ANALISIS DAN PERBANDINGAN STOPWORD TERHADAP AKURASI ANALISIS SENTIMEN TEKS DENGAN MENGGUNAKAN TF-IDF STUDI KASUS NLP

SKRIPSI

Diajukan sebagai Salah Satu Syarat untuk Mendapatkan

Gelar Sarjana Komputer (S.Kom) Program Studi Informatika

Damai Arsila Salsabila

105841107520

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS MUHAMMADIYAH MAKASSAR

2024

ANALISIS DAN PERBANDINGAN STOPWORD TERHADAP AKURASI ANALISIS SENTIMEN TEKS DENGAN MENGGUNAKAN TF-IDF STUDI KASUS NLP

Diajukan sebagai Salah Satu Syarat untuk Mendapatkan
Gelar Sarjana Komputer (S.Kom) Program Studi Informatika

Disusun dan Diajukan Oleh:

Damai Arsila Salsabila
105841107520

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MUHAMMADIYAH MAKASSAR
2024

UNIVERSITAS MUHAMMADIYAH MAKASSAR

FAKULTAS TEKNIK

GEDUNG MENARA IQRA LT. 3

Jl. Sultan Alauddin No. 259 Telp. (0411) 866 972 Fax (0411) 865 588 Makassar 90221
Website: www.unismuh.ac.id, e_mail: unismuh@gmail.com
Website: http://teknik.unismuh.makassar.ac.id

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

PENGESAHAN

Skripsi atas nama Damai Arsila Salsabila dengan nomor induk Mahasiswa 105 84 11075 20, dinyatakan diterima dan disahkan oleh Panitia Ujian Tugas Akhir/Skripsi sesuai dengan Surat Keputusan Dekan Fakultas Teknik Universitas Muhammadiyah Makassar Nomor : 110/05/A.5-VI/IV/45/2024, sebagai salah satu syarat guna memperoleh gelar Sarjana Komputer pada Program Studi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar pada hari Sabtu tanggal 24 Agustus 2024.

Makassar, 19 Safar 1446 H / 24 Agustus 2024 M

Panitia Ujian :

1. Pengawas Umum
   a. Rektor Universitas Muhammadiyah Makassar
      Dr. Ir. H. Abd. Rakhim Nanda, ST., MT., IPU. : .....
   b. Dekan Fakultas Teknik Universitas Hasanuddin
      Prof. Dr. Eng. Muhammad Isran Ramli, ST., MT. : .....
2. Penguji
   a. Ketua : Dr. Ir. Hj. Hafsah Nirwana, ST., MT. : .....
   b. Sekertaris : Dr. Ir. Ridwang, S.Kom., MT. : .....
3. Anggota : 1. Muhyiddin A.M. Hayat, S.Kom., M.T. : .....
   2. Rizki Yusliana Bakti ST., MT. : .....
   3. Lukman Anas, S.Kom., MT. : .....

Mengetahui :

Pembimbing I	Pembimbing II
Fahrim Irhamna Rahman S.Kom., M.T	**Titin Wahyuni S.Pd., MT**

Dekan

Dr. Ir. Hj. Nurnawaty, ST., MT., IPM.
NBM : 795 108

UNIVERSITAS MUHAMMADIYAH MAKASSAR

FAKULTAS TEKNIK

GEDUNG MENARA IQRA LT. 3

Jl. Sultan Alauddin No. 259 Telp. (0411) 866 972 Fax (0411) 865 588 Makassar 90221

Website: www.unismuh.ac.id, e_mail: unismuh@gmail.com

Website: http://teknik.unismuh.makassar.ac.id

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

HALAMAN PENGESAHAN

Tugas Akhir ini diajukan untuk memenuhi syarat ujian guna memperoleh gelar Sarjana Komputer (S.Kom) Program Studi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar.

Judul Skripsi : **ANALISIS DAN PERBANDINGAN STOPWORD TERHADAP AKURASI ANALISIS SENTIMEN TEKS DENGAN MENGGUNAKAN TF-IDF STUDI KASUS NLP**

Nama : DAMAI ARSILA SALSABILA

Stambuk : 105841107520

Makassar, 28 Agustus 2024

Telah Diperiksa dan Disetujui
Oleh Dosen Pembimbing;

Pembimbing I	Pembimbing II
Fahrim Irhamna Rahman S.Kom., MT.	**Titin Wahyuni, S.Pd., M.T.**

Mengetahui,
Ketua Program Studi Informatika

Muhyiddin A M Hayat, S.Kom., MT.
NBM : 1504 577

ABSTRAK

Damai Arsila Salsabila. ANALISIS DAN PERBANDINGAN STOPWORD TERHADAP AKURASI ANALISIS SENTIMEN TEKS DENGAN MENGGUNAKAN TF-IDF STUDI KASUS NLP (dibimbing oleh Fahrim Irhamna Rachman S.Kom., M.T. dan Titin Wahyuni S.Pd., M.T).

Dalam era digital yang berkembang pesat, jumlah data teks online meningkat signifikan, mencakup ulasan produk, komentar media sosial, dan artikel berita. Analisis sentimen penting untuk memahami opini masyarakat. Penelitian ini bertujuan untuk mengembangkan daftar stopwords yang lebih relevan menggunakan algoritma TF-IDF untuk meningkatkan representasi teks dalam analisis sentimen. serta mengevaluasi dan membandingkan pengaruh penggunaan stopwords yang dihasilkan oleh algoritma TF-IDF terhadap akurasi model analisis sentimen, dibandingkan dengan penggunaan stopwords Sastrawi. Hasil menunjukkan TF-IDF membantu mengidentifikasi kata-kata kurang penting, namun stopwords Sastrawi lebih baik mengenali konteks. Evaluasi dengan rasio pembagian data yang berbeda (90:10, 80:20, 70:30) menunjukkan akurasi tertinggi sebesar 0.789 pada rasio 80:20, meskipun ada ruang untuk peningkatan. Studi ini diharapkan dapat meningkatkan kinerja model analisis sentimen dengan daftar stopwords yang lebih sesuai.

Kata Kunci : Analisis Sentimen, TF-IDF, NLP, Stopwords

ABSTRACT

Damai Arsila Salsabilah. *ANALYSIS AND COMPARISON OF STOPWORDS ON TEXT SENTIMENT ANALYSIS ACCURACY USING TF-IDF: A CASE STUDY IN NLP (supervised by* Fahrim Irhamna Rahman S.Kom., M.T. *and* Titin Wahyuni S.Pd., M.T*).*

In the rapidly evolving digital era, the amount of online text data has significantly increased, encompassing product reviews, social media comments, and news articles. Sentiment analysis is crucial for understanding public opinion. This research aims to develop a more relevant stopword list using the TF-IDF algorithm to enhance text representation in sentiment analysis. Additionally, it evaluates and compares the impact of using stopwords generated by the TF-IDF algorithm on the accuracy of sentiment analysis models, compared to using Sastrawi stopwords. The results show that TF-IDF helps identify less important words, but Sastrawi stopwords are better at recognizing context. Evaluation with different data split ratios (90:10, 80:20, 70:30) showed the highest accuracy of 0.789 at the 80:20 ratio, although there is room for improvement. This study is expected to improve the performance of sentiment analysis models with a more suitable stopword list.

Keywords*: Sentiment Analysis, TF-IDF, NLP, Stopwords.*

KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa, atas segala rahmat dan karunia-Nya sehingga kami dapat menyelesaikan penelitian dengan judul "Analisis dan Perbandingan Stopword terhadap Akurasi Analisis Sentimen Teks dengan Menggunakan TF-IDF: Studi Kasus NLP" ini dengan baik dan tepat waktu.

Penelitian ini disusun sebagai salah satu syarat untuk menyelesaikan program studi Informatika. Dalam penelitian ini, kami membahas tentang bagaimana penggunaan stopword dapat mempengaruhi akurasi dalam analisis sentimen teks. Kami menggunakan pendekatan *Term Frequency-Inverse Document Frequency* (TF-IDF) sebagai metode untuk mengukur bobot kata-kata dalam teks, yang kemudian digunakandalam proses klasifikasi sentimen.

Studi kasus yang kami ambil adalah analisis sentimen terhadap teks dalam konteks Pemrosesan Bahasa Alami (*Natural Language Processing*/NLP). Kami berharap bahwa hasil dari penelitian ini dapat memberikan kontribusi yang berarti bagi perkembangan teknologi NLP, khususnya dalam hal optimalisasi analisis sentimen.

Kami menyadari bahwa penelitian ini tidak akan terwujud tanpa bantuan dan dukungan dari berbagai pihak. Oleh karena itu, kami ingin menyampaikan terima kasihyang sebesar-besarnya kepada:

1. Bapak, ayah dan ibu tercinta, yang selalu memberikan doa dan semangat tanpa henti.
2. Suami saya terima kasih atas cinta, pengertian, dan dukungan yang telah kamu berikan.
3. Bapak Prof. Dr. H. Ambo Asse, M.Ag., sebagai Rektor Perguruan Tinggi Universitas Muhammadiyah Makassar

4. Ibu Dr.Hj.Ir. Nurnawaty S.T., M.T Selaku Dekan Fakultas Teknik Universitas Muhammadiyah Makassar
5. Bapak Muhydin A.M. Hayat S.Kom, M.T Selaku Ketua Prodi Informatika, Fakultas Teknik Universitas Muhammadiyah Makassar.
6. Bapak Fahrim Irhamna Rachman, S.kom., M.T selaku Dosen Pembimbing I dan Ibu Titin Wahyuni S.Pd., M.T selaku Dosen Pembimbing II yang senantiasa meluangkan waktu dan pikirannya untuk membimbing dan mengarahkan penulis dalam penyusunan skripsi ini.
7. Teman – teman kelas C, yang selalu memberikan dukungan moral dan berbagi pengetahuan selama proses penelitian ini.

Kami menyadari bahwa penelitian ini masih memiliki keterbatasan dan kekurangan. Oleh karena itu, kami membuka diri terhadap kritik dan saran yang membangun demi penyempurnaan penelitian ini di masa mendatang. Akhir kata, semoga penelitian ini dapat memberikan manfaat bagi semua pihak yang membutuhkan, dan menjadi sumbangsih bagi perkembangan ilmu pengetahuan di bidang NLP.

Makassar Agustus 2024

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

DAFTAR ISTILAH

Machine learning	:	*Machine learning* adalah disiplin ilmu yang terletak dibawah payung *Artificial Intelligence* (AI), dimana pendekatan ini menggabungkan konsep - konsep dari ilmu komputer dan statistik.
Naïve Bayes	:	*Naïve Bayes* adalah metode klasifikasi yang paling dasar dan paling sering digunakan. Model klasifikasi ini mengestimasikan probabilitas posterior suatu kelas berdasarkan distribusi kata suatu dokumen, dengan menggunakan representasi dokumen yang sederhana sebagai *Bag of Words*.
NLP	:	NLP atau Natural Language Processing merupakan salah satu disiplin ilmu dalam bidang kecerdasan buatan yang berfokus pada pemahaman dan pengolahan bahasa manusia oleh komputer.
NLTK	:	NLTK atau *Natural Language ToolKit* adalah sebuah perpustakaan yang berguna untuk membantu pengguna dlam mengelolah teks.
Stopword	:	Stopword merupakan jenis kata yang sering dianggap tidak memberikan kontribusi signifikan dalam pemrosesan teks dalam bidang pemrosesan bahasa alamu.
TF-IDF	:	TF-IDF merupakan metode yang digunakan untuk mengintegrasikan dua teori berdasarkan keberadaan kata dalam dokumen.

BAB I

PENDAHULUAN

A. Latar Belakang Masalah

Dalam pesatnya pertumbuhan digital saat ini, jumlah data teks yang dihasilkan dan dibagikan secara online meningkat secara eksponen. Data ini mencangkup berbagai jenis informasi mulai dari ulasan produk, komentar media social, hingga artikel berita. Dalam hal ini analisis sentiment menjadi penting dalam memahami opini, sentiment, dan pandangan masyarakat terhadap berbagai topik dan produk. Analisis sentiment merupakan salah satu Teknik Natural Language Processing (NPL) yang popular digunakan untuk mengekstraksi informasi penting dari data teks tersebut dengan cara mengidentifikasi sentiment atau opini yang terkandung didalamnya(Septian etal., 2019).

Salah satu langkah penting dalam analisis sentimen adalah tahap praproses teks, di mana teks dibersihkan dari kata-kata yang dianggap tidak memiliki kontribusi signifikan terhadap makna atau sentimen dari teks tersebut. Kata-kata ini disebut sebagai stopwords (Kusumawardana, 2020). Stopwords umum seperti "yang", "dan", "di", seringkali dihapus untuk meningkatkan efisiensi dan akurasi model analisis teks. Namun, daftar stopwords yang digunakan biasanya bersifat umum dan mungkin tidak selalu optimal untuk semua jenis teks atau domain tertentu.

Salah satu pendekatan yang dapat digunakan untuk melakukan analisis sentimen adalah metode *Term Frequency-Inverse Document Frequency* (TF- IDF) Metode ini merupakan suatu metode pembobotan yang umum digunakandalam pemrosesan bahasa alami untuk menetapkan nilai bobot setiap kata dalam dokumen, yang didasarkan pada seberapa penting kata tersebut (Syahril Dwi Prasetyo et al., 2023). Metode ini juga dapat

membantu mengidentifikasi kata-kata yang lebih relevan atau signifikan dalam konteks tertentu dibandingkan dengan menggunakan daftar stopwords standar.

Beberapa pustaka umum yang digunakan dalam pemrosen bahasa alami (*Natural Language Processing* atau NLP) untuk bahasa indonesia adalah NLTK (*Natural Language Toolkit*) dan sastrawi. Nltk menyediakan beragam alat dan sumber daya untuk pemrosesan teks dalam bahasa yangberbeda, sementara satrawi adalah sebuah pustaka *Pyhton* yang dikembangkan khusus untuk bahasa Indonesia, termaksud modul untuk stemming (pemotongan akhir kata) dan penanganan *stopwords*. Saat ini, terdapat total 758 kata dalam bahasa Indonesia yang dikategorikan sebagai *stopwords* oleh NLTK.

Oleh karena itu, studi kasus ini bertujuan untuk menganalisis dan membandingkan pengaruh penggunaan *stopwords* sastrawi dengan *stopwords* yang dihasilkan menggunakan algoritma TF-IDF terhadap akurasi analisis sentimen teks. Dengan studi kasus pada berbagai dataset teks, penelitian ini diharapkan dapat menciptakan daftar stopwords baru yang lebih sesuai dan meningkatkan kinerja model analisis sentimen.

