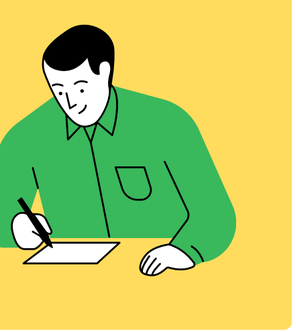
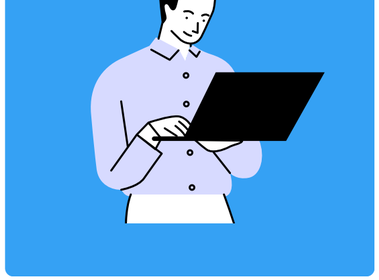
Text Summarization with BERT

NLP PROJECT BY TEAM A

Indonesia AI
AI for Everyone, AI for Indonesia

Spärck





Meet the Team

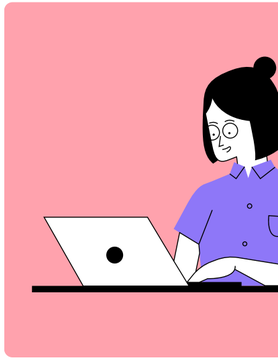
Alfandy Surya

Muhammad Habibullah

Efrad Galio

Anton Pranowo Medianto

Alfian Ali Murtadho



Spärck

Background & Problem Statement

Jumlah informasi yang tersedia melalui internet terus meningkat pesat di era digitalisasi saat ini. Pengguna sering kali mengalami kesulitan dalam menemukan dan mengkonsumsi informasi yang relevan bagi mereka. Portal berita online seperti **Liputan 6** menjadi salah satu sumber utama informasi dengan menyajikan berita-berita terkini dalam berbagai bidang.

Dengan banyaknya berita yang dipublikasikan setiap hari di portal berita tersebut, pengguna sering kali tidak memiliki waktu atau kesabaran untuk membaca seluruh artikel. Oleh karena itu, pada project kali ini akan dirancang sistem yang dapat merangkum konten berita menjadi ringkasan yang mencakup inti dari informasi yang disampaikan yang biasa disebut **text summarization**.

Spärck

Objectives & Scope

Objectives:

Melakukan eksperimen dengan membandingkan model-model pretrained BERT untuk text summarization menggunakan dataset Liputan 6.

Scope:

- Menggunakan data Liputan 6
- Menggunakan BERT language model
- Menggunakan V100 GPU/16 GB dan RAM 12.7 GB

Data Information

Dataset used: id_liputan6

Train & Val: canonical

Test: xtreme

Check null values:

```
Int64Index: 193883 entries, 0 to 193882
Data columns (total 5 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   id                 193883 non-null  int64
 1   url                193883 non-null  object
 2   clean_article      193883 non-null  object
 3   clean_summary      193883 non-null  object
 4   extractive_summary 193883 non-null  object
dtypes: int64(1), object(4)
memory usage: 8.9+ MB
```

Number of null rows: 0

Data Fields:

- **id:** id of the sample
- **url:** the url to the original article
- **clean_article:** the original article
- **clean_summary:** the abstractive summarization
- **extractive_summary:** index extractive summary

Language:

Indonesian

Duplicated Rows:

Number of duplicated rows: 0

Text Preprocessing

1. Convert number format
2. Remove parentheses
3. Remove specific word
4. Remove punctuation
5. Remove extra space
6. Case folding (lowercase)
7. Remove stopwords (EDA)
8. Remove single characters

words_to_remove = ["Liputan6.com', 'Liputan6', 'Liputan', Jakarta' , 'Surabaya', 'Bandung', 'Semarang', 'Medan', 'Makassar', 'Yogyakarta', 'Denpasar', 'Tangerang', 'Bogor']

Spärck

Article Example (Before & After)

