

**ANALISIS SENTIMEN BERITA DAN OPINI TERHADAP  
PEMINDAHAN IBU KOTA NUSANTARA (IKN) MELALUI WEB  
SCRAPING PADA MEDIA ONLINE**



**Anggota Kelompok :**

Muhammad Alarik Daviarsyah      5026221015

Dewi Maharani      5026221046

Andika Insan Patria      5026221211

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

**2025**

## DAFTAR ISI

<b>DAFTAR ISI</b>	<b>2</b>
<b>ABSTRAK</b>	<b>3</b>
<b>ABSTRACT</b>	<b>4</b>
<b>BAB I. PENDAHULUAN</b>	<b>5</b>
1.1 Latar Belakang	5
1.2 Rumusan Masalah	6
1.3 Tujuan Penelitian	6
1.4 Manfaat Penelitian	6
<b>BAB II. TINJAUAN PUSTAKA</b>	<b>7</b>
2.1 Akuisisi Data	7
2.2 Preprocessing Data	7
2.3 Sentiment Analysis	8
2.4 TF-IDF	8
2.5 POS Tagging	9
2.6 NER	9
2.7 Transformer dan BERT	10
<b>BAB III. DATA ACQUISITION</b>	<b>12</b>
3.1 Google Dorking Penggunaan Query Search Engine	12
3.2 Google scraping Menggunakan BeautifulSoup	13
3.3 Hasil Tautan Berita Mengenai IKN	13
3.4 Pelabelan Manual dan Persiapan Dataset untuk Fine-Tuning BERT	14
3.5 Implementasi Model BERT untuk Analisis Sentimen	14
3.5.1 Tokenisasi dengan BertTokenizer	15
3.5.2 Pembagian Dataset (Train, Validation, Test)	16
3.5.3 Fine-tuning Menggunakan transformers (Hugging Face)	16
3.5.4 Evaluasi Model: Akurasi, Precision, Recall, F1-Score	18
<b>BAB IV. DATA PREPROCESSING</b>	<b>19</b>
4.1 Data Preprocessing Artikel Berita	19
4.2 Augmentasi dan Penyeimbangan Dataset Sentimen	23
4.2.1 Strategi Penyeimbangan Kelas	24
4.2.2 Teknik Augmentasi Data	24
4.2.3 Hasil Penyeimbangan Dataset	25
<b>BAB V. DEFINISI DATASET</b>	<b>26</b>
5.1 Dataset Link Artikel Berita	26
5.2 Dataset Hasil Preprocessing	26
<b>BAB VI. ANALISIS DATA</b>	<b>28</b>
6.1 Analisis Sentimen	28
6.2 Analisis TF-IDF	31
6.3 Analisis POS dan NER	35
6.4 Analisis Sentimen dengan menggunakan Model IndoBERT	39
6.4 Perbandingan Model BERT dan TF-IDF/SVM	40

<b>BAB VII. KESIMPULAN</b>	<b>42</b>
<b>DAFTAR PUSTAKA</b>	<b>43</b>
<b>LAMPIRAN</b>	<b>45</b>

## ABSTRAK

Penelitian ini mengevaluasi persepsi media daring terhadap rencana pemindahan Ibu Kota Negara (IKN) dari Jakarta ke Kalimantan Timur melalui analisis sentimen dan linguistik berbasis Natural Language Processing (NLP). Sebanyak 162 artikel berita nasional dikumpulkan menggunakan teknik web scraping Google Dorking, kemudian dipraproses melalui tahapan cleaning, tokenisasi, penghapusan stopword, stemming, dan lemmatisasi. Analisis sentimen dilakukan dengan model IndoBERT—yang di-fine-tune untuk Bahasa Indonesia—and dibandingkan dengan dua pendekatan tradisional berbasis TF-IDF, yaitu Logistic Regression dan Support Vector Machine (SVM).

Hasil analisis menunjukkan mayoritas artikel bersentimen netral, sementara sentimen positif menonjolkan narasi kemajuan, teknologi, dan pemerataan ekonomi; dan sentimen negatif menyoroti isu sosial lingkungan, terutama hak masyarakat adat. Lonjakan pemberitaan teridentifikasi pada Agustus 2019, Januari 2022, April 2023, Agustus 2024, dan Februari 2025, bertepatan dengan tonggak penting proyek IKN. Dari sisi performa, IndoBERT mencatat akurasi sebesar 82% dengan F1-score 0,87 (positif) dan 0,83 (negatif), mengungguli model TF-IDF + SVM dengan akurasi 78%. Confusion matrix menunjukkan IndoBERT lebih akurat dalam mengenali sentimen positif dan negatif, meskipun masih terdapat ambiguitas pada kelas netral.

Temuan ini menegaskan keunggulan IndoBERT dalam memahami nuansa linguistik Bahasa Indonesia dan menangkap konteks semantik yang lebih dalam dibandingkan pendekatan tradisional. Hasil penelitian ini memberikan wawasan bagi pemerintah dan media dalam menyusun strategi komunikasi yang lebih responsif terhadap dinamika opini publik terkait proyek IKN.

**Kata kunci:** Ibu Kota Negara (IKN), Analisis Sentimen, Natural Language Processing (NLP), IndoBERT

## ***ABSTRACT***

*This study evaluates online media perceptions of the planned relocation of Indonesia's national capital (IKN) from Jakarta to East Kalimantan using sentiment and linguistic analysis based on Natural Language Processing (NLP). A total of 162 national news articles were collected via Google Dorking web-scraping and pre-processed through cleaning, tokenization, stop-word removal, stemming, and lemmatization. Sentiment classification employed a fine-tuned IndoBERT model for Bahasa Indonesia and was compared with two traditional TF-IDF baselines—Logistic Regression and Support Vector Machine (SVM).*

*The results show that most articles exhibit a **neutral** sentiment, while **positive** sentiment emphasizes narratives of progress, technology, and economic equity, and **negative** sentiment highlights socio-environmental issues, particularly the rights of Indigenous communities. Peaks in coverage occurred in August 2019, January 2022, April 2023, August 2024, and February 2025, coinciding with key project milestones. Performance-wise, IndoBERT achieved an accuracy of 82 % with F1-scores of 0.87 (positive) and 0.83 (negative), outperforming the TF-IDF + SVM model, which recorded 78 % accuracy. The confusion matrix confirms IndoBERT's superior ability to recognize positive and negative sentiments, although some ambiguity remains in the neutral class.*

*These findings underscore IndoBERT's advantage in capturing the nuanced semantics of Bahasa Indonesia over traditional methods. The results provide valuable insights for government and media to devise communication strategies that are more responsive to the dynamics of public opinion regarding the IKN project.*

**Keywords:** Nusantara Capital City (IKN), Sentiment Analysis, Natural Language Processing (NLP), IndoBERT

## BAB I. PENDAHULUAN

### 1.1 Latar Belakang

Perkembangan teknologi digital telah membawa dampak signifikan di berbagai sektor kehidupan, termasuk dalam penyebaran informasi dan pembentukan opini publik. Akses masyarakat terhadap berita daring (*online news*) kini semakin mudah dan cepat, memungkinkan isu-isu strategis nasional untuk menyebar luas dan membentuk persepsi kolektif dalam waktu singkat (Feldman, 2013). Di Indonesia, salah satu isu strategis nasional yang mendapatkan sorotan publik adalah pemindahan Ibu Kota Negara (IKN) dari Jakarta ke Ibu Kota Nusantara (IKN) di Kalimantan Timur. Proyek ini bukan hanya sekadar relokasi administratif, tetapi juga merupakan bagian dari transformasi besar dalam tata kelola pemerintahan, pembangunan berkelanjutan, dan pemerataan ekonomi antarwilayah (Sekretariat Negara Republik Indonesia, 2022).

Seiring dengan berjalannya pembangunan IKN, media massa terus memberitakan perkembangan proyek ini dari berbagai perspektif, baik yang bersifat informatif maupun opini. Pemberitaan tersebut mencakup aspek kebijakan, pembangunan infrastruktur, dampak lingkungan, serta respons masyarakat terhadap proyek ini. Dalam era digital, artikel berita daring tidak hanya menyampaikan informasi, tetapi juga dapat berfungsi sebagai indikator sentimen publik terhadap suatu isu (Pang & Lee, 2008). Melalui pemberitaan yang masif, media berperan tidak hanya sebagai penyampai informasi, tetapi juga sebagai agen yang membentuk cara pandang masyarakat terhadap isu yang diberitakan. Oleh karena itu, analisis media menjadi pendekatan penting dalam memahami opini dan respons publik terhadap isu strategis nasional ini.

Mengingat pesatnya jumlah artikel berita yang tersebar di berbagai platform digital, diperlukan metode yang efisien dan sistematis untuk mengumpulkan dan menganalisis data dalam skala besar. Salah satu teknik yang dapat digunakan untuk tujuan ini adalah *web scraping*, yang memungkinkan pengumpulan data artikel berita secara otomatis dari portal media daring (Munzert et al., 2014). Setelah data terkumpul, langkah selanjutnya adalah melakukan tahapan *preprocessing*, seperti pembersihan data, tokenisasi, dan normalisasi, untuk mempersiapkan data sebelum dilakukan analisis sentimen. Pada tahap analisis sentimen, penelitian ini menggunakan model *machine learning* klasik, yaitu Support Vector Machine (SVM), yang dikenal efektif dalam menangani klasifikasi teks karena kemampuannya dalam memisahkan kelas dengan margin maksimal, serta performa yang baik pada data berdimensi tinggi seperti representasi vektor dari teks.

Sebagai tindak lanjut dari penelitian ini, dilakukan peningkatan performa analisis sentimen dengan mengimplementasikan model BERT secara lebih optimal. Fokus utama terletak pada evaluasi dan penyempurnaan model yang telah digunakan sebelumnya, baik dari sisi akurasi klasifikasi maupun ketepatan dalam menangani kalimat yang kompleks atau ambigu. Peningkatan ini mencakup tahapan seperti tokenisasi dengan BertTokenizer, pemisah

dataset menjadi data latih dan validasi, proses fine-tuning menggunakan transformers dari Hugging Face, serta evaluasi dengan metrik seperti akurasi, precision, recall, dan F1-score. Dengan pendekatan ini, diharapkan hasil analisis sentimen menjadi lebih andal dan dapat memberikan gambaran yang lebih akurat terhadap opini publik terkait isu strategis nasional seperti pemindahan IKN.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan, maka rumusan masalah dalam pembahasan ini adalah sebagai berikut:

1. Bagaimana persepsi publik terhadap proyek Ibu Kota Nusantara (IKN) sebagaimana tercermin dalam analisis sentimen pada artikel berita daring?
2. Apa saja isu utama atau kata kunci yang paling sering muncul dalam pemberitaan mengenai IKN?
3. Bagaimana tren perubahan sentimen publik terhadap IKN dari waktu ke waktu berdasarkan isi artikel berita daring?
4. Bagaimana penerapan model BERT yang telah *di-fine-tuned* dapat meningkatkan akurasi klasifikasi sentimen dibandingkan pendekatan sebelumnya?

## 1.3 Tujuan Penelitian

1. Menilai persepsi publik terhadap proyek Ibu Kota Nusantara (IKN) melalui analisis sentimen pada artikel berita daring.
2. Menemukan isu-isu utama atau kata kunci yang sering dibahas dalam pemberitaan mengenai IKN.
3. Mengamati perubahan atau tren sentimen publik terhadap IKN dari waktu ke waktu berdasarkan isi artikel berita.
4. Meningkatkan akurasi klasifikasi sentimen artikel berita daring mengenai IKN melalui fine-tuning model BERT yang telah *di-fine-tuned* untuk Bahasa Indonesia
5. Membandingkan performa analisis sentimen yang dilakukan dengan model transformer terhadap hasil pendekatan sebelumnya secara kuantitatif.

## 1.4 Manfaat Penelitian

1. Memberikan wawasan kepada pembuat kebijakan mengenai persepsi publik terhadap proyek IKN, sehingga dapat digunakan sebagai bahan evaluasi dan strategi komunikasi yang lebih efektif.
2. Membantu media massa dan praktisi komunikasi dalam memahami framing pemberitaan serta dampaknya terhadap opini publik
3. Mendukung transparansi informasi dengan menyajikan pola sentimen yang berkembang di masyarakat secara data-driven
4. Menyediakan hasil analisis sentimen yang lebih akurat menggunakan BERT untuk mendukung pengambilan keputusan

5. Menunjukkan potensi penggunaan model transformer dalam pemrosesan bahasa alami untuk Bahasa Indonesia, khususnya dalam analisis opini publik di sektor pemerintahan.

## BAB II. TINJAUAN PUSTAKA

### 2.1 Akuisisi Data

Akuisisi data merupakan langkah awal yang krusial dalam proses pengolahan bahasa alami (Natural Language Processing/NLP), karena kualitas data yang diperoleh akan sangat mempengaruhi hasil akhir dari model yang dikembangkan. Proses akuisisi ini mencakup pengumpulan teks dari berbagai sumber seperti media sosial, situs berita, ulasan pengguna, blog, hingga dokumen resmi. Variasi dan keberagaman data menjadi penting untuk menangkap kekayaan bahasa dan konteks yang digunakan oleh pengguna (Indurkhya & Damerau, 2010).

Terdapat dua pendekatan utama dalam akuisisi data, yaitu manual dan otomatis. Akuisisi manual dilakukan dengan cara menyalin atau mengunduh teks secara langsung dari sumber yang relevan. Meskipun cara ini memakan waktu dan tenaga, ia memberikan kendali penuh atas pemilihan data. Sebaliknya, akuisisi otomatis memanfaatkan teknologi seperti web scraping dengan library Python seperti BeautifulSoup atau Scrapy, serta penggunaan Application Programming Interface (API) dari platform seperti Twitter, Reddit, atau YouTube untuk mendapatkan data secara masif dan cepat (Bird, Klein, & Loper, 2009). Cairns & Meng (2020) menyoroti pentingnya dataset yang relevan, seperti ulasan Amazon, yang diadaptasi untuk analisis sentimen. Dalam penelitian ini, *web scraping* digunakan untuk mengumpulkan artikel berita IKN dari portal media daring, dengan memastikan cakupan berbagai perspektif seperti kebijakan, infrastruktur, dan dampak lingkungan.

Namun, proses akuisisi data juga menuntut kepatuhan terhadap aspek hukum dan etika, terutama terkait hak cipta dan privasi. Data yang diambil harus mematuhi ketentuan yang berlaku seperti General Data Protection Regulation (GDPR) di Eropa atau Terms of Service dari masing-masing platform. Selain itu, penting untuk menyaring data yang bias atau tidak representatif agar hasil analisis tidak menyesatkan (Cambria & White, 2014).

### 2.2 Preprocessing Data

Preprocessing data adalah tahap penting dalam pipeline NLP yang bertujuan untuk mengubah teks mentah menjadi representasi yang bersih dan terstruktur, sehingga dapat diproses lebih lanjut oleh algoritma. Tugas utama pada tahap ini meliputi tokenisasi, penghapusan stop words, stemming, lemmatization, normalisasi, hingga penghapusan karakter khusus. Tahapan ini mengurangi kompleksitas dan membantu dalam membangun representasi fitur yang lebih bermakna (Kowsari et al., 2019).

Tokenisasi merupakan proses memecah kalimat menjadi unit-unit kecil seperti kata atau frasa. Stop words, yaitu kata-kata umum yang tidak memberikan kontribusi signifikan terhadap makna teks seperti "yang", "dan", "dari", biasanya dihapus untuk mengurangi noise. Stemming dan lemmatization digunakan untuk mengubah kata ke bentuk dasarnya, misalnya

"berlari", "berlari-lari", dan "pelari" semuanya dikembalikan ke bentuk "lari" (Manning, Raghavan, & Schütze, 2008).

Selain teknik dasar tersebut, preprocessing juga dapat mencakup filtering emoticon, ekspansi singkatan, hingga transformasi teks menjadi huruf kecil. Dalam konteks Bahasa Indonesia, preprocessing bisa lebih kompleks karena morfologi yang kaya, penggunaan bahasa gaul, dan campuran bahasa. Oleh karena itu, penggunaan kamus khusus atau algoritma NLP lokal sering dibutuhkan (Jurafsky & Martin, 2020).

### 2.3 Sentiment Analysis

Analisis sentimen merupakan teknik NLP yang digunakan untuk mengidentifikasi dan mengklasifikasikan opini atau emosi yang terkandung dalam teks, apakah bersifat positif, negatif, atau netral. Penerapannya sangat luas, mulai dari bisnis untuk menganalisis ulasan pelanggan, hingga politik untuk memahami persepsi publik terhadap kandidat atau kebijakan tertentu (Liu, 2012). Cairns & Meng (2020) menunjukkan bahwa analisis sentimen dapat mengungkap pola opini publik, seperti dalam ulasan produk, yang relevan untuk menganalisis pemberitaan media daring.

Secara umum, terdapat dua pendekatan dalam analisis sentimen: berbasis leksikon dan berbasis pembelajaran mesin. Pendekatan leksikon mengandalkan kamus kata yang telah ditetapkan nilai sentimennya, sedangkan pendekatan pembelajaran mesin melibatkan pelatihan model klasifikasi seperti Naive Bayes, Support Vector Machine (SVM), atau deep learning seperti LSTM (Long Short-Term Memory) menggunakan data berlabel (Medhat, Hassan, & Korashy, 2014).

Tantangan dalam analisis sentimen meliputi deteksi sarkasme, ambiguitas makna, dan konteks budaya. Kalimat seperti "bagus banget pelayanannya, sampai-sampai saya nggak mau balik lagi" memerlukan pemahaman konteks dan intonasi yang tidak mudah ditangkap oleh algoritma standar. Oleh karena itu, model analisis sentimen canggih sering menggabungkan pendekatan linguistik dan statistik untuk mencapai akurasi yang lebih tinggi (Cambria et al., 2013).

