Jurnal Komputer dan Elektro Sains, Vol.4 No.2, Sept 2026, 41-52
https://ejournal.sultanpublisher.com/komets

Received : 31 Juli, 2026; Revision: 22 Agustus 2026; Accepted: 22 Agustus 2026; Publication: 1 September 2026.

E-ISSN: 3021-8462, DOI: https://doi.org/10.58291/komets.v4i2.662

Fusi Metadata dan Masked Mean–Max Pooling dengan DeBERTa-v3 untuk Klasifikasi Tweet Bencana

Budi Mukhamad Mulyo[1], Hazna At Thooriqoh[1], Ardhon Rakhmadi[1], Devi Ambarwati Puspitasari[2], Bayu Permana Sukma[3]

[1]Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia
[2]Pusat Riset Bahasa, Sastra, dan Komunitas, Badan Riset dan Inovasi Nasional, Jakarta, Indonesia
[3]Lancaster University, Lancaster, England

Penulis Korespondensi :Budi Mukhamad Mulyo (e-mail: budi.m.mulyo.fasilkom@upnjatim.ac.id)

ABSTRAK

Klasifikasi tweet terkait bencana merupakan permasalahan penting dalam pemrosesan bahasa alami karena teks media sosial cenderung pendek, tidak baku, dan sering ambigu. Namun, pemanfaatan metadata dan optimasi ambang keputusan secara terpadu pada model transformer masih terbatas. Penelitian ini mengembangkan model klasifikasi biner dengan mengintegrasikan representasi kontekstual DeBERTa-v3, kata kunci, lokasi, dan metadata numerik tweet. Dataset terdiri atas 7.613 data latih dan 3.263 data uji. Teks dibentuk menjadi masukan terstruktur menggunakan penanda kata kunci, lokasi, dan isi tweet. Representasi token dipadatkan melalui masked mean–max pooling, kemudian digabungkan dengan fitur numerik. Model dilatih menggunakan stratified cross-validation, multi-sample dropout, dan optimasi threshold berdasarkan probabilitas out-of-fold. Kebaruan penelitian terletak pada demonstrasi empiris kontribusi metadata fusion, masked mean–max pooling, dan optimasi threshold berbasis out-of-fold dalam melengkapi representasi kontekstual DeBERTa-v3, yang diperkuat melalui evaluasi ablation terkontrol. Pengujian terhadap 15 konfigurasi menunjukkan bahwa DeBERTa-v3 metadata fusion dengan 3-fold out-of-fold ensemble menghasilkan skor F1 tertinggi sebesar 0,84063, melampaui DeBERTa-v3 metadata fusion dengan holdout sebesar 0,83971, DeBERTa-v3 mean–max sebesar 0,83695, DistilBERT sebesar 0,83450, dan GloVe–SVM sebesar 0,81581. Hasil ini menunjukkan bahwa integrasi metadata dan penentuan threshold berbasis out-of-fold efektif meningkatkan kinerja klasifikasi. Penelitian ini berpotensi mendukung penyaringan informasi bencana dari media sosial secara lebih akurat.

KATA KUNCI Klasifikasi Tweet Bencana; DeBERTa-v3; Metadata Fusion; Transformer; Natural Language Processing.

ABSTRACT

Disaster tweet classification is an important task in natural language processing because social media texts are often short, informal, and ambiguous. However, the integrated use of metadata and decision-threshold optimization in transformer-based models remains limited. This study develops a binary classification model by integrating DeBERTa-v3 contextual representations with tweet keywords, locations, and numerical metadata. The dataset contains 7,613 training instances and 3,263 test instances. Texts are transformed into structured inputs using keyword, location, and tweet-content markers. Token representations are aggregated through masked mean–max pooling and combined with numerical features. The model is trained using stratified cross-validation, multi-sample dropout, and out-of-fold threshold optimization. The novelty of this study lies in the empirical demonstration of the contributions of metadata fusion, masked mean–max pooling, and out-of-fold threshold optimization in complementing DeBERTa-v3 contextual representations, further supported by a controlled ablation study. Among 15 configurations, the 3-fold DeBERTa-v3 metadata-fusion ensemble achieved the highest F1 score of 0.84063, outperforming the holdout version at 0.83971, DeBERTa-v3 mean–max at 0.83695, DistilBERT at 0.83450, and GloVe–SVM at 0.81581. These results indicate that metadata integration and out-of-fold threshold optimization effectively improve classification performance. The proposed approach has the potential to improve the accuracy of filtering disaster-related information from social media.

KEYWORD Disaster Tweets Classification; DeBERTa-v3; Metadata Fusion; Transformer; Natural Language Processing.