B. Rumusan Masalah

Dengan merujuk pada latar belakang yang telah dijelaskan sebelumnya, maka dibentuklah permasalahan sebagai berikut

1. Bagaimana cara menciptakan daftar stopwords baru yang lebih relevan menggunakan algoritma TF-IDF?
2. Apakah penggunaan stopword yang dihasilkan oleh algoritma TD-IDF dapat meningkatkan akurasi analisis sentiment dibandingkan dengan penggunaan stopword sastrawi?

C. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah disajikan, tujuan penelitian ini adalah:

1. Mengembangkan daftar stopwords yang lebih relevan menggunakan algoritma TF-IDF untuk meningkatkan representasi teks dalam analisis sentimen.
2. Mengevaluasi dan membandingkan pengaruh penggunaan stopwords yang dihasilkan oleh algoritma TF-IDF terhadap akurasi model analisis sentimen, dibandingkan dengan penggunaan stopwords Sastrawi.

D. Manfaat Penelitian

Diharapkan pemelitian ini dapat memberikan maanfaat yang lebih luas tidak hanya pada penulis namun juga pada masyarakat. Dengan itu diharapkanpenelitian ini dapat memberikan harapan berupa:

1. Bagi Masyarakat

   Memberikan pandangan yang lebih akurat tentang sentiment masyarakat dengan pemahaman yang lebih baik tentang suatu opini atau pandangan.

2. Bagi Mahasiswa

   a. Memperluas pengetahuan dan pemahaman dalam bidang pemrosesan bahasa alami (NLP) dan analisis sentiment. Penilitian ini memberikan wawasan tentang penerapan metode TF-IDF dalam konteks analisis sentimenteks, serta menciptakan stopword baru.

   b. Memberikan kesempatan bagi mahasiswa untuk mempelajari dan

menerapkan algoritma TF-IDF dalam konteks nyata. Mahasiswa akan memahami cara kerja TF-IDF dalam mengidentifikasi kata-kata yang signifikan dalam dokumen dan bagaimana menggunakannya untuk mengoptimalkan daftar stopwords.

E. Ruang Lingkup Penelitian

1. Penelitian ini akan berfokus pada tempat wisata dimakassar. Wilayah ini dipilih karena dapat memberikan data yang representative untuk analisis sentiment.
2. Penelitian ini akan menggunakan data ulasan yang dikumpulkan melalui platform daring.
3. Penelitian ini akan mempertimbangkan ulasan-ulasan yang mengandung ulasan positif, negative dan netral terhadap tempat wisata yang diulah olehpengunjung.
4. Mengevaluasi penggunakan *stopwords* yang dihasilkan oleh algoritma TF-IDF terhadap akurasi analisi sentiment.

F. Sistematika Penulisan

Untuk memberikan gambaran umum dari seluruh penulisan ini, Adapunsistematika penulisan yaitu:

BAB I PENDAHULUAN

Bab ini mencangkup informasi mengenai latar belakang, rumusan masalah,tujuan, manfaat, batasan masalah, dan sistematika penulisan

BAB II TINJAUAN PUSTAKA

Bab ini mengandung teori-teori yang menjadi dasar dan pendukung penelitian yang dilakukan oleh penulis

BAB III METODE PENELITIAN

Bab ini menjelaskan tentang metode yang akan diterapkan dalam rancanganpembuatan sistem

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisikan uraian hasil penelitian berupa tampilan program yang dihasilkan beserta penjelasanya dan cara menggunakanya.

BAB V KESIMPULAN DAN SARAN

Bab ini berisikan simpulan dan saran-saran dari penelitan yang nantinya dapat digunakan untuk penelitian selanjutnya.

BAB II

LANDASAN TEORI

A. Landasan Teori

1. Natural Language Processing

NLP atau *Natural Language Processing* merupakan salah satu disiplin ilmu dalam bidang kecerdasan buatan yang berfokus pada pemahaman dan pengolahan bahasa manusia oleh komputer. Dengan menggunakan model – model komputasi ini, komunikasi antar manusia dan computer menjadi lebih efisien. Kemampuan ini sangat bermanfaat dalam memfasilitasi interaksi antar pengguna dan sistem computer terutama dalam hal pencarian informasi di internet atau dalam aplikasi lainya (Mulyatun etal., 2021).

NLP adalah bagian dari kecerdasan buatan yang berfokus pada pelatihan komputer agar mampu memahami, memproses, serta menghasilkan bahasa manusia dengan cara yang mirip dengan manusia. Teknologi ini memiliki peran penting dalam mendukung layanan mesin pencari, layanan penerjemah bahasa, serta asisten suara seperti *Siri*, *Alexsa*, atau *Google Home*. Penggunaan NLP semakin meluas dan telah menjadi bagian penting dalam kehidupan sehari – hari dengan menyediakan kemampuan kemampuan komunikasi antara manusia dan komputer yang semakin canggih dan intuitif (Rumaisa et al., 2021)

2. Stopword

Stopword merupakan jenis kata yang sering dianggap tidak memberikan kontribusi signifikan dalam pemrosesan teks dalam bidang pemrosesan bahasa alami. Proses penghilangan *stopword*, yang dikenal sebagai *stopword removal*, adalah suatu teknik yang digunakan untuk mengidentifikasikan dan menghapus kata – kata tersebut dari teks atau

dokumen. Teknik ini melibatkan pencocokan setiap kata dalam teks dengan daftar kata – kata *stopword* yang telah ditentukan sebelumnya. Jika kata – kata tersebut terdaftar dalam *stopwords*, maka kata tersebut akan dihapus dari teks atau dokumen untuk meningkatkan relevansi dan ketepatan analisis teks yang dilakukan (Wibawa et al., 2021).

3. Stopword NLTK Sastrawi

NLTK atau *Natural Language ToolKit* adalah sebuah perpustakaan yang berguna untuk membantu pengguna dlam mengelolah teks. Fasilitas yang disediakan oleh perpustakaan ini mencangkup berbagai fungsi seperti klasifikasi, tokenisasi, stemming, penandaan, penguraian, dan penalaran simantik. Namun salah satu kelemahan NLTK adalah kurangnya dukungan terhadap bahasa Indonesia. Untuk mengatasi hal tersebut, dapat menggunakan perpustakaan tambahan bernama sastrawi, yang khusus dikembangkan untuk pengolahan bahasa Indonesia (Widodod, 2021).

Library NLTK berperan penting dalam tahap tokenisasi, yang merupakan proses penting dalam pengolahan teks yang melibatkan pemisahan kalimat menjadi unit – unit kata individual. Hal ini lakukan agar teks dapat diolah secara terperinci. Di sisi lain library sastrawi memiliki peran khusus dalam tahap penghapusan *stopwords*, dimana kata – kata yang dianggap tidak memiliki relevensi atau makna khusus dalam konteks analisis teks dihilangkan untuk meningkatkan kualitas pemrosesan dan analisis selanjutnya (Duei Putri et al., 2022).

4. Machine Learning

Machine learning adalah disiplin ilmu yang terletak dibawah payung *Artificial Intelligence* (AI), dimana pendekatan ini menggabungkan konsep - konsep dari ilmu komputer dan statistik. Tujuanya adalah untuk mengembangkan model atau algoritm yang

mampu belajar dari data yang ada, sehingga dapat mengidentifikasi dan menangkap pola – pola yang terdapat dalam dataset tersebut. Dengan memanfaatkan teknik – teknik seperti pembelajaran basis data, optimisasi, dan generelasasi, mechine learning bertujuan untuk menciptakan sistem – sistem yang mampu beradaptasi dan membuat prediksi atau keputusan tanpa adanya intruksi langsung dari manusia (Giarsyani, 2020).

Mechine Learning merupakan bagian dari *artificial intelligence* atau kecerdasan buatan yang difokuskan pada pengembangan sistem yang mampu belajar sendiri tanpa intervensi manusia. Dengan konsep ini, *mechine lerning* memiliki kapabilitas untuk menyerap informasi dan pola yang terkandung dalam *database* yang diberika (Fahrizal et al., 2020). Dimana proses kerja *mechine lerning* melibatkan beberapa tahap, di antaranya adalah tahap pemilihan data yeng terdiri dari pemisahan dataset menjadi tiga bagian yang berbeda, yakni data yang akan digunakan sebagai pelatiahan (*training data*) untuk mengajar model, data validasi (*validation data*) untuk mengevaluasi dan menyempurnakan model, serta data uji (*testdata*) untuk menguji kinerja dan kemampuan prediktif model yang telah dikembangkan (Zailani et al., 2020).

5. Scikit Learn

Scikit learn adalah sebuah perpustakaan *open source* untuk analisis data yang telah dilakukan sebagai standar emas dalam dunia *mechine learning* dalam lingkup python. Perpustakaan ini menawarkan berbagai konsep dan fitur penting yang menjadi tulang punggung dalam pengembangan model *mechine learning* seperti metode pengambilan keputusan diantaranya:

a. Kalifikasi yang memungkinkan untuk mengidentifikasikan dan

mengktegorikan data berdasarkan pola yang ada.

b. Regresi yang memungkinkan untuk memprediksi dan memproyeksikan nilai – nilai data dengan menggunakan rata – rata daridata yang sudah ada dan data yang direncanakan
c. Pengelompokan yang memfasilitasi pengelompokan otomatis data yangserupa kedalam data set yang sesuai

Algoritma – algoritma yang mendukung analisis prediksi dapat bervariasi dari rgresi linear sederhana hingga pengenalan pola menggunakan jaringan saraf, dan matplotlib juga harus diperhatikan dalam mengembangkan solusianalisis data yang efektif (M, 2021)

6. TF – IDF

TF-IDF merupakan metode yang digunakan untuk mengintegrasikan dua teori berdasarkan keberadaan kata dalam dokumen. Dalam pendekatan ini, setiap kata memiliki satu atau lebih kata kunci dalam teks. Frekuensi kemunculan kata – kata tersebut dalam suatu dokumen menandakan tingkat signifikasi tertentu, sementara frekuensi kemunculan kata – kata tersebut di seluruh dokumen menggambarkan popularitas secara umum, dengan itu rumus Tf-IDF adalah (Syaifuddin & Ningsih, 2023) :

$$tf = 0.5 + 0.5x\frac{tf}{\max tf} \qquad (1)$$

$$i\,dft\ = log\left(\frac{d}{dft}\right) \qquad (2)$$

$$w_{d,t} = tf_{d,t} \text{ x } IDF_{d,t} \qquad (3)$$

Keterangan:

tf = jumlah kata yang dicari pada sebuah dokumen

D = total dokumen

df = jumlah dokumen yang berisi term t

idf = inversed dokumen frekuensi (log2)

d = dokumen ke-d

t = kata ke-t dari kata kunci

W = bobot dokumen ke-d terhadap kata ke-t

TF-IDF juga dikenal sebagai salah satu fitur pembobotan paling populer yang sering diterapkan dalam berbagai kontek, karena tingkat recall dan akurasi yang cukup tinggi. Dimana TF-IDF memperhitungkan frekuensi kemunculan kata dalam sebuah dokumen. Artinya semakin sering kata tersebut muncul dalam dokumen tertentu, semakin besar kontribusinya terhadap pembobotan, tetapi jika kata tersebut sering muncul dibanyak dokumen, maka kontribusinya akan lebih kecil. Pendekatan ini terdiri dari dua konsep utama, yaitu *Term Frequency* (TF) yang menghitung seberapa sering kata tersebut muncul dalam dokumen tertentu, dan *Inverse Document Frequency* (IDF) yang menggukur seberapa umum sebuah kata di seluruh koleksi dokumen (Yutika et al., 2021).

7. Naïve Bayes

Naïve Bayes adalah metode klasifikasi yang paling dasar dan

paling sering digunakan. Model klasifikasi ini mengestimasikan probabilitas posterior suatu kelas berdasarkan distribusi kata suatu dokumen, dengan menggunakan representasi dokumen yang sederhana sebagai *Bag of Words.* Pendekatan ini tidak memperhatikan posisi kata dalam dokumen, melainkan hanya mengekstraksi fitur *Bag of Words*. Metode ini memanfaatkan Teorema Bayes untuk memprediksi probabilita bahwa sekumpulan fitur yang diberikan akan terkait dengan label tertentu (Yulita,2021).

Naïve Bayes memiliki kemampuan untuk memproyeksikan probabilitas keanggotaan dalam kelas tertentu, seperti probabilitas inklusi dalam kelas yang spesifik, yang telah berguna dalam mengklasifikasikan data yang berasal dari twitter menjadi kelas positif, negative, dan juga netral. Metode ini didasarkan pada Teorema Bayes menggunakan pengalaman masa lalu untuk memproyeksikan kemungkinan kejadian dimasa depan. Salah satu karakteristik utama dari *Naïve Bayes* adalah asumsi yang kuat terhadap idenpendensi antara kondisi atau kejadian (Mas Pintoko & Muslim, 2018).

B. Penelitian Terkait

1. Okta Ihzan Gifari, Muh Adha, Ivan Rifky Hendrawan, Fernando Freddy Setlight Durrand 2022

   Pada tahun 2022, tim peneliti Okta Ihzan Gifari, Muh Adhan, Ivan Rifky Hendrawan, Fernando Setlight Durran melakukan penelitian dengan Judul "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vectory Machine" mereka menggumpulkan 200 ulasan film dari Twitter menggunakan crawling dan menyimpanya dengan format csv. Setelah data terkumpul, tim peneliti menerapkan metode TF-IDF untuk mengelolah ulasan film tersebut. Selanjutnya, mereka menggunakan SVM untuk mengklasifikasikan sentiment dari ulasan – ulasan tersebut.

Hasilnya menunjukan bahwa pengklasifikasian sentiment ulasan film berbahasa inidonesia dapat dilakukan dengan cukup baik dengan menggunakan metode TF-IDF dan algoritma SVM. Hal ini dapat dibuktikan dengan hasil penelitian yang menunjukan nilai akurasi sebesar 0.85, nilai presisi 1.0, nilai recall 0.7, dan nilai F1-Score 0.82. Ini menunjukan potensi Tf-IDF dan SVM dalam analisis sentiment film.

