

A high-angle, slightly blurred photograph of a business meeting. A wooden table is covered with several documents featuring colorful bar and pie charts. A person's hand is visible, pointing at a pie chart with a pen. Another person's hands are clasped in the foreground. A laptop is open on the left, and a tablet lies flat in the center. The scene is brightly lit, suggesting an office environment.

# Text Summarization

NLP - C  
Noam Chomsky

18 Oktober 2024

# Meet the Team



**Yuan Shafira**

 [linkedin.com/in/yuanshafira](https://www.linkedin.com/in/yuanshafira)



**Abdur Rozaq**

 [Abdur Rozaq](#)




**Erwin Duadja**

 [erwin-duadja-betha-sasana](#)




**Fadhillah Rojabhy**

 [Fadhillah Rojabhy](#)



**Aldythya Nugraha**

 [Aldythya Nugraha](#)

# Outline

- **Background & Problem Statement**
- **Objectives & Scope**
- **Data Preparation, Extraction**
- **Preparation (Preprocessing)**
- **Sample Artikel (Preprocessing)**
- **Exploratory Data Analysis**
- **Model Development, Training, Fine-Tuning**
- **Results**
- **Deployment**
- **Conlusion & Future Improvement**



# Background & Problem Statement

## Text Summarization:

- Metode komputasional yang digunakan untuk merangkum teks menjadi lebih singkat namun mengandung informasi penting.
- Teknik ini mengidentifikasi poin-poin kunci, frasa, dan konteks teks untuk menghasilkan ringkasan yang informatif.
- Text summarization sering digunakan dalam aplikasi analisis teks untuk membantu pengguna memahami isi dokumen atau artikel dengan lebih efisien.

## Project yang Dikembangkan:

- Mengembangkan sistem berbasis AI untuk Text Summarization, Dataset berasal dari website Liputan6.
- Terdapat 224.637 data artikel yang perlu dianalisis.

## Masalah Utama:

- Banyaknya data artikel yang perlu diolah
- Diperlukanya komputasi yang besar dalam training model.
- Perlu memilih teknik, algoritma dan model yang tepat untuk memaksimalkan performa.



# Objectives & Scope

## Tujuan Proyek:

- Melakukan eksperimen dengan algoritma BERT, dan algoritma lain.
- Optimasi performa model menggunakan hyperparameter tuning.

## Ruang Lingkup Proyek:

- Melakukan eksperimen dengan algoritma BERT dan algoritma lain
- Membandingkan performa algoritma melalui evaluasi model berdasarkan metrik akurasi dan lainnya.
- Melakukan optimasi model dengan menggunakan hyperparameter tuning.
- Publikasi hasil eksperimen dan model akhir ke platform Github dan Huggingface

## Langkah Selanjutnya:

- Menentukan algoritma terbaik melalui evaluasi performa model.
- Menyusun laporan akhir yang berisi hasil analisis dan kesimpulan proyek.
- Publikasi kode dan hasil Github dan Huggingface untuk digunakan oleh komunitas atau pengembangan lebih lanjut.



# Data Preparation, Extraction

- **Dataset : id\_liputan6 (2000 - 2010)**
- **Canonical**
  - Train
  - Test
  - Dev
- **Xtreme**
  - Test
  - Dev

Data	Jumlah
Jumlah Data Train	193.883
Jumlah Data Test	10.972
Jumlah Data Dev	10.972
Jumlah Data Xtreme Test	4.948
Jumlah Data Xtreme Dev	3.862
Total Data	224.637

- **id:** id dari artikel
- **url:** alamat url asli artikel
- **clean\_article:** isi teks asli artikel
- **clean\_summary:** abstraktif summary artikel
- **extractive\_summary:** ekstraktif summary artikel

Kata	Jumlah (juta)
Jumlah Kata Train	45 M
Jumlah Kata Test	2 M
Jumlah Kata Dev	2 M
Jumlah Kata Xtreme Test	1 M
Jumlah Kata Xtreme Dev	1 M
Total Kata	52 M

# Preprocessing

- **Teks Cleaning**

- **Konversi ke Lowercase:** Semua teks diubah menjadi huruf kecil.
- **kata berulang :** kata yang berulang (dengan tanda penghubung) menjadi terpisah ditambah spasi
- **Remove Spesific Text:**
  - hilangkan teks yang memiliki tanda kurun () dan [] / author-tanggal.
  - hilangkan awalan liputan6.com kemudian setelah itu tanda : Misal Liputan.com Jakarta:
  - menghilangkan teks liputan6.com, liputan6 kalau masih ada
- **Hapus Special Characters:** Karakter khusus seperti tanda baca dihapus.
- **Hapus Spasi Berlebih:** Menghilangkan spasi berlebih dalam teks.