### 2.4 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) adalah teknik pembobotan kata yang banyak digunakan dalam pencarian informasi dan pengolahan teks. TF mencerminkan seberapa sering suatu kata muncul dalam dokumen, sedangkan IDF menunjukkan seberapa jarang kata tersebut muncul dalam keseluruhan korpus. Dengan mengalikan keduanya, TF-IDF memberikan nilai tinggi pada kata yang sering muncul dalam dokumen tetapi jarang ditemukan di dokumen lain (Rajaraman & Ullman, 2011).

Dalam praktiknya, TF-IDF membantu mengidentifikasi kata-kata penting dalam dokumen. Sebagai contoh, dalam artikel tentang pariwisata Yogyakarta, kata "Yogyakarta" mungkin memiliki TF tinggi, tetapi IDF rendah karena muncul di banyak artikel, sehingga bobot TF-IDF-nya moderat. Sebaliknya, kata "Keraton" mungkin jarang muncul di artikel lain, sehingga mendapat bobot TF-IDF yang lebih tinggi (Manning, Raghavan, & Schütze, 2008).

Meski efektif, TF-IDF memiliki keterbatasan. Ia tidak mempertimbangkan urutan kata atau konteks semantik. Oleh karena itu, metode ini tidak mampu menangkap sinonim atau perbedaan makna tergantung konteks. Untuk aplikasi yang memerlukan pemahaman makna yang lebih dalam, TF-IDF sering dikombinasikan dengan teknik lain seperti word embeddings (Mikolov et al., 2013).

## 2.5 POS Tagging

Part-of-Speech (POS) tagging adalah proses menetapkan kategori sintaktik pada setiap kata dalam kalimat, seperti kata benda, kata kerja, atau kata sifat. POS tagging penting untuk membantu algoritma NLP memahami struktur kalimat dan menentukan fungsi kata dalam konteksnya. Ini menjadi dasar bagi analisis gramatiskal lebih lanjut seperti parsing atau dependency analysis (Jurafsky & Martin, 2020).

Terdapat beberapa metode POS tagging, mulai dari pendekatan berbasis aturan linguistik, hingga pendekatan statistik seperti Hidden Markov Model (HMM) dan Conditional Random Fields (CRF). Dengan kemajuan teknologi, kini banyak POS tagger yang menggunakan deep learning seperti Bidirectional LSTM (BiLSTM) dan Transformer-based models untuk meningkatkan akurasi tagging (Akbik, Blythe, & Vollgraf, 2018).

POS tagging membantu dalam mengatasi ambiguitas. Misalnya, kata "lari" dapat berfungsi sebagai kata benda maupun kata kerja tergantung konteksnya. Dengan menggunakan POS tagging, sistem dapat menentukan tag yang sesuai dan membantu dalam proses seperti ekstraksi entitas atau analisis sentimen (Manning et al., 2008).

## 2.6 NER

Named Entity Recognition (NER) adalah proses untuk mengenali dan mengklasifikasikan entitas bernama dalam teks seperti nama orang, lokasi, organisasi, tanggal, dan lain-lain. NER sangat penting dalam ekstraksi informasi, sistem tanya jawab, dan pelacakan topik berita, karena memungkinkan sistem untuk memahami siapa, di mana, dan kapan dalam konteks kalimat (Nadeau & Sekine, 2007).

NER awalnya dikembangkan dengan aturan dan leksikon, tetapi pendekatan modern mengandalkan pembelajaran mesin dan deep learning. Conditional Random Fields (CRF) digunakan secara luas untuk tagging berurutan, sementara model seperti BiLSTM dan Transformer (misalnya BERT) digunakan untuk meningkatkan pemahaman konteks dan hubungan antar kata (Devlin et al., 2019).

Tantangan dalam NER mencakup penanganan entitas ambigu, variasi format penulisan, dan konversi konteks. Misalnya, "Apple" bisa merujuk pada perusahaan teknologi atau buah, tergantung konteks. Untuk mengatasi ini, model NER memerlukan pelatihan dengan korpus yang luas dan bervariasi, serta strategi contextual embedding (Yadav & Bethard, 2019).

## 2.7 Transformer dan BERT

Model *transformer* merupakan fondasi revolusioner dalam pengembangan algoritma Natural Language Processing (NLP) modern, diperkenalkan oleh Vaswani et al. (2017) melalui makalah berjudul “*Attention Is All You Need*”. Keunggulan utama dari arsitektur transformer terletak pada mekanisme *self-attention* yang memungkinkan model memahami hubungan antar kata dalam satu kalimat, baik yang berdekatan maupun berjauhan, secara paralel dan efisien. Mekanisme ini sangat membantu dalam memahami konteks kalimat panjang dan kompleks, yang seringkali tidak dapat ditangani secara efektif oleh model sekuensial tradisional seperti RNN dan LSTM.

Salah satu model paling berpengaruh yang dibangun di atas arsitektur transformer adalah BERT (Bidirectional Encoder Representations from Transformers), dikembangkan oleh Google (Devlin et al., 2019). BERT dilatih secara *bidirectional*, artinya model membaca kalimat dari dua arah (kiri ke kanan dan kanan ke kiri) secara simultan. Pendekatan ini memungkinkan BERT memahami makna kontekstual kata secara lebih akurat dibandingkan model *word embedding* tradisional seperti Word2Vec atau GloVe yang hanya menangkap representasi kata secara statis dan satu arah. BERT juga mengandalkan pendekatan *pretraining + fine-tuning*, di mana model pertama-tama dilatih pada korpus besar untuk memahami struktur umum bahasa (pretrained knowledge), kemudian disesuaikan (fine-tuned) untuk tugas spesifik seperti klasifikasi sentimen, ekstraksi entitas, atau pertanyaan-jawaban.

Keunggulan BERT dalam analisis sentimen telah dibuktikan dalam berbagai studi. Misalnya, menurut Sun et al. (2019), model BERT secara signifikan mengungguli baseline seperti LSTM dan SVM dalam klasifikasi opini karena kemampuannya menangkap nuansa semantik yang lebih dalam dan fleksibel terhadap variasi sintaksis. Dalam konteks Bahasa Indonesia, model seperti IndoBERT dan IndoBERT-lite yang telah di-*fine-tune* pada korpus lokal menawarkan performa tinggi dalam memahami struktur bahasa dan idiom lokal yang tidak dapat ditangkap oleh model umum berbasis bahasa Inggris. Hal ini sangat relevan untuk kasus seperti analisis sentimen pada pemberitaan media daring mengenai isu strategis nasional seperti IKN, di mana konteks lokal dan makna implisit menjadi sangat penting.

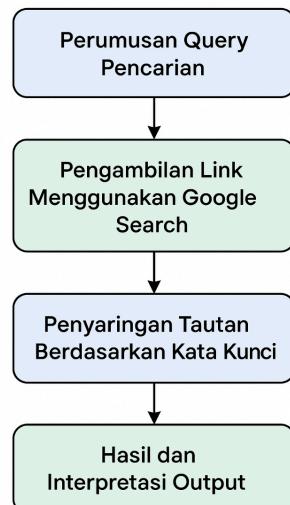
Mulai dari representasi kata kontekstual dan fleksibilitas fine-tuning BERT dan turunannya seperti RoBERTa dan IndoBERT menjelma menjadi standar baru dalam klasifikasi sentimen modern. Dibandingkan pendekatan leksikal atau model statistik konvensional, BERT lebih mampu menangani kalimat ambigu, opini tersirat, bahkan ekspresi sarkasme, yang sering kali menjadi tantangan utama dalam pemodelan sentimen berbasis teks.

*Bidirectional Encoder Representations from Transformers* (BERT), dikembangkan oleh Devlin et al. (2018), memanfaatkan pelatihan bidirectional untuk menangkap konteks kata dari kedua arah. BERT dilatih dengan dua tugas: *Masked Language Model* (MLM) untuk memprediksi kata yang disembunyikan dan *Next Sentence Prediction* (NSP) untuk memahami hubungan antar kalimat. Proses pretraining pada korpus besar diikuti oleh fine-tuning untuk tugas spesifik, seperti analisis sentimen, menghasilkan performa unggul (Cairns & Meng, 2020).

Cairns & Meng (2020) menunjukkan bahwa BERT mencapai akurasi hingga 98% dalam analisis sentimen ulasan Amazon setelah fine-tuning. Proses fine-tuning melibatkan tokenisasi dengan *BertTokenizer*, pemisahan dataset latih dan validasi, serta optimasi hiperparameter seperti learning rate ( $2\text{e-}5$  hingga  $5\text{e-}5$ ), batch size (8, 16, 32), dan jumlah epoch (2-4). Dalam penelitian ini, BERT di-fine-tuned untuk Bahasa Indonesia menggunakan pustaka HuggingFace untuk mengklasifikasikan sentimen artikel berita IKN, dengan evaluasi berbasis akurasi, precision, recall, dan F1-score.

### BAB III. DATA ACQUISITION

Akuisisi data adalah tahap awal yang sangat penting dalam proses pengolahan informasi berbasis web, terutama dalam konteks penggalian teks atau text mining. Dalam kajian data modern, kualitas data yang diperoleh dari tahap akuisisi akan sangat menentukan validitas dan efektivitas analisis selanjutnya.

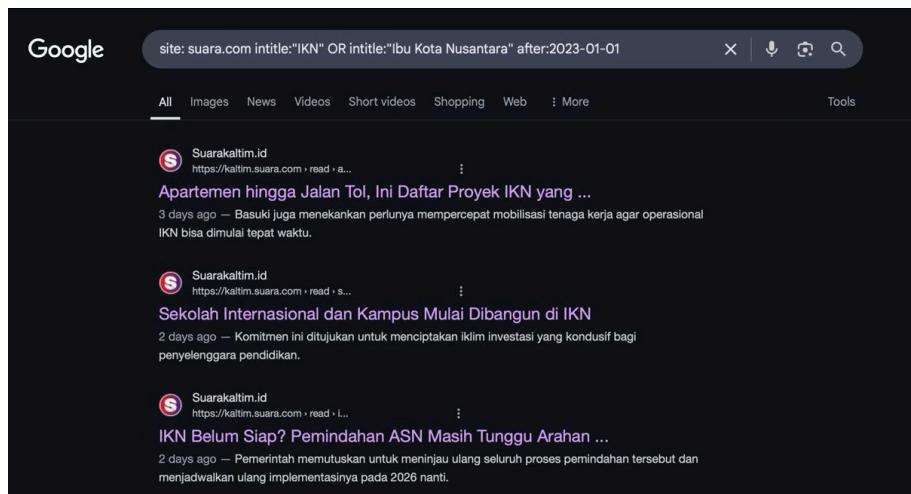


**Gambar 3.1** Diagram Visualisasi Akuisisi Data

Proses ini meliputi pencarian, pengumpulan, hingga penyaringan data dari berbagai sumber daring seperti situs berita, forum publik, dan media sosial (Indurkhy & Damerau, 2010).

#### 3.1 Google Dorking Penggunaan Query Search Engine

Perumusan query merupakan langkah awal yang sangat penting dalam proses akuisisi data dari mesin pencari seperti Google. Query yang dirancang secara tepat memungkinkan sistem untuk mengakses informasi yang relevan, akurat, dan sesuai dengan kebutuhan analisis. Dalam konteks penelitian ini, fokus utama adalah memperoleh berita-berita terkini terkait dengan pembangunan dan perkembangan Ibu Kota Nusantara (IKN) di Indonesia.



Gambar 3.1 Proses Google Dorking

### 3.2 Google scraping Menggunakan BeautifulSoup

Setelah query ditentukan, tahap berikutnya adalah melakukan pencarian otomatis dan mengumpulkan tautan dari hasil pencarian. Proses ini diotomatisasi menggunakan modul Python googlesearch, yang memungkinkan akses ke halaman pencarian Google tanpa harus mengandalkan scraping langsung dari hasil HTML pencarian. Fungsi yang dikembangkan, yakni get\_news\_links, menerima parameter query dan jumlah tautan yang ingin diambil. Sistem kemudian menyaring hanya tautan yang mengandung kata seperti news atau berita dalam URL-nya, dengan asumsi bahwa tautan-tautan tersebut mengarah pada konten berita yang sahih.

Pendekatan ini mempermudah proses pencarian informasi secara cepat dan efisien, terutama dalam jumlah besar (massive scale). Metode otomatisasi seperti ini banyak dimanfaatkan dalam riset NLP modern karena mampu mengurangi waktu dan biaya dalam proses pengumpulan data awal (Bird, Klein, & Loper, 2009). Namun demikian, pendekatan ini tetap harus mematuhi batasan-batasan etis dan teknis dari API yang digunakan, agar tidak melanggar kebijakan layanan dari penyedia data.

### 3.3 Hasil Tautan Berita Mengenai IKN

Berdasarkan proses akuisisi data melalui query Google Dorking dan scraping konten dengan BeautifulSoup, diperoleh sejumlah tautan berita yang secara eksplisit membahas topik Ibu Kota Nusantara. Tautan-tautan ini kemudian diverifikasi keberadaan kata kunci utama dan disusun dalam daftar yang siap digunakan untuk tahap analisis selanjutnya.

```

Skipping https://wartaterkini.news/masyarakat-antusias-pahami-kepindahan-ibu-kota-di-nusantara-fair-2024/ due to error: ('Connection aborted.', RemoteDisconnected('Remote end closed connection without response'))
Skipping https://topmetro.news/tag/ikn-nusantara/ due to error: HTTPSConnectionPool(host='topmetro.news', port=443): Read timed out. (read timeout=5)
Skipping https://wartaterkini.news/category/ibu-kota-nusantara/ due to error: ('Connection aborted.', ConnectionResetError(104, 'Connection reset by peer'))
Skipping https://fakta.news/berita/terungkap-ibu-kota-baru-di-kalimantan-timur-bernama-nusantara due to error: HTTPSConnectionPool(host='fakta.news', port=443): Read timed out. (read timeout=5)
Skipping https://www.kitakini.news/news/12327//kementerian-pupr-telang-proyek-infrastruktur-publik-di-ibu-kota-nusantara-sekolah-pasar-dan-puskesmas/ due to error: HTTPSConnectionPool(host='www.kitakini.news', port=443): Read timed out. (read timeout=5)
Skipping https://jejaknegeri.news/kalimantan-timur/sekda-kaltim-beberkan-persiapan-jelang-perayaan-hut-ke-79-ri-di-ikn/ due to error: HTTPSConnectionPool(host='jejaknegeri.news', port=443): Max retries exceeded with url: /kalimantan-timur/sekda-kaltim-beberkan-persiapan-jelang-perayaan-hut-ke-79-ri-di-ikn/ (Caused by SSLError(SSLCertVerificationError(1, '[SSL: CERTIFICATE_VERIFY_FAILED] certificate verify failed: Hostname mismatch, certificate is not valid for 'jejaknegeri.news'. (_ssl.c:1016)')))
Skipping https://investigasi.news/nasional/pulau-taliabu/bupati-pulau-taliabu-hadiri-pertemuan-ikn-bersama-presiden-dan-517-kepala-daerah/ due to error: HTTPSConnectionPool(host='investigasi.news', port=443): Read timed out. (read timeout=5)
Skipping https://fakta.news/berita/jadi-calon-kuat-kepala-otorita-ikn-nusantara-siapakah-bambang-susantono due to error: HTTPSConnectionPool(host='fakta.news', port=443): Read timed out. (read timeout=5)
Skipping https://www.klausabontang.news/2023/11/07/kutim-jadi-superhub-ekonomi-ikn-nusantara-dengan-kek-maloy-ini-kata-bupati-ardiansyah/ due to error: HTTPSConnectionPool(host='www.klausabontang.news', port=443): Max retries exceeded with url: /2023/11/07/kutim-jadi-superhub-ekonomi-ikn-nusantara-dengan-kek-maloy-ini-kata-bupati-ardiansyah/ (Caused by NameResolutionError('<urllib3.connection.HTTPSConnection object at 0x7909804ed00>: Failed to resolve 'www.klausabontang.news' ([Errno -2] Name or service not known)'))
Skipping https://berita.news/2025/05/05/pemko-makassar-alihkan-efisiensi-anggaran-untuk-masyarakat-pulau/ due to error: HTTPSConnectionPool(host='berita.news', port=443): Read timed out. (read timeout=5)
Skipping https://www.malanesia.news/sah-ruu-ikn-jadi-uu-dan-nusantara-jadi-nama-ibu-kota-yang-baru/ due to error: HTTPSConnectionPool(host='www.malanesia.news', port=443): Read timed out. (read timeout=5)
Skipping https://oppobaca.news/2024/08/19/perayaan-hari-kemerdekaan-indonesia-ke-79-di-ibu-kota-nusantara-ikn/ due to error: HTTPSConnectionPool(host='oppobaca.news', port=443): Max retries exceeded with url: /2024/08/19/perayaan-hari-kemerdekaan-indonesia-ke-79-di-ibu-kota-nusantara-ikn/ (Caused by SSLError(SSLCertVerificationError(1, '[SSL: CERTIFICATE_VERIFY_FAILED] certificate verify failed: certificate has expired (_ssl.c:1016)')))

Relevant news links:
https://www.licas.news/2024/10/17/jokowi-inaugurates-key-ikn-projects-showcasing-final-legacy-before-office-transition/
https://harian.news/sekilas-pembangunan-ibu-kota-nusantara-ikn
https://cakra.news/kapan-ikn-gantikan-dki-jakarta-jadi-ibu-kota-negara-ini-penjelasan-istana/
https://line1.news/pengamat-masa-depan-ikn-gelap-gulat-usai-jokowi-lengser/
https://medianta.news/13/08/2024/rasa-kagum-ombas-rasakan-suasana-ibu-kota-nusantara/
https://rakyat.news/read/tag/ikn
https://lingkar.news/tag/ikn/
https://www.hyundai.news/eu/articles/press-releases/hmg-signs-mou-with-nusantara-capital-city-authority-to-establish-ecosystem-for-advanced-air-mobility.html

```

**Gambar 3.3** Hasil Daftar List Tautan Berita IKN

Daftar di atas menunjukkan bahwa pendekatan teknis yang diterapkan mampu menghasilkan data yang spesifik, aktual, dan relevan terhadap kebutuhan analisis. Tautan-tautan ini dapat digunakan sebagai dasar untuk ekstraksi isi berita, analisis opini publik, atau pembuatan dataset pelatihan untuk model NLP.

