JSAI : Journal Scientific and Applied Informatics
Vol. 09, No. 2, *June* 2026, *pages*. 213~220
E-ISSN: 2614-3054; P-ISSN: 2614-3062, *accredited by Kemenristekdikti*, Sinta 4
DOI: 10.36085

Pengembangan Sistem Media *Intelligence* ESG Berbasis NLP Bahasa Indonesia Menggunakan TF-IDF dan IndoBERT

[1] Lukman Hakim Moeslich, [2] Cahyono Budy Santoso
[1,2]Universitas Pembangunan Jaya, Indonesia
[1]*Lukman.hakim@student.upj.ac.id* ; [2]*cahyono.budy@upj.ac.id*;

Article Info

Article history:

Received, 2026-05-28
Revised, 2026-06-01
Accepted, 2026-06-03

Kata Kunci:

ESG
media intelligence
analisis teks
IndoBERT
TF-IDF

ABSTRAK

Pemantauan *Environmental, Social, and Governance* (ESG) pada industri pertambangan nikel Indonesia semakin penting seiring meningkatnya tuntutan transparansi dan keberlanjutan industri. Namun, sistem otomatis untuk analisis ESG berbasis media berbahasa Indonesia masih terbatas. Penelitian ini bertujuan mengembangkan sistem media intelligence ESG berbasis *Natural Language Processing* (NLP) untuk menganalisis persepsi media terhadap PT *Indonesia Weda Bay Industrial Park* (IWIP) dan PT *Weda Bay Nickel* (WBN). Sistem dibangun melalui pipeline delapan tahap yang mencakup pengumpulan berita otomatis, praproses teks Bahasa Indonesia, pelabelan berbasis ontologi ESG, klasifikasi teks menggunakan TF-IDF + LinearSVC dan IndoBERT, serta analisis sentimen dan tren risiko ESG. Sebanyak 1.693 artikel berita periode Januari 2020–Mei 2026 berhasil dikumpulkan dan 1.320 artikel berhasil diberi label ESG. Hasil eksperimen menunjukkan TF-IDF terbaik memperoleh *Macro-F1* sebesar 0,7693, sedangkan IndoBERT mencapai 0,7698. Analisis media menunjukkan IWIP memiliki persepsi media yang dominan negatif pada aspek lingkungan dan sosial, sedangkan WBN relatif lebih positif pada aspek tata kelola. Penelitian ini berkontribusi pada pengembangan media intelligence ESG berbahasa Indonesia untuk industri pertambangan.

Keywords:

ESG
media intelligence
text analysis
IndoBERT
TF-IDF

ABSTRACT

Monitoring Environmental, Social, and Governance (ESG) issues in Indonesia's nickel mining industry has become increasingly important due to growing demands for transparency and sustainability. However, automated ESG media analysis for Indonesian-language news remains limited. This study aims to develop an ESG media intelligence system based on Natural Language Processing (NLP) to analyze media perception toward PT Indonesia Weda Bay Industrial Park (IWIP) and PT Weda Bay Nickel (WBN). The proposed system employs an eight-stage pipeline consisting of automated news collection, Indonesian text preprocessing, ontology-based ESG labeling, text classification using TF-IDF + LinearSVC and IndoBERT, as well as sentiment and ESG risk trend analysis. A total of 1,693 news articles published between January 2020 and May 2026 were collected, with 1,320 articles successfully labeled using an ontology-based weak supervision approach. Experimental results show that the best TF-IDF configuration achieved a Macro-F1 score of 0.7693, while IndoBERT achieved 0.7698. The findings indicate that TF-IDF remains competitive with transformer-based models on limited Indonesian ESG datasets. Media analysis revealed that IWIP received predominantly negative media perception on environmental and social issues, while WBN showed relatively more positive governance-related coverage. This research contributes to the development of Indonesian-language ESG media intelligence for the mining industry.

This is an open access article under the CC BY-SAlicense.

