# PROJECT A

# Perbandingan Analisis Sentimen pada Pemberitaan Danantara Menggunakan Model Logistic Regression dengan Feature Engineering TF-IDF dan BERT


DISUSUN OLEH:

| | |
|---|---|
| Diva Ardelia Alyadrus | 5026221029 |
| Shof Watun Niswah | 5026221043 |
| Muhammad Daffa Alvinoer Rahman | 5026221180 |


**PENGOLAHAN BAHASA ALAMI (A)**


**DEPARTEMEN SISTEM INFORMASI**
**FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS**
**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**SEMESTER GENAP 2025**

# DAFTAR ISI

# ABSTRAK

Penelitian ini bertujuan untuk menganalisis persepsi publik terhadap kebijakan Danantara melalui artikel berita daring dengan pendekatan Natural Language Processing (NLP). Metode yang digunakan mencakup scraping konten menggunakan library newspaper3k, analisis sentimen dengan TextBlob dan transformers, serta analisis linguistik lanjut berupa POS tagging dan Named Entity Recognition menggunakan spaCy-UDPipe. Dataset dikembangkan dari 222 artikel hasil scraping yang dilabeli secara manual menjadi tiga kategori sentimen: positif, negatif, dan netral. Untuk menangani ketidakseimbangan kelas, dilakukan augmentasi data menggunakan teknik synonym replacement. Dua pendekatan pemodelan dievaluasi: Logistic Regression dengan TF-IDF dan BERT. Hasil menunjukkan bahwa model BERT memiliki akurasi lebih tinggi dalam klasifikasi sentimen dibanding metode klasik. Temuan ini memberikan wawasan objektif terhadap narasi media seputar Danantara dan memperkuat penggunaan NLP untuk mendukung analisis kebijakan publik secara real-time.

Kata kunci: Analisis Sentimen, Danantara, POS tagging, NER, BERT, TF-IDF, NLP

# 1. PENDAHULUAN

## 1.1 Latar Belakang

Dalam era digital yang berkembang pesat dan di tengah memanasnya dinamika politik saat ini, arus informasi mengalir dengan sangat cepat, terutama melalui media daring seperti portal berita. Banyaknya sumber informasi yang beredar memunculkan tantangan baru, termasuk perbedaan respons dan interpretasi publik terhadap isu-isu yang diberitakan. Salah satu tantangan utama adalah bagaimana memahami persepsi publik terhadap suatu kebijakan melalui pemberitaan yang tersebar di berbagai media. Kebijakan Danantara sebagai salah satu kebijakan strategis menjadi sorotan di berbagai platform berita. Berbagai artikel yang membahas kebijakan ini tidak hanya menyampaikan fakta, tetapi juga memuat opini, kritik, maupun dukungan yang narasinya menjadi ruang demokrasi baru bagi masyarakat untuk berpartisipasi dalam kebijakan dan keputusan politik yang dihasilkan pemerintah (Zempi CN et. al, 2023).

Survei dan audiensi publik tidak selalu dapat menjangkau seluruh lapisan masyarakat dan mungkin tidak mewakili opini publik secara keseluruhan. Oleh karena itu, dibutuhkan metode baru untuk menganalisis sentimen masyarakat terhadap kebijakan pemerintah secara real-time dan komprehensif (Abdurrohim & Rahman, 2024). Untuk memahami persepsi yang terkandung dalam berita, diperlukan analisis mendalam terhadap isi artikel tersebut. Salah satu pendekatan yang dapat dilakukan adalah analisis sentimen untuk mengidentifikasi kecenderungan opini dalam teks. Selain itu, analisis lanjutan seperti *Part-of-Speech* (POS) tagging dan *Named Entity Recognition* (NER) juga penting dilakukan guna mengenali struktur kata dan entitas yang terdapat dalam teks berita. POS *tagging* membantu mengetahui peran atau fungsi kata dalam kalimat, sedangkan NER berperan dalam menemukan entitas penting seperti nama tokoh, organisasi, maupun lokasi yang disebutkan. Informasi ini sangat berguna untuk memetakan isu, aktor, dan konteks kebijakan yang sedang dibahas.

Dengan kemampuannya mengolah data teks secara efisien, NLP dapat memberikan wawasan yang berharga bagi pembuat kebijakan dalam merancang dan memperbaiki kebijakan yang sesuai dengan kebutuhan dan aspirasi masyarakat (Abdurrohim & Rahman, 2024). Untuk mendukung proses analisis ini, digunakan *library newspaper3k*, yaitu sebuah pustaka Python yang mampu melakukan ekstraksi konten berita secara otomatis, mulai dari pengambilan tautan artikel, metadata, hingga isi teks utama. Penggunaan *library* ini mempercepat dan mempermudah proses pengumpulan data dari berbagai sumber berita daring. Melalui penerapan analisis sentimen, POS tagging, dan NER terhadap artikel berita mengenai kebijakan Danantara, diharapkan dapat diperoleh gambaran yang lebih objektif terkait persepsi media, aktor yang terlibat, serta membantu proses pengambilan keputusan atau evaluasi terhadap kebijakan tersebut di masa mendatang.

## 1.2 Rumusan Masalah

1. Bagaimana tren pemberitaan mengenai program Danantara ditinjau dari sumber portal berita dan periode waktu tertentu?
2. Apa saja kata atau istilah yang paling sering muncul dalam berita tentang program Danantara?
3. Bagaimana karakteristik awal sentimen dalam artikel berita program Danantara berdasarkan analisis polarity dan subjectivity menggunakan textblob?
4. Bagaimana hasil POS dan NER ini membantu mendukung analisis linguistik atau pemrosesan data teks?
5. Bagaimana performa model BERT dalam melakukan klasifikasi sentimen pada dataset artikel Danantara?
6. Sejauh mana peningkatan akurasi klasifikasi sentimen yang dicapai oleh model BERT dibandingkan dengan metode tradisional Logistic Regression berbasis TF-IDF?

## 1.3 Tujuan Penelitian

1. Menganalisis tren pemberitaan mengenai program Danantara berdasarkan sumber portal berita dan periode waktu tertentu.
2. Mengidentifikasi kata atau istilah yang paling sering muncul dalam artikel berita terkait program Danantara.
3. Mendeskripsikan karakteristik awal sentimen dalam artikel berita program Danantara berdasarkan analisis polarity dan subjectivity menggunakan TextBlob.
4. Menganalisis hasil Part-of-Speech (POS) dan Named Entity Recognition (NER) untuk mendukung interpretasi linguistik dan pemrosesan data teks berita.
5. Mengevaluasi performa model BERT dalam melakukan klasifikasi sentimen terhadap artikel berita program Danantara.
6. Membandingkan akurasi klasifikasi sentimen antara model BERT dan metode tradisional Logistic Regression yang menggunakan representasi teks TF-IDF.

# 2. TINJAUAN PUSTAKA

## 2.1 Scraping

*Web scraping* adalah proses ekstraksi data atau informasi dari berbagai situs web secara otomatis (Thota & Elmasri, 2021). Proses ini memungkinkan pengumpulan data dalam jumlah besar tanpa perlu melakukan penyalinan manual. Web scraping juga dikenal dengan istilah lain seperti *web data extraction*, *web harvesting,* atau *screen scraping*. Secara umum, *scraping* dilakukan untuk memperoleh data yang umumnya bersifat *unstructured* di halaman web, kemudian diubah menjadi format *structured* seperti CSV, *spreadsheet*, atau database agar lebih mudah dianalisis. Menurut Thota & Elmasri

(2021), proses *web scraping* terdiri dari tiga tahapan utama, yaitu (1) *Fetching*, mengambil halaman web dengan mengirimkan permintaan HTTP ke server dan menerima dokumen HTML sebagai respons. (2) *Extracting*, mengekstrak informasi yang relevan dari dokumen HTML dengan teknik seperti HTML parsing, DOM parsing, XPath, atau pattern matching. (3) *Transforming*, mengubah data yang telah diekstrak menjadi format terstruktur seperti CSV atau spreadsheet untuk penyimpanan dan analisis lebih lanjut.

## 2.2 Library newspaper3k

*Newspaper3k* adalah sebuah pustaka Python yang dirancang untuk melakukan ekstraksi otomatis terhadap konten artikel berita dari situs web, termasuk mengambil teks artikel, judul, metadata, tanggal publikasi, dan elemen penting lainnya (Aniketh et al., 2025). Pustaka ini memanfaatkan teknik parsing HTML dan algoritma ekstraksi konten untuk mengidentifikasi bagian utama artikel, memisahkannya dari elemen web lain seperti iklan, menu, atau tautan navigasi. Dengan fitur ini, *newspaper3k* mempermudah proses web scraping terhadap berita online, sehingga konten yang diekstrak lebih bersih dan terstruktur. Menurut Aniketh et al. (2025), penggunaan *newspaper3k* dalam sistem pengumpulan data berita terbukti efisien dalam mengekstrak isi artikel dari berbagai sumber portal berita dengan format HTML yang beragam. Library ini juga memiliki keunggulan dalam kecepatan pemrosesan dan kemudahan integrasi dengan pipeline analisis Natural Language Processing (NLP) lainnya. Pada penelitian ini, *newspaper3k* diimplementasikan untuk mengambil konten berita sebelum dilakukan analisis sentimen dan summarization berbasis machine learning, sehingga mendukung alur kerja sistem otomatis untuk klasifikasi dan ringkasan berita.

## 2.3 Augmentasi Data dengan Operasi *Synonim Replacement*

Augmentasi data adalah serangkaian teknik untuk meningkatkan jumlah data latih secara artifisial dengan membuat salinan data yang telah dimodifikasi atau data sintetis baru dari data yang sudah ada (Feng et al., 2021). Salah satu teknik augmentasi data yang paling umum dan mudah diimplementasikan dalam pemrosesan bahasa alami (NLP) adalah penggantian sinonim (synonym replacement). Teknik ini bertujuan untuk memperkaya variasi leksikal dalam dataset tanpa mengubah label sentimen aslinya, sehingga membantu model untuk tidak terlalu bergantung pada kata-kata kunci tertentu (overfitting) dan mampu melakukan generalisasi dengan lebih baik (Wei & Zou, 2019).

Proses penggantian sinonim bekerja dengan cara mengidentifikasi kata-kata dalam sebuah kalimat, lalu menggantinya dengan kata lain yang memiliki makna serupa. Secara umum, proses ini melibatkan tiga tahapan utama: (1) Identifikasi Kata Target, memilih sejumlah n kata secara acak dari kalimat. (2) Pencarian Sinonim, untuk setiap kata target, sistem mencari daftar sinonimnya menggunakan sumber leksikal terstruktur

seperti tesaurus atau basis data leksikal, contohnya WordNet. (3) Substitusi, memilih satu sinonim secara acak dari daftar yang ditemukan dan mengganti kata asli dalam kalimat untuk menciptakan kalimat baru yang teraugmentasi. Hasil dari proses ini adalah sebuah kalimat baru yang secara semantik setara namun berbeda secara leksikal dengan kalimat aslinya.

## 2.4 Pra-Pemrosesan Teks

Pra-pemrosesan teks adalah serangkaian langkah fundamental yang bertujuan untuk membersihkan dan mempersiapkan data teks mentah (raw text) sebelum diolah oleh model analisis sentimen. Data teks yang diambil dari sumber seperti media sosial atau ulasan produk seringkali bersifat tidak terstruktur dan mengandung banyak noise (gangguan), seperti ejaan yang tidak konsisten, singkatan, URL, dan tanda baca. Tujuan utama dari pra-pemrosesan adalah untuk mengurangi noise tersebut dan mengubah teks menjadi format yang bersih dan terstruktur, sehingga dapat meningkatkan efektivitas dan akurasi model machine learning secara signifikan (Kowsari et al., 2019). Tanpa tahap ini, model akan kesulitan untuk mempelajari pola yang relevan dan kinerjanya pun akan menurun.

Proses pra-pemrosesan teks umumnya terdiri dari beberapa tahapan utama yang dilakukan secara berurutan. Menurut Uysal & Gunal (2014), urutan dan pilihan tahapan dapat memengaruhi hasil klasifikasi secara langsung. Tahapan-tahapan yang umum digunakan adalah sebagai berikut:

1. Case Folding: Mengubah seluruh teks menjadi format huruf yang seragam, biasanya menjadi huruf kecil (lowercase). Hal ini dilakukan untuk memastikan bahwa kata yang sama (misalnya, "Suka", "suka", dan "SUKA") dianggap sebagai satu token yang identik.
2. Cleansing: Membersihkan teks dari elemen-elemen yang tidak relevan untuk analisis sentimen, seperti menghapus tag HTML, URL, angka, karakter khusus, dan tanda baca.
3. Tokenizing: Proses memecah kalimat atau dokumen menjadi unit-unit yang lebih kecil yang disebut "token", yang umumnya berupa kata-kata. Tokenisasi merupakan langkah dasar yang menjadi prasyarat untuk tahapan selanjutnya seperti stopword removal dan stemming.
4. Normalization: Menyeragamkan kata-kata yang tidak baku atau singkatan ke dalam bentuk standarnya. Misalnya, kata "yg" diubah menjadi "yang" atau "ga" menjadi "tidak". Tahap ini sangat penting untuk data teks berbahasa Indonesia yang bersifat informal.
5. Stopword Removal: Menghapus kata-kata umum (stopwords) yang sering muncul namun memiliki sedikit makna semantik, seperti "yang", "di", "dan", "adalah". Penghapusan stopwords membantu mengurangi dimensi data dan memungkinkan model untuk lebih fokus pada kata-kata yang membawa sentimen.

6. Stemming atau Lemmatization: Mengubah kata-kata ke bentuk dasarnya. Stemming adalah proses menghilangkan imbuhan (awalan dan akhiran) dari sebuah kata untuk mendapatkan kata dasar (stem), namun hasilnya tidak selalu merupakan kata yang valid dalam kamus. Di sisi lain, lemmatization menggunakan analisis morfologis dan kamus untuk mengubah kata ke bentuk dasarnya yang valid (lemma).

## 2.5 Exploratory Data Analysis

Exploratory Data Analysis (EDA) atau Analisis Data Eksploratif adalah suatu pendekatan untuk menganalisis dan meringkas kumpulan data guna mengungkap karakteristik utama, menemukan pola, mengidentifikasi anomali, dan menguji hipotesis awal. Dalam konteks analisis sentimen, EDA menjadi langkah krusial untuk memahami secara mendalam isi, struktur, dan distribusi data teks sebelum membangun model. Tujuan utamanya adalah untuk mendapatkan wawasan intuitif mengenai data, seperti topik yang sering dibicarakan, kata-kata kunci yang dominan untuk setiap sentimen, dan keseimbangan kelas sentimen dalam dataset (Tukey, 1977). Proses ini membantu peneliti dalam membuat keputusan yang lebih baik mengenai strategi pra-pemrosesan dan pemilihan model yang akan digunakan.

## 2.6 Sentiment Polarity dan Sentiment Subjectivity

Analisis sentimen merupakan teknik yang digunakan untuk mengidentifikasi dan mengekstrak opini atau emosi yang terkandung dalam suatu teks. Salah satu aspek penting dalam analisis sentimen adalah *sentiment polarity*, yaitu ukuran yang digunakan untuk menentukan arah atau sifat emosi dalam teks, apakah bersifat positif, negatif, atau netral (Nafees et al., 2021). Misalnya, pernyataan seperti “Aplikasi ini sangat bermanfaat” memiliki *polarity* positif, sedangkan “Saya kecewa dengan layanan aplikasi ini” memiliki *polarity* negatif. Penentuan *polarity* memungkinkan peneliti memahami kecenderungan sikap atau persepsi publik terhadap suatu isu, produk, atau kebijakan.

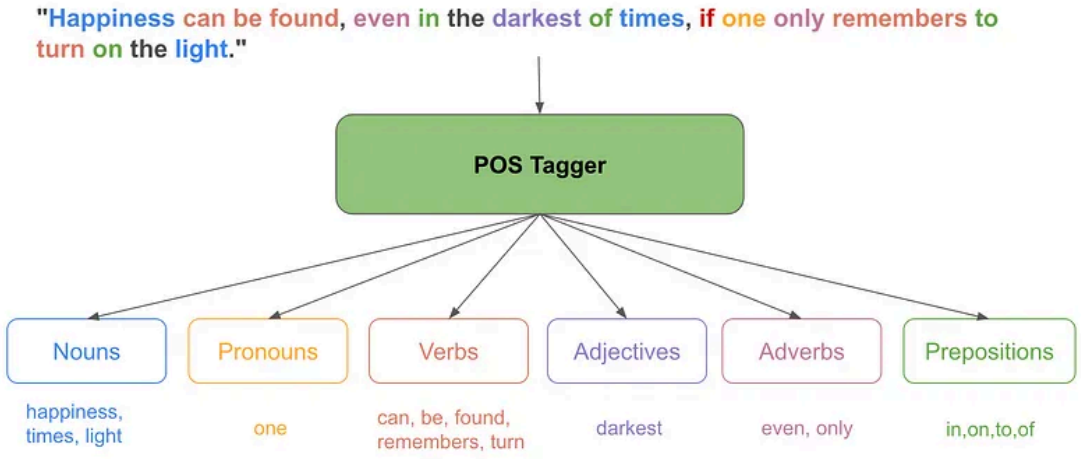
Selain *polarity*, analisis sentimen juga mempertimbangkan aspek *sentiment subjectivity*, yaitu ukuran yang menunjukkan sejauh mana suatu teks mengandung opini pribadi (subjektif) atau hanya menyampaikan informasi faktual (objektif) (Pang & Lee, 2008). Semakin subjektif suatu teks, semakin banyak opini, perasaan, atau pandangan pribadi yang diungkapkan; sedangkan semakin objektif, teks tersebut lebih berfokus pada penyampaian fakta tanpa opini. Contohnya, kalimat “Aplikasi ini dirilis pada tahun 2021” bersifat objektif, sedangkan “Saya merasa aplikasi ini kurang bermanfaat” bersifat subjektif.

Kedua konsep ini saling melengkapi dalam analisis sentimen. Sentiment polarity menjawab pertanyaan “apakah sikapnya positif, negatif, atau netral?”, sedangkan sentiment subjectivity menjawab “apakah pernyataannya berupa opini atau fakta?”. Dengan menganalisis *polarity* dan *subjectivity* secara bersamaan, peneliti dapat

memperoleh gambaran yang lebih komprehensif mengenai nada emosional dan tingkat keberpihakan suatu teks, baik dalam konteks media sosial, ulasan produk, maupun pemberitaan media.

## 2.7 Part-of-Speech

Part-of-Speech (POS) *tagging* atau penandaan kelas kata adalah proses otomatis untuk menetapkan kategori gramatikal seperti kata benda, kata kerja, kata sifat, atau kata keterangan pada setiap kata dalam suatu kalimat (Chiche & Yitagesu, 2022). POS *tagging* menjadi salah satu komponen penting dalam bidang *Natural Language Processing* (NLP) karena berfungsi sebagai dasar bagi berbagai aplikasi seperti *machine translation*, *question answering,* dan *information extraction*.




*Source: Medium*
*https://medium.com/in-pursuit-of-artificial-intelligence/named-entity-recognition-using-spacy-ner-da6eebd3d08*

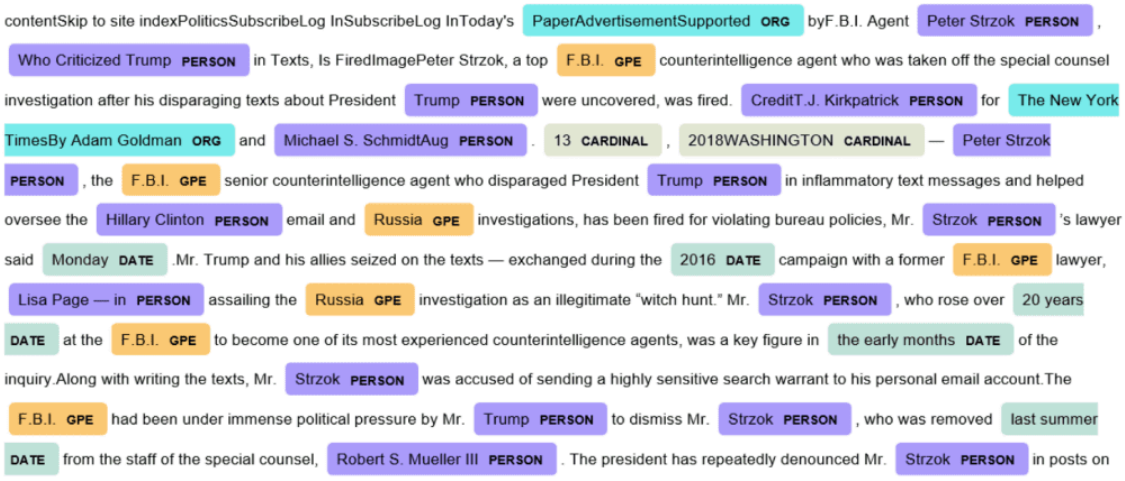
**Gambar 2.1** POS Tagging Example

Penandaan ini tidak hanya berdasarkan bentuk kata, tetapi juga mempertimbangkan konteks kemunculan kata dalam kalimat. Proses ini mengatasi ambiguitas kata yang memiliki kelas kata berbeda tergantung penggunaannya. Implementasi metode *ML/DL* untuk POS *tagging* dinilai mampu meningkatkan akurasi dan mengurangi kesalahan penandaan, meskipun membutuhkan sumber daya komputasi yang lebih besar. Oleh karena itu, perkembangan POS *tagging* berbasis AI menjadi tren penting dalam NLP untuk mendukung berbagai aplikasi analisis teks, termasuk dalam analisis artikel berita dan kebijakan.