2. Abdul Azis dan Fauziah 2022

Pada tahun 2022, Abdul Azis dan Fauziah melakukan penelitian dengan judul "Analisis Sentiment Identifikasi Opini Terhadap Produk, Layanan, dan Kebijakan Perusahaan Menggunakan Algorithma TF-IDF dan SentiStrength" yang bertujuan untuk melakukan analisis sentiment tanpa perlu proses labeling manual pada dataset yang digunakan. Metode penelitian yang digunakan adalah TF-IDF dan SentiStrenght. Proses pengambilan data dilakukan dengan crawling data melalui API Twitter untuk mengumpulkan data tweet. Hasil penelitian menunjukan bahwa dengan menggunakan sistem yang dibuat menggunakan algoritma TF-IDF dan SentiStrenght, didapatkan hasil sentiment positif sebesar 54%, negative 20%, dan netral 26%. Kemudian dilakukan perbandingan menggunakan Rapid Miner yang menggunakan algoritma Naïve Bayes dengan hasil sentiment positif 55%, negative 16%, metral 29% dan Decision tree dengan hasil sentiment positif sebesar 61%, negative 14%, dan netral 25%. Meskipun menggunakan algoritma yang berbeda, namun hasilnya menunjukan konsistensi dimana sentiment yang paling dominan adalah sentiment positi, disusul sentiment netral, dan yang paling rendah negative. Dengan demikian sistem yang dibuat menggunakan Alogitma TF-IDF dan SantiStrenght berhasil menjalankan fungsinya pada analisis analisis sentiment tanpa perlu melalukan tahap labeling manual seperti yang dilakukan pada algoritma Naïve Bayes dan Decision tree.

3. Dily Wardhani, Rika Astuti, dan dedi Dwi Saputra

   Pada tahun 2024, Diky Wardhani, Rika Astuti, dan Dedi Dwi Saputra melakukan penelitian dengan judul "Optimasi Featur Selection Text Mining: Stemming dan Stopword Untuk Sentimrn Analisis Aplikasi SatuSehat", tujuan penelitian ini dilakukan adalah membandingkan dan menggoptimalkan Featur Selection pada Text Mining untuk analisis sentiment aplikasi SatuSehat. Dalam penelitian ini metode yang digunakan adalah Naïve Bayes dan menggunakan 2000 data ulasan aplikasi yang diperoleh dari playstore dengan menggunakan metode Scrapping. Hasil penelitian menunjukan bahwa saat menggunakan feature selection stemming, akurasi yang diperoleh adalah 93,43% dengan presisi sebesar 88,42%. Sedangkan ssat menggunakan feature selection stopword, akurasi yang diperoleh adalah 89,19% dengan presisi sebesar 82,23%, jika menggunakan kedua feature secara bersamaan, akurasi yang diperoleh adalah 92,56% dengan presisi sebesar 95,46%. Dengan hal tersebut disimpulkan bahwa hasil terbesar diperoleh ketika menggunakan fitur stemming namun, hasil yang paling optimah untuk dataset adalh saat menggunakan keduanya.

4. Syahrih Dwi Prasetyo, Shofa Shofiah Hilabi,dan Fitri Nurapriani

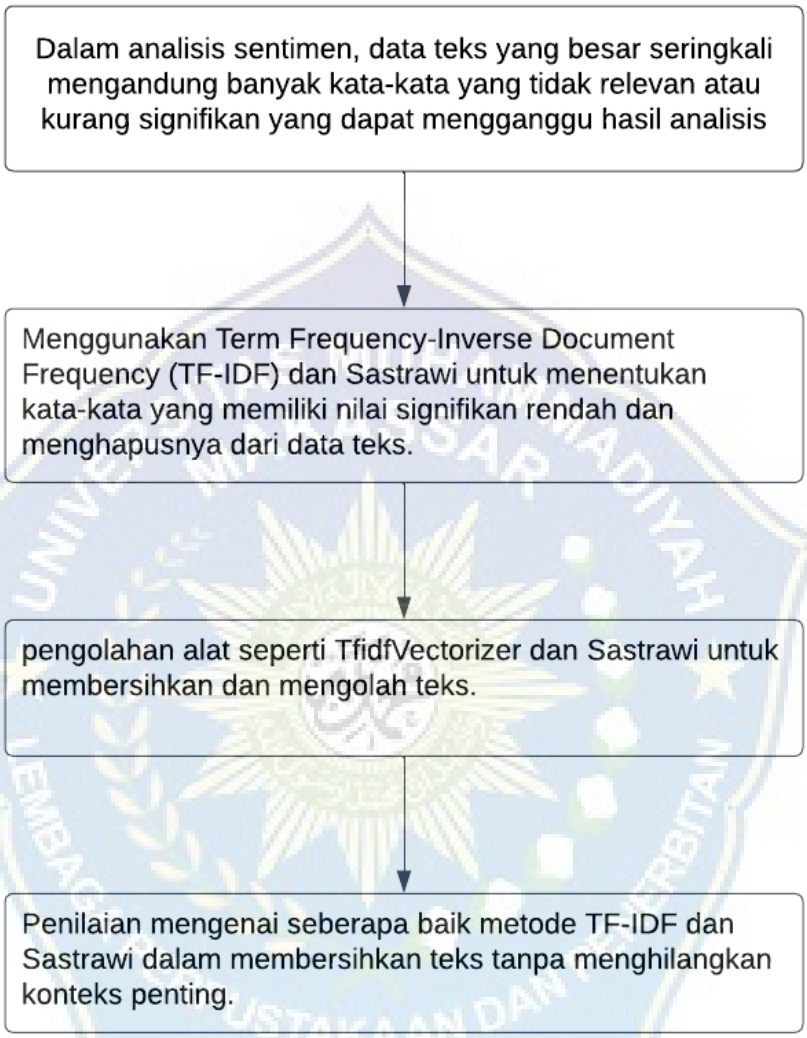
   Pada tahun 2023, Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, dan Fitri Nurapriani melakukan penelitian yang berjudul "Analisis Sentimen Relokasi Ibu Kota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN". Penlitian ini bertujuan untuk menyajikan hasil analisis yang tepat dan akurasi yang terkait dengan sentiment masyarakat terhadap pemindahan Ibu Kota Indonesia. Metode penelitian yang digunakan adalah Naïve Bayes dan KNN dengan proses pengambilan data melalui layanan microblogging Twitter yang disediakan oleh API Twitter. Hasil penelitian menunjukan bahwa analisis menggunakan metode Naïve Bayes memiliki tingkat Akurasi sebesar 82.27%, dengan nilai presisi sebesar

86.36% dan nilai recall sebesar 76,93%, sedangkan untuk metode KNN, tingkat akurasi yang diperoleh adalah 88,12%, dengan nilai presis sebesar 93,98% dan recall sebesar 81,53%. Berdasarkan nilai – nilai tersebut dapat disimpulkan bahwa metode LNN memiliki tingkat akurasi yang lebih unggul dibandingkan metode Naïve Bayes.

5. Ina Najiah dan Ifani Haryanto

   Pada tahun 2021, Ina Najiah dan Ifani Haryanto melokukan penelutian dengan judul "Sentimen Analisis COVID-19 dengan Mtode Probabilistic Neural Network dan TF-IDF. Tujuan penelitian ini adalah melakukan analisis sentiment dan mengklasifikasikan opini mengenai covid-19 menjadi 3 kelas yakni, positif, netral, dan negatif. Metode yang digunakan adlah TF-IDF dan NN dengan menggunakan data yang dieroleh dari Twitter, Facebook, dan Instagram yang berisi tema covid-19. Hasil penelitian ini menunjukan bahwa penelitian ini menghasilkan akurasi sebesar 86%, yang merupakan peningkatan dari penelitian sebelumnya yang memiliki akurasi 76%. Peningkatan ini disebabkan oleh jumlah dataset yang lebih besar dan proses pra-pemrosesan data yang lebih lengkap.

C. Kerangka Pikir



Gambar 1 Kerangka Pikir

BAB III

METODE PENELITIAN

A. Tempat dan Waktu Penelitian

1. Tempat Penelitian

   Penelitian ini dilakukan secara daring dengan menggumpulkan sejumlah ulasan mengenai wisata dari berbagai platform media sosial

2. Waktu Penelitian

   Penelitian ini akan berlangsung dari bulan Maret sampai Mei 2024

B. Alat dan Bahan

1. Kebutuhan Hardware (perangkat keras)
   a. Laptop Lenovo
2. Kebutuhan Software (perangkat lunak)
   a. Colab Google
   b. Excel
   c. Python
   d. Scikit-learn

C. Perancangan Sistem

Perancangan sistem memegang peranan krusial dalam pengembangan suatu sistem karena merincikan langkah – langkah dari perancangan hingga implementasi fungsi – fungsi yang diperlukan untuk menjalankan sistem tersebut. Tujuan utama dari perancangan sistem adalah untuk menjamin bahwa sistem yang dikembangkan mampu memberikan hasil yang diharapkan.

1. Flowchart Penelitian

Dalam penelitian ini, penulis akan melakukan studi litelatur yang komperhensif tentang analisis sentiment, penggunaan TF-IDF, dan penggaruh penghapusan stopword terhadap analisis sentiment. Selanjutnyadalam penelitian ini, penulis akan menggumpulkan data teks yang relevan dari berbagai sumber social media. Data ini mencangkup ulasan sentiment positif, negative, dan juga netral terhadap tempat wisata.

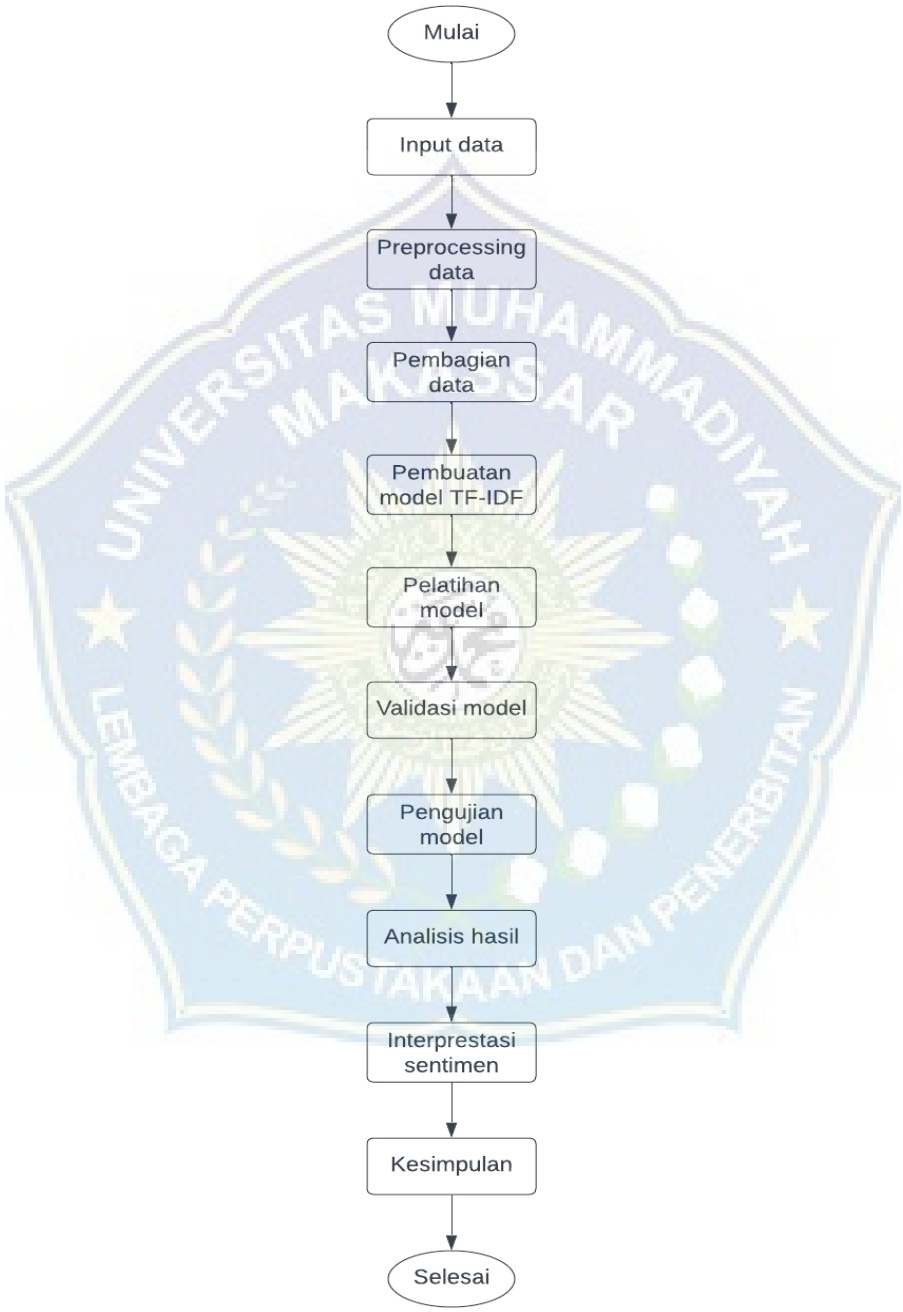
Setelah data terkumpul, penulis akan melakukan langkah – langkah pengolahan data. Ini termaksud perbersihan teks, seperti tokenisasi dan penghapusan karakter khusus, serta preprocessing seperti penghapusan stopword dan stemming. Penulis akan mempersiapkan data set yang bersihdan siap untuk digunakan dalam proses pengujian.

Penulis akan merancang sistem yang memanfaatkan algoritma TF-IDF untuk memiliki teks numerik. Sistem akan memiliki modul untuk menghapus stopword serta mekanisme untuk melatih model analisis sentiment. Penulis akan menggunakan teknik klasifikasi yang sesuai, seperti Naïve Bayes untuk melatih model. Sistem ini juga akan mencangkup evaluasi model untuk menggukur akurasi.

istem akan diuji menggunakan dataset yang telah disiapkan sebelumnya kemudian data set tersebut dibagi penjadi data latih dan data uji, dan melatih model dengan berbagai konfigurasi stopword. Selanjutnyapenulis akan mengevaluasi model menggunakan data uji dan membandingkan hasilnya untuk melihat pengaruh penghapusan stopword terhadap akurasi analis sentiment.