Penulis Korespondensi:

Cahyono Budy Santoso,
Program Studi Sistem Informasi ,
Universitas Pembangunan Jaya,
Email: cahyono.budy@upj.ac.id

1. PENDAHULUAN

Sektor pertambangan nikel Indonesia mengalami transformasi eksponensial dalam dekade terakhir, didorong oleh permintaan global terhadap baterai kendaraan listrik dan kebijakan hilirisasi mineral pemerintah. Kabupaten Halmahera Tengah, Provinsi Maluku Utara, kini menjadi salah satu episentrum industri nikel terpenting di dunia dengan kehadiran PT Indonesia Weda Bay Industrial Park (IWIP)—kawasan industri terpadu seluas sekitar tiga ribu hektar—dan PT Weda Bay Nickel (WBN) sebagai salah satu produsen nikel utama. Pertumbuhan industrial yang pesat ini membawa konsekuensi multidimensional: di satu sisi menggerakkan perekonomian daerah dan nasional, namun di sisi lain memunculkan tekanan lingkungan, sosial, dan tata kelola yang signifikan. Kompleksitas inilah yang menjadikan evaluasi keberlanjutan industri nikel sebagai isu yang mendesak untuk dikaji secara sistematis.

Kerangka *Environmental, Social, and Governance* (ESG) telah menjadi alat evaluasi global yang diakui untuk mengukur kinerja keberlanjutan perusahaan. ESG merupakan kerangka tiga pilar yang digunakan *investor*, *regulator*, dan pemangku kepentingan untuk menilai risiko dan peluang keberlanjutan, di mana pilar *Environmental* mencakup emisi karbon, pengelolaan limbah, dan konservasi air; pilar Social menyangkut hubungan dengan tenaga kerja, komunitas lokal, dan rantai pasok; serta pilar *Governance* meliputi transparansi, integritas manajemen, dan kepatuhan regulasi [1]. Di Indonesia, regulasi terkait semakin menguat melalui Peraturan Otoritas Jasa Keuangan (POJK) No. 51/POJK.03/2017 tentang penerapan keuangan berkelanjutan, serta adopsi standar pelaporan Global Reporting Initiative (GRI) oleh perusahaan-perusahaan yang terdaftar di Bursa Efek Indonesia.

Namun demikian, pengukuran ESG berbasis dokumen resmi perusahaan saja tidak mencerminkan persepsi publik yang sesungguhnya. Berbeda dengan laporan ESG resmi yang bersifat *self-reported* dan berpotensi bias, liputan media massa mencerminkan pandangan eksternal dan independen terhadap kinerja perusahaan [2]. Bose et al. (2022) menunjukkan bahwa sentimen media terhadap isu ESG secara signifikan memengaruhi persepsi investor dan nilai pasar perusahaan tambang di pasar berkembang [3], sementara Aerts & Cormier (2009) membuktikan bahwa tekanan media eksternal secara langsung memengaruhi tingkat dan kualitas pengungkapan informasi lingkungan korporat [4]. Fischbach et al. (2023) mengembangkan *ESG-Miner*, sebuah pipeline NLP *end-to-end* untuk menilai kinerja ESG secara otomatis dari data liputan media, dan membuktikan bahwa kanal non-korporat seperti berita merupakan pendorong utama transparansi ESG [5]. Sebelumnya, Srivastava & Pinto (2021) juga telah mengusulkan kerangka awal penilaian ESG otomatis dari teks berita berbasis NLP, meskipun masih terbatas pada corpus berbahasa Inggris [19]. Penting dicatat bahwa liputan media tidak selalu merepresentasikan kondisi ESG objektif perusahaan, melainkan lebih tepat dipahami sebagai representasi persepsi, framing, dan perhatian publik terhadap isu ESG pada periode tertentu.

Dari sisi metode pemrosesan teks, terdapat dua aliran pendekatan yang relevan. Pendekatan pertama berbasis fitur tradisional, yaitu *Term Frequency-Inverse Document Frequency* (TF-IDF) yang memberikan bobot tinggi pada kata yang sering muncul dalam dokumen tertentu tetapi jarang dalam korpus secara keseluruhan [6]. Dalam kombinasi dengan Linear *Support Vector Classifier* (LinearSVC), TF-IDF terbukti efektif untuk klasifikasi teks pendek hingga menengah pada bahasa Indonesia dengan keunggulan komputasi signifikan dibandingkan model transformer [7], khususnya ketika terminologi domain relatif eksplisit. Studi ablasi yang membandingkan representasi kata (unigram/bigram) versus karakter n-gram serta pengaruh stemming PySastrawi penting dilakukan untuk mengidentifikasi konfigurasi optimal pada domain spesifik [8]. Agustin et al. (2025) membuktikan bahwa kombinasi *Naïve Bayes*, TF-IDF, dan SMOTE efektif untuk analisis sentimen opini publik dengan akurasi 80% pada tiga kelas [22], sementara penelitian Tala (2003) tentang stemming bahasa Indonesia meletakkan fondasi algoritma Sastrawi yang penggunaannya pada domain pertambangan perlu dievaluasi karena risiko *over-stemming* istilah teknis [9].