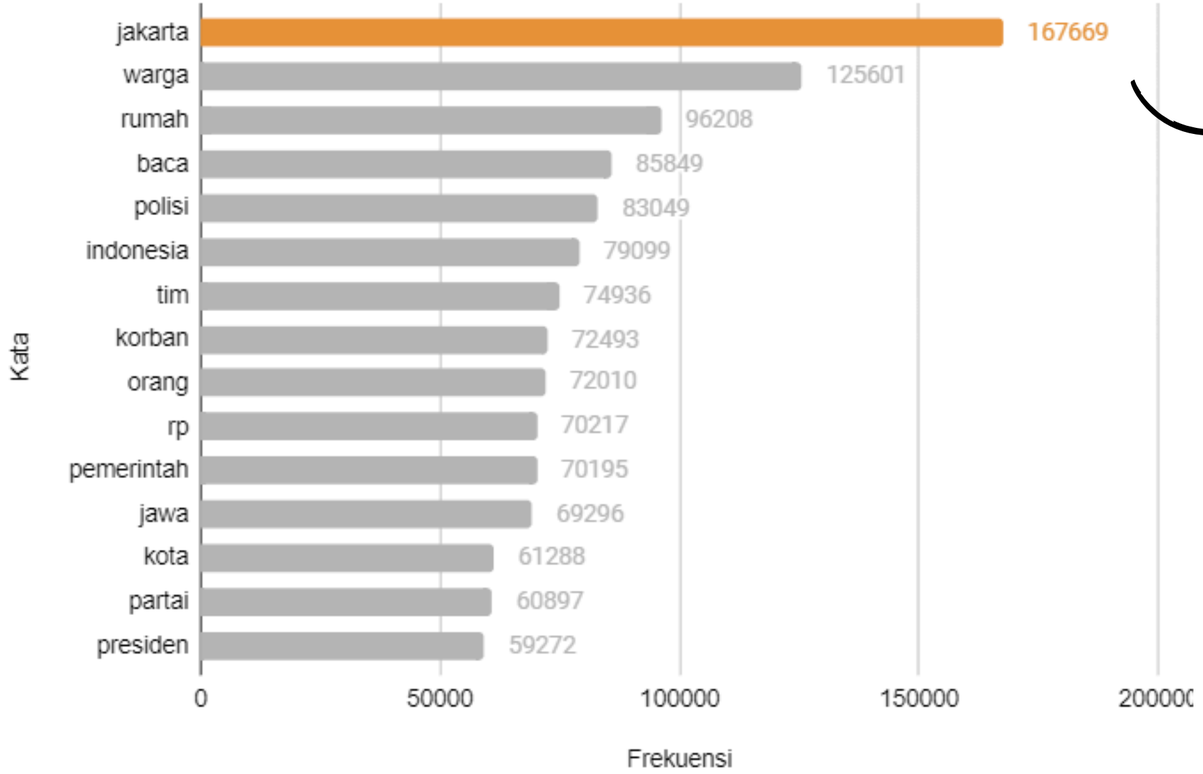
Text Preprocessing Stage	OUTPUT
CLEAN_ARTICLE	Liputan6.com, Jakarta: Kepolisian Daerah Riau bertekad memberantas pelaku penyelundupan kayu yang kerap terjadi di Riau. Selain itu, Polda setempat juga akan memberangus menipulasi dana reboisasi dan iuran hasil hutan. Demikian ditegaskan Kepala Polda Riau Brigadir Jenderal Polisi Johny Yodjana, seusai dilantik menjadi Kapolda Riau oleh Kepala Polri Jenderal Polisi Suroyo Bimantoro, di Jakarta, baru-baru ini. Menurut Johny, pelaku tindak kriminal yang kerap menjarah kayu di Riau akan ditindak tegas. "Saya tak akan pandang bulu," janji Johny. Selain itu, ia bertekad menyelidiki dugaan manipulasi dana reboisasi dan iuran hasil hutan sebesar Rp 680 miliar yang dilakukan sebuah perusahaan kayu di Riau. Sementara itu, selain melantik Johny Yodyana, Kapolri juga melantik Inspektur Jenderal Polisi Firman Gani menjadi Kapolda Sulawesi Selatan dan Brigjen Pol. Eddy Darnadi menjadi Kapolda Maluku. Selain itu, Bimantoro juga melantik Komisaris Besar Pol. Totok Soenarjo menjadi Kapolda Jambi, Brigjen Pol. Sugiri menjadi Kapolda Lampung, dan Brigjen Pol. Dwi Purwanto menjadi Kapolda Bengkulu. (ICH/Edi Priyono dan Andi Azril).
CLEAN_ARTICLE_PREP	kepolisian daerah riau bertekad memberantas pelaku penyelundupan kayu yang kerap terjadi di riau selain itu polda setempat juga akan memberangus menipulasi dana reboisasi dan iuran hasil hutan demikian ditegaskan kepala polda riau brigadir jenderal polisi johny yodjana seusai dilantik menjadi kapolda riau oleh kepala polri jenderal polisi suroyo bimantoro di jakarta baru baru ini menurut johny pelaku tindak kriminal yang kerap menjarah kayu di riau akan ditindak tegas saya tak akan pandang bulu janji johny selain itu ia bertekad menyelidiki dugaan manipulasi dana reboisasi dan iuran hasil hutan sebesar rp 680 miliar yang dilakukan sebuah perusahaan kayu di riau sementara itu selain melantik johny yodyana kapolri juga melantik inspektur jenderal polisi firman gani menjadi kapolda sulawesi selatan dan brigjen pol eddy darnadi menjadi kapolda maluku selain itu bimantoro juga melantik komisaris besar pol totok soenarjo menjadi kapolda jambi brigjen pol sugiri menjadi kapolda lampung dan brigjen pol dwi purwanto menjadi kapolda bengkulu