1. PENDAHULUAN

Media sosial menjadi salah satu sumber informasi awal ketika terjadi bencana karena pengguna dapat melaporkan kondisi lapangan, kerusakan, peringatan, kebutuhan bantuan, dan situasi keselamatan secara langsung [1], [2], [3], [4], [5]. Informasi tersebut bermanfaat untuk membangun kesadaran situasional, tetapi volume pesan yang tinggi membuat proses penyaringan manual sulit dilakukan [6], [7], [8]. Selain itu, tidak setiap tweet yang memuat istilah seperti fire, flood, crash, atau emergency benar-benar melaporkan kejadian bencana. Sebagian istilah digunakan secara metaforis, dalam percakapan sehari-hari, promosi, atau kutipan berita yang tidak menggambarkan peristiwa aktual. Oleh karena itu, sistem klasifikasi diperlukan untuk membedakan tweet bencana dan nonbencana secara cepat [9], [10], [11], [12].

Klasifikasi tweet memiliki karakteristik yang berbeda dari klasifikasi dokumen formal. Tweet cenderung pendek, mengandung singkatan, ejaan tidak baku, mention, hashtag, URL, simbol, dan konteks lokasi yang tidak selalu lengkap [10], [13], [14], [15]. Penelitian mengenai pesan krisis juga menunjukkan bahwa pesan singkat dan informal menimbulkan tantangan dalam parsing, pembentukan kategori, dan identifikasi informasi yang dapat ditindaklanjuti [5], [16], [17]. Pendekatan berbasis fitur dapat menangkap pola eksplisit, tetapi sering kesulitan memahami makna kontekstual dan ambiguitas suatu kata [9], [12], [18]. Sebaliknya, model bahasa pralatih mampu menghasilkan representasi kontekstual yang lebih baik, namun informasi tambahan seperti kata kunci, lokasi, dan pola permukaan tweet belum tentu dimanfaatkan secara optimal [8], [19], [20].

Arsitektur Transformer dan berbagai turunannya telah banyak diterapkan dalam analisis media sosial berbasis bencana karena mampu memodelkan hubungan antarkata secara kontekstual [14], [17], [21]. Model berbasis BERT juga menunjukkan peningkatan performa dibandingkan beberapa pendekatan pembelajaran mesin dan jaringan saraf konvensional dalam klasifikasi informasi kebencanaan [22], [23], [24], [25]. Pengembangan selanjutnya mencakup penggunaan Transformer yang lebih efisien, model yang disesuaikan dengan karakteristik media sosial, serta pendekatan yang mengoptimalkan pemanfaatan representasi kontekstual [19], [26], [27], [28]. Meskipun sejumlah penelitian telah menggunakan model Transformer untuk klasifikasi tweet terkait bencana, kajian yang secara sistematis mengintegrasikan metadata tweet yang ringan dengan representasi kontekstual sekaligus mengoptimalkan ambang keputusan menggunakan prediksi out-of-fold masih terbatas. Penelitian ini secara khusus menguji apakah metadata ringan yang tersedia langsung pada tweet dapat memberikan informasi komplementer terhadap representasi kontekstual DeBERTa-v3, serta apakah optimasi keputusan berbasis OOF dapat mempertahankan kinerja tanpa memerlukan sumber multimodal eksternal. Dalam penelitian ini, DeBERTa-v3 dipilih sebagai encoder utama karena kemampuannya menghasilkan representasi kontekstual yang kuat. Representasi tersebut tidak hanya diambil dari token pertama, tetapi diringkas dari seluruh *hidden state* dan dipadukan dengan fitur tambahan yang menggambarkan karakteristik tweet [15], [29], [30].

Penelitian ini mengusulkan model DeBERTa-v3 metadata fusion dengan empat komponen utama. Pertama, kata kunci, lokasi, dan teks tweet disusun sebagai masukan terstruktur menggunakan penanda khusus. Penyusunan ini bertujuan agar encoder dapat membedakan informasi utama dan atribut pendamping yang terdapat pada data [19], [20]. Kedua, hidden state seluruh token diringkas menggunakan masked mean pooling dan masked max pooling. Ketiga, representasi tekstual tersebut digabungkan dengan metadata numerik yang menggambarkan karakteristik permukaan tweet. Pendekatan fusi ini sejalan dengan penelitian yang menunjukkan bahwa penggabungan beberapa representasi dapat memperkaya informasi yang digunakan dalam klasifikasi bencana [15], [29], [31]. Keempat, prediksi diperkuat melalui multi-sample dropout, validasi silang terstratifikasi, dan optimasi ambang klasifikasi berdasarkan probabilitas out-of-fold. Penggunaan beberapa pembagian data dan strategi regularisasi diperlukan untuk mengurangi sensitivitas model terhadap ukuran dan komposisi data pelatihan [13], [28]. Kontribusi penelitian meliputi: (1) rancangan integrasi teks dan metadata untuk klasifikasi tweet bencana; (2) evaluasi bertahap terhadap pendekatan fitur klasik, word embedding, jaringan saraf, dan Transformer; serta (3) analisis pengaruh ambang klasifikasi dan jumlah fold terhadap generalisasi model. Kebaruan yang dievaluasi dalam penelitian ini tidak hanya terletak pada penggabungan beberapa komponen, tetapi pada demonstrasi empiris mengenai kontribusi metadata ringan, strategi pooling, dan optimasi keputusan berbasis OOF dalam satu rangkaian eksperimen klasifikasi tweet bencana.

