JURNAL INSY PRO Information System and Processing
Volume 10, Nomor 1, Mei 2025 P–ISSN 2597-419X, E-ISSN 2579-468X
Ojs : http://journal.uinalauddin.ac.id/index.php/insypro
Email : insypro@uin-alauddin.ac.id

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

# *ANALYSIS AND COMPARISON OF STOPWORDS ON TEXT SENTIMENT ANALYSIS ACCURACY USING TF-IDF: A CASE STUDY IN NLP*

**Titin Wahyuni[1], Damai Arsila Salsabila[2]**
[1,2]Program Studi Informatika Universitas Muhammadiyah Makassar
[1,2]Jl. Sultan Alauddin No 259, Kota Makassar, Sulawesi Selatan 90123

E-mail: Titinwahyuni@unismuh.ac.id [1)], 105841108520@student.unismuh.ac.id [2)]

**Abstrak** – Dalam era digital yang berkembang pesat, jumlah data teks online meningkat signifikan, mencakup ulasan produk, komentar media sosial, dan artikel berita. Analisis sentimen, yang penting untuk memahami opini masyarakat, membutuhkan praproses teks dengan menghapus stopwords tidak signifikan. Penelitian ini mengeksplorasi penggunaan algoritma TF-IDF untuk menghasilkan stopwords kontekstual dan membandingkan pengaruhnya terhadap akurasi model analisis sentimen dengan stopwords standar. 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*** *– 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.*

## PENDAHULUAN

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 et al., 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.

JURNAL INSY PRO Information System and Processing
Volume 10, Nomor 1, Mei 2025 P–ISSN 2597-419X, E-ISSN 2579-468X
Ojs : http://journal.uinalauddin.ac.id/index.php/insypro
Email : insypro@uin-alauddin.ac.id

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 digunakan dalam 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.

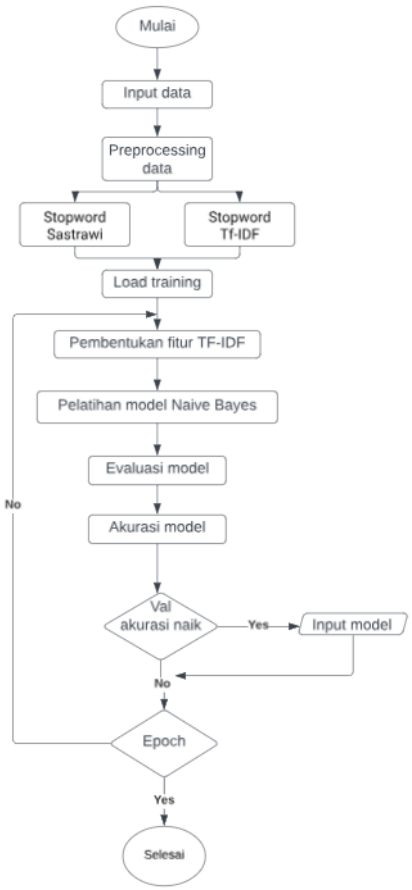
Oleh karena itu, studi kasus ini bertujuan untuk menganalisis dan membandingkan pengaruh penggunaan stopwords sastrawi dengan stopwords yang dihasilkan menggunakan algoritma TF-IDF terhadap akurasi analisis sentimen teks. Dengan studi kasus pada berbagai dataset teks, penelitian ini diharapkan dapat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

## METODOLOGI PENELITIAN

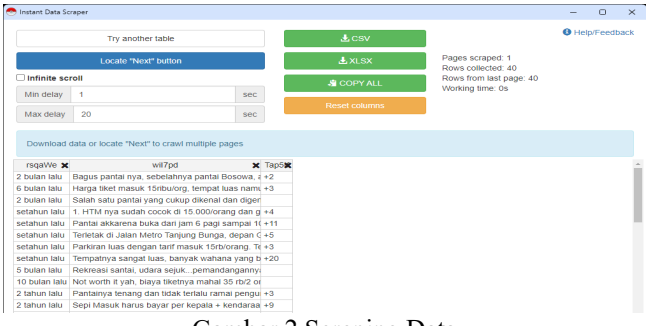
TF-IDF adalah metode yang digunakan untuk mengukur signifikansi kata dalam dokumen berdasarkan frekuensi kemunculannya (Syaifuddin & Ningsih, 2023). Pendekatan ini terdiri dari dua konsep utama, yaitu Term Frequency (TF) yang menghitung seberapa sering kata tersebut muncul dalam dokumen tertentu, dan Inverse Document Frequency (IDF) yang menggukur seberapa umum sebuah kata di seluruh koleksi dokumen (Yutika et al., 2021). Ini membantu meningkatkan kinerja model dalam analisis teks dan sentimen dengan memfokuskan pada kata-kata yang lebih penting.

Seperti pada gambar 1 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 pada kedua 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




Gambar 1 Perancangan Sistem




Gambar 2 Screping Data

hasil tersebut penulis dapat menarik kesimpulan terhadap pengaruh dalam penggunaan stopword.

