Tugas-A-1-Eksplorasi Analisis Sentimen - [Tugas Individu]

Nama : Alfa Renaldo Aluska
NRP : 5026221144

Kode dan hasil *output* bisa dilihat di:
https://github.com/renaldoaluska/pba2025gasal/blob/main/%23TugasA-1/Tugas_A_1_5026221144.ipynb

***Dataset* yang digunakan:**
https://github.com/renaldoaluska/pba2025gasal/blob/main/%23TugasA-1/7-garudaindonesia_news_cleaned_simple.csv

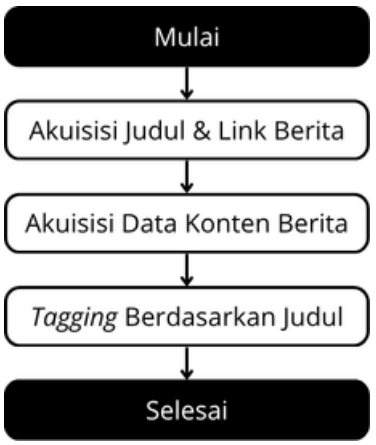
1. Persiapan *Dataset*

Menyiapkan *dataset* untuk tugas analisis sentimen adalah langkah krusial dalam proyek NLP, karena kualitas *dataset* sangat memengaruhi performa model. Proses ini mencakup pengumpulan, pembersihan, pelabelan, dan format data sehingga dapat digunakan untuk pelatihan dan evaluasi model.

a. Jelaskan upaya yang telah anda lakukan dalam menyiapkan *dataset* untuk tugas analisis sentimen.

Jawaban:

Tahap akuisisi data merupakan langkah awal dalam penelitian ini yang bertujuan untuk menghimpun, menyiapkan, dan menyusun data berita agar dapat diproses dan dianalisis secara lebih mendalam. Prosedur ini dilaksanakan secara sistematis melalui serangkaian tahapan berurutan guna memastikan bahwa data yang dikumpulkan bersifat relevan, terstruktur, dan siap untuk tahap pra-pemrosesan serta analisis. Secara garis besar, proses akuisisi dimulai dengan pengambilan judul dan tautan berita dari situs resmi daring, yang kemudian diikuti oleh ekstraksi isi berita berdasarkan tautan tersebut. Setelah seluruh konten berhasil dikumpulkan, dilakukan proses penandaan awal (*tagging*) untuk memberikan label kategori pada masing-masing berita, sesuai dengan kerangka penelitian yang telah ditentukan. Rangkaian lengkap tahapan akuisisi data ditampilkan dalam diagram berikut.



Gambar Diagram Alir Tahap Scraping Berita

a.1 Akuisisi Judul dan Tautan Berita

Pada tahap ini, dilakukan proses pengambilan data berita dari Google News secara otomatis menggunakan bahasa pemrograman Python dan pustaka Selenium. Langkah awal melibatkan instalasi sejumlah dependensi penting seperti wget, curl, unzip, serta pustaka Python seperti Selenium, Chromedriver-Autoinstaller, dan Dateparser. Untuk mendukung otomasi peramban dalam mode tanpa tampilan grafis (headless), juga diperlukan instalasi Google Chrome dan Chromedriver. Pengambilan artikel dibatasi hingga tanggal 30 September 2025, dengan total 1194 artikel berhasil dikumpulkan. *Dataset* yang telah diperoleh kemudian menjalani tahap pra-pemrosesan dan disiapkan untuk analisis, sebagaimana dijelaskan pada tabel berikut.

Tabel Jumlah artikel dalam *dataset* per tahapan

No.		Tahapan	Dihapus	Jumlah Akhir
1	Hanya judul dan tautan berita	Akuisisi judul dan tautan berita	-0	1194
2		Hapus baris *scraping* judul dan tautan berita kosong karena gagal di-*scrap*	-191	1003
3		Filter bahasa Inggris judul berita dengan formula Google Speadsheets, lalu dihapus dengan Python	-389	614
4	Beserta dengan isi konten berita	*Scraping* isi konten berdasarkan tautan yang telah di-*scrap* sebelumnya,	-0	614
5		Penghapusan baris hasil *scraping* isi konten yang kosong	-81	533
6		Penghapusan baris hasil *scraping* isi konten berbahasa Inggris dengan fungsi detect_language()	-38	495
7		Filter dan hapus konten berbahasa Inggris manual melalui Microsoft Excel	-26	469

Setelah seluruh dependensi terpasang, dilakukan konfigurasi opsi Chrome agar dapat dijalankan secara aman di lingkungan kerja. Selanjutnya, dibuat struktur data berupa *set* untuk menyimpan hasil scraping agar tidak terjadi duplikasi tautan.

Fungsi utama yang digunakan adalah scrape_google_news_link(). Fungsi ini bertugas untuk mengakses hasil pencarian berita di Google News berdasarkan kata kunci tertentu, yaitu “Garuda Indonesia” dan “Pesawat Garuda”. Melalui Selenium, *browser* otomatis membuka halaman-halaman hasil pencarian, kemudian mengambil sejumlah artikel per batch (setiap 10 artikel), dan mengekstrak beberapa komponen penting dari masing-masing berita, meliputi: tautan berita (*link*), judul artikel, tanggal publikasi, dan nama portal berita. Data yang telah dikumpulkan disimpan dalam variabel all_articles yang berisi kumpulan data berita unik.

```
Mengakses URL: https://www.google.com/search?q=garuda+indonesia
Halaman artikel sudah tidak tersedia
Proses selesai. Total link artikel unik yang berhasil diambil: 1194
------------------------------------------------------------
Mengakses URL: https://www.google.com/search?q=pesawat+garuda
Halaman artikel sudah tidak tersedia
Proses selesai. Total link artikel unik yang berhasil diambil: 1194
------------------------------------------------------------
```

Gambar Hasil proses *scraping*

Setelah proses pengumpulan selesai, sebagian data ditampilkan ke layar sebagai contoh untuk memastikan hasil *scraping* berjalan dengan baik. Bagian ini memperlihatkan beberapa entri awal berupa tautan, judul, tanggal, dan portal berita.

```
('https://www.travelandtourworld.com/news/article/garuda-indonesia-now-expands-flight-network-to-bali-yogyakarta-surabaya-and
-other-main-tourism-hub-by-2029-heres-what-you-need-to-know/', 'Garuda Indonesia Now Expands Flight Network to Bali, Yogyakar
ta, Surabaya, and Other Main Tourism Hub by 2029: Here's What You Need To Know', '2025-07-31', 'Travel And Tour World')
('https://www.ft.com/content/45ff1892-b4fe-4caf-9d07-2e6a0a189fed', 'Islamic bonds come under microscope after Garuda Indones
ia default', '2021-08-17', 'Financial Times')
('https://ulasan.co/beredar-video-ban-pesawat-garuda-menggelinding-di-landasan-pacu-tanjungpinang/', 'Beredar Video Ban Pesaw
at Garuda Menggelinding di Landasan Pacu Tanjungpinang', '2025-04-16', 'Ulasan.co')
('https://www.cnnindonesia.com/ekonomi/20240824164025-92-1137068/135-penumpang-pesawat-tangki-bocor-garuda-sudah-diterbangkan
-kembali', '135 Penumpang Pesawat Tangki Bocor Garuda Sudah Diterbangkan Kembali', '2024-08-24', 'CNN Indonesia')
('https://tirto.id/cara-pilih-kursi-garuda-terbaru-dan-besaran-tarifnya-g65a', 'Cara Pilih Kursi Garuda Terbaru dan Besaran T
arifnya', '2024-12-31', 'Tirto.id')
```

Gambar Entri awal hasil *scraping*

Tahap berikutnya adalah mengonversi data hasil scraping ke dalam bentuk DataFrame menggunakan pustaka Pandas. DataFrame ini memiliki empat kolom utama yaitu *link*, *judul*, *tanggal*, dan *portal*. Kolom tanggal kemudian dikonversi ke format DateTime dan diurutkan dari tanggal terbaru ke tanggal terlama agar memudahkan proses analisis di tahap berikutnya.