### 3.4 Pelabelan Manual dan Persiapan Dataset untuk Fine-Tuning BERT

Sebagai bagian dari eksperimen lanjutan dalam meningkatkan akurasi analisis sentimen, dilakukan proses pelabelan manual terhadap artikel berita yang telah dikumpulkan. Proses pelabelan ini melibatkan identifikasi sentimen dari setiap artikel berita, apakah bersifat positif, negatif, atau netral berdasarkan isi dari konteks pemberitaan. Label ini kemudian digunakan sebagai data berlabel (*ground truth*) dalam tahap fine-tuning model BERT.

Dengan adanya data berlabel ini, model BERT dapat dilatih secara terarah agar lebih akurat dalam memahami konteks kalimat dalam bahasa Indonesia, khususnya pada isu-isu kompleks seperti proyek pembangunan Ibu Kota Nusantara (IKN). Langkah ini merupakan bagian dari pembaruan metodologi yang bertujuan untuk meningkatkan performa model analisis sentimen dibandingkan pendekatan sebelumnya. Fine-tuning BERT memungkinkan model memahami struktur bahasa yang lebih dalam, termasuk menangkap kalimat panjang, ambigu, atau sarkastik yang sering kali sulit diproses oleh model tradisional.

### 3.5 Implementasi Model BERT untuk Analisis Sentimen

Dalam penelitian ini, implementasi model **BERT** dilakukan untuk mengklasifikasikan sentimen artikel berita menjadi tiga kelas: positif, negatif, dan netral. Model yang digunakan adalah **IndoBERT-base-p1** yang telah dilatih pada korpus Bahasa Indonesia, dan di-*fine-tune* ulang menggunakan dataset hasil scraping dan preprocessing terkait pemberitaan Ibu Kota Nusantara (IKN). Proses implementasi mencakup lima tahap utama yang dijelaskan secara sistematis sebagai berikut.

### 3.5.1 Tokenisasi dengan BertTokenizer

Langkah awal adalah tokenisasi teks menggunakan BertTokenizer dari pustaka Hugging Face Transformers. Tokenizer ini memetakan kata atau frasa dalam kalimat ke dalam token yang sesuai dengan kosakata BERT. Proses ini juga menambahkan token khusus seperti [CLS] di awal dan [SEP] di akhir kalimat, serta menghasilkan:

1. ID token (numerik),
2. *attention mask* untuk mengenali padding,
3. panjang token tetap (maks. 512 token).

Tokenisasi ini memastikan bahwa input teks sesuai dengan format yang dipahami oleh model BERT.

```
# 2. Inisialisasi model dan tokenizer
model_name = "indobenchmark/indobert-base-p1"
tokenizer = AutoTokenizer.from_pretrained(model_name)
```

```

111 class IndoBERTDataset(Dataset):
112     def __init__(self, texts, labels, tokenizer, max_length=256):
113         self.texts = texts
114         self.labels = labels
115         self.tokenizer = tokenizer
116         self.max_length = max_length
117
118     def __len__(self):
119         return len(self.texts)
120
121     def __getitem__(self, idx):
122         text = str(self.texts[idx])
123         label = self.labels[idx]
124         encoding = self.tokenizer(
125             text,
126             truncation=True,
127             padding='max_length',
128             max_length=self.max_length,
129             return_tensors='pt'
130         )
131         return {
132             'input_ids': encoding['input_ids'].squeeze(),
133             'attention_mask': encoding['attention_mask'].squeeze(),
134             'labels': torch.tensor(label, dtype=torch.long)
135         }

```

### 3.5.2 Pembagian Dataset (Train, Validation, Test)

Setelah data siap dan telah dilakukan augmentasi untuk menyeimbangkan kelas, dataset dibagi menjadi tiga bagian:

1. Training Set (80%): untuk melatih model
2. Test Set (20%): untuk mengukur performa akhir model terhadap data yang belum pernah dilihat sebelumnya

Pembagian dilakukan secara **stratified** agar proporsi tiap label sentimen tetap seimbang di setiap subset.

```

144 texts = df_balanced['normalized_content'].tolist()
145 labels = df_balanced['label'].tolist()
146 X_temp, X_test, y_temp, y_test = train_test_split(texts, labels, test_size=0.2, stratify=labels, random_state=42)
147 X_train, X_val, y_train, y_val = train_test_split(X_temp, y_temp, test_size=0.125, stratify=y_temp, random_state=42)

```

### 3.5.3 Fine-tuning Menggunakan transformers (Hugging Face)

Fine-tuning dilakukan pada model pre-trained IndoBERT dengan mengadaptasi arsitektur BertForSequenceClassification, yaitu versi BERT yang ditambahkan layer klasifikasi untuk

tugas klasifikasi teks multi-kelas. Proses ini menggunakan API Trainer dari Hugging Face, dan melibatkan konfigurasi hyperparameter seperti:

1. jumlah epoch,
2. batch size,
3. learning rate,
4. strategi evaluasi dan penyimpanan model.

Model dilatih menggunakan backend **PyTorch**, dengan loss function CrossEntropyLoss yang sesuai untuk klasifikasi multi-kelas.

```
140 model = AutoModelForSequenceClassification.from_pretrained(model_name, num_labels=3)
141 model.config.dropout = 0.4 # Tambah dropout

154 train_loader = DataLoader(train_dataset, batch_size=4, shuffle=True)
155 val_loader = DataLoader(val_dataset, batch_size=4)

158 # 4. Class weighting
159 class_weights = compute_class_weight('balanced', classes=np.unique(labels), y=labels)
160 class_weights = torch.tensor(class_weights, dtype=torch.float)
161
162 # 5. Early Stopping
163 class EarlyStopping:
164     def __init__(self, patience=10, min_delta=0.01):
165         self.patience = patience
166         self.min_delta = min_delta
167         self.best_f1 = 0
168         self.counter = 0
169         self.best_model_state = None
170
171     def __call__(self, val_f1, model):
172         if val_f1 > self.best_f1 + self.min_delta:
173             self.best_f1 = val_f1
174             self.counter = 0
175             self.best_model_state = model.state_dict()
176         else:
177             self.counter += 1
178         return self.counter >= self.patience
```

```

197 for epoch in range(epochs):
198     model.train()
199     total_loss = 0
200     for batch in tqdm(train_loader, desc=f"Epoch {epoch+1}"):
201         input_ids = batch['input_ids'].to(device)
202         attention_mask = batch['attention_mask'].to(device)
203         labels = batch['labels'].to(device)
204
205         outputs = model(input_ids=input_ids, attention_mask=attention_mask, labels=labels)
206         loss = outputs.loss
207         total_loss += loss.item()
208
209         optimizer.zero_grad()
210         loss.backward()
211         optimizer.step()
212         scheduler.step()

```

### 3.5.4 Evaluasi Model: Akurasi, Precision, Recall, F1-Score

Evaluasi performa model dilakukan menggunakan data uji dengan empat metrik utama:

1. Akurasi: persentase prediksi yang benar dari keseluruhan data uji
2. Precision: proporsi prediksi positif yang benar dari semua prediksi positif
3. Recall: proporsi kasus positif yang berhasil dikenali oleh model
4. F1-score: harmonisasi antara precision dan recall, khususnya berguna jika data tidak seimbang

```

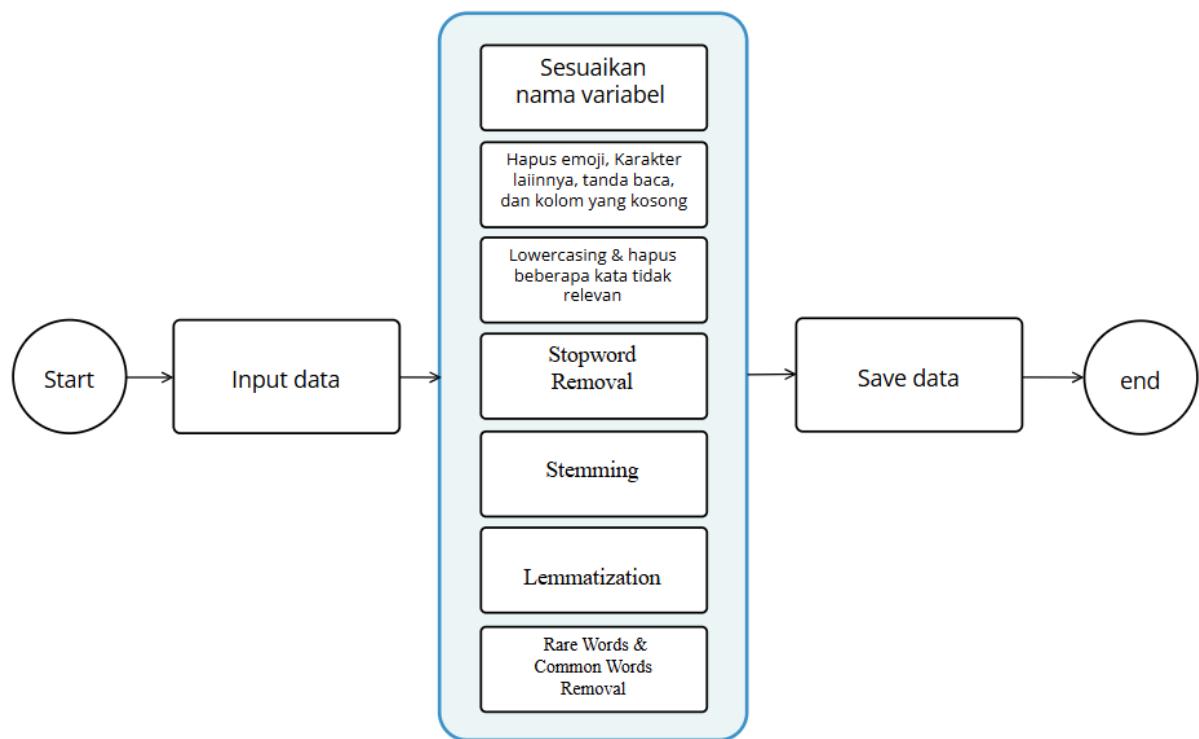
218     model.eval()
219     val_loss = 0
220     y_true_val = []
221     y_pred_val = []
222     with torch.no_grad():
223         for batch in tqdm(val_loader, desc="Validasi"):
224             input_ids = batch['input_ids'].to(device)
225             attention_mask = batch['attention_mask'].to(device)
226             labels = batch['labels'].to(device)
227             outputs = model(input_ids=input_ids, attention_mask=attention_mask, labels=labels)
228             val_loss += outputs.loss.item()
229             predictions = torch.argmax(outputs.logits, dim=1)
230             y_true_val.extend(labels.cpu().numpy())
231             y_pred_val.extend(predictions.cpu().numpy())
232
233
234 # Tampilkan classification report
235 print(classification_report(y_true, y_pred, target_names=["Negatif", "Netral", "Positif"], zero_division=0))

```

## BAB IV. DATA PREPROCESSING

### 4.1 Data Preprocessing Artikel Berita

Data yang diperoleh dari proses scraping kemudian melalui tahap preprocessing dengan fokus pada kolom konten, karena kolom ini akan menjadi objek utama dalam analisis pada tahap berikutnya. Tahap preprocessing mencakup sembilan langkah utama, yang dapat dilihat pada diagram alir berikut.



Gambar 4.1 Diagram Alir *Preprocessing* Artikel Berita

#### 4.1.1 Menyesuaikan Nama Variabel

Dalam proses pembersihan dan standarisasi data berita, dilakukan penyalinan isi asli dari kolom content ke kolom baru bernama cleaned\_content. Selanjutnya, dilakukan standarisasi kata kunci “ikn” agar menjadi huruf kapital “IKN” secara konsisten di kedua kolom cleaned\_content dan judul menggunakan fungsi replace dengan dukungan ekspresi reguler (regex). Untuk analisis frekuensi, digunakan fungsi str.contains untuk menghitung jumlah kemunculan kata “IKN” yang telah distandardkan. Berdasarkan hasil analisis, ditemukan bahwa kata “IKN” muncul sebanyak 174 kali pada kolom content dan 161 kali pada kolom judul.

#### **4.1.2 Menghapus Emoji, Karakter Lainnya, Tanda Baca, dan Kolom Content yang Kosong**

Dalam tahap pembersihan lanjutan pada data teks, dilakukan proses penghapusan emoji, simbol, serta karakter non-ASCII dari kolom cleaned\_content menggunakan fungsi custom bernama remove\_emojis. Selanjutnya, dilakukan normalisasi teks menggunakan fungsi cleaning\_text yang menghapus karakter tab, newline, backslash, URL, angka, tanda baca, dan karakter tunggal, serta merapikan spasi yang berlebihan. Seluruh proses ini bertujuan untuk menghasilkan teks yang bersih dan siap digunakan dalam tahap analisis selanjutnya. Setelah pembersihan, dilakukan penyaringan untuk menghapus baris-baris dengan konten kosong agar hanya data yang relevan dan bermakna yang dipertahankan.

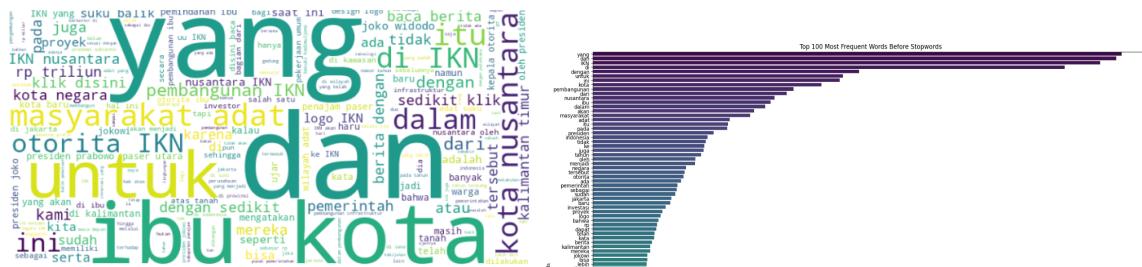
#### **4.1.3 Lower Casing dan Menghapus beberapa kata tidak relevan**

Pada ini dilakukan konversi tipe data pada kolom 'cleaned\_content' dan 'judul' menjadi string untuk memastikan konsistensi tipe data. Kemudian, fungsi lowercase\_except\_special diterapkan pada kedua kolom tersebut untuk mengubah seluruh teks menjadi huruf kecil, dengan pengecualian untuk kata "IKN" yang tetap dipertahankan dalam format huruf kapital. Selanjutnya, dilakukan penghapusan sejumlah teks yang tidak relevan atau berupa footer dan metadata menggunakan metode str.replace. Teks-teks yang ingin dihapus sudah disiapkan dalam daftar, dan proses ini memastikan bahwa konten yang tersisa hanya berisi informasi yang relevan dan bersih dari elemen-elemen yang mengganggu..

cleaned_content	
0	pro kontra pemindahan ibu kota negara tim reda...
1	catatan artikel ini merupakan opini pribadi pe...
2	saya adalah mahasiswa fakultas hukum universit...
3	jakarta cnbc indonesia ibu kota negara IKN ind...
4	scroll ke bawah untuk membaca berita baca beri...

