Seminar Nasional Sains Data 2025 (SENADA 2025)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Pipeline ETL Terdistribusi untuk Klasifikasi Berita *Clickbait* dan Topik Berita

Gesang Nur Zamroji, Rafly Anugrah Syahputra, Sofia Zahira Rohman, Yuliani Puji Astuti, Ibnu Febry Kurniawan
Departemen Sains Data, Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya
{gesang.23145, rafly.23195, sofia.23197}@mhs.unesa.ac.id, yulianipuji@unesa.ac.id
Corresponding author email: ibnufebry@unesa.ac.id

Abstract: *This research aims to design and implement an efficient distributed ETL pipeline in the classification of clickbait news and news topics, with a focus on managing large-scale news data in the digital era. The developed system utilizes RabbitMQ as a message broker and Celery as a task queue manager for parallel and asynchronous processing. The ETL process includes data extraction from CSV files, data transformation through preprocessing, clickbait and non-clickbait news classification, and topic classification using TF-IDF feature engineering and Random Forest model. The results show that the system can classify news into clickbait and topic categories efficiently while handling large volumes of data. The developed system is also capable of storing the classification results in a PostgreSQL database, enabling further analysis and monitoring. This research contributes to the development of modular and distributed ETL-based systems and introduces an approach that can be used for large-scale analysis of digital news data.*
Keywords: *data pipeline, insight analysis, distributed processing, text processing, clickbait analysis*

Abstrak: Penelitian ini bertujuan untuk merancang dan mengimplementasikan *pipeline ETL* terdistribusi yang efisien dalam klasifikasi berita *clickbait* dan topik berita, dengan fokus pada pengelolaan data berita dalam skala besar di era digital. Sistem yang dikembangkan memanfaatkan *RabbitMQ* sebagai *message broker* dan *Celery* sebagai *task queue manager* untuk pemrosesan paralel dan *asynchronous*. Proses *ETL* meliputi ekstraksi data dari file CSV, transformasi data melalui tahap *pre-processing*, klasifikasi berita *clickbait* dan *non-clickbait*, serta klasifikasi topik menggunakan *feature engineering TF-IDF* dan model *Random Forest*. Hasil penelitian menunjukkan bahwa sistem ini dapat mengelompokkan berita ke dalam kategori *clickbait* dan topik dengan efisien, sambil menangani volume data yang besar. Sistem yang dikembangkan juga mampu menyimpan hasil klasifikasi ke dalam *database PostgreSQL*, memungkinkan analisis dan pemantauan lebih lanjut. Penelitian ini memberikan kontribusi pada pengembangan sistem berbasis *ETL* yang modular dan terdistribusi, serta memperkenalkan pendekatan yang dapat digunakan untuk analisis data berita digital dalam skala besar.
Kata kunci: *data pipeline, insight analysis, distributed processing, text processing, clickbait analysis*

I. PENDAHULUAN

Dalam era informasi digital saat ini, konsumsi berita *online* mengalami peningkatan yang sangat pesat. Seiring dengan perkembangan tersebut, fenomena penyebaran berita yang bersifat *clickbait* juga semakin meluas. *Clickbait* adalah judul konten yang dibuat untuk menarik perhatian pembaca untuk mengklik tautan dan mengarahkannya ke artikel terkait [1]. Penelitian menunjukkan bahwa judul *clickbait* secara signifikan dapat menurunkan kredibilitas berita di mata pembaca, dengan pengaruh yang bervariasi tergantung pada faktor usia dan tingkat rasa ingin tahu pembaca [2].

Sebaliknya, karena volume data berita yang terus meningkat, diperlukan sistem pengelolaan data yang efektif dan otomatis, khusunya untuk tujuan integrasi dan penyimpanan data skala besar. [3]. Karakteristik data berita yang tidak terstruktur, dinamis, dan terus berubah memperkuat masalah ini. Oleh karena itu, pendekatan berbasis *Extract*, *Transform*, and *Load* (ETL) diperlukan, yang memiliki kemampuan untuk menangani data secara efisien dan siap untuk digunakan dalam berbagai jenis analisis lanjutan.