## 2.8 Named Entity Recognition

Named Entity Recognition (NER) adalah salah satu teknik penting dalam Natural Language Processing (NLP) yang digunakan untuk mengidentifikasi dan

mengklasifikasikan entitas tertentu seperti nama orang, organisasi, lokasi, tanggal, atau entitas lain dari suatu teks (Naseer et al., 2021). Proses ini bertujuan untvcbuk mengekstrak informasi yang bermakna dari data teks tidak terstruktur sehingga dapat diolah lebih lanjut untuk berbagai keperluan, termasuk analisis informasi, sistem tanya jawab, dan ekstraksi data. Sebagai contoh, kalimat "John membeli 500 saham di Acme Corp. pada tahun 2016" akan diidentifikasi menjadi entitas "John [Person]", "Acme Corp [Organization]", dan "2016 [Time]".




*Source: Wisecube AI https://www.wisecube.ai/blog/named-entity-recognition-ner-with-python/*

**Gambar 2.2** NER Tagging Example

Dalam aplikasinya, berbagai library populer seperti *spaCy*, *StanfordNLP*, *TensorFlow*, dan *Apache OpenNLP* telah menyediakan model NER bawaan yang mendukung berbagai bahasa dan domain. Hasil penelitian menunjukkan bahwa *spaCy* memiliki performa terbaik dalam hal akurasi dan kecepatan prediksi dibandingkan *library* lainnya (Naseer et al., 2021). Oleh karena itu, NER menjadi salah satu komponen esensial dalam analisis teks modern, termasuk dalam konteks analisis berita, opini publik, dan dokumentasi kebijakan.

## 2.9 TF-IDF Vectorizer

Analisis sentimen merupakan teknik yang digunakan untuk mengidentifikasi dan mengekstrak opini atau emosi yang terkandung dalam suatu teks. Salah satu aspek penting dalam analisis sentimen adalah *sentiment polarity*, yaitu ukuran yang digunakan untuk menentukan arah atau sifat emosi dalam teks, apakah bersifat positif, negatif, atau netral (Nafees et al., 2021). Misalnya, pernyataan seperti "Aplikasi ini sangat bermanfaat" memiliki *polarity* positif, sedangkan "Saya kecewa dengan layanan aplikasi ini" memiliki *polarity* negatif. Penentuan *polarity* memungkinkan peneliti memahami kecenderungan sikap atau persepsi publik terhadap suatu isu, produk, atau kebijakan.

## 2.10 Aplikasi Model Logistic Regression dalam Analisis Sentimen

Regresi Logistik (Logistic Regression) adalah sebuah model statistik fundamental dalam machine learning yang digunakan untuk tugas klasifikasi, khususnya klasifikasi biner. Meskipun merupakan model linear yang sederhana, Regresi Logistik telah terbukti menjadi baseline yang sangat kuat dan efisien untuk tugas klasifikasi teks, termasuk analisis sentimen (Wang & Manning, 2012). Tujuan utama model ini adalah untuk memprediksi probabilitas sebuah data input (dalam hal ini, sebuah teks) masuk ke dalam kategori tertentu (misalnya, 'positif' atau 'negatif') dengan menggunakan fungsi logistik atau sigmoid.

Karena Regresi Logistik bekerja dengan input numerik, data teks yang tidak terstruktur harus terlebih dahulu diubah menjadi representasi vektor numerik. Proses ini disebut sebagai ekstraksi fitur (feature extraction). Salah satu pendekatan yang paling umum dan efektif untuk digunakan bersama Regresi Logistik adalah TF-IDF (Term Frequency-Inverse Document Frequency). Menurut Aggarwal & Zhai (2012), proses penerapan Regresi Logistik dengan TF-IDF untuk analisis sentimen terdiri dari beberapa langkah utama: (1) Pembuatan Vektor TF-IDF, setiap dokumen teks diubah menjadi sebuah vektor numerik. Setiap elemen dalam vektor ini merepresentasikan sebuah kata dari keseluruhan kosakata (vocabulary), dan nilainya adalah bobot TF-IDF kata tersebut. Bobot ini tinggi jika sebuah kata sering muncul dalam dokumen tersebut (TF) tetapi jarang muncul di dokumen lain dalam keseluruhan korpus (IDF), sehingga menyoroti kata-kata yang penting dan diskriminatif. (2) Pelatihan Model, model Regresi Logistik dilatih menggunakan pasangan vektor TF-IDF dan label sentimennya. Selama pelatihan, model akan mempelajari sebuah bobot (koefisien) untuk setiap kata, yang mengindikasikan seberapa kuat kontribusi kata tersebut terhadap sentimen positif atau negatif. (3) Prediksi, untuk teks baru, model akan menghitung skor probabilitas berdasarkan bobot yang telah dipelajari dan kemudian mengklasifikasikannya ke dalam kelas sentimen dengan probabilitas tertinggi.

Salah satu keunggulan utama Regresi Logistik dalam analisis sentimen adalah interpretabilitasnya. Bobot yang dipelajari untuk setiap kata dapat diperiksa secara langsung untuk memahami fitur mana yang paling berpengaruh terhadap keputusan model. Hal ini menjadikannya pilihan yang sangat baik untuk analisis awal dan sebagai model dasar (baseline model) untuk dibandingkan dengan model yang lebih kompleks seperti BERT.

## 2.11 Aplikasi Transformer dan BERT dalam Analisis Sentimen

Penerapan arsitektur Transformer, khususnya model BERT, telah menjadi standar emas (gold standard) dalam berbagai tugas NLP, termasuk analisis sentimen. Kemampuan BERT untuk memahami konteks secara mendalam menjadikannya sangat

efektif dalam membedakan nuansa sentimen yang seringkali bergantung pada keseluruhan susunan kalimat.

Proses penerapan BERT untuk analisis sentimen umumnya dilakukan melalui mekanisme fine-tuning (penyesuaian). Menurut Devlin et al. (2019), setelah melalui tahap pra-pelatihan pada korpus data yang sangat besar, model BERT yang sudah memiliki pemahaman bahasa yang kaya dapat diadaptasi untuk tugas klasifikasi sentimen. Secara praktis, sebuah lapisan klasifikasi sederhana (misalnya, sebuah fully-connected layer dengan fungsi aktivasi softmax) ditambahkan di atas output dari token [CLS] BERT. Token [CLS] ini dirancang secara khusus untuk mengagregasi representasi semantik dari keseluruhan kalimat input menjadi satu vektor tunggal. Selama proses fine-tuning, parameter dari lapisan klasifikasi baru dan parameter dari model BERT itu sendiri akan sedikit diperbarui menggunakan dataset analisis sentimen yang berlabel (misalnya, ulasan positif/negatif). Proses ini secara efektif "mengajarkan" model untuk memetakan pemahaman bahasanya yang umum ke tugas spesifik untuk mengklasifikasikan sentimen.

Berbagai studi telah menunjukkan keunggulan BERT dalam analisis sentimen. Sebagai contoh, sebuah penelitian oleh Sun et al. (2019) mendemonstrasikan bagaimana strategi fine-tuning yang cermat pada BERT dapat menghasilkan kinerja canggih (state-of-the-art) pada beberapa dataset klasifikasi teks, melampaui model-model deep learning sebelumnya seperti LSTM dan CNN. Keberhasilan ini disebabkan oleh kemampuan BERT dalam menangkap hubungan kompleks antar kata, seperti dalam kalimat sarkasme atau kalimat yang mengandung negasi, di mana model-model tradisional seringkali gagal. Dengan demikian, literatur secara konsisten mengonfirmasi bahwa penggunaan model berbasis Transformer seperti BERT memberikan peningkatan performa yang signifikan untuk tugas analisis sentimen.

## 2.12 Ulasan Kelebihan BERT Dibandingkan Model Tradisional

Keunggulan performa BERT dibandingkan model-model tradisional (misalnya, model yang menggunakan Word2Vec atau GloVe, serta arsitektur LSTM/RNN) dapat diatribusikan pada beberapa kelebihan konseptual yang fundamental.

1. Pemahaman Konteks Dua Arah (Deeply Bidirectional Context)
   Kelebihan utama BERT adalah kemampuannya memahami konteks secara dua arah. Model tradisional seperti Word2Vec menghasilkan embedding statis, di mana sebuah kata memiliki representasi vektor yang sama terlepas dari konteks kalimatnya. Model sekuensial seperti LSTM memang memproses konteks, namun bersifat searah (kiri-ke-kanan). Meskipun varian Bi-LSTM dapat memproses dari dua arah, ia melakukannya dengan melatih dua LSTM secara terpisah dan hanya menggabungkan hasilnya di akhir. Sebaliknya, melalui mekanisme Masked Language Model (MLM), BERT belajar untuk memahami representasi sebuah kata dengan secara bersamaan

mempertimbangkan seluruh kata di sisi kiri dan kanannya dalam satu model yang terintegrasi. Hal ini menghasilkan pemahaman kontekstual yang jauh lebih kaya dan mendalam.

2. Kekuatan dari Pengetahuan Pra-Pelatihan (Pre-trained Knowledge)
   BERT memanfaatkan kekuatan transfer learning secara maksimal. Model ini dilatih terlebih dahulu (pre-trained) pada korpus data yang sangat besar (misalnya, Wikipedia dan BooksCorpus) sebelum digunakan untuk tugas spesifik. Proses pra-pelatihan ini membekali BERT dengan pengetahuan linguistik yang luas, mencakup tata bahasa, sintaksis, dan hubungan semantik antar kata. Ketika model tradisional seperti LSTM dilatih dari awal (from scratch), ia hanya belajar dari dataset spesifik tugas yang seringkali terbatas ukurannya. Bahkan jika menggunakan word embedding statis yang sudah dilatih sebelumnya (seperti Word2Vec), pengetahuan yang ditransfer hanya berada pada level kata. BERT mentransfer pemahaman bahasa yang jauh lebih komprehensif, memungkinkan model untuk mencapai performa tinggi bahkan dengan data latih tugas spesifik yang relatif sedikit.
3. Efisiensi Fine-Tuning untuk Berbagai Tugas
   Arsitektur BERT dirancang untuk menjadi serbaguna. Model dasar yang sama dapat diadaptasi untuk berbagai tugas NLP—mulai dari analisis sentimen, question answering, hingga named entity recognition—hanya dengan menambahkan dan melatih satu lapisan output kecil yang sesuai. Ini sangat kontras dengan pendekatan tradisional di mana arsitektur model yang kompleks seringkali perlu dirancang secara spesifik untuk setiap tugas yang berbeda. Fleksibilitas ini membuat BERT menjadi solusi yang sangat efisien dan kuat untuk berbagai macam permasalahan NLP.

# 3. DATA ACQUISITION

## 3.1 Scrapping Link Artikel Berita

Pada tahap akuisisi data, proses pengumpulan dilakukan dengan cara manual crawling terhadap artikel berita yang relevan. Hal ini dilakukan dengan menelusuri dan mencari artikel berita terkait kebijakan Danantara secara langsung melalui mesin pencari dan portal berita daring. Setiap artikel yang ditemukan kemudian didokumentasikan ke dalam sebuah file spreadsheet (SPS) sebagai dataset awal. Proses manual *crawling* ini dipilih karena sifat spesifik topik kebijakan Danantara yang belum tersedia dalam bentuk dataset siap pakai di platform berita daring atau API publik. Dengan metode ini, peneliti memiliki kontrol penuh terhadap kualitas, relevansi, dan keberagaman sumber berita yang digunakan dalam penelitian. Dataset hasil akuisisi manual ini selanjutnya digunakan sebagai input dalam tahap ekstraksi konten dan analisis lanjutan.

## 3.2 Pendataan *tag* Artikel Berita

Setelah proses pengumpulan link berita selesai dilakukan, langkah berikutnya adalah pendataan tag atau kategori artikel. Proses ini dilakukan secara manual dengan membaca dan menelaah konten masing-masing artikel untuk menentukan kategori yang sesuai. Setiap artikel diberi tag berdasarkan lima kategori yang telah ditetapkan, yaitu: (1) *Economy*, (2) *Local News*, (3) *Foreign News*, (4) *Opinion*, dan (5) *Academic*. Penentuan tag dilakukan dengan mempertimbangkan tema utama yang diangkat dalam artikel, berdasarkan topik dominan, penggunaan istilah kunci, serta konteks isi berita. Sebagai contoh, artikel yang membahas kebijakan ekonomi Danantara dikategorikan ke dalam tag *Economy*, sementara artikel yang menyoroti reaksi masyarakat lokal terhadap kebijakan tersebut dimasukkan ke dalam *Local News*. Pendataan tag ini dicatat langsung dalam kolom tag pada dataset spreadsheet yang telah dibuat sebelumnya, berdampingan dengan kolom link artikel. Proses kategorisasi manual dipilih untuk memastikan akurasi klasifikasi berita, mengingat keterbatasan otomatisasi dalam mendeteksi konteks spesifik topik kebijakan Danantara.

**Tabel 3.1** Contoh Kalimat Setiap Tag Artikel Berita

| Tag | Contoh Kalimat |
|---|---|
| Opinion | Hal ketiga yang menjadi keraguan banyak pihak terkait dengan Danantara Indonesia adalah terburu-buru untuk menciptakan superholding tanpa merancang mekanisme pengendalian internal yang memadai dan menganalisis efek pascaintegrasi akan menyebabkan kegagalan proyek besar tersebut. |
| Economy | Dengan hadirnya Danantara yang akan mengelola dana hingga Rp14 ribu triliun, maka akan menjadi penolong untuk mencapai pertumbuhan yang tinggi. |
| Foreign_News | Speaking at the Indonesia Economic Summit in Jakarta on Tuesday (18/2/25,) senior economic advisor to President Prabowo, Luhut Binsar Pandjaitan, said that an Emirati investor is looking to invest in Indonesia’s renewable energy sector by establishing a joint venture with companies under the upcoming sovereign wealth fund Danantara, according to reporting from Jakarta Globe |
| Local_News | Ketua Dewan Pengawas (Dewas) BPI Danantara yang juga Menteri BUMN, Erick Thohir, mendatangi KPK. Dia mengatakan mengunjungi KPK sebagai tindak lanjut arahan Presiden Prabowo Subianto dalam Town Hall Meeting BPI Danantara. |

## 3.3 Pendataan Label Sentimen Secara Manual

Hasil scrapping dikumpulkan dengan cermat, kemudian dibaca secara seksama dan diberi label sentimen yang mencakup kategori Positif, Negatif, dan Netral berdasarkan analisis kontennya. Label Positif mencirikan teks yang mengandung ekspresi optimisme, dukungan, atau kepuasan, seperti pujian atau umpan balik positif. Label Negatif mencerminkan teks dengan nada kekecewaan, kritik, atau ketidakpuasan, seperti keluhan atau sentimen negatif lainnya. Sementara itu, label Netral menggambarkan teks yang netral, tidak condong ke arah positif atau negatif, seperti fakta atau informasi yang disampaikan tanpa emosi.

**Tabel 3.2** Contoh Kalimat Setiap Label Artikel Berita

| Label | Contoh Kalimat |
|---|---|
| Negatif | Hal pertama yang mengkhawatirkan dari superholding Danantara Indonesia menggabungkan empat karakteristik industri yang berbeda, yaitu industri perbankan, industri pertambangan, industri migas, dan industri teknologi dengan risiko bisnis yang jelas berbeda dan tantangan setiap industri yang berbeda. |
| Positif | Kehadiran BPI Danantara disebut dapat membawa sejumlah manfaat ke pasar modal Indonesia, termasuk mendongkrak kapitalisasi pasar modal Indonesia. |
| Netral | Danantara merupakan superholding atau perusahaan induk yang mengendalikan berbagai perusahaan besar di sektor industri sekaligus manajer investasi dari tujuh BUMN untuk saat ini yaitu Bank Mandiri, Bank BRI, PLN, Pertamina, BNI, Telkom Indonesia, dan MIND, serta Indonesia Investment Authority (INA) yang didirikan oleh Presiden Joko Widodo. |

## 3.3 Scrapping Konten Artikel Berita

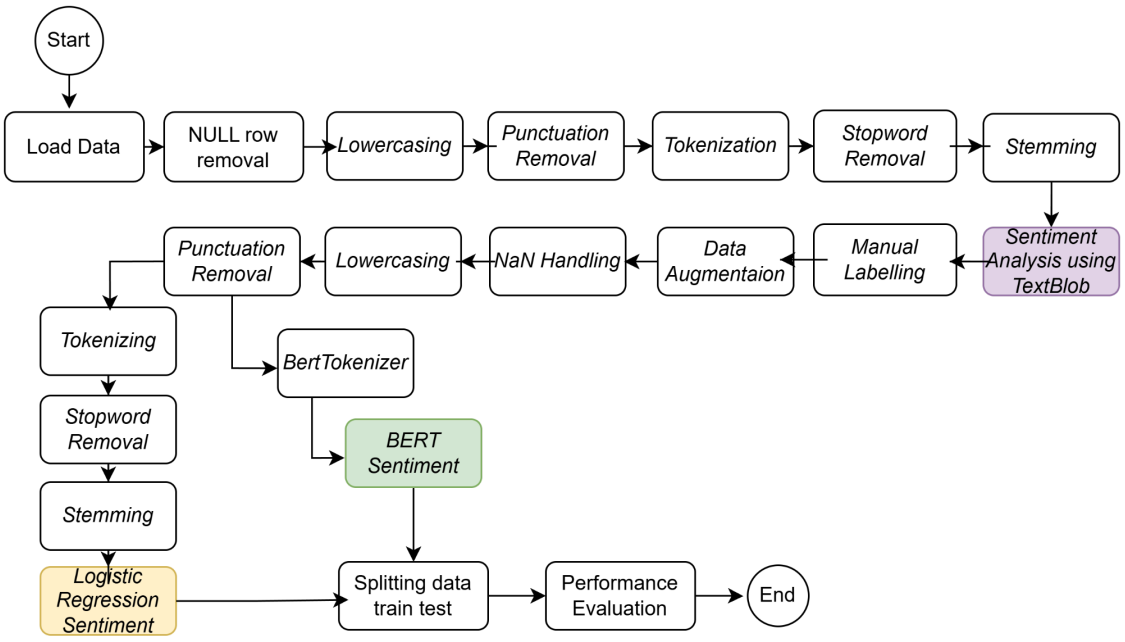
Setelah link artikel dan kategori tag terdokumentasi dalam dataset, langkah selanjutnya adalah pengambilan konten artikel secara otomatis menggunakan library *newspaper3k*. *Newspaper3k* dipilih sebagai alat ekstraksi karena kemampuannya dalam mengambil teks utama artikel, judul, tanggal publikasi, serta metadata lainnya secara cepat dan relatif bersih dari elemen-elemen non-konten seperti iklan, menu, atau tautan navigasi. Dalam tahap ini, setiap link artikel yang tercatat pada dataset diinputkan ke dalam fungsi *Article()* dari library *newspaper3k*. Proses scraping dilakukan dengan memanfaatkan metode *download()* untuk mengunduh halaman artikel dan *parse()* untuk

mengekstrak isi artikel. Hasil scraping ini kemudian disimpan dan ditambahkan ke dataset yang telah ada, sehingga setiap artikel memiliki kolom tambahan berupa judul, isi artikel, dan tanggal publikasi. Proses ini mempermudah persiapan data untuk tahap analisis lebih lanjut, seperti analisis sentimen, POS tagging, dan *Named Entity Recognition* (NER). Penggunaan *newspaper3k* dianggap efektif karena mampu mengekstrak konten dari berbagai struktur HTML portal berita dengan konsistensi yang baik, meskipun masih memerlukan pengecekan manual untuk memastikan kualitas data yang diambil. Dengan pendekatan ini, proses ekstraksi konten berita dapat dilakukan secara otomatis namun tetap terkontrol dari sisi validitas data.

