode *Silhouette Index* dan akan mendapatkan nilai validasi SI dengan ketentuan nilai yang diperoleh yaitu apakah mendekati nilai angka -1 atau angka 1.

3.9 Visualisasi Data

Hasil dari analisis sentimen akan di visualisasikan ke dalam bentuk tabel, grafik batang, grafik lingkaran, serta dalam bentuk *wordcloud*. Visualisai hasil penelitian ini bertujuan untuk informasi dari hasil sentimen dapat dipahami oleh pembaca.

3.10 Contoh Penerapan

Adapun contoh penerapan dengan menggunakan dataset yang memiliki data berisi komentar tentang produk pakaian, sebagai berikut :

Id	Comment_Body
1	I love, love, love this jumpsuit. it's fun, flirty, and fabulous!
2	3 tags sewn in, 2 small (about 1" long) and 1 huge (about 2" x 3"). very itchy so i cut them out.
3	Gorgeous sweater - such a simple design, but really stunning on.
4	So soft!! such beautiful fabric!!
5	The fit is not as shown on the website. will be returning.
6	I am 5"7" and weigh 123 lbs. i tried on both these dresses in medium, and they both fit.
7	Why do designers keep making crop tops??!! i can't imagine this would be flattering on anyone
8	I... love... this dress. the fit is incredibly flattering
9	Love the pattern. the pants are really cute!
10	This is so thin and poor quality. especially for the price.

Pada *datasets* diatas, kemudian akan dilakukan tahapan *preprocessing,* sebagai berikut :

a. *Cleaning*

Sebelum	Sesudah
I love, love, love this jumpsuit. it's fun, flirty, and fabulous!	I love love love this jumpsuit its fun flirty and fabulous
3 tags sewn in, 2 small (about 1" long) and 1 huge (about 2" x 3"). very itchy so i cut them out.	tags sewn in small about long and huge about very itchy so i cut them out
Gorgeous sweater - such a simple design, but really stunning on.	Gorgeous sweater such a simple design but really stunning on
So soft!! such beautiful fabric!!	So soft such beautiful fabric
The fit is not as shown on the website. will be returning.	The fit is not as shown on the website will be returning

I am 5"7" and weigh 123 lbs. i tried on both these dresses in medium, and they both fit.	I am and weigh lbs i tried on both these dresses in medium and they both fit
Why do designers keep making crop tops??!! i can't imagine this would be flattering on anyone	Why do designers keep making crop tops??!! i can't imagine this would be flattering on anyone
I... love... this dress. the fit is incredibly flattering	I love this dress the fit is incredibly flattering
Love the pattern. the pants are really cute!	Love the pattern. the pants are really cute
This is so thin and poor quality. especially for the price.	This is so thin and poor quality especially for the price

b. *Case Folding*

Sebelum	Sesudah
I love love love this jumpsuit its fun flirty and fabulous	i love love love this jumpsuit its fun flirty and fabulous
tags sewn in small about long and huge about very itchy so i cut them out	tags sewn in small about long and huge about very itchy so i cut them out
Gorgeous sweater such a simple design but really stunning on	gorgeous sweater such a simple design but really stunning on
So soft such beautiful fabric	so soft such beautiful fabric
The fit is not as shown on the website will be returning	the fit is not as shown on the website will be returning
I am and weigh lbs i tried on both these dresses in medium and they both fit	i am and weigh lbs i tried on both these dresses in medium and they both fit
Why do designers keep making crop tops??!! i can't imagine this would be flattering on anyone	why do designers keep making crop tops i can't imagine this would be flattering on anyone

I love this dress the fit is incredibly flattering	i love this dress the fit is incredibly flattering
Love the pattern. the pants are really cute	love the pattern the pants are really cute
This is so thin and poor quality especially for the price	this is so thin and poor quality especially for the price

c. *Tokenizing*

Sebelum	Sesudah
I love love love this jumpsuit its fun flirty and fabulous	[i] [love] [love] [love] [this] [jumpsuit] [its] [fun] [flirty] [and] [fabulous]
tags sewn in small about long and huge about very itchy so i cut them out	[tags] [sewn] [in] [small] [about] [long] [and] [huge] [about] [very] [itchy] [so] [i] [cut] [them] [out]
Gorgeous sweater such a simple design but really stunning on	[gorgeous] [sweater] [such] [a] [simple] [design] [but] [really] [stunning] [on]
So soft such beautiful fabric	[so] [soft] [such] [beautiful] [fabric]
The fit is not as shown on the website will be returning	[the] [fit] [is] [not] [as] [shown] [on] [the] [website] [will] [be] [returning]
I am and weigh lbs i tried on both these dresses in medium and they both fit	[i] [am] [and] [weigh] [lbs] [i] [tried] [on] [both] [these] [dresses] [in] [medium] [and] [they] [both] [fit]
Why do designers keep making crop tops i cant imagine this would be flattering on anyone	[why] [do] [designers] [keep] [making] [crop] [tops] [i] [cant] [imagine] [this] [would] [be] [flattering] [on] [anyone]
I love this dress the fit is incredibly flattering	[i] [love] [this] [dress] [the] [fit] [is] [incredibly] [flattering]