	link	judul	tanggal	portal
0	https://kumparan.com/kumparanbisnis/garuda-ind...	Garuda Indonesia Kembali RUPSLB di Tengah Isu ...	2025-09-30	Kumparan
1	https://voi.id/en/economy/519004	Commission V DPR Will Investigate Allegations ...	2025-09-29	VOI.ID
2	https://www.prnewswire.com/news-releases/garud...	Garuda Indonesia Goes Digital: Air Cargo Capac...	2025-09-29	PR Newswire
3	https://in.investing.com/news/company-news/gar...	Garuda Indonesia adds air cargo capacity to We...	2025-09-29	Investing.com India
4	https://jambi.pikiran-rakyat.com/info-data/pr-...	Jadwal Kedatangan Pesawat di Bandara Sultan Th...	2025-09-29	Jambian
...	...	...	...	...
1189	https://www.thejakartapost.com/indonesia/2024/...	Garuda Indonesia flight makes emergency landin...	NaT	The Jakarta Post
1190	https://finance.detik.com/bursa-dan-valas/d-81...	Gelar RUPSLB Lagi, Anak Usaha Garuda Tunjuk Di...	NaT	detikFinance
1191	https://djsaviation.net/garuda-indonesia-aircr...	Garuda Indonesia Aircraft Are Grounded	NaT	Dj's Aviation
1192	https://www.aviacionline.com/qatar-airways-and...	Qatar Airways and Garuda Indonesia expand part...	NaT	Aviacionline
1193	https://www.flightglobal.com/airlines/garuda-s...	Garuda secures key payment from Indonesian gov...	NaT	FlightGlobal

1194 rows × 4 columns

Gambar *Preview dataset* hasil *scraping*

Hasil akhir dari proses ini disimpan ke dalam dua format file, yaitu CSV dan Excel, dengan nama *link_berita_garudaindonesia.csv* dan *link_berita_garudaindonesia.xlsx*. File

tersebut berisi kumpulan tautan berita lengkap dengan judul, tanggal publikasi, serta nama portal berita yang menjadi hasil akhir tahap akuisisi.

Setelah *scraping* menggunakan Python, didapatkan data sebanyak 1004 baris, dengan 1003 artikel dan 1 *header*. Akan tetapi, pada data tersebut, terdapat beberapa judul berita dalam bahasa Inggris. Berita dalam bahasa Inggris kemudian di-*detect* melalui Google Spreadsheets dan dilabeli dengan kode 'EN'. Setelah itu, dilakukan *filtering* untuk menghapus semua baris berita berbahasa Inggris.

```
=IF(AND(LEN(TRIM(A2))>=10, DETECTLANGUAGE(A2)="en"), "EN","KEEP")
```

Gambar Formula deteksi bahasa Inggris

a.2 *Tagging* Berdasarkan Judul

Penentuan *tag* dilakukan berdasarkan kecocokan ekspresi reguler (*regex*) dari judul berita (setelah diubah menjadi huruf kecil) dengan daftar kata kunci yang telah ditentukan menggunakan fungsi =ARRAYFORMULA() di Google Spreadsheets.

```
=ARRAYFORMULA(
IF(A2:A="","",
  IFS(
    REGEXMATCH(LOWER(A2:A),"(laba|rugi|pendapatan|utang|restrukturisasi|pkpu|obligasi|kuartal|laporan keuangan)"),"Keuangan",
    REGEXMATCH(LOWER(A2:A),"(rute|penerbangan baru|buka rute|tutup rute|frekuensi|penerbangan|operasional|insiden|delay|batal|keselamatan)"),"Rute/Operasional",
    REGEXMATCH(LOWER(A2:A),"(direktur|komisaris|manajemen|rups|ceo|pergantian|pengurus)"),"Manajemen",
    REGEXMATCH(LOWER(A2:A),"(kasus|pengadilan|kpk|dugaan|suap|hukum|sidang|sanksi|denda|izin)"),"Hukum/Regulasi",
    TRUE,"Lainnya"
  )
))
```

Gambar Formula untuk *tagging* berdasarkan judul

Keterangan:

- Tag "Keuangan" diberikan jika teks mengandung salah satu kata berikut: *laba*, *rugi*, *pendapatan*, *utang*, *restrukturisasi*, *pkpu*, *obligasi*, *kuartal*, atau *laporan keuangan*.
- Tag "Rute/Operasional" diberikan jika teks mengandung salah satu kata berikut: *rute*, *penerbangan baru*, *buka rute*, *tutup rute*, *frekuensi penerbangan*, *operasional*, *insiden*, *delay*, *batal*, atau *keselamatan*.
- Tag "Manajemen" diberikan jika teks mengandung salah satu kata berikut: *direktur*, *komisaris*, *manajemen*, *rups*, *ceo*, *pergantian*, atau *pengurus*.
- Tag "Hukum/Regulasi" diberikan jika teks mengandung salah satu kata berikut: *kasus*, *pengadilan*, *kpk*, *dugaan*, *suap*, *hukum*, *sidang*, *sanksi*, *denda*, atau *izin*.
- Tag "Lainnya" diberikan jika teks tidak memenuhi kriteria kata kunci dari salah satu kategori di atas.

a.3 Akuisisi Isi Konten Berita

Tahap ini bertujuan untuk mengekstrak isi berita dari daftar tautan yang telah diperoleh pada tahap sebelumnya. Data awal berupa berkas data_link_berita.csv sudah berisi kolom judul dan tautan berita. Pada tahap ini, sistem hanya berfokus untuk mengunduh serta mengekstrak teks utama dari setiap tautan tersebut.

Proses diawali dengan instalasi beberapa pustaka pendukung seperti Pandas, Tqdm, Requests, Newspaper3k, Trafilatura, dan Readability-LXML. Pustaka-pustaka ini digunakan untuk membaca data, melakukan permintaan ke situs berita, serta mengekstrak isi utama halaman web secara otomatis.

Setiap tautan dalam *dataset* diproses satu per satu menggunakan beberapa metode ekstraksi yang disusun secara berlapis. Metode pertama adalah Newspaper3k, yang berusaha mengambil judul dan isi artikel secara langsung. Jika hasilnya terlalu pendek atau gagal, sistem beralih ke Trafilatura, yang mengekstrak teks berdasarkan struktur HTML halaman. Bila kedua metode tersebut masih belum menghasilkan teks yang memadai, digunakan Readability-LXML untuk menyaring bagian teks yang paling relevan dari halaman tersebut.

Selama proses pengambilan konten, sistem menggunakan mekanisme *timeout*, *retry* otomatis, serta jeda acak antar-*request* untuk mencegah kegagalan akibat pembatasan akses (error 429). Setiap teks hasil ekstraksi juga dibersihkan dari spasi berlebih dan disusun ulang agar tampil lebih rapi.

Setelah seluruh tautan berhasil diproses, hasil ekstraksi berupa konten berita ditempatkan bersebelahan dengan kolom judul di dalam dataframe. Data akhir kemudian disimpan kembali dalam berkas keluaran berformat .xlsx, dengan nama yang sama seperti berkas input namun ditambahkan akhiran _with_content.xlsx. Hasil akhir tahap ini adalah dataset yang telah berisi teks lengkap dari masing-masing berita, siap digunakan pada tahap pra-pemrosesan teks berikutnya.

```
Mengambil konten artikel: 100%|████████| 614/614 [42:42<00:00,  4.17s/it]
Selesai! Tersimpan: data_link_berita_with_content.csv
```

Gambar Hasil scraping isi konten berita

Meskipun sudah dilakukan tahapan *filtering* judul artikel berbahasa Inggris, beberapa artikel berbahasa Inggris tetap masuk dalam *dataset*. Oleh karena itu, dilakukan tahapan untuk membersihkan dan memfilter data hasil *scraping* sebelumnya agar hanya tersisa data yang valid dan relevan untuk proses analisis berikutnya. Proses diawali dengan membaca *file* hasil akuisisi, yaitu *data_link_berita_with_content.csv*, menggunakan pustaka Pandas. File tersebut berisi data mentah berupa tautan, judul, tanggal, portal, serta konten berita yang telah berhasil dikumpulkan dari Google News.