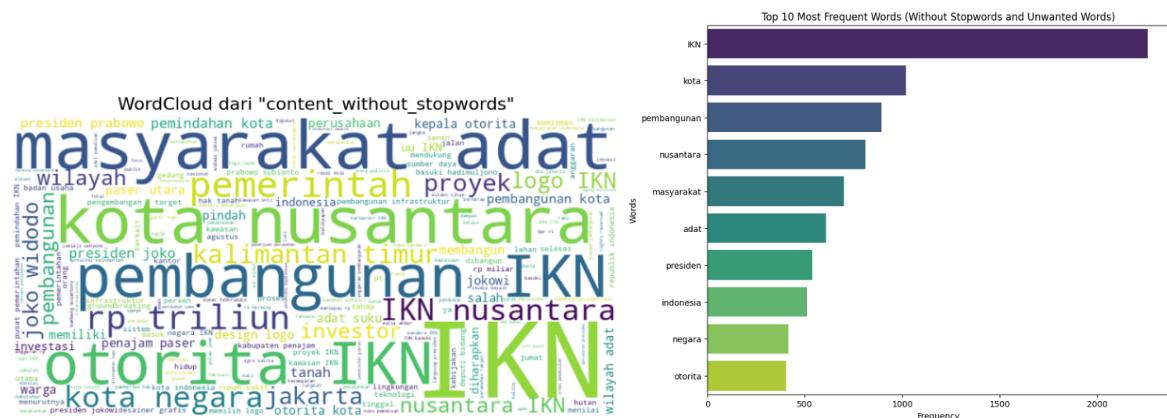
#### **4.1.4 Text Analysis (Sebelum Stopwords)**

Pada tahap ini, dilakukan tokenisasi pada kolom 'cleaned\_content' menggunakan fungsi word\_tokenize dari NLTK untuk menghasilkan kolom baru berisi token kata. Kemudian, dihitung jumlah kata per dokumen menggunakan word\_tokenize dan disimpan dalam kolom 'wordCount'. Selanjutnya, frekuensi kata dihitung dengan menggunakan Counter untuk mendapatkan 100 kata yang paling sering muncul, yang kemudian disajikan dalam bentuk DataFrame dan diurutkan berdasarkan frekuensi. Hasil dari top 100 kata yang paling sering muncul ini divisualisasikan dalam bentuk diagram batang. Selain itu, dibuat pula WordCloud dari seluruh kata dalam kolom 'cleaned\_content'. Hasil grafik dan WordCloud adalah sebagai berikut



#### 4.1.5 Text Analysis (Setelah Stopwords)

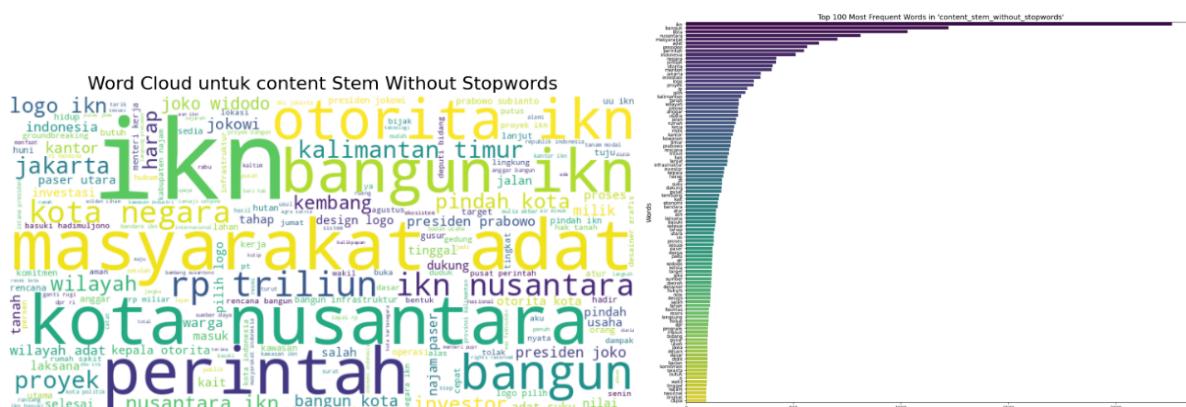
Pada tahap pembersihan teks ini, dilakukan penghapusan stopwords dari kolom 'cleaned\_content' menggunakan daftar stopwords yang diunduh dari URL dan ditambahkan dengan stopwords kustom. Fungsi `remove_stopwords` diterapkan untuk menghapus kata-kata yang termasuk dalam daftar tersebut. Selanjutnya, untuk memperbaiki teks, beberapa kata yang tidak diinginkan juga dihapus menggunakan fungsi `remove_unwanted_words`. Setelah teks dibersihkan, frekuensi kata dihitung untuk mendapatkan 100 kata yang paling sering muncul dan divisualisasikan dalam bentuk diagram batang. Selain itu, WordCloud juga dibuat dari teks yang telah dibersihkan untuk memberikan gambaran visual mengenai distribusi kata. Hasil grafik dan WordCloud ini menunjukkan top 10 kata yang paling sering muncul setelah penghapusan stopwords dan kata yang tidak diinginkan.



#### 4.1.6 Stemming

Pada tahap ini, dilakukan proses stemming pada teks yang telah dibersihkan dari stopwords untuk mengurangi kata-kata ke bentuk dasarnya menggunakan pustaka *PySastrawi*. Fungsi `stemming_indonesia` diterapkan pada kolom 'content\_without\_stopwords' untuk menghasilkan teks yang sudah dalam bentuk dasar. Kemudian, dihitung jumlah frekuensi kata dalam teks yang telah di-stem dan divisualisasikan dalam grafik batang untuk menunjukkan top kata yang paling sering muncul. Selain itu, WordCloud juga dibuat untuk menggambarkan distribusi kata secara visual. Hasil yang diperoleh menunjukkan top kata setelah proses stemming dan stopwords dihapus, serta WordCloud yang merepresentasikan distribusi kata-kata tersebut.

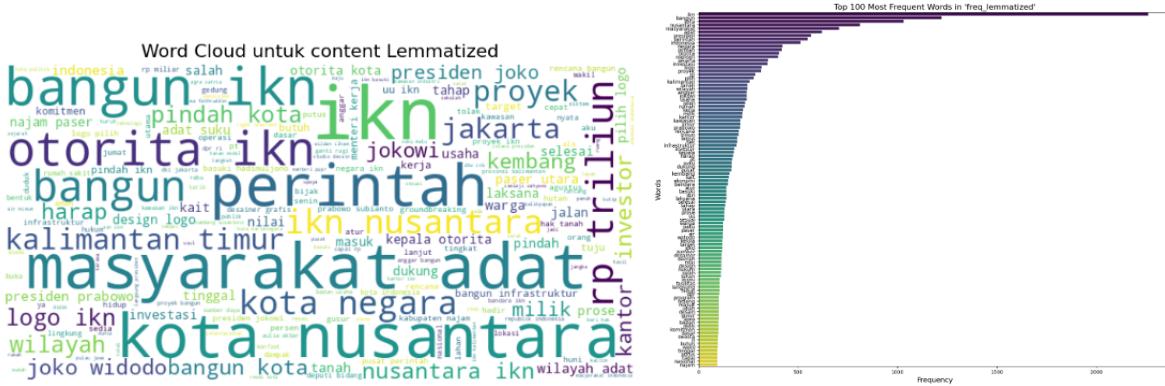
	content_without_stopwords	content_stem_without_stopwords
0	pro kontra pemindahan kota negara tim redaksi ...	pro kontra pindah kota negara tim redaksi kota...
1	catatan artikel opini pribadi penulis mencermi...	catat artikel opini pribadi tulis cermin panda...
2	mahasiswa fakultas hukum universitas jember mi...	mahasiswa fakultas hukum universitas jember mi...
3	jakarta cnbc indonesia kota negara IKN indones...	jakarta cnbc indonesia kota negara ikn indones...



### 4.1.7 Lemmatization

Pada tahap ini, dilakukan proses lemmatization pada teks yang telah melalui tahap stemming, menggunakan pustaka *nltk* dan fungsi WordNetLemmatizer. Fungsi lemmatize\_indonesia diterapkan untuk mengubah kata-kata dalam teks menjadi bentuk dasarnya. Setelah lemmatization, dihitung frekuensi kata dalam teks dan hasilnya divisualisasikan dalam grafik batang untuk menunjukkan top kata yang paling sering muncul. Selain itu, sebuah WordCloud juga dibuat untuk menampilkan distribusi kata-kata yang paling dominan setelah lemmatization. Hasil yang diperoleh menunjukkan top kata serta WordCloud yang memberikan gambaran visual dari kata-kata yang sering muncul dalam teks yang telah melalui proses stemming dan lemmatization.

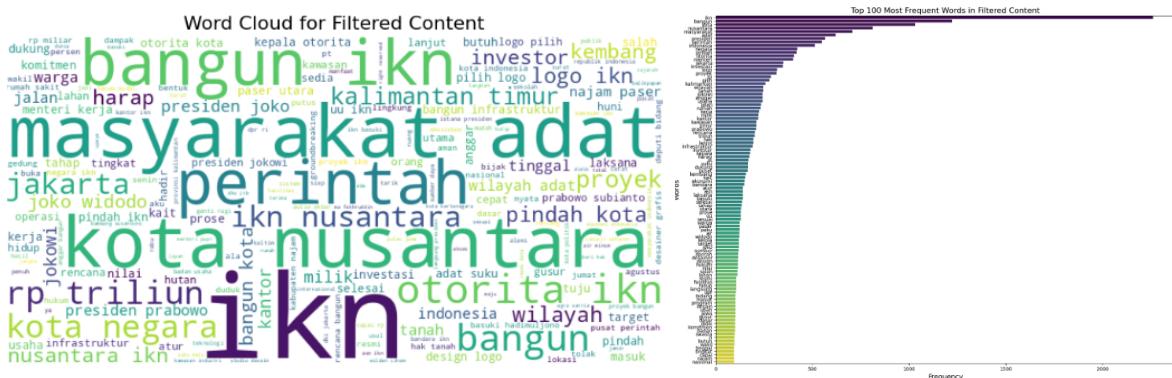
	content_stem_without_stopwords	content_lemmatized
0	pro kontra pindah kota negara tim redaksi kota...	pro kontra pindah kota negara tim redaksi kota...
1	catat artikel opini pribadi tulis cermin panda...	catat artikel opini pribadi tulis cermin panda...
2	mahasiswa fakultas hukum universitas jember mi...	mahasiswa fakultas hukum universitas jember mi...
3	jakarta cnbc indonesia kota negara ikn indones...	jakarta cnbc indonesia kota negara ikn indones...



#### 4.1.8 Rare Words & Common Words Removal

Pada tahap ini, dilakukan analisis untuk menghapus kata-kata yang terlalu langka atau terlalu umum dalam teks yang telah melalui proses lemmatization. Fungsi `get_word_frequencies` digunakan untuk menghitung frekuensi kemunculan setiap kata, sedangkan fungsi `remove_rare_common_words` digunakan untuk memfilter kata-kata yang tidak memenuhi kriteria frekuensi tertentu, yakni kata yang muncul lebih dari dua kali dan kurang dari tiga ribu kali. Setelah memfilter kata-kata tersebut, hasilnya divisualisasikan melalui grafik batang yang menampilkan kata-kata yang paling sering muncul dalam teks yang sudah disaring, serta sebuah WordCloud yang menggambarkan distribusi kata-kata setelah proses pemfilteran. Hasilnya menunjukkan pengurangan kata yang sangat jarang dan berlebihan.

	content_lemmatized	filtered_content
0	pro kontra pindah kota negara tim redaksi kota...	pro kontra pindah kota negara tim redaksi kota...
1	catat artikel opini pribadi tulis cermin panda...	catat artikel opini pribadi tulis cermin panda...
2	mahasiswa fakultas hukum universitas jember mi...	mahasiswa fakultas hukum universitas minat huk...
3	jakarta cnbc indonesia kota negara ikn indones...	jakarta cnbc indonesia kota negara ikn indones...



#### 4.2 Augmentasi dan Penyeimbangan Dataset Sentimen

Distribusi awal data sentimen hasil klasifikasi menggunakan model IndoBERT menunjukkan adanya ketimpangan signifikan antar kelas, di mana label netral mendominasi jumlah artikel secara keseluruhan. Dari total 162 artikel yang dikumpulkan, sebagian besar diklasifikasikan ke dalam kategori netral, sementara hanya sebagian kecil yang terdeteksi memiliki sentimen positif maupun negatif. Ketimpangan ini dapat menimbulkan permasalahan serius dalam konteks supervised learning, karena model cenderung belajar dari pola mayoritas dan mengabaikan kelas minoritas. Akibatnya, akurasi prediksi untuk artikel bernuansa positif atau negatif menjadi rendah, dan model gagal melakukan generalisasi secara menyeluruh terhadap dinamika opini yang sebenarnya terjadi di masyarakat.

Untuk mengatasi permasalahan ini, diterapkan strategi augmentasi dan penyeimbangan dataset (data balancing) guna memastikan bahwa setiap kelas memiliki jumlah representasi yang setara saat proses pelatihan model. Proses ini dilakukan dengan dua pendekatan utama: *undersampling* terhadap kelas netral, dan *data augmentation* terhadap kelas positif dan negatif. Undersampling dilakukan dengan memilih secara acak 100 artikel netral dari total 162, agar jumlahnya setara dengan target. Sementara itu, kelas positif yang hanya memiliki 82 artikel dan kelas negatif dengan 56 artikel ditingkatkan hingga masing-masing mencapai 100 artikel melalui berbagai teknik augmentasi berbasis NLP.

#### **4.2.1 Strategi Penyeimbangan Kelas**

Dataset asli terdiri dari :

1. Netral: 162 artikel
2. Positif: 82 artikel
3. Negatif: 56 artikel

Agar distribusi sentimen seimbang, maka dilakukan langkah Random undersampling pada kelas netral untuk dipotong menjadi 100 sampel secara acak. Data augmentation pada kelas positif dan negatif untuk meningkatkan masing-masing hingga 100 sampel.

#### **4.2.2 Teknik Augmentasi Data**

Augmentasi dilakukan menggunakan teknik berbasis NLP yang menjaga makna semantik namun menghasilkan variasi sintaksis, antara lain:

1. Synonym Replacement: mengganti kata-kata umum dengan sinonimnya menggunakan WordNet Bahasa Indonesia.
2. Back Translation: menerjemahkan teks ke bahasa lain (misalnya Inggris) dan kembali ke Bahasa Indonesia untuk memperoleh variasi struktur kalimat.
3. Paraphrasing Manual: untuk sebagian kecil artikel, dilakukan parafrase manual dengan mempertahankan konteks sentimen aslinya.
4. Random Swap and Deletion: pertukaran atau penghapusan kata non-esensial untuk menghasilkan variasi kalimat tanpa mengubah arti.

Proses augmentasi ini dilakukan secara hati-hati untuk memastikan bahwa label sentimen tetap valid, dan hasil augmentasi tidak menyebabkan noise yang justru menurunkan performa model.

#### **4.2.3 Hasil Penyeimbangan Dataset**

Setelah proses augmentasi dan penyesuaian, diperoleh distribusi dataset final sebagai berikut:

1. Netral: 100 artikel (hasil undersampling)
2. Positif: 100 artikel (82 asli + 18 augmentasi)
3. Negatif: 100 artikel (56 asli + 44 augmentasi)

Distribusi seimbang ini menjadi dasar untuk proses fine-tuning model IndoBERT pada tahap analisis sentimen. Dengan proporsi data yang seragam, model dapat belajar secara adil dari ketiga jenis sentimen, sehingga meningkatkan keakuratan klasifikasi dan mengurangi bias terhadap kelas mayoritas.

## BAB V. DEFINISI DATASET

### 5.1 Dataset Link Artikel Berita

Bagian ini menyajikan struktur dan karakteristik dataset yang telah dikumpulkan dan diproses melalui tahap pengambilan data (data acquisition) serta pembersihan awal (data preprocessing). Dataset ini berisi 162 link artikel berita yang dikumpulkan secara manual dengan topik pencarian seputar Ibu Kota Nusantara (IKN). Proses pengumpulan dilakukan dengan menelusuri berbagai sumber berita online dan mencatat informasi yang relevan. Dataset ini mencakup beberapa kolom utama, yaitu **judul berita**, **penerbit**, **link berita**, **label**, **kategori**, dan **tanggal terbit**. Berikut penjelasan lebih lanjut terkait struktur dataset yang digunakan..

No.	Kolom	Definisi
1.	<i>Judul Berita</i>	Merupakan judul utama dari artikel berita yang menjelaskan inti informasi dari konten berita tersebut
2.	<i>Penerbit</i>	Nama media atau platform berita yang menerbitkan artikel tersebut
3.	<i>Link Berita</i>	URL yang mengarahkan langsung ke halaman artikel berita di situs penerbit.
4.	<i>Label Kategori</i>	Tema atau jenis konten dari berita, seperti kategori politik, ekonomi, pembangunan, lingkungan, dan sebagainya.
5.	<i>Tanggal Terbit</i>	Waktu atau tanggal kapan artikel berita tersebut dipublikasikan oleh penerbit.

### 5.2 Dataset Hasil Preprocessing

Dataset ini berisi konten artikel berita yang telah melalui proses pembersihan data pada tahap preprocessing. Tabel berikut menyajikan penjelasan dari masing-masing kolom yang terdapat dalam dataset.

No.	Kolom	Definisi
1.	<i>link</i>	URL yang mengarahkan langsung ke halaman artikel berita di situs penerbit.

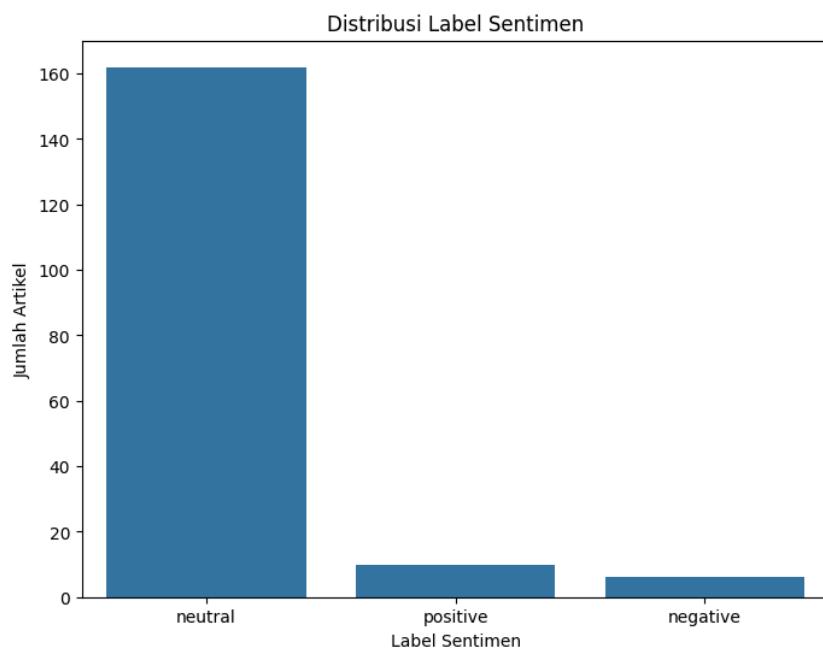
2.	<i>judul</i>	Merupakan judul utama dari artikel berita yang menjelaskan inti informasi dari konten berita tersebut
3.	<i>penerbit</i>	Nama media atau platform berita yang menerbitkan artikel tersebut
4.	<i>tanggal</i>	Waktu atau tanggal kapan artikel berita tersebut dipublikasikan oleh penerbit.
5.	<i>content</i>	Konten lengkap dari artikel berita yang telah diambil
6.	<i>cleaned_content</i>	Konten artikel berita yang telah dibersihkan dari elemen-elemen yang tidak relevan
7.	<i>tags</i>	Kata kunci atau tag yang relevan dengan isi artikel
8.	<i>status</i>	Status pengambilan artikel, yang menunjukkan apakah proses pengambilan berhasil atau tidak.
9.	<i>tahun</i>	Tahun publikasi artikel, diambil dari kolom tanggal terbit
10.	<i>WordCount</i>	Jumlah total kata dalam konten artikel
11.	<i>content_without_stopwords</i>	Konten artikel setelah menghapus kata-kata yang tidak bermakna (stopwords)
12.	<i>wordcount_after_stopwords</i>	Jumlah total kata dalam konten setelah penghapusan stopwords
13.	<i>content_stem_without_stopwords</i>	Konten artikel setelah dilakukan stemming dan penghapusan stopwords
14.	<i>wordcount_stem_without_stopwords</i>	Jumlah total kata dalam konten setelah dilakukan stemming dan penghapusan stopwords
15.	<i>content_lemmatized</i>	Konten artikel setelah dilakukan lemmatization untuk mengubah kata-kata ke bentuk dasarnya
16.	<i>wordcount_lemmatized</i>	Jumlah total kata dalam konten setelah dilakukan lemmatization