Kajian komparatif terbaru menunjukkan bahwa klasifikasi tweet bencana telah dikembangkan menggunakan fitur leksikal, n-gram, seleksi fitur, algoritma pembelajaran mesin, jaringan saraf, serta model Transformer [9], [10], [11], [12], [14], [16], [18]. Pendekatan berbasis fitur masih relevan karena efisien, memiliki kebutuhan komputasi yang relatif rendah, dan lebih mudah diinterpretasikan. Namun, tinjauan terbaru menunjukkan bahwa metode tersebut menghadapi keterbatasan ketika digunakan pada teks media sosial yang ambigu, informal, dan terus berubah [13], [17]. Representasi sparse seperti TF-IDF juga tidak secara langsung memahami polisemi, hubungan semantik

antarkata, atau perubahan makna suatu istilah berdasarkan konteks kalimat [9], [12], [18].

Model berbasis Transformer mengatasi sebagian keterbatasan tersebut dengan menghasilkan representasi yang berubah sesuai konteks kalimat. Berbagai penelitian menunjukkan bahwa BERT dan turunannya memberikan hasil yang kompetitif dalam klasifikasi tweet bencana, identifikasi kerusakan, dan penyaringan pesan yang informatif [14], [22], [23], [24], [26], [28]. Model yang disesuaikan dengan bahasa media sosial, termasuk pendekatan berbasis RoBERTa, juga dapat meningkatkan kemampuan sistem dalam memproses bentuk bahasa informal yang umum ditemukan pada tweet [22], [27]. DeBERTa-v3 mengembangkan representasi kontekstual melalui mekanisme perhatian terpisah dan strategi pralatih yang berbeda dari BERT standar. Dalam penelitian ini, seluruh hidden state digunakan agar informasi dari berbagai token tetap membentuk representasi tweet.

Selain pemodelan teks, penelitian terkini memperlihatkan bahwa penggabungan beberapa modalitas, kanal fitur, atau sumber data tambahan dapat meningkatkan kemampuan klasifikasi pesan krisis [32], [33], [34], [35], [36], [37]. Informasi visual, spasial, semantik, dan karakteristik tambahan dari unggahan dapat membantu model ketika teks tidak menyediakan konteks yang cukup. Namun, pendekatan multimodal sering membutuhkan citra, koordinat geografis, atau sumber eksternal yang tidak selalu tersedia pada setiap tweet.

Metadata sederhana, seperti kata kunci pencarian, lokasi pengguna, keberadaan URL, jumlah mention, panjang teks, penggunaan huruf kapital, tanda baca, dan pola karakter, menjadi alternatif yang lebih ringan dibandingkan integrasi multimodal penuh [8], [15], [38], [39]. Fitur-fitur tersebut dapat memberikan petunjuk tambahan mengenai bentuk, sumber, dan kemungkinan konteks suatu tweet. Meskipun demikian, penggunaan metadata harus dirancang secara hati-hati agar tidak menimbulkan kebocoran informasi, ketergantungan pada karakteristik dataset tertentu, atau penurunan generalisasi ketika diterapkan pada kejadian bencana yang berbeda [3], [5], [7], [39], [40]. Oleh karena itu, penelitian ini memisahkan pengolahan representasi teks dan metadata sebelum menggabungkan keduanya pada lapisan fusi, sehingga kontribusi masing-masing sumber informasi dapat dianalisis secara lebih terkontrol.

2. METODE

2.1. Dataset dan Pembagian Data

Dataset penelitian diperoleh dari repositori publik Kaggle pada *Natural Language Processing with Disaster Tweets*. Dataset tersebut dirancang untuk klasifikasi biner tweet dengan membedakan tweet yang menggambarkan kejadian bencana nyata dan tweet yang tidak berkaitan dengan kejadian bencana nyata[41].