Love the pattern the pants are really cute	[love] [the] [pattern] [the] [pants] [are] [really] [cute]
This is so thin and poor quality especially for the price	[this] [is] [so] [thin] [and] [poor] [quality] [especially] [for] [the] [price]

d. *Stopword Removal*

Sebelum	Sesudah
[i] [love] [love] [love] [this] [jumpsuit] [its] [fun] [flirty] [and] [fabulous]	[love] [love] [love] [jumpsuit] [fun] [flirty] [fabulous]
[tags] [sewn] [in] [small] [about] [long] [and] [huge] [about] [very] [itchy] [so] [i] [cut] [them] [out]	[tags] [see] [small] [long] [huge] [very] [itchy] [cut]
[gorgeous] [sweater] [such] [a] [simple] [design] [but] [really] [stunning] [on]	[gorgeous] [sweater] [such] [simple] [design] [really] [stunning]
[so] [soft] [such] [beautiful] [fabric]	[soft] [beautiful] [fabric]
[the] [fit] [is] [not] [as] [shown] [on] [the] [website] [will] [be] [returning]	[fit] [shown] [website] [returning]
[i] [am] [and] [weigh] [lbs] [i] [tried] [on] [both] [these] [dresses] [in] [medium] [and] [they] [both] [fit]	[weigh] [lbs] [tried] [dresses] [medium] [fit]
[why] [do] [designers] [keep] [making] [crop] [tops] [i] [cant] [imagine] [this] [flattering] [on] [anyone]	[designers] [keep] [making] [crop] [tops] [imagine] [flattering] [anyone]
[i] [love] [this] [dress] [the] [fit] [is] [incredibly] [flattering]	[love] [dress]] [fit] [incredibly] [flattering]
[love] [the] [pattern] [the] [pants] [are] [really] [cute]	[love] [pattern] [pants] [really] [cute]

[this] [is] [so] [thin] [and] [poor] [quality] [especially] [for] [the] [price]	[thin] [poor] [quality] [especially] [price]

e. *Stemming*

Sebelum	Sesudah
[love] [love] [love] [jumpsuit] [fun] [flirty] [fabulous]	[love] [love] [love] [jumpsuit] [fun] [flirty] [fabulous]
[tags] [see] [small] [long] [huge] [very] [itchy] [cut]	[tag] [see] [small] [long] [huge] [very] [itch] [cut]
[gorgeous] [sweater] [such] [simple] [design] [really] [stunning]	[gorgeous] [sweater] [such] [simple] [design] [really] [stunning]
[soft] [beautiful] [fabric]	[soft] [beautiful] [fabric]
[fit] [shown] [website] [returning]	[fit] [show] [website] [return]
[weigh] [lbs] [tried] [dresses] [medium] [fit]	[weigh] [lbs] [try] [dress] [medium] [fit]
[designers] [keep] [making] [crop] [tops] [imagine] [flattering] [anyone]	[design] [keep] [make] [crop] [top] [imagine] [flatter] [anyone]
[love] [dress] [fit] [incredibly] [flattering]	[love] [dress] [fit] [incredible] [flattering]
[love] [pattern] [pants] [really] [cute]	[love] [pattern] [pants] [really] [cute]
[thin] [poor] [quality] [especially] [price]	[thin] [poor] [quality] [special] [price]

Kemudian dilanjutkan ke tahapan menghitung nilai polaritas serta skor komposit pada tiap data komentar berdasarkan dari kamus kata *Lexicon VADER* :

Komentar	Nilai Polaritas	Normalisasi Skor Komposit	Jenis Sentimen

[love] [love] [love] [jumpsuit] [fun] [flirty] [fabulous]	[3,2] [3,2] [3,2] [0] [,3] [0,6] [2,4]	$\frac{14,9}{\sqrt{14,9^2+15}} = \frac{14,9}{29,9}$ = 0,49	Positif
[tag] [sew] [small] [long] [huge] [very] [itchy] [cut]	[0] [0] [0] [0] [1,3] [0] [-1,1] [-1,1]	$\frac{-0,9}{\sqrt{-0,9^2+15}} = \frac{-0,9}{3,76}$ = -0,23	Negatif
[gorgeous] [sweater] [such] [simple] [design] [really] [stunning]	[3,0] [0] [0] [0] [0] [0] [1,6]	$\frac{4,6}{\sqrt{4,6^2+15}} = \frac{4,6}{6,01}$ = 0,76	Positif
[soft] [beautiful] [fabric]	[0] [2,9] [0]	$\frac{2,9}{\sqrt{2,9^2+15}} = \frac{2,9}{4,83}$ = 0,6	Positif
[fit] [show] [website] [return]	[0] [0] [0] [0]	= 0	Netral
[weigh] [lbs] [try] [dress] [medium] [fit]	[0] [0] [0] [0] [0] [0]	= 0	Netral
[design] [keep] [make] [crop] [top] [imagine] [flatter] [anyone]	[0] [0] [0] [0] [0,8] [0] [0,4] [0]	$\frac{1,2}{\sqrt{1,2^2+15}} = \frac{1,2}{4,05}$ = 0,29	Positif
[love] [dress] [fit] [incredible] [flattering]	[3,2] [0] [0] [0] [1,3]	$\frac{4,5}{\sqrt{4,5^2+15}} = \frac{4,5}{5,93}$ = 0,75	Positif

[love] [pattern] [pants] [really] [cute]	[3,2] [0] [0] [0] [2.0]	$\frac{5,2}{\sqrt{5,2^2 + 15}} = \frac{5,2}{6,48}$ = 0,8	Positif
[thin] [poor] [quality] [special] [price]	[0] [-2,1] [0] [0] [1,7] [0]	$\frac{-0,4}{\sqrt{-0,4^2 + 15}} = \frac{-0,4}{3,85}$ = -0,10	Negatif

Jadi, hasil analisis sentimen dari 10 data komentar suatu produk pakaian yaitu sentimen positif sebanyak 6 komentar, sentimen netral sebanyak 2 komentar dan sentimen negatif sebanyak 2 komentar. Dengan kesimpulan komentar mengenai suatu produk pakaian memiliki sentimen yang positif.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Pada penelitian ini, peneliti melakukan analisis sentimen terhadap data komentar tentang ChatGPT yang diutarakan oleh pengguna di situs Reddit. Dengan data yang digunakan yaitu berasal dari situs penyedia *dataset* yaitu Kaggle.com. Jumlah data awal pada *dataset* berjumlah 52.416 data. Setelah melalui tahapan-tahapan penelitian seperti *data reduction* dan *text preprocessing* data yang sesuai dengan permasalahan pada penelitian berjumlah 27.296 data.