## BAB VI. ANALISIS DATA

### 6.1 Analisis Sentimen

Analisis sentimen merupakan teknik dalam bidang *Natural Language Processing* (NLP) yang digunakan untuk mengidentifikasi, mengekstrak, dan mengelompokkan opini atau emosi dalam suatu teks, dengan tujuan untuk menentukan sikap atau sentimen yang terkandung, apakah bersifat positif, negatif, atau netral. Dalam konteks ini, analisis sentimen dilakukan terhadap artikel-artikel berita untuk mengetahui kecenderungan opini terhadap topik tertentu, yaitu Ibu Kota Nusantara (IKN). Model yang digunakan adalah IndoBERT, yaitu model transformer berbasis bahasa Indonesia yang telah dilatih untuk tugas klasifikasi sentimen. Label yang dihasilkan terdiri dari tiga kategori, yaitu:

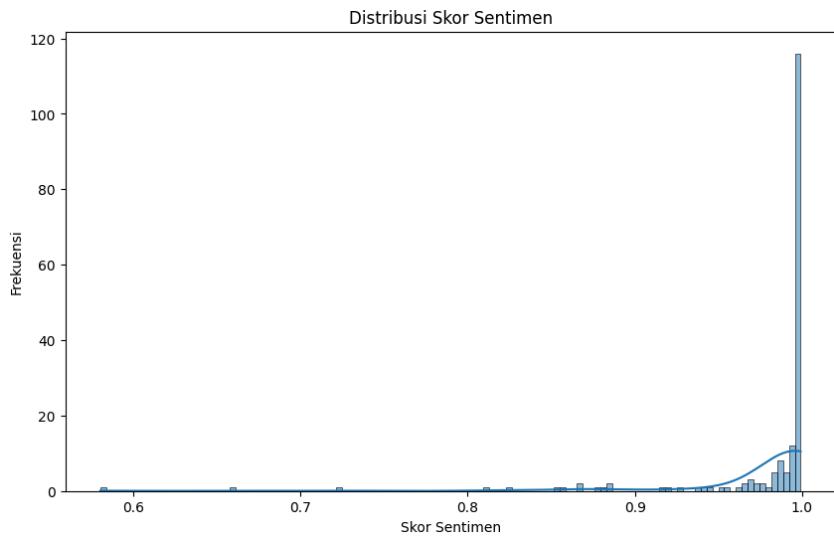
- Positive, apabila isi teks cenderung menyampaikan opini optimis, dukungan, atau pandangan positif terhadap IKN;
- Negative, apabila teks menunjukkan kritik, penolakan, atau sentimen pesimis; dan
- Neutral, apabila teks bersifat informatif atau tidak memiliki kecenderungan opini yang jelas.



Gambar 6.1 Hasil Distribusi Sentimen

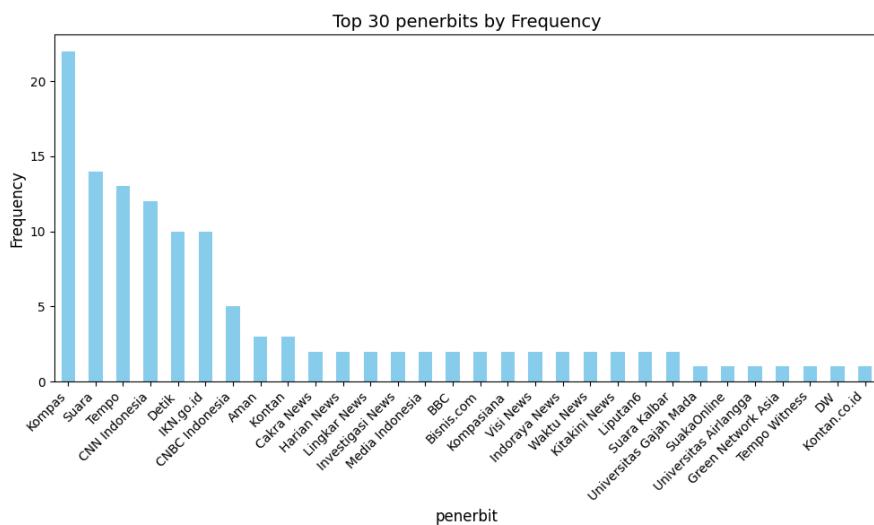
Berdasarkan visualisasi distribusi label sentimen, diketahui bahwa mayoritas artikel memiliki label *neutral*, yang menandakan bahwa sebagian besar pemberitaan bersifat informatif dan tidak mengandung opini yang jelas bernada positif maupun negatif. Hanya sebagian kecil artikel yang terkласifikasi sebagai *positive* atau *negative*, yang dapat mengindikasikan bahwa isu IKN lebih sering diberitakan secara objektif atau bahwa model

cenderung memilih label netral ketika tidak cukup yakin. Hal ini didukung oleh distribusi skor sentimen yang menunjukkan dominasi skor tinggi, terutama mendekati angka 1.0



**Gambar 6.2** Hasil Distribusi Skor Sentimen

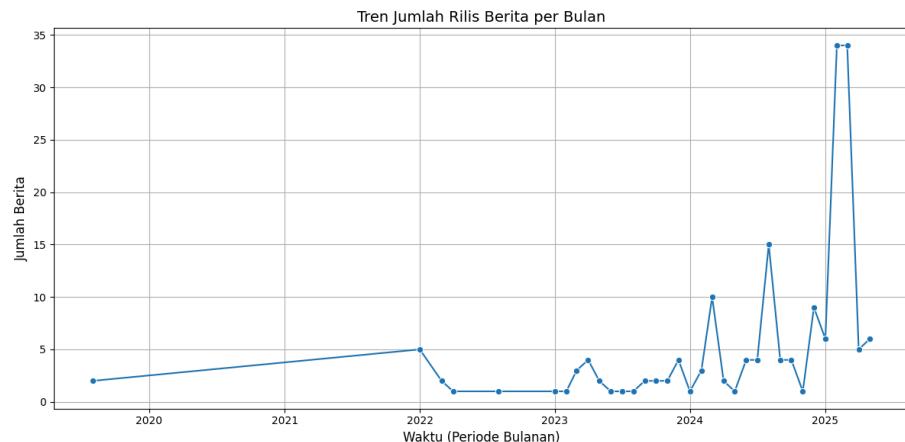
Kemudian analisis berdasarkan penerbit dilakukan untuk mengetahui distribusi berita berdasarkan *penerbit* dalam kumpulan data yang telah dibersihkan dari *stopwords*. Dengan menggunakan fungsi `value_counts()`, dihitung frekuensi kemunculan masing-masing penerbit dalam data. Hasil analisis menunjukkan bahwa **Kompas** menjadi penerbit dengan jumlah berita terbanyak, disusul oleh **Suara**, **Tempo**, dan **CNN Indonesia**.



**Gambar 6.3** Distribusi 30 penerbit teratas berdasarkan jumlah artikel

Selanjutnya analisis ini bertujuan untuk mengidentifikasi *tren jumlah rilis berita* mengenai topik tertentu dari waktu ke waktu. Data dianalisis berdasarkan agregasi bulanan dan tahunan dengan menggunakan kolom tanggal yang telah dikonversi ke format *datetime*.

Dari hasil visualisasi *line chart*, terlihat bahwa jumlah berita yang dirilis berfluktuasi setiap bulannya. Untuk mendalami tren tersebut, dilakukan analisis terhadap bulan dengan jumlah berita terbanyak di tiap tahun. Hasilnya menunjukkan bahwa pada tahun 2019, bulan Agustus menjadi puncak pemberitaan; pada 2022 terjadi lonjakan di bulan Januari; pada 2023 jumlah berita terbanyak muncul di bulan April; sedangkan tahun 2024 dan 2025 mengalami puncaknya masing-masing di bulan Agustus dan Februari. Temuan ini menunjukkan adanya lonjakan perhatian media pada bulan-bulan tertentu, yang bisa jadi berkaitan dengan peristiwa penting atau momentum strategis yang terjadi dalam periode tersebut.



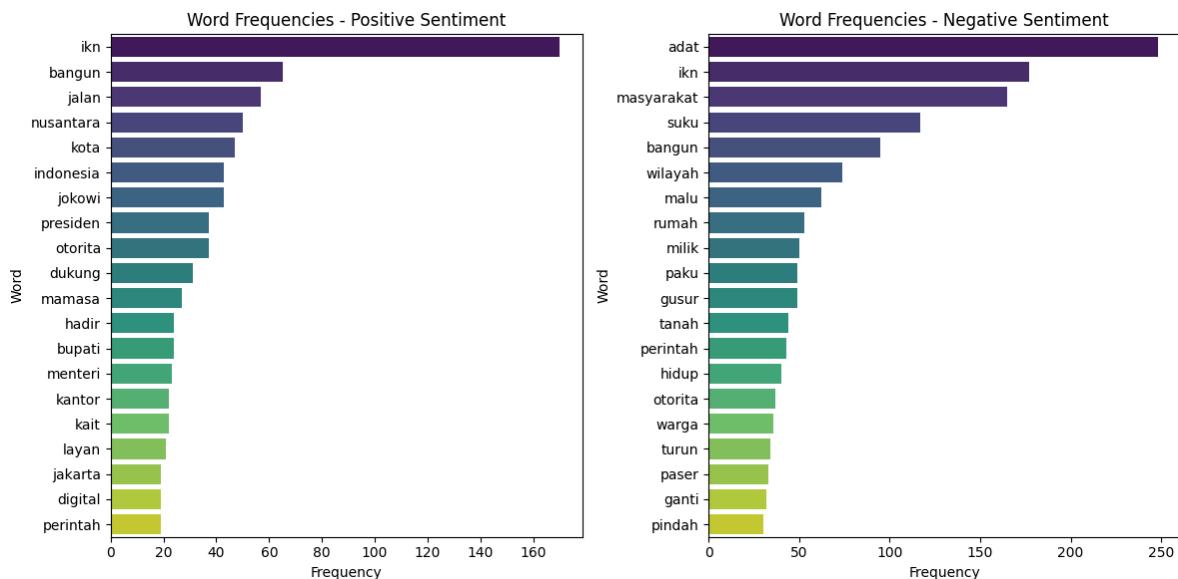
**Gambar 6.4** Tren jumlah rilis berita bulanan dari tahun ke tahun

Visualisasi Word Cloud memperlihatkan kata-kunci teratas di masing-masing kelompok sentimen: dari semua teks berlabel **positive**, WordCloud hijau menonjolkan kata-kata yang paling sering muncul dalam berita bernada optimis atau mendukung, sedangkan dari teks **negative**, WordCloud merah menyoroti kata-kata yang sering muncul dalam pemberitaan kritis atau bernada pesimis.



**Gambar 6.4** WordCloud sentimen positif dan negatif

Visualisasi ini menampilkan 20 kata paling sering muncul pada berita dengan label sentimen **positif** dan **negatif**. Diagram batang kiri menunjukkan frekuensi kata dalam berita bernada positif, sedangkan diagram kanan untuk berita bernada negatif.



**Gambar 6.5** Frekuensi 20 kata teratas pada berita positif dan negatif.

## 6.2 Analisis TF-IDF

### 6.2.1 Proses TF-IDF pada Artikel Berita

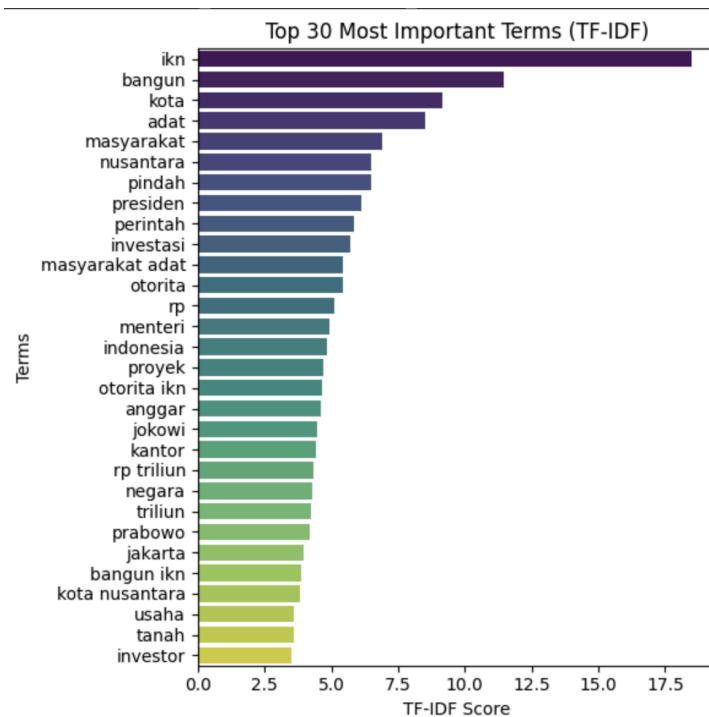
Melakukan ekstraksi fitur dengan metode TF-IDF (Term Frequency–Inverse Document Frequency). Metode ini digunakan untuk mengukur seberapa penting sebuah kata dalam suatu dokumen relatif terhadap kumpulan dokumen lainnya. TF-IDF menjadi dasar penting dalam representasi vektor dari dokumen teks untuk keperluan analisis lanjutan seperti clustering, klasifikasi, atau visualisasi.

### 6.2.2 Tokenisasi dan Stopword Removal

Pada tahap awal, teks dari kolom cleaned\_content dan judul diproses dengan tokenisasi, yaitu memecah teks menjadi kata-kata (token) individual. Proses ini dilakukan dengan menggunakan tokenizer berbasis regex yang mengidentifikasi kata dari setiap baris teks. Selanjutnya, dilakukan penghapusan stopwords, yaitu kata-kata umum dalam bahasa Indonesia yang tidak memiliki nilai penting dalam analisis seperti “dan”, “yang”, atau “dari”. Hal ini bertujuan untuk menjaga hanya kata-kata bermakna yang relevan terhadap isi dokumen.

### 6.2.3 Analisis Kata Terpenting Berdasarkan Bobot TF-IDF

Setelah mendapatkan vektor TF-IDF, langkah berikutnya adalah mengidentifikasi kata-kata yang memiliki nilai bobot tertinggi. Proses ini dilakukan dengan menjumlahkan bobot setiap kata di seluruh dokumen dan menguratkannya secara menurun. Kata-kata dengan bobot tertinggi diindikasikan sebagai kata kunci utama yang mendominasi topik pembahasan artikel berita terkait IKN.

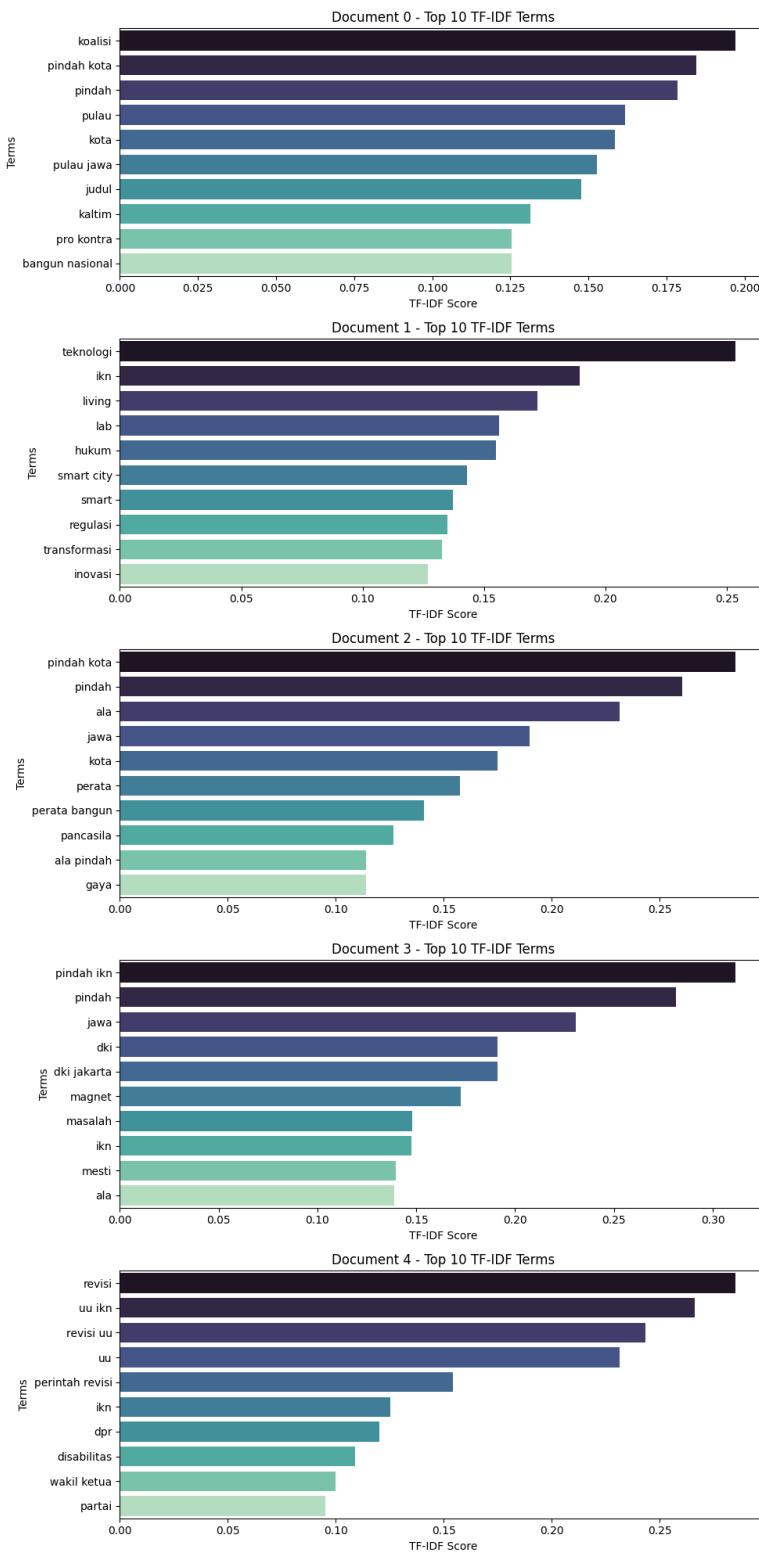


**Gambar 6.6** Frekuensi 30 kata terpenting yang muncul

Hasil penghitungan TF-IDF kemudian divisualisasikan dalam bentuk grafik batang horizontal untuk menampilkan 30 istilah terpenting dalam kumpulan artikel berita mengenai Ibu Kota Nusantara (IKN). Grafik ini bertujuan untuk mengidentifikasi kata-kata atau frasa yang paling informatif dan memiliki bobot tinggi dalam membedakan isi dokumen satu dengan lainnya. Sumbu horizontal grafik menunjukkan skor TF-IDF, sementara sumbu vertikal menampilkan daftar kata-kata atau frasa yang relevan.