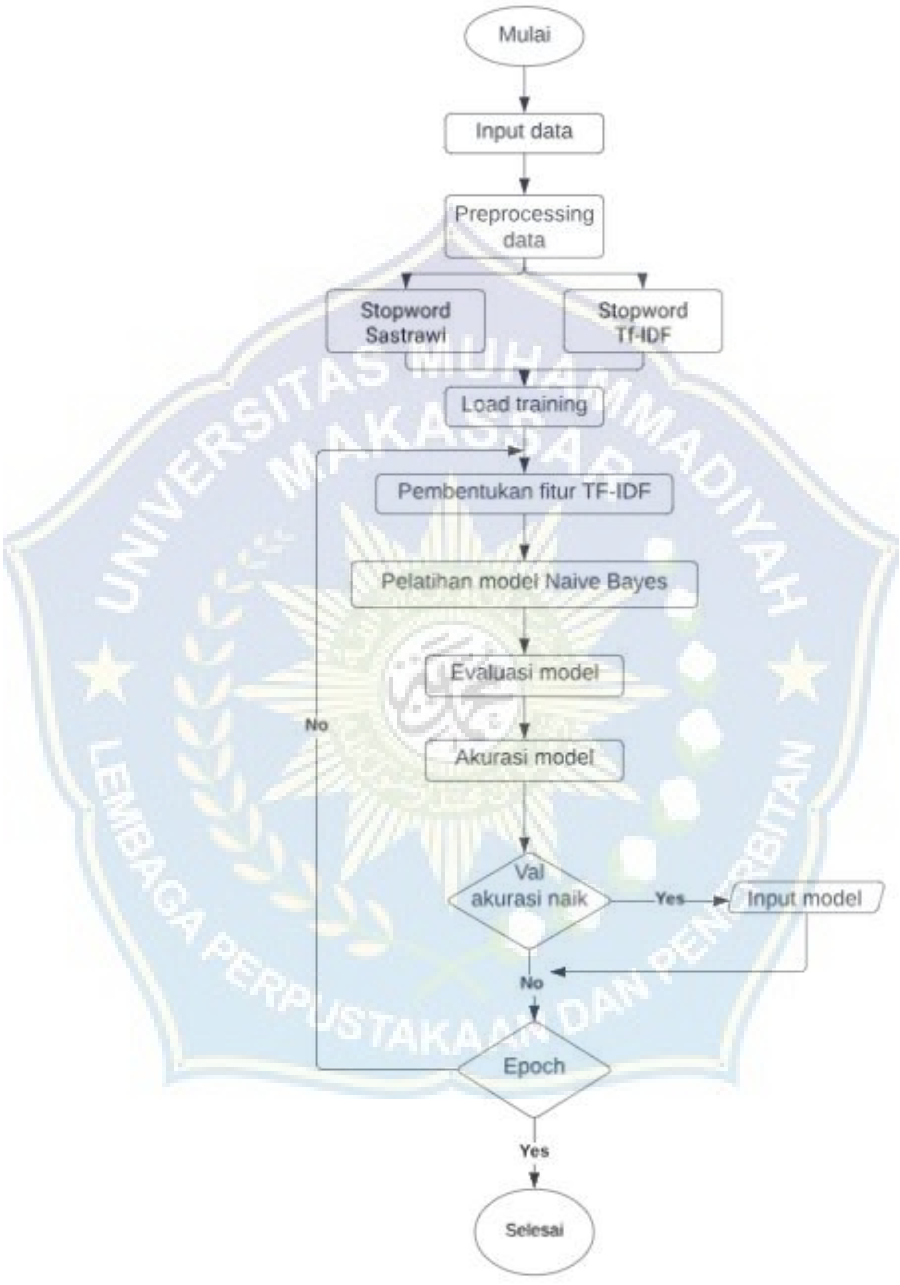
Akhirnya, penulis akan menganalisis hasil pengujian untuk menarik kesimpulan tentang pengaruh stopword terhadap analisis sentiment teks menggunakan TF-IDF, penulis akan membahas temuan ini dan implikasinya terhadap praktik analisis sentiment di masa depan di masa depan. Seperti yang digambarkan pada gambr 2 dibawah ini.

1. Flowchart penelitian



Gambar 2 Diagram proses penelitizn

2. Flowchart sistem



Gambar 3 Diagram Sistem

Seperti pada gambar 3 diatas penelitian ini dimulai dengan preprocessing pada dataset dengan melakukan normalisasi teks, selanjutnya data dibagi menjadi kelompok stopword sastrawi dan stopword TF-IDF. Setelah preprosesing selesai, penulis melakukanektraksi fitur dengan menggunakan TF-IDF untuk mengubah teks menjadi vector numerik.

Dilanjutkan dengan melatih model dengan klasifikasi Naïve Bayes padakedua kelompok data. kemudian hasil klasifikasi dievaluasi dengan perbandingan akurasi, presisi, recall, dan F1-score dari kedua model. Penulis membandingkan apakah penggunaan stopword dapat meningkatkan atau mengurangi pengaruhi hasil analisis sentiment. Dari hasil tersebut penulis dapat menarik kesimpulan terhadap pengaruh dalam penggunaan stopword.

D. Teknik Pengujian Sistem

Pengujian dimulai dengan memuat dataset yang terdiri dari berbagai ulasan produk yang telah dilabeli dengan sentimen positif, negatif, atau netral.Setelah dataset dimuat, langkah selanjutnya adalah melatih model TF-IDF menggunakan data pelatihan yang telah disiapkan. Proses pelatihan dilakukan dengan membagi dataset menjadi bagian pelatihan dan bagian validasi untuk memonitor performa model selama pelatihan.

Setelah model dilatih, langkah berikutnya adalah melakukan pengujian menggunakan dataset pengujian. Model akan menerima input ulasan produk dan menghasilkan prediksi sentimen. Pengujian ini akan mengukur akurasi, presisi, recall, dan F1-score dari sistem analisis sentimen yang dikembangkan.

Hasil dari skenario pengujian ini akan digunakan untuk mengevaluasi performa pengaruh penggunaan stopword terhadap analisis

sentimen menggunakan TF-IDF dan menentukan apakah model tersebut siap untuk diimplementasikan dalam lingkungan produksi atau memerlukan penyesuaian dan perbaikan lebih lanjut.

E. Te**knik Analisis Data**

Proses analisis data adalah serangkaian langkah sistem untuk menyusun dan memperoleh makna dari informasi yang dikumpulkan melalui wawancara, observasi, atau sumber data digital. Langkah – langkah tersebut meliputi pengorganisasian data ke dalam kategori, pembagian menjadi unit – unit yang lebih kecil, sintesis, pembentukan pola, pemilihan informasi yang relevan, serta penarikan kesimpulan. Tujuan dari analisis data adalah untuk mempermudah pemahaman informasi, baik oleh peneliti maupun oleh pohak lain yang membaca hasil analisis. Untuk mencapai tujuan tersebut, peneliti menjalankan serangkaian tahapan analisis yang terdiri dari:

1. Pengumpulan Data

   Proses pengumpulan data merupakan suatu data yang diperoleh dari berbagai sumber yang relevan dengan penelitian, seperti survei, wawancara,atau data digital.

2. Preprocessing

   Setelah itu, dilakukan tehapan preprocessing data, dimana disesuaikan agarsiap untuk diperoses. Hal ini melibatkan langkah – langkah penting sepertipembersihan data, tokenisasi, stemming, dan penghapusan kata – kata stopwords.

3. Display Data

Penyajian data yang telah direduksi secara teestruktur dan sistematis oleh peneliti. Tujuan dari proses ini adalah untuk memudahkan pemahaman informasi dalam data.

5. Kesimpulan

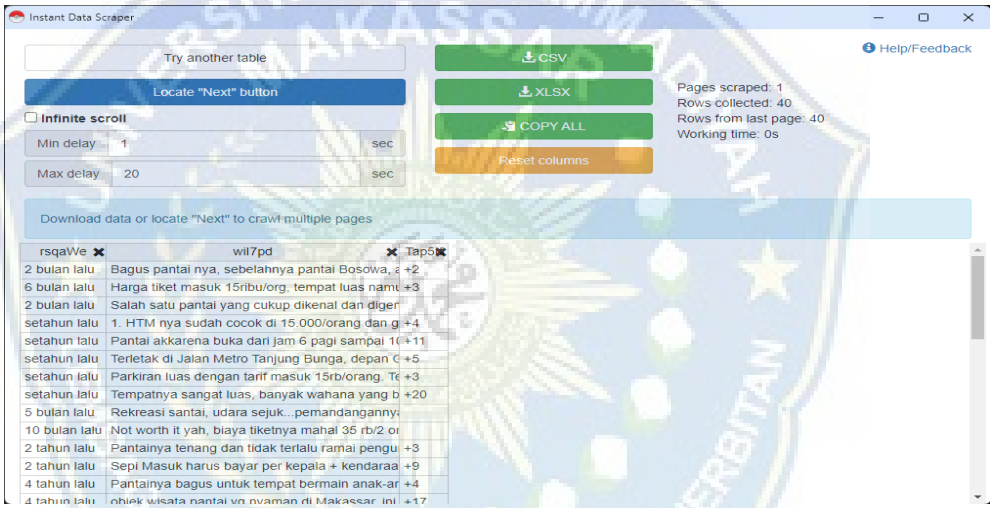
Pada tahap ini, peneliti menyajikan kesimpulan berdasarkan data yang diperoleh, menjadikan peneliti sebagai usaha untuk memberikan jawaban terhadap permasalahan yang dihadapi

BAB IV

HASIL DAN PEMBAHASAN

A. Pengambilan Data

Penelitian ini menggunakan data ulasan dari Google Maps yang berisi pendapat pengguna mengenai pemandangan, fasilitas, pelayanan, keramaian, dan kebersihan suatu tempat. Data tersebut diakses dalam bentuk file berekstensi .csv dengan total 4.500 data ulasan, yang masing-masing terbagi menjadi 1.500 data untuk tiga label sentimen: positif, negatif, dan netral.



Gambar 4 Screping Data

Setelah mengumpulkan data ulasan dari Google Maps mengenai tempat wisata dengan menggunakan data instan, peneliti menyimpan hasilnya dalam format Excel. Berikut adalah hasil analisis ulasan yang telah kami simpan dalam file Excel, mencakup berbagai aspek seperti pelayanan, fasilitas, dan pengalaman keseluruhan di tempat wisata tersebut.

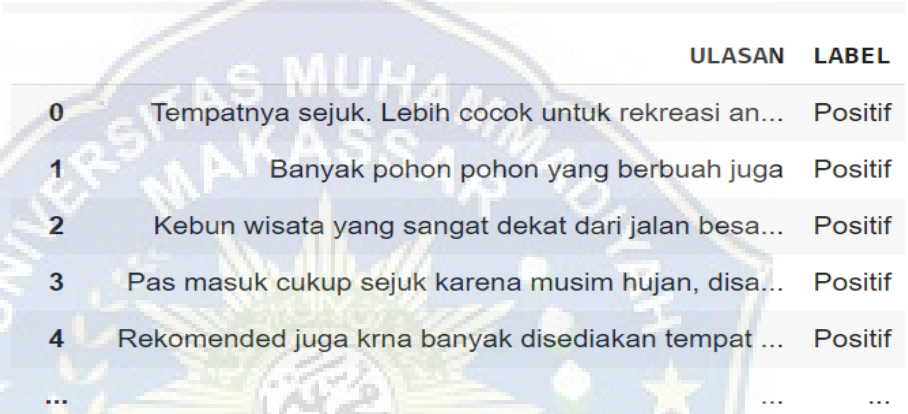
Table 1 Data Ulasan

Tempatnya sejuk. Lebih cocok untuk rekreasi anak anak. Tiket weekdays 15.000 / orang. Tempat duduk duduk banyak. Ada pemancingan, wahana bermain anak, sewa sepeda listrik atau ATV, juga kolam renang	Positif
Banyak pohon pohon yang berbuah juga	Positif
Kebun wisata yang sangat dekat dari jalan besar sehingga mudah dijangkau berbagai jenis kendaraan, dengan perjalanan sekitar 25-35 menit dari kota makassar. Tiket pada hari biasa 15 k, dan pada Sabtu-Minggu 25 k.	Positif
Pas masuk cukup sejuk karena musim hujan, disambut kolam dan penginapan. Ada 3 kolam besar yang terbagi dengan sekat, kolam dalam, sedang dan untuk anak-anak. Di lokasi ada tempat beli cemilan makanan ringan dan sewa alat renang.	Positif
Rekomended juga krna banyak disediakan tempat duduk dan berteduh di pinggiran kolam. Di sini juga disediakan aula. Sangat cocok buat liburan bareng keluarga	Positif
Karena sekolah anak sy adakan outbond di wisata kebun jd setelah sekian purnama akhirnya bisa lg ke wisata kebun rame² bareng dgn orgtua siswa yg lainnya,,, pas masuk ke area parkir kesan pertama ya , area parkirnya luas , trs ke bagian loket u/ beli karcis masuk , karyawannya ramah bnget .	Positif
Begitu udah masuk ke dlm area bener² kereeenn banget , semua tempat dlm area wiskeb cucok banget buat foto² , pokoknya memori hp bakalan full	Positif
Anak sy betah dgn kegiatan oubondnya setelah outbond berenang deh,,	Positif
Di wisata kebun itu harga tiket terjangkau, fasilitas oke , area bermain anak banyak , kantin jg ada , kolam pancing ada , gasebo² gratis bersih juga	Positif
Tiket masuknya perorang Rp.15.000, ada kebun durian, rambutan. Yang mau berenang juga disediakan 4 kolam renang untuk anak-anak & dewasa, ada penginapan, kantin, tempat mancing, banyak spot foto-foto yang bagus juga. Kemarin nyobain kereta keliling wisata kebun 2x hanya Rp.5.000	Positif

B. Pelabelan Data

Proses labeling dilakukan guna untuk menentukan sentimen masing-masing ulasan. Terdapat beberapa teknik pelabelan yang umum digunakan seperti Transformer, TextBlob, dan VADER yang mampu mengklasifikasikan

sentimen secara otomatis. Namun, untuk meningkatkan akurasi dan memastikan keakuratan hasil, proses labeling dilakukan secara manual oleh tim ahli yang terlatih dalam analisis sentimen. Metode manual ini memungkinkan penilaian yang lebih kontekstual dan tepat terhadap nuansa sentimen dalam ulasan, sehingga menghasilkan dataset yang lebih andal untuk analisis lebih lanjut.



	ULASAN	LABEL
0	Tempatnya sejuk. Lebih cocok untuk rekreasi an...	Positif
1	Banyak pohon pohon yang berbuah juga	Positif
2	Kebun wisata yang sangat dekat dari jalan besa...	Positif
3	Pas masuk cukup sejuk karena musim hujan, disa...	Positif
4	Rekomended juga krna banyak disediakan tempat ...	Positif
...	...	...

Gambar 5 Pelabelan Data

Data ulasan dibagi menjadi tiga kategori sentimen, masing-masing terdiri dari 1.500 ulasan, kemudian dibagi lagi menjadi data latih dan data uji menggunakan library dari scikit-learn untuk memastikan pemisahan yang konsisten dan acak. Proporsi pembagian ditetapkan dengan 90% digunakan sebagai data latih, sementara 10% sisanya digunakan sebagai data uji. Pendekatan ini memastikan bahwa model dapat diujikan secara efektif dengan data yang tidak terlihat selama proses pelatihan, sehingga memberikan gambaran akurat tentang kemampuan generalisasi model.