Pendekatan kedua berbasis model bahasa kontekstual. BERT yang diperkenalkan Devlin et al. (2019) merevolusi pemrosesan bahasa alami melalui *pre-training* kontekstual bidireksional pada korpus skala besar [10]. IndoBERT yang dikembangkan Wilie et al. (2020) sebagai bagian dari IndoNLU *Benchmark* merupakan model BERT yang di-pre-train khusus pada korpus bahasa Indonesia berukuran 23,4 GB [11]. *Fine-tuning* IndoBERT untuk tugas klasifikasi domain-spesifik terbukti menghasilkan performa superior dibandingkan TF-IDF pada teks yang memerlukan pemahaman kontekstual jangka panjang [12]; Kuncoro et al. (2022) berhasil menerapkannya untuk klasifikasi sentimen berita keuangan Indonesia dengan Macro-F1 hingga 0,82 [13], dan survei Putra & Winatmoko (2022) mengonfirmasi IndoBERT sebagai model terkuat untuk sebagian besar tugas klasifikasi teks berbahasa Indonesia [21]. Tantangan utama penggunaannya adalah kebutuhan komputasi tinggi dan ketersediaan data berlabel yang terbatas.

Selain klasifikasi, implementasi sistem intelijen media memerlukan dukungan teknik pengumpulan dan pemodelan data yang andal. Pengumpulan berita otomatis melalui web scraping menghadapi tantangan heterogenitas struktur HTML antar-portal, *paywall*, deteksi bot, dan inkonsistensi metadata [14], sehingga

pendekatan berbasis kata kunci yang dikombinasikan dengan mesin telusur atau API berita terbukti lebih *scalable* dibandingkan *crawling* langsung [15]. Deduplikasi konten menjadi tantangan kritis mengingat praktik re-publikasi dan sindikasi artikel, dan penggunaan *cosine similarity* berbasis TF-IDF dengan *threshold* 0,90 merupakan pendekatan tervalidasi untuk mendeteksi artikel hampir-duplikat [16]. Pendekatan analisis teks untuk intelijen institusional yang sebelumnya diterapkan di domain keuangan [20] diadaptasi dalam penelitian ini untuk konteks ESG media berbahasa Indonesia. Pada tahap pelabelan, pendekatan berbasis ontologi menawarkan keunggulan interpretabilitas dan kontrol semantik dibandingkan anotasi berbasis pembelajaran mesin semata, karena ontologi mendefinisikan hierarki konseptual yang menghubungkan *pilar*, *subtheme*, dan kata kunci spesifik domain [17], serta memungkinkan pemodelan hubungan antar-konsep dan penanganan sinonim yang tidak dapat dilakukan kamus kata kunci sederhana [18]. Untuk mengeksplorasi tema laten secara tak-tersupervisi, *Latent Dirichlet Allocation* (LDA) yang diperkenalkan Blei et al. (2003) mengasumsikan setiap dokumen sebagai campuran topik laten [23]; implementasinya menggunakan *library* Gensim [24] dan telah terbukti efektif mengidentifikasi tema dalam *corpus* berita berskala besar [25].

Berdasarkan tinjauan tersebut, kesenjangan utama yang teridentifikasi adalah ketiadaan sistem otomatis untuk pemantauan ESG berbasis media berbahasa Indonesia, khususnya bagi kegiatan industri dan pertambangan nikel di kawasan Halmahera Tengah. Sistem pemantauan yang ada umumnya bersifat manual, tidak terstruktur, dan tidak mampu memproses volume berita yang terus bertumbuh secara efisien. Di samping itu, karakteristik unik bahasa Indonesia—morfologi yang kaya, penggunaan afiks yang kompleks, dan keragaman dialek regional—menyulitkan penerapan langsung model NLP berbasis bahasa Inggris. Sistem media intelligence ESG yang diusulkan tidak dimaksudkan menggantikan ESG rating formal, melainkan sebagai instrumen pelengkap untuk mendeteksi sinyal risiko, sinyal reputasi, dan isu sosial secara lebih dinamis.