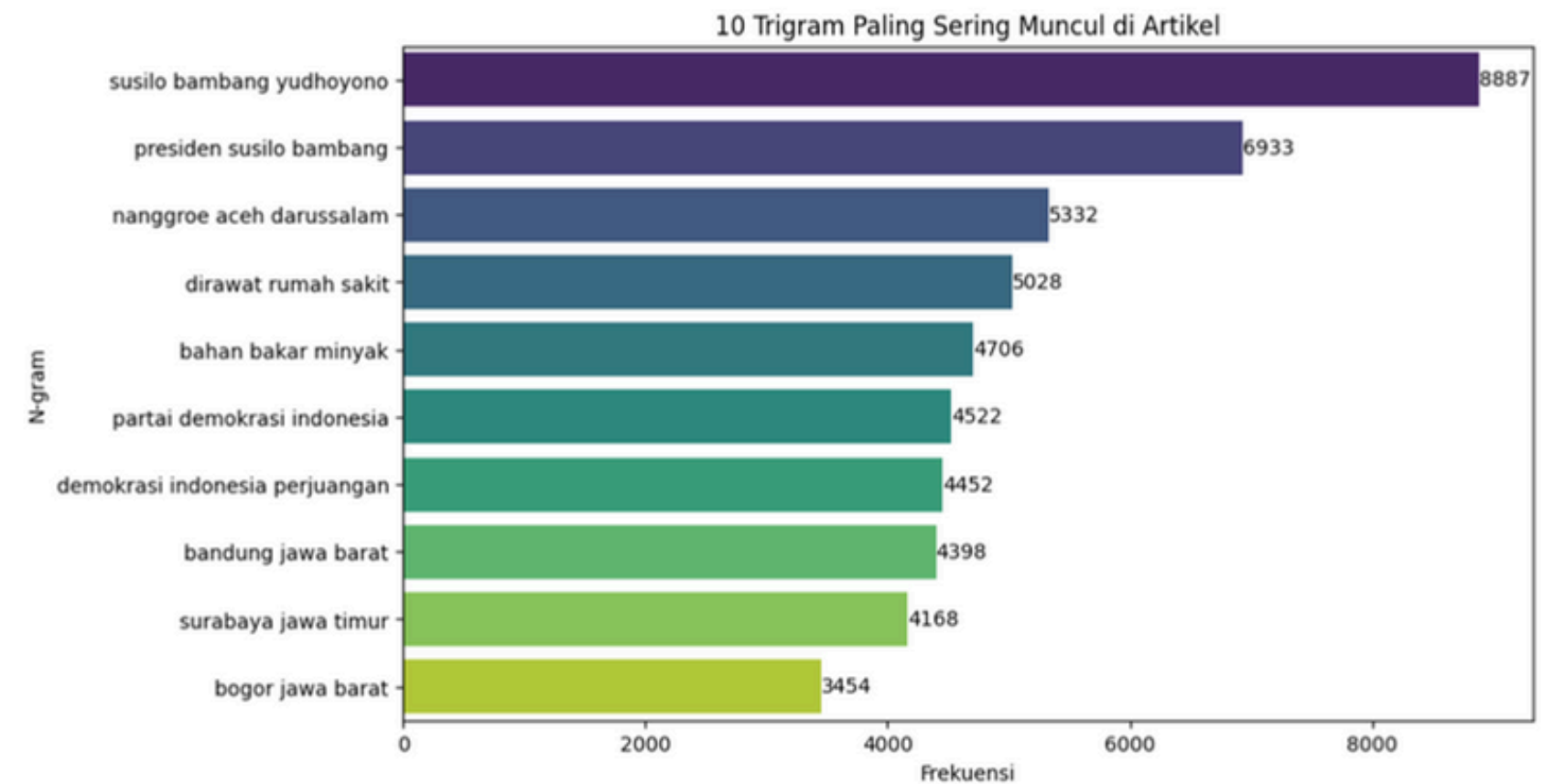
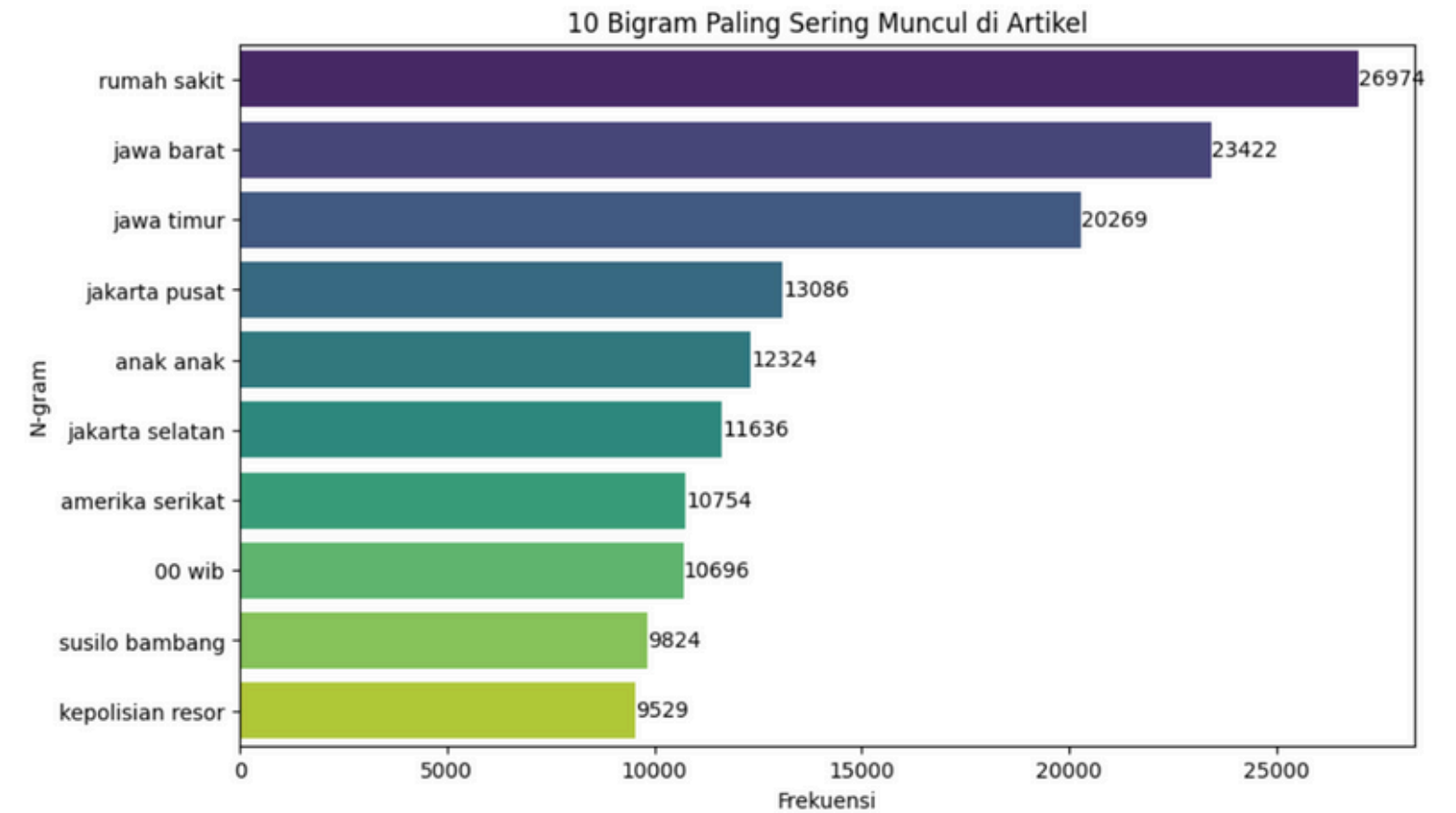
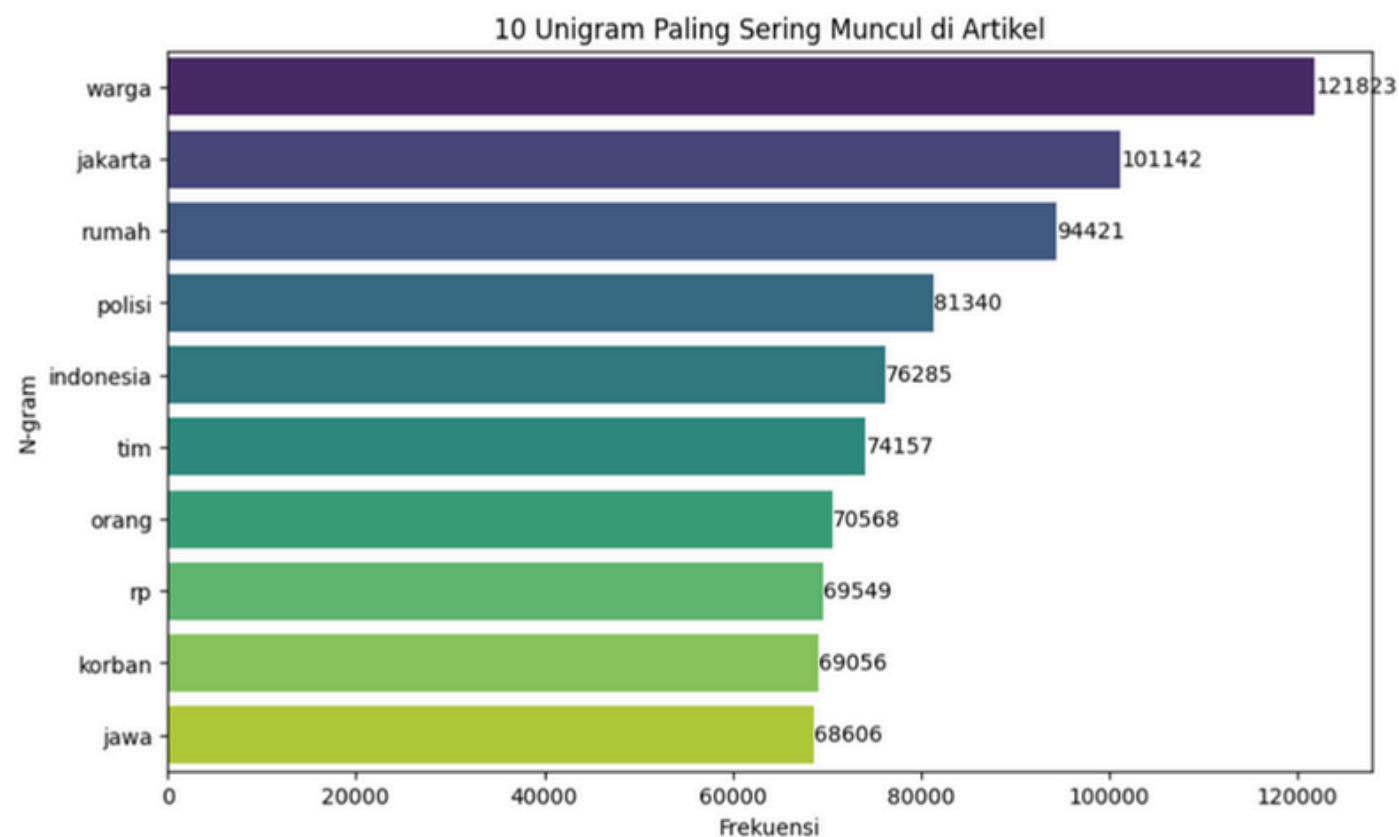
# Sample Artikel (Preprocessing)