C. **Data Preprocessing**

Data yang telah dipilih kemudian dipraproses untuk memastikan kualitas dan konsistensi. Proses praproses ini terdiri dari beberapa tahap

penting yang bertujuan untuk membersihkan dan menyiapkan data sebelum digunakan dalam analisis sentimen. Tahapan praproses meliputi:

1. Pembersihan Data

   Tahap pembersihan data melibatkan penghapusan tanda baca seperti koma, titik, tanda tanya, tanda seru, bintang, dan pagar dari teks atau data. Langkah ini penting karena karakter-karakter tersebut biasanya tidak memberikan kontribusi signifikan dalam analisis data atau pemrosesan teks. Sebagai contoh, dalam analisis teks, tanda baca sering kali dianggap sebagai noise yang tidak relevan dan dihilangkan untuk memusatkan perhatian pada informasi utama yang terkandung dalam teks.

Table 2 Pembersihan Data

Sebelum	**Sesudah**
Tempat wisata yang menyenangkan…	Tempat wisata yang menyenangkan
Kualitas air terlihat kotor dan tidak sebanding dengan harga tiket.	Kualitas air terlihat kotor dan tidak sebanding dengan harga tiket
BUGIS WATERPARK tempat rekreasi yang juga sekaligus memperkenalkan baik Bahasa Bugis dan budaya Bugis	BUGIS WATERPARK tempat rekreasi yang juga sekaligus memperkenalkan baik Bahasa Bugis dan budaya Bugis

2. Pengisian Nilai kosong

   Program `df['ULASAN'].fillna('', inplace=True)` digunakan untuk mengisi nilai-nilai kosong (NaN) di dalam kolom 'ULASAN' pada DataFrame `df` dengan string kosong (''). Penggunaan `inplace=True` menandakan bahwa perubahan akan diterapkan secara langsung pada DataFrame `df`, tanpa perlu menyimpan hasilnya ke variabel baru. Dengan melakukan pengisian

nilai kosong ini, program memastikan bahwa data yang akan diproses selanjutnya tidak mengalami gangguan atau kesalahan karena keberadaan nilai-nilai yang kosong.

```
df['ULASAN'].fillna('', inplace=True)
```

3. Penghapusan Tanda Baca

```
def remove_punctuation(text):
            translator     =     str.maketrans('',     '',
string.punctuation)
    return text.translate(translator)
df['ULASAN_BERSIH']                                    =
df['ULASAN'].apply(remove_punctuation)
```

Program yang diberikan bertujuan untuk membersihkan teks dari tanda baca pada sebuah DataFrame. Proses dimulai dengan definisi sebuah fungsi bernama `remove_punctuation` yang menggunakan Python untuk menghapus semua tanda baca dari teks yang diberikan. Fungsi ini menggunakan metode `str.maketrans('', '', string.punctuation)` untuk membuat sebuah translator yang mengatur penghapusan tanda baca. Parameter `text` yang diterima oleh fungsi ini akan diubah dengan menggunakan metode `translate(translator)` untuk menghilangkan semua tanda baca yang ada dalam teks.

Setelah fungsi `remove_punctuation` didefinisikan, langkah berikutnya adalah menerapkannya pada DataFrame `df`. Dalam konteks ini, kolom `'ULASAN'` dari DataFrame tersebut digunakan sebagai input untuk fungsi `remove_punctuation`. Hasil dari pemrosesan ini disimpan kembali dalam kolom baru yang disebut `'ULASAN_BERSIH'`.

4. Penghapusan stopword

```
factory = StopWordRemoverFactory()
stopword = factory.create_stop_word_remover()
def remove_stopwords(text):
    return stopword.remove(text)
```

```
df['ULASAN_BERSIH'] =
df['ULASAN'].apply(remove_stopwords)
```

Program ini dirancang untuk membersihkan teks yang terdapat dalam kolom 'ULASAN' dari stopwords. Stopwords adalah kata-kata umum yang sering tidak memiliki makna khusus dalam analisis teks, seperti "dan", "atau", "yang", dan sebagainya. Proses dimulai dengan mengimpor sebuah library atau modul yang menyediakan fasilitas untuk menghapus stopwords, yang terlihat dari penggunaan `StopWordRemoverFactory`.

Setelah library diimpor, program melakukan inisialisasi `StopWordRemoverFactory()` untuk membuat sebuah factory yang akan digunakan untuk menciptakan objek remover stopwords. Langkah selanjutnya adalah membuat objek remover stopwords itu sendiri dengan menggunakan factory yang sudah dibuat, dan objek ini disimpan dalam variabel `stopword`.

Untuk mengimplementasikan penghapusan stopwords pada teks, sebuah fungsi bernama `remove_stopwords` didefinisikan. Fungsi ini menerima satu parameter `text`, yang merupakan teks yang akan dibersihkan dari stopwords. Di dalam fungsi, teks `text` diproses dengan memanggil metode `.remove()` dari objek `stopword`, yang mengembalikan teks tanpa stopwords.

Terakhir, fungsi `remove_stopwords` diterapkan ke setiap nilai dalam kolom 'ULASAN' dari DataFrame (`df`) menggunakan metode `.apply()`. Hasilnya disimpan dalam kolom baru 'ULASAN_BERSIH', yang akan berisi teks dari kolom 'ULASAN' yang telah dibersihkan dari stopwords.

D. Pembangunan Model

1. Tranformasi Teks dengan TF-IDF

```
vectorizer = TfidfVectorizer()
x = vectorizer.fit_transform(df['ULASAN_BERSIH'])
y = df['LABEL']
```

Program ini menggunakan TfidfVectorizer dari library scikit-learn untuk melakukan ekstraksi fitur dari teks dalam kolom 'ULASAN_BERSIH' pada sebuah DataFrame. Kemudian, objek TfidfVectorizer diinisialisasi tanpa parameter tambahan.

TfidfVectorizer ini digunakan untuk mengonversi kumpulan dokumen teks menjadi matriks fitur TF-IDF, yang merupakan representasi numerik dari teks berdasarkan frekuensi kemunculan kata dan seberapa penting kata tersebut dalam dokumen.

Selanjutnya, transformasi data dilakukan dengan memanggil metode fit_transform pada objek vectorizer, menggunakan data dari kolom 'ULASAN_BERSIH' dalam DataFrame df. Metode fit_transform melakukan dua fungsi utama:

- Fit: Membangun kamus internal dari kata-kata yang ditemukan dalam 'ULASAN_BERSIH' dan menghitung frekuensi kemunculan setiap kata.
- Transform: Mengubah setiap dokumen (ulasan) dalam 'ULASAN_BERSIH' menjadi vektor numerik berdasarkan skema TF-IDF.

Hasil dari transformasi ini disimpan dalam variabel x, yang berisi matriks fitur TF-IDF. Setiap baris dalam matriks ini mewakili satu dokumen (ulasan), dan setiap kolom mewakili kata-kata unik yang ditemukan dalam semua ulasan. Nilai dalam setiap sel menunjukkan bobot TF-IDF dari kata tersebut dalam dokumen yang sesuai.

```
tfidf_scores = np.asarray(x.mean(axis=0)).flatten()
```

Setelah matriks TF-IDF terbentuk, kemudian menghitung rata-rata

nilai TF-IDF dari setiap kata di seluruh dokumen dalam dataset. Ini dilakukan dengan memanggil x.mean(axis=0), yang mengonversi hasil perhitungan ini menjadi array satu dimensi. Nilai rata-rata TF-IDF ini dapat digunakan untuk mengidentifikasi kata-kata yang kurang signifikan dalam keseluruhan dokumen dan mungkin dihapus untuk meningkatkan kualitas data yang akan dianalisis lebih lanjut.

```
threshold = 0.003

indikasi_Kata_dibawah_threshold = np.where(tfidf_scores < threshold)[0]

kata_dibawah_Threshold = [vectorizer.get_feature_names_out()[idx] for idx in indikasi_Kata_dibawah_threshold]

print("Kata-kata dengan nilai TF-IDF di bawah ambang batas:")
print(kata_dibawah_Threshold)
```

Penetapan ambang batas nilai TF-IDF sebesar 0.003 untuk mengidentifikasi dan menghapus kata-kata yang kurang signifikan. Dengan menghitung nilai rata-rata TF-IDF untuk setiap kata, kemudian menemukan indeks kata-kata yang nilai TF-IDF-nya di bawah ambang batas ini menggunakan fungsi 'np.where'. Kata-kata tersebut kemudian diambil dari daftar fitur yang dihasilkan oleh *TfidfVectorizer* dan dicetak ke layar untuk mengidentifikasi mana saja yang tidak signifikan. Proses ini membantu menyaring teks sehingga hanya kata-kata yang memberikan kontribusi penting yang dipertahankan.

2. Pembagian Dataset

```
x_train, x_test, y_train, y_test, train_indices, test_indices = train_test_split(x, y, df.index, random_state=0, test_size=0.1)
```

Proses pembagian dataset menjadi data pelatihan dan data pengujian merupakan langkah krusial dalam pengembangan model pembelajaran mesin. Untuk mencapai tujuan ini, fungsi *train_test_split* dari pustaka *scikit-learn* digunakan. Pada baris kode yang diberikan, dataset lengkap yang terdiri dari ulasan (x) dan label (y) dibagi menjadi beberapa bagian dengan cara yang terstruktur.

Pada langkah awal, data ulasan dan label masing-masing dibagi menjadi dua subset. satu subset untuk melatih model (x_train dan y_train) dan subset lainnya untuk menguji model (x_test dan y_test). Selain itu, indeks dari dataframe asli (df) juga dibagi menjadi dua bagian yaitu train_indices yang berisi indeks untuk data pelatihan, dan test_indices yang berisi indeks untuk data pengujian.

Pembagian ini dilakukan dengan parameter random_state yang ditetapkan ke nilai 0, yang memastikan bahwa proses pengacakan dalam pembagian data akan menghasilkan hasil yang konsisten setiap kali kode dijalankan. Selain itu, parameter *test_size* diatur ke 0.1, yang berarti bahwa 10% dari data lengkap akan dialokasikan untuk data pengujian, sementara sisanya 90% digunakan untuk pelatihan.

3. Pelatihan Model

```
model = MultinomialNB()
model.fit(x_train, y_train)
```

Pada tahap ini, dilakukan implementasi model pembelajaran mesin menggunakan algoritma *Multinomial Naïve Bayes* (MultinomialNB) untuk tugas klasifikasi. Langkah pertama dalam proses ini adalah inisialisasi model. Peneliti memulai dengan membuat *instance* dari kelas `MultinomialNB`, yang kemudian disimpan dalam variabel `model`. Algoritma *Multinomial Naïve Bayes* sangat cocok untuk menangani data yang didistribusikan secara multinomial, seperti dokumen teks yang telah

diubah menjadi representasi *bag-of-words* atau TF-IDF.

Setelah model diinisialisasi, langkah selanjutnya adalah melatih model tersebut menggunakan data pelatihan. Penggunaan metode `fit` dari objek `model` untuk melakukan proses pelatihan. Dalam hal ini, `x_train` adalah matriks fitur yang berisi representasi data input, sementara `y_train` adalah vektor label yang berisi kategori atau kelas target dari data pelatihan tersebut. Proses pelatihan ini memungkinkan model untuk mempelajari pola-pola dalam data sehingga model dapat membuat prediksi yang akurat pada data baru yang belum pernah dilihat sebelumnya.

E. Evaluasi Model

1. Hasil Akurasi

Setelah model dilatih, dilakukan evaluasi menggunakan data uji. Evaluasi model dilakukan dengan menghitung akurasi, confusion matrix, dan classification report. Akurasi adalah metrik yang menunjukkan persentase prediksi yang benar. Confusion matrix menunjukkan jumlah prediksi benar dan salah untuk setiap kelas. Classification report memberikan informasi lebih mendetail tentang precision, recall, dan F1-score.

```
accuracy = accuracy_score(y_test, y_predict)
print("Accuracy:", accuracy)

conf_matrix = confusion_matrix(y_test, y_predict)
print("Confusion Matrix:\n", conf_matrix)

class_report = classification_report(y_test, y_predict)
print("Classification Report:\n", class_report)
```

Program ini bertujuan untuk mengevaluasi kualitas sebuah model klasifikasi pada data yang belum pernah dilihat sebelumnya, yang disebut sebagai data uji. Proses evaluasi dimulai dengan melakukan prediksi

menggunakan model yang sudah dilatih sebelumnya terhadap data uji (`x_test`). Hasil prediksi ini kemudian dibandingkan dengan label sebenarnya dari data uji (`y_test`).

Langkah pertama adalah menghitung akurasi model, yang mencerminkan seberapa baik model dapat memprediksi label dengan benar. Selanjutnya, dibuat matriks kebingungan (confusion matrix) yang memberikan gambaran tentang jumlah prediksi yang benar (true positive dan true negative) serta yang salah (false positive dan false negative) untuk setiap kelas yang ada.