ETL merupakan tahapan sangat penting dalam proses integrasi data, yang mencakup ekstraksi data dari berbagai sumber, transformasi untuk pembersihan dan penyesuaian, serta pemuatan ke dalam sistem penyimpanan terstruktur seperti *Data Warehouse* [5]. Dalam konteks big data, arsitektur ETL konvensional yang berjalan secara terpusat seringkali menghadapi kendala dalam performa dan

skalabilitas. Oleh sebab itu, dibutuhkan sistem ETL yang bersifat terdistribusi dan mendukung pemrosesan paralel, agar proses pengelolaan data tetap efisien meskipun beban kerja terus meningkat [4].

Penelitian-penelitian sebelumnya cenderung hanya berfokus pada penerapan *machine learning* untuk klasifikasi konten berita seperti *clickbait* atau berita palsu [6,7]. Namun, aspek pengelolaan data secara menyeluruh mulai dari akuisisi, pembersihan, hingga penyimpanan sering kali belum mendapat perhatian yang memadai dalam kerangka arsitektur sistem yang utuh dan skalabel. Hal ini menciptakan kesenjangan dalam pengembangan sistem berbasis data yang modular dan siap operasional dalam skala besar.

Oleh karena itu, penelitian ini mengembangkan sistem ETL terdistribusi yang dapat menangani data berita secara otomatis, paralel, dan efisien. Sistem ini dibangun menggunakan RabbitMQ sebagai *message broker* dan Celery sebagai *task queue manager*, dua komponen yang umum digunakan dalam arsitektur sistem terdistribusi modern. Dengan pendekatan ini, proses ETL dibagi menjadi tugas-tugas independen yang dapat dieksekusi secara *asynchronous*, sehingga setiap tahapan dapat dilakukan secara paralel dan skalabel. Hal ini memungkinkan sistem untuk menangani data berita dalam jumlah besar dan terus berkembang, sekaligus menyediakan fondasi yang kuat untuk analisis konten lanjutan. Dengan fokus pada perancangan *pipeline* ETL yang modular, terdistribusi, dan efisien, penelitian ini diharapkan dapat memberikan kontribusi penting dalam pengelolaan data berita digital. Sistem ini juga membuka peluang penerapan lebih luas untuk berbagai kebutuhan *data-driven*, seperti pelaporan berbasis data, sistem intelijen media, dan analitik konten dalam skala besar.

II. METODE PENELITIAN

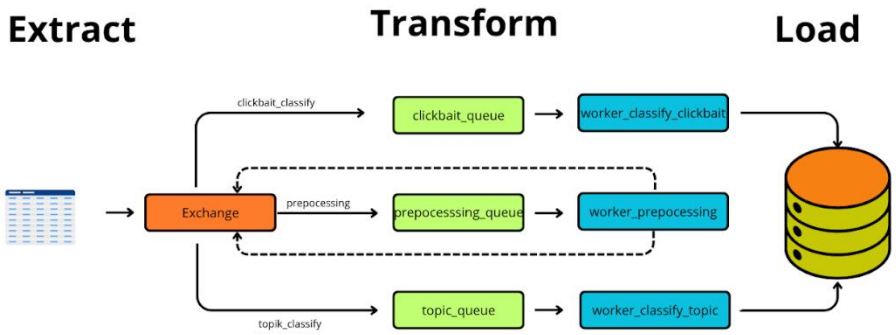
Penelitian ini menggunakan pendekatan kuantitatif dengan metode rekayasa sistem untuk merancang dan menguji *pipeline* ETL terdistribusi dalam klasifikasi berita *clickbait* dan topik berita. Sistem ini dibangun menggunakan RabbitMQ sebagai *message broker* dan Celery sebagai *task queue manager* untuk mengelola proses ETL secara *asynchronous* dan paralel. Proses ETL mencakup tahap *extract* dari file CSV hasil scraping, *transform* yang terdiri dari *preprocessing* teks, klasifikasi *clickbait*, dan klasifikasi topik menggunakan *feature engineering* TF-IDF untuk vektorisasi teks dan model Random Forest, serta tahap *load* untuk menyimpan hasil *transform* ke dalam PostgreSQL. Evaluasi dilakukan dengan mengamati aliran data dan performa task guna memastikan sistem mampu menangani data berita secara efisien dan skalabel.

II.1. Sumber Data

Penelitian ini menggunakan tiga dataset utama: Pertama, dataset untuk melatih model klasifikasi *clickbait* yang berasal dari sumber *open source* kaggle yaitu *CLICK-ID: A novel dataset for Indonesian clickbait headlines*. Kedua, dataset pelatihan model klasifikasi topik yang bersumber dari github. Terakhir, data untuk melakukan simulasi pada sistem yang didapat dari hasil scraping pada tiga *platform* berita yaitu detik.com, CNN indonesia dan Kompas.com.