Metode yang digunakan pada penelitian ini yaitu metode *Lexicon Based* dan menggunakan kamus kata VADER. Dengan tahapan-tahapan penelitian yang dilakukan yaitu pengumpulan data, data reduction, text preprocessing meliputi *case folding*, *text normalization, cleaning*, *tokenizing*, *stopword removal*, dan juga *stemming*, implementasi metode *Lexicon* VADER, hasil analisis sentimen, dan juga visualisasi data. Hasil dari penelitian ini yaitu data komentar yang memiliki jenis sentimen positif berjumlah 14.250 komentar dengan presentase sebanyak 53,2%, sentimen netral berjumlah 6.767 komentar dengan presentase sebanyak 24.8% dan sentimen negatif berjumlah 6.009 dengan presentase sebanyak 22.0% dari total keseluruhan jumlah data komentar sebanyak 27.296 data.

Validasi pada penelitian ini menggunakan metode validasi *Silhouette Index* dengan nilai *Silhouette Index* yang dihasilkan dari penelitian ini yaitu 0.532. Dengan nilai *Silhouette Index* mendekati angka 1, maka dapat diketahui

bahwasannya pada penelitian dengan mengimplementasikan metode *Lexicon Based* memiliki kualitas klasterisasi yang sudah baik.

Hasil penelitian ini menunjukkan bahwa komentar pengguna Reddit terhadap ChatGPT memiliki sentimen yang positif dan klasterisasi sentimen menggunakan metode *Lexicon Based* memiliki kualitas yang sudah baik. Sehingga, dapat disimpulkan bahwa pengguna Reddit menyambut hadirnya ChatGPT dengan komentar-komentar yang memiliki muatan positif.

5.2 Saran

Adapun saran dari penelitian ini yang peneliti utraikan sebagai berikut :

1. Penelitian ini hanya menggunakan metode *Lexicon Based* dan Kamus Kata VADER dan penelitian dapat dikembangkan dengan metode atau algoritma lain seperti *logistic regression*, *random forest*, dan *support vector machine.*
2. Penelitian ini dapat mengguanakan kamus kata atau *lexicon* yang lain seperti Liu-Hu, SentiWordNet, AFINN, dan kamus lexicon lainnya.
3. Tahapan *text preprocessing* dapat dilakukan lebih baik lagi dengan menambahkan beberapa tahapan seperti *spelling correction* dan *emoticon normalization*.
4. Menambahkan lebih banyak jenis sentimen tidak hanya positif, netral, serta negatif seperti semi positif, maupun semi negatif.
5. Menggunakan dataset dengan jumlah data yang lebih banyak agar dapat mengetahui polaritas sentimen lebih banyak.

DAFTAR PUSTAKA

Abimanyu, D. (2022). Analisis Sentimen Akun Twitter Apex Legends Menggunakan VADER. *Analisis Sentimen Akun Twitter Apex Legends Menggunakan VADER*, *5*(03), 423–431.

Alfarizi, M. R. S., Al-farish, M. Z., Taufiqurrahman, M., Ardiansah, G., & Elgar, M. (2023). Python Sebagai Bahasa Pemrograma Penggunaan Python Sebagai Bahasa Pemrograman Untuk Machine Learning Dan Deep Learning. *Karimah Tauhid*, *2*(1).

Alrajak, M. S., Ernawati, I., & Nurlaili, I. (2020). Analisis Sentimen Terhadap Pelayanan PT. PLN Di Jakarta Pada Twitter Dengan Algoritma K-Nearest Neighbor (K-NN). *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer Dan Aplikasinya*, *1*(2), 110–122.

Amaliah, F., & Nuryana, I. K. D. (2022). Perbandingan Akurasi Metode Lexicon Based Dan Naive Bayes Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter. *Journal of Informatics and Computer Science (JINACS)*, *3*(03), 384–393.

Aripiyanto, S., Tukino, T., Sufyan, A., & Nandaputra, R. (2022). Sentimen Analisis Twitter Ibu Kota Negara Nusantara Menggunakan Long Short-Term Memory dan Lexicon Based. *EXPERT: Jurnal Manajemen Sistem Informasi Dan Teknologi*, *12*(2), 119–125.

Ashari, H., Arifianto, D. A., & Al Faruq, H. A. (2020). Kinerja Algoritma Multinomial Naïve Bayes (Mnb), Multivariate Bernoulli Dan Rocchio Algorithm Dalam Klasifikasi Konten Berita Hoax Berbahasa Indonesia Dengan Jupyter Notebook. *JASIE: Jurnal Aplikasi Sistem Informasi Dan Elektronika*, *2*(2), 52–65.

Asri, Y., Suliyanti, W. N., Kuswardani, D., & Fajri, M. (2022). Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile. *Vol*, *15*, 264–275.

Birks, D., Coleman, A., & Jackson, D. (2020). Unsupervised Identification of Crime Problems from Police Free-text Data. *Crime Science*, *9*. https://doi.org/10.1186/s40163-020-00127-4

Cindo, M., Rini, D. P., & Ermatita, E. (2019). Literatur Review: Metode Klasifikasi Pada Sentimen Analisis. *Seminar Nasional Teknologi Komputer \& Sains (SAINTEKS)*, *1*(1).

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, *15*(1), 131–145.