Kolom *keyword* dan *location* merupakan atribut yang telah disediakan bersama dataset dan tidak dibentuk dari nilai *target* dalam proses penelitian ini. Dokumentasi dataset mendeskripsikan *keyword* sebagai kata tertentu yang berkaitan dengan tweet dan *location* sebagai informasi lokasi yang menyertai tweet; kedua atribut tersedia pada data latih maupun data uji, sedangkan *target* hanya tersedia pada data latih. Dokumentasi publik tidak menjelaskan prosedur ekstraksi *keyword* secara lebih rinci, sehingga atribut tersebut diperlakukan sebagai metadata bawaan dataset, bukan fitur yang diturunkan dari label. Selain itu, *location* tidak diasumsikan sebagai lokasi kejadian bencana yang telah diverifikasi, tetapi digunakan sesuai nilai yang tersedia pada dataset.

Dataset penelitian berisi 7.613 tweet berlabel dan 3.263 tweet tanpa label yang digunakan sebagai data uji. Pada data latih, setiap record terdiri atas id, keyword, location, text, dan target, sedangkan data uji tidak memiliki atribut target. Target 0 menunjukkan tweet nonbencana, sedangkan target 1 menunjukkan tweet yang benar-benar berkaitan dengan bencana. Distribusi data latih terdiri atas 4.342 data kelas 0 (57,03%) dan 3.271 data kelas 1 (42,97%). Distribusi tersebut tidak sangat timpang, tetapi stratifikasi tetap digunakan agar setiap fold mempertahankan proporsi kelas yang serupa.

Audit tambahan terhadap data menunjukkan adanya duplikasi pada dataset. Setelah normalisasi URL, mention, kapitalisasi, dan spasi, diperoleh 6.944 teks unik dari 7.613 data, sehingga terdapat 669 baris redundan. Sebanyak 305 kelompok teks memiliki lebih dari satu kemunculan dan 69 kelompok memiliki label yang tidak seragam. Pada pembagian StratifiedKFold yang digunakan, 230 kelompok duplikat tersebar pada fold yang berbeda dan melibatkan 814 data. Analisis kemiripan berbasis karakter juga menunjukkan bahwa 597 data memiliki nearest-neighbor pada fold berbeda dengan similarity ≥0,95. Dataset tidak menyediakan identitas event, source, atau author, sehingga pemisahan berdasarkan kejadian atau sumber tidak dapat diverifikasi secara langsung. Kondisi ini diperlakukan sebagai keterbatasan dataset dan dipertimbangkan dalam interpretasi hasil validasi.

Tabel 1. Distribusi dataset penelitian

Komponen	Jumlah	Proporsi
Data latih	7.613	-
Kelas nonbencana	4.342	57,03%
Kelas bencana	3.271	42,97%
Data uji	3.263	-

2.2. Prapemrosesan

Prapemrosesan dilakukan secara ringan agar informasi semantik dan gaya penulisan tidak hilang. Nilai kosong pada keyword dan location diganti dengan token none. URL dinormalisasi menjadi HTTPURL dan mention menjadi @USER. Spasi berulang diringkas, sedangkan hashtag, tanda seru, tanda tanya, kapitalisasi, dan angka dipertahankan karena digunakan sebagai metadata. Masukan transformer dibangun dalam format [KEYWORD] keyword [LOCATION] location [TEXT] tweet. Tokenizer menggunakan tokenizer bawaan

checkpoint microsoft/deberta-v3-base melalui AutoTokenizer dengan mode fast. Tiga token khusus, yaitu [KEYWORD], [LOCATION], dan [TEXT], ditambahkan sebagai *additional special tokens*. Dengan demikian, kosakata bertambah sebanyak tiga token dan ukuran *token embedding* pada encoder disesuaikan dengan ukuran tokenizer yang baru.

Tokenisasi menggunakan panjang maksimum 112 token. Padding tidak disertakan dalam pooling. Token khusus [CLS], [SEP], [PAD], serta penanda segmen juga dikeluarkan dari mask pooling agar representasi terutama berasal dari token informatif. Strategi ini berbeda dari penggunaan representasi satu token khusus karena seluruh token valid berkontribusi pada vektor akhir.

2.3. Ekstraksi Metadata Numerik

Sebanyak 12 fitur numerik diekstraksi dari setiap record, yaitu keberadaan keyword, keberadaan location, panjang karakter, jumlah kata, jumlah hashtag, URL, mention, tanda seru, tanda tanya, rasio huruf kapital, kesesuaian keyword terhadap isi tweet, dan jumlah digit. Fitur hitungan ditransformasi menggunakan log(1+x) untuk mengurangi dominasi nilai ekstrem. Pada setiap fold, StandardScaler hanya dipelajari dari bagian data latih, kemudian diterapkan pada data validasi dan data uji. Prosedur per-fold ini mencegah informasi statistik data validasi masuk ke proses pelatihan.