E-ISSN 2808-5841
P-ISSN 2808-7283

II.2. Desain dan Arsitektur Sistem



Gambar 1. Desain dan Arsitektur Sistem

Diagram sistem di atas menggambarkan alur pemrosesan data berita untuk keperluan klasifikasi *clickbait* dan klasifikasi topik berita dengan menggunakan RabbitMQ sebagai *message broker*. Proses dimulai ketika data berita dikirimkan ke RabbitMQ melalui *direct exchange* dengan *routing key* "preprocessing", yang kemudian diteruskan ke *preprocessing_queue* dan diproses oleh worker_processing yang bertugas membersihkan data berita. Selanjutnya, data berita yang telah dibersihkan diteruskan ke dua queue yang berbeda, yaitu *clickbait_queue* dengan *routing key* "clickbait_classify" dan topik_queue dengan *routing key* "topik_classify". Data pada masing-masing *queue* kemudian diproses oleh *worker* yang sesuai untuk menghasilkan label klasifikasi *clickbait* dan topik berita.Terakhir, data berita asli, data yang telah dibersihkan, serta label hasil klasifikasi *clickbait* dan topik akan disimpan ke dalam database.

II.3. Proses ETL

2.3.1. Extract

Pada tahap ini, sistem melakukan proses ekstraksi data dari file CSV untuk dimasukkan ke dalam *pipeline* ETL. Proses ekstraksi dimulai dengan membaca file CSV secara baris per baris menggunakan skrip Python. Setiap baris mewakili satu entri berita yang kemudian dikemas daam format *JSON*. Data yang telah dikemas kemudian akan dikirim ke *message broker* (Rabbitmq) melalui *direct exchange* menggunakan *routing key* yang telah ditentukan sebelumnya, yakni ”preprocessing”. Data ini kemudian akan masuk ke dalam *preprocessing_queue* untuk diproses lebih lanjut.

2.3.2. Transform

Setelah data diekstraksi, data akan melalui tiga tahapan utama dalam proses transformasi. Proses ini mencakup langkah-langkah untuk mempersiapkan data, mengklasifikasikan *clickbait*, dan mengidentifikasi topik berita. Berikut adalah tahapan yang dijelaskan lebih rinci:

2.3.2.1. Pre-Processing

Pada tahap ini, data berita yang sudah diekstraksi melalui proses pembersihan dan persiapan agar siap untuk diproses lebih lanjut. Langkah-langkah utama dalam *pre-processing* adalah:

- Lowercasing: Mengubah semua huruf menjadi non-kapital
- Remove Digits: Menghapus angka dari teks

E-ISSN 2808-5841
P-ISSN 2808-7283

- Remove Non-ASCII Karakter: Menghapus karakter yang tidak termasuk dalam set karakter ASCII standar.
- Remove punctuation : Menghapus tanda baca , seperti "!#@!"
- Reduce repeated characters: Mengurangi karakter yang terulang secara berlebihan menjadi satu, contohnya "sayaaaaaaaa" diubah menjadi "saya".
- Tokenize: Teks berita dipecah menjadi token, yaitu kata-kata atau frasa yang memiliki makna. Tokenisasi ini penting agar model machine learning dapat menganalisis kata-kata atau frasa secara terpisah.
- Stopword Removal: Menghapus kata penghubung atau kata sambung, seperti di, ke, dari.
- Stemming: Mengubah kata menjadi bentuk dasarnya, misalnya mengubah kata "berlari" menjadi "lari", untuk mengurangi kompleksitas kata yang digunakan dalam teks.

Setelah tahap *Pre-processing* ini selesai, data siap untuk dibagi ke dalam dua jalur pemrosesan berikutnya.

2.3.2.2. Klasifikasi *Clickbait*

Pada jalur pertama, data yang telah diproses melalui *pre-processing* akan dianalisis untuk menentukan apakah berita tersebut merupakan *clickbait* atau tidak.Proses ini menggunakan model *machine learning* yang telah dilatih sebelumnya dengan teknik *TF-IDF* untuk vektorisasi teks dan *Random Forest* untuk klasifikasi, pada Tabel 1 dibawah disajikan metrik evaluasi performa model hasil training model.

