Natural Language Processing for Insight Extraction from Food Safety Reports: A PESTLE Perspective at Badan Gizi Nasional

1st Maharani Padma Utami
School of Industrial Engineering
Telkom University
Bandung, Indonesia
maharanipadmautami@student.telkomuniversity.ac.id

2nd Alfi Inayati
School of Industrial Engineering
Telkom University
Bandung, Indonesia
alfiinayatiai@student.telkomuniversity.ac.id

***Abstrak*— Tata kelola keamanan pangan bergantung pada berbagai dokumen laporan dan kebijakan institusional yang bersifat tidak terstruktur dan berjumlah besar, sehingga menyulitkan analisis manual secara sistematis. Penelitian ini mengusulkan pendekatan Natural Language Processing (NLP) berbasis *unsupervised learning* untuk mengekstrak insight kebijakan dari laporan keamanan pangan menggunakan kerangkan PESTLE (Political, Economic, Social, Technological, Legal, Environmental). Dataset terdiri dari laporan keamanan pangan dan dokumen kebijakan strategis yang diterbitkan oleh Badan Gizi Nasional, yang diproses melalui ekstraksi teks dan *optical character recognition (OCR).* Metode TF-IDF dan Latent Dirichlet Allocation (LDA) digunakan untuk ekstraksi kata kunci dan penemuan topik, diikuti dengan klasifikasi PESTLE berbasis aturan pada Tingkat paragraph. Hasil penelitian menunjukkan bahwa perspektif Technological dan Political mendominasi narasi keamanan pangan, sementara dimensi Economic dan Environmental relative kurang menonjol. Temuan ini menunjukkan potensi NLP sebagai alat bantu analisis kebijakan serta mengindikasikan adanya ruang penguatan kebijakan keamanan pangan yang lebih komperehensif.**

***Keywords*— keamanan pangan; natural language processing; analisis kebijakan, PESTLE, OCR; TF-IDF; LDA; text mining; BGN.**

I. INTRODUCTION

Keamanan pangan merupakan isu strategis nasional yang berkaitan erat dengan kesehatan masyarakat, stabilitas sosial, dan keberlanjutan pembangunan yang didefinisikan sebagai kondisi dan upaya yang diperlukan untuk mencegah pangan dari cemaran biologis, kimia, maupun benda lain yang berpotensi membahayakan kesehatan manusia [1]. Pemerintah Indonesia secara rutin menghasilkan berbagai laporan resmi terkait keamanan pangan, pengawasan pangan olahan, kejadian luar biasa (KLB) keracunan pangan, serta kebijakan ketahanan pangan dan gizi. Laporan-laporan tersebut diproduksi oleh berbagai institusi, antara lain Badan Pengawas Obat dan Makanan (BPOM), Direktorat Registrasi Pangan Olahan, dan instansi yang berperan dalam program keamanan pangan dan gizi nasional .

Seiring meningkatnya volume dan kompleksitas laporan keamanan pangan, tantangan utama yang dihadapi oleh pembuat kebijakan adalah bagaimana mengekstraksi insight strategis secara cepat, konsisten, dan komprehensif. Analisis manual terhadap dokumen teks yang panjang, heterogen, dan tidak terstruktur berpotensi menimbulkan bias interpretasi serta keterlambatan dalam pengambilan keputusan. Hal ini menjadi semakin krusial dalam konteks penguatan kebijakan berbasis bukti, khususnya bagi institusi seperti Badan Gizi Nasional yang memerlukan ringkasan isu lintas sektor secara sistematis.

Natural Language Processing (NLP) menawarkan pendekatan komputasional untuk menganalisis teks tidak terstruktur dan mengekstraksi informasi penting secara otomatis. NLP telah digunakan dalam pengelolaan dokumen institusional dan analisis laporan formal untuk meningkatkan efisiensi dan kualitas pengambilan informasi

Namun, pemanfaatan NLP untuk mengekstraksi insight kebijakan dari laporan keamanan pangan di Indonesia masih relatif terbatas. Di sisi lain, kerangka PESTLE (Political, Economic, Social, Technological, Legal, Environmental) banyak digunakan dalam analisis kebijakan publik untuk memetakan faktor eksternal yang memengaruhi suatu sektor secara multidimensional.