Fathonah, F., & Herliana, A. (2021). Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid-19 Menggunakan Metode NaÃ{\={}} ve Bayes.

Jurnal Sains Dan Informatika, *7*(2), 155–164.

Fitriani, K., Isbandi, I., & Amaliyah, A. (2021). Perancangan Sistem Manajemen Dokumen Dengan Menggunakan Metode Text Mining Di Kantor Kelurahan Sekejati. *Telematika*, *3*(1), 45–59.

Fitriyah, N., Warsito, B., & Di Asih, I. M. (2020). Analisis Sentimen Gojek Pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (SVM). *Jurnal Gaussian*, *9*(3), 376–390.

Guntara, R. G. (2023a). Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7. *Jurnal Teknologi Dan Sistem Informasi Bisnis*, *5*(1), 55–60.

Guntara, R. G. (2023b). Visualisasi Data Laporan Penjualan Toko Online Melalui Pendekatan Data Science Menggunakan Google Colab. *ULIL ALBAB: Jurnal Ilmiah Multidisiplin*, *2*(6), 2091–2100.

Habibah, N. (2023). Analisis Sentimen Mengenai Penggunaan E-Wallet Pada Google Play Menggunakan Lexicon Based Dan K-Nearest Neighbor. *Analisis Sentimen Mengenai Penggunaan E-Wallet Pada Google Play Menggunakan Lexicon Based Dan K-Nearest Neighbor*, *1*(1), 192–200.

Hernikawati, D., & others. (2021). Kecenderungan Tanggapan Masyarakat Terhadap Vaksin Sinovac Berdasarkan Lexicon Based Sentiment Analysis (The Trend of Public Response to Sinovac Vaccine Based on Lexicon Based Sentiment Analysis). *Jurnal Iptekkom (Jurnal Ilmu Pengetahuan \& Teknologi Informasi)*, *23*(1), 21–31.

Hidayati, A. R., Fitrani, A. S., & Rosid, M. A. (2023). Analisa Sentimen Pemilu 2019 Pada Judul Berita Online Menggunakan Metode Logistic Regression. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer Dan Manajemen)*, *4*(2), 298–305.

Ismail, A. R., & Hakim, R. B. F. (2023). Implementasi Lexicon Based Untuk Analisis Sentimen Dalam Menentukan Rekomendasi Pantai Di DI Yogyakarta Berdasarkan Data Twitter: Implementasi Lexicon Based. *Emerging Statistics and Data Science Journal*, *1*(1), 37–46.

Julianto, I. T. (2022). Analisis Sentimen Terhadap Sistem Informasi Akademik Institut Teknologi Garut. *Jurnal Algoritma*, *19*(1), 458–468.

Krohn, R., & Weninger, T. (2022). Subreddit Links Drive Community Creation and User Engagement on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 536–547.

Long, S., Lucey, B., Xie, Y., & Yarovaya, L. (2023). “I just like the stock”: The role of Reddit sentiment in the GameStop share rally. *Financial Review*, *58*(1), 19–37.

Misnawati, M. (2023). ChatGPT: Keuntungan, Risiko, Dan Penggunaan Bijak Dalam Era Kecerdasan Buatan. *Prosiding Seminar Nasional Pendidikan, Bahasa, Sastra, Seni, Dan Budaya*, *2*(1), 54–67.

Nahdliyah, M. A., Widiharih, T., & Prahutama, A. (2019). Metode K-Medoids Clustering dengan Validasi Silhouette Index dan C-Index (Studi Kasus Jumlah Kriminalitas Kabupaten/Kota di Jawa Tengah Tahun 2018). *Jurnal Gaussian*, *8*(2), 161–170.

Pabubung, M. R. (2021). Epistemologi Kecerdasan Buatan (AI) dan Pentingnya Ilmu Etika dalam Pendidikan Interdisipliner. *Jurnal Filsafat Indonesia*, *4*(2), 152–159.

Putri, L. R., Sudarsono, S. C., & Wardani, M. M. S. (2021). Kekerasan Verbal Dalam Kolom Komentar di Akun Instagram Garudarevolution Pada Bulan September 2019. *Sintesis*, *15*(1), 32–56.

Rachmawaty, A. (2021). Optimasi Media Sosial Dalam Meningkatkan Penjualan di Masa Pembatasan Sosial Berskala Besar. *Tematik: Jurnal Teknologi Informasi Komunikasi (e-Journal)*, *8*(1), 29–44.

Reinaldi, Y., Ulinnuha, N., & Hafiyusholeh, M. (2021). Comparison of Single Linkage, Complete Linkage, and Average Linkage Methods on Community Welfare Analysis in Cities and Regencies in East Java. *Jurnal Matematika, Statistika Dan Komputasi*, *18*(1), 130–140.

Rendragraha, A. D., Bijaksana, M. A., & Romadhony, A. (2021). Pendekatan Metode Transformers Untuk Deteksi Bahasa Kasar Dalam Komentar Berita Online Indonesia. *EProceedings of Engineering*, *8*(2).

Rosyadi, H. E., Amrullah, F., Marcus, R. D., & Affandi, R. R. (2020). Rancang bangun chatbot informasi lowongan pekerjaan berbasis Whatsapp dengan metode NLP (Natural Language Processing). *Briliant: Jurnal Riset Dan Konseptual*, *5*(3), 619–626.

Setiawan, A., & Luthfiyani, U. K. (2023). Penggunaan ChatGPT Untuk Pendidikan di Era Education 4.0: Usulan Inovasi Meningkatkan Keterampilan Menulis. *JURNAL PETISI (Pendidikan Teknologi Informasi)*, *4*(1), 49–58.