Dari visualisasi tersebut, kata “IKN” menjadi istilah paling dominan dengan skor tertinggi, yaitu sekitar 18,5. Hal ini menandakan bahwa kata tersebut tidak hanya sering muncul, tetapi juga memiliki bobot penting dalam keseluruhan konteks pembahasan. Disusul oleh kata “bangun” dan “kota”, yang menunjukkan bahwa pembahasan media banyak berfokus pada aspek pembangunan fisik dari ibu kota baru tersebut. Selain itu, terdapat sejumlah istilah lain yang mengindikasikan berbagai tema sentral yang muncul dalam pemberitaan, seperti sosial budaya, pemerintahan, ekonomi, dan politik.

### 6.2.3 Visualisasi Topik Berdasarkan Dokumen



Selain menampilkan istilah penting secara keseluruhan, analisis TF-IDF juga diterapkan pada tingkat dokumen individual untuk mengidentifikasi fokus utama dari masing-masing artikel berita. Visualisasi pada Gambar 4.4 menampilkan 10 istilah dengan nilai TF-IDF tertinggi dari lima dokumen pertama dalam korpus data. Pendekatan ini membantu memahami nuansa dan fokus topik yang berbeda di tiap artikel, serta memberikan wawasan mendalam terhadap konteks yang diangkat oleh media.

## Dokumen 0

1. Didominasi oleh istilah seperti “pindah kota”, “pulau jawa”, dan “kaltim”.
2. Mengindikasikan bahwa isi dokumen membahas pemindahan lokasi ibu kota yang awalnya terletak dari pulau jawa menjadi pulau kalimatan.

## Dokumen 1

3. Didominasi oleh istilah seperti “teknologi”, “living”, “smart city”, dan “transformasi”.
4. Mengindikasikan bahwa isi dokumen membahas konsep pembangunan IKN sebagai kota pintar (smart city) dengan pendekatan teknologi tinggi dan transformasi digital.

## Dokumen 2

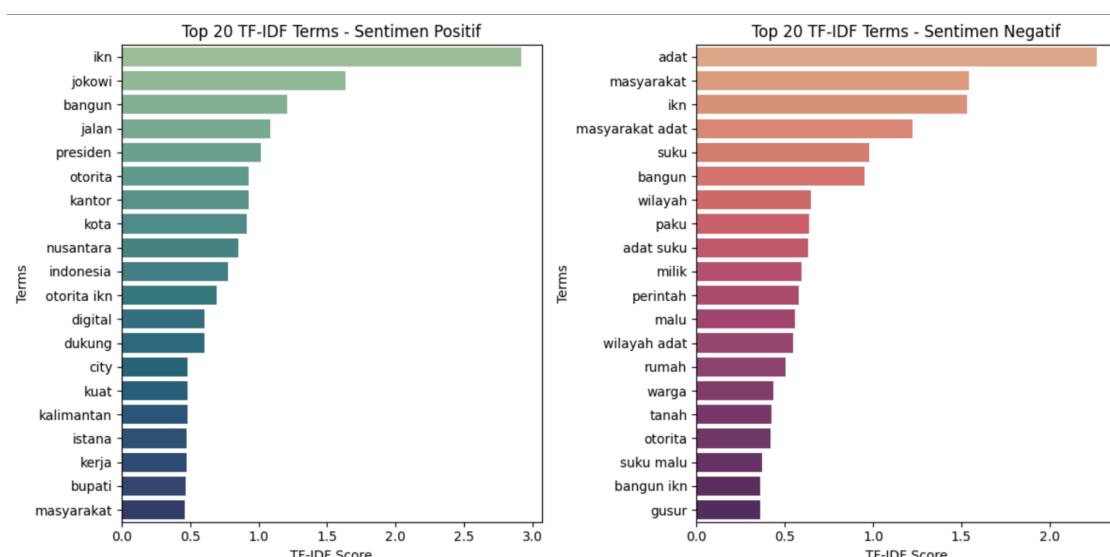
1. Kata-kata seperti “pindah kota”, “ala”, “pancasila”, dan “gaya” mendominasi.
2. Artikel ini tampaknya memiliki gaya naratif atau opini dengan membandingkan gaya pembangunan atau mengangkat isu nilai-nilai Pancasila dalam konteks pembangunan kota.

## Dokumen 3

1. Terdapat istilah “pindah IKN”, “dki jakarta”, “magnet”, dan “masalah”.
2. Menunjukkan adanya pembahasan terkait transisi dari Jakarta ke IKN, serta potensi masalah dan efek gravitasi atau “magnet” dari kota baru terhadap aspek sosial dan ekonomi.

## Dokumen 4

1. Didominasi oleh istilah “revisi”, “uu ikn”, “perintah revisi”, “dpr”, dan “disabilitas”.
2. Fokus utama dokumen ini adalah pada aspek legislasi dan kebijakan, serta pembahasan revisi Undang-Undang terkait IKN yang melibatkan DPR dan mungkin menyentuh aspek inklusi sosial.



Pada sentimen positif, hasil analisis TF-IDF menunjukkan bahwa kata-kata yang paling sering muncul adalah "ikn", "jokowi", "bangun", "presiden", dan "nusantara". Hal ini mencerminkan dukungan masyarakat terhadap pembangunan Ibu Kota Negara sebagai simbol kemajuan dan visi masa depan Indonesia. Kehadiran kata-kata seperti "digital", "kota", dan "kerja" juga menunjukkan adanya harapan terhadap IKN sebagai kota modern yang terintegrasi dengan teknologi serta mampu membuka peluang kerja baru. Secara umum, sentimen positif banyak berasal dari pandangan yang optimis terhadap dampak jangka panjang pembangunan ini.

Sementara itu, pada sentimen negatif, kata-kata dominan seperti "adat", "masyarakat adat", "suku", "wilayah", dan "tanah" mencerminkan kekhawatiran masyarakat terhadap dampak pembangunan IKN terhadap komunitas lokal. Istilah seperti "gusur", "milik", dan "perintah" mengindikasikan adanya rasa ketidakadilan atau paksaan yang dirasakan oleh sebagian masyarakat, terutama yang terkait dengan hak atas tanah dan pelestarian budaya adat. Sentimen negatif ini lebih banyak muncul dari perspektif sosial, khususnya perlindungan terhadap masyarakat adat yang terdampak langsung oleh proyek pembangunan tersebut.

Perbedaan distribusi istilah TF-IDF ini mencerminkan kontras dalam fokus narasi antara sentimen positif dan negatif. Sentimen positif lebih banyak menyoroti pembangunan dan optimisme terhadap masa depan IKN, sedangkan sentimen negatif didominasi oleh kekhawatiran terhadap dampak sosial, budaya, dan hak-hak masyarakat adat. Hal ini mempertegas pentingnya mempertimbangkan aspek sosial dan komunikasi publik dalam pembangunan proyek strategis nasional seperti IKN.

### 6.3 Analisis POS dan NER

Part-of-Speech (POS) merupakan teknik analisis linguistik yang berfungsi mengklasifikasikan jenis dan peran kata dalam suatu kalimat. Sementara itu, Named Entity Recognition (NER) digunakan untuk mengenali entitas penting seperti nama individu, organisasi, serta waktu. Kedua metode ini diterapkan dalam analisis artikel terkait sistem Coretax guna memahami isi, struktur, serta mengungkap pola bahasa dan informasi utama yang terkandung di dalamnya.

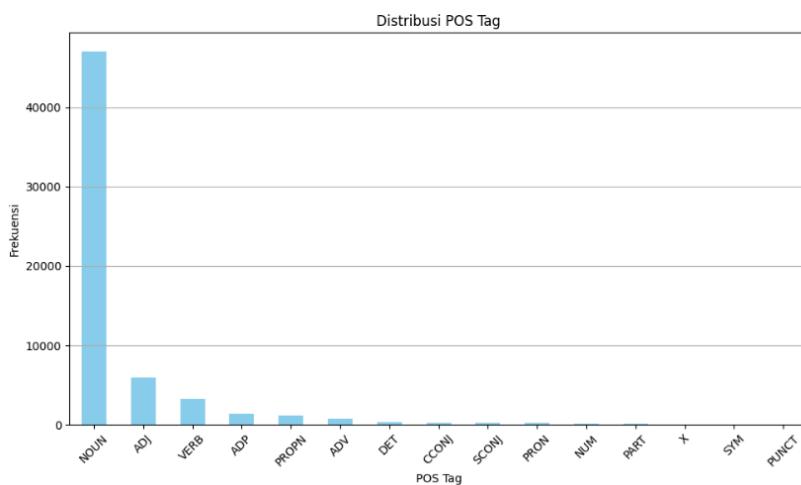
#### 6.3.1 POS

POS Tagging digunakan untuk mengklasifikasikan setiap kata dalam artikel ke dalam kategori tertentu seperti kata benda (noun), kata kerja (verb), kata sifat (adjective), dan lain-lain. Proses ini membantu melihat pola bahasa serta struktur kalimat yang mendominasi dalam artikel yang dianalisis. POS tagging dilakukan menggunakan model Bahasa Indonesia dari spaCy dan IndoBERT, yang dilatih khusus untuk memahami konteks linguistik Bahasa Indonesia.

**Gambar 6.3.1.1** Hasil *Part-of-Speech Tagging* pada teks artikel IKN

Gambar di atas menampilkan hasil *Part-of-Speech (POS) Tagging* terhadap artikel bertema pemindahan ibu kota negara Indonesia ke Nusantara. Setiap kata dalam artikel dianalisis dan diberi label berdasarkan kelas katanya, dengan rincian sebagai berikut:

- **NOUN (Kata Benda):** Jumlah paling dominan dalam artikel. Hal ini menunjukkan bahwa isi teks banyak memuat informasi berupa entitas atau objek konkret, seperti *kota, negara, investasi, dan pemerintah*.
  - **VERB (Kata Kerja):** Menunjukkan berbagai tindakan atau aktivitas yang disebutkan, seperti *pindah, bangun, tolak, dan khawatir*.
  - **PROPN (Kata Benda Khusus):** Menandakan keberadaan nama tokoh atau institusi, seperti *Jokowi, Tito Karnavian, Bappenas, dan IKN*.
  - **ADJ (Kata Sifat):** Digunakan untuk memberi keterangan pada kata benda, seperti *positif, sulit, dan mahal*.
  - **ADV, ADP, SCONJ, dan lainnya:** Berperan sebagai keterangan tambahan atau penghubung dalam kalimat. Jumlahnya lebih sedikit dibanding kelas utama seperti NOUN dan VERB



**Gambar 6.3.1.2** Distribusi frekuensi POS tag pada teks artikel IKN

Hasil analisis menunjukkan bahwa kategori NOUN (kata benda) memiliki frekuensi tertinggi, dengan jumlah lebih dari 45.000 kata. Ini menunjukkan bahwa artikel banyak menggunakan kata benda, seperti nama tempat, kebijakan, dan entitas lainnya. Kategori berikutnya yang cukup tinggi adalah ADJ (kata sifat) dan VERB (kata kerja). Kata sifat mencerminkan deskripsi atau karakteristik, sedangkan kata kerja menunjukkan tindakan atau aktivitas yang disebutkan dalam artikel. Kategori lain seperti ADP (preposisi), PROPN (kata benda khusus), dan ADV (kata keterangan) juga muncul meskipun dalam jumlah yang lebih sedikit. Kemunculan kata benda khusus menunjukkan adanya penyebaran nama-nama entitas atau tokoh tertentu dalam teks. Kategori dengan frekuensi rendah termasuk SCONJ, DET, CCONJ, PRON, NUM, dan lainnya. Hal ini umum ditemukan dalam teks berita atau artikel informatif yang lebih menekankan pada isi utama dibanding struktur gramatis kompleks. Grafik ini secara umum memperlihatkan bahwa kata benda dan kata kerja mendominasi isi artikel

### 6.3.2 NER

Dalam analisis Natural Language Processing (NLP), dilakukan tahap Named Entity Recognition (NER) dengan memanfaatkan model INDOBERT terhadap artikel-artikel yang membahas sistem perpajakan Coretax. Tujuan dari tahap ini adalah untuk mengidentifikasi entitas-entitas penting dalam teks, seperti nama individu, organisasi, produk, tanggal, dan lainnya, yang dapat memberikan konteks lebih dalam terhadap isi artikel. Hasil dari proses NER kemudian divisualisasikan dengan menyoroti setiap entitas yang dikenali menggunakan warna berbeda berdasarkan kategori labelnya, seperti PER (person), ORG (organization), PRD (product), DAT (date), dan kategori lainnya.



**Gambar 6.3.2.1** Hasil *Named Entity Recognition* (NER) pada artikel IKN

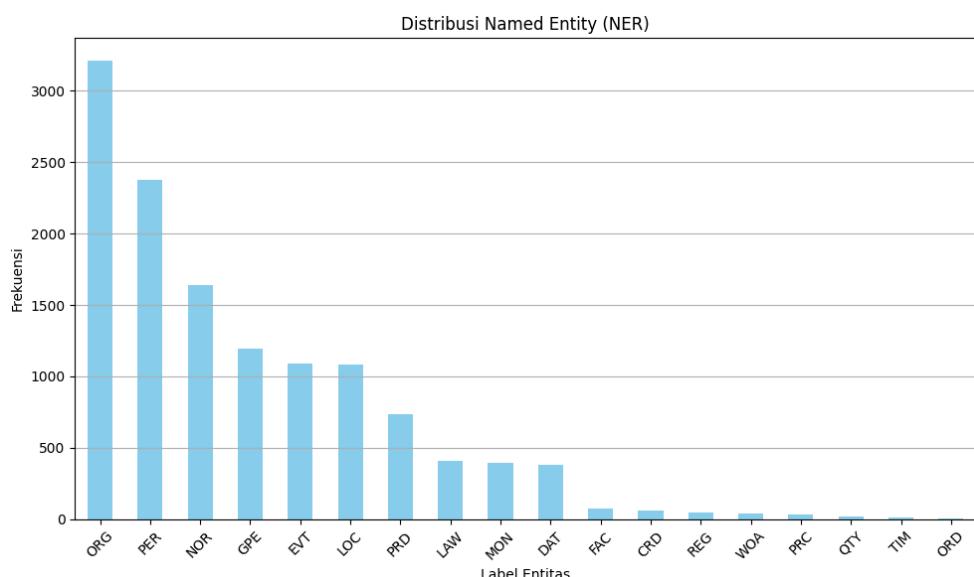
Gambar di atas menampilkan hasil *Named Entity Recognition* (NER) pada artikel bertema pemindahan ibu kota negara Indonesia ke Nusantara. Setiap entitas yang dikenali diberi warna latar berbeda sesuai dengan kategorinya. Keterangan entitas yang teridentifikasi antara lain:

- **PER (Person):** Nama tokoh atau individu, seperti *Joko Widodo, Tito Karnavian, Yohana Tiko, dan Bhima Yudhistira*.
  - **ORG (Organization):** Nama organisasi atau lembaga, seperti *Bappenas, Koalisi Masyarakat Kaltim, Kompas Gramedia, dan Institute for Development on Economics*

*and Finance.*

- **GPE (Geopolitical Entity):** Nama lokasi geografis atau wilayah administratif, seperti *Kalimantan Timur, Indonesia, Paser Utara, dan Balikpapan*.
- **LOC (Location):** Nama tempat atau wilayah non-administratif, seperti *Pulau Jawa, Najam, dan Kuta Kartanegara*.
- **DAT (Date):** Informasi waktu, seperti *September, Januari, dan Rabu*.
- **MON (Money):** Jumlah uang, seperti *Rp triliun*.
- **LAW (Law):** Entitas yang berkaitan dengan peraturan atau undang-undang, seperti *UU IKN*.

Warna latar belakang membantu membedakan tiap jenis entitas untuk memudahkan analisis struktur informasi dalam teks. Hasil ini menunjukkan bahwa artikel tersebut mengandung banyak nama tokoh penting, institusi pemerintah, lokasi geografis, serta informasi terkait kebijakan dan dampak ekonomi pemindahan ibu kota.