Apabila *file* berhasil dimuat, program akan menampilkan beberapa baris pertama untuk memastikan struktur data telah sesuai. Jika file tidak ditemukan, sistem akan memberikan pesan kesalahan agar pengguna dapat memperbaiki jalur file yang digunakan.

```
Successfully loaded data from data_link_berita_with_content.csv
```

	link	judul	konten	tanggal	portal	tag
0	https://kumparan.com/kumparanbisnis/garuda-ind...	Garuda Indonesia Kembali RUPSLB di Tengah Isu ...	Garuda Indonesia Kembali RUPSLB di Tengah Isu ...	2025-09-30 00:00:00	Kumparan	Manajemen
1	https://www.bloombergtechnoz.com/detail-news/8...	Garuda Gelar RUPSLB di Tengah Isu Masuknya Dir...	Garuda Gelar RUPSLB di Tengah Isu Masuknya Dir...	2025-09-29 00:00:00	Bloomberg Technoz	Manajemen
2	https://voi.id/ekonomi/519004/komisi-v-dpr-bak...	Komisi V DPR Bakal Dalami Dugaan Mafia Jual Be...	JAKARTA - Ketua Komisi V DPR Lasarus mengataka...	2025-09-29 00:00:00	VOI.ID	Rute/Operasional
3	https://in.investing.com/news/company-news/gar...	Garuda Indonesia adds air cargo capacity to We...	NaN	2025-09-29 00:00:00	Investing.com India	Lainnya
4	https://www.kompasiana.com/zainularifin2714/68...	Rencana Merger Garuda Indonesia - Pelita Air: ...	Latar Belakang\nPada pertengahan 2023, wacana ...	2025-09-29 00:00:00	Kompasiana.com	Lainnya

Gambar *Preview* hasil *scraping* mentah

Langkah pertama dalam pembersihan data adalah menghapus baris yang tidak memiliki nilai pada kolom “konten”. Hal ini dilakukan menggunakan fungsi dropna() agar hanya berita yang memiliki isi lengkap yang tersisa dalam dataset. Berdasarkan hasil eksekusi, jumlah data awal sebanyak 614 baris berkurang menjadi 533 baris setelah pembersihan, yang berarti terdapat 81 data tanpa isi konten yang dihapus.

```
Original DataFrame shape: (614, 6)
Cleaned DataFrame shape: (533, 6)
```

	link	judul	konten	tanggal	portal	tag
0	https://kumparan.com/kumparanbisnis/garuda-ind...	Garuda Indonesia Kembali RUPSLB di Tengah Isu ...	Garuda Indonesia Kembali RUPSLB di Tengah Isu ...	2025-09-30 00:00:00	Kumparan	Manajemen
1	https://www.bloombergtechnoz.com/detail-news/8...	Garuda Gelar RUPSLB di Tengah Isu Masuknya Dir...	Garuda Gelar RUPSLB di Tengah Isu Masuknya Dir...	2025-09-29 00:00:00	Bloomberg Technoz	Manajemen
2	https://voi.id/ekonomi/519004/komisi-v-dpr-bak...	Komisi V DPR Bakal Dalami Dugaan Mafia Jual Be...	JAKARTA - Ketua Komisi V DPR Lasarus mengataka...	2025-09-29 00:00:00	VOI.ID	Rute/Operasional
4	https://www.kompasiana.com/zainularifin2714/68...	Rencana Merger Garuda Indonesia - Pelita Air: ...	Latar Belakang\nPada pertengahan 2023, wacana ...	2025-09-29 00:00:00	Kompasiana.com	Lainnya
5	https://www.cnnindonesia.com/ekonomi/202509292...	Dony Oskaria Pastikan Merger Pelita Air-Garuda...	--\nPlt Menteri Badan Usaha Milik Negara (BUMN...	2025-09-29 00:00:00	CNN Indonesia	Lainnya

Gambar *Preview* hasil *scraping* setelah penghapusan baris kosong

Selanjutnya dilakukan deteksi bahasa pada kolom “judul” menggunakan pustaka LangDetect. Fungsi khusus detect_language() dibuat untuk mengidentifikasi bahasa setiap judul berita dengan menangani kemungkinan error, misalnya jika teks kosong atau bukan bertipe string. Setelah bahasa terdeteksi, data yang berbahasa Inggris (kode “en”) dihapus agar hanya berita berbahasa Indonesia yang dipertahankan dalam dataset.

```
df_cleaned['detected_language'] = df_cleaned['judul'].apply(detect_language)
```

	link	judul	konten	tanggal	portal	tag
0	https://kumparan.com/kumparanbisnis/garuda-ind...	Garuda Indonesia Kembali RUPSLB di Tengah Isu ...	Garuda Indonesia Kembali RUPSLB di Tengah Isu ...	2025-09-30 00:00:00	Kumparan	Manajemen
1	https://www.bloombergtechnoz.com/detail-news/8...	Garuda Gelar RUPSLB di Tengah Isu Masuknya Dir...	Garuda Gelar RUPSLB di Tengah Isu Masuknya Dir...	2025-09-29 00:00:00	Bloomberg Technoz	Manajemen
2	https://voi.id/ekonomi/519004/komisi-v-dpr-bak...	Komisi V DPR Bakal Dalami Dugaan Mafia Jual Be...	JAKARTA - Ketua Komisi V DPR Lasarus mengataka...	2025-09-29 00:00:00	VOI.ID	Rute/Operasional
4	https://www.kompasiana.com/zainularifin2714/68...	Rencana Merger Garuda Indonesia - Pelita Air: ...	Latar Belakang\nPada pertengahan 2023, wacana ...	2025-09-29 00:00:00	Kompasiana.com	Lainnya
5	https://www.cnnindonesia.com/ekonomi/202509292...	Dony Oskaria Pastikan Merger Pelita Air-Garuda...	--\nPlt Menteri Badan Usaha Milik Negara (BUMN...	2025-09-29 00:00:00	CNN Indonesia	Lainnya

Gambar *Preview dataset* setelah penghapusan artikel berbahasa Inggris

Data berita yang berbahasa Inggris kemudian disimpan dalam *dataframe* terpisah. Tujuannya adalah untuk menampilkan daftar berita yang terdeteksi menggunakan bahasa Inggris untuk diperiksa sebelum dihapus dari *dataset* utama.

Rows removed (English titles):

	link	judul	konten	tanggal	portal	tag	detected_languag
19	https://www.ch-aviation.com/news/158639-citili...	Citilink, Garuda Indonesia to reactivate more ...	Aviation Intelligence for your everyday use\nO...	2025-09-27 00:00:00	ch-aviation	Lainnya	en
33	https://www.ch-aviation.com/news/158613-garuda...	Garuda Indonesia says Pelita Air merger in ear...	Aviation Intelligence for your everyday use\nO...	2025-09-25 00:00:00	ch-aviation	Lainnya	en
48	https://www.kompas.id/artikel/en-berat-sebelah...	The One-Sided Merger of Garuda Indonesia and P...	The issue of merging PT Garuda Indonesia (Pers...	2025-09-24 00:00:00	Kompas.id	Lainnya	en
62	https://en.tempo.co/read/2050959/garuda-indone...	Garuda Indonesia to Launch New Halim-Palembang...	TEMPO.CO, Jakarta - PT Garuda Indonesia (Perse...	2025-09-23 00:00:00	Tempo.co	Lainnya	en

Gambar *Preview dataset* berbahasa Inggris yang dihapus

Setelah proses penyaringan selesai, data bersih disimpan ke dalam file baru dengan nama *data_link_berita_with_content_cleaned.csv*. File ini berisi berita dengan konten lengkap dan berbahasa Indonesia yang siap digunakan pada tahap analisis berikutnya.

Sebagai tahap akhir, dilakukan analisis deskriptif awal terhadap dataset hasil pembersihan. Analisis ini mencakup informasi statistik dasar, jumlah nilai unik pada setiap kolom kategorikal, serta distribusi data berdasarkan portal berita dan tag yang ada. Selain itu, kolom *tanggal* dikonversi ke format waktu standar untuk melihat rentang waktu publikasi berita yang berhasil dikumpulkan. Hasil analisis menunjukkan persebaran artikel dari berbagai portal dengan periode publikasi tertentu yang akan menjadi dasar untuk analisis lanjutan.

Descriptive Statistics:

	link	judul	konten	tanggal	portal	tag
count	495	495	495	495	495	495
unique	495	494	494	191	104	5
top	https://www.tempo.co/ekonomi/tahun-depan-garud...	Danantara injects US$405 million into Garuda I...	Aviation Intelligence for your everyday use\nO...	2025-06-10 00:00:00	Kompas.com	Lainnya
freq	1	2	2	19	66	347