2.4. Arsitektur Model yang Diusulkan

Checkpoint microsoft/deberta-v3-base digunakan sebagai encoder melalui implementasi Hugging Face Transformers. Seluruh parameter encoder dan lapisan tambahan pada metadata fusion dilatih selama fine-tuning; tidak terdapat lapisan encoder yang dibekukan. Hidden state terakhir dinyatakan sebagai H={h1,h2,…,hn}. Mask m_i bernilai 1 untuk token valid dan 0 untuk padding atau token khusus. Masked mean pooling menghitung rata-rata representasi token valid, sedangkan masked max pooling memilih aktivasi maksimum pada setiap dimensi.

$$h\_mean = \frac{(\sum_i m_i h_i)}{(\sum_i m_i)} \quad (1)$$

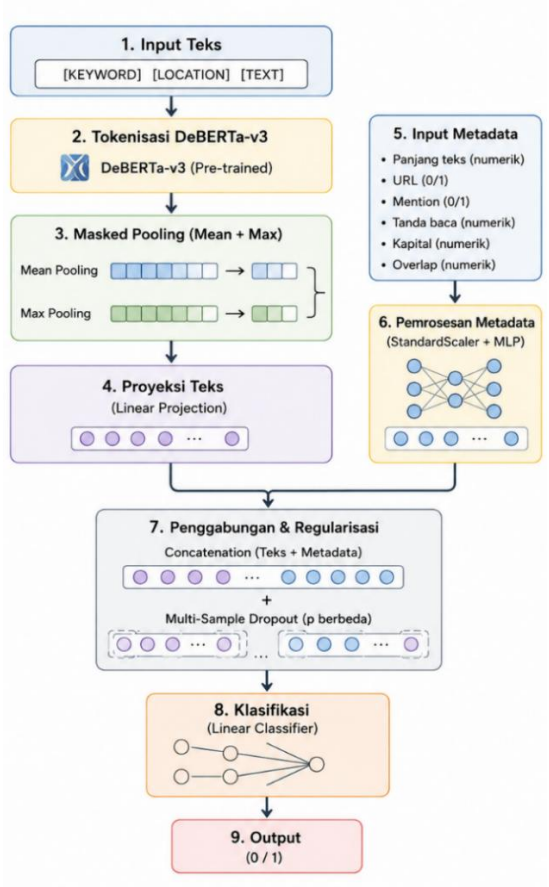
$$h\_max = max_i(h_i | m_i = 1) \quad (2)$$

Kedua vektor digabungkan menjadi h_text=[h_mean;h_max], dinormalisasi dengan LayerNorm, dan diproyeksikan ke 256 dimensi menggunakan lapisan linear dan GELU. Metadata numerik diproyeksikan ke 32 dimensi melalui linear, LayerNorm, GELU, dan dropout. Vektor teks dan metadata selanjutnya dikonkatenasi dan diproyeksikan kembali ke 256 dimensi. Multi-sample dropout menerapkan lima tingkat dropout, yaitu 0,10; 0,20; 0,30; 0,40; dan 0,50, pada vektor fusi yang sama. Semua cabang menggunakan classifier linear bersama, kemudian logits dirata-ratakan. Mekanisme ini menghasilkan beberapa gangguan regularisasi dalam satu forward pass tanpa membuat encoder tambahan.

2.5. Pelatihan dan Evaluasi

Seluruh eksperimen dijalankan pada lingkungan Google Colab menggunakan GPU NVIDIA Tesla T4 dengan memori GPU sekitar 14,56 GB dan RAM sistem sekitar 12,67 GB. Lingkungan komputasi menggunakan CUDA versi 12.8 untuk mendukung proses pelatihan model berbasis GPU. Implementasi utama menggunakan PyTorch 2.11.0+cu128 dan Hugging Face Transformers 4.57.1. Random seed utama ditetapkan sebesar 42 untuk pembagian data dan inisialisasi eksperimen. Pada konfigurasi 3-fold utama, seed pelatihan setiap fold ditetapkan secara deterministik berdasarkan seed utama dan nomor fold. Seluruh parameter model dipertahankan dalam FP32, sedangkan operasi komputasi menggunakan automatic mixed precision.

Model dilatih menggunakan AdamW dengan learning rate $1{,}5\times10^{-5}$ untuk encoder dan 1×10^{-4} untuk classification head. Weight decay ditetapkan 0,01, batch size 8, gradient accumulation 2, dan maksimum lima epoch. Linear scheduler menggunakan warmup 10% dari seluruh langkah optimasi. Gradien dibatasi pada norma maksimum 1,0. Fungsi kerugian yang digunakan adalah BCEWithLogitsLoss. Mixed precision diterapkan pada operasi komputasi, sedangkan parameter model dipertahankan dalam FP32.