Tabel 1. Tabel Hasil Klasifikasi Clickbait

Label Berita	Precision	Recall	F1-Score
clickbait	0.76	0.87	0.81
Non-clickbait	0.78	0.63	0.70

2.3.2.3. Klasifikasi Topik Berita

Jalur kedua dari transformasi adalah mengidentifikasi topik berita yang telah diproses sebelumnya. Proses ini membantu mengkategorikan berita ke dalam topik-topik yang lebih spesifik. Proses ini menggunakan model *machine learning* yang telah dilatih sebelumnya dengan teknik *TF-IDF* untuk vektorisasi teks dan *Random Forest* untuk klasifikasi, pada tabel 2 dibawah disajikan metrik evaluasi performa model.

Tabel 2. Tabel Hasil Klasifikasi Topik Berita

Label Berita	Precission	Recall	F1-Score
Pendidikan	0.96	0.95	0.96
Entertainment	0.95	0.95	0.95
Olahraga	0.93	0.91	0.93
Ekonomi	0.97	0.97	0.97

Politik	0.97	0.97	0.97
Pariwisata	0.93	0.90	0.93
Teknologi	0.98	0.97	0.98
Internasional	0.86	0.93	0.86
Bisnis	1.00	1.00	1.00
Kesehatan	0.92	0.92	0.95

2.3.3. Load

Setelah proses klasifikasi selesai, hasilnya dimuat (*load*) ke dalam *PostgreSQL* untuk disimpan. Hasil klasifikasi *clickbait* dan klasifikasi topik disimpan dalam dua tabel terpisah di *database*, yaitu tabel untuk berita *clickbait* dan tabel untuk topik. Penyimpanan data ini memungkinkan analisis lebih lanjut, visualisasi, dan pemantauan hasil secara berkala.

II.4. Penggunaan Tools

- Rabbitmq: sebagai *message broker* yang mendistribusikan data CSV ke beberapa *queue* untuk diproses lebih lanjut.
- Celery: sebagai *job scheduler* yang menangani pemrosesan sebagai *job scheduler* yang menangani pemrosesan data secara *asynchronous* dan paralel
- PostgreSQL: sebagai sistem basis data relasional yang menyimpan output akhir.
- Flower: sebagai pengawas sistem Celery secara *real time*.
- Python: sebagai *producer* dan inti dari sistem logika pemrosesan data.

III. HASIL DAN PEMBAHASAN

III.1. Konfigurasi Sistem *Asynchronous* dengan Celery

- Exchange :
  - Exchange_pipeline , type *direct* , mengarah ke tiga *queue*
- Queue dan routing key:
  - Prepocessing_queue : *prepocessing*
  - Clickbait_queue : *clickbait_clasify*
  - Topic_queue : *topik_classify*
- Worker
  - Worker_prepocessing , bertugas melakukan pembersihan data
  - Worker_classify_clickbait, bertugas melakukan klasifikasi berita *clickbait*
  - Worker_classify_topik, bertugas melakukan klasifikasi topik berita

III.2. Flower

Untuk memastikan proses *transform* berjalan dengan baik, dilakukan pemantauan menggunakan Flower, yakni sebuah *tool monitoring* untuk *task queue* pada Celery.

3.2.1. Hasil Flower untuk *Pre-Processing*

Tabel berikut menunjukkan hasil eksekusi salah satu proses *pre-processing* berita yang berhasil dijalankan.

Tabel 3. Tabel Hasil *Pre-processing*

Field	Value
Name	app.tasks.preprocess_task
UUID	bf1f2ae5-07a4-4b07-8e57-89802c5ee5e8
State	SUCCESS
args	('Relawan Prabowo Ancam Laporkan Sekjen PDI-P hingga Penyebar Hoaks Isu "Tampar dan Cekik" Wamen',)
Result	{'news': 'Relawan Prabowo Ancam Laporkan Sekjen PDI-P hingga Penyebar Hoaks Isu "Tampar dan Cekik" Wamen', 'cleaned_news': 'rawan prabowo ancam lapor sekjen pdi p hingga sebar hoaks isu tampar dan cekik wamen', 'processed_at': 1748676888.26189, 'status': 'processed'}
Retries	0
Worker	worker_preprocessing@LAPTOP-OSP1BNSP
Runtime	0.9220000000204891