Selain itu, program juga menghasilkan laporan klasifikasi yang detail. Laporan ini mencakup presisi (precision), recall, nilai F1-score, dan dukungan (support) untuk setiap kelas dalam data uji. Presisi mengindikasikan seberapa sering model dapat memprediksi suatu kelas dengan benar dari semua prediksi yang dilakukan untuk kelas tersebut. Recall menunjukkan seberapa baik model dapat menemukan kembali semua contoh kelas yang sebenarnya. Nilai F1-score adalah rata-rata harmonik dari presisi dan recall, memberikan gambaran komprehensif tentang performa klasifikasi.

a. Stopword TF-IDF

- Pembagian Data 90:10

```
Accuracy: 0.7866666666666666
Confusion Matrix:
 [[119  23  10]
 [ 12  95  24]
 [ 12  15 140]]
Classification Report:
              precision    recall  f1-score   support

     Negatif       0.83      0.78      0.81       152
      Netral       0.71      0.73      0.72       131
     Positif       0.80      0.84      0.82       167

    accuracy                           0.79       450
   macro avg       0.78      0.78      0.78       450
weighted avg       0.79      0.79      0.79       450
```

Gambar 6 Akuraasi TF-IDF Rasio 90:10

- Pembagian Data 80:20

```
Accuracy: 0.7888888888888889
Confusion Matrix:
 [[248  42  18]
 [ 31 210  50]
 [ 21  28 252]]
Classification Report:
              precision    recall  f1-score   support

     Negatif       0.83      0.81      0.82       308
      Netral       0.75      0.72      0.74       291
     Positif       0.79      0.84      0.81       301

    accuracy                           0.79       900
   macro avg       0.79      0.79      0.79       900
weighted avg       0.79      0.79      0.79       900
```

Gambar 7 Akuraasi TF-IDF Rasio 80:20

- Pembagian Data 70:30

```
Accuracy: 0.78
Confusion Matrix:
 [[361  62  22]
 [ 58 311  73]
 [ 37  45 381]]
Classification Report:
              precision    recall  f1-score   support

     Negatif       0.79      0.81      0.80       445
      Netral       0.74      0.70      0.72       442
     Positif       0.80      0.82      0.81       463

    accuracy                           0.78      1350
   macro avg       0.78      0.78      0.78      1350
weighted avg       0.78      0.78      0.78      1350
```

Gambar 8 Akuraasi TF-IDF Rasio 70:30

b. Stopword Sastrawi

- Pembagian data 90:10

```
Accuracy: 0.7666666666666667
Confusion Matrix:
 [[118  22  12]
 [ 14  87  30]
 [ 10  17 140]]
Classification Report:
              precision    recall  f1-score   support

     Negatif       0.83      0.78      0.80       152
      Netral       0.69      0.66      0.68       131
     Positif       0.77      0.84      0.80       167

    accuracy                           0.77       450
   macro avg       0.76      0.76      0.76       450
weighted avg       0.77      0.77      0.77       450
```

Gambar 9 Akuraasi Sastrawi Rasio 90:10

- Pembagian data 80:10

```
Accuracy: 0.7733333333333333
Confusion Matrix:
 [[243  40  25]
 [ 36 197  58]
 [ 21  24 256]]
Classification Report:
              precision    recall  f1-score   support

     Negatif       0.81      0.79      0.80       308
      Netral       0.75      0.68      0.71       291
     Positif       0.76      0.85      0.80       301

    accuracy                           0.77       900
   macro avg       0.77      0.77      0.77       900
weighted avg       0.77      0.77      0.77       900
```

Gambar 10 Akuraasi Sastrawi Rasio 80:20

- Pembagian Data 70:30

```
Accuracy: 0.7533333333333333
Confusion Matrix:
 [[342  68  35]
 [ 60 293  89]
 [ 33  48 382]]
Classification Report:
              precision    recall  f1-score   support

     Negatif       0.79      0.77      0.78       445
      Netral       0.72      0.66      0.69       442
     Positif       0.75      0.83      0.79       463

    accuracy                           0.75      1350
   macro avg       0.75      0.75      0.75      1350
weighted avg       0.75      0.75      0.75      1350
```

Gambar 11 Akuraasi Sastrawi Rasio 70:30

Dalam melakukan evaluasi terhadap model klasifikasi sentimen pada data ulasan, kami membagi data menjadi tiga percobaan dengan rasio pembagian yang berbeda: 90:10, 80:20, dan 70:30. Tujuan utama dari evaluasi ini adalah untuk memahami

bagaimana performa model dapat bervariasi tergantung pada pembagian data yang digunakan.

Pada percobaan pertama dengan pembagian data 90:10, kami mencapai akurasi sebesar 0.786 untuk TF-DF dan 0.766 untuk sastrawi. Hasil ini menunjukkan bahwa model mampu mengklasifikasikan ulasan dengan tingkat keakuratan yang baik, namun masih ada ruang untuk peningkatan.

Percobaan kedua dengan pembagian data 80:20 memberikan akurasi yang hampir sama, yaitu 0.788 untuk TF-IDF dan 0.773 untuk sastrawi. Hal ini menunjukkan bahwa pembagian data ini memberikan hasil yang stabil dan konsisten dalam mempertahankan performa model.

Sementara itu, pada percobaan dengan pembagian data 70:30, meskipun akurasi tetap tinggi dengan nilai 0.766 untuk TF- IDF dan 0.753 untuk sastrawi, terlihat sedikit penurunan dibandingkan dengan percobaan sebelumnya. Ini menyarankan bahwa proporsi data latih yang lebih kecil dapat mempengaruhi kemampuan model dalam mengeneralisasi data uji.

Confusion matrix dari ketiga percobaan menunjukkan distribusi prediksi yang benar dan salah untuk masing-masing kelas sentimen (Negatif, Netral, Positif). Evaluasi lebih lanjut dengan menggunakan classification report memberikan insight tentang precision, recall, dan f1-score untuk setiap kelas. Hal ini membantu kami untuk memahami seberapa baik model dalam mengklasifikasikan ulasan ke dalam kategori sentimen yang sesuai.

Berdasarkan hasil evaluasi ini, peneliti merekomendasikan pembagian data 80:20 sebagai pilihan yang optimal. Pembagian ini memberikan akurasi yang baik serta mempertahankan stabilitas performa model dalam menghadapi data uji yang baru. Dengan

demikian, kami dapat memastikan bahwa model dapat diandalkan dalam memprediksi sentimen berdasarkan ulasan dengan tingkat keakuratan yang tinggi dan konsisten.

2. Hasil Analisis

Pada hasil prediksi menggunakan stopword dengan metode TF-IDF, terdapat 110 prediksi yang salah dari total 450 data yang diuji. Sementara itu, dengan menggunakan metode Sastrawi, terdapat 105 prediksi yang salah dari jumlah data yang sama.

a. TF-IDF

Table 3 Hasil Prediksi TF-IDF

Original Text	Cleaned Text	Predicted Sentiment	Actual Sentiment	Result
View nggk bagus	View bagus	Positif	Negatif	SALAH
Lumayan, cuma tiketnya mahal. Wahananya juga harus ganti-ganti nyalanya.	Lumayan cuma tiketnya mahal Wahananya juga harus	Netral	Negatif	SALAH
Tempatnya bagus, untuk masa pandemi protokol kesehatan betul-betul diterapkan di tempat ini,	Tempatnya bagus untuk di tempat ini ada banyak tempat dan	Positif	Netral	SALAH

ada banyak tempat cuci tangan, handsanitizer, dan alat pengukur suhu.				
Arena permainannya cukup memuaskan.	Arena permainannya cukup memuaskan	Positif	Netral	SALAH

Pada contoh di atas, kesalahan terjadi karena metode TF-IDF dengan stopword tidak dapat menangkap konteks negatif dalam kalimat seperti "View nggk bagus". Kata "nggk" yang artinya "tidak" mungkin tidak dikenali dengan baik oleh model atau tidak diberi bobot yang cukup besar dalam vektor fitur. Selain itu, teks dengan sentimen campuran atau yang memiliki banyak detail sering kali salah diklasifikasikan, seperti terlihat pada kalimat "Tempatnya bagus, untuk masa pandemi protokol kesehatan betul-betul diterapkan di tempat ini, ada banyak tempat cuci tangan, handsanitizer, dan alat pengukur suhu". Teks ini memiliki bagian yang positif ("Tempatnya bagus") dan bagian yang negatif ("protokol kesehatan sangat minim"), dan model mungkin lebih fokus pada bagian positif. Disamping utu juga terdapat kesalahan pada kata umum yag mengandung sentiment seperti pada kalimat "Arena permainanya cukup memuaskan". Kata "cukup" dalam konteks ini tidak terlalu kuat untuk memberikan sentimen positif, namun model mungkin menganggapnya demikian.

b.Sastrawi

Table 4 Hasil Prediksi Sastrawi

Original Text	Cleaned Text	Predicted Sentiment	Actual Sentiment	Result
View nggk bagus	View nggk bagus	Negatif	Negatif	BENAR
Lumayan, cuma tiketnya mahal. Wahananya juga harus ganti-ganti nyalanya.	Lumayan cuma tiketnya mahal Wahananya juga harus	Netral	Negatif	SALAH
Tempatnya bagus, untuk masa pandemi protokol kesehatan betul-betul diterapkan di tempat ini, ada banyak tempat cuci tangan, handsanitizer, dan alat pengukur suhu.	Tempatnya bagus untuk di tempat ini ada banyak tempat dan	Netral	Netral	BENAR
Arena permainannya cukup	Arena permainannya cukup	Positif	Netral	SALAH

memuaskan. memuaskan

Pada metode Sastrawi, terdapat beberapa perbaikan dalam pengenalan konteks, seperti pada kalimat “View nggk bagus” yang berhasil diklasifikasikan sebagai negatif. Namun, beberapa teks dengan sentimen campuran masih sulit untuk diklasifikasikan dengan benar. Misalnya, kalimat “Lumayan, cuma tiketnya mahal. Wahananya juga harus ganti-ganti nyalanya”. tetap salah diklasifikasikan sebagai netral.

Disamping dari perbedaan hasil prediksi, terdapat juga perbedaan cara pembersihan teks dari penggunaan dua metode digunakan, metode berbasis nilai TF-IDF dan metode Sastrawi. Berikut perbandingan hasil pembersihan teks menggunakan kedua metode tersebut:

- Pada kalimat "View nggk bagus," metode TF-IDF menghapus kata "nggk" karena dianggap tidak signifikan berdasarkan nilai TF-IDF-nya, sehingga hasil teks bersih menjadi "View bagus." Sebaliknya, metode Sastrawi mempertahankan kata "nggk," menghasilkan teks bersih yang sama dengan teks asli, yaitu "View nggk bagus." Hal ini menunjukkan bahwa metode TF-IDF lebih agresif dalam menghapus kata-kata yang dianggap tidak memberikan kontribusi penting.
- Untuk kalimat "Lumayan, cuma tiketnya mahal. Wahananya juga harus ganti-ganti nyalanya.," kedua metode pembersihan menghasilkan teks bersih yang sama: "Lumayan cuma tiketnya mahal Wahananya juga harus." Ini menunjukkan bahwa kata-kata yang dihapus oleh metode TF-IDF dalam kasus ini juga dianggap sebagai kata-kata tidak penting oleh metode Sastrawi.
- Pada kalimat yang lebih panjang dan kompleks seperti "Tempatnya bagus, untuk masa pandemi protokol kesehatan betul-betul diterapkan di tempat

ini, ada banyak tempat cuci tangan, handsanitizer, dan alat pengukur suhu.," kedua metode pembersihan menghasilkan teks bersih yang identik: "Tempatnya bagus untuk di tempat ini ada banyak tempat dan." Namun, dalam proses pembersihan menggunakan TF-IDF, kata-kata tertentu yang memiliki nilai TF-IDF rendah mungkin telah dihapus, yang menunjukkan efektivitas kedua metode dalam mempertahankan kata-kata penting meskipun terdapat beberapa perbedaan dalam pendekatan mereka.

- Terakhir, pada kalimat "Arena permainannya cukup memuaskan.," kedua metode pembersihan menghasilkan teks bersih yang sama: "Arena permainannya cukup memuaskan." Ini menunjukkan bahwa pada kalimat yang lebih sederhana, hasil pembersihan kedua metode cenderung serupa.

Secara keseluruhan, perbandingan ini menunjukkan bahwa metode pembersihan menggunakan TF-IDF cenderung menghapus kata-kata yang dianggap tidak signifikan berdasarkan nilai TF-IDF mereka, sementara metode Sastrawi lebih cenderung mempertahankan kata-kata asli. Meskipun pendekatan berbasis TF-IDF dapat membantu meningkatkan fokus pada kata-kata yang lebih penting dalam analisis sentimen, penggunaannya harus dilakukan dengan hati-hati untuk menghindari penghilangan konteks penting dalam teks ulasan.

BAB V

KESIMPILAN DAN SARAN

A. Kesimpulan

1. Penelitian ini dapat mengembangkan daftar stopwords yang lebih relevan menggunakan algoritma TF-IDF untuk meningkatkan representasi teks dalam analisis sentimen. Kata-kata dengan nilai TF-IDF rendah akan dimasukkan ke dalam daftar stopword baru sperti, ‘aisyah’, ‘aja’, ‘ajak’, ‘yg’, guna meningkatkan relevansi analisis sentimen.
2. Hasil pengujian dilakukan dengan berbagai skenario pembagian dataset, namun pada percobaan dengan rasio 80:20, kedua metode mencapai akurasi tertinggi, yaitu 0,788 untuk TF-IDF dan 0,773 untuk Sastrawi. Hasil ini menunjukkan bahwa model TF-IDF mampu mengklasifikasikan ulasan dengan tingkat akurasi yang cukup baik dibandingkan Sastrawi, meskipun masih ada ruang untuk perbaikan.

B. Saran

Saran penelitian berfokus pada penggunaan lebih lanjut dari metode TF-IDF dalam analisis teks dan perlunya penelitian lebih lanjut untuk mengoptimalkan penggunaan stopwords kontekstual dalam berbagai aplikasi analisis sentiment.

DAFTAR PUSTAKA

Duei Putri, D., Nama, G. F., & Sulistiono, W. E. (2022). Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier. *Jurnal Informatika Dan Teknik Elektro Terapan*, *10*(1), 34–40.https://doi.org/10.23960/jitet.v10i1.2262

Fahrizal, Reynaldi, F. O., & Hikmah, N. (2020). Implementasi Machine Learning padaSistem PETS Identification Menggunakan Python Berbasis UBuntu. *Journal of Information System, Informatics and Computing*, *4*(1), 86–91. http://journal.stmikjayakarta.ac.id/index.php/jisicom/article/view/212

Giarsyani, N. (2020). Komparasi Algoritma Machine Learning dan Deep Learning untuk Named Entity Recognition : Studi Kasus Data Kebencanaan. *Indonesian Journal of Applied Informatics*, *4*(2), 138. https://doi.org/10.20961/ijai.v4i2.41317

Kusumawardana, A. (2020). *Stopwords Bahasa Indonesiao Title*. Humas UKDW.https://www.ukdw.ac.id/stopwords-bahasa-indonesia-karya-ananda- kusumawardana/

M, R. (2021). *What is Scikit-Learn in Python?* ActiveState.https://www.activestate.com/resources/quick-reads/what-is-scikit-learn-in- python/

Mas Pintoko, B., & Muslim, K. (2018). Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naïve Bayes Classifier. *E-Proceeding of Engineering*, *5*(3), 8121–8130.