Gambar Distribusi hasil *scraping* konten setelah *cleaning*

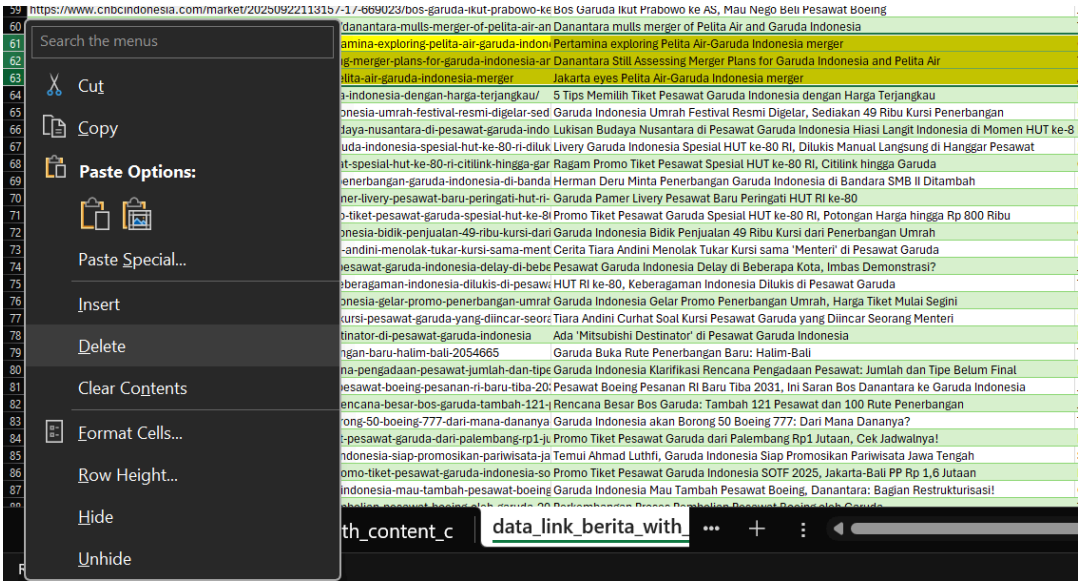
portal	count
Kompas.com	66
Tempo.co	34
Bisnis.com	30
Liputan6.com	26
CNBC Indonesia	26
CNN Indonesia	25
detikFinance	25
ANTARA News	25
detikTravel	13
Suara.com	12

tag	count
Lainnya	347
Rute/Operasional	97
Keuangan	22
Hukum/Regulasi	19
Manajemen	10

```
Date Range of Articles:
Earliest Date: 2012-12-07 00:00:00
Latest Date: 2025-09-30 00:00:00
```

Gambar Distribusi hasil *scraping* konten setelah *cleaning* per portal dan tag

Selanjutnya, isi *file* diperiksa secara manual melalui Microsoft Excel. Dari hasil pemeriksaan, ditemukan bahwa, meskipun judul-judul bahasa Inggris sudah dihapus sebelum dilakukan akuisisi isi konten berita, beberapa judul dan konten berbahasa Inggris masih saja tetap masuk ke dalam *dataset*. Oleh karena itu, dilakukan pembersihan secara manual dengan penandaan dan penghapusan dalam Microsoft Excel.



Gambar Flag manual dengan Excel

Setelah pembersihan awal tersebut dilakukan, dilakukan pemeriksaan statistika deskriptif dasar terhadap data. Dari hasil pemeriksaan tersebut, didapatkan pengurangan jumlah baris data dari yang awalnya berjumlah 495 menjadi 469.

Descriptive Statistics:

	link	judul	konten	tanggal	portal	tag
count	469	469	469	469	469	469
unique	469	469	469	176	90	5
top	https://www.tempo.co/ekonomi/tahun-depan-garud...	Tahun Depan Garuda Datangkan 24 Pesawat Baru	TEMPO.CO, Jakarta - Maskapai penerbangan pelat...	10/6/2025	Kompas.com	Lainnya
freq	1	1	1	19	66	322

Gambar Distribusi hasil *scraping* konten setelah *cleaning manual*

portal	count
Kompas.com	66
Bisnis.com	30
Tempo.co	30
CNBC Indonesia	26
Liputan6.com	26
detikFinance	25
CNN Indonesia	25
ANTARA News	21
detikTravel	13
MetroTVNews.com	12

tag	count
Lainnya	322
Rute/Operasional	97
Keuangan	22
Hukum/Regulasi	19
Manajemen	9

Date Range of Articles:
Earliest Date: 2012-12-07 00:00:00
Latest Date: 2025-09-30 00:00:00

Gambar Distribusi hasil *scraping* konten setelah *cleaning manual* per portal dan tag

a.4 *Tagging* dengan OpenAI API

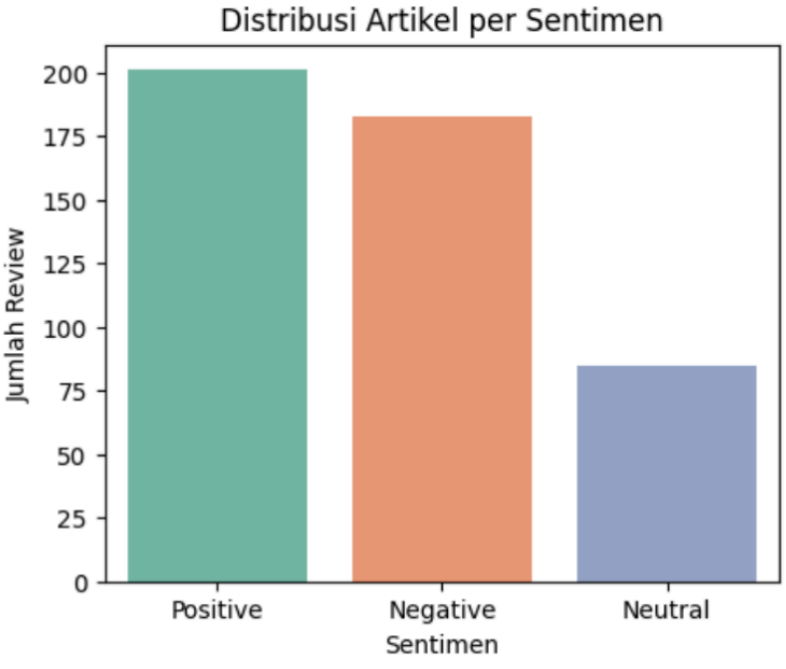
Untuk meningkatkan akurasi klasifikasi, dilakukan proses penandaan ulang (*re-tagging*) dengan bantuan model bahasa besar (LLM) melalui OpenAI API. Model yang digunakan adalah GPT-4.1, dengan pertimbangan keseimbangan antara harga dan efektivitas pemrosesan teks dalam skala besar.

Selain menghasilkan kategori topik baru (tag_new), tahap ini juga menambahkan dimensi analisis sentimen, yaitu mengidentifikasi polaritas emosi dari teks (*positive*, *neutral*, atau *negative*). Proses pelabelan sentimen dilakukan menggunakan model bahasa GPT-4.1 yang diintegrasikan melalui API OpenAI pada Google Colab. Sebelum menjalankan pelabelan, dilakukan proses instalasi dependensi seperti openai, pandas, dan google-colab, serta inisialisasi variabel API agar koneksi dapat dijalankan secara otomatis. Setelah itu, dilakukan pengaturan nama file input dan output, penentuan kolom teks yang akan dianalisis, serta daftar kategori berita yang valid, yaitu *Kinerja & Keuangan*, *Operasional & Pelayanan*, *Regulasi & Kebijakan*, *Krisis & Kontroversi*, dan *Industri & Pariwisata Nasional*.

```
system_instruction = (
"Kamu adalah AHLI analisis sentimen yang sangat AKURAT. "
"Untuk setiap teks, kamu harus mengeluarkan output dalam format JSON lengkap seperti contoh berikut:\n\n"
'Contoh:\n'
'{"sentiment": "Positive", "tag_new": "Operasional & Pelayanan"}\n'
'{"sentiment": "Neutral", "tag_new": "Kinerja & Keuangan"}\n'
'{"sentiment": "Negative", "tag_new": "Krisis & Kontroversi"}\n\n'
"Selalu isi kedua field dengan nilai yang sesuai.\n"
"Nilai sentiment hanya boleh: Positive, Neutral, atau Negative.\n"
"Nilai tag hanya boleh salah satu dari lima kategori berikut:\n"
"1. Kinerja & Keuangan\n"
"2. Operasional & Pelayanan\n"
"3. Regulasi & Kebijakan\n"
"4. Krisis & Kontroversi\n"
"5. Industri & Pariwisata Nasional\n\n"
"Jangan kosongkan field apapun. Jangan tulis apapun di luar JSON."
)
```

Gambar *Prompt labelling* API OpenAI

Fungsi utama dalam tahap ini adalah analyze_sentiment_openai(). Fungsi ini menerima input berupa teks berita, kemudian mengirimkannya ke model GPT-4.1 untuk dianalisis. Model diminta menghasilkan output dalam format JSON yang berisi dua atribut, yaitu sentiment dan tag_new. Atribut sentiment hanya memiliki tiga kemungkinan nilai, yakni *Positive*, *Neutral*, atau *Negative*, sedangkan tag_new menentukan kategori berita sesuai konteks isi teks. Untuk memastikan hasil yang valid, fungsi ini juga mencakup proses validasi dan *fallback* agar sistem tetap memberikan hasil standar meskipun terjadi kesalahan pemrosesan. Kemudian, didapatkan distirbusi artikel per sentimen sebagai berikut:




```
sentiment
Positive    201
Negative    183
Neutral      85
Name: count, dtype: int64
```