Penerapan PESTLE masih didominasi pendekatan kualitatif manual dan belum terintegrasi dengan analisis berbasis NLP. Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menerapkan NLP dalam mengekstraksi insight dari laporan keamanan pangan dan memetakannya ke dalam perspektif PESTLE.

II. RELATED WORK

Beberapa penelitian sebelumnya telah menerapkan Natural Language Processing (NLP) untuk analisis dokumen kebijakan dan laporan institusional. Blei et al. [2] memperkenalkan Latent Dirichlet Allocation (LDA) sebagai metode untuk mengekstrak topik laten dari korpus teks. Studi lain memanfaatkan NLP untuk analisis kebijakan publik dan kesehatan, namun umumnya berfokus pada klasifikasi tematik tanpa kerangka analisis strategis yang eksplisit. Dalam konteks keamanan pangan, sebagian besar penelitian menekankan aspek teknis dan risiko kesehatan, sementara integrasi NLP dengan kerangka PESTLE masih terbatas.

Pemanfaatan Natural Language Processing (NLP) menunjukkan peningkatan yang signifikan dalam beberapa tahun terakhir, seiring dengan melimpahnya dokumen kebijakan dan laporan resmi dalam bentuk teks tidak terstruktur. NLP telah digunakan untuk membantu analisis kebijakan, pengelolaan dokumen institusional, serta identifikasi isu strategis yang sebelumnya dianalisis secara manual. Di sisi lain, kerangka analisis kebijakan seperti PESTLE banyak digunakan untuk memahami faktor eksternal yang memengaruhi kebijakan publik secara multidimensional. Namun, kajian yang mengintegrasikan NLP dan PESTLE, khususnya dalam konteks laporan keamanan pangan, masih relatif terbatas. Oleh karena itu, bagian ini membahas

penelitian-penelitian terdahulu yang relevan sebagai landasan konseptual dan metodologis bagi penelitian ini.

A. NLP untuk Analisis Dokumen

Penelitian terkait NLP pada dokumen institusional menunjukkan bahwa NLP efektif untuk mengekstraksi informasi dari dokumen formal yang memiliki struktur semi-baku, seperti laporan tahunan dan dokumen kebijakan. Pendekatan preprocessing, tokenisasi, dan ekstraksi fitur linguistik terbukti mampu meningkatkan efisiensi pengelolaan informasi dokumen.

NLP banyak digunakan untuk analisis laporan kesehatan, surveilans penyakit, dan deteksi risiko kesehatan. Namun, sebagian besar studi berfokus pada data klinis atau media sosial, bukan pada laporan resmi keamanan pangan.

B. Analisis Laporan Keamanan Pangan

Laporan tahunan BPOM dan Direktorat Registrasi Pangan Olahan memuat informasi strategis mengenai pengawasan pre-market, perizinan, pengaduan masyarakat, serta kepatuhan terhadap standar keamanan pangan . Selain itu, laporan penyelidikan epidemiologi KLB keracunan pangan memberikan gambaran rinci tentang risiko kesehatan akibat pangan yang tidak aman. Meskipun kaya akan data tekstual, laporan-laporan ini umumnya dianalisis secara manual dan deskriptif.