Mulyatun, S., Utama, H., & Mustopa, A. (2021). Pendekatan Natural LanguageProcessing Pada Aplikasi Chatbot Sebagai Alat Bantu Customer Service. *Journal of Information System Management (JOISM)*, *3*(1), 12–17.https://doi.org/10.24076/joism.2021v3i1.404

Rumaisa, F., Puspitarani, Y., Rosita, A., Zakiah, A., & Violina, S. (2021). Penerapan Natural Language Processing (NLP) di bidang pendidikan. *Jurnal Inovasi Masyarakat*, *1*(3), 232–235. https://doi.org/10.33197/jim.vol1.iss3.2021.799

Sari, Y. (2017). Pengenalan Natural Language Toolkit (NLTK). *3 September2019*, *September*, 1–5. https://code.tutsplus.com/id/tutorials/introducing-the-natural- language-toolkit-nltk--cms-28620

Septian, J. A., Fachrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. *Journal of Intelligent System and Computation*, *1*(1), 43–49. https://doi.org/10.52985/insyst.v1i1.36

Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, & Fitri Nurapriani. (2023). Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes danKNN. *Jurnal KomtekInfo*, *10*, 1–7.ht

Syaifuddin, A., & Ningsih, M. (2023). Penerapan Metode Content-Based Filtering Dalam Strategi Komunikasi Pemasaran Pada Marketplace Tokopedia. *Jurnal Responsif*, *5*(2), 185–194. https://ejurnal.ars.ac.id/index.php/jti

Wibawa, A. P., Miftahuddin, F., & Suyono. (2021). *K-MEDOIDS CLUSTERING UNTUK PEMBENTUKAN DATABASE STOPWORD*

BAHASA JAWA. *10*(2),261–269.

Widodod, F. (2021). *NLP Sederhana Dengan Python*. Sites Unpad. https://sites.unpad.ac.id/widodo/2021/03/09/nlp-dengan-python/

Yulita, W. (2021). Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier. *Jurnal Data Mining Dan Sistem Informasi*, *2*(2), 1. https://doi.org/10.33365/jdmsi.v2i2.1344

Yutika, C. H., Adiwijaya, A., & Faraby, S. Al. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes. *Jurnal Media Informatika Budidarma*, *5*(2), 422. https://doi.org/10.30865/mib.v5i2.2845

Zailani, A. U., Perdananto, A., Nurjaya, & Sholihin. (2020). Pengenalan Sejak Dini Siswa Smp Tentang Machine Learning Untuk Klasifikasi Gambar Dalam Menghadapi Revolusi 4.0. *KOMMAS: Jurnal Pengabdian Kepada Masyarakat*, *244*(1), 7–15. *http://openjournal.unpam.ac.id/index.php/kommas/articLampiran 1le/view/4599*

LAMPIRAN

Lampiran 1 Hasil plagiasi

MAJELIS PENDIDIKAN TINGGI PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN
Alamat kantor: Jl.Sultan Alauddin N0.259 Makassar 90221 Tlp.(0411) 866972,881593, Fax.(0411) 865588

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

SURAT KETERANGAN BEBAS PLAGIAT

UPT Perpustakaan dan Penerbitan Universitas Muhammadiyah Makassar,
Menerangkan bahwa mahasiswa yang tersebut namanya di bawah ini:

Nama : Damai Arsila Salsabila
Nim : 105841107520
Program Studi : Teknik Informatika

Dengan nilai:

No	Bab	Nilai	Ambang Batas
1	Bab 1	9 %	10 %
2	Bab 2	9 %	25 %
3	Bab 3	10 %	10 %
4	Bab 4	7 %	10 %
5	Bab 5	3 %	5 %

Dinyatakan telah lulus cek plagiat yang diadakan oleh UPT- Perpustakaan dan Penerbitan Universitas Muhammadiyah Makassar Menggunakan Aplikasi Turnitin.

Demikian surat keterangan ini diberikan kepada yang bersangkutan untuk dipergunakan seperlunya.

Makassar, 14 Agustus 2024
Mengetahui,

Kepala UPT- Perpustakaan dan Pernerbitan,

Nursinah, S.Ag., M.I.P
NBM. 964 591

Jl. Sultan Alauddin no 259 makassar 90222
Telepon (0411)866972,881 593,fax (0411)865 588
Website: www.library.unismuh.ac.id
E-mail : perpustakaan@unismuh.ac.id

BAB I DAMAI ARSILA SALSABILA 105841107520

by TahapTutup

Submission date: 13-Aug-2024 01:02PM (UTC+0700)
Submission ID: 2431407420
File name: bab_1_bila.docx (12.57K)
Word count: 749
Character count: 5049

BAB I DAMAI ARSILA SALSABILA 105841107520

ORIGINALITY REPORT

9%	8%	0%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	kartika1902.blogspot.com Internet Source	2%
2	digilibadmin.unismuh.ac.id Internet Source	2%
3	repository.radenintan.ac.id Internet Source	2%
4	id.123dok.com Internet Source	2%
5	Submitted to Universitas Brawijaya Student Paper	2%

Exclude quotes Off
Exclude bibliography Off
Exclude matches < 2%

BAB II DAMAI ARSILA SALSABILA 105841107520

by TahapTutup

Submission date: 13-Aug-2024 12:59PM (UTC+0700)
Submission ID: 2431406263
File name: bab_2_bila.docx (44.39K)
Word count: 1815
Character count: 11731

BAB II DAMAI ARSILA SALSABILA 105841107520

ORIGINALITY REPORT

9%	7%	5%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repository.uin-suska.ac.id Internet Source	3%
2	ejurnal.ars.ac.id Internet Source	3%
3	Rafi Rahmadani, Abdul Rahim, Rudiman Rudiman. "ANALISIS SENTIMEN ULASAN "OJOL THE GAME" DI GOOGLE PLAY STORE MENGGUNAKAN ALGORITMA NAIVE BAYES DAN MODEL EKSTRAKSI FITUR TF-IDF UNTUK MENINGKATKAN KUALITAS GAME", Jurnal Informatika dan Teknik Elektro Terapan, 2024 Publication	2%
4	journal.shantibhuana.ac.id Internet Source	2%

Exclude quotes Off

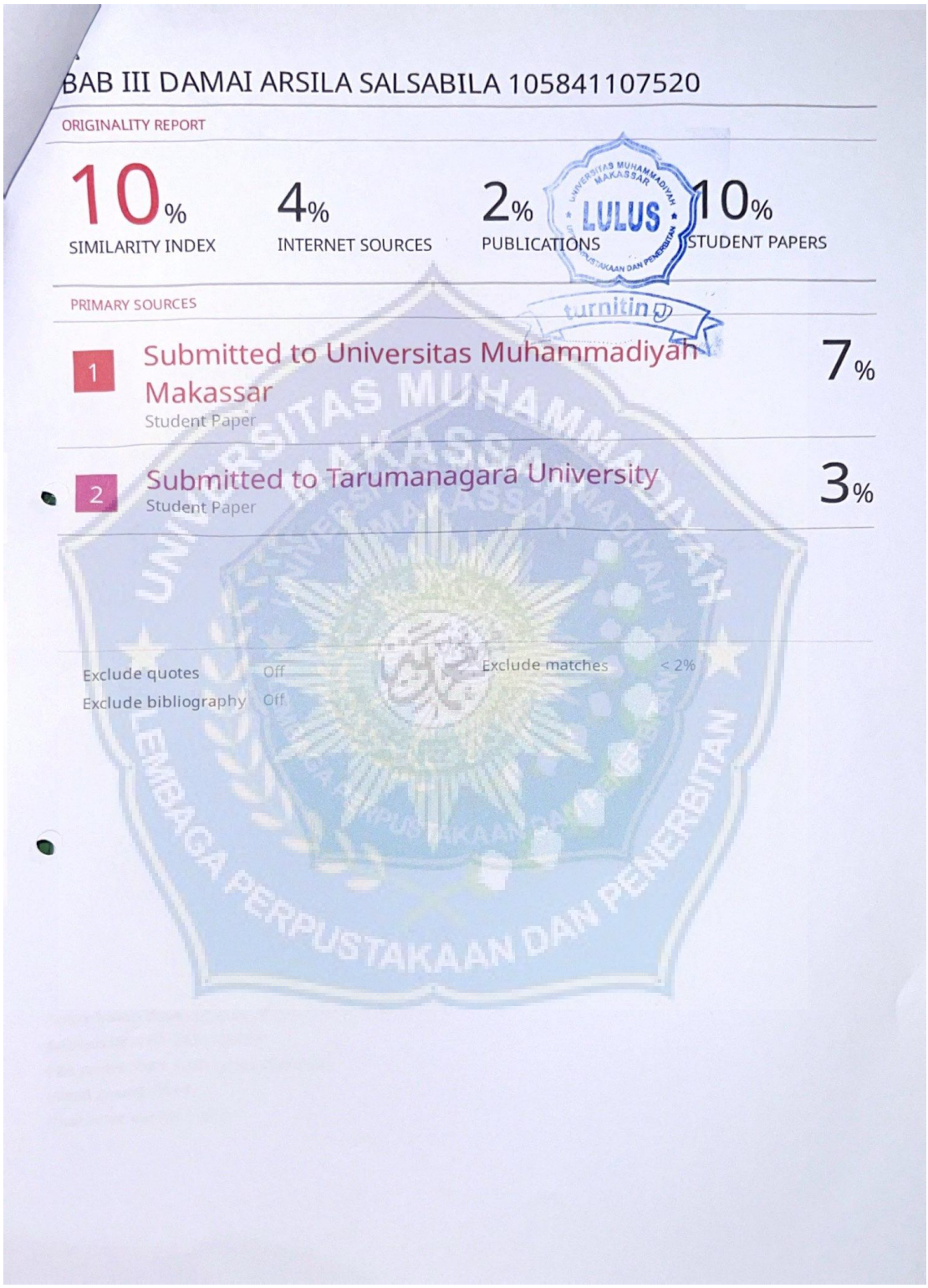
Exclude bibliography Off

Exclude matches < 2%

BAB III DAMAI ARSILA SALSABILA 105841107520

by TahapTutup

Submission date: 13-Aug-2024 01:00PM (UTC+0700)
Submission ID: 2431406471
File name: bab_3_bila.docx (44.79K)
Word count: 762
Character count: 5208



BAB III DAMAI ARSILA SALSABILA 105841107520

ORIGINALITY REPORT

10%	4%	2%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Muhammadiyah Makassar Student Paper	7%
2	Submitted to Tarumanagara University Student Paper	3%

Exclude quotes Off
Exclude bibliography Off
Exclude matches < 2%

BAB IV DAMAI ARSILA SALSABILA 105841107520

by TahapTutup

Submission date: 13-Aug-2024 01:01PM (UTC+0700)
Submission ID: 2431406853
File name: bab_4_bila.docx (525.59K)
Word count: 2644
Character count: 17596

BAB IV DAMAI ARSILA SALSABILA 105841107520

ORIGINALITY REPORT

7%	7%	0%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	digilibadmin.unismuh.ac.id Internet Source	7%

Exclude quotes Off
Exclude bibliography Off
Exclude matches < 2%

BAB V DAMAI ARSILA SALSABILA 105841107520

by TahapTutup

Submission date: 13-Aug-2024 01:01PM (UTC+0700)
Submission ID: 2431407048
File name: bab_5_bila.docx (10.41K)
Word count: 241
Character count: 1597

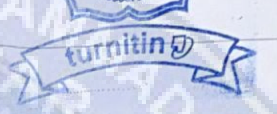
BAB V DAMAI ARSILA SALSABILA 105841107520

ORIGINALITY REPORT

3%	3%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

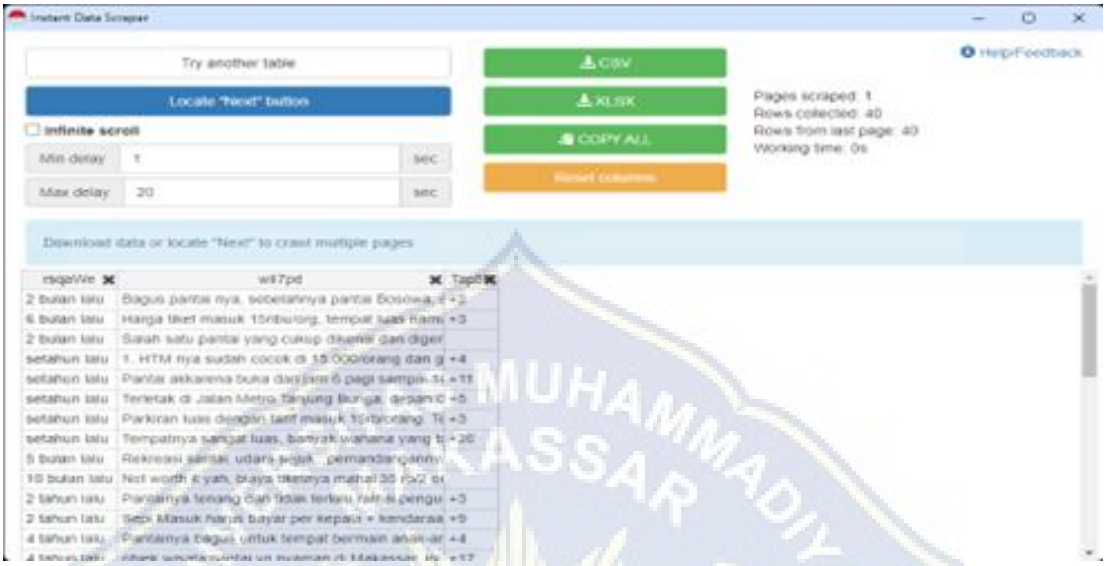
PRIMARY SOURCES

1	eprints.poltekkesjogja.ac.id Internet Source	3%



Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Lampiran 2 Pengambilan Data



Lampiran 3 Data Ulasan

Lampiran 4 Source Code TF-IDF

```
[1] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[2] import pandas as pd
    import string
    import numpy as np
    from sklearn.feature_extraction.text import TfidfVectorizer
    from sklearn.model_selection import train_test_split
    from sklearn.naive_bayes import MultinomialNB
    from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
```

```
[5] #Import manajemen dataset
    df = pd.read_excel('/content/drive/MyDrive/Damai Arsila/DATABARU.xlsx',sheet_name="Sheet1")
```

```
[7] df
```

[7]

	ULASAN	LABEL
0	Tempatnya sejuk. Lebih cocok untuk rekreasi an...	Positif
1	Banyak pohon pohon yang berbuah juga	Positif
2	Kebun wisata yang sangat dekat dari jalan besa...	Positif
3	Pas masuk cukup sejuk karena musim hujan, disa...	Positif
4	Rekomended juga krna banyak disediakan tempat ...	Positif
...	...	...
4495	Tempat ini sangat ramai setelah lebaran. Mungk...	Netral
4496	Harga tiket masuk yang disebutkan oleh petugas...	Netral
4497	Meskipun bersih dan bagus, akses jalan menuju ...	Netral
4498	Pantai ini sangat bersih dan pasirnya lembut.	Netral
4499	Salah satu pantai yang indah dengan pasir puti...	Netral

4500 rows × 2 columns

```
[8] df.describe()
```

	ULASAN	LABEL
count	4500	4500
unique	4400	3
top	Bersih	Positif
freq	4	1500

```
[9] df
```

[9]