Gambar Distribusi Artikel per Sentimen

b. *Pre-processing* apa saja yang Anda terapkan. Jelaskan alasan masing-masing *pre-process* tersebut mengapa dibutuhkan pada tugas ini.
Jawaban:

Berikut ini langkah-langkah yang akan dilakukan pada tahapan persiapan dan praproses data:

1. Mulai: Data mentah masuk ke dalam proses.
2. *Lowercasing*: Seluruh teks diubah menjadi huruf kecil (*lowercase*) untuk menyeragamkan data dan menghindari perbedaan yang tidak perlu antar kata (misalnya, "Data" dan "data" dianggap sama).
3. Hapus Emoji dan Karakter: Karakter non-teks seperti emoji, simbol, atau karakter khusus yang tidak relevan dengan analisis dihapus.
4. Koreksi Ejaan: Dilakukan perbaikan kesalahan pengetikan atau ejaan untuk memastikan konsistensi dan akurasi kata.
5. Selesai: Data teks telah bersih dan siap untuk tahap analisis berikutnya.

b.1 *Lowercasing* dan Hapus Emoji-Karakter

Pada tahap ini, dilakukan proses pemrosesan awal terhadap teks konten berita untuk memastikan bahwa seluruh data berada dalam format yang seragam dan bebas dari karakter yang tidak relevan. Proses ini mencakup beberapa langkah utama yaitu konversi huruf menjadi huruf kecil (*lowercasing*), penghapusan emoji serta karakter non-alfanumerik, dan perlindungan angka yang mengandung makna finansial agar tidak terhapus selama pembersihan.

Langkah pertama dimulai dengan memuat dataset hasil pembersihan sebelumnya, yaitu *data_link_berita_with_content_cleaned_manual.csv*, menggunakan pustaka Pandas. *Dataset* ini memuat berita yang telah diseleksi dan siap untuk tahap pra-pemrosesan teks lebih lanjut.

Selanjutnya ditentukan beberapa pola (*regular expression*) untuk mendeteksi dan menghapus karakter yang tidak diinginkan, meliputi emoji dan karakter "*zero-width*", yang sering muncul akibat salinan teks dari laman berita atau media sosial; dan tanda baca dan simbol non-alfanumerik, agar teks yang tersisa hanya terdiri atas huruf, angka, dan spasi.

Namun, oleh karena data ini mencakup konteks finansial (misalnya laporan laba rugi, nilai saham, atau angka mata uang), maka disusun pula pola proteksi numerik. Pola ini menjaga agar angka-angka penting yang berkaitan dengan istilah finansial seperti "Rp 2 miliar", "naik 5%", atau "laba 120 juta" tidak terhapus selama proses pembersihan. Perlindungan dilakukan dengan membungkus angka-angka tersebut menggunakan placeholder khusus yang kemudian dikembalikan ke bentuk aslinya setelah seluruh proses selesai.

Setelah semua pola ditetapkan, dibuat dua fungsi utama, yakni protect_financial_numbers(), berfungsi membungkus angka-angka finansial dengan placeholder agar tidak ikut terhapus; dan clean_text_with_numeric_rules(), berfungsi menghapus emoji, tanda baca, dan angka yang tidak relevan, menormalisasi huruf menjadi huruf kecil, serta merapikan spasi ganda. Kedua fungsi tersebut diterapkan pada kolom "konten", lalu hasilnya disimpan ke kolom baru bernama "konten_hapus_karakter". Dengan demikian, setiap berita kini memiliki dua versi: versi asli dan versi yang telah dibersihkan dari karakter berlebih namun tetap mempertahankan informasi numerik penting.

Dataset hasil pembersihan ini kemudian disimpan dalam *file* baru bernama *garudaindonesia_news_lower_hapus_karakter.csv* untuk digunakan pada tahap analisis berikutnya.

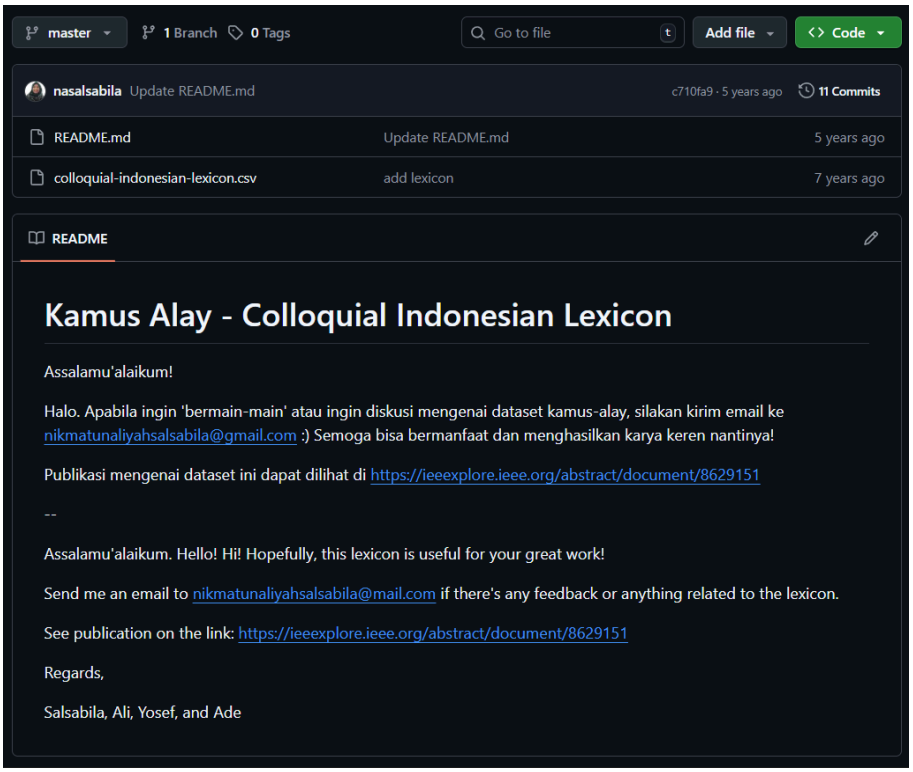
	konten	konten_hapus_karakter
0	Garuda Indonesia Kembali RUPSLB di Tengah Isu ...	garuda indonesia kembali rupslb di tengah isu ...
1	Garuda Gelar RUPSLB di Tengah Isu Masuknya Dir...	garuda gelar rupslb di tengah isu masuknya dir...
2	JAKARTA - Ketua Komisi V DPR Lasarus mengataka...	jakarta ketua komisi v dpr lasarus mengatakan ...
3	Latar Belakang\nPada pertengahan 2023, wacana ...	latar belakang pada pertengahan wacana konsoli...
4	--\nPlt Menteri Badan Usaha Milik Negara (BUMN...	plt menteri badan usaha milik negara bumn seka...

Gambar *Preview Dataset* hasil pembersihan karakter

b.2 Koreksi Ejaan

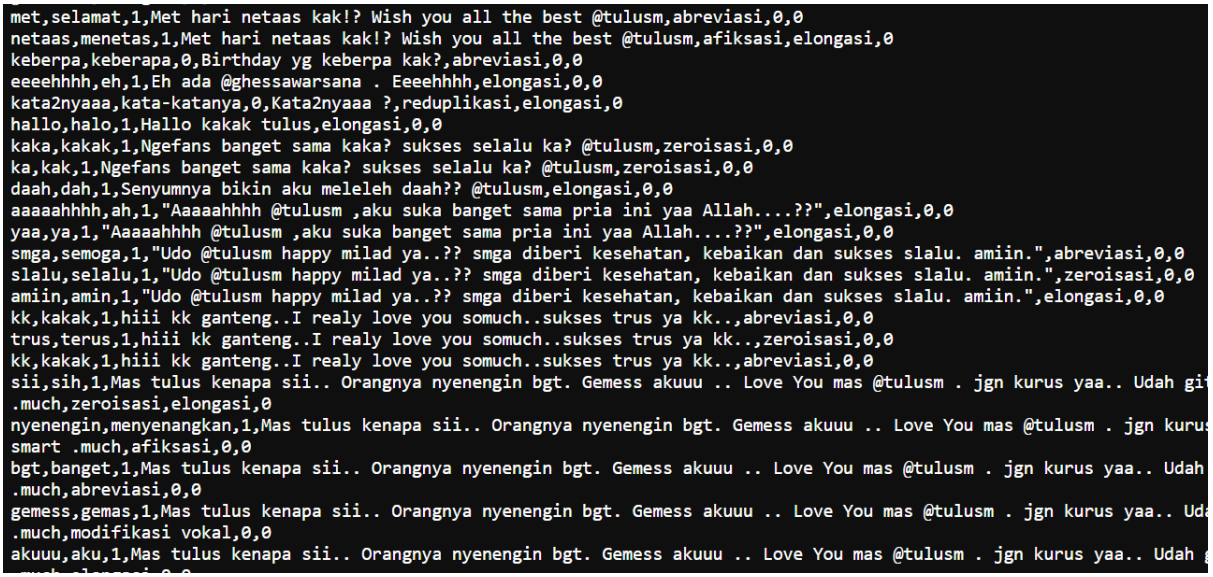
Tahap ini bertujuan untuk melakukan koreksi atau normalisasi ejaan terhadap teks konten berita agar bentuk katanya menjadi lebih baku dan konsisten. Proses ini penting karena data hasil scraping sering kali mengandung bahasa tidak formal, singkatan, atau bentuk *slang* yang perlu disesuaikan sebelum analisis lebih lanjut dilakukan.