Summary Example (Before & After)

Text Preprocessing Stage	OUTPUT
CLEAN_SUMMARY	Kapolda Riau baru Brigjen Pol. Johny Yodjana bertekad memberantas pelaku penyelundupan kayu di Riau. Ia berjanji akan menindak tegas pelaku tanpa pandang bulu.
CLEAN_SUMMARY_PREP	kapolda riau baru brigjen pol johny yodjana bertekad memberantas pelaku penyelundupan kayu di riau ia berjanji akan menindak tegas pelaku tanpa pandang bulu

Exploratory Data Analysis – 1

Most Used Word (no stopwords):



dapat dilihat **Jakarta** menjadi kata dengan frekuensi kemunculan paling tinggi. Namun, pada data train yang digunakan, lokasi berita seringkali terletak di awal berita, Jika kita mengambil kata pertama dari berita, dapat dilihat lokasi **Jakarta** menempati urutan pertama dengan frekuensi kemunculan di awal kalimat mencapai 62 ribu kali. Oleh karena itu first word yang mengandung kata -kata di samping akan diremove.

Top 10 News Location :

Lokasi	Frekuensi Kemunculan
Jakarta	62906
Surabaya	3442
Bandung	3044
Semarang	2977
Medan	2857
Makassar	2204
Yogyakarta	2123
Denpasar	2110
Tangerang	1948
Bogor	1926

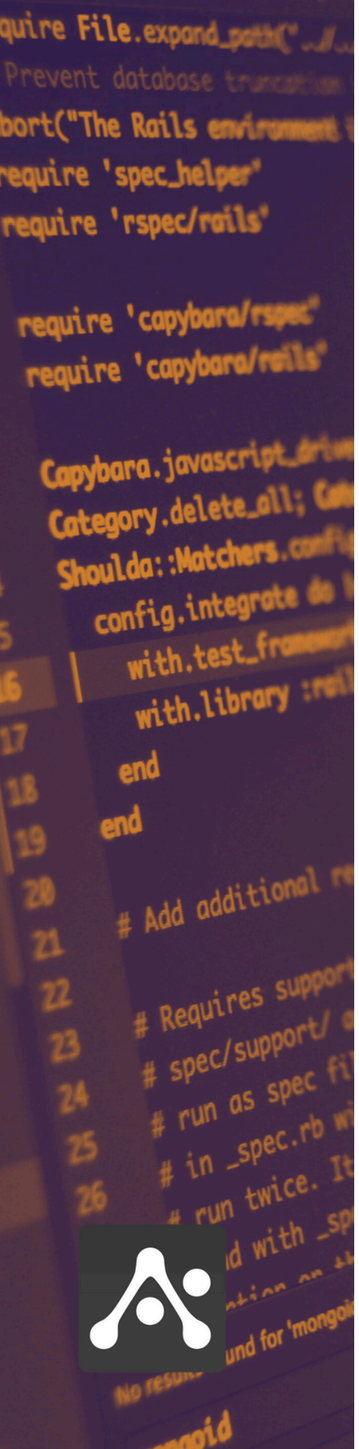
Exploratory Data Analysis - 2

Bi-gram Frequency

Bigram	Frekuensi
(rumah, sakit)	27146
(jawa, barat)	23482
(jawa, timur)	20374
(jakarta, pusat)	13174
(anak, anak)	12388
(jakarta, selatan)	11664
(amerika, serikat)	10858
(susilo, bambang)	9847
(bambang, yudhoyono)	9672
(kepolisian, resor)	9533

Tri-gram Frequency

Trigram	Frekuensi
(susilo, bambang, yudhoyono)	9644
(presiden, susilo, bambang)	6948
(nanggroe, aceh, darussalam)	5364
(dirawat, rumah, sakit)	5041
(bahan, bakar, minyak)	4713
(partai, demokrasi, indonesia)	4565
(demokrasi, indonesia, perjuangan)	4463
(bandung, jawa, barat)	4413
(surabaya, jawa, timur)	4192
(bogor, jawa, barat)	3463



Model: Xtreme Dataset

Data Splitting Strategy:

68,7% train	17,5% validation	13,7% test
193.883 rows	**4948 rows**	**3862 rows**
original data		

19.388 rows — 10% random sampled data

Menyesuaikan dengan Resource agar dapat di training dengan cepat.