- 1. **Index:** 176761
- 2. **Url:** <https://www.liputan6.com/news/read/71898/warga-masih-bingung-saat-simulasi-pemilu>
- 3.1. **Clean Article:**
  - Liputan6.com, Bandung: Ratusan warga Kecamatan Andir, Kota Madya Bandung, Jawa Barat, antusias mengikuti sosialisasi teknik pencoblosan Pemilihan Umum 2004 yang digelar Komisi Pemilihan Umum Daerah Bandung, Senin(9/2). Walau secara keseluruhan tak ada kesulitan, warga tetap mengaku bingung. Soalnya ukuran kertas suara lebih besar dibanding ukuran bilik suara. Dalam sosialisasi teknis pencoblosan, KPU Bandung memang langsung memperagakan dengan menggunakan bilik suara asli serta kertas suara sesuai ukuran yang sebenarnya. Hal ini dimaksudkan agar dalam pelaksanaan pemilu nanti para calon pemilih tak mengalami kesulitan lagi. Seperti diketahui, kertas suara berukuran 48 X 84 sentimeter. Sementara bilik suara hanya berukuran 50 X 50 sentimeter dengan tinggi 60 sentimeter [baca: Pemilu Sekarang Memang Berbeda].(ICH/Patria Hidayat dan Taufik Hidayat).
- 3.2. **Prep Clean Article:**
  - ratusan warga kecamatan andir kota madya bandung jawa barat antusias mengikuti sosialisasi teknik pencoblosan pemilihan umum 2004 yang digelar komisi pemilihan umum daerah bandung senin walau secara keseluruhan tak ada kesulitan warga tetap mengaku bingung soalnya ukuran kertas suara lebih besar dibanding ukuran bilik suara dalam sosialisasi teknis pencoblosan kpu bandung memang langsung memperagakan dengan menggunakan bilik suara asli serta kertas suara sesuai ukuran yang sebenarnya hal ini dimaksudkan agar dalam pelaksanaan pemilu nanti para calon pemilih tak mengalami kesulitan lagi seperti diketahui kertas suara berukuran 48 x 84 sentimeter sementara bilik suara hanya berukuran 50 x 50 sentimeter dengan tinggi 60 sentimeter
- 3.4. **Words Length:** 114
- 4.1. **Clean Summary:**
  - Walau secara keseluruhan tak ada kesulitan, warga Kecamatan Andir, Bandung, Jawa Barat, mengaku tetap bingung. Soalnya ukuran kertas suara lebih besar dibanding ukuran bilik suara.
- 4.2. **Prep Clean Summary:**
  - walau secara keseluruhan tak ada kesulitan warga kecamatan andir bandung jawa barat mengaku tetap bingung soalnya ukuran kertas suara lebih besar dibanding ukuran bilik suara
- 4.3. **Words Length:** 25
- 5.1. **Extractive Summary:**
  - Walau secara keseluruhan tak ada kesulitan, warga tetap mengaku bingung. Soalnya ukuran kertas suara lebih besar dibanding ukuran bilik suara.
- 5.2. **Prep Extractive Summary:**
  - walau secara keseluruhan tak ada kesulitan warga tetap mengaku bingung soalnya ukuran kertas suara lebih besar dibanding ukuran bilik suara
- 5.3. **Words Length:** 20
- 6.1. **Ratio Kata:**
  - \*. **Clean Article - Clean Summary:**  $114/25 = 4.56$
  - \*. **Clean Article - Extractive Summary:**  $114/20 = 5.7$
- 6.2. **Lokasi Berita:**
  - Bandung
- 6.3. **Tanggal Berita:**
  - Senin(9/2)
- 6.4. **Penulis Berita:**
  - ICH/Patria Hidayat dan Taufik Hidayat
    -