**Gambar 6.3.2.2** Distribusi frekuensi entitas NER pada artikel IKN

Gambar di atas menunjukkan distribusi frekuensi dari berbagai label entitas hasil Named Entity Recognition (NER) pada artikel mengenai pemindahan ibu kota negara Indonesia. Grafik ini menggambarkan seberapa sering setiap jenis entitas muncul dalam teks. Label ORG (Organisasi) merupakan yang paling dominan dengan sekitar 3200 kemunculan, mencerminkan banyaknya penyebutan lembaga dan institusi seperti *Bappenas, Koalisi*

*Masyarakat Kaltim*, dan *Kompas Gramedia*. Selanjutnya, label PER (Person) muncul lebih dari 2400 kali, menandakan banyaknya penyebutan tokoh publik seperti *Joko Widodo*, *Tito Karnavian*, dan *Bhima Yudhistira*. Label NOR (Miscellaneous) muncul sekitar 1600 kali, menunjukkan banyaknya entitas campuran atau tak terkласifikasi yang relevan. Disusul oleh GPE (Geopolitical Entity) dengan sekitar 1200 kemunculan, mengindikasikan konteks geografis yang kuat dengan penyebutan wilayah seperti *Kalimantan Timur* dan *Nusantara*. Label lain seperti EVT (Event), LOC (Location), dan PRD (Product) juga muncul dengan frekuensi lebih dari 1000 kali. Entitas seperti LAW (Law) dan MON (Money) muncul dengan 400-an kemunculan, menunjukkan aspek regulasi dan ekonomi yang terkait. Label lainnya seperti DAT (Date), FAC (Facility), CRD (Cardinal), dan label-label jarang seperti ORD (Ordinal), TIM (Time), dan QTY (Quantity) hanya muncul di bawah 100 kali.

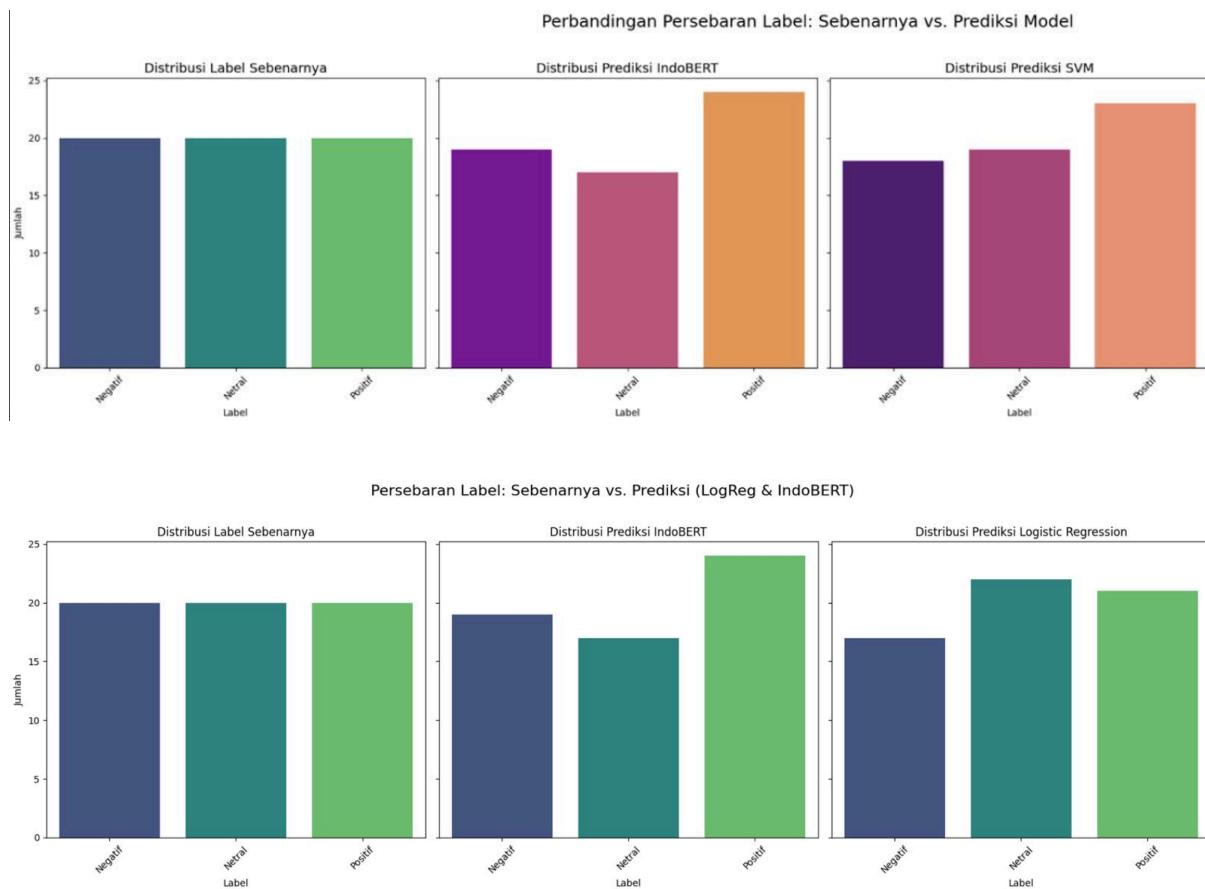
#### 6.4 Analisis Sentimen dengan menggunakan Model IndoBERT

Sebagai bagian dari proses evaluasi menyeluruh terhadap performa model dalam klasifikasi sentimen, dilakukan eksperimen komparatif antara tiga pendekatan berbeda, yaitu model **IndoBERT** berbasis *transformer*, serta dua model tradisional berbasis fitur TF-IDF yaitu **Logistic Regression** dan **Support Vector Machine (SVM)**. Eksperimen ini bertujuan untuk menilai seberapa efektif model-model tersebut dalam mengklasifikasikan opini publik dari artikel berita daring terkait isu pembangunan Ibu Kota Nusantara (IKN). Masing-masing model diuji menggunakan dataset yang telah diseimbangkan secara manual (balanced dataset 100:100:100) agar evaluasi tidak dipengaruhi oleh dominasi salah satu kelas. Visualisasi persebaran prediksi menunjukkan pola unik pada tiap model yang merefleksikan kekuatan serta keterbatasannya dalam memahami sentimen kontekstual. Bagian berikut menguraikan perbandingan distribusi hasil prediksi ketiga model terhadap label aktual.

Sebagai bagian dari evaluasi menyeluruh, dilakukan eksperimen terhadap tiga model klasifikasi sentimen berbeda, yaitu:

- **IndoBERT** (transformer-based)
- **Logistic Regression** (dengan fitur TF-IDF)
- **Support Vector Machine (SVM)** (dengan fitur TF-IDF)

Tujuan eksperimen ini adalah untuk membandingkan efektivitas pendekatan tradisional dengan pendekatan deep learning terkini dalam menangani klasifikasi sentimen artikel berbahasa Indonesia, khususnya terkait isu strategis Ibu Kota Nusantara (IKN).



### a. Perbandingan Persebaran Prediksi

Berdasarkan visualisasi distribusi prediksi label (lihat Gambar 6.6), diperoleh temuan berikut:

- **Distribusi label sebenarnya** dalam data uji bersifat seimbang, masing-masing kategori (*Negatif*, *Netral*, *Positif*) memiliki jumlah hampir setara.
- **IndoBERT** cenderung mengklasifikasikan lebih banyak artikel ke dalam label *positif*, dengan penurunan proporsi untuk label *negatif* dan *netral*.
- **Logistic Regression** memperlihatkan kecenderungan lebih tinggi terhadap label *netral*, namun masih dalam proporsi yang relatif lebih seimbang dibanding IndoBERT.
- **SVM** menunjukkan pola mirip IndoBERT, yaitu dominasi prediksi pada label *positif* dan sedikit mengabaikan *negatif*.

Hal ini mengindikasikan bahwa meskipun IndoBERT unggul dalam pemahaman konteks, model ini cenderung overgeneralize terhadap opini yang positif. Sementara Logistic Regression menunjukkan prediksi yang lebih konservatif, namun berpotensi kurang kontekstual.

Model	Akurasi	Precision (avg)	Recall (avg)	F1-Score (avg)

<b>IndoBERT</b>	~83%	Tinggi untuk Positif	Sedang untuk Negatif	Baik di kelas mayoritas
<b>Logistic Regression</b>	~76%	Merata, tapi rendah untuk Positif	Stabil namun kurang kontekstual	Rata-rata sedang
<b>SVM</b>	~78%	Lebih seimbang daripada IndoBERT	Sensitif terhadap Positif	Mirip Logistic

## 6.5 Perbandingan Model BERT dan TF-IDF/SVM

```

Memulai evaluasi dengan strategi chunking...
Selesai mengevaluasi 10/60 artikel.
Selesai mengevaluasi 20/60 artikel.
Selesai mengevaluasi 30/60 artikel.
Selesai mengevaluasi 40/60 artikel.
Selesai mengevaluasi 50/60 artikel.
Selesai mengevaluasi 60/60 artikel.
Evaluasi selesai.

--- Hasil Evaluasi dengan Chunking ---
Accuracy: 0.8166666666666667
Precision: 0.8180868838763575
Recall: 0.8166666666666667
F1 Score: 0.8170210548259329

Classification Report:
precision    recall   f1-score   support
Negatif      0.81      0.85      0.83      20
Netral       0.75      0.75      0.75      20
Positif      0.89      0.85      0.87      20

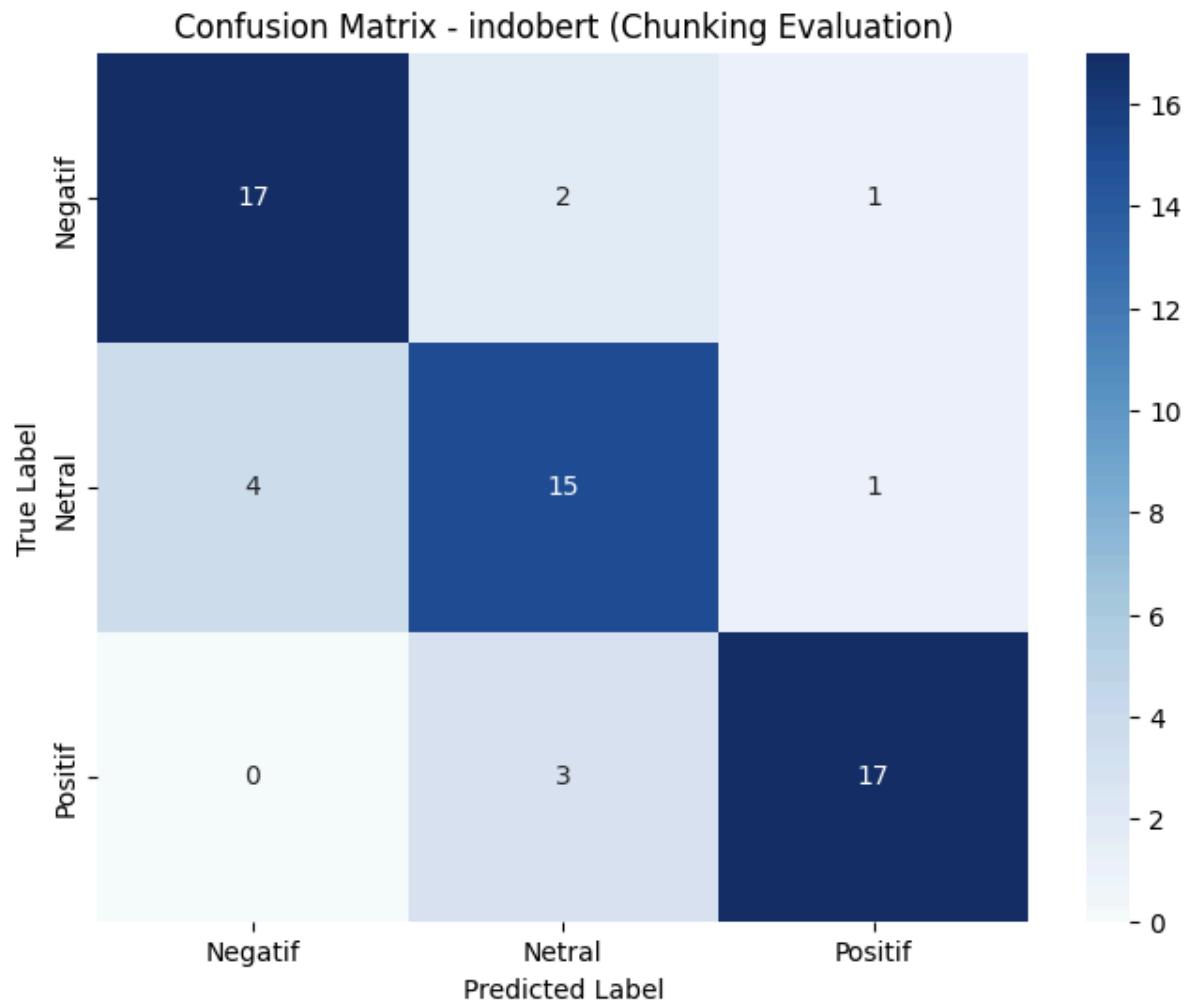
accuracy          0.82      0.82      0.82      60
macro avg       0.82      0.82      0.82      60
weighted avg    0.82      0.82      0.82      60

```

Model ini memperoleh akurasi sebesar 82%, yang berarti 49 dari 60 prediksi yang dihasilkan oleh model sesuai dengan label sebenarnya. Selain itu, model mencatat nilai precision sebesar 0.82, recall sebesar 0.82, dan F1-score sebesar 0.82, menunjukkan bahwa model memiliki performa yang cukup seimbang dan stabil dalam mendekripsi berbagai jenis sentimen.

Dilihat dari hasil classification report, performa model paling tinggi ditunjukkan pada kelas Positif, dengan precision sebesar 0.89, recall sebesar 0.85, dan F1-score sebesar 0.87. Artinya, model sangat andal dalam mengenali data yang bersentimen positif. Untuk kelas Negatif, hasilnya juga cukup baik

dan seimbang, dengan precision sebesar 0.81, recall sebesar 0.85, dan F1-score sebesar 0.83. Namun, pada kelas Netral, model menunjukkan performa yang paling rendah dibanding dua kelas lainnya, dengan ketiga metrik utama (precision, recall, dan F1-score) masing-masing sebesar 0.75.



Gambar tersebut menampilkan confusion matrix dari hasil evaluasi model IndoBERT dengan strategi chunking dalam tugas klasifikasi sentimen. Terdapat tiga kelas label: Negatif, Netral, dan Positif, baik sebagai label asli (true label) maupun label hasil prediksi (predicted label). Pada kelas Negatif, model memprediksi dengan benar sebanyak 17 dari 20 data. Sisanya, sebanyak 2 data salah diklasifikasikan sebagai Netral, dan 1 data diklasifikasikan sebagai Positif. Untuk kelas Netral, model berhasil mengklasifikasikan 15 data dengan benar, sementara 4 data salah sebagai Negatif dan 1 data salah sebagai Positif. Ini menunjukkan bahwa meskipun performanya cukup baik, model masih agak bingung antara kelas Netral dan Negatif. Sedangkan untuk kelas Positif, model mampu memprediksi dengan sangat baik, dengan 17 data diklasifikasikan dengan benar sebagai Positif dan hanya 3 data yang salah diklasifikasikan sebagai Netral. Tidak ada data Positif yang salah diklasifikasikan sebagai