Gambar 1. Arsitektur DeBERTa-v3 metadata fusion

Eksperimen utama menggunakan StratifiedKFold tiga bagian. Setiap fold menghasilkan probabilitas validasi untuk indeks yang tidak digunakan pada pelatihan dan probabilitas data uji. Seluruh probabilitas validasi digabungkan menjadi out-of-fold (OOF), sedangkan probabilitas data uji dirata-ratakan dari tiga model. Threshold tidak ditetapkan langsung pada 0,50, tetapi dicari pada rentang 0,20–0,80 dengan interval 0,0025 berdasarkan F1 OOF. Metrik utama adalah F1 karena menggabungkan precision dan recall.

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (3)$$

Pemilihan threshold dilakukan hanya menggunakan probabilitas OOF dan label data latih. Data uji tidak digunakan untuk mempelajari scaler, memperbarui parameter model, memilih checkpoint, maupun menentukan threshold. Pada setiap fold, probabilitas data uji hanya dihasilkan melalui inferensi dan selanjutnya dirata-ratakan antar-model untuk memperoleh prediksi akhir. Karena label data uji tidak tersedia bagi peneliti, tidak terdapat informasi label data uji yang masuk ke proses pelatihan atau optimasi threshold. Skor evaluasi eksternal digunakan untuk melaporkan dan membandingkan hasil 15 konfigurasi, tetapi tidak digunakan dalam perhitungan gradient, pemilihan checkpoint, ataupun optimasi threshold pada masing-masing konfigurasi. Dalam pelaporan penelitian ini, konfigurasi dengan skor F1 eksternal tertinggi ditetapkan sebagai konfigurasi terbaik di antara 15 konfigurasi yang dievaluasi. Karena evaluasi eksternal dilakukan terhadap beberapa konfigurasi, data uji tersebut tidak diperlakukan sebagai single-use holdout dan potensi model-selection bias dipertimbangkan dalam interpretasi hasil.

Perbandingan antarkonfigurasi menggunakan skor F1 pada data uji eksternal yang diperoleh melalui sistem evaluasi kompetisi, karena label data uji tidak tersedia bagi peneliti. Sedangkan evaluasi diagnostik model akhir dilakukan menggunakan prediksi out-of-fold pada data berlabel. Evaluasi OOF mencakup accuracy, precision, recall, F1, ROC-AUC, dan confusion matrix. Karena berasal dari skema evaluasi yang berbeda, F1 pengujian dan F1 OOF tidak dibandingkan secara langsung.

Untuk mengevaluasi perbedaan kinerja yang relatif kecil antarkonfigurasi, analisis statistik dilakukan pada sampel berlabel yang sama. Perbedaan F1 dievaluasi menggunakan paired bootstrap sebanyak 10.000 resampling dengan interval kepercayaan 95%, sedangkan perbedaan pola kesalahan klasifikasi diuji menggunakan exact McNemar test pada taraf signifikansi 0,05. Analisis ini digunakan sebagai evaluasi tambahan terhadap perbedaan numerik skor pengujian dan tidak menggantikan F1 sebagai metrik utama pemeringkatan konfigurasi.

Efisiensi komputasi dievaluasi dengan mencatat waktu pelatihan setiap fold, total waktu pelatihan, waktu inferensi, throughput, dan penggunaan memori GPU maksimum. Pengukuran inferensi dilakukan pada batch size 32 terhadap ensemble tiga model. Sinkronisasi CUDA dilakukan sebelum dan sesudah pengukuran waktu agar pencatatan mencerminkan penyelesaian operasi pada GPU.

Setiap konfigurasi utama dieksekusi satu kali menggunakan seed yang telah ditetapkan. Pada konfigurasi berbasis cross-validation, satu eksperimen terdiri atas pelatihan satu model pada setiap fold; dengan demikian konfigurasi utama 3-fold menghasilkan tiga model, sedangkan konfigurasi 5-fold menghasilkan lima model. Eksperimen tidak menggunakan pengulangan multi-seed, sehingga variasi akibat seed tetap menjadi salah satu keterbatasan penelitian.

Untuk mengidentifikasi kontribusi masing-masing kelompok metadata, dilakukan studi ablation terkontrol menggunakan pembagian holdout yang sama dengan konfigurasi metadata fusion, yaitu 1.142 data validasi dengan random seed 42. Setiap varian menghilangkan satu kelompok fitur dengan mempertahankan arsitektur, hyperparameter, dan prosedur optimasi yang sama. Untuk keyword dan location, informasi dihilangkan baik dari masukan terstruktur maupun fitur numerik terkait, sedangkan kelompok metadata lainnya dinonaktifkan pada jalur metadata numerik. Nilai F1 pada studi ablation digunakan untuk mengevaluasi perubahan relatif antarvarian pada holdout terkontrol dan tidak dibandingkan secara langsung dengan skor F1 eksternal pada Tabel 2.