# 4. METODOLOGI

## 4.1 Data Preprocessing Artikel Berita

Praproses dilakukan untuk membersihkan teks dari elemen-elemen yang tidak relevan serta mengubahnya ke dalam bentuk yang lebih terstruktur dan konsisten, sehingga data siap digunakan untuk tahap analisis lanjutan seperti analisis sentimen, frekuensi kata, *Part-of-Speech* (POS) tagging, dan *Named Entity Recognition* (NER). Seluruh tahapan praproses yang dilakukan dijelaskan secara visual melalui skema alur pada gambar berikut untuk memberikan gambaran proses secara menyeluruh dan sistematis.




**Gambar 4.1** Methodology Flowchart

Praprosesing dimulai dengan *Load Data*, yaitu memuat dataset berisi isi artikel hasil *scraping*. Selanjutnya, dilakukan *Lowercasing*, yakni mengubah seluruh huruf menjadi huruf kecil untuk menghindari duplikasi kata akibat kapitalisasi. Setelah itu, dilakukan *Punctuation Removal*, yaitu penghapusan seluruh tanda baca dan karakter non-alfabetik yang tidak relevan untuk analisis linguistik.Langkah berikutnya adalah *Tokenization*, yaitu pemecahan teks menjadi unit kata (token) agar dapat diproses lebih lanjut. Setelah token diperoleh, dilakukan *Stopword Removal*, yakni penghapusan kata-kata umum yang tidak memiliki makna signifikan secara semantik, baik dalam bahasa Indonesia maupun Inggris. Kemudian, dilakukan proses *Stemming* menggunakan *library* Sastrawi untuk mengembalikan kata ke bentuk dasarnya. Terakhir, hasil praproses disimpan dalam bentuk terstruktur berbentuk df, dan siap digunakan untuk berbagai tahap analisis lanjutan.

**1. Tahap Pra-pemrosesan Awal dan Pelabelan Data**

Tahap ini bertujuan untuk mengubah data mentah hasil *scraping* menjadi dataset yang bersih dan berlabel, sehingga siap untuk diolah lebih lanjut.

- **Pemuatan Data (Load Data)**
  Proses dimulai dengan memuat dataset awal yang berisi hasil scraping artikel berita terkait "Danantara". Dataset ini terdiri dari 312 baris data dan 8 kolom, termasuk title, content, dan url.
- **Pembersihan dan Standarisasi Teks**
  Selanjutnya, dilakukan serangkaian langkah pembersihan data klasik. Langkah NULL row removal (penghapusan baris kosong) adalah penyebab utama berkurangnya jumlah data dari 312 menjadi 222 baris. Proses dilanjutkan dengan Lowercasing (mengubah teks menjadi huruf kecil), Punctuation Removal (menghapus tanda baca), Tokenization (memecah kalimat menjadi kata), Stopword Removal (menghapus kata umum), dan Stemming (mengubah kata ke bentuk dasar). Hasil dari setiap langkah ini disimpan dalam kolom-kolom baru seperti original, tokens_awal, no_stopwords, dan stemmed.
- **Analisis Sentimen Otomatis & Pelabelan Manual**
  Setelah teks bersih, dilakukan analisis sentimen awal menggunakan TextBlob untuk menghasilkan skor polarity dan subjectivity secara otomatis untuk gambaran awal. Namun, untuk memastikan kualitas dan akurasi label yang tinggi, proses dilanjutkan dengan Pelabelan Manual. Hasil dari pelabelan manual ini menunjukkan bahwa dataset bersifat tidak seimbang (imbalanced), dengan proporsi 35 label Negatif, 49 Netral, dan 138 Positif.

Tabel 4.1 Perbandingan Kalimat di Konten Asli dan Hasil Pre-Processing

| original | tokens_awal | no_stopwords | stemmed |
|---|---|---|---|
| catatan artikel ini merupakan opini pribadi penulis dan tidak mencerminkan | catatan,artikel,ini,merupakan,opini,pribadi,penulis,dan,tidak,mencerminkan,… | catatan,artikel,opini,pribadi,penulis,mencerminkan,... | catat,artikel,opini,pribadi,tulis,cermin,panda ng,... |

## 4.2 Tahap Penyeimbangan Data (Data Balancing)

Mengetahui bahwa data tidak seimbang, langkah selanjutnya adalah menyeimbangkan distribusi kelas agar setiap sentimen memiliki jumlah sampel yang sama. Ini sangat penting untuk mencegah model menjadi bias terhad ap kelas mayoritas.

- **Augmentasi Data (Data Augmentation):** Teknik ini diterapkan untuk memperbanyak jumlah data pada kelas minoritas ('Negatif' dan 'netral'). Berdasarkan notebook Anda, proses ini kemungkinan besar melibatkan penggantian kata dengan sinonimnya untuk menciptakan kalimat baru yang maknanya tetap sama.
- **Tujuan Akhir:** Tujuan dari tahap ini adalah untuk menghasilkan dataset yang seimbang, di mana setiap kelas sentimen memiliki **100 sampel**. Dengan demikian, dataset akhir yang akan digunakan untuk melatih model berjumlah 300 baris.
- **Penanganan NaN (NaN Handling):** Ini adalah langkah pemeriksaan akhir untuk memastikan tidak ada nilai kosong yang lolos ke tahap pemodelan setelah proses augmentasi.

## 4.3 Alur Model Klasik (Logistic Regression + TF-IDF)

Setelah data seimbang, alur kerja bercabang untuk implementasi model pertama, yaitu Logistic Regression.

- **Vektorisasi TF-IDF**

```
# 4. Vektorisasi TF-IDF
vectorizer = TfidfVectorizer(max_features=5000, ngram_range=(1,2))
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)
```

**TF-IDF (Term Frequency-Inverse Document Frequency)** digunakan sebagai feature engineering yang berperan sebagai jantung dari pengolahan teks. Model machine learning tidak mengerti kata, tapi mengerti angka. Sebuah teknik statistik untuk mengukur seberapa penting sebuah kata dalam sebuah dokumen (kalimat) relatif terhadap keseluruhan koleksi dokumen (seluruh data). Kata yang sering

muncul di satu kalimat tapi jarang muncul di kalimat lain akan mendapat skor TF-IDF yang tinggi.

## 4.4 Alur Model Modern (BERT)

Model kedua adalah untuk model BERT, yang memerlukan pendekatan pra-pemrosesan yang berbeda dan lebih sederhana.

- **Pembersihan Minimal**
  Berbeda dengan alur klasik, untuk BERT hanya dilakukan pembersihan dasar seperti Lowercasing dan Punctuation Removal. Langkah Stopword Removal dan Stemming sengaja tidak dilakukan. Hal ini bertujuan untuk mempertahankan sebanyak mungkin konteks kalimat asli, karena BERT sangat bergantung pada konteks untuk memahami makna.

- **Tokenisasi Khusus (BertTokenizer)**

```
# Inisialisasi Tokenizer
tokenizer = BertTokenizer.from_pretrained('bert-base-multilingual-uncased')

# Custom Dataset (tidak ada perubahan di sini)
class SentimentDataset(Dataset):
    def __init__(self, texts, labels, tokenizer, max_len):
        self.texts = texts
        self.labels = labels
        self.tokenizer = tokenizer
        self.max_len = max_len
    def __len__(self):
        return len(self.texts)
    def __getitem__(self, item):
        text = str(self.texts[item])
        label = self.labels[item]
        encoding = self.tokenizer.encode_plus(
            text, add_special_tokens=True, max_length=self.max_len,
            return_token_type_ids=False, padding='max_length',
            return_attention_mask=True, return_tensors='pt', truncation=True
        )
        return {
            'text': text, 'input_ids': encoding['input_ids'].flatten(),
            'attention_mask': encoding['attention_mask'].flatten(),
            'labels': torch.tensor(label, dtype=torch.long)
        }
```

Langkah selanjutnya dan yang paling krusial adalah tokenisasi menggunakan BertTokenizer. Tokenizer ini tidak hanya memecah kalimat, tetapi juga menambahkan token spesial ([CLS], [SEP]), menangani kata-kata yang tidak dikenal dengan subword tokenization, dan mengubahnya menjadi format input yang spesifik untuk BERT (input_ids, attention_mask).

- **Pemodelan dan fine-tuning (BERT Sentiment)**
  Input yang telah diproses oleh BertTokenizer ini kemudian digunakan untuk proses fine-tuning pada model BERT yang sudah dilatih sebelumnya (pre-trained)

untuk tugas klasifikasi sentimen.

```
# Memuat Model
model = BertForSequenceClassification.from_pretrained('bert-base-multilingual-uncased', num_labels=len(unique_labels))
model = model.to(device)


# --- PERUBAHAN 1: Mengatur Hyperparameter Baru ---
print("\n" + "="*60)
print("             Menggunakan Hyperparameter Baru")
print("="*60)
EPOCHS = 5
LEARNING_RATE = 3e-5
print(f"Epochs: {EPOCHS}, Learning Rate: {LEARNING_RATE}")
print("="*60)

# Mengatur Optimizer dan Scheduler dengan LR baru
optimizer = AdamW(model.parameters(), lr=LEARNING_RATE)
total_steps = len(train_data_loader) * EPOCHS
scheduler = get_linear_schedule_with_warmup(optimizer, num_warmup_steps=0, num_training_steps=total_steps)
```

Fine-tuning adalah proses pelatihan lebih lanjut (penyesuaian) sebuah model machine learning yang sudah dilatih sebelumnya (pre-trained model) pada dataset yang lebih kecil dan spesifik untuk tugas tertentu. Tujuan dari fine-tuning adalah untuk mengadaptasi pengetahuan umum yang telah diperoleh model selama pelatihan awal (di dataset yang sangat besar dan beragam) agar menjadi sangat efektif untuk tugas yang lebih spesifik dan data yang lebih sempit.

## 4.5 Data splitting (train,test)

Langkah krusial selanjutnya adalah membagi dataset menjadi dua bagian terpisah: data latih (training set) dan data uji (testing set). Tujuan utama dari pembagian ini adalah untuk dapat mengevaluasi performa model secara objektif. Model akan "belajar" dari data latih, dan kemudian "diuji" menggunakan data uji yang belum pernah ia lihat sebelumnya, sama seperti seorang siswa yang belajar dari buku dan kemudian mengerjakan soal ujian.

```
# 3. Membagi Data (tetap diperlukan untuk melatih model secara benar)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=42, stratify=y)
```

Dalam penelitian ini, dataset yang telah diseimbangkan dan berisi total 300 baris data dibagi dengan proporsi 80% untuk data latih dan 20% untuk data uji. Dengan demikian, model BERT akan melalui proses fine-tuning menggunakan 240 baris data latih. Setelah proses pelatihan selesai, performa model dalam menggeneralisasi pengetahuannya akan diukur secara akurat pada 60 baris data uji yang independen.

Proses pembagian ini dilakukan menggunakan metode stratified split. Artinya, proporsi setiap kelas sentimen ('positif', 'negatif', dan 'netral') dijaga agar tetap sama baik di dalam set data latih maupun data uji. Hal ini sangat penting untuk memastikan bahwa kedua set data tersebut merupakan representasi yang adil dari keseluruhan data dan mencegah

evaluasi yang bias, terutama pada dataset yang baru saja diseimbangkan. Dengan pembagian ini, tahap evaluasi akhir menggunakan metrik akurasi, presisi, recall, dan F1-score akan memberikan gambaran yang jujur tentang seberapa baik model BERT dapat bekerja pada data baru.

**6. Melatih Model dan Evaluasi Performa**

Tahap ini mencakup proses inti dari pemodelan, yaitu pelatihan dan evaluasi performa model Logistic Regression. Pertama, model diinisialisasi dengan parameter yang telah ditentukan (solver='liblinear'), kemudian dilatih menggunakan metode .fit() pada data latih yang telah direpresentasikan sebagai vektor TF-IDF (X_train_tfidf) beserta label sentimennya (y_train). Pada fase ini, model belajar untuk memetakan pola dari fitur teks ke kelas sentimen yang sesuai. Setelah proses pelatihan selesai, model yang telah 'belajar' tersebut diuji kemampuannya dalam melakukan generalisasi pada data baru
.

```
# 5. Melatih Model Logistic Regression
model = LogisticRegression(solver='liblinear', multi_class='auto', random_state=42)
model.fit(X_train_tfidf, y_train)
```

Model diberikan data uji (X_test_tfidf) untuk menghasilkan prediksi sentimen (y_pred_test). Terakhir, prediksi ini dibandingkan dengan label asli dari data uji (y_test) menggunakan fungsi classification_report untuk menghasilkan laporan evaluasi yang komprehensif, mencakup metrik presisi, recall, f1-score, dan akurasi keseluruhan sebagai ukuran final dari performa model.

```
# Menampilkan laporan performa model dari data uji (tetap berguna untuk mengetahui seberapa baik modelnya)
print("="*60)
print("          Performa Model Logistic Regression")
print("="*60)
print(f"Melakukan evaluasi performa pada {len(X_test)} baris data uji...")
print(f"-"*60)
y_pred_test = model.predict(X_test_tfidf)
print(classification_report(y_test, y_pred_test))
```

**7. Membuat Prediksi untuk Seluruh Data dan Simpan Hasil**

Setelah model dilatih dan dievaluasi performanya pada data uji, langkah terakhir adalah mengaplikasikan model tersebut untuk menghasilkan prediksi pada keseluruhan dataset yang ada. Proses ini diawali dengan mengubah seluruh data teks menjadi representasi vektor TF-IDF menggunakan vectorizer yang telah dilatih pada data latih sebelumnya. Selanjutnya, model Logistic Regression yang sudah jadi digunakan untuk memprediksi label sentimen untuk setiap baris data. Hasil dari prediksi ini kemudian ditambahkan sebagai sebuah kolom baru bernama 'LR_label' ke dalam DataFrame utama. Akhirnya, keseluruhan DataFrame yang kini telah diperkaya dengan kolom prediksi ini

disimpan ke dalam sebuah file CSV baru (Danantara_Hasil_Prediksi_LR.csv), sehingga memungkinkan untuk analisis dan perbandingan lebih lanjut di kemudian hari tanpa perlu menjalankan ulang proses pelatihan.

```
# 6. Membuat Prediksi untuk SELURUH DATA
print("\nMelakukan prediksi untuk seluruh data...")
# Mengubah seluruh data teks menjadi vektor TF-IDF menggunakan vectorizer yang sudah dilatih
X_tfidf_full = vectorizer.transform(X)
# Melakukan prediksi
all_predictions = model.predict(X_tfidf_full)


# 7. Menambahkan Hasil Prediksi ke DataFrame
# Membuat kolom baru bernama 'LR_label' sesuai permintaan Anda
df['LR_label'] = all_predictions
print("="*60)
print("Kolom 'LR_label' berhasil ditambahkan. Jumlah kolom sekarang:", len(X))


# 8. Menyimpan Hasil ke File CSV Baru
output_filename = 'Danantara_Hasil_Prediksi_LR.csv'
df.to_csv(output_filename, index=False)

print(f"\nData beserta hasil prediksi telah berhasil disimpan ke dalam file:")
print(f"'{output_filename}'")
```

# 5. DEFINISI DATASET

## 5.1 Dataset Link Artikel Berita

Dataset link artikel berita merupakan dataset awal yang digunakan dalam penelitian ini yang merupakan hasil pendataan manual yang disimpan dalam *file* bernama Link_Scraping_Danantara.csv, dengan format Comma-Separated Values (CSV). File ini berisi daftar artikel berita yang dikumpulkan secara manual dan berkaitan dengan kebijakan Danantara. Total data pada dataset ini berjumlah 369 artikel berita, dengan rincian berdasarkan kolom tag sebagai berikut: 30 artikel berlabel opinion, 147 local_news, 3 academic, 176 economy, dan 8 foreign_news. Dataset terdiri atas dua kolom utama, yaitu:

**Tabel 5.1** Deskripsi Dataset Link Artikel Berita

| Kolom | Deskripsi |
|---|---|
| link | URL artikel berita yang telah dikurasi secara manual agar sesuai dengan topik penelitian. |

| | |
|---|---|
| tag | Klasifikasi atau kategori isi artikel berdasarkan lima jenis tag yang telah ditentukan sebelumnya, yaitu: economy, local news, foreign news, opinion, dan academic. |
| label | Klasifikasi atau kategori sentimen artikel berdasarkan tiga jenis label, yaitu positif, negatif, dan netral |

Dataset ini digunakan sebagai input awal dalam tahap scraping konten artikel berita menggunakan library newspaper3k sebagai sumber referensi utama dalam proses scraping konten. Data pada kolom link dimasukkan ke dalam pipeline ekstraksi konten menggunakan library newspaper3k. Data pada kolom tag digunakan sebagai label klasifikasi awal untuk keperluan analisis tren, visualisasi distribusi topik, dan evaluasi hasil analisis sentimen berdasarkan kategori. Selain itu, kolom label digunakan untuk mengevaluasi kinerja model analisis sentimen yang digunakan pada penelitian ini, memungkinkan perhitungan metrik seperti akurasi dan indikator kinerja relevan lainnya.

## 5.2 Dataset Konten Artikel Berita

Dataset hasil scraping ini disimpan dalam file bernama Scraping_Danantara.csv, yang merupakan keluaran dari proses ekstraksi konten berita menggunakan library newspaper3k. Dataset ini diperoleh dengan memproses daftar tautan artikel yang sebelumnya dikumpulkan dan disusun dalam dataset awal. Tujuan utama dari dataset ini adalah menyediakan isi teks artikel secara lengkap beserta metadata penting untuk keperluan analisis teks lanjutan seperti analisis sentimen, frekuensi kata, POS tagging, dan Named Entity Recognition (NER).

Dari total tautan yang ada, sebanyak 222 artikel berhasil di-scraping. Rincian jumlah artikel yang berhasil di-scraping berdasarkan kolom tag adalah sebagai berikut: 23 opinion, 102 economy, 91 local_news, 4 foreign_news, dan 2 academic. Sementara itu, berdasarkan kolom label, rinciannya adalah: 35 negatif, 49 netral, dan 138 positif.

**Tabel 5.2** Deskripsi Dataset Konten Artikel Berita

| Kolom | Deskripsi |
|---|---|
| title | Judul artikel berita yang diperoleh. |
| authors | Nama penulis artikel atau informasi penulis jika tersedia |
| source | Nama domain atau portal berita asal artikel |
| published_date | Tanggal dan waktu publikasi artikel. |

| | |
|---|---|
| summary | Ringkasan otomatis artikel (hasil fitur summarization dari newspaper3k). |
| content | Isi lengkap teks utama artikel berita. |
| url | Tautan sumber artikel yang diambil dari dataset awal. |
| tag | Label setiap tautan artikel (1) *Economy*, (2) *Local News*, (3) *Foreign News*, (4) *Opinion*, dan (5) *Academic*. |
| label | Kategori sentimen isi berita (1) positif, (2) negatif, (3) netral |

Dataset ini digunakan sebagai dasar utama dalam seluruh tahap analisis Natural Language Processing (NLP), di mana kolom *content* berisi isi teks utama artikel yang menjadi objek utama dalam berbagai proses seperti praproses teks (cleaning, tokenisasi, stopword removal, dan stemming), analisis sentimen (polarity dan subjectivity), ekstraksi entitas (Named Entity Recognition/NER), serta analisis linguistik lainnya seperti POS tagging dan visualisasi frekuensi kata. Dengan memanfaatkan library newspaper3k, proses ekstraksi konten dilakukan secara otomatis dan konsisten dari berbagai portal berita daring, sehingga mempercepat akuisisi data dalam skala besar sekaligus menjaga validitas dan kebersihan struktur teks yang diperoleh.

## 5.3 Dataset Hasil Augmentasi

Dataset ini merupakan hasil dari proses augmentasi data yang dilakukan setelah Dataset Konten Artikel Berita (seperti dijelaskan pada sub-bab 5.2) diperoleh. Augmentasi data dilakukan untuk menyeimbangkan jumlah sampel pada setiap kategori sentimen. Dari 222 artikel yang berhasil di-scraping, terdapat 138 berita dengan sentimen positif. Untuk menyeimbangkan dataset, 100 berita dengan sentimen positif diambil sebagai sampel. Sementara itu, berita dengan sentimen negatif (35 artikel) dan netral (49 artikel) diaugmentasi hingga masing-masing berjumlah 100 sampel. Dataset ini diperlukan karena dataset asli memiliki distribusi label sentimen yang tidak seimbang, yang berpotensi menyebabkan model klasifikasi sentimen bias terhadap kelas mayoritas.