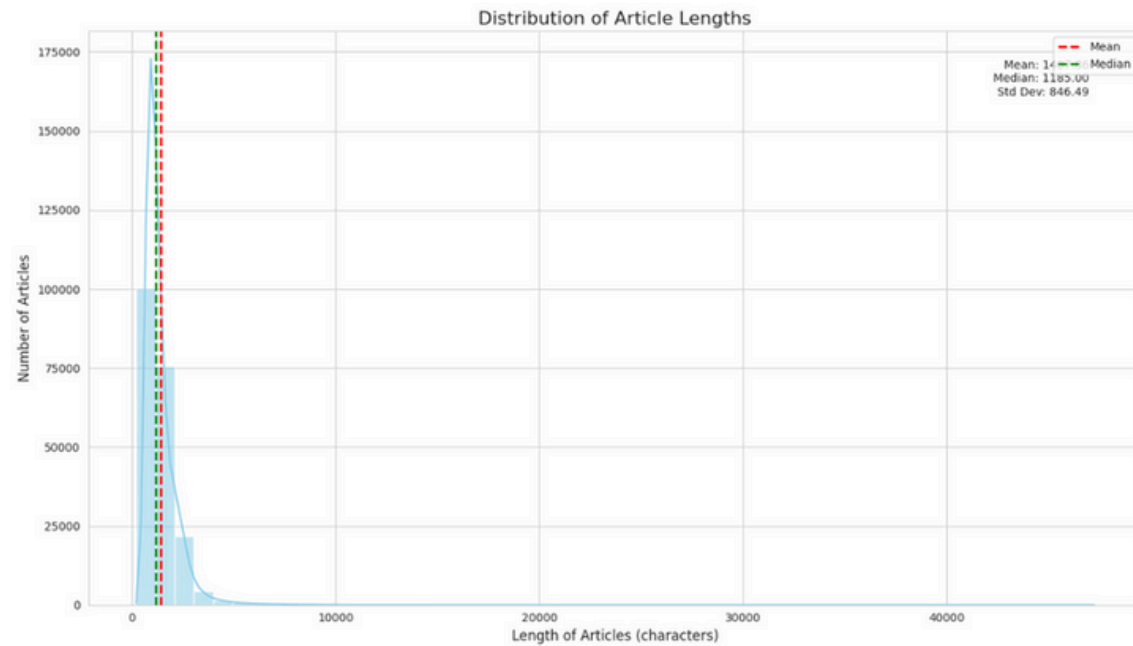


# Exploratory Data Analysis

- Unigram, kata warga frekuensi tertinggi
- Bigram, kata Rumah sakit frekuensi tertinggi
- Trigram, Kata susilo bambang yudhoyono frekuensi tertinggi
- karena ini data tahun 2000 - 2010, wajar jika waktu itu SBY sebagai presiden menjadi salah 1 yang paling sering di sebut di dataset, baik di bigram maupun trigram

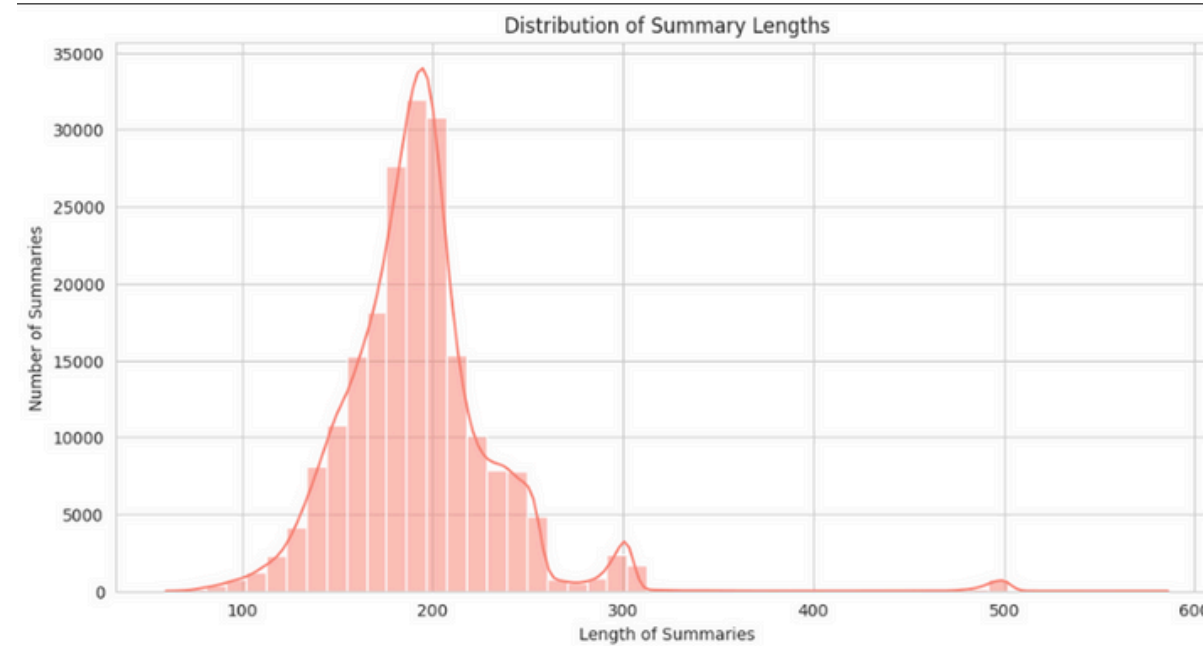


# Exploratory Data Analysis



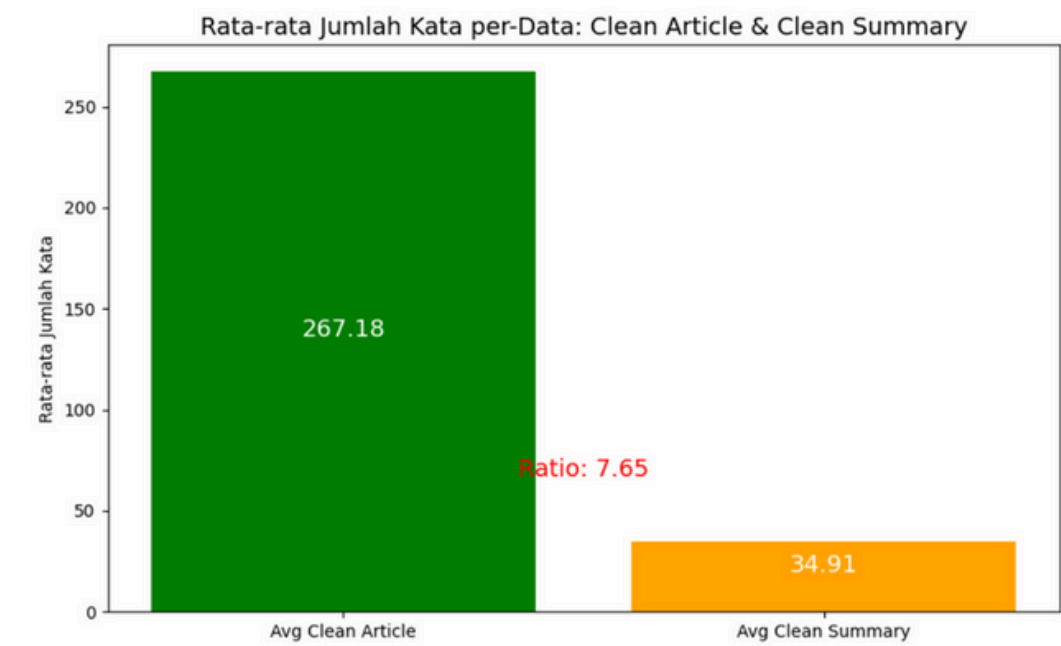
- Kemiringan: 4.42
- Kurtosis: 72.91
- Articles longer than 10,000 characters: 83 (0.04%)
- Articles shorter than 1,000 characters: 70526 (34.43%)

distribusi panjang artikel sangat miring ke kanan dengan ekor panjang yang signifikan, sementara sebagian besar artikel relatif pendek.