3. HASIL DAN PEMBAHASAN

3.1. Perbandingan Model

Evaluasi dilakukan secara bertahap untuk melihat kontribusi jenis representasi, arsitektur, metadata, ensemble, dan skema validasi. Tabel 2 menyajikan skor F1 dari 15 konfigurasi. TF-IDF dengan Logistic Regression menjadi baseline terkuat pada kelompok fitur sparse, sedangkan penggunaan SVM tanpa representasi kontekstual belum memberikan peningkatan. GloVe–SVM mencapai 0,81581 dan melampaui CNN maupun BiLSTM-Attention pada embedding yang sama. Hasil ini menunjukkan bahwa model yang lebih kompleks tidak otomatis lebih baik ketika ukuran dataset terbatas.

Tabel 2. Perbandingan skor F1 seluruh konfigurasi

No.	Konfigurasi	F1
1	TF-IDF + Logistic Regression	0,79895
2	TF-IDF + SVM	0,78577
3	TF-IDF + SVM (manual tuning)	0,79160
4	TF-IDF + SVM (Optuna)	0,79344
5	GloVe Twitter + SVM	0,81581
6	GloVe Twitter + CNN	0,79803
7	GloVe + BiLSTM-Attention	0,78577
8	DistilBERT + masked mean pooling	0,83450
9	BERTweet + masked mean pooling	0,83358
10	DeBERTa-v3 mean–max + MSD	0,83695
11	DeBERTa-v3 metadata fusion (holdout)	0,83971
12	Weighted ensemble model 10 + 11	0,83604
13	Metadata fusion 3-fold OOF	0,84063
14	5-fold snapshot + LLRD	0,83787

No.	Konfigurasi	F1
15	Metadata fusion 5-fold murni	0,82500

Transformer memberikan peningkatan yang jelas dibandingkan pendekatan sebelumnya. DistilBERT memperoleh F1 0,83450 dan BERTweet 0,83358. Kinerja BERTweet yang sedikit lebih rendah menunjukkan bahwa kesesuaian domain pralatih terhadap tweet tidak menjadi satu-satunya faktor. Tujuan pralatih, kapasitas model, strategi pooling, metadata, dan proses fine-tuning turut menentukan hasil. DeBERTa-v3 dengan mean–max pooling dan multi-sample dropout mencapai 0,83695, atau meningkat 0,00245 terhadap DistilBERT.

Penambahan keyword, location, dan metadata numerik meningkatkan skor menjadi 0,83971 pada pembagian holdout. Peningkatan 0,00276 terhadap model DeBERTa-v3 tanpa metadata menunjukkan bahwa fitur tambahan memberikan informasi komplementer. Keyword dapat memperjelas topik yang tidak eksplisit pada teks, sedangkan location membantu membedakan penyebutan kejadian aktual dari ungkapan umum. Fitur gaya penulisan seperti URL, kapitalisasi, tanda baca, dan panjang teks juga berperan sebagai sinyal pendukung, tetapi tidak digunakan secara mandiri sehingga keputusan tetap didominasi representasi kontekstual.

3.2. Pengaruh Threshold

Threshold standar 0,50 pada model metadata fusion hanya menghasilkan F1 0,82899, sedangkan threshold yang diperoleh dari validasi menghasilkan 0,83971. Selisih 0,01072 menunjukkan bahwa probabilitas model belum terkalibrasi untuk langsung dipisahkan pada 0,50. Pada data dengan tujuan optimasi F1, threshold optimal bergantung pada keseimbangan precision dan recall. Threshold yang terlalu tinggi menurunkan recall karena sebagian tweet bencana diklasifikasikan sebagai nonbencana, sedangkan threshold terlalu rendah menambah false positive.

Optimasi threshold harus dilakukan pada data yang tidak dipakai memperbarui parameter model. Pada holdout, threshold hanya bergantung pada 15% data sehingga berisiko sensitif terhadap komposisi sampel. Penggunaan OOF memperluas dasar pemilihan threshold karena setiap data latih diprediksi oleh model yang tidak melihat data tersebut saat pelatihan. Dengan demikian, estimasi threshold lebih stabil dan mendekati kondisi inferensi. Seluruh proses pencarian threshold tersebut dilakukan tanpa menggunakan label maupun skor evaluasi data uji.