Penelitian ini menganalisis 4.500 data ulasan dari Google Maps mengenai tempat wisata, mencakup aspek pemandangan, fasilitas, pelayanan, keramaian, dan kebersihan. Data ulasan tersebut diklasifikasikan ke dalam tiga label sentimen—positif, negatif, dan netral—masing-masing dengan 1.500 data. Setelah pengumpulan dan analisis, hasilnya disimpan dalam format Excel, yang merangkum berbagai aspek layanan, 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 |

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.

| 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 |

Proses pengolahan data dimulai dengan pengisian nilai kosong dalam kolom ulasan pada dataframe. Pada tahap ini, nilai kosong diisi dengan string kosong secara langsung pada dataframe, tanpa perlu disimpan ke dalam variabel baru. Langkah ini dilakukan untuk memastikan bahwa data yang akan diproses selanjutnya tidak terganggu oleh keberadaan nilai kosong yang bisa menyebabkan kesalahan dalam analisis.

Selanjutnya, dilakukan penghapusan tanda baca melalui fungsi bernama `remove_punctuation`. Fungsi ini menggunakan metode `str.maketrans` dalam Python untuk membuat translator yang bertugas menghapus semua tanda baca dari teks yang diberikan. Dengan menghilangkan tanda baca, teks menjadi lebih bersih dan siap untuk dianalisis lebih lanjut.

Proses berikutnya adalah penghapusan stopword, yaitu kata-kata umum seperti "dan", "atau", "yang" yang sering tidak memberikan makna khusus dalam analisis teks. Untuk menghapus stopwords, library khusus diimpor, seperti *stopword remover factory*. Objek remover stopwords dibuat dan digunakan untuk memproses teks, menghapus kata-kata yang tidak relevan, sehingga menghasilkan teks yang lebih bersih dan fokus untuk analisis.

## HASIL DAN PEMBAHASAN

Penulis melakukan evaluasi model klasifikasi sentimen pada data ulasan dengan tiga percobaan menggunakan rasio pembagian data yang berbeda: 90:10, 80:20, dan 70:30. Tujuan dari evaluasi ini adalah

untuk memahami variasi performa model berdasarkan pembagian data yang digunakan.

a. Percobaan 90:10

Akurasi yang dicapai adalah 0.786 untuk TF-IDF dan 0.766 untuk Sastrawi. Model menunjukkan performa yang baik, tetapi masih ada ruang untuk peningkatan.

b. Percobaan 80:20

Akurasi hampir sama dengan percobaan sebelumnya, yaitu 0.788 untuk TF-IDF dan 0.773 untuk Sastrawi. Pembagian ini memberikan hasil yang stabil dan konsisten.

c. Percobaan 70:30

Akurasi sedikit menurun menjadi 0.766 untuk TF-IDF dan 0.753 untuk Sastrawi, menunjukkan bahwa proporsi data latih yang lebih kecil dapat mempengaruhi kemampuan model dalam menggeneralisasi data uji.

Confusion matrix dan classification report menunjukkan distribusi prediksi yang benar dan salah serta memberikan wawasan tentang precision, recall, dan f1-score untuk setiap kelas sentimen. Berdasarkan hasil evaluasi ini, pembagian data 80:20 direkomendasikan sebagai pilihan optimal karena memberikan akurasi yang baik dan stabilitas performa model.

Untuk Hasil akurasi dari 450 data yang diuji, prediksi menggunakan stopword dengan metode TF-IDF menghasilkan 110 kesalahan, sementara metode Sastrawi menghasilkan 105 kesalahan.

a. Metode TF-IDF:

- Kesalahan terjadi karena model tidak dapat menangkap konteks negatif, seperti pada kalimat "View nggk bagus" yang diprediksi sebagai positif karena kata "nggk" tidak dikenali dengan baik.
- •Teks dengan sentimen campuran, seperti "Tempatnya bagus, untuk masa pandemi protokol kesehatan betul-betul diterapkan," sering salah diklasifikasikan karena model lebih fokus pada bagian positif.

b. Metode Sastrawi:

- Model berhasil mengklasifikasikan beberapa konteks dengan lebih baik, seperti "View nggk bagus" yang diprediksi sebagai negatif.
- Namun, beberapa teks dengan sentimen campuran masih salah diklasifikasikan, seperti "Lumayan, cuma tiketnya mahal. Wahananya juga harus ganti-ganti nyalanya."

## KESIMPULAN

Penelitian ini mengeksplorasi metode dan teknik dalam menghasilkan stopword baru, serta memberikan pemahaman tentang efektivitas penggunaan TF-IDF untuk menghapus kata-kata yang tidak signifikan. Kata-kata dengan nilai TF-IDF rendah akan dimasukkan ke dalam daftar stopword baru, guna meningkatkan relevansi analisis sentimen.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. 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

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

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 dan KNN. *Jurnal KomtekInfo*, *10*, 1–7. https://doi.org/10.35134/komtekinfo.v10i1.330

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

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