Berdasarkan Tabel 3, proses *pre-processing* dengan *task* bernama app.tasks.preprocess_task berhasil dijalankan, ditandai dengan status SUCCESS. Data yang diproses berupa judul berita yang mengandung unsur clickbait. Judul asli yang tercantum dalam parameter args dan Result menunjukkan hasil pemrosesan berupa teks yang telah dibersihkan dari karakter dan kata yang tidak relevan, seperti tanda baca dan huruf kapital . Tidak ada percobaan ulang dalam proses ini (Retries: 0), dan proses dijalankan oleh worker bernama worker_preprocessing@LAPTOP-OSP1BNSP dengan waktu eksekusi sekitar 0.92 detik. Hal ini menunjukkan bahwa sistem *pre-processing* sudah berjalan dengan efisien dan sesuai dengan yang diharapkan.

3.2.2. Hasil Flower untuk Klasifikasi *Clickbait*

Setelah proses *pre-processing* selesai dilakukan, langkah selanjutnya adalah klasifikasi untuk menentukan apakah judul berita tergolong *clickbait* atau bukan. Proses ini dilakukan secara otomatis melalui *task* yang dijalankan dalam sistem klasifikasi, dan hasilnya dapat dimonitor menggunakan Flower. Tabel berikut menunjukkan hasil eksekusi salah satu proses klasifikasi *clickbait* berita yang berhasil dijalankan.

Tabel 4. Tabel Hasil Klasifikasi *clickbait*

Field	Value
Name	app.tasks.clickbait_task
UUID	da184baf-052b-4042-84c8-0d5df6d609cc
State	SUCCESS
args	{'news': 'Relawan Prabowo Ancam Laporkan Sekjen PDI-P hingga Penyebar Hoaks Isu "Tampar dan Cekik" Wamen', 'cleaned_news': 'rawan prabowo ancam lapor sekjen pdi p hingga sebar hoaks isu tampar dan cekik wamen', 'processed_at': 1748676888.26189, 'status': 'processed'}
Result	{'news': 'Relawan Prabowo Ancam Laporkan Sekjen PDI-P hingga Penyebar Hoaks Isu "Tampar dan Cekik" Wamen',

	'cleaned_news': 'rawan prabowo ancam lapor sekjen pdi p hingga sebar hoaks isu tampar dan cekik wamen', 'label': 'non-clickbait'}
Retries	0
Worker	worker_classify_clickbait@LAPTOP-OSP1BNSP
Runtime	0.40600000007543713

Tabel 4 menunjukkan hasil dari *task* app.tasks.clickbait_task yang berhasil dijalankan dengan status SUCCESS. Data yang dianalisis berupa judul berita, baik dalam bentuk aslinya (news) maupun yang telah dibersihkan (cleaned_news). Hasil klasifikasi (*Result*) menunjukkan bahwa judul berita tersebut dikategorikan sebagai "non-clickbait", artinya tidak mengandung unsur yang menyesatkan atau sensasional. Proses ini dijalankan oleh worker worker_classify_clickbait@LAPTOP-OSP1BNSP tanpa perlu pengulangan (Retries: 0) dan memerlukan waktu sekitar 0.406 detik untuk diselesaikan. Hal ini mencerminkan bahwa sistem klasifikasi *clickbait* sudah berjalan dengan baik dan efisien dalam mengidentifikasi jenis judul berita.

3.2.3. Hasil Flower untuk Klasifikasi Topik

Setelah proses *pre-processing*, sistem juga melakukan klasifikasi topik untuk menentukan kategori isi berita berdasarkan konteks dan kata kunci yang terdapat pada judul. Proses ini dijalankan secara otomatis dan diawasi melalui Flower, sebagaimana ditunjukkan pada hasil eksekusi salah satu *task* klasifikasi topik berikut.