Soen, G. I. E., Marlina, M., & Renny, R. (2022). Implementasi Cloud Computing dengan Google Colaboratory pada Aplikasi Pengolah Data Zoom Participants. *JITU: Journal Informatic Technology And Communication*, *6*(1), 24–30.

Sumitro, P. A., Mulyana, D. I., Saputro, W., & others. (2021). Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode Lexicon Based. *Jurnal Informatika Dan Teknologi Komputer (J-ICOM)*, *2*(2), 50–56.

Utama, H. S., Rosiyadi, D., Prakoso, B. S., Ariadarma, D., & others. (2019). Analisis sentimen sistem ganjil genap di tol Bekasi menggunakan algoritma Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 243–250.

Vindua, R., & Zailani, A. U. (2023). Analisis Sentimen Pemilu Indonesia Tahun 2024 Dari Media Sosial Twitter Menggunakan Python. *JURIKOM (Jurnal Riset Komputer)*, *10*(2), 479–487.

Yunefri, Y., Fadrial, Y. E., & Sutejo, S. (2021). Chatbot Pada Smart Cooperative Oriented Problem Menggunakan Natural Language Processing dan Naive Bayes Classifier. *INTECOMS: Journal of Information Technology and Computer Science*, *4*(2), 131–140.

LAMPIRAN

1. Lampiran Hasil Plagiasi

turnitin

Similarity Report ID: oid:29477:58024774

PAPER NAME
SKRIPSI 198160044.pdf

AUTHOR
Pagas Sukarno Putra

WORD COUNT
9822 Words

CHARACTER COUNT
59513 Characters

PAGE COUNT
59 Pages

FILE SIZE
1.4MB

SUBMISSION DATE
Apr 22, 2024 8:42 AM GMT+7

REPORT DATE
Apr 22, 2024 8:43 AM GMT+7

19% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 16% Internet database
- 2% Publications database
- Crossref database
- Crossref Posted Content database
- 16% Submitted Works database

Excluded from Similarity Report

- Small Matches (Less then 10 words)

Summary

2. Lampiran SK Pembimbing Tugas Akhir

UNIVERSITAS MEDAN AREA
FAKULTAS TEKNIK

Kampus I : Jalan Kolam Nomor 1 Medan Estate/Jalan PBSI Nomor 1☎(061) 7366878, 7360168, 7364348, 7366781, Fax.(061) 7366998 Medan 20223
Kampus II : Jalan Setiabudi Nomor 79 / Jalan Sei Serayu Nomor 70 A, ☎ (061) 8225602, Fax. (061) 8226331 Medan 20122
Website: www.teknik.uma.ac.id E-mail: univ_medanarea@uma.ac.id

Nomor : 760/FT.6/01.10/X/2023 30 Oktober 2023
Lamp : -
H a l : **Perpanjang SK Pembimbing Tugas Akhir**

Yth. Pembimbing Tugas Akhir
Andre Hasudungan Lubis, S. Ti, MSc
di
Tempat

Dengan hormat,
Sehubungan telah berakhirnya waktu masa berlaku SK pembimbing nomor 311/FT.6/01.10/IV/2023 tertanggal 13 April 2023 maka perlu diterbitkan kembali SK Pembimbing Skripsi baru atas nama mahasiswa berikut :

N a m a : Pagas Sukarno Putra
N P M : 198160044
Jurusan : Teknik Informatika

Oleh karena itu kami mengharapkan kesediaan saudara :

Andre Hasudungan Lubis, S. Ti, MSc (Sebagai Pembimbing)

Adapun Tugas Akhir Skripsi berjudul :

"Analisis Sentimen tentang *ChatGPT* pada Kolom Komentar Situs *Reddit* dengan Metode *Lexicon Based*"

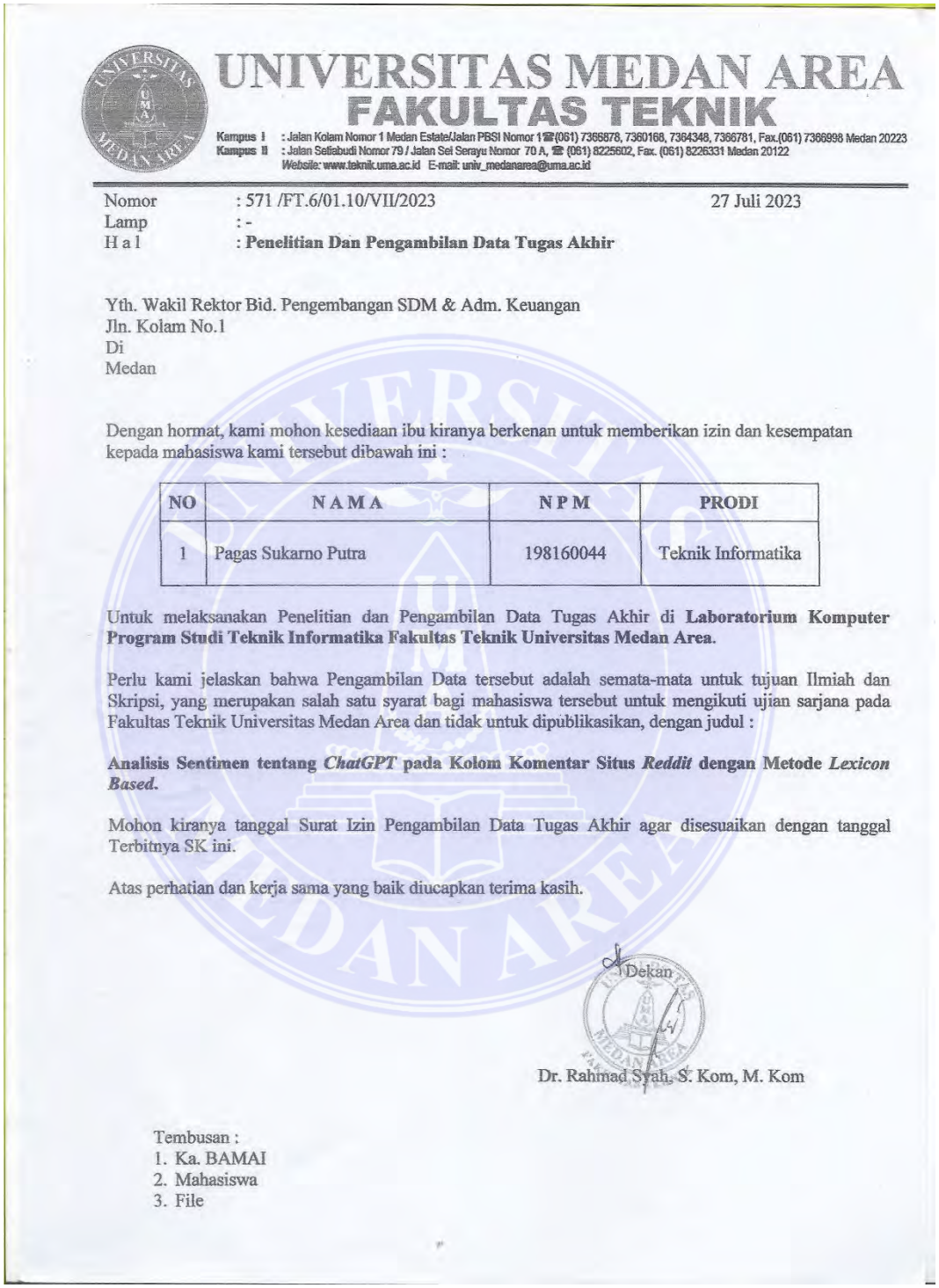
SK Pembimbing ini berlaku selama enam bulan terhitung sejak SK ini diterbitkan. Jika proses pembimbing melebihi batas waktu yang telah ditetapkan, SK ini dapat ditinjau ulang.