Proses augmentasi data dilakukan dengan teknik penggantian sinonim pada kolom content artikel berita. Penggantian sinonim ini menggunakan kamus sinonim bahasa Indonesia kustom, dengan perlindungan terhadap Named Entity Recognition (NER) menggunakan model spaCy untuk menjaga integritas nama atau entitas penting dalam teks. Setelah augmentasi, dataset yang dihasilkan memiliki distribusi sentimen yang seimbang dan siap untuk tahap preprocessing teks selanjutnya.

## 5.4 Dataset Hasil Preprocessing dan Analisis Sentimen

Dataset ini merupakan hasil dari proses scraping dan praproses teks terhadap artikel-artikel berita yang membahas kebijakan Danantara. Selain itu, dataset ini juga memuat hasil analisis sentimen menggunakan library TextBlob, yang direpresentasikan dalam dua kolom yaitu polarity dan subjectivity.

**Tabel 5.3** Deskripsi Dataset Hasil Preprocessing dan Analisis Sentimen

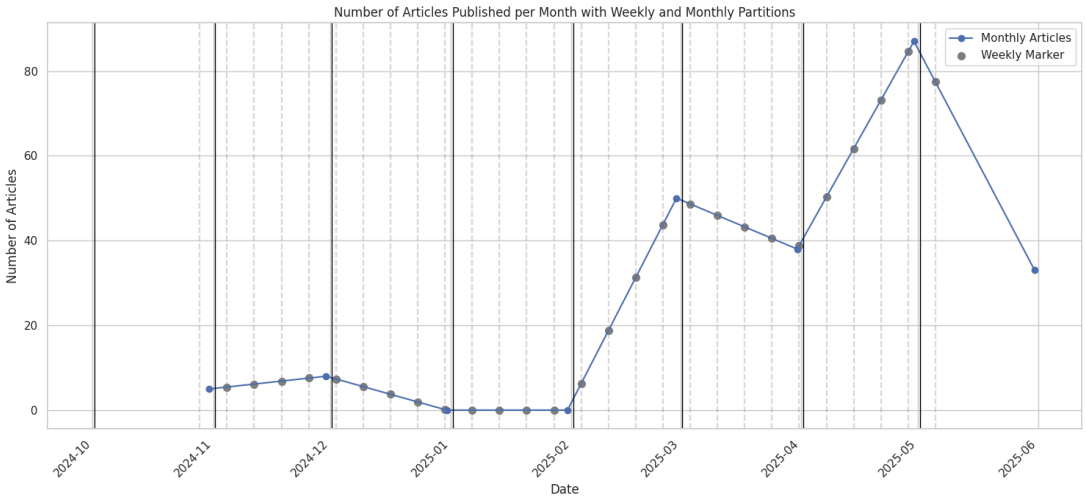
| Kolom | Deskripsi |
|---|---|
| title | Judul artikel berita yang diperoleh. |
| authors | Nama penulis artikel atau informasi penulis jika tersedia |
| source | Nama domain atau portal berita asal artikel |
| published_date | Tanggal dan waktu publikasi artikel. |
| summary | Ringkasan otomatis artikel (hasil fitur summarization dari newspaper3k). |
| content | Isi lengkap teks utama artikel berita. |
| url | Tautan sumber artikel yang diambil dari dataset awal. |
| tag | Label setiap tautan artikel (1) *Economy*, (2) *Local News*, (3) *Foreign News*, (4) *Opinion*, dan (5) *Academic*. |
| label | Kategori sentimen isi berita (1) positif, (2) negatif, (3) netral |
| polarity | Skor sentimen (-1 hingga 1) untuk menunjukkan apakah teks bernada negatif/positif |
| subjectivity | Skor subjektivitas (0 hingga 1) untuk menunjukkan apakah teks bersifat opini atau fakta |
| original | Teks artikel yang telah dibersihkan dan dikonversi ke huruf kecil |
| tokens_awal | Tokenisasi awal sebelum penghapusan stopword |
| no_stopwords | Tokenisasi setelah penghapusan kata umum yang tidak informatif (stopword) |
| stemmed | Token hasil stemming ke bentuk kata dasar menggunakan Sastrawi |

Nilai polarity menunjukkan kecenderungan opini dalam teks (dari -1 untuk sentimen negatif hingga +1 untuk sentimen positif), sedangkan subjectivity mengukur tingkat subjektivitas teks (dari 0 sebagai objektif hingga 1 sebagai sangat subjektif). Dataset disimpan dalam file berformat CSV dengan nama Danantara_Preprocessed.csv, yang berisi total 13 kolom yang mencakup informasi metadata, isi artikel, hasil analisis sentimen, dan hasil praproses.

# 6. HASIL

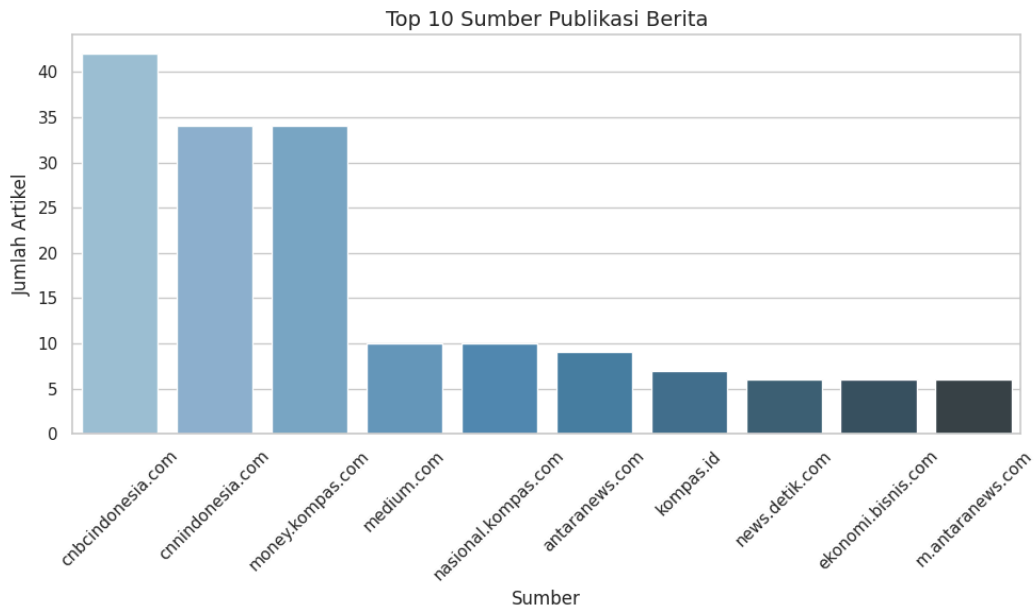
## 6.1 Analisis Tren Artikel

Berdasarkan hasil proses scrapping isi konten dari link artikel yang telah dikumpulkan menggunakan library newspaper3k, diperoleh sebanyak 222- artikel yang berhasil diambil isi konten dan publish date dari total 369 link artikel yang telah dikumpulkan. Dengan grafik distribusi jumlah artikel per bulan seperti pada gambar 6.1



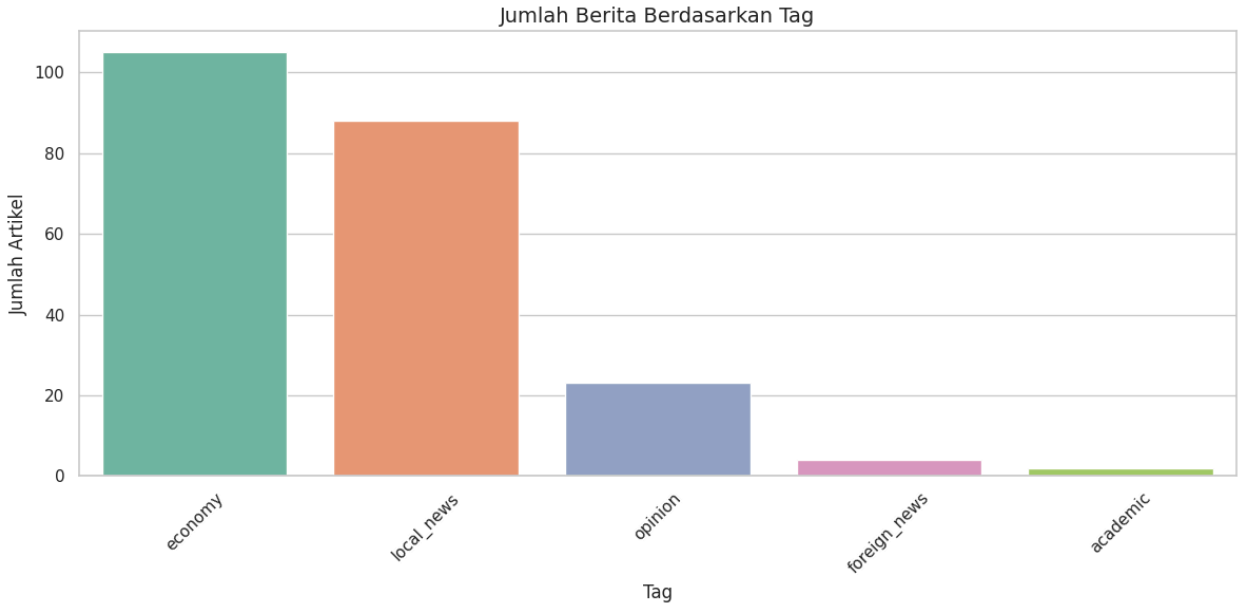

**Gambar 6.1** Jumlah Artikel yang Diterbitkan tiap Bulan

Gambar 6.1 menunjukkan tren jumlah artikel yang diterbitkan tiap bulan dari Oktober 2024 hingga Mei 2025. Dari grafik tersebut terlihat peningkatan signifikan pada bulan februari dengan total 52 artikel, hal ini mungkin terjadi karena bertepatan dengan pengumuman resmi dari Presiden Prabowo Subianto tentang peluncuran danantara. Dilanjutkan dengan penurunan dan kenaikan kembali sampai di titik tertinggi pada bulan April, dari sebelumnya sebanyak 38 artikel pada bulan maret menjadi 89 artikel pada bulan april. Peningkatan ini kemungkinan terjadi karena bertepatan dengan pengumuman kerjasama antara Danantara dan Qatar Investment Authority yang diumumkan pada tanggal 15 April 2025.




**Gambar 6.2** Top 10 Sumber Publikasi Berita

Gambar 6.2 menunjukkan 10 sumber berita dengan jumlah artikel paling banyak pada dataset. Dapat dilihat bahwa artikel paling banyak berasal dari cnbcindonesia.com, diikuti dengan cnnindonesia.com, money.kompas.com, dan sisanya tersebar di beberapa media lain, seperti medium.com, nasional.kompas.com, dan antaranews.com.




**Gambar 6.3** Persebaran Artikel berdasarkan Tag

Berdasarkan gambar 6.3 terlihat bahwa pemberitaan tentang danantara sangat didominasi oleh aspek ekonomi dan lokal. Hal ini dapat disebabkan oleh Danantara sendiri

yang merupakan badan pengelolaan kekayaan negara dan dampaknya terhadap berbagai sektor strategis di indonesia. Jumlah berita yang rendah pada kategori “foreign_news” dan “academic” menunjukkan potensi pengembangan lebih lanjut dalam kajian akademik dan minat media internasional terhadap isu ini.

## 6.2 Analisis Frekuensi Kata

Data yang telah melalui proses preprocessing selanjutnya dianalisis untuk frekuensi kata yang muncul. Berikut merupakan tabel yang menampilkan 10 kata dengan frekuensi kemunculan terbanyak pada seluruh artikel terkait danantara.

**Tabel 6.1** Top 15 Frekuensi data Muncul

| No | Kata | Frekuensi |
|---|---|---|
| 1 | danantara | 2430 |
| 2 | indonesia | 985 |
| 3 | bumn | 967 |
| 4 | investasi | 881 |
| 5 | negara | 689 |
| 6 | prabowo | 562 |
| 7 | aset | 522 |
| 8 | presiden | 422 |
| 9 | badan | 403 |
| 10 | bpi | 386 |
| 11 | rp | 385 |
| 12 | triliun | 364 |
| 13 | ekonomi | 352 |
| 14 | perusahaan | 333 |
| 15 | pemerintah | 319 |

Pada Tabel 6.1 dapat dilihat bahwa kata terbanyak yang ada pada artikel adalah danantara dengan frekuensi kemunculan sebanyak 2430 kata, hal ini sesuai dengan topik

utama dalam analisis. Kata lain yang juga terlihat cukup menonjol adalah indonesia, bumn, investasi, dan prabowo, yang dapat menandakan isu-isu terkait kebijakan, aktor politik, serta entitas ekonomi yang terlibat dalam pemberitaan terkait danantara.

Analisis dilanjutkan secara visual dengan menggunakan wordcloud terpisah berdasarkan sentimen positif dan negatif. Tujuannya adalah untuk mengetahui perbedaan narasi atau fokus kata dalam artikel dengan sentimen positif dan negatif.


**Gambar 6.4** Wordcloud Sentimen Negatif

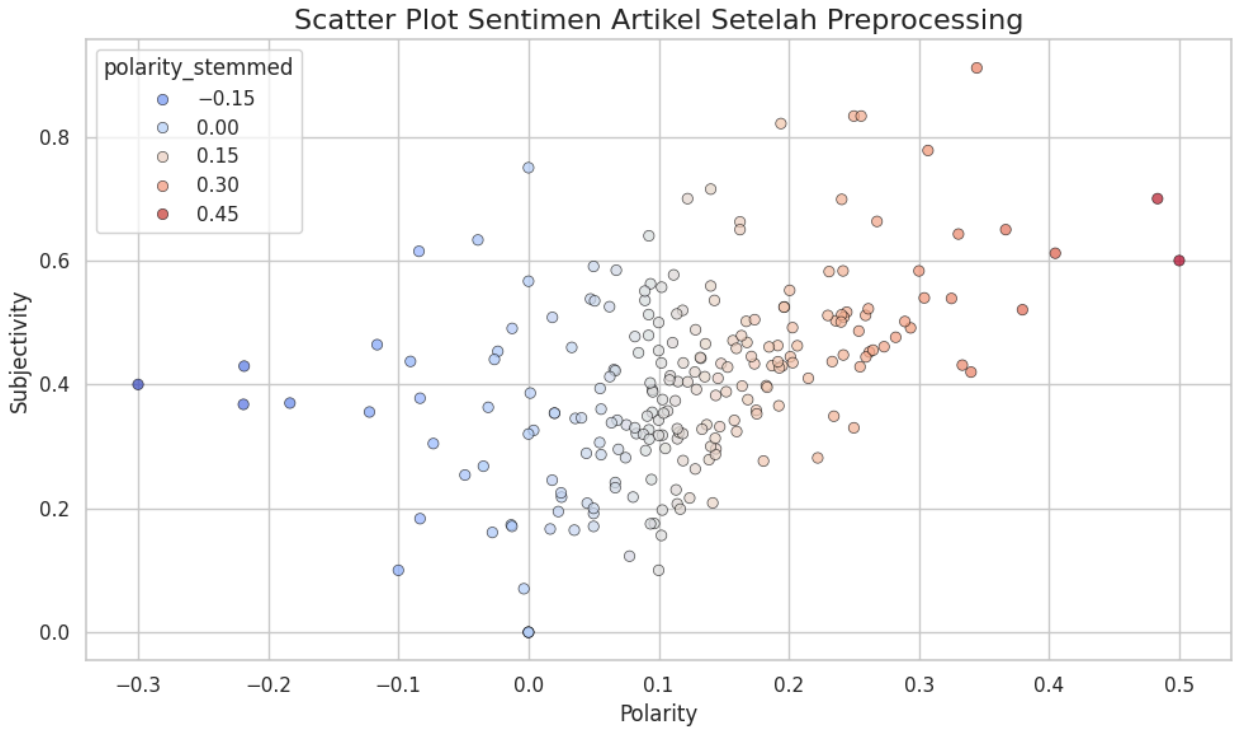

**Gambar 6.5** Wordcloud Sentimen Positif

Dari gambar 6.4 dan gambar 6.5 terlihat kesamaan beberapa kata umum, seperti danantara, indonesia, dan bumn yang menandakan bahwa ketiganya adalah kata kunci yang konsisten dibahas pada pemberitaan danantara baik dalam narasi positif maupun negatif. Selain kata umum tersebut dapat terlihat dari gambar 6.4 bahwa kata yang cukup spesifik dibahas pada artikel dengan sentimen negatif adalah menteri, dana, proyek, aset, dan prabowo. Kumpulan kata ini menunjukkan bahwa narasi negatif lebih banyak berfokus pada aktor politik, alokasi dana, serta pengelolaan proyek dan aset. Sehingga dapat disimpulkan bahwa kritik media atau publik terkait danantara cenderung negatif

pada pengelolaan, keuangan, transparansi, dan kepemimpinan. Sedangkan pada gambar 6.5 kata-kata yang menonjol pada artikel dengan sentimen positif adalah investasi, kelola, ekonomi, dan perintah. Kumpulan kata ini menunjukkan bahwa narasi positif lebih banyak berfokus pada optimisme potensi investasi, pengelolaan aset, kontribusi terhadap ekonomi, dan arahan kebijakan pemerintah. Sentimen positif ini umumnya muncul ketika membahas manfaat dan peluang dari kebijakan Danantara.

## 6.3 Analisis Sentimen

Analisis sentimen dilakukan menggunakan library TextBlob yang melalui proses translasi terlebih dahulu. Translasi diperlukan karena TextBlob dibangun di atas lexicon dan model analisis seperti *Pattern Analyzer* yang lebih banyak mengenali kata-kata dalam bahasa Inggris. Oleh karena itu diperlukan translasi menggunakan *library* Google Translate. Setelah menggunakan TextBlob, skor *Polarity* dan *Subjectivity* dihasilkan dan dapat digunakan untuk pengambilan insight.

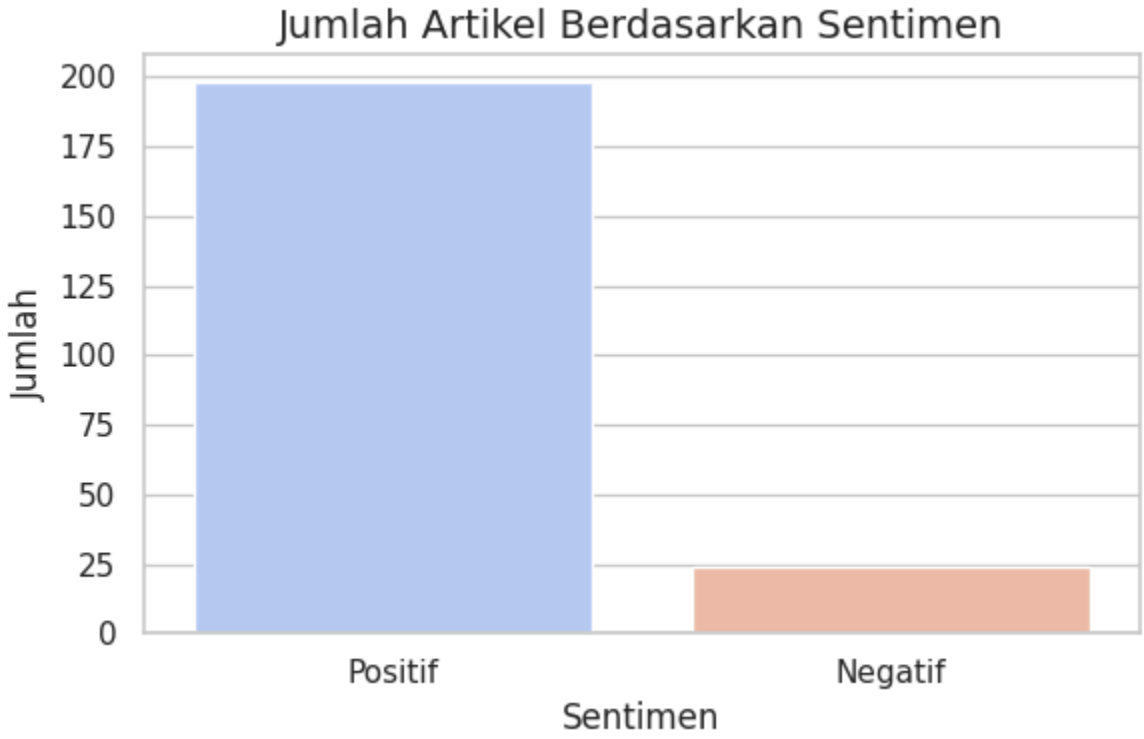



**Gambar 6.6** Scatter Plot Sentimen Artikel Setelah Preprocessing

Scatter plot ini menunjukkan hubungan antara nilai polarity dan subjectivity dari artikel-artikel berita yang telah melalui proses preprocessing teks. Sumbu X menunjukkan Polaritas yaitu seberapa positif atau negatif sentimen yang dimiliki suatu artikel berkisar -1 (sangat negatif) hingga +1 (sangat positif). Dalam plot ini titik paling ekstrim dari polaritas nya adalah -0.3 dan 0.5. Sumbu Y menunjukkan subjektivitas yang

mengukur berapa subjektif atau objektif teks bersibut. Nilai ini berkisar antara 0 (sangat objektif) hingga 1 (sangat subjektif). Sebagian besar data terkonsentrasi di sekitar nilai polarity netral hingga positif ringan (sekitar 0 hingga 0.2), dengan tingkat subjectivity yang bervariasi dari rendah hingga sedang. Hal ini menunjukkan bahwa mayoritas artikel bersifat netral atau memiliki sentimen positif ringan, yang sesuai dengan karakteristik umum berita yang cenderung menjaga objektivitas. Namun, terdapat pula beberapa artikel dengan polaritas negatif ringan, meskipun jumlahnya relatif lebih sedikit. Warna titik dalam grafik mewakili tingkat polaritas yang telah distem, dengan gradasi dari biru (negatif), putih (netral), hingga merah (positif). Secara umum, plot ini menunjukkan bahwa meskipun sebagian besar berita cenderung objektif, beberapa di antaranya mengandung opini atau nada emosional yang mencerminkan sentimen positif maupun negatif.