C. PESTLE

PESTLE digunakan untuk memetakan faktor makro yang mempengaruhi kebijakan:

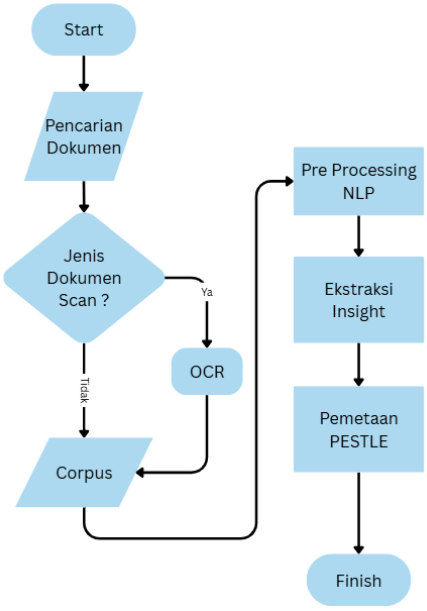
- *Political*: agenda pemerintah, koordinasi lintas lembaga, prioritas program, dan tata kelola.
- *Economic*: biaya pengawasan, stabilitas harga, distribusi, logistik, dan ketahanan pasokan.
- *Social*: perilaku konsumsi, literasi, risiko kesehatan publik, kepercayaan masyarakat.
- *Technological*: sistem informasi, digitalisasi, laboratorium, monitoring, integrasi data.
- Legal: regulasi, standar, sertifikasi, perizinan, kepatuhan, sanksi.
- *Environmental*: sanitasi, kontaminasi, limbah, air, dan risiko iklim.

Dalam konteks keamanan pangan, PESTLE membantu mengubah hasil *text mining* yang "berupa kata" menjadi ringkasan kebijakan yang lebih mudah dipakai untuk perencanaan dan evaluasi.

III. METHODOLOGY

Penelitian ini menggunakan pendekatan *text mining* dan Natural Language Processing (NLP) untuk mengekstraksi *insight* kebijakan dari laporan keamanan pangan. Metodologi dirancang untuk mengolah dokumen teks tidak terstruktur yang bersumber dari laporan resmi pemerintah, sehingga memungkinkan analisis yang sistematis, objektif, dan berbasis data. Kerangka analisis yang digunakan mengombinasikan teknik NLP dengan perspektif PESTLE (*Political, Economic, Social, Technological, Legal, dan Environmental*) guna memetakan isu keamanan pangan secara multidimensional.

Secara umum, tahapan metodologi penelitian ini meliputi: (1) pengumpulan dan seleksi dokumen laporan keamanan pangan, (2) *preprocessing* teks menggunakan teknik NLP, (3) ekstraksi dan representasi fitur teks, (4) pemetaan tematik berbasis perspektif PESTLE, serta (5) analisis dan interpretasi hasil untuk menghasilkan insight kebijakan. Pendekatan ini memungkinkan pengolahan data teks dalam skala besar sekaligus mempertahankan konteks kebijakan yang relevan bagi institusi publik, khususnya Badan Gizi Nasional.



Gambar 1. Alur Proses NLP untuk Ekstraksi Insight

A. Data dan Sumber Data

Data penelitian berupa PDF berbasis teks maupun PDF hasil *scan*, yaitu kumpulan dokumen kebijakan strategis dan laporan resmi terkait keamanan pangan, meliputi:

- Kebijakan Strategis Ketahanan Pangan dan Gizi 2019
- Laporan Kinerja BPOM 2024
- Laporan Kinerja Deputi Bidang Pengawasan Pangan Olahan (BPOM) 2023
- Laporan Penyelidikan Epidemiologi Sementara Keracunan Makanan di Wilayah Desa Kampung Baru Kecamatan Banda 2022
- Laporan Tahunan Balai Besar POM di Jakarta 2023
- Laporan Tahunan Balai Besar POM di Padang 2023
- Laporan Tahunan BPOM 2021
- Laporan Tahunan Direktorat Registrasi Pangan Olahan 2022
- Laporan Tahunan Direktorat Registrasi Pangan Olahan 2023
- Laporan Tahunan Direktorat Registrasi Pangan Olahan 2024
- Laporan Tahunan Layanan Informasi Publik (BPOM) 2016
- Laporan Tahunan Layanan Informasi Publik (BPOM) 2017
- Laporan Tahunan Layanan Informasi Publik (BPOM) 2019
- Laporan Tahunan Layanan Informasi Publik (BPOM) 2023
- Laporan Tahunan Layanan Informasi Publik (BPOM) 2024