Demikian kami sampaikan, atas kesediaan saudara diucapkan terima kasih.

Dekan,

Dr. Rahmad Syah, S. Kom, M. Kom

3. Lampiran Surat Penelitian dan Pengambilan Data Tugas Akhir



UNIVERSITAS MEDAN AREA
FAKULTAS TEKNIK
Kampus I : Jalan Kolam Nomor 1 Medan Estate/Jalan PBSI Nomor 1☎(061) 7366878, 7360168, 7364348, 7366781, Fax.(061) 7366998 Medan 20223
Kampus II : Jalan Setiabudi Nomor 79 / Jalan Sei Serayu Nomor 70 A, ☎ (061) 8225602, Fax. (061) 8226331 Medan 20122
Website: www.teknik.uma.ac.id E-mail: univ_medanarea@uma.ac.id

Nomor : 571 /FT.6/01.10/VII/2023 27 Juli 2023
Lamp : -
H a l : **Penelitian Dan Pengambilan Data Tugas Akhir**

Yth. Wakil Rektor Bid. Pengembangan SDM & Adm. Keuangan
Jln. Kolam No.1
Di
Medan

Dengan hormat, kami mohon kesediaan ibu kiranya berkenan untuk memberikan izin dan kesempatan kepada mahasiswa kami tersebut dibawah ini :

NO	N A M A	N P M	PRODI
1	Pagas Sukarno Putra	198160044	Teknik Informatika

Untuk melaksanakan Penelitian dan Pengambilan Data Tugas Akhir di **Laboratorium Komputer Program Studi Teknik Informatika Fakultas Teknik Universitas Medan Area.**

Perlu kami jelaskan bahwa Pengambilan Data tersebut adalah semata-mata untuk tujuan Ilmiah dan Skripsi, yang merupakan salah satu syarat bagi mahasiswa tersebut untuk mengikuti ujian sarjana pada Fakultas Teknik Universitas Medan Area dan tidak untuk dipublikasikan, dengan judul :

Analisis Sentimen tentang *ChatGPT* pada Kolom Komentar Situs *Reddit* dengan Metode *Lexicon Based*.

Mohon kiranya tanggal Surat Izin Pengambilan Data Tugas Akhir agar disesuaikan dengan tanggal Terbitnya SK ini.

Atas perhatian dan kerja sama yang baik diucapkan terima kasih.

Dekan

Dr. Rahmad Syah, S. Kom, M. Kom

Tembusan :
1. Ka. BAMAI
2. Mahasiswa
3. File

4. Lampiran Surat Keterangan Selesai Penelitian

UNIVERSITAS MEDAN AREA

Kampus I : Jalan Kolam Nomor 1 Medan Estate ☎ (061) 7360168, 7366878, 7364348 🖷 (061) 7368012 Medan 20223
Kampus II : Jalan Setiabudi Nomor 79 / Jalan Sei Serayu Nomor 70 A ☎ (061) 8225602 🖷 (061) 8226331 Medan 20122
Website: www.uma.ac.id *E-Mail:* univ_medanarea@uma.ac.id

SURAT KETERANGAN
Nomor : 2114 /UMA/B/01.7/XI/2023

Rektor Universitas Medan Area dengan ini menerangkan bahwa :

Nama	: Pagas Sukarno Putra
No. Pokok Mahasiswa	: 198160044
Program Studi	: Teknik Informatika
Fakultas	: Teknik

Benar telah selesai Pengambilan Data Tugas Akhir di Laboratorium Komputer Universitas Medan Area dengan Judul Skripsi "**Analisis Sentimen Tentang ChatGPT Pada Kolom Komentar Situs Reddit Denga Metode *Lexicon Based***".