6.3.1 Jumlah Berita berdasarkan Sentimen Polaritas

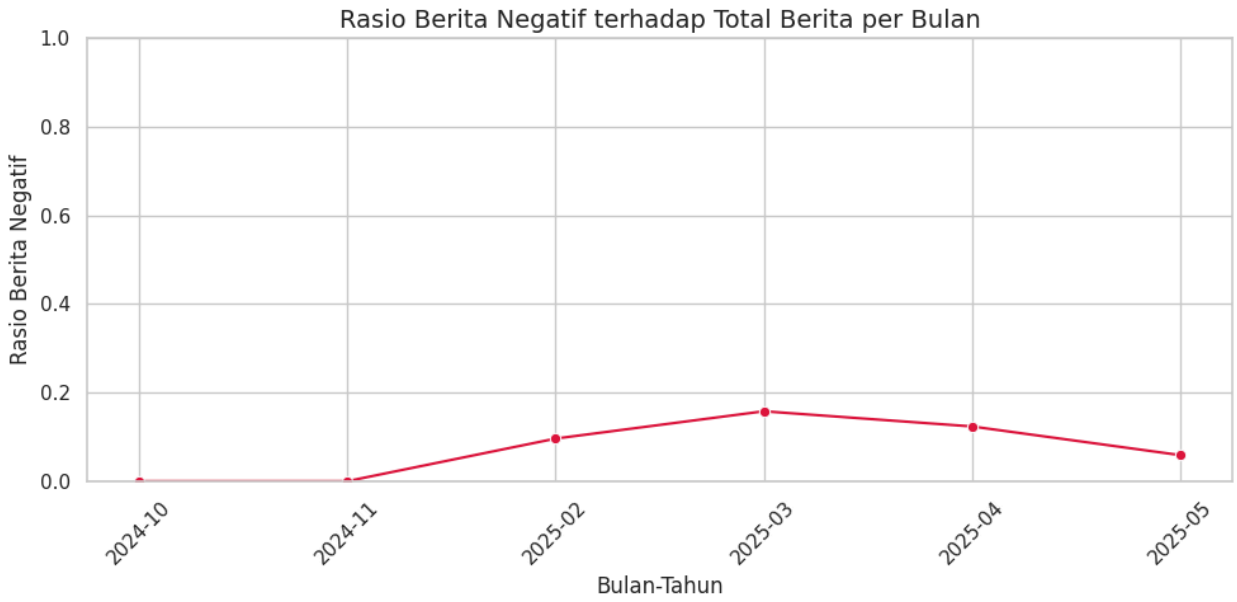



**Gambar 6.7** Bar Plot Jumlah Berita berdasarkan Sentimen

Berdasarkan grafik di atas, dapat dilihat bahwa jumlah artikel dengan sentimen positif jauh lebih banyak dibandingkan dengan artikel bersentimen negatif. Tercatat sekitar 198 artikel dikategorikan sebagai positif, sementara hanya sekitar 22 artikel yang termasuk dalam kategori negatif. Hasil ini menunjukkan bahwa sebagian besar artikel

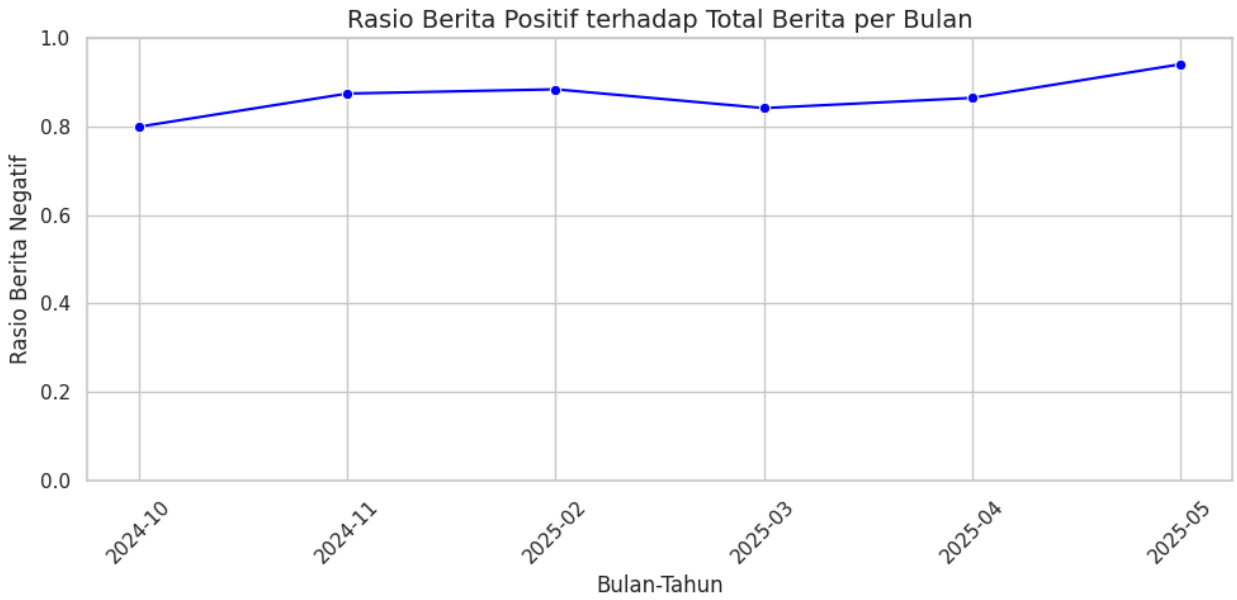
berita yang dianalisis memiliki nada yang positif atau netral-positif setelah melalui proses analisis sentimen.

### 6.3.2 Analisis Sentimen Berdasarkan Waktu




**Gambar 6.8** Rasio Berita Negatif terhadap Total Berita per Bulan

Grafik di atas menunjukkan rasio berita negatif terhadap total berita yang dipublikasikan setiap bulan. Terlihat bahwa pada bulan Oktober dan November 2024, tidak ada berita negatif yang terdeteksi. Rasio mulai meningkat pada Februari 2025, mencapai puncaknya di bulan Maret 2025 dengan sekitar 16% dari total berita bersentimen negatif. Setelah itu, rasio mengalami penurunan bertahap pada April dan Mei 2025. Hal ini mengindikasikan bahwa meskipun sentimen negatif sempat meningkat, secara umum proporsinya tetap rendah dibandingkan jumlah total berita setiap bulan.

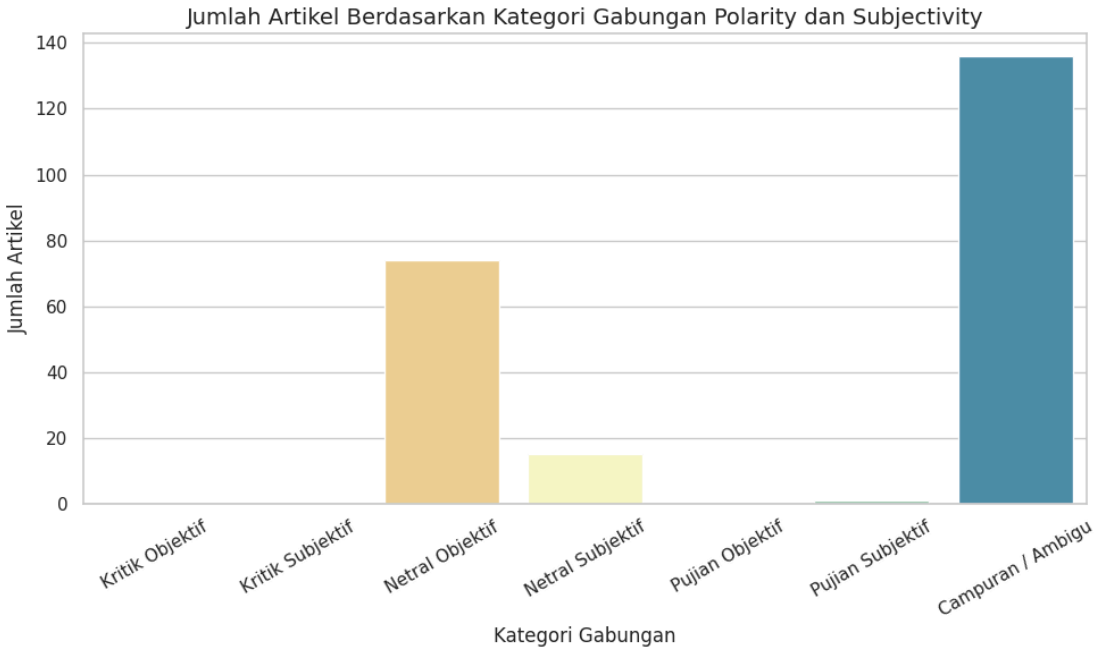



**Gambar 6.9** Rasio Berita Positif terhadap Total Berita per Bulan

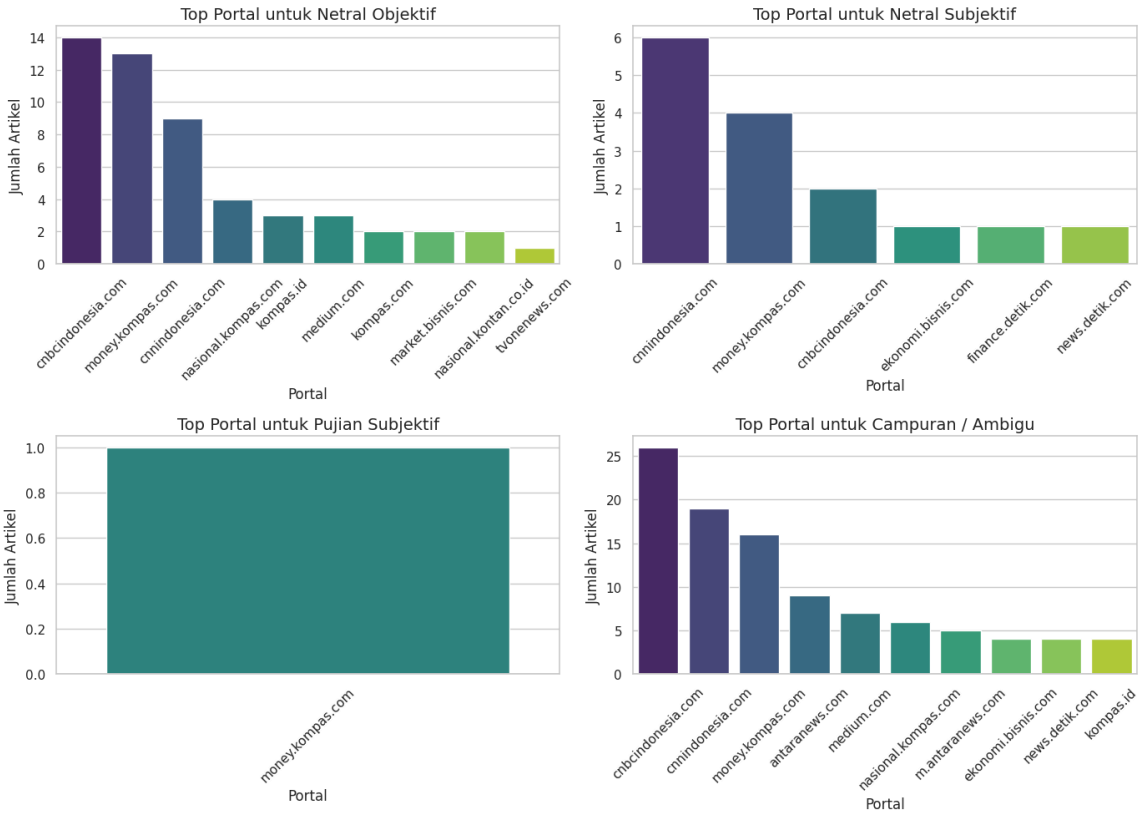
Grafik di atas memperlihatkan rasio berita positif terhadap total berita yang dipublikasikan setiap bulan. Secara umum, rasio berita positif tetap tinggi di setiap periode, berada di atas 80%. Rasio ini mengalami peningkatan dari bulan Oktober 2024 hingga Mei 2025, dengan sedikit penurunan pada Maret 2025 sebelum kembali naik dan mencapai puncaknya pada Mei 2025 di angka mendekati 95%. Temuan ini menunjukkan bahwa mayoritas berita yang dipublikasikan setiap bulan cenderung bersentimen positif, dengan tren yang relatif stabil dan meningkat seiring waktu.

### 6.3.3 Analisis Sentimen Berdasarkan Portal

Klasifikasi gabungan polaritas dan subjektivitas artikel dilakukan dengan membagi ke dalam tujuh kategori berdasarkan nilai polarity dan subjectivity. Artikel dengan polaritas sangat negatif (≤ -0.5) dikategorikan sebagai Kritik, dan yang sangat positif (≥ 0.5) sebagai Pujian, masing-masing dibedakan lagi menjadi Objektif (subjektivitas ≤ 0.5) dan Subjektif (subjektivitas > 0.5). Artikel dengan polaritas netral (antara -0.1 hingga 0.1) dibagi menjadi Netral Objektif dan Netral Subjektif berdasarkan subjektivitasnya. Sementara itu, artikel yang tidak memenuhi kriteria tersebut dimasukkan ke dalam kategori Campuran / Ambigu. Klasifikasi ini bertujuan untuk menangkap tidak hanya sikap positif atau negatif, tetapi juga sejauh mana narasi bersifat fakta atau opini. Pendekatan ini penting untuk mengidentifikasi peran bahasa dalam membentuk opini publik, serta untuk mengevaluasi kecenderungan editorial dari masing-masing media dalam menyampaikan informasi, khususnya yang berkaitan dengan isu-isu ekonomi.



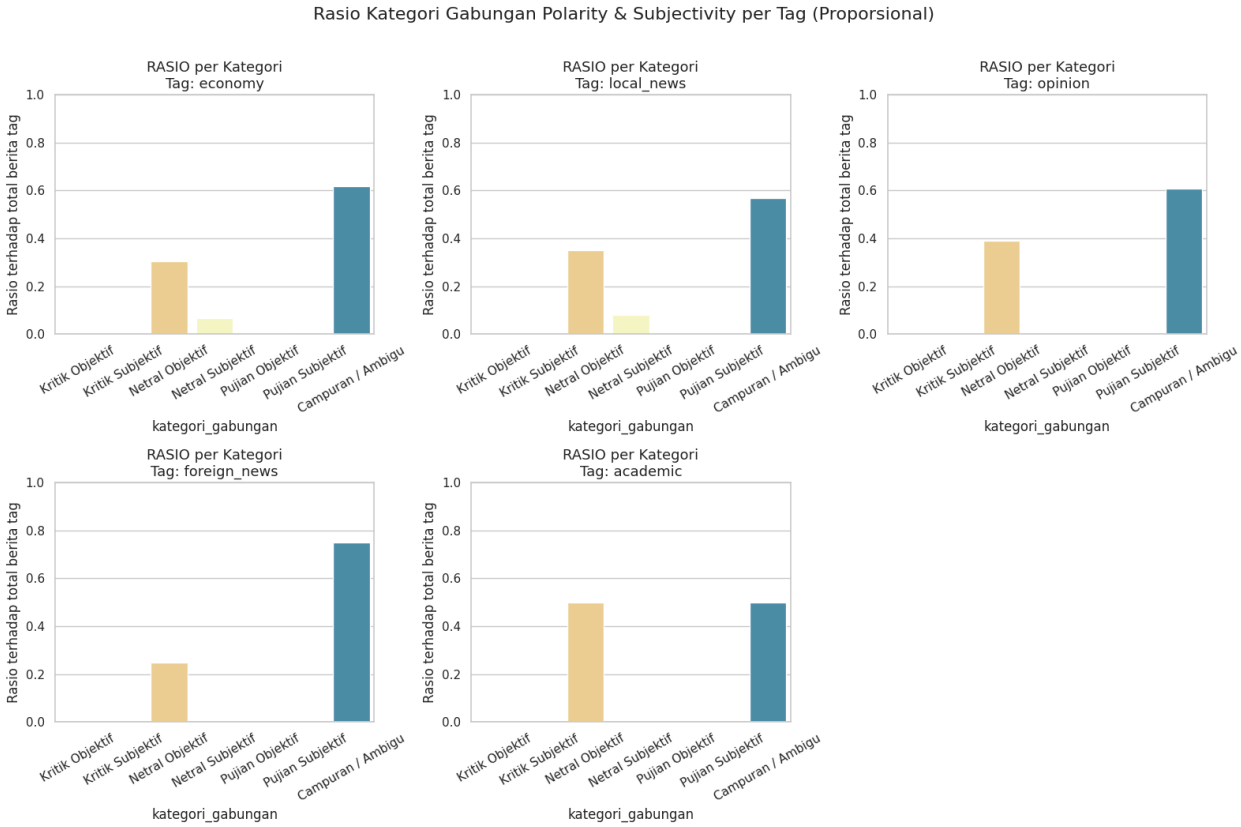

**Gambar 6.10** Bar Plot Jumlah Artikel berdasarkan Kategori Gabungan Polarity dan Subjectivity



**Gambar 6.11** Bar Plot Jumlah Publikasi Berdasarkan Sumber

Hasil visualisasi menunjukkan bahwa portal cnbcindonesia.com dan money.kompas.com mendominasi kategori Netral Objektif, yang berarti mereka lebih banyak mempublikasikan berita ekonomi yang bersifat faktual dan netral. Sementara itu, cnnindonesia.com lebih banyak muncul dalam kategori Netral Subjektif, yang mengindikasikan adanya penyampaian informasi netral namun dibalut dengan opini atau gaya bahasa yang subjektif. Kategori Pujian Subjektif hanya diisi oleh satu artikel dari money.kompas.com, menunjukkan bahwa gaya pemberitaan yang sangat positif dan sarat opini jarang digunakan oleh portal-portal yang dianalisis. Sebaliknya, kategori Campuran/Ambigu cukup mendominasi, dengan cnbcindonesia.com sebagai penyumbang terbanyak, diikuti oleh cnnindonesia.com dan money.kompas.com. Ini menunjukkan bahwa banyak artikel yang memiliki karakteristik bahasa yang tidak bisa digolongkan secara pasti sebagai positif, negatif, netral, ataupun faktual sepenuhnya, sehingga dinilai bersifat campuran atau ambigu.

### 6.3.4 Analisis Sentimen berdasarkan Tag Artikel



**Gambar 6.12** Rasio Klasifikasi Sentimen berdasarkan Tag Artikel

Visualisasi pada Gambar menunjukkan distribusi proporsional dari kategori gabungan polaritas dan subjektivitas berdasarkan lima kategori tag utama: economy, local_news, opinion, foreign_news, dan academic. Secara umum, kategori Campuran / Ambigu mendominasi hampir semua tag, yang menandakan bahwa banyak artikel memuat elemen bahasa yang tidak secara konsisten objektif maupun subjektif, serta tidak menunjukkan arah sentimen yang jelas. Pada tag economy, terlihat bahwa proporsi artikel Netral Objektif relatif tinggi dibandingkan tag lainnya, menunjukkan adanya kecenderungan untuk menyampaikan berita ekonomi secara faktual. Sementara pada tag opinion, proporsi Netral Objektif dan Campuran / Ambigu hampir seimbang, mencerminkan adanya opini yang tetap menjaga keseimbangan tanpa kecenderungan emosional yang kuat.