Tabel 5. Tabel Hasil Klasifikasi Topik

Field	Value
Name	app.tasks.topik_task
UUID	f353f7af-8907-44e8-9943-01c0b3380783
State	SUCCESS
args	{'news': 'Relawan Prabowo Ancam Laporkan Sekjen PDI-P hingga Penyebar Hoaks Isu "Tampar dan Cekik" Wamen', 'cleaned_news': 'rawan prabowo ancam lapor sekjen pdi p hingga sebar hoaks isu tampar dan cekik wamen', 'processed_at': 1748676888.26189, 'status': 'processed'}
Result	{'news': 'Relawan Prabowo Ancam Laporkan Sekjen PDI-P hingga Penyebar Hoaks Isu "Tampar dan Cekik" Wamen', 'cleaned_news': 'rawan prabowo ancam lapor sekjen pdi p hingga sebar hoaks isu tampar dan cekik wamen', 'label': 'Politik'}
Retries	0
Worker	worker_classify_topik@LAPTOP-OSP1BNSP
Runtime	0.4529999999795109

Tabel 5 menunjukkan bahwa *task* app.tasks.topik_task berhasil dijalankan dengan status SUCCESS. Data yang diproses berupa judul berita yang telah dibersihkan dari unsur tidak relevan (cleaned_news). Hasil klasifikasi pada bagian Result menunjukkan bahwa judul tersebut masuk dalam kategori topik Politik. Proses ini dilakukan tanpa pengulangan (Retries: 0) oleh worker worker_classify_topik@LAPTOP-OSP1BNSP dan selesai dalam waktu sekitar 0.45 detik. Hasil

E-ISSN 2808-5841
P-ISSN 2808-7283

ini mengindikasikan bahwa sistem klasifikasi topik sudah berjalan secara optimal dalam mengelompokkan berita sesuai dengan konteksnya.

III.3. Hasil *Output Sistem*

3.3.1. Klasifikasi *Clickbait*

Tabel 6. Tabel Output Klasifikasi Clickbait

Text Berita	Text Berita Clean	Label
Aturan Baru Sri Mulyani: Gaji Minimal Rp 5 Juta Kena Pajak 5 Persen"	atur baru sri mulyani gaji minimal rp juta kena pajak persen	Non-clickbait
Menangis Saat Dihubungi Sambo, Kuat Ma'ruf: Bohong Mulu Capek Wat, Kamu Siap Ya Dipenjara	menang saat hubung sambo kuat ma ruf bohong mulu capek wat kamu siap ya penjara	Clickbait

Tabel 6 memperlihatkan hasil klasifikasi berita berdasarkan penggunaan *clickbait* pada judul. Proses klasifikasi ini dilakukan untuk mengidentifikasi sejauh mana praktik *clickbait* digunakan dalam penyajian berita. Hasil tersebut diharapkan dapat memberikan pemahaman mengenai pola dan karakteristik judul yang bersifat menarik perhatian secara berlebihan namun tidak selalu mencerminkan isi berita secara akurat.

Tabel 7. Tabel Hasil Klasifikasi Clickbait

Label Berita	Jumlah
clickbait	943
Non-clickbait	2.920

Tabel 7 memperlihatkan hasil klasifikasi judul berita ke dalam dua kategori, yaitu *clickbait* dan *non-clickbait*, berdasarkan model yang telah dilatih sebelumnya. Dari total 3.863 judul berita yang dianalisis, sebanyak 943 berita (24,4%) termasuk dalam kategori *clickbait*, sementara 2.920 berita (75,6%) tergolong *non-clickbait*. Proporsi ini dihitung berdasarkan perbandingan jumlah berita pada masing-masing kategori terhadap total keseluruhan data yang diproses. Temuan ini menunjukkan bahwa meskipun mayoritas berita tidak menggunakan judul *clickbait*, praktik tersebut masih cukup signifikan. Sekitar seperempat dari total berita mengandung judul yang dirancang untuk menarik perhatian secara berlebihan, namun tidak selalu mencerminkan isi berita secara akurat, yang mencerminkan bahwa strategi *clickbait* masih digunakan secara aktif di berbagai platform media digital.

E-ISSN 2808-5841
P-ISSN 2808-7283

3.3.2. Klasifikasi Topik

Tabel 8. Tabel Output Klasifikasi Clickbait

Text Berita	Text Berita Clean	Label Topik
Link Live Streaming Semifinal Piala AFF 2022, Leg 2 Vietnam Vs Indonesia	link live streaming semifinal piala aff leg vietnam vs indonesia	Olahraga
Pilihan Berat Christine Hakim Saat Dapat Tawaran Casting Serial The Last of Us	pilih berat christine hakim saat dapat tawar casting serial the last of us	Entertainment
Jadi Bakal Capres Favorit Partai Ummat, Anies Diundang ke Rakernas	jadi bakal capres favorit partai ummat anies undang ke rakernas	Politik

Proses ini bertujuan untuk memberikan pemetaan yang lebih terstruktur terhadap jenis berita yang ada, sehingga memudahkan dalam pemantauan tren berita yang terjadi. Setelah proses klasifikasi selesai, hasilnya disimpan dalam database *PostgreSQL*, seperti yang dapat dilihat pada Tabel 8, yang menunjukkan contoh data berita yang telah diproses dan diklasifikasikan .