- Laporan Tahunan Pejabat Pengelola Informasi dan Dokumentasi (BPOM) 2018
- Laporan Tahunan Pejabat Pengelola Informasi dan Dokumentasi (BPOM) 2020
- Laporan Tahunan Pengelolaan Informasi dan Dokumentasi (BPOM) 2021
- Laporan Tahunan Pengelolaan Informasi dan Dokumentasi (BPOM) 2022
- Laporan Keamanan Pangan - Program MBG-BGN 2025
- Laporan Tahunan Badan POM 2023

B. Preprocessing, Ekstraksi dan Analisis Teks

Untuk memastikan keterbacaan data, tahap ini bertujuan membangun corpus yang bersih dan siap dianalisis dari dokumen keamanan pangan dengan format campuran (PDF text-based, PDF hasil scan, DOCX, TXT).

1. Inventarisasi dokumen multi-format

   Seluruh file dalam folder data dibaca otomatis dengan filter ekstensi .pdf, .docx, .txt, lalu disusun menjadi daftar paths dan tabel df_files. Tujuannya memastikan seluruh sumber data terdokumentasi dan bisa ditelusuri.

2. Deteksi PDF *scanned* vs *text-based*

   Untuk setiap PDF, sistem menghitung text_chars per halaman menggunakan page.extract_text(). Jika proporsi halaman dengan teks sangat rendah (misalnya < 50 karakter) melebihi ambang (misalnya ≥ 70%), dokumen ditandai LIKELY_SCANNED. Mekanisme ini penting agar OCR hanya diterapkan pada dokumen yang memang tidak memiliki layer teks.

3. OCR selektif untuk dokumen *scanned* (Tesseract)

   PDF yang terdeteksi scanned diproses OCR per halaman dengan resolusi tertentu (mis. 250 dpi), lalu hasil OCR disimpan sebagai cache TXT (*_OCR.txt). Caching ini membuat eksperimen lebih efisien karena OCR tidak perlu diulang pada eksekusi berikutnya.

4. Pembangunan corpus final (penggabungan sumber teks)

   Seluruh teks digabung menjadi df_corpus dengan sumber:

   - PDF_TEXT untuk PDF text-based,
   - OCR_TEXT untuk PDF scanned,
   - TXT untuk file teks,
   - DOCX_TEXT (bila fungsi pembaca docx diaktifkan).

   Setiap dokumen disertai metadata dasar (doc_name, doc_path, source, status, n_chars) untuk *traceability*.

5. Pembersihan konten (*content* cleaning) dan pemisahan lampiran

   Agar analisis tidak bias oleh noise, dilakukan pembersihan bertahap:

   - penghapusan baris daftar isi (TOC-like),
   - penghapusan penanda halaman (“Halaman X”, “Page X”, “3/40”),
   - penghapusan header/footer berulang berbasis frekuensi baris (remove_repeated_lines), opsional penghapusan blok tabel teks (ASCII table-like),
   - pemisahan lampiran dari isi utama (split_main_and_annex),
   - normalisasi whitespace.

   Tahap ini memastikan fitur yang diekstrak benar-benar merepresentasikan isi substantif, bukan artefak formatting.

6. Segmentasi paragraf dan metadata tagging

   Teks utama dipecah menjadi paragraf menggunakan adaptive splitter: default double newline, dan jika paragraf terlalu sedikit maka fallback ke penggabungan baris. Setiap paragraf disimpan di df_paras dengan atribut id_dokumen, par_id, paragraf_raw, serta metadata heuristik seperti tanggal/tahun, unit, dan jenis_laporan.

7. Pra-pemrosesan Bahasa Indonesia (Sastrawi)

   Setiap paragraf diproses menjadi paragraf_clean dengan: *lowercase*, penghapusan URL, normalisasi karakter, *stopword* minimal, dan *stemming* Bahasa Indonesia (Sastrawi). Output ini menjadi input utama analisis statistik teks.