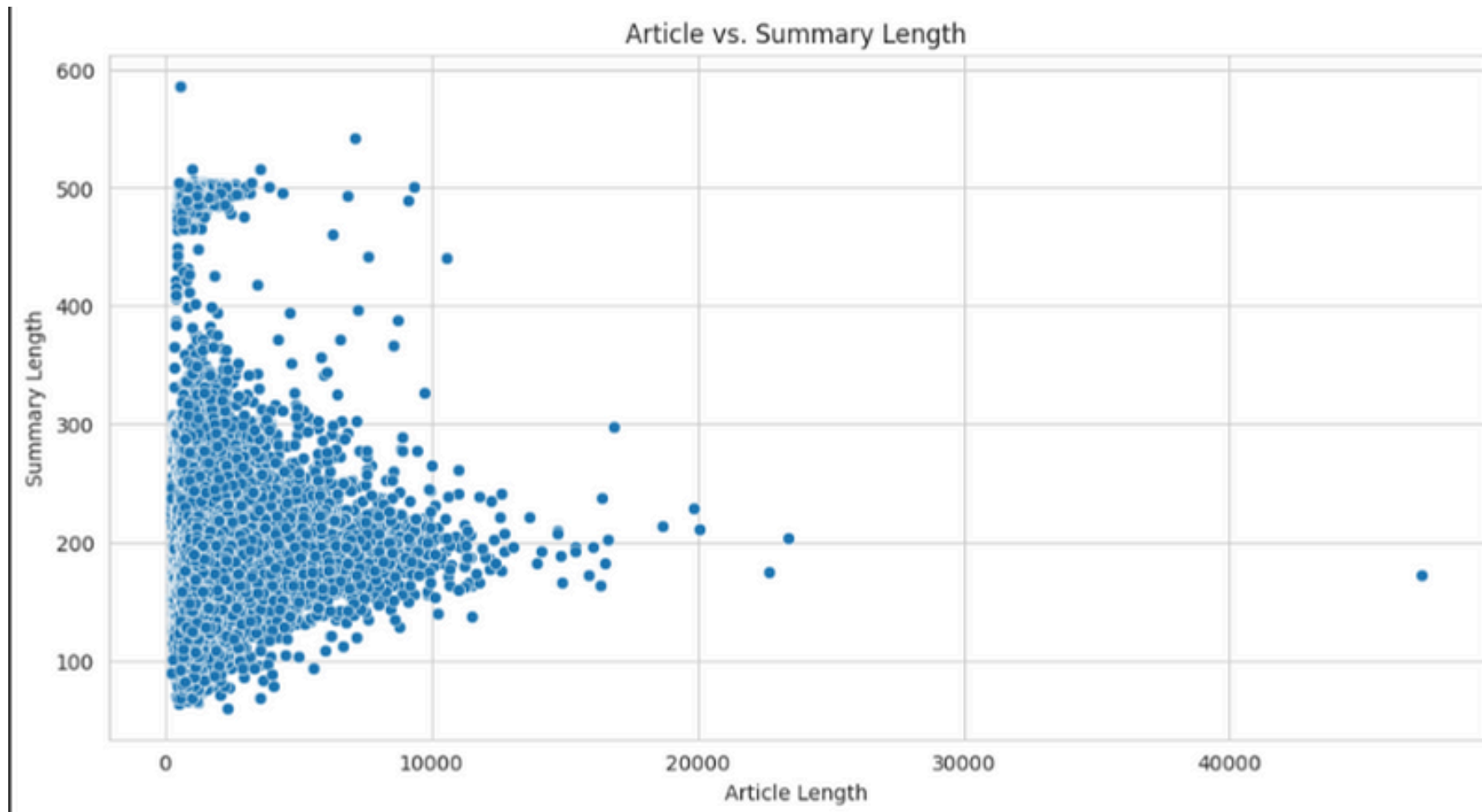
- Kemiringan: 2.22
- Kurtosis: 13.87
- Summaries longer than 1,000 characters: 0 (0.00%)
- Summaries shorter than 100 characters: 1003 (0.49%)

distribusi ringkasan artikel menunjukkan bahwa sebagian besar ringkasan sangat pendek. Kemiringan yang positif dan kurtosis yang tinggi menunjukkan bahwa ada konsentrasi ringkasan yang lebih banyak di sekitar panjang yang lebih pendek, dan tidak ada variasi panjang yang signifikan



- Rata-rata panjang kata article 267.18,
- rata-rata panjang kata summary 34,91
- ratio perbandingan : 7.65

# Exploratory Data Analysis



- Koefisien korelasi sebesar 0.13 menunjukkan hubungan yang sangat lemah dan positif antara panjang artikel dan panjang ringkasan.
- Rasio kompresi rata-rata sebesar 0.17 berarti rata-rata panjang ringkasan adalah sekitar 17% dari panjang artikel.

# Model Development, Training, Fine-Tuning

Train : 193.883

Val : 10.972

Test : 10.972

Schema	Train Data	Learning-rate	Batch Size
GPU	20000	5e-5	8
CPU	5000	3e-5	4


Pretrained Model
cahya/bert-base-indonesian-1.5G, cahya/gpt2-small-indonesian-522M
cahya/bert2bert-indonesian-summarization
panggi/t5
cahya/bert2gpt-indonesian-summarization