Dan kami harapkan Data tersebut kiranya dapat membantu yang bersangkutan dalam penyusunan skripsi dan dapat bermanfaat bagi mahasiswa khususnya Fakultas Teknik

Demikian surat ini diterbitkan untuk dapat digunakan seperlunya.

Medan, 20 November 2023.
a.n Rektor
Wakil Rektor Bidang Pengembangan SDM &
Administrasi Keuangan,

Prof. Dr. Ir. Suswati, MP

Tembusan :
1. Mahasiswa Ybs
2. File

5. Lampiran Source Code Google Colaboratory

 # DATA REDUCTION

Reddit_ChatGPT = Reddit_ChatGPT.loc[Reddit_ChatGPT['subreddit'] ==
'r/ChatGPT']

Reddit_ChatGPT_df = pd.DataFrame(Reddit_ChatGPT['comment_body'])

#TEXT PREPROCESSING : CASE FOLDING

Reddit_ChatGPT_df['case_folding'] =
Reddit_ChatGPT_df['comment_body'].str.lower()

#TEXT PREPROCESSING : NORMALIZATION

singkatan={ "what's":"what is", "what're":"what are", "who's":"who is",
"who're":"who are", "where's":"where is", "where're": "where re","when's":"when
is", "when're":"when are", "how's":"how is", "how're":"how are", "i'm":"i am",
"we're":"we are", "you're":"you are", "they're":"they are", "it's":"it is", "he's":"he
is", "she's":"she is", "that's":"that is", "there's":"there is", "there're":"there
are","i've":"i have", "we've":"we have", "you've":"you have", "they've":"they
have", "who've":"who have", "would've":"would have", "not've":"not have",
"i'll":"i will", "we'll":"we will", "you'll":"you will", "he'll":"he will", "she'll":"she
will", "it'll":"it will", "they'll":"they will", "isn't":"is not", "wasn't":"was not",
"aren't":"are not", "weren't":"were not", "can't":"can not", "couldn't":"could not",
"don't":"do not", "didn't":"did not", "shouldn't":"should not", "wouldn't":"would
not", "doesn't":"does not", "haven't":"have not", "hasn't":"has not", "hadn't":"had
not", "won't":"will not", "ngl":"not gonna lie", "lol":"laughing out loud",

```
"wtf":"what the fuck", "omg":"oh my god", "aka":"also known as", "ily":"i love
you","omfg":"oh my fucking god", "btw":"by the way", "fyi":"for your
information", "jk":"just kidding", "wth":"what the hell", "ty":"thank
you","lmao":"laughing my ass off",}
Reddit_ChatGPT_df['normalization'] =
Reddit_ChatGPT_df['case_folding'].replace(singkatan,regex=True)
```

#TEXT PREPROCESSING : CLEANING

```
# menghapus tag yang berisi u/ yaitu username dan r/ yaitu subreddit
Reddit_ChatGPT_df['cleaning'] =
Reddit_ChatGPT_df['normalization'].str.replace('r/|/r/\S+|r/\S+|/u/\S+|u/\S+|gif\S*'
, '', regex=True)
# menghapus tag akun yang berisi awalan @ serta email serta tagar
Reddit_ChatGPT_df['cleaning'] =
Reddit_ChatGPT_df['cleaning'].str.replace('\S*@\S*\s?|\S*#\S*\s?', '',
regex=True)
# menghapus url link http dan https pada data
Reddit_ChatGPT_df['cleaning'] =
Reddit_ChatGPT_df['cleaning'].str.replace('http\S+|www.\S+', '', regex=True)
# menghapus data yang berisi [deleted] dan [removed] pada data
Reddit_ChatGPT_df['cleaning'] =
Reddit_ChatGPT_df['cleaning'].str.replace('\[deleted]|\[removed]', '', regex=True)
# membersihkan tanda baca yang tidak diperlukan
```

```
Reddit_ChatGPT_df['cleaning'] = Reddit_ChatGPT_df['cleaning'].str.replace('-', '
', regex=True)
Reddit_ChatGPT_df['cleaning'] =
Reddit_ChatGPT_df['cleaning'].str.replace('[^A-Za-z ]', '', regex=True)
#membersihkan baris yang kosong dan juga baris dengan duplikat data
Reddit_ChatGPT_df.drop(Reddit_ChatGPT_df[Reddit_ChatGPT_df.cleaning ==
''].index, inplace=True)
Reddit_ChatGPT_df.drop_duplicates(subset=['cleaning'], keep ='first',
inplace=True)
```

#TEXT PREPROCESSING : TOKENIZING

```
from nltk.tokenize import word_tokenize
Reddit_ChatGPT_df['tokenizing'] =
Reddit_ChatGPT_df['cleaning'].apply(word_tokenize)
```

#TEXT PREPROCESSING : STOPWORD REMOVAL

```
from nltk.corpus import stopwords
stop_words = set(stopwords.words("english"))
new_stopwords =
['ai','artificial','intellegence','machine','learning','chatgpt','gpt','google','googles','ad
obe','amazon','email','program','photoshop','twitter','reddit','subreddit','facebook','ti
ktok','youtube','assistant','chatbot','robot','bot','bots','game','games','language','vs','i
nternet','model','models','users','user','prompt','syntax','command','commands','com
```

```
ment','comments','post','posts','posted','topic','context','contexts','system','systems','
video','text','texts','texting']
stop_words = stop_words.union(new_stopwords)
Reddit_ChatGPT_df['stopword_removal'] =
Reddit_ChatGPT_df['tokenizing'].apply(lambda stopword: [words for words in
stopword if words not in stop_words])
```