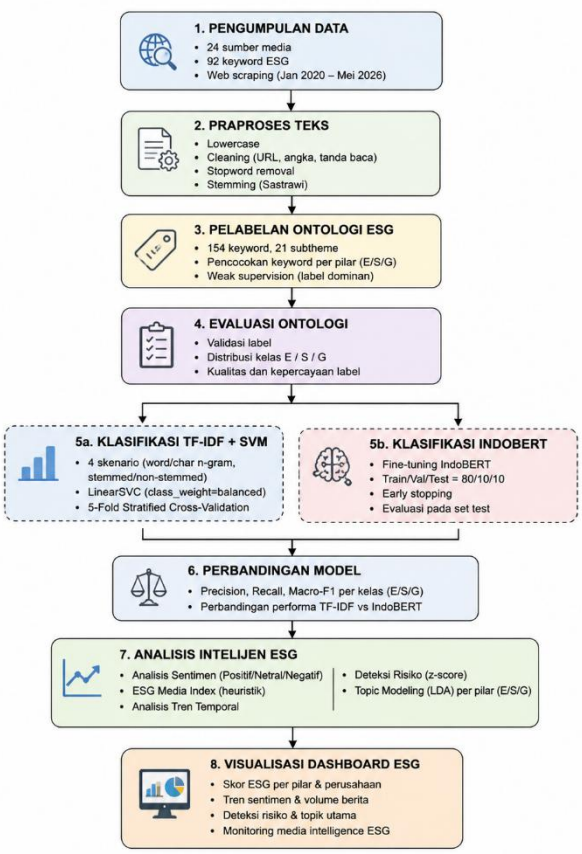
Dalam penelitian ini, media *intelligence* ESG didefinisikan sebagai proses sistematis untuk mengumpulkan, menganalisis, dan menginterpretasikan informasi media massa guna mengidentifikasi persepsi, sentimen, dan sinyal risiko terkait ESG perusahaan menggunakan pendekatan NLP. Sebagian besar penelitian *automated* ESG *assessment* masih berfokus pada corpus berbahasa Inggris dan belum banyak diterapkan pada industri pertambangan Indonesia menggunakan corpus berita Bahasa Indonesia.

Penelitian ini bertujuan untuk: (1) membangun pipeline otomatis pengumpulan berita multi-sumber dari 24 portal media Indonesia yang mencakup kategori Regional-Maluku Utara, Regional-Lainnya, Nasional, dan ESG-Specialist; (2) mengembangkan sistem pelabelan ESG berbasis ontologi dengan 154 kata kunci yang mencakup tiga pilar (E/S/G) dan 21 sub-tema; (3) membandingkan performa klasifikasi teks TF-IDF + LinearSVC (studi ablasi empat kondisi) dengan IndoBERT yang telah disetel spesifik (fine-tuned) untuk domain pertambangan nikel; dan (4) menghasilkan skor ESG media, analisis tren temporal, deteksi sinyal risiko, dan visualisasi dashboard berbasis Python untuk PT IWIP dan PT WBN. Kontribusi penelitian ini terletak pada tiga aspek orisinalitas, yaitu fokus domain spesifik pada industri nikel Indonesia dengan data berbahasa Indonesia, arsitektur alur proses menyeluruh yang mengintegrasikan pengumpulan data, pelabelan, klasifikasi, dan intelijen dalam satu sistem kohesif, serta evaluasi empiris dual-model (TF-IDF vs. IndoBERT) pada corpus domain-spesifik yang belum pernah dilakukan sebelumnya untuk konteks pertambangan nikel di Maluku Utara.

Penelitian ini tidak hanya melakukan klasifikasi ESG, tetapi juga mengintegrasikan pengumpulan berita otomatis, ontologi ESG, analisis sentimen, deteksi risiko temporal, dan *topic modeling* dalam satu *pipeline* terintegrasi.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan pipeline sekuensial yang terdiri atas beberapa tahapan utama untuk membangun sistem media intelligence ESG berbasis *Natural Language Processing* (NLP). Pipeline dirancang secara modular agar setiap tahap dapat dijalankan secara independen sehingga mendukung reproduktibilitas penelitian. Implementasi sistem menggunakan Python 3.10 dengan beberapa library utama, yaitu BeautifulSoup4 untuk *web scraping*, PySastrawi untuk praproses Bahasa Indonesia, *Scikit-learn* untuk klasifikasi TF-IDF + LinearSVC, HuggingFace *Transformers* untuk IndoBERT, serta Matplotlib dan *Seaborn* untuk visualisasi hasil.