Experimentation (Abstractive Model):

Pretrained Model	Fine-Tuning Pretrained: cahya/bert2bert
• cahya/bert2bert-indonesian-summarization • Alfahluzi/bert2bert-extreme	• Learning Rate (1e-5 & 5e-5) • Batch Size (2 & 4)

Spärck

Fine Tune Experiment
Train Result

Spärck

using cahya/bert2bert-indonesian-summarization models

Model	Tokenizer	Data Collator?	Freeze Encoder?	Train Runtime (s)	Train Loss	Validation Loss	Epoch	FLOPS
• Batch = 4 • LR = 5e-5	cahya/bert 2bert	Yes	Yes	2406	1,13	3,52	2,06	2,45E+28
• Batch = 2 • LR = 5e-5	cahya/bert 2bert	Yes	Yes	1697	1,46	3,48	1,03	1,23E+28
• Batch = 4 • LR = 1e-5	cahya/bert 2bert	Yes	Yes	2440	1,27	3,50	2,06	2,45E+28
• Batch = 2 • LR = 1e-5	cahya/bert 2bert	Yes	Yes	1785	1,41	3,49	1,03	1,23E+28

Dapat dilihat dari train summary di atas, batch = 4 membutuhkan waktu train dan kebutuhan resource (dilihat dari train runtime dan FLOPS) yang relatif lebih besar dibandingkan batch = 2. Learning rate disini relatif tidak memberikan pengaruh yang signifikan terhadap hasil train dari batch 2 dan cukup berpengaruh terhadap batch 4.

ROUGE Result

Spärck

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) dihitung menggunakan 50 random sample test untuk mempercepat proses komputasi.
Nilai di bawah ini adalah rata-rata rouge untuk 50 random sample test.

Model	Type	Rouge1	Rouge2	RougeL	RougeLSUM
cahya/bert2bert-indonesian-summarization	Pretrained	0.312	0.111	0.252	0.225
Alfahluzi/bert2bert-extreme	Pretrained	0.065	0.00	0.049	0.038
model_batch_4_lr_5e-5	Pretrained + Fine Tuned	0.333	0.120	0.251	0.235
model_batch_2_lr_5e-5	Pretrained + Fine Tuned	0.328	0.117	0.250	0.232
model_batch_4_lr_1e-5	Pretrained + Fine Tuned	0.320	0.117	0.245	0.227
model_batch_2_lr_1e-5	Pretrained + Fine Tuned	0.342	0.126	0.262	0.243

Model cahya/bert2bert yang di-finetune menggunakan batch 2 dan learning rate 5e-5 menghasilkan overall rouge yang lebih baik dibandingkan dengan model lainnya.

Inferencing

Spärck

Generated Summary	Reference Summary
menurut pengamat ekonomi didiek rachbini, bank indonesia masih akan menghadapi situasi sulit kendati bank sentral amerika serikat terus menurunkan tingkat suku bunga.	kendati bank sentral as menurunkan suku bunganya namun bi dinilai masih akan menemui masa sulit suku bunga bank sentral as akan diturunkan menjadi empat persen
penghapusan beberapa pasal menyangkut hak buruh dalam keputusan menakertrans no. 78 tahun 2001 bukan semata mata pembelaan kepada kelompok pengusaha revisi justru untuk menyelesaikan masalah pengangguran.	revisi kepmennaker nomor 78 tahun 2001 dinilai bukan semata mata untuk membela kepentingan para pengusaha para buruh diminta jangan mementingkan diri sendiri
operasi sadar jaya yang dilancarkan selasa malam mengejutkan pengunjung diskotik millenium yang berlokasi di jalan gajah mada, jakpus. sebanyak 200 petugas gabungan dari polres metro jakpus dan kesatuan brigade mobil polda metro jaya menggeledah seluruh pengunjuk rasa.	polisi menangkap 32 pengunjung diskotik milenium karena tertangkap basah membawa pil ekstasi penggerebekan ini adalah bagian operasi sadar jaya yang bertujuan memberantas peredaran narkoba di ibu kota