#TEXT PREPROCESSING : STEMMING

```
from krovetzstemmer import Stemmer
stemmer = Stemmer()
Reddit_ChatGPT_df['stemming'] =
Reddit_ChatGPT_df['stopword_removal'].apply(lambda x: [stemmer.stem(y) for
y in x])
# Mengembalikan data text yang sebelumnya di tokenizer ke dalam bentuk awal
from mosestokenizer import MosesDetokenizer
detokenizer = MosesDetokenizer('en')
Reddit_ChatGPT_df['preprocessing']=Reddit_ChatGPT_df['stemming'].apply(lam
bda x: detokenizer(x))

# Mengecek kembali apakah masih ada baris yang kosong dan data yang duplikat,
Reddit_ChatGPT_df.drop(Reddit_ChatGPT_df[Reddit_ChatGPT_df.preprocessin
g == ''].index, inplace=True)
Reddit_ChatGPT_df.drop_duplicates(subset=['preprocessing'], keep ='first',
inplace=True)
```

```
# mereset index pada dataframe sehingga nomor menjadi berurutan kembali
Reddit_ChatGPT_df.reset_index(drop=True, inplace=True)
Reddit_ChatGPT_df.index = np.arange(1, len(Reddit_ChatGPT_df) + 1)
preprocessing_df = pd.DataFrame(Reddit_ChatGPT_df['preprocessing'])

# IMPLEMENTASI VADER LEXICON
from nltk.sentiment.vader import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
# Syntax untuk proses mencari nilai polaritas pada kata di dataset
preprocessing_df['Positif'] = [analyzer.polarity_scores(x)['pos'] for x in
preprocessing_df['preprocessing']]
preprocessing_df['Netral'] = [analyzer.polarity_scores(x)['neu'] for x in
preprocessing_df['preprocessing']]
preprocessing_df['Negatif'] = [analyzer.polarity_scores(x)['neg'] for x in
preprocessing_df['preprocessing']]
preprocessing_df['Compound Score'] = [analyzer.polarity_scores(x)['compound']
for x in preprocessing_df['preprocessing']]
preprocessing_df['Jenis Sentimen']=''
preprocessing_df.loc[preprocessing_df['Compound Score'] >=0.05,'Jenis
Sentimen']='Positif'
preprocessing_df.loc[(preprocessing_df['Compound Score'] >-0.05) &
(preprocessing_df['Compound Score'] <0.05),'Jenis Sentimen']='Netral'
preprocessing_df.loc[preprocessing_df['Compound Score'] <=-0.05,'Jenis
Sentimen']='Negatif'
```

```
# Mengitung jumlah data pada tiap jenis sentimern
jumlah_sentimen = preprocessing_df['Jenis Sentimen'].value_counts()
jumlah_sentimen = jumlah_sentimen.to_frame(name='jumlah')
jumlah_sentimen = jumlah_sentimen.reset_index()
jumlah_sentimen.index = np.arange(1, len(jumlah_sentimen) + 1)

# Validasi dengan Silhouette Index (SI)
from sklearn.metrics import silhouette_score
positif = np.array(positif_df['Compound Score'][:6000])
netral = np.array(netral_df['Compound Score'][:6000])
negatif = np.array(negatif_df['Compound Score'][:6000])
data = np.concatenate([negatif, positif, netral])
labels = [0]*len(positif) + [1]*len(netral) + [2]*len(negatif)
score = silhouette_score(data.reshape(-1, 1), labels)
print("Silhouette Score: ", score)
# Visualisasi
# Diagram Batang
import matplotlib.pyplot as plt
data = preprocessing_df['Jenis Sentimen'].value_counts()
ax = plt.axes()
ax.bar(data.index, data.values, width=0.4, color='#429FFD')
plt.title('Sentimen Tentang ChatGPT di Situs Reddit')
plt.xlabel('Jenis Sentimen')
plt.ylabel('Jumlah')
```

```
plt.xticks(rotation=50)
for p in ax.patches:
    x = p.get_x() + p.get_width() / 2
    y = p.get_y() + p.get_height() / 2
    value = int(p.get_height())
    ax.text(x, y, value, ha='center', va='center')
plt.show()
```

#Diagram Lingkaran

```
import matplotlib.pyplot as plt
data = preprocessing_df['Jenis Sentimen'].value_counts()
colors = ['#279EFF', '#FFE867', '#F05454']
ax = plt.axes()
ax.pie(data, autopct='%1.1f%%', colors=colors)
plt.title('Sentimen Tentang ChatGPT di Situs Reddit')
ax.legend(data.index, loc='upper right', bbox_to_anchor=(1.2, 1))
plt.show()
```

 # Wordcloud untuk data dengan sentimen positif

```
import wordcloud
import matplotlib.pyplot as plt
wc_positif = " ".join(positif_df['preprocessing'])
wc = wordcloud.WordCloud(stopwords=wordcloud.STOPWORDS,
background_color="white", width=800, height=400).generate(wc_positif)
```

```
plt.axis ("off")
plt.imshow(wc)
plt.show()
```

 # Wordcloud untuk data dengan sentimen netral

```
import wordcloud
import matplotlib.pyplot as plt
wc_netral = " ".join(netral_df['preprocessing'])
wc = wordcloud.WordCloud(stopwords=wordcloud.STOPWORDS,
background_color="white", width=800, height=400).generate(wc_netral)
plt.axis ("off")
plt.imshow(wc)
plt.show()
```

 # Wordcloud untuk data dengan sentimen negatif

```
import wordcloud
import matplotlib.pyplot as plt
wc_negatif = " ".join(negatif_df['preprocessing'])
wc = wordcloud.WordCloud(stopwords=wordcloud.STOPWORDS,
background_color="white", width=800, height=400).generate(wc_negatif)
plt.axis ("off")
plt.imshow(wc)
plt.show()
```