Models 15


indo

Full-text search


Sort: Most downloads

 cahya/bert2bert-indonesian-summarization

Summarization • Updated Jan 29, 2021 • 582 • 4

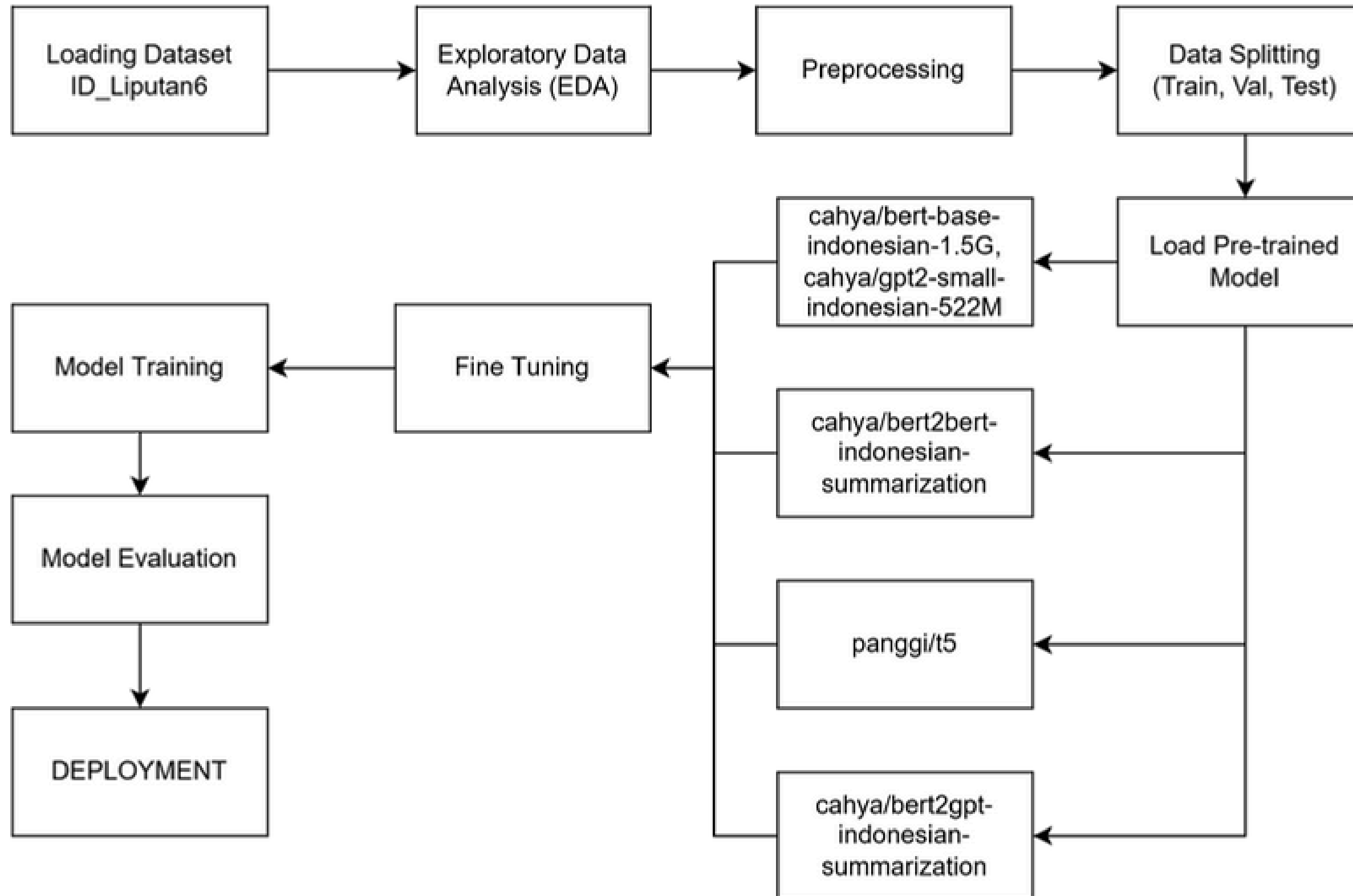
 panggi/t5-base-indonesian-summarization-cased

Summarization • Updated Jun 23, 2021 • 503 • 5

 cahya/bert2gpt-indonesian-summarization

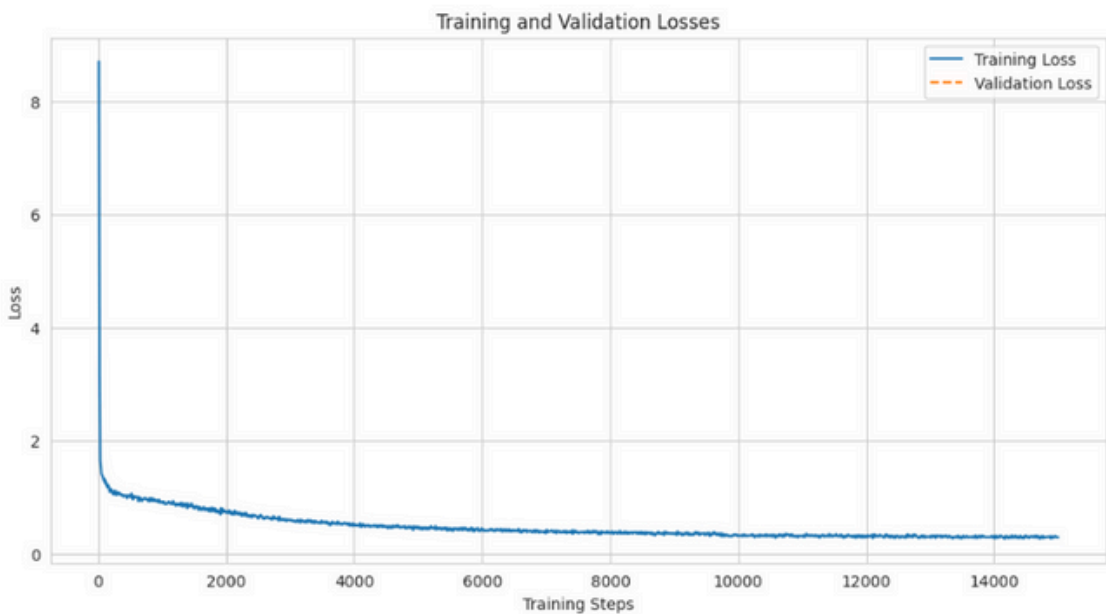
Summarization • Updated Feb 8, 2021 • 468 • 7

# Flow Training



# Results #Train

Model (Train using GPU)	Trainin Loss	Training Runtime
cahya/bert-base-indonesian-1.5G	0.21560	4 jam 10 menit



Model (Train using CPU)	Trainin Loss	Eval Loss	Training Runtime	Eval Runtime	Total FLOS
BERT2BERT-Cahya	0.97342	2.98323	14 jam 31 menit	2 jam 57 menit	5854044104245248
T5-Panggi	2.23624	2.83412	2 jam 46 menit	23 menit	1837523298539520
BERT2GPT-Cahya	2.01565	3.22243	14 jam 40 menit	6 jam 19 menit	2354744629094400

BERT2BERT-Cahya memiliki training loss terendah, yang menunjukkan bahwa model ini belajar paling baik selama pelatihan, tetapi memiliki waktu pelatihan yang lebih lama dan lebih berat dari sisi komputasi (FLOS tertinggi).