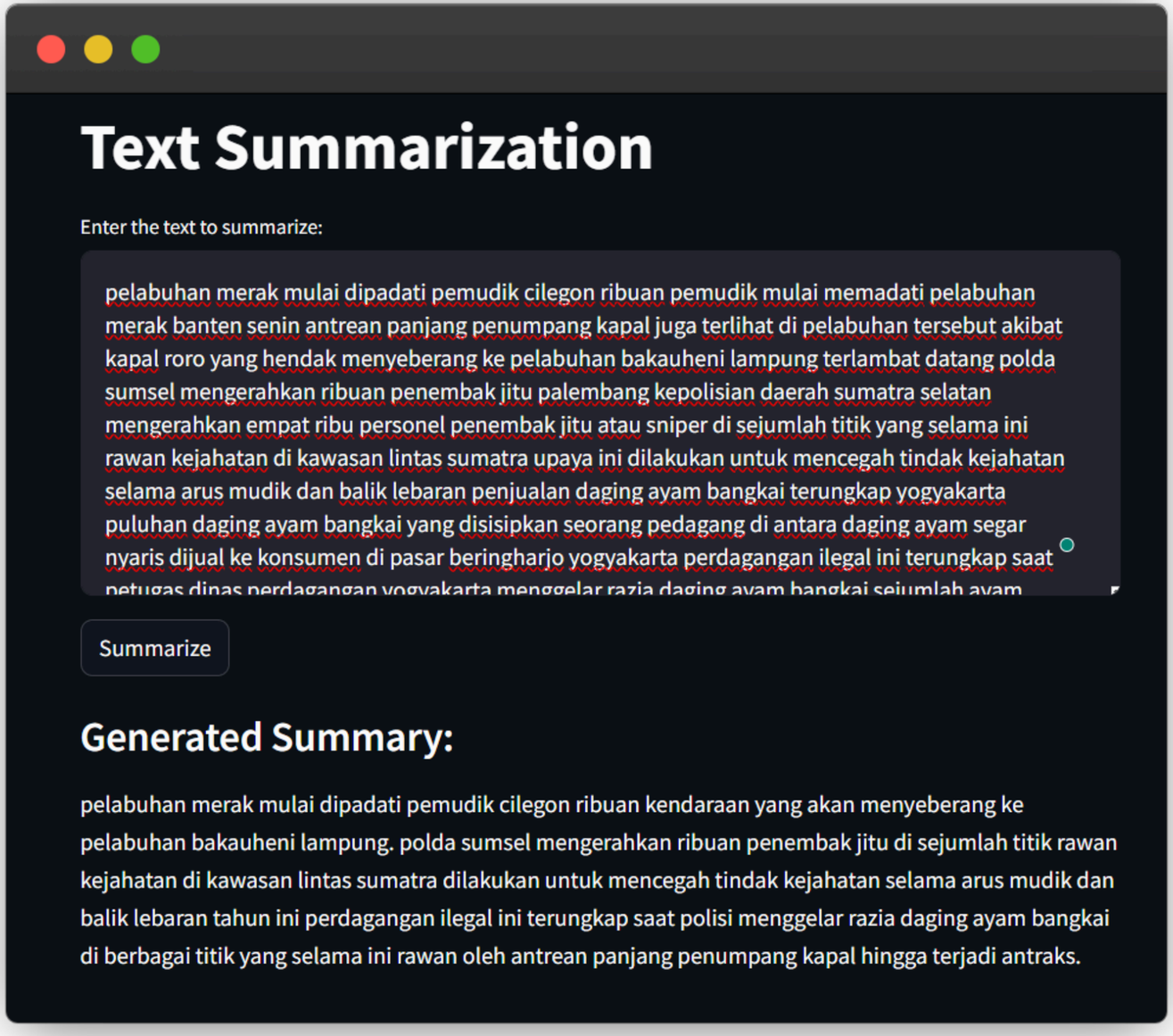
App Deployment

Spärck

Proses deployment masih berada pada tahap **local deployment**. Berikut ini adalah screenshot dari prototype aplikasi

Back-end: Flask

Front-end: Streamlit



Summary

- Model **(model_batch_2_lr_1e-5)** pretrained **cahya/bert2bert-indonesian-summarization** yang di-*finetune* menggunakan 10% data id_liputan6, batch = 2, dan learning rate = 1e-5 menghasilkan performa ROUGE paling baik dibandingkan model lainnya.
- Selain dari performa prediksi, performa proses training dari model ini juga sangat baik karena training duration hanya membutuhkan waktu 1697 seconds dengan kebutuhan resource yang relatif rendah.

Future Improvements

- Dikarenakan dataset yang besar dan model yang cukup kompleks, maka dibutuhkan resource yang besar untuk projek selanjutnya.
- Project ini menggunakan 10% dari total train dataset, maka projek selanjutnya menggunakan seluruh dataset untuk train fine tuning dengan resource yang mendukung.
- Melakukan error analysis.
- Menggunakan model text-summarization yang *up-to-date* (SOTA)

Thank you, any question?