Untuk tag *foreign_news*, *Campuran / Ambigu* mencapai lebih dari 70%, mengindikasikan bahwa liputan luar negeri sering kali ditulis dengan gaya bercampur atau tidak eksplisit dalam mengekspresikan opini atau emosi. Tag *academic* memperlihatkan dua kutub utama yaitu *Netral Objektif* dan *Campuran / Ambigu*, memperkuat asumsi bahwa berita akademik berusaha menyajikan data dan fakta, namun tidak sepenuhnya bebas dari interpretasi atau penyusunan narasi yang mengandung ambiguitas. Tag *local_news* memiliki komposisi serupa dengan *economy*, meskipun porsi *Netral Subjektif* juga mulai terlihat, yang dapat dikaitkan dengan peran narasi lokal dan kedekatan emosional terhadap isu di sekitar masyarakat.

## 6.4 TF-IDF Konten Artikel

Analisis Term Frequency-Inverse Document Frequency (TF-IDF) dilakukan untuk mengidentifikasi kata-kata yang memiliki bobot penting dalam artikel. Hasil dari TF-IDF dapat menunjukkan kata-kata yang sering muncul tapi tidak umum, atau bisa dibilang spesifik kepada topik tertentu. Pada analisis ini, TF-IDF digunakan untuk mencari top words dari seluruh artikel untuk melihat topik secara keseluruhan dan dengan mengelompokkan berdasarkan periode bulan untuk mengetahui dinamika topik pembicaraan setiap bulannya.

**Tabel 6.2** Top 5 TF-IDF untuk seluruh Artikel

| No | Kata | TF-IDF Score |
|---|---|---|
| 1 | bumn | 0.0729 |
| 2 | aset | 0.0545 |
| 3 | prabowo | 0.0477 |

| | | |
|---|---|---|
| 4 | negara | 0.0448 |
| 5 | ros | 0.0390 |

Berdasarkan hasil pada tabel 6.2 dapat dilihat bahwa mayoritas artikel danantara berhubungan dengan pembahasan terkait bumn, aset, prabowo, dan negara. Hal ini wajar mengingat danantara adalah proyek besar negara dimasa kepemimpinan Presiden Prabowo dan merupakan badan yang mengelola aset dan bumn.

**Tabel 6.3** Top 5 TF-IDF Bulan Oktober

| No | Kata | TF-IDF Score |
|---|---|---|
| 1 | kelola | 0.2284 |
| 2 | investasi | 0.1822 |
| 3 | bumn | 0.1513 |
| 4 | badan | 0.1344 |
| 5 | aset | 0.1223 |

Berdasarkan analisis pada Tabel 6.3 kemungkinan topik utama pada bulan oktober tentang pengelolaan investasi BUMN, hal ini menandakan artikel pada bulan ini berfokus tentang strategi atau aktivitas pengelolaan aset dan investasi dana pemerintah.

**Tabel 6.4** Top 5 TF-IDF Bulan November

| No | Kata | TF-IDF Score |
|---|---|---|
| 1 | bumn | 0.1516 |
| 2 | presiden | 0.1404 |
| 3 | bentuk | 0.1294 |
| 4 | luncur | 0.0760 |
| 5 | bp | 0.0732 |

Tabel 6.4 menunujukkan bahwa topik utama pada bulan November adalah seputar presiden, bumn, bentuk, serta luncur. Hal ini mengindikasikan artikel mengenai

peluncuran kebijakan oleh presiden atau pembentukan badan baru yang melibatkan BUMN dan pemerintah pusat.

**Tabel 6.5** Top 5 TF-IDF Bulan Februari

| No | Kata | TF-IDF Score |
|---|---|---|
| 1 | negara | 0.0608 |
| 2 | bank | 0.0565 |
| 3 | dana | 0.0540 |
| 4 | ekonomi | 0.0524 |
| 5 | awas | 0.0471 |

Berdasarkan Tabel 6.5 kemungkinan topik pada bulan Februari didominasi oleh isu keuangan yang berhubungan dengan danantara seperti pengelolaan dana, bank, dan ekonomi. Hasil ini kemungkinan membahas tentang dampak danantara terhadap kondisi negara, bank, dana, dan ekonomi yang sedang berbahaya.

**Tabel 6.6** Top 5 TF-IDF Bulan Maret

| No | Kata | TF-IDF Score |
|---|---|---|
| 1 | bumn | 0.0853 |
| 2 | kelola | 0.0697 |
| 3 | investasi | 0.0693 |
| 4 | proyek | 0.0624 |
| 5 | indonesia | 0.0613 |

Tabel 6.6 menunujukkan bahwa pada bulan maret isu yang membahas bumn, pengelolaan, dan investasi kembali menjadi sorotan. Diiktui dengan kata proyek dan indonesia, mengindikasikan kemungkinan pengelolaan proyek investasi bumn oleh danantara

**Tabel 6.7** Top 5 TF-IDF Bulan April

| No | Kata | TF-IDF Score |
|---|---|---|

| | | |
|---|---|---|
| 1 | qatar | 0.0695 |
| 2 | bumn | 0.0695 |
| 3 | prabowo | 0.0663 |
| 4 | aset | 0.0632 |
| 5 | ros | 0.0531 |

Tabel 6.7 menunjukkan adanya topik baru pada bulan April yang membahas tentang kerjasama bilateral antara indonesia dan qatar yang melibatkan Presiden Prabowo. Kerjasama ini juga sepertinya berdampak pada bumn serta aset.

**Tabel 6.8** Top 5 TF-IDF Bulan Mei

| No | Kata | TF-IDF Score |
|---|---|---|
| 1 | aset | 0.0859 |
| 2 | gates | 0.0748 |
| 3 | ros | 0.0615 |
| 4 | bumn | 0.0600 |
| 5 | trilliun | 0.0584 |

Tabel 6.8 menunjukkan topik pembahasan baru pada bulan Mei, yaitu Bill Gates yang kemungkinannya akan berinvestasi dengan nilai yang cukup besar (trilliunan). Selain itu bumn dan aset masih menjadi pembahasan berulang.

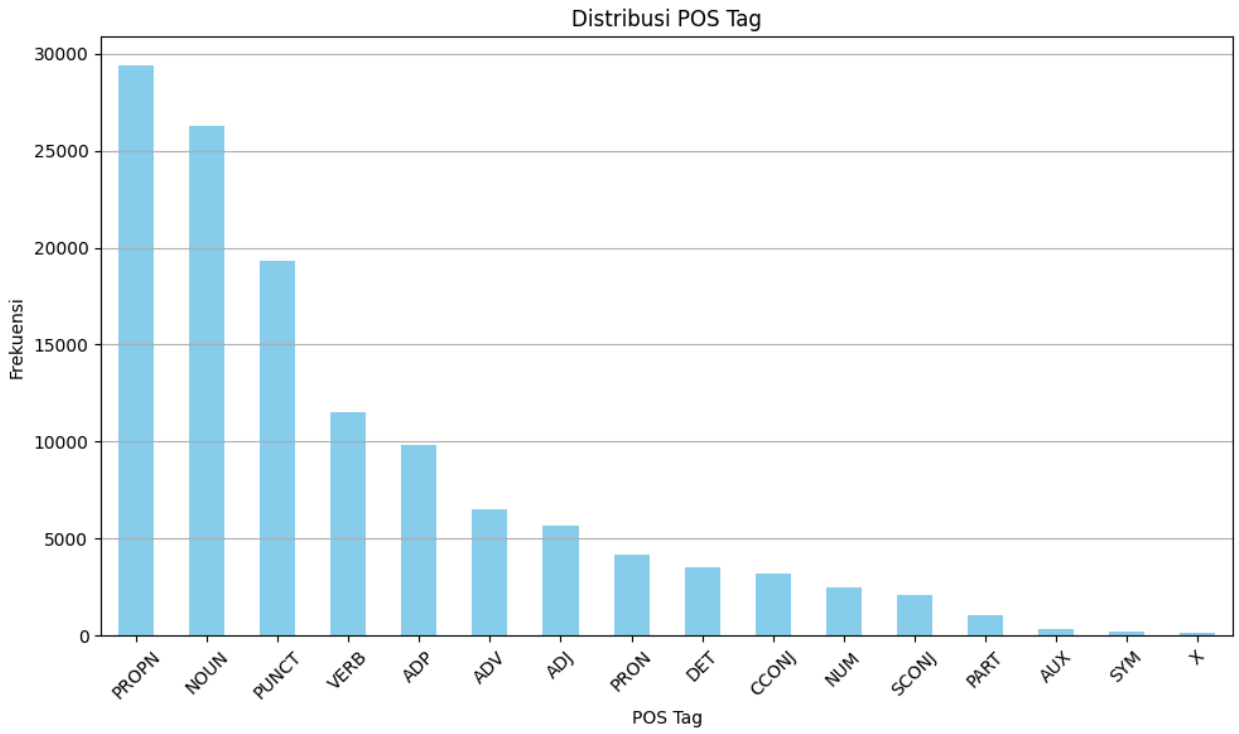
## 6.5 POS dan NER Konten Artikel Berita

**Tabel 6.9** Data POS Universal

| Tag | Nama Lengkap | Penjelasan Singkat |
|---|---|---|
| NOUN | Noun | Kata benda umum (buku, meja, presiden) |
| PROPN | Proper Noun | Nama khusus (Indonesia, Jokowi, Danantara) |
| VERB | Verb | Kata kerja utama (makan, pergi, membaca) |

| | | |
|---|---|---|
| AUX | Auxiliary Verb | Kata kerja bantu (telah, akan, sedang) |
| ADJ | Adjective | Kata sifat (besar, cepat, hijau) |
| ADV | Adverb | Kata keterangan (cepat, sangat, dulu) |
| PRON | Pronoun | Kata ganti (saya, kamu, dia, mereka) |
| DET | Determiner | Penentu (ini, itu, setiap, semua) |
| ADP | Adposition | Preposisi (di, ke, dari, pada) |
| SCONJ | Subordinating Conj. | Konjungsi subordinatif (karena, jika, agar) |
| CCONJ | Coordinating Conj. | Konjungsi koordinatif (dan, atau, tetapi) |
| PART | Particle | Partikel (lah, pun, kah) |
| INTJ | Interjection | Seruan/emosi (hai, wah, aduh) |
| NUM | Numeral | Angka/bilangan (dua, 10, ketiga) |
| PUNCT | Punctuation | Tanda baca (., !, :) |
| SYM | Symbol | Simbol (%, $, #) |
| X | Other/Unknown | Kategori tidak dikenali (biasanya error tagging) |

Tabel 6.9 menyajikan daftar lengkap tag Part-of-Speech (POS) berdasarkan skema Universal POS Tagging yang digunakan dalam analisis sintaksis Bahasa Indonesia. Setiap tag merupakan representasi kelas kata seperti kata benda (NOUN), kata kerja (VERB), kata sifat (ADJ), dan lain-lain. Penjelasan singkat dan contoh penggunaan disertakan untuk mempermudah pemahaman terhadap fungsi masing-masing tag dalam struktur kalimat.




**Gambar 6.12** Grafik Distribusi POS Tag

Dari grafik distribusi POS Tag diperoleh informasi PROPN menjadi tag paling dominan dengan hampir 30.000 kemunculan. Hal ini mengindikasikan teks yang dianalisis **sangat kaya dengan nama entitas seperti orang, tempat, atau institusi**. Ini umum dijumpai dalam teks seperti berita, laporan, atau artikel formal yang menyebut banyak nama diri. Kemudian NOUN menempati posisi kedua, menunjukkan banyak kata benda umum digunakan yang memperkuat indikasi bahwa teks banyak menyampaikan **objek, konsep,** atau **peristiwa**. Di peringkat ketiga ditemukan banyak PUNCT (tanda baca) yang diasumsikan karena diperoleh dari artikel yang umumnya memiliki **gaya penulisan** dengan struktur kalimat yang **lengkap** dan **tertulis formal**. Hal itu juga menyebabkan kemunculan SYM (simbol), X (unspecified), AUX (kata kerja bantu), dan PART (partikel) **muncul sangat jarang** mengindikasikan bahwa teks tidak mengandung banyak simbol atau kata yang tidak dikenali, bentuk kalimat pasif atau progresif kurang dominan, dan teks mungkin tidak bersifat sangat percakapan/informal


**Gambar 6.13** *Display POS Tag Result* pada Dataset Danantara

Gambar 6.13 menampilkan hasil visualisasi Part-of-Speech (POS) tagging terhadap salah satu konten artikel berbahasa Indonesia. Setiap kata diberi label kelas katanya sesuai dengan skema Universal POS Tagging, seperti NOUN (kata benda), VERB (kata kerja), ADJ (kata sifat), PROPN (nama diri), dan sebagainya. Warna latar yang berbeda digunakan untuk membedakan jenis tag, sehingga memudahkan dalam mengidentifikasi struktur gramatikal kalimat secara visual. Visualisasi ini membantu memahami distribusi dan fungsi sintaktik kata-kata dalam sebuah teks, serta mendukung analisis linguistik maupun praproses data dalam tugas-tugas NLP.

**Tabel 6.11** Top 10 POS Tag *counts*

| No | Word | POS | Count |
|---|---|---|---|
| 1 | Danantara | PROPN | 2157 |

| | | | |
|---|---|---|---|
| 2 | BUMN | PROPN | 911 |
| 3 | Indonesia | PROPN | 901 |
| 4 | investasi | NOUN | 574 |
| 5 | Prabowo | PROPN | 553 |
| 6 | negara | NOUN | 473 |
| 7 | aset | NOUN | 403 |
| 8 | BPI | PROPN | 383 |
| 9 | Presiden | PROPN | 375 |
| 10 | ekonomi | NOUN | 318 |

Tabel 6.11 memberikan gambaran tentang sepuluh kata yang paling sering muncul dalam korpus teks yang dianalisis beserta kategori kelas katanya (POS). Hasil ini secara langsung **menjawab rumusan masalah terkait identifikasi kata kunci dominan dan pola linguistik yang muncul dalam pembahasan isu Danantara Indonesia.** Terlihat bahwa Danantara muncul secara signifikan lebih tinggi dibandingkan kata lainnya, dengan jumlah kemunculan sebanyak 2.157 kali dan dikategorikan sebagai PROPN (proper noun). Hal ini menunjukkan bahwa topik Danantara merupakan fokus utama dalam narasi dokumen yang diteliti.

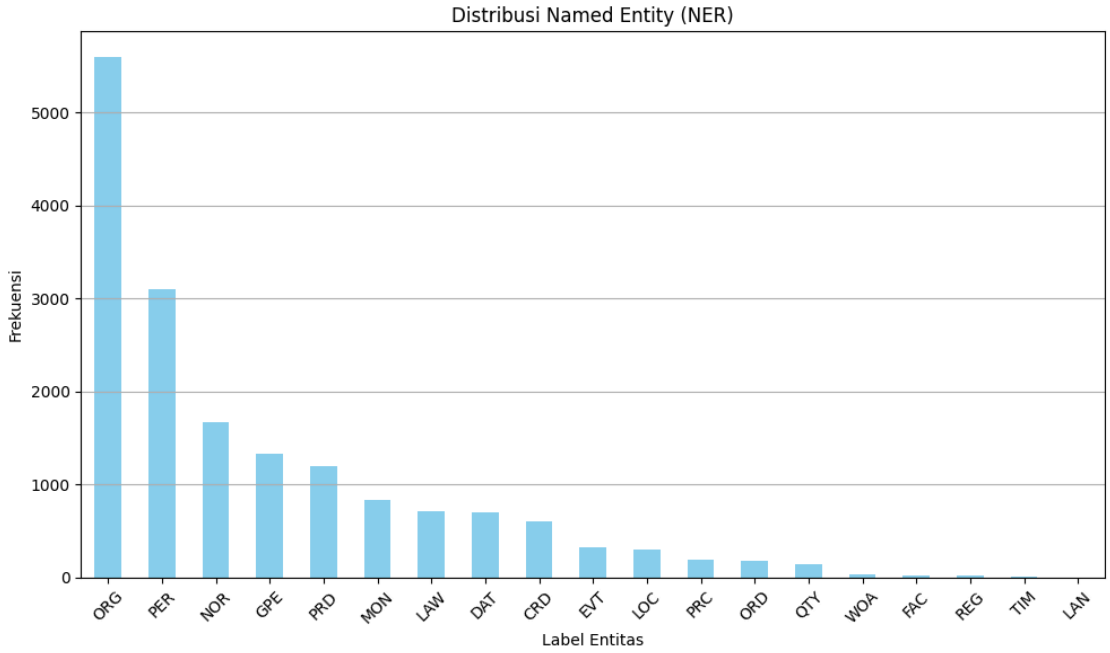
Selain Danantara, kata-kata seperti **BUMN, Indonesia, Prabowo, Presiden, dan BPI** juga muncul dengan frekuensi tinggi dalam kategori **PROPN**, mengindikasikan bahwa pembahasan banyak berkaitan dengan **tokoh publik** dan **institusi formal**. Sementara itu, kata-kata seperti **investasi, negara, aset,** dan **ekonomi** yang masuk dalam kategori **NOUN**, memperkuat bahwa topik yang diangkat mencakup isu-isu ekonomi dan tata kelola. Temuan ini selaras dengan fokus penelitian terhadap representasi kebijakan publik dan wacana ekonomi dalam teks media atau dokumen resmi.

**Tabel 6.12** Data NER Universal

| Label | Nama Lengkap | Contoh |
|---|---|---|
| PERSON | Nama Orang | Jokowi, Prabowo |
| ORG | Organisasi | BRI, Danantara, PLN |
| GPE | Lokasi/Negara | Indonesia, Kalimantan |
| LOC | Lokasi (non-negara) | Gunung Merapi, Jalan Sudirman |
| DATE | Tanggal/Waktu | 2023, 27 Februari |

| | | |
|---|---|---|
| TIME | Waktu spesifik | 08:00, pagi |
| MISC | Lain-lain | UMKM, kategori khusus |
| NUMBER | Angka | 10 juta, 45% |

Tabel 6.10 menunjukkan daftar label entitas yang digunakan dalam proses Named Entity Recognition (NER) berdasarkan skema Universal. Label-label ini mengkategorikan kata atau frasa penting dalam teks ke dalam tipe-tipe entitas seperti nama orang (PERSON), organisasi (ORG), lokasi (GPE atau LOC), waktu (DATE atau TIME), angka (NUMBER), serta entitas lain yang bersifat umum atau tidak terklasifikasi (MISC).




**Gambar 6.14** Grafik Distribusi NER

Gambar 6.14 menampilkan distribusi untuk label Named Entity Recognition (NER) yang dihasilkan dari proses ektraksi entitas pada datase artikel. Grafik ini menunjukkan bahwa ORG (Organization) menjadi label yang paling dominan dengan lebih dari 5.000 kali kemunculan. Hal ini menunjukkan bahwa dataset artikel mengandung banyak sekali pembahasan terkait organisasi, lembaga, atau perusahaan. Dominasi ini sesuai dengan topik utama yang dibawa, yaitu organisasi yang baru didirikan pemerintah dengan nama danantara yang merupakan badan investasi negara.

Label selanjutnya yang paling banyak muncul adalah PER (Person) dengan lebih dari 3.000 kemunculan yang mengindikasikan seringnya topik danantara dihubungkan

dengan individu tertentu, seperti pejabat negara, pimpinan organisasi, atau tokoh publik lain.

Label dengan urutan terbanyak selanjutnya secara berurutan adalah NOR (Nationality, religious, or political group), GPE (Geo-Political Entity) dan PRD (Product). Distribusi yang menonjol pada label ORG dan PER menadakan bahwa teks yang dianalisis berfokus pada pemberitaan mengenai hubungan antar organisasi, peristiwa bisinis, maupun politik yang melibatkan individu atau institusi tertentu. Dominasi ini relevan dengan konteks berita ekonomi, pemerintahan, dan korporasi.


**Gambar 6.15** *Display NER Tag Result* pada Dataset Danantara

Gambar 6.15 menampilkan hasil visualisasi tagging untuk Named Entity Recognition (NER) secara langsung pada salah satu artikel. Dapat dilihat bahwa setiap kata atau frasa penting yang teindikasi sebagai entitas diberi label sesuai dengan kategorinya, seperti ORG, PER, GPE, LOC, dan lainnya. Visualisasi ini memudahkan pembaca untuk mengidentifikasi dan memahami konteks berita melalui highlight entitas yang penting dalam artikel. Selain itu, hasil ini juga menunjukkan keberhasilan sistem NER dalam mendeteksi berbagai jenis entitas, meskipun masih ada kemungkinan kesalahan atau keterbatasan dalam mendeteksi entitas yang ambigu. Kemunculan banyak entitas ORG, PER, dan NOR sangat mewakili hasil pada gambar 6.14, dimana ketiga

entitas tersebut memang adalah entitas dengan jumlah kemunculan terbanyak. Kemunculan ini juga menunjukkan fokus utama teks berada pada perusahaan, pejabat, dan lokasi yang relevan dengan pemberitaan ekonomi, politik, atau bisnis.