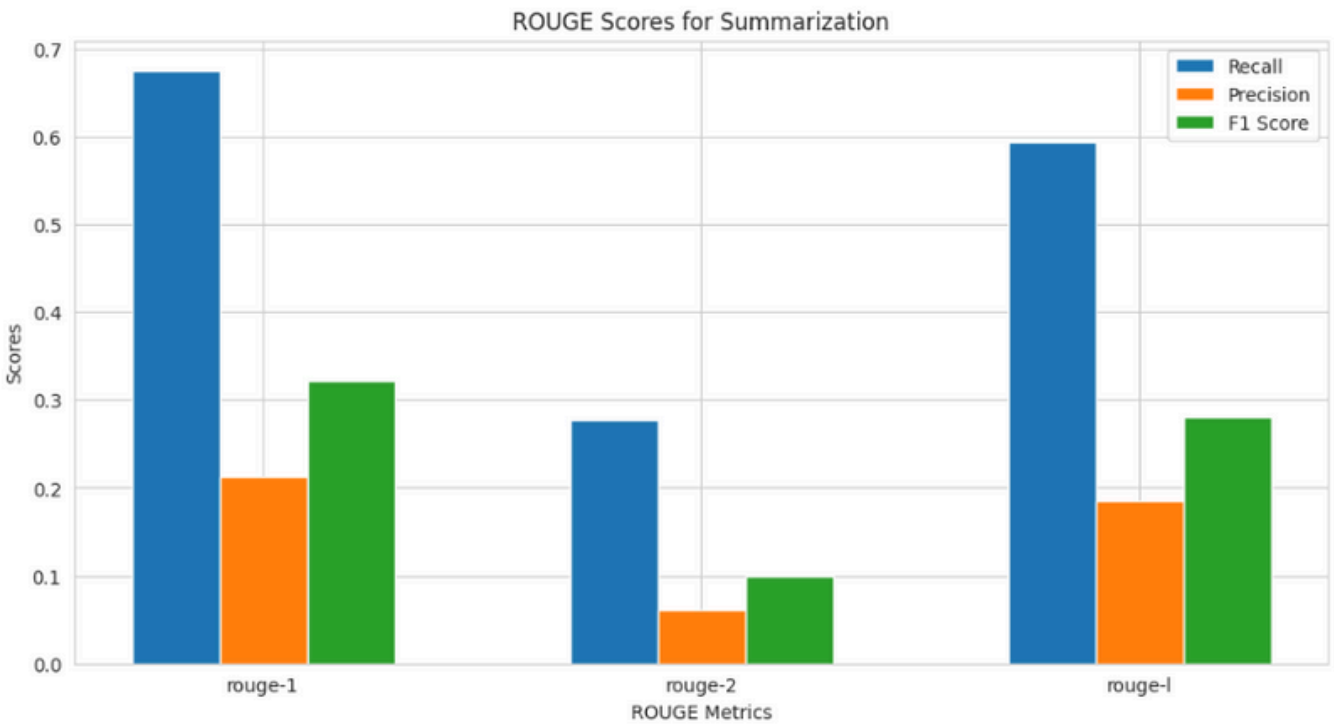
T5-Panggi memiliki eval loss terendah dan runtime paling efisien untuk pelatihan dan evaluasi, tetapi training loss-nya lebih tinggi, yang menunjukkan model ini mungkin tidak belajar dengan baik pada data pelatihan, meskipun tampil lebih baik pada data evaluasi.

BERT2GPT-Cahya memiliki waktu pelatihan yang cukup panjang, eval loss tertinggi, dan juga lebih berat dalam hal komputasi, tetapi masih bisa digunakan tergantung pada kriteria performa yang diinginkan.



# Results #Rouge

Model (GPU)	Rouge1	Rouge2	RougeL
cahya/bert-base-indonesian-1.5G	0.6754	0.2774	0.5936



Model (CPU)	Rouge1	Rouge2	RougeL	RougeLSum
BERT2BERT-Cahya	0.4681	0.3134	0.428	0.4681
T5-Panggi	0.2616	0.1231	0.2255	0.2616
BERT2GPT-Cahya	0.4822	0.333	0.4441	0.482