Tabel 9. Tabel Hasil Klasifikasi Topik

Topik Berita	Jumlah
Pendidikan	1.285
Entertainment	490
Olahraga	475
Ekonomi	461
Politik	461
Pariwisata	229
Teknologi	161
Internasional	160
Bisnis	107
Kesehatan	34

Berdasarkan Tabel 9, topik berita yang paling banyak muncul adalah topik Pendidikan dengan jumlah 1.285 berita, disusul oleh *Entertainment* sebanyak 490 berita dan Olahraga sebanyak 475 berita. Topik Ekonomi dan Politik memiliki jumlah yang hampir sama, yaitu masing-masing 461 berita. Sementara itu, topik dengan jumlah berita paling sedikit adalah Kesehatan dengan hanya 34 berita, diikuti oleh Bisnis sebanyak 107 berita dan *Internasional* sebanyak 160 berita. Hal ini menunjukkan bahwa fokus utama pemberitaan lebih condong ke bidang pendidikan, hiburan, dan olahraga dibandingkan dengan topik-topik lain.

IV. KESIMPULAN

Penelitian ini berhasil mengimplementasikan *pipeline ETL* terdistribusi untuk klasifikasi berita clickbait dan topik berita dengan menggunakan sistem berbasis *RabbitMQ* dan *Celery*. Sistem ini mampu menangani volume data berita yang besar secara efisien dan paralel, memanfaatkan kemampuan pemrosesan terdistribusi untuk mengelola data dalam skala besar. Proses ETL yang terdiri dari tahap *Extract*, *Transform*, dan *Load* berhasil dilakukan dengan baik, menghasilkan data yang telah diklasifikasikan dan disimpan dalam *database PostgreSQL*. Secara keseluruhan, penelitian ini memberikan kontribusi penting dalam pengelolaan dan analisis data berita digital, serta menyediakan dasar untuk pengembangan lebih lanjut dalam sistem *intelijen media* dan *analitik konten* skala besar.

REFERENSI

1. A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop Clickbait: Detecting and preventing clickbaits in online news media," in Proc. 2016 IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining (ASONAM), Aug. 2016, pp. 9–16, doi: 10.1109/ASONAM.2016.7752207.
2. V. Kaushal and K. Vemuri, "Clickbait—Trust and credibility of digital news," IEEE Transactions on Technology and Society, vol. 2, no. 3, pp. 146–154, Apr. 2021, doi: 10.1109/TTS.2021.3089201.
3. A. Yoraeni, P. Handayani, S. N. Rakhmah, J. Siregar, D. Y. Al Afghani, H. Rianto, F. Riza, A. Yuswanto, E. P. Saputra, E. Prayitno, M. Muharrom, T. Muryanto, R. Damayanti, D. Febrianto, dan A. Nurrohman, Sistem Informasi Manajemen, S. J. Al Din, Ed. Jakarta: PT. Scifintech Andrew Wijaya, 2023.
4. D. Seenivasan, "Distributed ETL Architecture for Processing and Storing Big Data," International Journal of Current Science (IJCSPUB), vol. 12, no. 3, July 2022, ISSN: 2250-1770.
5. A. Simitsis, S. Skiadopoulos, and P. Vassiliadis, "The History, Present, and Future of ETL Technology," in International Workshop on Data Warehousing and OLAP (DOLAP), 2023.
6. R. Yunanto, A. P. Purfini, dan A. Prabuwisesa, "Survei Literatur: Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning," Jurnal Manajemen Informatika (JAMIKA), vol. 11, no. 2, pp. 118–130, 2021, doi: 10.34010/jamika.v11i2.493.
7. N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, "Fake News Classification using transformer based enhanced LSTM and BERT," Int. J. Cogn. Comput. Eng., vol. 3, pp. 98–105, 2022, doi: 10.1016/j.ijcce.2022.03.003.