8. Analisis tematik: TF-IDF dan LDA

   TF-IDF digunakan untuk mengekstrak kata kunci dominan global dan mendukung interpretasi yang transparan (unigram dan bigram).

   LDA Topic Modeling digunakan untuk menemukan topik laten (mis. 6 topik) dan memberikan topic_id serta topic_score per paragraf.

   Kedua teknik ini dipilih karena lebih mudah dijelaskan ke pengambil kebijakan dibanding model embedding kontekstual pada studi eksploratif tanpa data berlabel.

C. Pemetaan Perspektif PESTLE

Tahap ini bertujuan mengubah hasil *text mining* (kata kunci/topik) menjadi *insight* kebijakan yang terstruktur melalui kerangka PESTLE. Pemetaan dilakukan sebagai baseline yang transparan dan dapat diaudit.

Setiap dimensi PESTLE didefinisikan melalui daftar kata kunci domain-relevan:

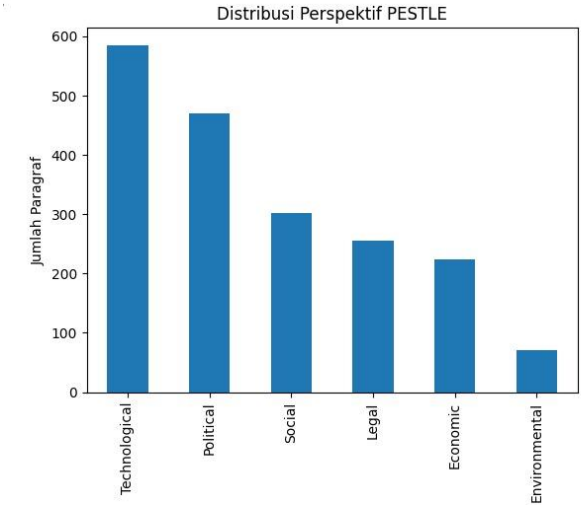
- *Political*: kebijakan, program, pemerintah, koordinasi, strategi, anggaran, nasional, daerah
- *Economic*: harga, biaya, pasokan, distribusi, rantai, supply, inflasi, subsidi, logistik
- *Social*: masyarakat, konsumen, kesehatan, edukasi, literasi, perilaku, komunitas, anak, sekolah
- *Technological*: teknologi, sistem, digital, aplikasi, data, laboratorium, monitoring, otomasi, platform
- Legal: regulasi, aturan, undang, peraturan, standar, sanksi, kepatuhan, sertifikasi, perizinan
- *Environmental*: lingkungan, sanitasi, limbah, iklim, cuaca, kontaminasi, mikroba, higien, air

Setiap paragraf_clean dihitung kecocokan kata kunci terhadap masing-masing dimensi. Label PESTLE dipilih berdasarkan skor tertinggi. Jika tidak ada kecocokan, paragraf diberi label Uncategorized. Pendekatan paragraf (bukan dokumen keseluruhan) dipilih karena kebijakan/temuan sering

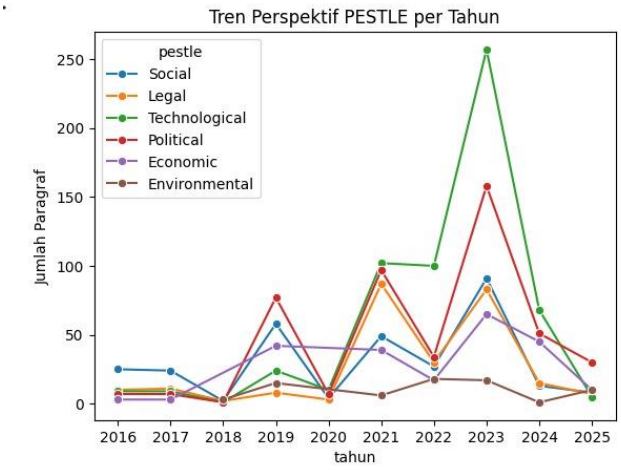
muncul sebagai unit argumentasi dalam paragraf, sehingga lebih granular dan informatif untuk evidence-based insight.