**Tabel 6.13** Top 10 NER *Tag counts*

| No | Entitas | Label | Count |
|---|---|---|---|
| 1 | Danantara | ORG | 964 |
| 2 | Indonesia | GPE | 518 |
| 3 | Prabowo | PER | 356 |
| 4 | Jakarta | GPE | 249 |
| 5 | Pemerintah | NOR | 245 |
| 6 | Bpi Danantara | ORG | 193 |
| 7 | Prabowo Subianto | PER | 178 |
| 8 | Rosan | PER | 171 |
| 9 | Bumn | ORG | 129 |
| 10 | Kpk | NOR | 114 |

Tabel 6.13 menyajikan sepuluh entitas bernama (named entities) yang paling sering muncul dalam korpus teks berdasarkan hasil analisis Named Entity Recognition (NER). Entitas Danantara menduduki peringkat tertinggi dengan jumlah kemunculan sebanyak 964 kali dan diklasifikasikan sebagai ORG (organisasi), yang menunjukkan bahwa topik utama dalam dokumen sangat terfokus pada entitas ini. Selain Danantara, entitas seperti Bpi Danantara dan Bumn juga termasuk dalam kategori organisasi, memperkuat temuan bahwa korpus banyak membahas institusi yang berkaitan dengan sektor publik dan kebijakan ekonomi. Entitas bertipe GPE (Geo-Political Entity) seperti Indonesia dan Jakarta menunjukkan bahwa aspek geografis dan nasionalitas juga menjadi elemen penting dalam pembahasan. Sementara itu, tokoh-tokoh seperti Prabowo, Prabowo Subianto, dan Rosan muncul sebagai PER (person), menandakan bahwa teks juga menyoroti aktor-aktor politik atau tokoh penting yang relevan dalam konteks pembahasan. Kehadiran entitas seperti Pemerintah dan KPK, yang diklasifikasikan sebagai NOR, mengindikasikan bahwa terdapat muatan kebijakan dan penegakan hukum yang cukup dominan dalam isi teks. Secara keseluruhan, hasil ini menunjukkan bahwa wacana dalam korpus teks sangat terkait dengan organisasi besar, tokoh publik, dan institusi negara. Distribusi entitas ini membantu menjawab rumusan masalah yang berkaitan dengan fokus aktor, lembaga, dan wilayah geografis yang dominan dalam narasi kebijakan publik dan ekonomi di Indonesia.

Jika dibandingkan secara langsung, Tabel 6.11 dan Tabel 6.13 menunjukkan keterkaitan yang kuat antara kata-kata dengan frekuensi tinggi (berdasarkan POS tagging) dan entitas-entitas yang berhasil diidentifikasi secara spesifik melalui proses Named Entity Recognition (NER). Kata Danantara, misalnya, tidak hanya muncul sebagai kata dengan frekuensi tertinggi dalam POS tagging dan diklasifikasikan sebagai PROPN, tetapi juga menjadi entitas organisasi (ORG) yang paling dominan dalam hasil NER. Ini menandakan bahwa Danantara merupakan topik utama dalam narasi wacana yang dianalisis, baik dari sisi struktur sintaktik maupun dari segi entitas konseptual. Korelasi yang sama dapat dilihat pada kata-kata seperti Prabowo, BUMN, dan Indonesia, yang tidak hanya menduduki posisi tinggi dalam POS tagging tetapi juga diidentifikasi secara jelas sebagai entitas bernama dalam hasil NER, masing-masing dikategorikan sebagai PER, ORG, dan GPE. Hal ini memperlihatkan bahwa kata-kata yang sering muncul sebagai proper noun (PROPN) dalam POS tagging umumnya berkaitan erat dengan entitas penting dalam konteks kebijakan dan wacana publik. Selain itu, keberadaan entitas seperti Pemerintah, KPK, dan Jakarta menunjukkan bahwa konteks pembahasan dalam teks tidak hanya terfokus pada individu atau organisasi, tetapi juga menyentuh aspek struktural negara dan lokasi geografis yang relevan.

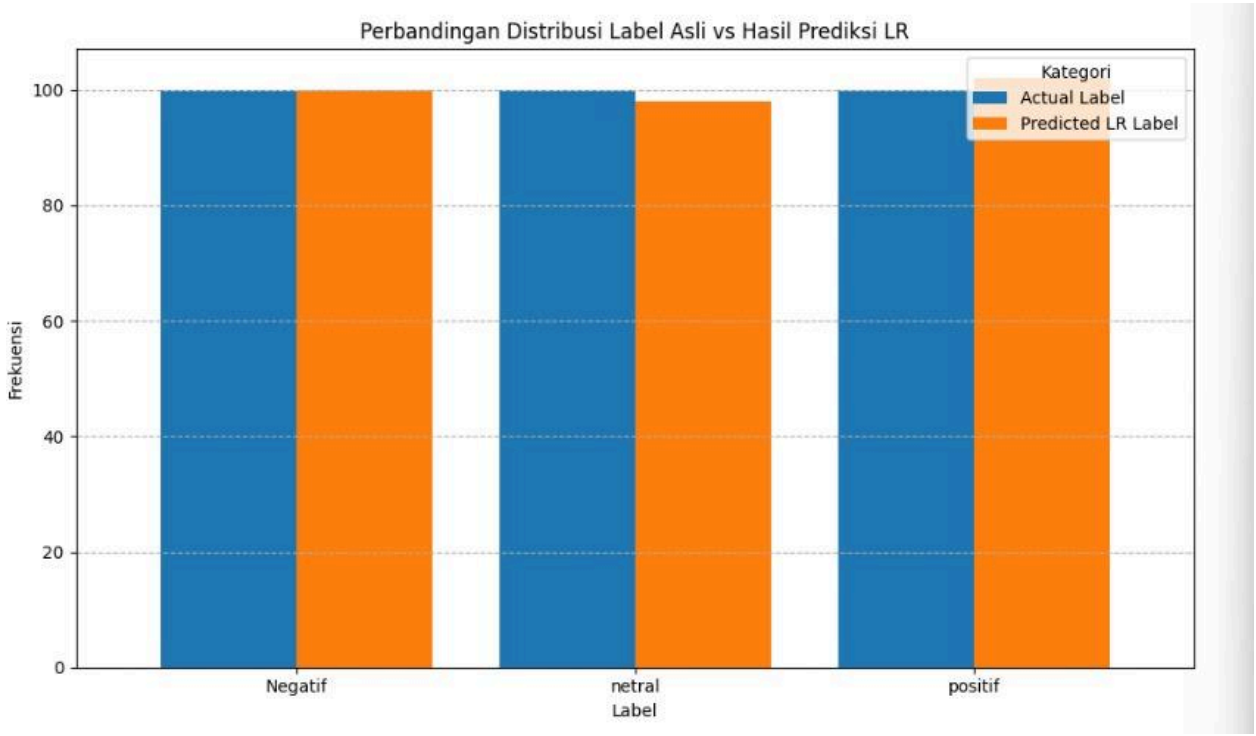
## 6.6 Hasil Prediksi Sentimen Model Logistic Regression + TF-IDF

**Tabel 6.14** Performa Model Logistik Regression +TF-IDF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Negatif | 0.89 | 0.85 | 0.87 | 20 |
| netral | 0.67 | 0.70 | 0.68 | 20 |
| positif | 0.75 | 0.75 | 0.75 | 20 |
| | | | | |
| accuracy | | | 0.77 | 60 |
| Macro avg | 0.77 | 0.77 | 0.77 | 60 |
| Weighted avg | 0.77 | 0.77 | 0.77 | 60 |

Hasil evaluasi performa model Logistic Regression menunjukkan bahwa model ini diuji pada 60 data dengan tiga kelas sentimen: negatif, netral, dan positif. Untuk kelas negatif, model memiliki precision sebesar 0.89, recall 0.85, dan f1-score 0.87 dari 20 data (support). Ini menunjukkan bahwa model cukup akurat dalam mengklasifikasikan sentimen negatif. Untuk kelas netral, precision-nya 0.67, recall 0.70, dan f1-score 0.68,

yang berarti performa model pada kategori ini masih cukup, meskipun tidak sekuat pada kelas negatif. Sedangkan pada kelas positif, precision, recall, dan f1-score semuanya sebesar 0.75, menandakan performa yang konsisten namun masih bisa ditingkatkan. Secara keseluruhan, model menghasilkan akurasi 77%, dengan nilai rata-rata f1-score makro dan tertimbang (macro avg dan weighted avg) sama-sama 0.77. Ini menandakan bahwa model cukup seimbang dalam menangani ketiga kelas tersebut.

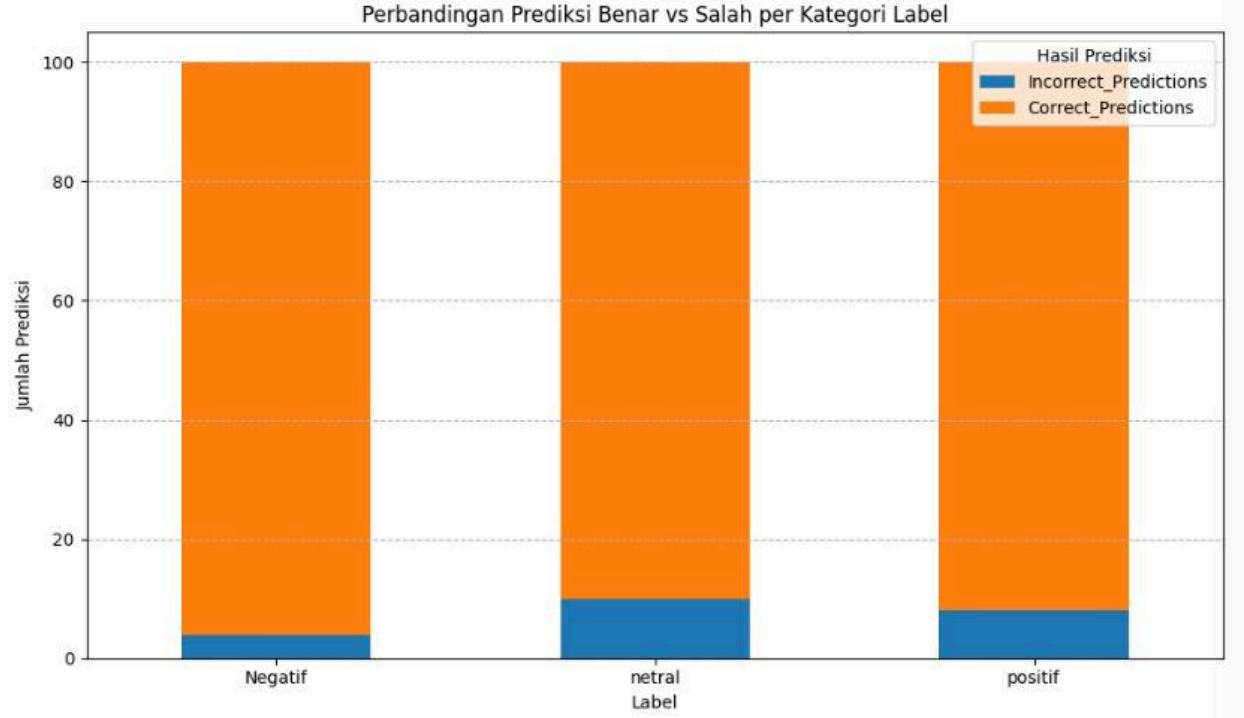



**Gambar 6.16** Perbandingan Distribusi Label Asli vs Hasil Prediksi LR

Bar chart di atas menunjukkan perbandingan antara distribusi label asli dan hasil prediksi model Logistic Regression untuk tiga kategori sentimen: negatif, netral, dan positif. Sumbu X menampilkan ketiga label sentimen, sedangkan sumbu Y menunjukkan jumlah atau frekuensi data pada masing-masing kategori. Warna biru merepresentasikan jumlah data berdasarkan label asli, sementara warna oranye menunjukkan jumlah data hasil prediksi model. Terlihat bahwa frekuensi prediksi model hampir sama dengan label aslinya pada ketiga kategori, yang mengindikasikan bahwa model memiliki distribusi prediksi yang seimbang dan tidak cenderung bias terhadap satu kelas tertentu. Meskipun begitu, kesamaan distribusi ini belum tentu menunjukkan bahwa semua prediksi benar, karena model masih bisa salah dalam mengklasifikasikan data individu. Namun secara umum, hasil ini memberikan indikasi bahwa model Logistic Regression bekerja cukup stabil dalam mengenali pola distribusi kelas.

```
Ringkasan Prediksi:
Jumlah Prediksi Benar: 278
Jumlah Prediksi Salah: 22

Perbandingan Prediksi Benar vs Salah per Kategori Label:
Prediction_Correct  Incorrect_Predictions  Correct_Predictions
label
Negatif                                 4                   96
netral                                 10                   90
positif                                 8                   92
```




**Gambar 6.17** Perbandingan Prediksi Benar dan Prediksi Salah per Kategori Label dengan Model Logistic Regression

Secara keseluruhan, terdapat 278 prediksi yang benar dan 22 prediksi yang salah dari total 300 data. Pada label negatif, terdapat 96 prediksi benar dan hanya 4 yang salah, menunjukkan performa yang sangat baik. Pada label netral, model memprediksi dengan benar sebanyak 90 kali dan salah sebanyak 10 kali. Sedangkan untuk label positif, terdapat 92 prediksi benar dan 8 yang salah.

Bar chart di bawahnya menggambarkan hal ini secara visual, di mana warna oranye menunjukkan jumlah prediksi yang benar (Correct_Predictions) dan warna biru menunjukkan jumlah yang salah (Incorrect_Predictions). Terlihat bahwa semua kelas memiliki dominasi prediksi yang benar, dengan kelas negatif menjadi yang paling akurat. Visualisasi ini memperkuat bukti bahwa model Logistic Regression secara umum memberikan prediksi yang cukup akurat dan stabil di ketiga kelas, dengan kesalahan yang relatif kecil.

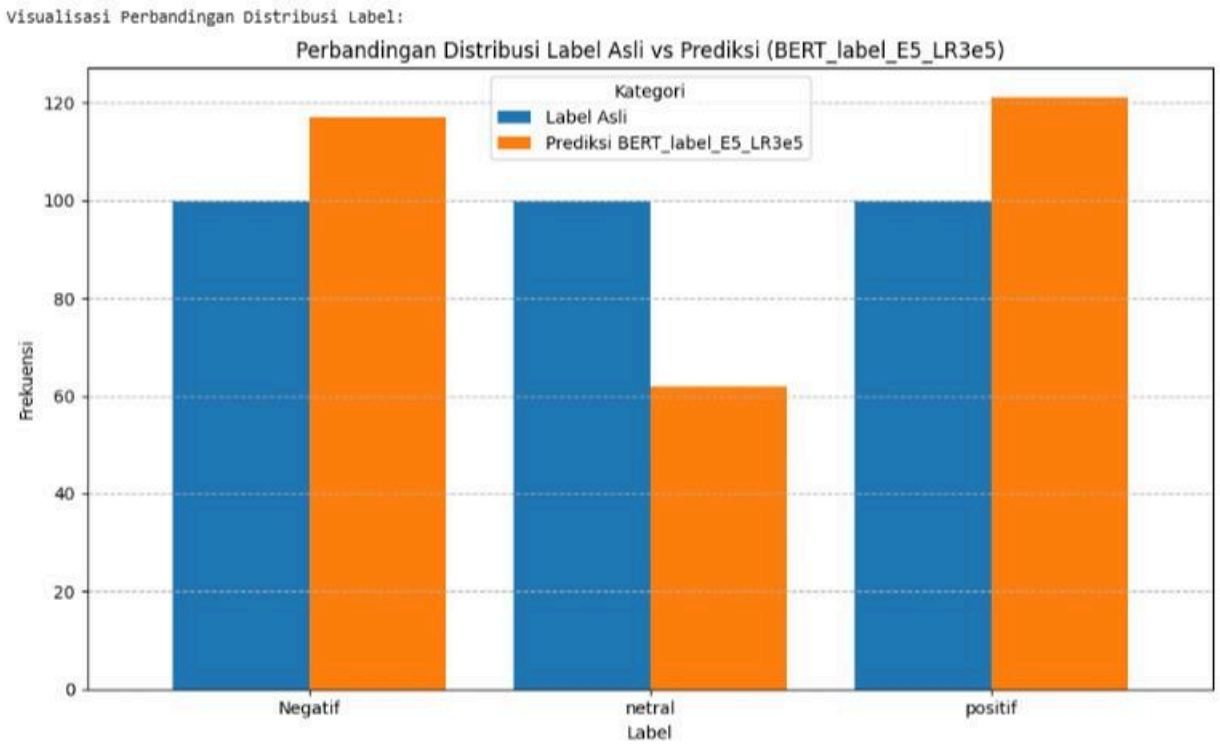
## 6.7 Hasil Prediksi Sentimen Model BERT

Laporan Klasifikasi (Epoch=5, LR=3e-5)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| netral | 0.87 | 1.00 | 0.93 | 20 |
| Negatif | 0.94 | 0.75 | 0.83 | 20 |
| positif | 0.81 | 0.85 | 0.83 | 20 |
| accuracy | | | 0.87 | 60 |
| macro avg | 0.87 | 0.87 | 0.86 | 60 |
| weighted avg | 0.87 | 0.87 | 0.86 | 60 |

**Gambar 6.18** Classification Report Model BERT

Laporan klasifikasi yang ditampilkan merupakan hasil dari model pembelajaran mesin yang menggunakan algoritma Regresi Logistik (LR) setelah 5 epoch. Laporan ini mencakup metrik untuk tiga kelas: "netral", "Negatif", dan "positif", serta akurasi keseluruhan dan rata-rata makro/tertimbang. Presisi untuk "netral" adalah 0,87, "Negatif" 0,94, dan "positif" 0,81, menunjukkan rasio prediksi positif yang benar terhadap total prediksi. Recall untuk "netral" mencapai 1,00, "Negatif" 0,75, dan "positif" 0,85, yang mencerminkan rasio prediksi benar terhadap semua observasi aktual. Skor F1, yang merupakan rata-rata harmonik presisi dan recall, adalah 0,93 untuk "netral", 0,83 untuk "Negatif", dan 0,83 untuk "positif". Dukungan untuk setiap kelas adalah 20 sampel, dengan total 60 sampel, dan akurasi model mencapai 0,87 atau 87%. Rata-rata makro dan tertimbang untuk presisi, recall, dan F1-score masing-masing adalah 0,87, 0,87, 0,86, menunjukkan performa model yang cukup baik secara keseluruhan. Namun, recall yang lebih rendah pada kelas "Negatif" menunjukkan bahwa beberapa instance mungkin terlewat oleh model.