Langkah pertama adalah memuat dua *file* utama, yaitu *garudaindonesia_news_lower_hapus_karakter.csv* yang berisi kolom *konten_lowercase*, dan *colloquial-indonesian-lexicon.csv* dari repositori github.com/nasalsabila/kamus-alay yang berisi daftar padanan kata tidak baku (*slang*) dengan bentuk formalnya. Kedua *file* ini digunakan sebagai dasar untuk melakukan proses substitusi kata.



Gambar GitHub nasalsabila/kamus-alay

Selanjutnya dilakukan pembuatan kamus konversi dari kata slang ke bentuk formal. Kolom pada file leksikon dinormalisasi ke huruf kecil (*lowercase*) dan dihapus karakter kosong di awal atau akhir string menggunakan fungsi *strip()*. Jika terdapat duplikasi pada kolom *slang*, hanya entri pertama yang digunakan agar tidak terjadi konflik dalam pemetaan. Hasilnya adalah sebuah kamus Python (slang2formal) yang berisi pasangan kata tidak baku dan padanan formalnya.



```
met,selamat,1,Met hari netaas kak!? Wish you all the best @tulusm,abreviasi,0,0
netaas,menetas,1,Met hari netaas kak!? Wish you all the best @tulusm,afiksasi,elongasi,0
keberpa,keberapa,0,Birthday yg keberpa kak?,abreviasi,0,0
eeeehhhh,eh,1,Eh ada @ghessawarsana . Eeeehhhh,elongasi,0,0
kata2nyaaa,kata-katanya,0,Kata2nyaaa ?,reduplikasi,elongasi,0
hallo,halo,1,Hallo kakak tulus,elongasi,0,0
kaka,kakak,1,Ngefans banget sama kaka? sukses selalu ka? @tulusm,zeroisasi,0,0
ka,kak,1,Ngefans banget sama kaka? sukses selalu ka? @tulusm,zeroisasi,0,0
daah,dah,1,Senyumnya bikin aku meleleh daah?? @tulusm,elongasi,0,0
aaaaahhhh,ah,1,"Aaaaahhhh @tulusm ,aku suka banget sama pria ini yaa Allah....??",elongasi,0,0
yaa,ya,1,"Aaaaahhhh @tulusm ,aku suka banget sama pria ini yaa Allah....??",elongasi,0,0
smga,semoga,1,"Udo @tulusm happy milad ya..?? smga diberi kesehatan, kebaikan dan sukses slalu. amiin.",abreviasi,0,0
slalu,selalu,1,"Udo @tulusm happy milad ya..?? smga diberi kesehatan, kebaikan dan sukses slalu. amiin.",zeroisasi,0,0
amiin,amin,1,"Udo @tulusm happy milad ya..?? smga diberi kesehatan, kebaikan dan sukses slalu. amiin.",elongasi,0,0
kk,kakak,1,hiii kk ganteng..I realy love you somuch..sukses trus ya kk..,abreviasi,0,0
trus,terus,1,hiii kk ganteng..I realy love you somuch..sukses trus ya kk..,zeroisasi,0,0
kk,kakak,1,hiii kk ganteng..I realy love you somuch..sukses trus ya kk..,abreviasi,0,0
sii,sih,1,Mas tulus kenapa sii.. Orangnya nyenengin bgt. Gemess akuuu .. Love You mas @tulusm . jgn kurus yaa.. Udah gi
.much,zeroisasi,elongasi,0
nyenengin,menyenangkan,1,Mas tulus kenapa sii.. Orangnya nyenengin bgt. Gemess akuuu .. Love You mas @tulusm . jgn kurus
smart .much,afiksasi,0,0
bgt,banget,1,Mas tulus kenapa sii.. Orangnya nyenengin bgt. Gemess akuuu .. Love You mas @tulusm . jgn kurus yaa.. Udah
.much,abreviasi,0,0
gemess,gemas,1,Mas tulus kenapa sii.. Orangnya nyenengin bgt. Gemess akuuu .. Love You mas @tulusm . jgn kurus yaa.. Uda
.much,modifikasi vokal,0,0
akuuu,aku,1,Mas tulus kenapa sii.. Orangnya nyenengin bgt. Gemess akuuu .. Love You mas @tulusm . jgn kurus yaa.. Udah
```

Gambar Daftar padanan tidak baku dari nasalsabila/kamus-alay

Langkah berikutnya adalah membangun pola ekspresi reguler (*regex*) untuk mendeteksi kata-kata yang akan diganti. Daftar kata dalam kamus diurutkan berdasarkan panjang karakter dari yang terpanjang ke terpendek agar sistem dapat memprioritaskan pencocokan frasa panjang terlebih dahulu. Setiap kata di-*escape* agar karakter khusus dapat dibaca dengan benar oleh regex, kemudian digabung menjadi satu pola pencarian yang siap digunakan.

Fungsi normalize_text() kemudian dibuat untuk melakukan substitusi kata dalam teks berdasarkan kamus tersebut. Setiap kata yang cocok dengan pola regex akan digantikan dengan bentuk formalnya menggunakan pemetaan dari kamus *slang2formal*.

Setelah fungsi siap, proses normalisasi diterapkan pada kolom konten_hapus_karakter untuk menghasilkan kolom baru bernama konten_normalized. Proses ini dilakukan secara otomatis untuk setiap baris teks. Setelah normalisasi selesai, teks juga dirapikan dengan menghapus spasi ganda dan memastikan tidak ada karakter kosong di awal maupun akhir kalimat.

	konten_hapus_karakter	konten_normalized
0	garuda indonesia kembali rupslb di tengah isu ...	garuda indonesia kembali rupslb di tengah isu ...
1	garuda gelar rupslb di tengah isu masuknya dir...	garuda gelar rupslb di tengah isu masuknya dir...
2	jakarta ketua komisi v dpr lasarus mengatakan ...	jakarta ketua komisi v dpr lasarus mengatakan ...
3	latar belakang pada pertengahan wacana konsoli...	latar belakang pada pertengahan wacana konsoli...
4	plt menteri badan usaha milik negara bumn seka...	plt menteri badan usaha milik negara bumn seka...

Gambar *Preview dataset* setelah normalisasi ejaan

Langkah terakhir adalah menyimpan hasil normalisasi ke dalam file baru bernama *garudaindonesia_news_normalized.csv* agar dapat digunakan pada tahap analisis selanjutnya. *File* ini berisi versi teks yang sudah bersih dan telah dikoreksi secara ejaan menggunakan leksikon bahasa Indonesia. Kolom konten_normalized inilah yang akan digunakan sebagai data pelatihan dan pengujian model.

b.3 Definisi *dataset*

Dataset yang digunakan:
https://github.com/renaldoaluska/pba2025gasal/blob/main/%23TugasA-1/7-garudaindonesia_news_cleaned_simple.csv