	ULASAN	LABEL
0	Tempatnya sejuk. Lebih cocok untuk rekreasi an...	Positif
1	Banyak pohon pohon yang berbuah juga	Positif
2	Kebun wisata yang sangat dekat dari jalan besa...	Positif
3	Pas masuk cukup sejuk karena musim hujan, disa...	Positif
4	Rekomended juga krna banyak disediakan tempat ...	Positif
...	...	...
4495	Tempat ini sangat ramai setelah lebaran. Mungk...	Netral
4496	Harga tiket masuk yang disebutkan oleh petugas...	Netral
4497	Meskipun bersih dan bagus, akses jalan menuju ...	Netral
4498	Pantai ini sangat bersih dan pasirnya lembut.	Netral
4499	Salah satu pantai yang indah dengan pasir puti...	Netral

4500 rows × 2 columns

```
[10] # Membersihkan nilai np.nan pada kolom 'ULASAN'
     df['ULASAN'].fillna('', inplace=True)  # Mengganti np.nan dengan string kosong
```

```
[11] x = df['ULASAN']
     y = df['LABEL']
```

```
[12] x.describe()
```

```
count       4500
unique      4400
top       Bersih
freq         4
Name: ULASAN, dtype: object
```

```
[13] y
```

```
0       Positif
1       Positif
2       Positif
3       Positif
4       Positif
```

```
[14] # Fungsi untuk menghapus tanda baca dari teks
     def remove_punctuation(text):
         translator = str.maketrans('', '', string.punctuation)
         return text.translate(translator)
```

```
[15] # Terapkan penghapusan tanda baca pada kolom teks
     df['ULASAN_BERSIH'] = df['ULASAN'].apply(remove_punctuation)
```

```
[28] # Inisialisasi TfidfVectorizer
     vectorizer = TfidfVectorizer()
```

```
[29] # Transformasikan teks menjadi fitur
     x = vectorizer.fit_transform(df['ULASAN_BERSIH'])
     y = df['LABEL']
```

```
[30] # Hitung nilai rata-rata TF-IDF untuk setiap kata
     tfidf_scores = np.asarray(x.mean(axis=0)).flatten()
```

```
[31] # Ambil daftar kata-kata (fitur) dari vectorizer
     terms = vectorizer.get_feature_names_out()
```

```
[34] # Tentukan ambang batas untuk kata-kata yang akan ditampilkan
     threshold = 0.003

     # Dapatkan indeks kata-kata yang berada di bawah ambang batas
     indikasi_Kata_dibawah_threshold = np.where(tfidf_scores < threshold)[0]

     kata_dibawah_Threshold = [vectorizer.get_feature_names_out()[idx] for idx in indikasi_Kata_dibawah_threshold]

     # Cetak kata-kata yang memiliki nilai TF-IDF di bawah ambang batas
     print("Kata-kata dengan nilai TF-IDF di bawah ambang batas:")
     print(kata_dibawah_Threshold)
```

```
Kata-kata dengan nilai TF-IDF di bawah ambang batas:
['10', '100', '10000', '100000', '1004jam', '100k4', '100rb', '100ribu', '10k', '10rb', '10rborg', '10ribu', '12', '120', '120200', '13', '14', '15',
```

```
[35] # Simpan daftar kata-kata ke dalam DataFrame
     data = {'STOP WORD DARI TF-IDF': kata_dibawah_Threshold}
     df_words = pd.DataFrame(data)

     # Simpan DataFrame ke dalam file Excel
     df_words.to_excel('kata_kunci.xlsx', index=False)
```

```
[38] # Fungsi untuk menghapus stopword berdasarkan nilai TF-IDF
     def remove_low_tfidf_words(text):
         return ' '.join([word for word in text.split() if word not in kata_dibawah_Threshold])
```

```
[39] # Terapkan penghapusan stopword berdasarkan nilai TF-IDF pada kolom teks
     df['ULASAN_BERSIH'] = df['ULASAN_BERSIH'].apply(remove_low_tfidf_words)
```

```
[40] # Transformasikan teks yang telah dibersihkan menjadi fitur
     x = vectorizer.fit_transform(df['ULASAN_BERSIH'])
```

```
[41] # Splitting data dan simpan indeks dari x_test
     x_train, x_test, y_train, y_test, train_indices, test_indices = train_test_split(x, y, df.index, random_state=0, test_size=0.2)
```

```
[42] # Inisialisasi model Multinomial Naive Bayes
     model = MultinomialNB()
```

```
[43] # Latih model dengan data training
     model.fit(x_train, y_train)
```

```
MultinomialNB
MultinomialNB()
```

```
[44] # Prediksi pada data test
     y_predict = model.predict(x_test)
```

```
[45] # Evaluasi model
     accuracy = accuracy_score(y_test, y_predict)
     print("Accuracy:", accuracy)

     conf_matrix = confusion_matrix(y_test, y_predict)
     print("Confusion Matrix:\n", conf_matrix)

     class_report = classification_report(y_test, y_predict)
     print("Classification Report:\n", class_report)
```

```
Accuracy: 0.7388888888888889
Confusion Matrix:
 [[244  49  15]
 [ 47 180  64]
 [ 27  33 241]]
Classification Report:
              precision    recall  f1-score   support

     Negatif       0.77      0.79      0.78       308
      Netral       0.69      0.62      0.65       291
     Positif       0.75      0.80      0.78       301

    accuracy                           0.74       900
   macro avg       0.74      0.74      0.74       900
weighted avg       0.74      0.74      0.74       900
```

```
[46] # Pastikan panjang array sama
     print(f"Length of y_test: {len(y_test)}")
     print(f"Length of y_predict: {len(y_predict)}")
     print(f"Length of test_indices: {len(test_indices)}")
```

```
Length of y_test: 900
Length of y_predict: 900
Length of test_indices: 900
```

```
[47] # Tambahkan label hasil prediksi (Benar/Salah) berdasarkan Confusion Matrix
     result_labels = ['BENAR' if actual == predicted else 'SALAH' for actual, predicted in zip(y_test, y_predict)]

     # Membuat DataFrame dengan hasil prediksi dan nilai aktual menggunakan indeks asli
     results = pd.DataFrame({
         'Original Text': df.loc[test_indices, 'ULASAN'].values,
         'Cleaned Text': df.loc[test_indices, 'ULASAN_BERSIH'].values,
         'Predicted Sentiment': y_predict,
         'Actual Sentiment': y_test.values,
         'Result': result_labels
     })
```

```
[47] # Tampilkan DataFrame dengan teks dan hasil prediksinya
     print(results)

     # Simpan DataFrame ke dalam file Excel
     results.to_excel('HASIL PREDIKSI DENGAN TF-IDF.xlsx', index=False)
```

```
                                  Original Text  \
0                    Arena bermainnya luas dan nyaman.
1                              Tidak suka sama sekali
2                  Sangat buruk sebagai tempat berlibur
3    Sering dibicarakan org tempat ini buruk .. gk ...
4    Ada fasilitas kereta anak yang sudah tidak ber...
..                                                 ...
895                            jam segini masi buka?
896  Kurang bagus dan tidak bersih dengan tiket mas...
897  Tolong untuk penjadwalan wahana dipajang agar ...
898  Salah satu tempat wisata di Makassar yang sang...
899           Sangat bisa untuk membawa makanan masuk.

                                       Cleaned Text Predicted Sentiment  \
0                              Arena luas dan nyaman             Positif
1                             Tidak suka sama sekali             Negatif
2                         Sangat buruk tempat berlibur             Negatif
3                       Sering tempat ini buruk kesini             Negatif
4              Ada fasilitas anak yang sudah tidak lagi             Positif
```

Lampiran 5 Source Code Sastrawi

```
[ ] pip install sastrawi
```

```
Collecting sastrawi
  Downloading Sastrawi-1.0.1-py2.py3-none-any.whl (209 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 209.7/209.7 kB 1.7 MB/s eta 0:00:00
Installing collected packages: sastrawi
Successfully installed sastrawi-1.0.1
```

```
[ ] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[ ] import pandas as pd
    import string
    from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
    from sklearn.feature_extraction.text import TfidfVectorizer
    from sklearn.model_selection import train_test_split
    from sklearn.naive_bayes import MultinomialNB
    from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
```

```
[ ] #Import manajemen dataset
    df = pd.read_excel('/content/drive/MyDrive/Damai Arsila/DATABARU.xlsx',sheet_name="Sheet1")
```

```
[ ] df
```

	ULASAN	LABEL
0	Tempatnya sejuk. Lebih cocok untuk rekreasi an...	Positif
1	Banyak pohon pohon yang berbuah juga	Positif
2	Kebun wisata yang sangat dekat dari jalan besa...	Positif
3	Pas masuk cukup sejuk karena musim hujan, disa...	Positif
4	Rekomended juga krna banyak disediakan tempat ...	Positif
...	...	...
4495	Tempat ini sangat ramai setelah lebaran. Mungk...	Netral
4496	Harga tiket masuk yang disebutkan oleh petugas...	Netral
4497	Meskipun bersih dan bagus, akses jalan menuju ...	Netral
4498	Pantai ini sangat bersih dan pasirnya lembut	Netral

```
[ ] df.describe()
```

	ULASAN	LABEL
count	4500	4500
unique	4400	3
top	Bersih	Positif
freq	4	1500

```
[ ] df
```

	ULASAN	LABEL
0	Tempatnya sejuk. Lebih cocok untuk rekreasi an...	Positif
1	Banyak pohon pohon yang berbuah juga	Positif
2	Kebun wisata yang sangat dekat dari jalan besa...	Positif
3	Pas masuk cukup sejuk karena musim hujan, disa...	Positif
4	Rekomended juga krna banyak disediakan tempat ...	Positif

```
[ ] # Membersihkan nilai np.nan pada kolom 'ULASAN'
    df['ULASAN'].fillna('', inplace=True)  # Mengganti np.nan dengan string kosong
```

```
[ ] x = df['ULASAN']
    y = df['LABEL']
```

```
[ ] x.describe()
```

```
count       4500
unique      4400
top       Bersih
freq           4
Name: ULASAN, dtype: object
```

```
[ ] y
```

```
0       Positif
1       Positif
2       Positif
3       Positif
4       Positif
```

```
[ ] # Inisialisasi stopword remover bahasa Indonesia
    factory = StopWordRemoverFactory()
    stopword = factory.create_stop_word_remover()
```

```
[ ] # Fungsi untuk menghapus tanda baca dari teks
    def remove_punctuation(text):
        translator = str.maketrans('', '', string.punctuation)
        return text.translate(translator)
```

```
[ ] # Fungsi untuk menghapus stopword dari teks
    def remove_stopwords(text):
        return stopword.remove(text)
```

```
[ ] # Terapkan penghapusan tanda baca dan stopword pada kolom teks
    df['ULASAN_BERSIH'] = df['ULASAN'].apply(remove_punctuation).apply(remove_stopwords)
```

```
[ ] # Inisialisasi TfidfVectorizer
    vectorizer = TfidfVectorizer()
```

```
[ ] # Transformasikan teks menjadi fitur
    x = vectorizer.fit_transform(df['ULASAN_BERSIH'])
    y = df['LABEL']
```

```
[ ] # Splitting data dan simpan indeks dari x_test
    x_train, x_test, y_train, y_test, train_indices, test_indices = train_test_split(x, y, df.index, random_state=0, test_size=0.3)
```

```
[ ] # Inisialisasi model Multinomial Naive Bayes
    model = MultinomialNB()
```

```
[ ] # Latih model dengan data training
    model.fit(x_train, y_train)
```

```
▾ MultinomialNB
MultinomialNB()
```

```
[ ] # Prediksi pada data test
    y_predict = model.predict(x_test)
```

```
[ ] # Evaluasi model
    accuracy = accuracy_score(y_test, y_predict)
    print("Accuracy:", accuracy)

    conf_matrix = confusion_matrix(y_test, y_predict)
    print("Confusion Matrix:\n", conf_matrix)

    class_report = classification_report(y_test, y_predict)
    print("Classification Report:\n", class_report)
```

```
Accuracy: 0.7533333333333333
Confusion Matrix:
 [[342  68  35]
 [ 60 293  89]
 [ 33  48 382]]
Classification Report:
              precision    recall  f1-score   support

     Negatif       0.79      0.77      0.78       445
      Netral       0.72      0.66      0.69       442
     Positif       0.75      0.83      0.79       463

    accuracy                           0.75      1350
   macro avg       0.75      0.75      0.75      1350
weighted avg       0.75      0.75      0.75      1350
```

```
[ ] # Pastikan panjang array sama
    print(f"Length of y_test: {len(y_test)}")
    print(f"Length of y_predict: {len(y_predict)}")
    print(f"Length of test_indices: {len(test_indices)}")
```

```
Length of y_test: 1350
Length of y_predict: 1350
Length of test_indices: 1350
```

```
[ ] # Tambahkan label hasil prediksi (Benar/Salah) berdasarkan Confusion Matrix
    result_labels = ['BENAR' if actual == predicted else 'SALAH' for actual, predicted in zip(y_test, y_predict)]

    # Membuat DataFrame dengan hasil prediksi dan nilai aktual menggunakan indeks asli
    results = pd.DataFrame({
        'Original Text': df.loc[test_indices, 'ULASAN'].values,
        'Cleaned Text': df.loc[test_indices, 'ULASAN_BERSIH'].values,
        'Predicted Sentiment': y_predict,
        'Actual Sentiment': y_test.values,
        'Result': result_labels
    })
```

```
[ ] # Tampilkan DataFrame dengan teks dan hasil prediksinya
    print(results)

    # Simpan DataFrame ke dalam file Excel
    results.to_excel('HASIL PREDIKSI DENGAN SASTRAWI.xlsx', index=False)
```

```
                                      Original Text  \
0                    Arena bermainnya luas dan nyaman.
1                                Tidak suka sama sekali
2                   Sangat buruk sebagai tempat berlibur
3     Sering dibicarakan org tempat ini buruk .. gk ...
4     Ada fasilitas kereta anak yang sudah tidak ber...
...                                                 ...
1345  Tempatnya bagus untuk olahraga pagi, dan tidak...
1346  Bisa membawa makanan dari luar, bisa menggunak...
1347  Resor ini tidak cocok untuk berkumpul bersama ...
1348  Senang sekali melihat kolam renang dan pohon-p...
1349                                  kolam penuh lumut

                                           Cleaned Text Predicted Sentiment  \
0                          Arena bermainnya luas nyaman             Positif
1                              Tidak suka sama sekali               Negatif
2                            Sangat buruk tempat berlibur             Negatif
3     Sering dibicarakan org tempat buruk  gk bakal ...             Negatif
4     Ada fasilitas kereta anak sudah berfungsi sehi...             Negatif
```