Gambar 1. Alur Penelitian

Tahap awal penelitian dilakukan melalui pengumpulan berita dari 24 sumber media lokal dan nasional pada periode Januari 2020 hingga Mei 2026 menggunakan 92 *keyword* terkait ESG dan industri nikel. Proses pengumpulan dilengkapi dengan filter kualitas berupa panjang minimum artikel, deteksi bahasa otomatis, dan deduplikasi menggunakan *cosine similarity* TF-IDF. Dari proses tersebut diperoleh 1.693 artikel unik yang selanjutnya diproses menggunakan tahapan NLP Bahasa Indonesia, meliputi *lowercase*, pembersihan teks, stopword removal, dan stemming menggunakan algoritma Sastrawi.

Pelabelan artikel dilakukan menggunakan pendekatan ontologi ESG berbasis *keyword* yang terdiri atas 154 *keyword* dan 21 *subtheme* untuk mengklasifikasikan artikel ke dalam kategori *Environmental* (E), *Social* (S), dan *Governance* (G). Ontologi disusun berdasarkan literatur ESG dan berita pertambangan Indonesia, kemudian divalidasi secara manual pada sebagian sampel artikel untuk memastikan kesesuaian konteks. Pelabelan menggunakan pendekatan *weak supervision* sehingga label yang dihasilkan merepresentasikan kategorisasi semi-otomatis berbasis ontologi dan belum sepenuhnya menggantikan anotasi manual oleh pakar domain.

Tahap klasifikasi dilakukan menggunakan dua pendekatan, yaitu TF-IDF + LinearSVC dan IndoBERT. Eksperimen TF-IDF menggunakan empat skenario kombinasi *word/character* n-gram dan stemming/non-stemming dengan evaluasi 5-fold stratified *cross-validation* menggunakan metrik *Macro Precision*, *Recall*, dan *F1-score*. Ketidakseimbangan distribusi kelas ditangani menggunakan parameter class_weight=balanced pada LinearSVC. Sementara itu, model IndoBERT di-fine-tune menggunakan skema train/validation/test split karena keterbatasan komputasi pada proses transformer berbasis CPU. Evaluasi IndoBERT menggunakan Macro-F1 sebagai metrik utama sehingga hasil kedua model diposisikan sebagai indikasi performa relatif dan bukan perbandingan absolut.

Tahap akhir berupa analisis intelijen ESG yang mencakup analisis sentimen berbasis leksikon, penghitungan *heuristic* ESG media *index*, analisis tren temporal, deteksi risiko berbasis *z-score*, dan *topic modeling* menggunakan Latent *Dirichlet Allocation* (LDA). Seluruh hasil analisis divisualisasikan dalam bentuk dashboard ESG untuk memantau persepsi media terhadap perusahaan pertambangan. Dalam penelitian ini, skor ESG yang dihasilkan tidak dimaksudkan sebagai ESG rating formal, melainkan sebagai indikator persepsi media terhadap isu keberlanjutan perusahaan berdasarkan pemberitaan media massa.

3. HASIL DAN PEMBAHASAN

Proses pengumpulan data berhasil memperoleh 1.693 artikel berita unik terkait PT IWIP dan PT WBN dari 24 sumber media lokal dan nasional periode Januari 2020 hingga Mei 2026. Mayoritas artikel berasal dari media lokal Maluku Utara (51,8%), diikuti media regional lain dan media nasional. Dominasi media lokal menunjukkan bahwa isu pertambangan di Halmahera Tengah lebih banyak diberitakan oleh media daerah yang memiliki kedekatan langsung dengan kondisi lapangan. Setelah proses pelabelan berbasis ontologi ESG, sebanyak 1.320 artikel berhasil diklasifikasikan ke dalam kategori *Environmental* (E), *Social* (S), dan *Governance* (G). Pilar sosial menjadi kategori paling dominan, yang menunjukkan bahwa isu ketenagakerjaan, konflik sosial, demonstrasi buruh, dan hubungan masyarakat merupakan fokus utama pemberitaan media terkait industri nikel di kawasan tersebut.