Tabel Definisi *dataset*

No.	Nama Kolom	Deskripsi
1	link	URL atau tautan unik yang merujuk langsung ke sumber artikel berita asli.
2	judul	Judul asli dari artikel berita yang diambil dari sumbernya.
3	konten	Isi atau teks lengkap dari artikel berita yang menjadi objek utama analisis.
4	tanggal	Tanggal publikasi artikel berita, menunjukkan kapan berita tersebut diterbitkan.
5	portal	Nama media atau portal berita yang mempublikasikan artikel.
6	tag	Hasil *tagging* manual berdasarkan judul berita. Penentuan *tag* dilakukan berdasarkan kecocokan ekspresi reguler (*regex*) dari teks (setelah diubah menjadi huruf kecil) dengan daftar kata kunci yang telah ditentukan menggunakan fungsi =ARRAYFORMULA() di Google Spreadsheets. (Lihat bagian a.2) ● Tag "Keuangan" diberikan jika teks mengandung salah satu kata berikut: *laba*, *rugi*, *pendapatan*, *utang*, *restrukturisasi*, *pkpu*, *obligasi*, *kuartal*, atau *laporan keuangan*. ● Tag "Rute/Operasional" diberikan jika teks mengandung salah satu kata berikut: *rute*, *penerbangan baru*, *buka rute*, *tutup rute*, *frekuensi penerbangan*, *operasional*, *insiden*, *delay*, *batal*, atau *keselamatan*. ● Tag "Manajemen" diberikan jika teks mengandung salah satu kata berikut: *direktur*, *komisaris*, *manajemen*, *rups*, *ceo*, *pergantian*, atau *pengurus*. ● Tag "Hukum/Regulasi" diberikan jika teks mengandung salah satu kata berikut: *kasus*, *pengadilan*, *kpk*, *dugaan*, *suap*, *hukum*, *sidang*, *sanksi*, *denda*, atau *izin*. ● Tag "Lainnya" diberikan jika teks tidak memenuhi kriteria kata kunci dari salah satu kategori di atas.
7	konten_hapus_karakter	Hasil pemrosesan kolom *konten* dengan simbol, angka, atau karakter khusus yang telah dihapus.
8	**konten_normalized**	Hasil pemrosesan kolom *konten_hapus_karakter* dengan penyeragaman istilah-istilah gaul (*slang*). **Kolom ini digunakan untuk *training* dan *testing*.**
9	konten_nostop	Hasil pemrosesan kolom *konten_normalized* dengan *stopwords* (kata yang frekuensinya tinggi namun tidak memiliki makna penting) yang telah dihapus.

10	konten_stem	Hasil pemrosesan kolom *konten_nostop* dengan pengonversian setiap katanya ke dalam bentuk kata dasar/akar.
11	**sentiment**	Label hasil analisis polaritas sentimen kalimat: Positive, Neutral, atau Negative hasil klasifikasi GPT-4.1 berdasarkan konteks isi berita (Lihat bagian a.4) **Kolom ini digunakan untuk patokan kelas label sentimen.**
12	tag_new	Kategori topik baru hasil klasifikasi GPT-4.1 berdasarkan konteks isi berita.

2. Baseline Model

Berdasarkan *dataset* tersebut, pilih model yang akan dianggap sebagai *baseline*. Model meliputi algoritma ekstraksi fitur (misal Bag of Words) dan *machine learning* yang digunakan (misal Naïve Bayes).

a. Jelaskan algoritma ekstraksi yang digunakan pada tahap ini.

Jawaban:

Ekstraksi fitur dilakukan menggunakan pendekatan **Bag of Words (BoW)**, yaitu metode representasi teks yang menghitung frekuensi kemunculan kata-kata dalam dokumen tanpa memperhatikan urutan atau konteksnya. Meskipun sederhana, BoW cukup efektif untuk menangkap pola umum dalam teks dan sering digunakan sebagai baseline dalam tugas klasifikasi.

Metode Bag of Words tetap memiliki kelemahan utama karena tidak mempertimbangkan urutan kata maupun konteks makna dalam kalimat sehingga informasi semantik yang penting bisa hilang. Setiap kata diperlakukan sebagai fitur yang berdiri sendiri (*independent*), tanpa membedakan apakah kata tersebut muncul dalam bentuk negasi, ironi, atau dalam struktur kalimat yang memengaruhi makna. Selain itu, BoW menghasilkan vektor berdimensi tinggi dan *sparsity* yang tinggi, yang dapat menurunkan efisiensi model dan membuatnya rentan terhadap *noise* dari kata-kata tidak penting.

b. Jelaskan algoritma *machine learning* yang digunakan.

Jawaban:

Untuk klasifikasi, digunakan algoritma **Multinomial Naïve Bayes**. Algoritma ini mengasumsikan bahwa fitur (kata) bersifat independen satu sama lain dan mengikuti distribusi multinomial. Naïve Bayes dikenal efisien dan cocok untuk data teks yang bersifat diskrit seperti hasil ekstraksi BoW.

c. Bagaimana kinerja yang diperoleh dari model ini?

Jawaban:

Dataset dibagi menggunakan metode *stratified train-test split* dengan rasio 80:20 sehingga distribusi label tetap proporsional di kedua *subset*. Model *baseline* dilatih menggunakan data latih (*training set*) dan dievaluasi terhadap data uji (*test set*) yang tidak mengalami augmentasi atau modifikasi.

Berikut tabel distribusi label sebelum dan sesudah pembagian train-test:

Kelas	Jumlah Awal	Jumlah *Train*	Jumlah *Test*
Positive	201	161	40
Negative	183	146	37
Neutral	85	68	17

Hasil evaluasi model *baseline* (BoW + NB) terhadap data uji ditunjukkan pada tabel berikut:

Kelas	Precision	Recall	F1-Score	Support
Positive	0.88	0.95	0.92	40
Negative	0.82	0.86	0.84	37
Neutral	0.75	0.53	0.62	17

Macro Avg F1-Score: 0.79
Weighted Avg F1-Score: 0.83

Secara umum, model menunjukkan akurasi yang tergolong tinggi, terutama pada kelas *Positive* dan *Negative*. Namun, performa pada kelas *Neutral* relatif rendah, dengan F1-score hanya 0.62. Hal ini mengindikasikan adanya ketidakseimbangan dalam prediksi, dimana model cenderung bias terhadap kelas mayoritas. Temuan ini menjadi dasar untuk melakukan augmentasi data dan eksplorasi pendekatan lain.

3. Data Augmentation and Improvement

Data augmentation adalah proses memperbanyak dan memvariasikan data pelatihan dengan memodifikasi atau menghasilkan data teks baru dari data yang sudah ada, tanpa mengubah makna intinya. Teknik ini berguna ketika jumlah data pelatihan terbatas, tidak seimbang antar kelas, atau saat ingin meningkatkan kemampuan generalisasi model seperti analisis sentimen, klasifikasi, NER, atau *chatbot*.

a. Jelaskan bagaimana anda melakukan data augmentasi untuk mendapatkan *dataset* yang lebih baik. Tunjukkan distribusi akhir dari *dataset*

Jawaban:

Data augmentation dilakukan untuk memperbaiki ketidakseimbangan distribusi label dalam data latih. Teknik yang digunakan adalah ***contextual embedding*** dengan model IndoBERT melalui *library* nlpaug. Proses ini mengganti kata-kata dalam kalimat berdasarkan konteks semantik sehingga menghasilkan variasi teks yang tetap relevan secara makna. Langkah-langkah augmentasi:

- Augmentasi dilakukan hanya pada data latih, setelah proses *train-test split*, untuk menjaga integritas evaluasi.
- Kelas yang di-augmentasi adalah *Neutral* (total: 85) dan *Negative* (total: 183) karena jumlahnya lebih sedikit dibanding *Positive* (total: 201).

- Target distribusi adalah 161 data per kelas, disamakan dengan jumlah data *Positive* di training set.
- Augmentasi dilakukan secara iteratif hingga jumlah target tercapai, dengan substitusi kata berbasis konteks IndoBERT.

Distribusi data *train* sebelum dan setelah augmentasi, distribusi data *test*, dan distribusi keseluruhan data per kelas:

Kelas	**Jumlah total**	**Jumlah *train* Sebelum**	**Jumlah *train* Sesudah**	**Jumlah *test***
Positive	201	161	161	40
Neutral	85	68	161	37
Negative	183	146	161	17

b. Berdasarkan *dataset* yang telah diperbaiki, lakukan analisis sentimen dengan menggunakan mode *machine learning* pada *baseline*. Jelaskan apakah terjadi perbaikan kinerja?

Jawaban:

Setelah augmentasi, dilakukan pelatihan ulang model *baseline* menggunakan fitur BoW dan algoritma Naïve Bayes. Hasil evaluasi menunjukkan adanya peningkatan kinerja, terutama pada kelas *Neutral* yang sebelumnya memiliki F1-score rendah. Perbandingan kinerja BoW + Naïve Bayes:

Kelas	**Asli**			**Augmented**		
	F1		**Accuracy**	**F1**		**Accuracy**
Negative	0.84	0.7928	0.8404	0.85	0.8345	0.8617
Neutral	0.62			0.73		
Positive	0.92			0.93		

Terjadi peningkatan signifikan pada kelas *Neutral* dan skor keseluruhan, menunjukkan bahwa augmentasi berhasil membantu model mengenali pola dari kelas minoritas.