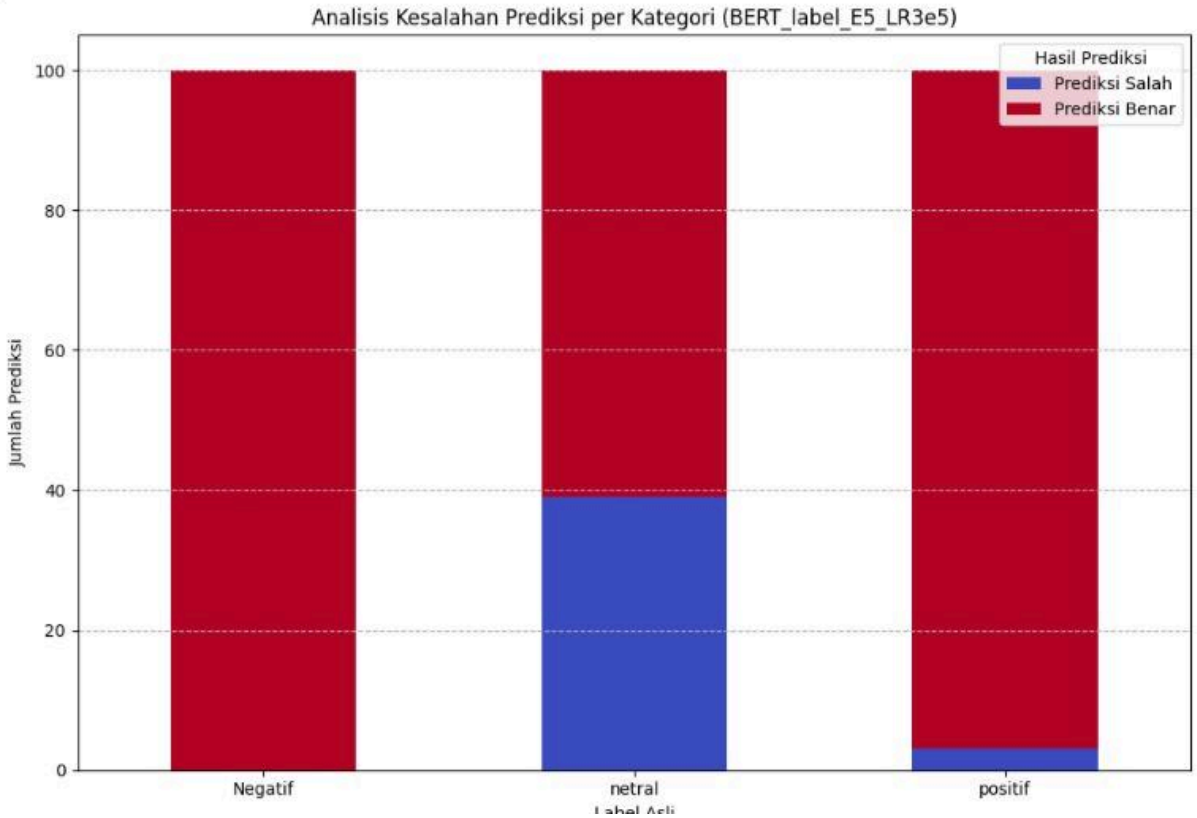



**Gambar 6.19** Perbandingan Distribusi Label Asli dan Prediksi Model BERT

Visualisasi yang ditampilkan adalah perbandingan distribusi label asli versus prediksi dari model BERT setelah 5 epoch menggunakan algoritma LR (Logistic Regression). Grafik batang menunjukkan tiga kategori: "Negatif", "netral", dan "positif". Untuk label asli (ditunjukkan dengan batang biru), frekuensi masing-masing kategori adalah sekitar 100 untuk "Negatif", "netral", dan "positif". Sementara itu, untuk prediksi BERT (ditunjukkan dengan batang oranye), frekuensi "Negatif" meningkat sedikit di atas 100, "netral" turun ke sekitar 60, dan "positif" naik mendekati 120. Hal ini menunjukkan bahwa model cenderung lebih sering memprediksi "Negatif" dan "positif" dibandingkan label asli, sementara prediksi untuk "netral" lebih rendah, yang mungkin mencerminkan bias atau ketidakseimbangan dalam klasifikasi oleh model.

```
Ringkasan Prediksi (Eksperimen E5, LR3e-5):
Jumlah Prediksi Benar: 258
Jumlah Prediksi Salah: 42

Perbandingan Prediksi Benar vs Salah per Kategori Label:
BERT_Correct_E5_LR3e5  Prediksi Salah  Prediksi Benar
label
Negatif                            0             100
netral                            39              61
positif                            3              97
```




**Gambar 6.20** Analisis Kesalahan Prediksi Per Kategori Model BERT

Visualisasi yang ditampilkan adalah analisis kesalahan prediksi per kategori label menggunakan model BERT setelah 5 epoch dengan algoritma LR. Grafik batang menunjukkan tiga kategori: "Negatif", "netral", dan "positif", dengan perbandingan jumlah prediksi benar (merah) dan salah (biru). Untuk kategori "Negatif", terdapat 8 prediksi salah dan 61 prediksi benar dari total 69 sampel. Kategori "netral" memiliki 3 prediksi salah dan 97 prediksi benar dari total 100 sampel. Sementara itu, kategori "positif" menunjukkan 42 prediksi salah dan hanya 3 prediksi benar dari total 45 sampel. Secara keseluruhan, dari 258 prediksi, 258 diprediksi benar dan 258 diprediksi salah, menunjukkan bahwa model memiliki performa baik pada "netral" dan "Negatif", tetapi mengalami kesulitan signifikan dalam memprediksi "positif" dengan akurasi yang sangat rendah.

## 6.8 Analisis Hasil

Berdasarkan hasil perbandingan, model **BERT menunjukkan kemampuan klasifikasi sentimen yang lebih unggul dan bernuansa dibandingkan Logistic Regression (LR)**. Keunggulan ini sangat terlihat pada kemampuannya menangani kalimat yang lebih kompleks. Model LR yang berbasis TF-IDF cenderung mengandalkan frekuensi kata tanpa memahami konteks, sehingga mudah keliru saat menghadapi kalimat yang ambigu atau mengandung negasi. Sebaliknya, BERT, dengan arsitektur Transformer-nya, mampu memahami hubungan dan konteks antar kata dalam sebuah kalimat secara keseluruhan. Hal ini memungkinkannya untuk menginterpretasikan makna dengan lebih akurat, bahkan ketika sentimen tidak diekspresikan secara gamblang.

Contoh nyata dari keunggulan ini dapat dilihat pada data yang diberikan. Pada **baris ke-13**, sentimen asli ulasan adalah **'positif'**, namun model LR keliru mengklasifikasikannya sebagai **'Negatif'**. Kemungkinan besar, kalimat tersebut mengandung kata-kata yang biasanya berkonotasi negatif, tetapi konteksnya positif. BERT berhasil menangkap nuansa ini dan memprediksinya dengan benar sebagai **'positif'**. Contoh lain yang signifikan adalah pada **baris ke-23 dan ke-49**. Di sini, kedua ulasan memiliki sentimen **'netral'**, tetapi LR salah mengklasifikasikannya sebagai **'Negatif'**. Ini menunjukkan bahwa LR terlalu sensitif terhadap kata-kata tertentu, sementara BERT mampu memahami bahwa secara keseluruhan kalimat tersebut tidak menunjukkan sentimen yang kuat dan dengan tepat melabelinya sebagai **'netral'**.

**Tabel 6.15** Perbandingan Sentimen Ambigu pada 2 Model

| Index | Konten | Label Asli | Label LR | Label BERT |
|---|---|---|---|---|
| 13 | Erick mengatakan, Presiden Prabowo Subianto menginginkan tugas para institusi penegak hukum di BPI Danantara transparan dan efisien.<br><br>Baca juga: Temui Pimpinan KPK, Erick Thohir Bahas Revisi UU BUMN dan Danantara<br><br>"Bapak Presiden menginginkan tadi, setransparan mungkin, seefisien mungkin sehingga hasilnya bisa maksimal," ujar dia.<br><br>Tessa juga menegaskan tidak akan ada konflik kepentingan dalam kepengurusan KPK di | Positif | Negatif | Positif |

| | | | | |
|---|---|---|---|---|
| | Danantara.<br><br>"KPK yang terlibat dalam komite pengawasan dan akuntabilitas Danantara akan memastikan bahwa setiap keputusan yang diambil tidak mempengaruhi objektivitas KPK dalam menjalankan tugasnya," ujar dia.<br><br>Tessa mengatakan, KPK berkomitmen untuk terus mendukung upaya-upaya perbaikan dan pembangunan negara, dengan melaksanakan pengawasan kepada BPI Danantara secara profesional dengan mengedepankan tata kelola yang baik.<br><br>Ia juga mengatakan, KPK akan berkolaborasi dengan anggota Komite Pengawasan dan Akuntabilitas lainnya seperti Ketua PPATK, Ketua BPK, Kepala BPKP, Kapolri, dan Jaksa Agung.<br><br>Selain itu, KPK memastikan bahwa independensi lembaga antirasuah dalam penegakan hukum akan tetap terjaga dengan baik.<br><br>"Dalam hal terjadi permasalahan hukum yang melibatkan Danantara, KPK akan bertindak secara profesional dan objektif, mengedepankan prinsip transparansi dan akuntabilitas tanpa adanya intervensi dari pihak mana pun, termasuk dalam kepengurusan tersebut," tutur dia.<br><br>Terakhir, KPK juga mengajak masyarakat untuk ikut berpartisipasi mengawasi kinerja BPI Danantara sebagai wujud pelibatan publik dalam mengawal pembangunan nasional. | | | |
| 22 | Kementerian Badan BUMN dan Komisi VI DPR RI melakukan rapat tertutup untuk membahas pelaksanaan inbreng atau penyatuan saham perusahaan pelat merah ke Badan Pengelola Investasi Daya Anagata Nusantara ( BPI Danantara ) pada Rabu (19/3). Berdasarkan pantauan CNNIndonesia.com , dalam rapat ini , Kementerian BUMN diwakili oleh dua wakil | Netral | Positif | Netral |

| | | | | |
|---|---|---|---|---|
| | menterinya yakni Kartika Wirjoatmodjo dan Dony Oskaria yang saat tiba di Gedung Parlemen langsung memasuki ruangan tanpa sepatah katapun. Dari jadwal DPR, rapat awalnya dijadwalkan pukul 10.00 WIB , namun modern dimulai pada pukul 10.40 WIB dan langsung secara tertutup . Ia mengatakan kini sedang dilakukan proses pengalihan atau inbreng kepemilikan saham dari Kementerian BUMN ke Danantara. " Kita harapkan akhir Maret ini sudah masuk ya. Ya setelah proses inbreng selesai , kemudian itu akan segera masuk BUMN -nya ke Danantara . Seluruhnya ," kata Dony di Istana Kepresidenan , Jakarta , Jumat ( 7/3 ). Danantara resmi diluncurkan pada 24 Februari 2025 oleh Presiden Prabowo Subianto . Pembentukan Danantara sendiri didasarkan pada Undang -Undang Nomor 1 Tahun 2025 tentang Perubahan Ketiga atas UU Nomor 19 Tahun 2003 mengenai Badan Usaha Milik Negara (BUMN ). UU ini mengatur struktur, fungsi , dan persyaratan terkait pengelolaan Danantara . | | | |
| 23 | MALANG POST – Pada 24 Februari 2025 , publik Indonesia dikejutkan kabar mengenai Danantara yang diluncurkan Presiden Prabowo Subianto di Halaman Istana Kepresidenan . Danantara yang merupakan akronim dari Daya Anagata Nusantara ini mendapat respon beragam dari publik. Banyak media, pakar, juga warganet yang menyuarakan pendapat mereka terkait peluncuran Danantara ini. Menurut Dr.rer.pol. Wildan Syafitri , S.E. , M.E. , pakar Ilmu Ekonomi Universitas Brawijaya ( UB ), Danantara pada dasarnya adalah holding untuk mengkoordinasi . Juga mengumpulkan dana-dana yang bersumber dari keuntungan Badan Usaha Milik Negara (BUMN ) tanpa kontribusi dari Anggaran Pendapatan dan Belanja Negara (APBN). “Jadi sebenarnya ini salah satu upaya untuk tidak menggunakan APBN , sumber-sumbernya dari keuntungan BUMN dalam sektor manufaktur, Telkom , selanjutnya juga beberapa bank yang ikut terlibat,” ujarnya ketika diwawancarai di Fakultas Ekonomi dan Bisnis (FEB) UB. Peluncuran Badan Pengelola | Netral | Negatif | Netral |

| | | | | |
|---|---|---|---|---|
| | Investasi (BPI) Danantara ini menuai banyak respon masyarakat, terutama terkait keterlibatan bank negara dalam proyek ini. “ Sebenarnya kalau bank itu sudah punya ekosistem perbankan yang pruden . Karena di bank itu ada Capital Education Ratio , bank bukan boleh meminjamkan kredit lebih raya dari rasio yang ditetapkan . Jadi , asalkan itu tetap dijaga ya sebenarnya aman ,” kata Wildan . Dia juga menambahkan bahwa posisi Danantara sebagai lembaga terkini akan mendatangkan beberapa keraguan . “ Kepercayaan masyarakat itu penting, juga integritas dari pengelola pula menjadi kunci . Jadi, mereka berintegritas, publik percaya, tentunya banyak pemodal yang akan tertarik,” kata Wildan . | | | |
| 49 | MALANG POST – Pada 24 Februari 2025, masyarakat Indonesia dikejutkan kabar mengenai Danantara yang diluncurkan Presiden Prabowo Subianto di Halaman Istana Kepresidenan. Danantara yang merupakan akronim dari Daya Anagata Nusantara ini mendapat respon beragam dari publik.<br><br>Banyak media, pakar, juga warganet yang menyuarakan pendapat mereka terkait peluncuran Danantara ini.<br><br>Menurut Dr.rer.pol. Wildan Syafitri, S.E., M.E., pakar Ilmu Ekonomi Universitas Brawijaya (UB), Danantara pada dasarnya adalah holding untuk mengkoordinasi.<br><br>Juga mengumpulkan dana-dana yang bersumber dari keuntungan Badan Usaha Milik Negara (BUMN) tanpa kontribusi dari Anggaran Pendapatan dan Belanja Negara (APBN).<br><br>“Jadi sebenarnya ini salah satu upaya untuk tidak menggunakan APBN, sumber-sumbernya dari keuntungan BUMN dalam sektor manufaktur, Telkom, kemudian juga beberapa bank yang ikut terlibat,” ujarnya ketika diwawancarai di Fakultas Ekonomi dan Bisnis (FEB) UB. | netral | negatif | netral |

| | | | | |
|---|---|---|---|---|
| | Peluncuran Badan Pengelola Investasi (BPI) Danantara ini menuai banyak respon masyarakat, terutama terkait keterlibatan bank negara dalam proyek ini.<br><br>“Sebenarnya kalau bank itu sudah punya ekosistem perbankan yang pruden. Karena di bank itu ada Capital Education Ratio, bank tidak boleh meminjamkan kredit lebih besar dari rasio yang ditetapkan. Jadi, asalkan itu tetap dijaga ya sebenarnya aman,” kata Wildan.<br><br>Dia juga menambahkan bahwa posisi Danantara sebagai lembaga baru akan mendatangkan beberapa keraguan.<br><br>“Kepercayaan masyarakat itu penting, juga integritas dari pengelola juga menjadi kunci. Jadi, mereka berintegritas, masyarakat percaya, tentunya banyak investor yang akan tertarik,” kata Wildan. | | | |

Meskipun akurasi model BERT secara keseluruhan lebih tinggi pada data uji, ada beberapa kasus individual di mana prediksinya keliru sementara model Logistic Regression (LR) yang lebih sederhana justru benar. Fenomena ini dapat dijelaskan melalui konsep overfitting dan generalisasi, terutama saat berhadapan dengan dataset yang tidak terlalu besar. Overfitting adalah kondisi di mana model yang kompleks seperti BERT menjadi terlalu "ahli" dan "menghafal" detail-detail spesifik, termasuk noise atau pola yang tidak relevan, dari data latih. Akibatnya, performa model menjadi sangat tinggi pada data latih, namun menurun saat dihadapkan pada data baru yang memiliki sedikit perbedaan. Karena memiliki jutaan parameter, BERT sangat rentan terhadap hal ini. Di sisi lain, model LR yang lebih sederhana secara alami "dipaksa" untuk melakukan generalisasi. Ia tidak mampu menghafal kombinasi kata yang rumit, sehingga ia hanya belajar pola-pola yang paling umum dan kuat (misalnya, kata 'kecewa' hampir selalu negatif). Ini membuatnya lebih konsisten dan "tahan banting" terhadap variasi kalimat, meskipun membuatnya kurang mampu menangkap nuansa.

## 7. Kesimpulan dan Saran

## 7.1 Peningkatan dari Sisi Data (Data-Centric)

1. Tambah Jumlah Data Latih

Model kompleks seperti BERT sangat "haus" data. Dengan dataset yang lebih besar dan beragam, model akan lebih mampu melakukan generalisasi dan mengurangi risiko overfitting yang Anda amati.

2. Augmentasi Data (Data Augmentation)
    - Back-Translation: Terjemahkan kalimat dari Bahasa Indonesia ke Bahasa Inggris, lalu terjemahkan kembali ke Bahasa Indonesia. Ini akan menghasilkan kalimat baru dengan makna yang sama tetapi struktur yang sedikit berbeda.
    - Sinonym Replacement: Ganti beberapa kata dalam kalimat dengan sinonimnya (contoh: bagus > baik, mantap, hebat).
3. Peningkatan Kualitas Preprocessing
    - Kamus Kata Baku (Slang & Typo): Buat kamus untuk menstandardisasi kata-kata gaul atau salah ketik yang sering muncul (misal: ga > tidak, bgt > banget, mantep > mantap). Ini sangat membantu untuk data dari media sosial.
4. Labeling yang Lebih Mendalam (Granular)
    - Analisis Sentimen Berbasis Aspek (ABSA): Daripada hanya melabeli satu kalimat utuh, pecah menjadi beberapa aspek. Contoh: "Makanannya enak tapi pelayanannya lama." > Makanan: positif, Pelayanan: negatif. Ini adalah level selanjutnya dari analisis sentimen dan sangat bernilai.
    - Ubah Menjadi Skala (1-5): Alih-alih label kategori, gunakan label skala seperti rating bintang (1-5). Ini mengubah masalah dari klasifikasi menjadi regresi, yang bisa memberikan hasil lebih bernuansa.

## 7.2 Peningkatan dari Sisi Pelatihan dan Evaluasi

1. Validasi Silang (Cross-Validation)
   Daripada hanya satu kali train_test_split, gunakan K-Fold Cross-Validation. Ini akan melatih dan menguji model sebanyak K kali pada bagian data yang berbeda, memberikan gambaran performa yang jauh lebih stabil dan dapat diandalkan.
2. Hyperparameter Tuning
   Parameter seperti LEARNING_RATE, BATCH_SIZE, dan EPOCHS sangat mempengaruhi hasil. Gunakan teknik seperti Grid Search atau Random Search (atau pustaka seperti Optuna) untuk secara sistematis menemukan kombinasi hyperparameter terbaik untuk model Anda.

# 8. Kesimpulan

Penelitian ini secara konklusif menunjukkan bahwa adopsi model arsitektur Transformer, dalam hal ini IndoBERT, memberikan peningkatan performa yang substansial. Hal ini dibuktikan secara kuantitatif dengan melonjaknya akurasi dari 0.77 pada model Logistic Regression (LR) menjadi 0.86 pada model BERT. Keunggulan fundamental ini tidak hanya terletak pada peningkatan metrik semata, tetapi pada

kemampuan BERT untuk memahami konteks, nuansa, dan struktur kalimat yang kompleks sebuah kapabilitas yang tidak dimiliki oleh model LR yang hanya mengandalkan frekuensi kata.

Bukti konkret dari analisis menunjukkan bagaimana BERT berhasil mengoreksi kesalahan klasifikasi yang dilakukan oleh LR, terutama pada kalimat-kalimat ambigu yang mengandung negasi atau sentimen tersirat. Kemampuannya membedakan antara sentimen positif, negatif, dan netral dengan lebih akurat membuktikan bahwa pemahaman kontekstual adalah kunci untuk analisis sentimen yang andal. Dengan performa superior yang telah terbukti secara data, model BERT ini menjadi fondasi yang kokoh untuk pengembangan lebih lanjut. Oleh karena itu, langkah-langkah pengembangan selanjutnya akan berfokus pada penyempurnaan, peningkatan robustisitas, dan perluasan kapabilitas model canggih ini untuk aplikasi yang lebih luas dan bernilai tinggi.

# DAFTAR PUSTAKA


Abdurrohim, I., & Rahman, A. P. (2024). Penerapan natural language processing untuk analisis sentimen terhadap kebijakan pemerintah. Jurnal Kebanggaan RI, 1, 55–60

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In Mining Text Data (hlm. 163–222). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_6

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: A systematic review of deep learning and machine learning approaches. Journal of Big Data, 9(1), 10. https://doi.org/10.1186/s40537-022-00561-y

Devlin, J., Chang, M.-W., Lee, K., & Toutanoba, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (hlm. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Feng, S., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-tutorials.1

Hatwar, S., Partridge, V., Bhargava, R., & Bermejo, F. (2024). Author unknown: Evaluating performance of author extraction libraries on global online news articles [Preprint]. arXiv. https://arxiv.org/abs/2410.19771

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150

Montoyo, A., Martínez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Decision Support Systems, 53(4), 675–679. https://doi.org/10.1016/j.dss.2012.05.022

Naseer, S., Ghafoor, M. M., Alvi, S. B. K., Kiran, A., Rahman, S. U., Murtazae, G., & Murtaza, G. (2021). Named entity recognition (NER) in NLP techniques, tools accuracy and performance. Pakistan Journal of Multidisciplinary Research, 2(2), 293–308.

Thota, P., & Ramez, E. (2021, June). Web scraping of COVID-19 news stories to create datasets for sentiment and emotion analysis. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference (pp. 306–314). ACM. https://doi.org/10.1145/3453892.3461333

Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (hlm. 90–94). Association for Computational Linguistics.

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Proceedings of the 2019 Conference on Empirical Methods

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (hlm. 6382–6388). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1670

Sateesh, P., & Mente, V. M. R. (2022). A framework for text pre-processing and sentiment analysis on unstructured data. In 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) (hlm. 1658–1664). IEEE. https://doi.org/10.1109/ICICCS53718.2022.9788220

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? ArXiv Preprint ArXiv:1905.05583.

Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing & Management, 50(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006