Tabel 1. Statistik Dataset Penelitian

Kategori	Jumlah
Total artikel unik	1.693
Artikel berlabel ESG	1.320
Pilar *Social* (S)	557
Pilar *Governance* (G)	323
Pilar *Environmental* (E)	155
Multi-pilar ESG	285

Eksperimen klasifikasi menunjukkan bahwa pendekatan TF-IDF + LinearSVC maupun IndoBERT memiliki performa yang relatif kompetitif pada tugas klasifikasi ESG berbahasa Indonesia. Hasil terbaik TF-IDF diperoleh pada konfigurasi character n-gram dengan stemming yang mencapai *Macro-F1* sebesar 0,7693, sedangkan IndoBERT memperoleh *Macro-F1* sebesar 0,7698. Perbedaan performa keduanya relatif kecil, yang mengindikasikan bahwa pendekatan berbasis fitur tradisional masih efektif pada dataset ESG berukuran terbatas. Selain itu, stemming menggunakan PySastrawi terbukti meningkatkan performa klasifikasi secara konsisten karena mampu menormalisasi variasi morfologis Bahasa Indonesia. Namun, evaluasi kedua model menggunakan skema yang berbeda, yaitu *cross-validation* pada TF-IDF dan hold-out split pada IndoBERT, sehingga hasil tidak sepenuhnya dapat dibandingkan secara langsung.

Tabel 2. Perbandingan Performa Model

Model	Macro-F1
TF-IDF + LinearSVC	0,7693
IndoBERT	0,7698

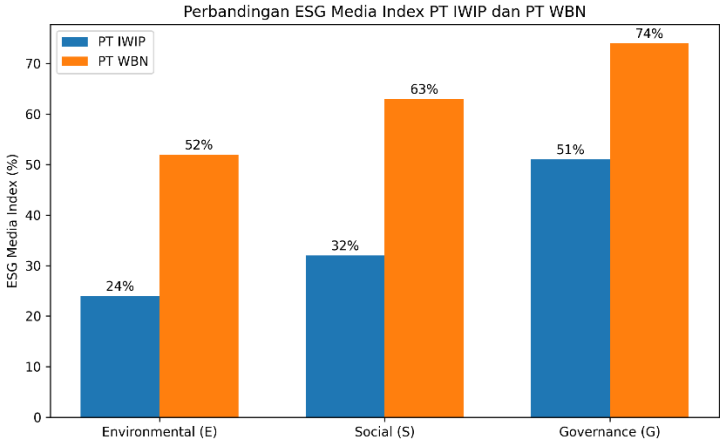
Analisis lebih lanjut menunjukkan bahwa kelas Environmental memiliki performa klasifikasi yang relatif lebih rendah dibandingkan Social karena jumlah data yang lebih sedikit dan variasi isu lingkungan yang lebih kompleks. Sebaliknya, kelas Social lebih mudah dikenali karena banyak menggunakan istilah eksplisit seperti demonstrasi, buruh, PHK, dan masyarakat. Temuan ini menunjukkan bahwa ketidakseimbangan distribusi kelas masih menjadi tantangan dalam klasifikasi ESG berbahasa Indonesia.

Analisis persepsi media menunjukkan perbedaan profil ESG antara PT IWIP dan PT WBN. PT IWIP memperoleh skor rendah pada aspek lingkungan dan sosial akibat dominasi pemberitaan terkait pencemaran lingkungan, banjir industri, konflik tenaga kerja, dan demonstrasi buruh. Sebaliknya, PT WBN memperoleh persepsi media yang relatif lebih positif, terutama pada aspek tata kelola dan program sosial perusahaan. Dalam penelitian ini, skor ESG diposisikan sebagai heuristic ESG media *index* atau indikator persepsi media, bukan sebagai ESG rating formal perusahaan.

Tabel 3. ESG Media Index

Perusahaan	E	S	G
PT IWIP	0,24	0,32	0,51
PT WBN	0,52	0,63	0,74

Analisis sentimen menunjukkan bahwa pilar sosial memiliki volume berita terbesar dengan dominasi sentimen negatif yang berkaitan dengan isu ketenagakerjaan dan konflik sosial. Pilar lingkungan juga didominasi sentimen negatif, terutama terkait pencemaran Sungai Sagea dan dampak ekologis aktivitas tambang. Sementara itu, pilar tata kelola menunjukkan distribusi sentimen yang lebih seimbang karena pemberitaan mencakup aspek regulasi, penghargaan perusahaan, dan hubungan dengan pemerintah daerah. Namun demikian, pendekatan sentimen berbasis leksikon masih memiliki keterbatasan dalam memahami konteks ironi, negasi, dan istilah teknis pertambangan.