Untuk meningkatkan kualitas interpretasi, notebook menghitung pestle_confidence sebagai jumlah kata kunci yang cocok pada label terpilih. Skor ini dipakai untuk: mengambil paragraf "bukti" dengan *confidence* tertinggi per PESTLE (evidence_by_pestle), meminimalkan bias interpretasi dengan menampilkan kutipan yang paling kuat secara leksikal. Label PESTLE selanjutnya diagregasi: global, per dokumen, per unit, dan per tahun. Output ini memungkinkan: pemetaan fokus kebijakan antar-unit, deteksi tren tema keamanan pangan dari waktu ke waktu, identifikasi gap (marker "belum/perlu") dan penyebab (marker "karena/kendala").

IV. RESULTS AND DISCUSSION



Hasil distribusi PESTLE menunjukkan bahwa perspektif Technological dan Political mendominasi narasi keamanan pangan. Hal ini mengindikasikan bahwa keamanan pangan diposisikan terutama sebagai isu tata kelola sistem dan kebijakan publik, sementara dimensi Economic dan Environmental relatif kurang mendapat perhatian.



Tren temporal menunjukkan peningkatan signifikan dimensi Technological dan Political pada periode 2021–2023, yang mengindikasikan fase penguatan kebijakan dan sistem pengawasan. Sebaliknya, dimensi Environmental cenderung stabil pada level rendah, menandakan keterbatasan integrasi isu lingkungan dalam narasi keamanan pangan.

Pendekatan NLP dengan kerangka PESTLE berhasil mengungkap bahwa diskursus keamanan pangan didominasi oleh perspektif teknologi dan kebijakan publik. Temuan ini menunjukkan kekuatan tata kelola sistem, namun sekaligus menyoroti kebutuhan untuk memperkuat dimensi ekonomi dan lingkungan guna membangun kebijakan keamanan pangan yang lebih berkelanjutan.

V. CONCULISION

Penelitian ini menunjukkan bahwa NLP berbasis unsupervised learning yang dikombinasikan dengan kerangka PESTLE dapat mengekstrak insight kebijakan dari laporan keamanan pangan. Hasil analisis mengungkap dominasi perspektif teknologi dan politik, sekaligus mengidentifikasi area yang relatif kurang mendapat perhatian.

Sebagai rekomendasi, penelitian selanjutnya dapat mengintegrasikan validasi ahli, data tambahan, serta pendekatan *supervised* untuk memperkuat temuan. Pendekatan yang diusulkan berpotensi menjadi alat pendukung pengambilan keputusan berbasis bukti dalam perumusan kebijakan keamanan pangan.

REFERENCES

[1] Undang-Undang Republik Indonesia Nomor 18 Tahun 2012 tentang Pangan. Jakarta, Indonesia: Sekretariat Negara Republik Indonesia, 2012. [Online]. *Available*: https://peraturan.bpk.go.id

[2] Kowsari, K., et al. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[3] Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data. Frontiers in Artificial Intelligence, 3, 42.

[4] Pemerintah Republik Indonesia. (2004). Peraturan Pemerintah No. 28 Tahun 2004 tentang Keamanan, Mutu, dan Gizi Pangan.

[5] Pemerintah Republik Indonesia. (2025). Peraturan Presiden No. 115 Tahun 2025 tentang Tata Kelola Penyelenggaraan Program Makan Bergizi Gratis.

[6] Kementerian Kesehatan Republik Indonesia. (2025). Publikasi/berita terkait penguatan pengawasan keamanan pangan dalam Program MBG.

[7] Badan Pangan Nasional. (2024). Peraturan Badan/ketentuan terkait pengawasan keamanan pangan dan informasi fungsi kelembagaan pada portal resmi.