4. Pertimbangkan sejumlah cara untuk meningkatkan kinerja dari model analisis sentimen!
Misal dengan menggunakan *feature engineering* yang lain dan atau menerapkan algoritma *machine learning* lain yang lebih bagus.

a. Bandingkan kinerja dari sejumlah model atau cara yang dipilih pada soal sebelumnya, dengan *dataset* asli (*not augmented*) sebagai *baseline*. Model atau teknik mana yang memberikan kinerja paling baik?
Jawaban:

Berbagai kombinasi fitur dan algoritma diuji untuk membandingkan performa dengan dan tanpa augmentasi. Berikut ringkasan hasil terbaik:

Algoritma Ekstraksi: Bag of Words (Keseluruhan)

Algoritma ML	Data	Accuracy	Macro F1
Naïve Bayes	Asli	0.8404	0.7928
	Augmented	**0.8617**	**0.8345**
Logistic Regression	Asli	0.7872	0.7095
	Augmented	0.7872	0.7307
Support Vector Machine	Asli	0.7766	0.7012
	Augmented	0.7872	0.7227

Algoritma Ekstraksi: Bag of Words (Per Kelas)

Algoritma ML	Data	F1 Positive	F1 Neutral	F1 Negative
Naïve Bayes	Asli	0.92	0.62	0.84
	Augmented	**0.93**	**0.73**	**0.85**
Logistic Regression	Asli	0.85	0.43	0.85
	Augmented	0.84	0.50	0.85
Support Vector Machine	Asli	0.84	0.43	0.84
	Augmented	0.84	0.48	0.85

Algoritma Ekstraksi: TF-IDF (Keseluruhan)

Algoritma	Data	Accuracy	Macro F1
Naïve Bayes	Asli	0.7447	0.5459
	Augmented	**0.8404**	**0.8083**
Logistic Regression	Asli	0.7340	0.5377
	Augmented	0.8085	0.7353
Support Vector Machine	Asli	0.7979	0.6912
	Augmented	0.8191	0.7437

Algoritma Ekstraksi: TF-IDF (Per Kelas)

Algoritma	Data	F1 Positive	F1 Neutral	F1 Negative
Naïve Bayes	Asli	0.82	0.00	0.82
	Augmented	**0.93**	**0.67**	**0.83**
Logistic Regression	Asli	0.81	0.00	0.80
	Augmented	0.88	0.50	0.82
Support Vector Machine	Asli	0.87	0.36	0.84
	Augmented	0.89	0.50	0.84

Dari tabel-tabel di atas, dapat disimpulkan bahwa model terbaik secara keseluruhan adalah **BoW + Naïve Bayes dengan augmentasi**.

b. Berdasarkan soal sebelumnya, rencanakan upaya untuk meningkatkan performance dari masing-masing model atau teknik.

Jawaban:

Untuk meningkatkan performa lebih lanjut, beberapa strategi dapat diterapkan untuk percobaan lanjutan:

1. **Feature Engineering:**
    - Gabungkan BoW dan TF-IDF sebagai *hybrid feature*
    - Tambahkan n-gram dan *stopword filtering*
    - Gunakan *domain-specific lexicon* atau *sentiment scores*
2. **Modeling:**
    - Coba algoritma lain seperti Random Forest, XGBoost, atau *ensemble voting*

- Fine-tune IndoBERT langsung dengan *transfer learning*

3. **Data Strategy:**
   - Tambahkan augmentasi untuk kelas *Positive* agar variasi tetap seimbang
   - Lakukan *cleaning* dan *stemming* tambahan untuk *noise reduction* (saat ini baru dicoba eksplorasi di *dataset* yang sudah dinormalisasi saja, tetapi belum di-*stemming* agar augmentasinya tidak hilang konteks karena datanya sudah di-praproses sejak kelas sebelumnya/sebelum ETS)
4. **Evaluation:**
   - Gunakan stratified *cross-validation* untuk hasil yang lebih stabil
   - Analisis *error* pada *confusion matrix* untuk perbaikan *targeted*

Note: Jangan lupa sertakan *dataset* dan *snippet* atau program yang digunakan

LAMPIRAN

Kode dan hasil *output* bisa dilihat di (agar tidak menumpuk di sini):
https://github.com/renaldoaluska/pba2025gasal/blob/main/%23TugasA-1/Tugas_A_1_5026221144.ipynb

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Accuracy} = \frac{\text{Jumlah prediksi benar}}{\text{Total jumlah data}}$$

=== TABEL NEGATIVE ===

Fitur	Algoritma	Data	Precision	Recall	F1 Negative	Support
BoW	Naïve Bayes	Asli	0.82	0.86	0.84	37
BoW	Naïve Bayes	Augmented	0.86	0.84	0.85	37
BoW	Logistic Reg.	Asli	0.79	0.92	0.85	37
BoW	Logistic Reg.	Augmented	0.79	0.92	0.85	37
BoW	SVM	Asli	0.77	0.92	0.84	37
BoW	SVM	Augmented	0.79	0.92	0.85	37
TF-IDF	Naïve Bayes	Asli	0.78	0.86	0.82	37
TF-IDF	Naïve Bayes	Augmented	0.86	0.81	0.83	37
TF-IDF	Logistic Reg.	Asli	0.74	0.86	0.80	37
TF-IDF	Logistic Reg.	Augmented	0.77	0.89	0.82	37
TF-IDF	SVM	Asli	0.77	0.92	0.84	37
TF-IDF	SVM	Augmented	0.77	0.92	0.84	37

=== TABEL NEUTRAL ===

Fitur	Algoritma	Data	Precision	Recall	F1 Neutral	Support
BoW	Naïve Bayes	Asli	0.75	0.53	0.62	17
BoW	Naïve Bayes	Augmented	0.75	0.71	0.73	17
BoW	Logistic Reg.	Asli	0.55	0.35	0.43	17
BoW	Logistic Reg.	Augmented	0.53	0.47	0.50	17
BoW	SVM	Asli	0.55	0.35	0.43	17
BoW	SVM	Augmented	0.58	0.41	0.48	17
TF-IDF	Naïve Bayes	Asli	0.00	0.00	0.00	17
TF-IDF	Naïve Bayes	Augmented	0.63	0.71	0.67	17
TF-IDF	Logistic Reg.	Asli	0.00	0.00	0.00	17
TF-IDF	Logistic Reg.	Augmented	0.86	0.35	0.50	17
TF-IDF	SVM	Asli	0.80	0.24	0.36	17
TF-IDF	SVM	Augmented	0.86	0.35	0.50	17

=== TABEL POSITIVE ===

Fitur	Algoritma	Data	Precision	Recall	F1 Positive	Support
BoW	Naïve Bayes	Asli	0.88	0.95	0.92	40
BoW	Naïve Bayes	Augmented	0.90	0.95	0.93	40
BoW	Logistic Reg.	Asli	0.85	0.85	0.85	40
BoW	Logistic Reg.	Augmented	0.89	0.80	0.84	40
BoW	SVM	Asli	0.85	0.82	0.84	40
BoW	SVM	Augmented	0.85	0.82	0.84	40
TF-IDF	Naïve Bayes	Asli	0.72	0.95	0.82	40
TF-IDF	Naïve Bayes	Augmented	0.93	0.93	0.93	40
TF-IDF	Logistic Reg.	Asli	0.73	0.93	0.81	40
TF-IDF	Logistic Reg.	Augmented	0.84	0.93	0.88	40
TF-IDF	SVM	Asli	0.82	0.93	0.87	40
TF-IDF	SVM	Augmented	0.86	0.93	0.89	40

=== TABEL OVERALL METRIK ===

Fitur	Algoritma	Data	Accuracy	Macro F1
BoW	Naïve Bayes	Asli	0.8404	0.7928
BoW	Naïve Bayes	Augmented	0.8617	0.8345
BoW	Logistic Reg.	Asli	0.7872	0.7095
BoW	Logistic Reg.	Augmented	0.7872	0.7307
BoW	SVM	Asli	0.7766	0.7012
BoW	SVM	Augmented	0.7872	0.7227
TF-IDF	Naïve Bayes	Asli	0.7447	0.5459
TF-IDF	Naïve Bayes	Augmented	0.8404	0.8083
TF-IDF	Logistic Reg.	Asli	0.7340	0.5377
TF-IDF	Logistic Reg.	Augmented	0.8085	0.7353
TF-IDF	SVM	Asli	0.7979	0.6912
TF-IDF	SVM	Augmented	0.8191	0.7437