Gambar 2. Grafik ESG Media

Sistem deteksi risiko berhasil mengidentifikasi beberapa periode dengan lonjakan sentimen negatif menggunakan pendekatan *z-score*. Risiko lingkungan tertinggi terdeteksi pada September 2023 saat meningkatnya pemberitaan mengenai pencemaran Sungai Sagea. Pada aspek sosial, lonjakan risiko terjadi pada Juli 2024 dan sepanjang tahun 2025 akibat demonstrasi buruh dan isu PHK. Sementara itu, aspek tata kelola mengalami lonjakan signifikan pada akhir 2025 yang berkaitan dengan isu pajak dan investigasi regulasi perusahaan. Temuan ini menunjukkan bahwa analisis temporal mampu membantu identifikasi dini terhadap peningkatan risiko reputasi perusahaan di media.

Topic modeling menggunakan *Latent Dirichlet Allocation* (LDA) menunjukkan bahwa isu lingkungan didominasi topik pencemaran sungai, kerusakan lingkungan, dan konflik lahan. Pada pilar sosial, topik utama berkaitan dengan masyarakat adat, hubungan industrial, keselamatan kerja, dan pembangunan ekonomi daerah. Sementara itu, topik tata kelola berfokus pada regulasi pertambangan, perizinan, perpajakan, dan kebijakan hilirisasi industri nikel. Hasil *topic modeling* juga menunjukkan konsistensi dengan ontologi ESG yang dirancang, sehingga mendukung validitas kerangka pelabelan yang digunakan dalam penelitian.

Dominasi media lokal Maluku Utara memberikan implikasi penting terhadap hasil analisis. Media lokal cenderung memiliki kedekatan dengan isu lapangan sehingga mampu menangkap dinamika sosial dan lingkungan secara lebih detail dibandingkan media nasional. Namun, dominasi tersebut juga berpotensi menghasilkan framing yang lebih kuat terhadap konflik sosial dan isu lingkungan tertentu. Selain itu, keterbatasan jumlah media ESG spesialis menunjukkan bahwa ekosistem media ESG berbahasa Indonesia pada sektor pertambangan masih relatif terbatas dan perlu dikembangkan lebih lanjut pada penelitian mendatang.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem media intelligence ESG berbasis Python untuk menganalisis persepsi media terhadap PT IWIP dan PT WBN di Halmahera Tengah melalui *pipeline* yang mencakup pengumpulan data, praproses NLP Bahasa Indonesia, pelabelan ontologi ESG, klasifikasi teks, dan analisis intelijen ESG. Sistem berhasil memproses 1.693 artikel unik dari 24 sumber media, dengan dominasi media lokal Maluku Utara sebesar 51,8% yang menunjukkan kuatnya pengaruh perspektif lokal dalam pemberitaan industri pertambangan. Hasil klasifikasi menunjukkan bahwa IndoBERT memperoleh performa terbaik dengan Macro-F1 sebesar 76,98%, sedikit lebih tinggi dibandingkan TF-IDF + LinearSVC sebesar 76,93%, sehingga pendekatan TF-IDF tetap kompetitif pada dataset ESG berukuran terbatas dengan biaya komputasi lebih rendah. Analisis ESG media menunjukkan bahwa PT IWIP memiliki persepsi media negatif pada aspek lingkungan (24%) dan sosial (32%), sedangkan PT WBN memperoleh persepsi yang relatif lebih positif terutama pada aspek tata kelola (74%). Sistem deteksi risiko juga berhasil mengidentifikasi beberapa periode peningkatan sentimen negatif yang berkaitan dengan isu lingkungan, konflik sosial, dan tata kelola perusahaan. Meskipun demikian, penelitian ini masih memiliki keterbatasan pada jumlah dataset berlabel, penggunaan sentiment analysis berbasis leksikon, validasi ontologi yang belum sepenuhnya manual, serta evaluasi model yang menggunakan skema berbeda. Oleh karena itu, penelitian selanjutnya disarankan menggunakan dataset yang lebih besar, validasi *multi-annotator*, dan model bahasa Indonesia yang lebih mutakhir untuk meningkatkan akurasi analisis ESG berbasis media.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada seluruh portal media yang informasinya digunakan sebagai data penelitian ini. Penulis juga berterima kasih kepada komunitas pengembang IndoBERT (*indobenchmark*) atas

tersedianya model bahasa Indonesia pre-trained secara terbuka, serta kepada komunitas PySastrawi dan Scikit-learn atas kontribusi library open-source yang digunakan dalam pipeline ini. Penelitian ini dilaksanakan secara mandiri tanpa pendanaan institusi eksternal.