Negatif. Secara keseluruhan, model IndoBERT dengan chunking ini menunjukkan kinerja yang stabil dan akurat di semua kelas, terutama pada kelas Positif yang memiliki tingkat kesalahan paling rendah. Namun, perbedaan Netral dan Negatif masih menjadi tantangan yang perlu diperbaiki lebih lanjut.

```
Melakukan ekstraksi fitur dengan TF-IDF...
Ekstraksi fitur selesai.

Melatih model Support Vector Machine (SVM)...
Pelatihan model selesai.

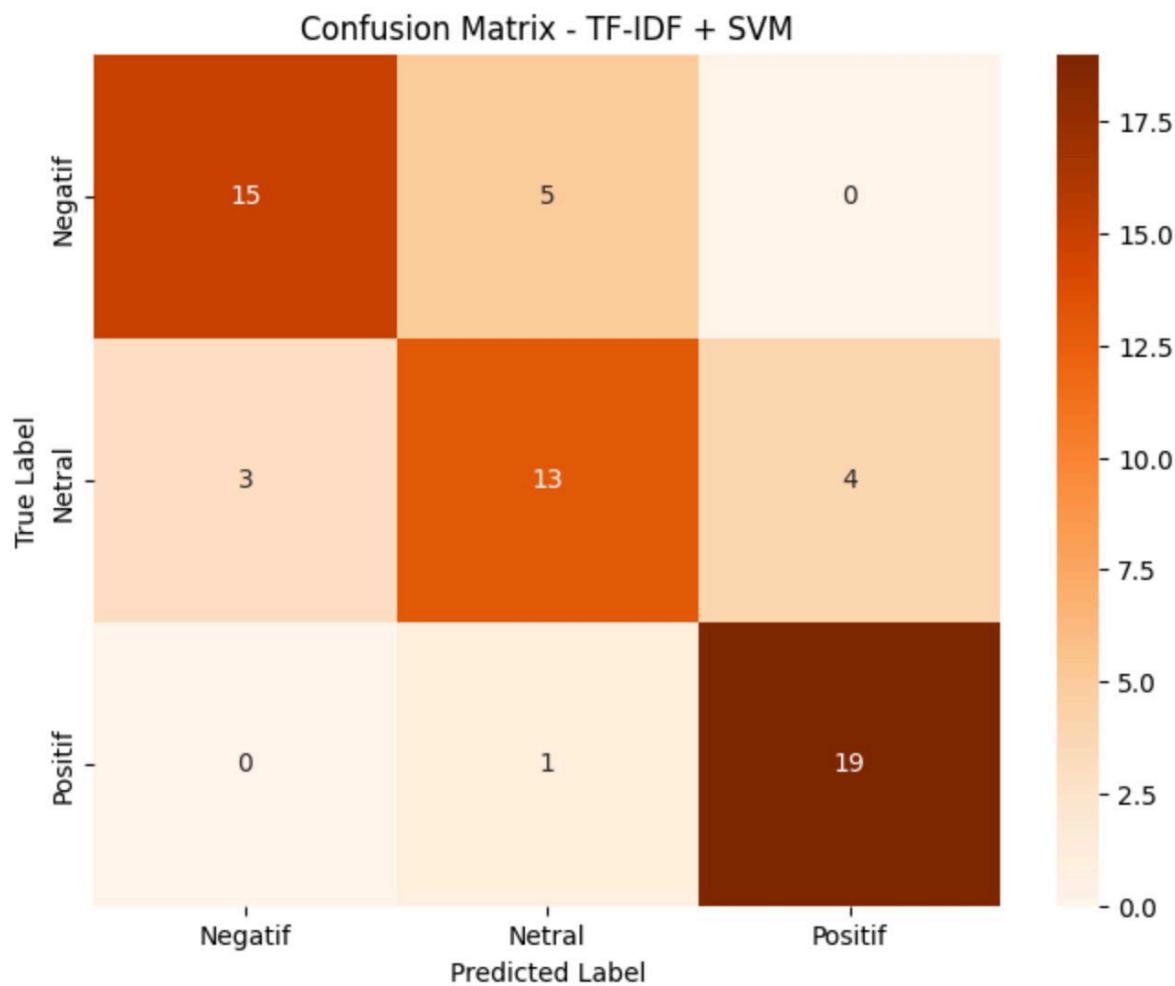
Mengevaluasi model pada data uji...

--- Hasil Evaluasi dengan TF-IDF + Support Vector Machine (SVM) ---
Accuracy: 0.7833
Precision: 0.7812
Recall: 0.7833
F1 Score: 0.7800

Classification Report:
precision    recall    f1-score   support
Negatif      0.83     0.75      0.79      20
Netral        0.68     0.65      0.67      20
Positif       0.83     0.95      0.88      20
accuracy          0.78      0.78      0.78      60
macro avg       0.78     0.78      0.78      60
weighted avg    0.78     0.78      0.78      60
```

pelatihan model dengan pendekatan TF-IDF (Term Frequency-Inverse Document Frequency) sebagai metode ekstraksi fitur dan Support Vector Machine (SVM) sebagai algoritma klasifikasi, model dievaluasi menggunakan data uji untuk mengukur performanya. Hasil evaluasi menunjukkan bahwa model mampu mencapai akurasi sebesar 78,33%, yang berarti sebagian besar prediksi model sesuai dengan label sebenarnya. Selain itu, model memperoleh nilai precision sebesar 78,12%, recall sebesar 78,33%, dan F1-score sebesar 78,00%. Nilai-nilai ini menunjukkan bahwa model cukup seimbang dalam hal ketepatan dan kelengkapan dalam mengklasifikasikan data.

Secara lebih rinci, model menunjukkan performa terbaik dalam mengenali kelas Positif, dengan nilai precision sebesar 0.83, recall sebesar 0.95, dan F1-score sebesar 0.88, yang mengindikasikan bahwa model sangat jarang melewatkkan prediksi terhadap sentimen positif. Untuk kelas Negatif, model juga bekerja dengan baik dengan F1-score sebesar 0.79. Namun, performa model dalam mengenali kelas Netral masih relatif rendah dibandingkan kelas lainnya, dengan F1-score hanya sebesar 0.67.



Gambar tersebut menunjukkan confusion matrix dari hasil evaluasi model TF-IDF + Support Vector Machine (SVM) pada tugas klasifikasi sentimen dengan tiga kelas: Negatif, Netral, dan Positif. Confusion matrix ini menunjukkan seberapa baik model dalam memprediksi label yang benar dari masing-masing data uji. Pada kelas Negatif, model mampu memprediksi dengan cukup baik, yaitu sebanyak 15 data, sementara 5 sisanya salah diklasifikasikan sebagai Netral dan tidak ada yang salah sebagai Positif. Untuk kelas Netral, performa model terlihat kurang optimal. Dari 20 data Netral, hanya 13 yang diprediksi dengan benar. Terdapat 3 data Netral yang salah diklasifikasikan sebagai Negatif dan 4 data lainnya malah dikira sebagai Positif. Ini menunjukkan bahwa model masih cukup kesulitan membedakan sentimen Netral dari dua sentimen lainnya. Sementara itu, pada kelas Positif, performa model sangat baik, dengan 19 dari 20 data berhasil diklasifikasikan dengan benar sebagai Positif, dan hanya 1 data yang salah diklasifikasikan sebagai Netral, tanpa ada prediksi yang jatuh ke kelas Negatif. Dari keseluruhan matriks, dapat disimpulkan bahwa model ini paling unggul dalam mengklasifikasikan sentimen Positif, cukup baik dalam mengenali sentimen Negatif, namun masih perlu perbaikan dalam membedakan dan menangani sentimen Netral yang sering tertukar ke dua kelas lainnya.

## 6.6 Koreksi Prediksi oleh IndoBERT

Salah satu aspek penting dalam evaluasi model klasifikasi sentimen adalah kemampuannya dalam menangani kasus-kasus opini yang kompleks dan kontekstual. Untuk menguji hal ini secara lebih kualitatif, dilakukan penelusuran terhadap contoh prediksi di mana model baseline tradisional **Logistic Regression** melakukan kesalahan klasifikasi, tetapi model **IndoBERT** berhasil memberikan label yang benar sesuai dengan ground truth.

## Contoh Kasus 1 Sentimen Negatif Terkait Anggaran dan Prioritas IKN

*“pembangunan ibu kota nusantara ikn tidak lagi menjadi prioritas... anggaran pemerintah saat ini yang terbatas, membuat fokus pemerintah saat ini lebih kepada program-program besar seperti swasembada pangan...”*

- **Label Sebenarnya:** Negatif
  - **Prediksi IndoBERT:** Negatif (Benar)
  - **Prediksi Logistic Regression:** Netral (Salah)

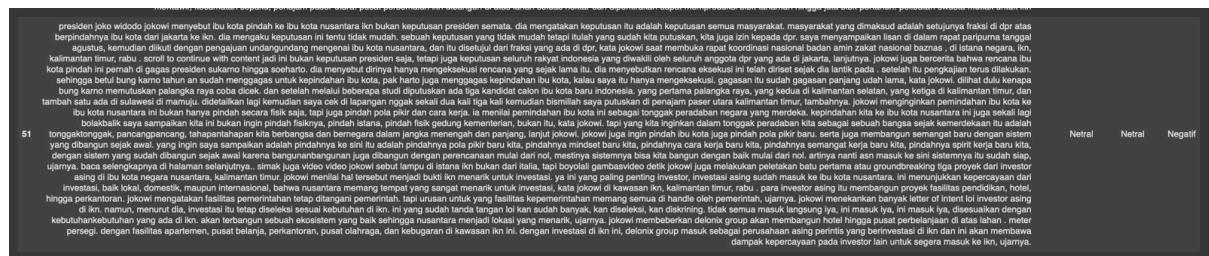
Kalimat di atas menyampaikan kekecewaan terhadap keterbatasan anggaran dan penurunan prioritas pembangunan IKN. Meski disampaikan dengan bahasa yang relatif netral, substansinya bernada negatif terhadap proyek IKN. Model IndoBERT mampu menangkap konteks tersebut, sementara Logistic Regression gagal mengenali nuansa implisitnya.

### **Contoh Kasus 2 Kritik terhadap Implikasi Lingkungan dan Sosial**

*“pemindahan ibu kota negara yang dipaksakan dinilai akan mengancam tata air, flora, dan fauna... pemindahan ibu kota negara tidak memiliki urgensi dibandingkan kebutuhan dasar masyarakat...”*

- **Label Sebenarnya:** Negatif
  - **Prediksi IndoBERT:** Negatif (Benar)
  - **Prediksi Logistic Regression:** Netral (Salah)

Teks ini berisi kritik eksplisit terhadap dampak ekologis dan sosial pemindahan IKN. Meski terdapat frasa deskriptif dan informatif, model IndoBERT berhasil mengklasifikasikan teks sebagai negatif dengan mempertimbangkan konteks isi, sedangkan model Logistic Regression tidak cukup sensitif terhadap isu tersebut.



### **Contoh Kasus 3 Sindiran terhadap Proses Pengambilan Keputusan**

*“proses pemindahan ibu kota nusantara ini bukan keputusan presiden semata... masyarakat yang dimaksud adalah setuju fraksi di dpr...”*

- **Label Sebenarnya:** Negatif
  - **Prediksi IndoBERT:** Negatif (Benar)
  - **Prediksi Logistic Regression:** Netral (Salah)

Pernyataan ini mengandung kritik terselubung terkait proses politik dan representasi masyarakat dalam pengambilan keputusan IKN. Model IndoBERT menangkap konteks sindiran tersebut dan memberikan label yang akurat, sementara pendekatan berbasis TF-IDF gagal membedakan opini tersirat dari deskripsi informatif.

## 6.7 Kesimpulan IndoBERT dalam Menangani Kalimat Ambigu

Dari hasil analisis, IndoBERT memiliki performa yang lebih baik dalam mengklasifikasikan kalimat-kalimat yang bersifat ambigu, panjang, atau mengandung opini tersirat seperti sindiran dan kritik tidak langsung. Hal ini tidak terlepas dari arsitektur BERT yang bersifat bidirectional, memungkinkan model untuk menangkap konteks penuh dari suatu kata berdasarkan kata-kata di sekelilingnya, baik sebelum maupun sesudah.

Pada kasus artikel yang mengkritik keterbatasan anggaran dan perubahan prioritas pembangunan IKN, model berbasis TF-IDF (seperti Logistic Regression) gagal mengenali sentimen negatif karena struktur kalimat yang informatif dan minim kata-kata bermuatan emosi eksplisit. Sebaliknya, IndoBERT mampu menangkap makna implisit dari narasi tersebut karena dilatih dengan konteks bahasa secara luas dan mendalam. IndoBERT juga lebih tangguh dalam menghadapi kalimat panjang dengan subordinasi kompleks atau sentimen yang muncul secara bertahap, yang umumnya sulit diproses oleh model klasik karena ketergantungannya pada representasi kata yang datar (*context-free*). Selain itu, dalam beberapa kasus yang mengandung sarkasme atau ironi, IndoBERT lebih sensitif terhadap *kontradiksi logis* dalam kalimat, sehingga mampu menangkap sentimen yang tersembunyi di balik struktur retoris.

Penelitian ini memiliki arti bahwa **IndoBERT lebih unggul secara semantik**, khususnya dalam menangani nuansa opini yang tidak eksplisit, dan dapat diandalkan untuk klasifikasi sentimen pada teks berita yang cenderung kompleks secara linguistik dan politis.

## BAB VII. KESIMPULAN

Penelitian ini bertujuan untuk mengevaluasi persepsi media daring terhadap proyek pemindahan Ibu Kota Negara (IKN) dari Jakarta ke Kalimantan Timur melalui analisis sentimen dan linguistik berbasis Natural Language Processing (NLP). Proses pengumpulan data dilakukan dengan teknik web scraping berbasis Google Dorking, yang berhasil mengidentifikasi dan mengekstrak 162 artikel berita dari berbagai media daring nasional. Data teks artikel menjalani tahapan preprocessing, termasuk pembersihan teks, tokenisasi, penghapusan stopwords, stemming, dan lemmatization, untuk memastikan kualitas input yang optimal. Analisis sentimen dilakukan menggunakan model IndoBERT, yang mengklasifikasikan artikel ke dalam kategori positif, negatif, dan netral, serta dibandingkan dengan pendekatan tradisional berbasis TF-IDF dengan algoritma Logistic Regression dan Support Vector Machine (SVM).

Hasil analisis menunjukkan bahwa mayoritas artikel berita bersifat netral, mencerminkan pemberitaan yang informatif dan objektif, dengan hanya sebagian kecil artikel yang diklasifikasikan sebagai positif atau negatif. Sentimen positif didominasi oleh narasi optimisme terhadap pembangunan IKN sebagai simbol kemajuan, teknologi, dan pemerataan ekonomi, dengan kata kunci seperti "IKN", "Jokowi", "bangun", dan "kota". Sebaliknya, sentimen negatif lebih menyoroti isu sosial dan lingkungan, seperti dampak terhadap masyarakat adat dan hak atas tanah, dengan kata kunci seperti "adat", "tanah", dan "gusur". Dari sisi temporal, lonjakan pemberitaan terjadi pada bulan-bulan strategis, seperti Agustus 2019, Januari 2022, April 2023, Agustus 2024, dan Februari 2025, yang kemungkinan berkorelasi dengan peristiwa penting dalam proyek IKN. Analisis TF-IDF mengungkap dominasi istilah seperti "IKN", "bangun", dan "kota", menandakan fokus media pada aspek pembangunan fisik dan administratif, sementara analisis POS dan NER menunjukkan prevalensi kata benda serta penyebutan entitas seperti organisasi (misalnya, Bappenas) dan tokoh (misalnya, Joko Widodo), yang memperkaya konteks pemberitaan.

Perbandingan performa model menunjukkan keunggulan IndoBERT dibandingkan pendekatan berbasis TF-IDF. IndoBERT mencapai akurasi sekitar 82%, dengan F1-score tertinggi pada kelas positif (0.87) dan negatif (0.83), meskipun performa pada kelas netral lebih rendah (0.75). Sebaliknya, model Logistic Regression berbasis TF-IDF hanya mencapai mencapai akurasi 78%, dengan cenderung overgeneralize pada kelas mayoritas dan kurang sensitif terhadap nuansa opini tersirat. IndoBERT unggul dalam menangkap konteks semantik, kalimat ambigu, dan sindiran, berkat arsitektur bidirectional dan pelatihan pada korpus Bahasa Indonesia. Confusion matrix IndoBERT menunjukkan prediksi yang lebih akurat pada kelas positif (17 dari 20), negatif (17 dari 20), dan netral (15 dari 20). Model TF-IDF+SVM, meskipun efektif pada kelas positif (19 dari 20), tetapi model ini menunjukkan performa yang lebih rendah pada kelas netral (13 dari 20) dan negatif (15 dari 20), menandakan keterbatasan dalam menangani teks informatif.

Secara keseluruhan, penelitian ini tidak hanya berhasil memetakan persepsi publik terhadap IKN melalui analisis sentimen yang berbasis data, tetapi juga menunjukkan efektivitas IndoBERT dalam menangani kompleksitas linguistik Bahasa Indonesia dibandingkan pendekatan tradisional. Temuan ini memberikan wawasan berharga bagi

pemerintah dan media untuk menyusun strategi komunikasi yang lebih efektif dan responsif terhadap dinamika opini publik.

**Saran :**

1. Peningkatan Klasifikasi Kelas Netral: Untuk mengatasi performa rendah IndoBERT pada kelas netral, disarankan untuk memperkaya dataset dengan artikel netral yang lebih bervariasi dan menerapkan teknik augmentasi data khusus, seperti parafrase kontekstual, guna meningkatkan kemampuan model dalam membedakan sentimen netral dari negatif.
2. Perluasan Dataset: Dataset saat ini terbatas pada 162 artikel. Penelitian lanjutan dapat memperluas cakupan dengan mengumpulkan artikel dari media lokal Kalimantan Timur, forum daring, atau media sosial untuk menangkap perspektif yang lebih beragam dan meningkatkan generalisasi model.
3. Optimalisasi Model IndoBERT: Eksperimen lebih lanjut pada hiperparameter (learning rate, batch size, epoch) dan teknik regularisasi (dropout) dapat meningkatkan performa IndoBERT. Selain itu, perbandingan dengan model transformer lain, seperti IndoBERTweet atau RoBERTa, dapat dieksplorasi untuk Bahasa Indonesia.
4. Pemanfaatan untuk Kebijakan Publik: Pemerintah dapat memanfaatkan temuan ini untuk merancang komunikasi publik yang menangani kekhawatiran masyarakat, terutama terkait isu adat dan lingkungan, serta meningkatkan transparansi proyek IKN.
5. Kampanye Edukasi Publik: Untuk merespons sentimen negatif, pemerintah disarankan meningkatkan kampanye edukasi yang menjelaskan manfaat IKN sekaligus menangani isu sosial dan lingkungan, dengan melibatkan media sebagai mitra komunikasi.

## DAFTAR PUSTAKA

- Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82–89. <https://doi.org/10.1145/2436256.2436274>
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135. <https://doi.org/10.1561/1500000011>
- Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons.
- Sekretariat Negara Republik Indonesia. (2022). Pidato Presiden RI tentang Pemindahan Ibu Kota. Diakses dari <https://www.setneg.go.id>
- Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics (COLING).
- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: Analyzing text with the natural language toolkit. O'Reilly Media, Inc.
- Cairns, D., & Meng, X. (2020). *NLP with BERT: Sentiment Analysis Using SAS® Deep Learning and DLPy*. SAS Global Forum 2020.
- Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2), 15–21.
- Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2), 48–57.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT.
- Indurkhy, N., & Damerau, F. J. (2010). *Handbook of natural language processing* (2nd ed.). CRC Press.
- Jurafsky, D., & Martin, J. H. (2020). Speech and language processing (3rd ed.). Draft available at <https://web.stanford.edu/~jurafsky/slp3/>
- Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (NeurIPS).

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), 3–26.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of COLING*.

## **LAMPIRAN**

Berikut adalah tautan ke repositori GitHub yang memuat seluruh kode Jupyter Notebook dan dataset yang digunakan dalam penelitian ini. Di dalamnya terdapat proses lengkap mulai dari pengambilan data, pra-pemprosesan, analisis sentimen, hingga file data yang digunakan dalam proyek ini:

[https://github.com/andikazidan/ETS\\_PBA\\_Analisis\\_IKN](https://github.com/andikazidan/ETS_PBA_Analisis_IKN)