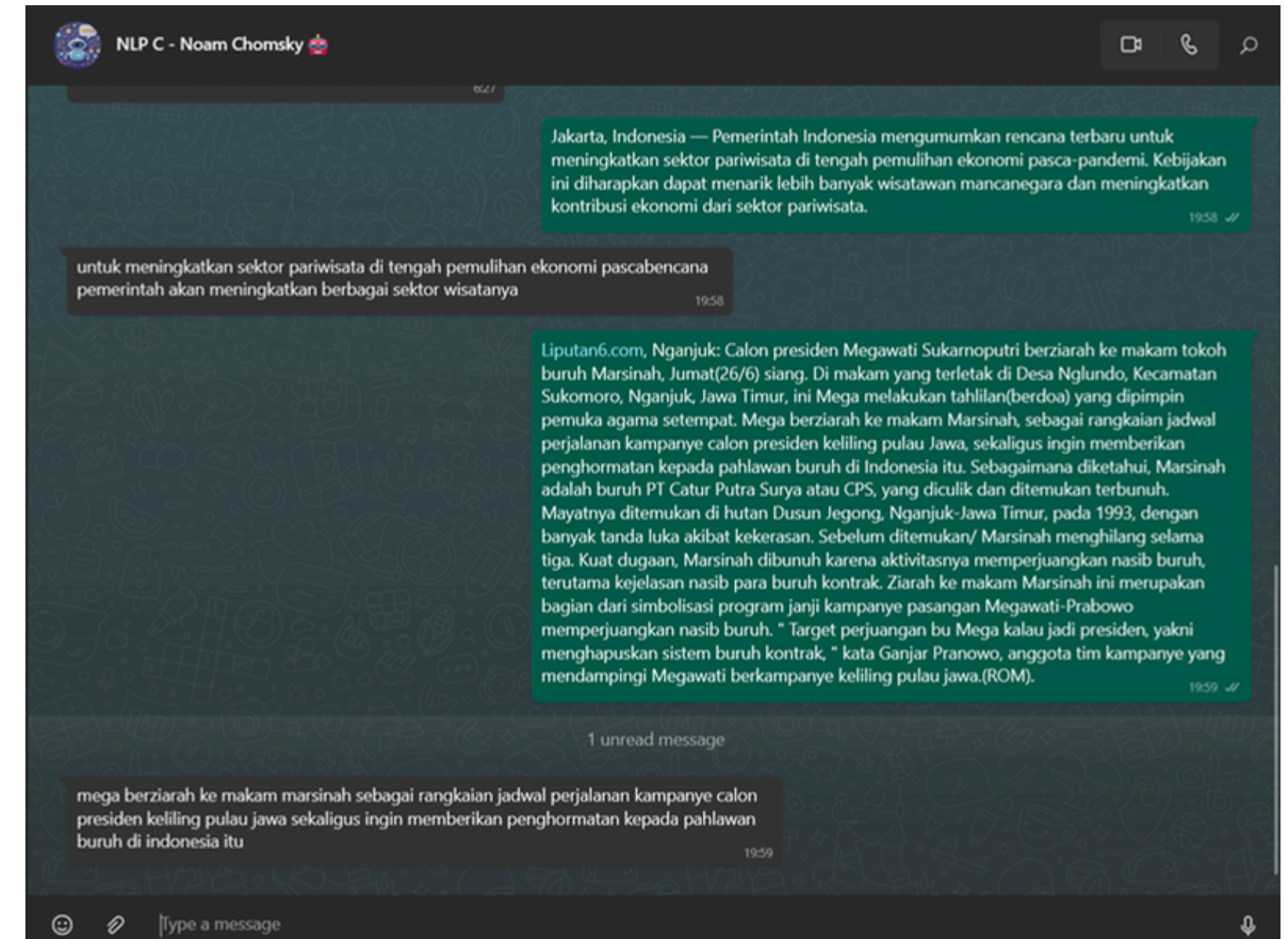
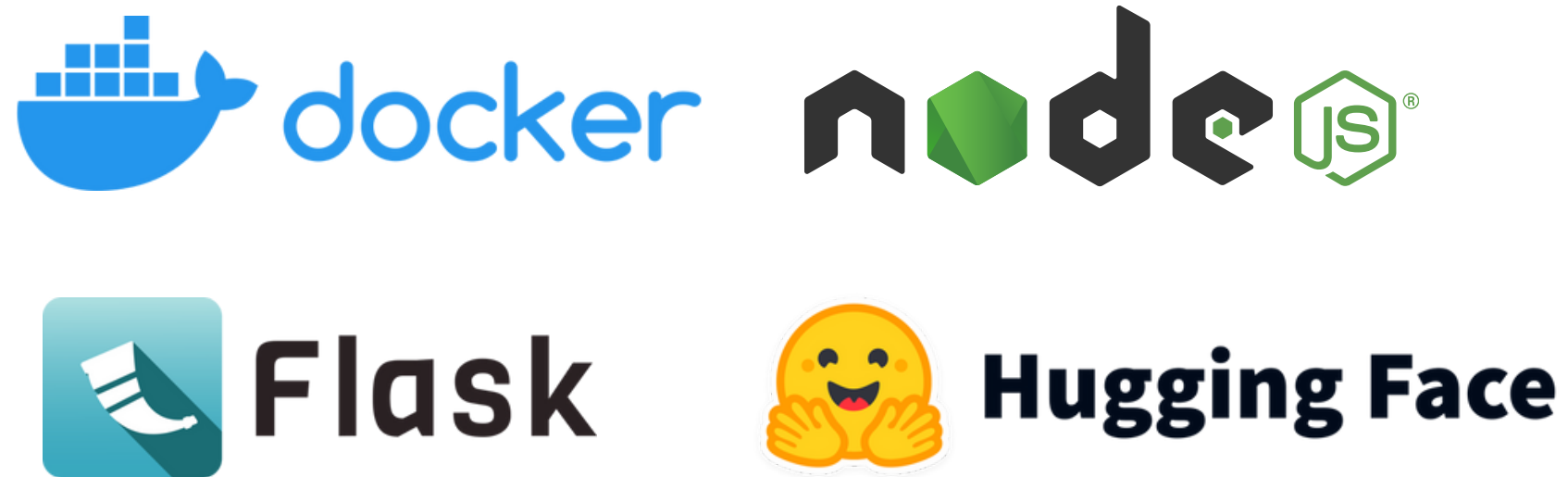
- BERT2GPT-Cahya memiliki skor terbaik di semua metrik (ROUGE-1, ROUGE-2, ROUGE-L, dan ROUGE-LSum), yang berarti model ini menghasilkan ringkasan yang lebih baik dibandingkan model lainnya dalam menangkap kesamaan unigram, bigram, dan urutan kata dengan ringkasan referensi.
- BERT2BERT-Cahya berada di urutan kedua dengan skor yang sedikit lebih rendah dari BERT2GPT-Cahya, tetapi masih jauh lebih baik daripada model T5-Panggi.
- T5-Panggi menunjukkan hasil yang jauh lebih rendah di semua metrik, yang berarti model ini kurang efektif dalam menghasilkan ringkasan yang akurat dibandingkan dua model lainnya.

# Results #Sample

ID Article	Reference Clean Summary	Generate Clean Summary (from Model)
27049	tiga anggota keluarga di jalan lumba lumba kendari sultra ditemukan tewas dengan luka di kepala polisi menyita sebuah parang dengan bekas darah dan beberapa benda mencurigakan lain	polresta kendari sultra menemukan luka di kepala nadir abola dan herlina sang istri serta seorang anak berusia enam tahun diduga tewas sejak beberapa hari silam korban terakhir kali melihat korban saat meminta kembali pacul yang dipinjam pada Selasa sore pekan silam
53576	fosil tulangbelulang gajah purba ditemukan di blora jateng para ahli penggalian fosil memperkirakan binatang ini hidup di zaman prasejarah	fosil tulang gajah purba yang masuk dalam spesies elephant hisudin atau mirip gajah sumatra ditemukan peneliti mengatakan binatang ini hidup di zaman prasejarah
62186	puluhan toko di pusat kota tondano minahasa sulut terbakar tidak ada korban jiwa dalam peristiwa itu namun kerugian materi diperkirakan mencapai miliaran rupiah polisi masih menyelidiki penyebab kebakaran	puluhan toko di pusat kota tondano minahasa sulawesi utara terbakar tak ada korban jiwa dalam peristiwa ini kerugian materi diperkirakan mencapai miliaran rupiah

# Deployment

- Aplikasi berupa chatbot whatsapp,
  - +62 877-8761-1391



<https://github.com/rowjak/text-summarization-liputan6>

 [rowjak/bert-indonesian-news-summarization](https://github.com/rowjak/bert-indonesian-news-summarization)

 Summarization • Updated 1 day ago • ↓ 9

 [rowjak/bert2gpt-indonesian-news-summarization](https://github.com/rowjak/bert2gpt-indonesian-news-summarization)

Updated 1 day ago • ↓ 2

# Conclusion

- Dikarenakan dataset yang besar dan komplek, maka membutuhkan resource yang besar untuk melakukan training
- untuk saat ini training hanya dilakukan dengan 20ribu data, namun hasilnya sudah cukup baik
- perbandingan training menggunakan CPU dan GPU sangat jauh. GPU dengan 20ribu data hanya dalam waktu 4 jam, sementara CPU dengan 5ribu data dalam waktu 14 jam
- Finetuning dengan dataset yang serupa menghasilkan hasil yang lebih baik daripada finetuning dengan dataset yang berbeda. Contoh menggunakan model yang di cahya. itu menggunakan dataset liputan6, sementara model T5-Panggi menggunakan dataset indosum
- Melakukan freeze encoder jika dataset berbeda akan mengakibatkan kurangnya adaptasi model dengan data baru jika data berbeda, dan model juga cenderung akan lebih bergantung pada model sebelumnya

# Future Improvement

- Dapat menggunakan seluruh dataset yang ada. setelah menggunakan dataset canon dan mendapatkan hasil yang baik, menguji kembali dengan dataset Xtreme
- Menggunakan model text-summarization lain yang up to date. karena model yang digunakan dalam pelatihan kali ini rata-rata rilis di tahun 2021
- Melakukan error analysis



**Thank you, any question ?**

**Semangat!!!**