REFERENSI

[1] Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), 210–233. https://doi.org/10.1080/20430795.2015.1118917

[2] Serafeim, G., & Yoon, A. (2022). Which corporate ESG news does the market react to? Financial Analysts Journal, 78(1), 59–78. https://doi.org/10.1080/0015198X.2021.1973879

[3] Bose, S., Dong, G., & Simpson, A. (2019). The financial ecosystem: The role of finance in achieving sustainability. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-05624-7

[4] Aerts, W., & Cormier, D. (2009). Media legitimacy and corporate environmental communication. Accounting, Organizations and Society, 34(1), 1–27. https://doi.org/10.1016/j.aos.2008.02.005.

[5] Fischbach, J., Adam, M., Dzhagatspanyan, V., Mendez, D., Frattini, J., Kosenkov, O., & Elahidoost, P. (2023). Automatic ESG assessment of companies by mining and evaluating media coverage data: NLP approach and tool. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), pp. 2823–2830. https://doi.org/10.1109/BIGDATA59044.2023.10386488

[6] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

[7] Murfi, H., Suhartono, D., & Heryadi, Y. (2022). Indonesian news text classification using TF-IDF and support vector machine. In Proceedings of the 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), pp. 1–6. https://doi.org/10.1109/ICIMCIS56813.2022.10083201

[8] Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94 (pp. 161–175).

[9] Tala, F. Z. (2003). A study of stemming effects on information retrieval in Bahasa Indonesia (Master's thesis). Universiteit van Amsterdam.

[10] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

[11] Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., He, X., Ghifari, Aky., ... & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In Proceedings of AACL-IJCNLP 2020 (pp. 843–857).

[12] Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., He, X., Ghifari Aky., ... & Purwarianti, A. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In Proceedings of EMNLP 2021 (pp. 8875–8898).

[13] Andika, F., Nugraheni, T. E., & Susanto, A. (2022). Analisis sentimen ulasan berbahasa Indonesia menggunakan fine-tuning IndoBERT dan R-CNN. ILKOM Jurnal Ilmiah, 14(3), 238–247. https://doi.org/10.33096/ilkom.v14i3.1195

[14] Mitchell, R. (2018). Web scraping with Python: Collecting more data from the modern web (2nd ed.). O'Reilly Media.

[15] Hamborg, F., Meuschke, N., Breitinger, C., & Gipp, B. (2017). news-please: A generic news crawler and extractor. In Proceedings of the 15th International Symposium of Information Science (ISI 2017), pp. 218–223. https://doi.org/10.18452/1447

[16] Broder, A. Z. (2000). Identifying and filtering near-duplicate documents. In Proceedings of CPM 2000, Lecture Notes in Computer Science, 1848, 1–10.

[17] Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05.

[18] Kang, C., Du, X., Wang, H., & Luo, L. (2022). ESG ontology for structured ESG reporting. In Proceedings of the 31st International Conference on Information and Knowledge Management (pp. 4105–4109). https://doi.org/10.1145/3511808.3557416

[19] Srivastava, A., & Pinto, M. (2021). Towards automated ESG scoring of corporates from news. In Proceedings of the 2021 ACM SIGKDD Workshop on Machine Learning for Finance (pp. 1–8).

[20] Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text mining for central banks (Staff Working Paper No. 550). Bank of England.

[21] Putra, G. D., & Winatmoko, Y. (2022). Indonesian language natural language processing: State of the art and future directions. TELKOMNIKA, 20(3), 661–672.

[22] Agustin, Y. H., Mulyani, N. C., & Prasetya, W. S. (2025). Analisis sentimen opini publik menggunakan algoritma Naive Bayes dan TF-IDF. Jurnal Algoritma, 22(2), 1373–1384. https://doi.org/10.33364/algoritma/v.22-2.2671

[23] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

[24] Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45–50). European Language Resources Association.

[25] Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism, 4(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271