Evaluasi diagnostik pada konfigurasi percobaan nomor 13 menunjukkan pola yang konsisten pada skema OOF. Penggunaan threshold standar 0,50 menghasilkan accuracy sebesar 0,82727, precision 0,80055, recall 0,79639, dan F1-score 0,79847. Setelah threshold dioptimalkan menjadi 0,7175 berdasarkan seluruh probabilitas OOF, accuracy meningkat menjadi 0,84198, precision menjadi 0,85582, dan F1-score menjadi 0,80525, sedangkan recall menurun menjadi 0,76032. Perubahan threshold tersebut mengubah 348 dari 7.613 keputusan klasifikasi. Hasil ini menunjukkan bahwa optimasi threshold terutama meningkatkan precision dengan konsekuensi penurunan recall, sehingga pemilihan batas keputusan berperan dalam menentukan keseimbangan antara ketepatan prediksi kelas bencana dan kemampuan mendeteksi seluruh tweet bencana.

3.3. Pengaruh Cross-Validation dan Ensemble

Model 3-fold OOF menghasilkan skor tertinggi 0,84063, meningkat 0,00092 terhadap metadata fusion holdout. Selisih tersebut bersifat marginal sehingga belum cukup untuk menunjukkan keunggulan yang konsisten tanpa evaluasi statistik pada prediksi berlabel yang sama. Setiap model dilatih menggunakan sekitar dua pertiga data dan menghasilkan pola kesalahan berbeda akibat pembagian fold. Rata-rata probabilitas mengurangi variansi tanpa mencampurkan model yang secara arsitektural terlalu berbeda.

Untuk menilai apakah perbedaan tersebut juga tercermin pada sampel berlabel yang dapat dibandingkan secara langsung, konfigurasi nomor 11 dan konfigurasi nomor 13 dievaluasi pada 1.142 data yang sama. Pada subset ini, F1 konfigurasi nomor 11 sebesar 0,81728 dan konfigurasi 13 sebesar 0,80724, dengan selisih -0,01003. Paired bootstrap sebanyak 10.000 resampling menghasilkan interval kepercayaan 95% untuk selisih F1 sebesar -0,02843 hingga 0,00760, sehingga interval tersebut mencakup nol. Exact McNemar test terhadap 70 pasangan prediksi yang berbeda juga menghasilkan $p = 0,07224$. Hasil ini menunjukkan bahwa perbedaan kinerja kedua konfigurasi belum signifikan secara statistik pada taraf 0,05. Dengan demikian, skor pengujian percobaan konfigurasi 13 yang sedikit lebih tinggi lebih tepat diinterpretasikan sebagai keunggulan numerik yang marginal, bukan sebagai bukti keunggulan statistik yang konklusif.

Weighted ensemble antara model DeBERTa-v3 tanpa metadata dan model metadata fusion menghasilkan 0,83604. Penurunan ini dapat disebabkan korelasi prediksi yang tinggi karena kedua model menggunakan backbone dan pembagian data serupa. Penambahan model yang lebih lemah tidak selalu meningkatkan ensemble; keragaman kesalahan harus diimbangi kualitas masing-masing anggota. Hasil ini menegaskan bahwa soft voting perlu didasarkan pada OOF yang sebanding dan bukan hanya pencarian bobot pada satu validation split.

Peningkatan jumlah fold menjadi lima juga tidak otomatis memperbaiki hasil. Konfigurasi 5-fold dengan layer-wise learning-rate decay, label smoothing, cosine scheduler, dan snapshot ensemble memperoleh 0,83787. Ketika komponen tambahan dihilangkan, 5-fold murni turun menjadi 0,82500. Pada lima fold, setiap model memang menerima data latih lebih banyak, tetapi ukuran validasi per fold lebih kecil dan threshold global dapat bergeser. Selain itu, perubahan seed serta variasi epoch terbaik dapat menghasilkan probabilitas yang berbeda. Hasil tersebut menunjukkan bahwa jumlah fold merupakan hyperparameter eksperimental, bukan jaminan peningkatan performa.

Evaluasi OOF pada 7.613 data berlabel memperkuat bahwa pengaruh jumlah fold tidak seragam pada seluruh metrik. Exp.13 dengan 3-fold memperoleh accuracy 0,84198 dan precision 0,85582, sedangkan Exp.14 dengan 5-fold snapshot dan LLRD menghasilkan recall 0,79150, F1 0,80667, dan ROC-AUC 0,89277, sedikit lebih tinggi secara numerik daripada F1 0,80525 dan ROC-AUC 0,88944 pada Exp.13. Sementara itu, konfigurasi 5-fold murni menghasilkan F1 OOF 0,80354. Temuan ini menunjukkan bahwa peningkatan jumlah fold dapat mengubah keseimbangan antar-metrik, tetapi tidak menghasilkan peningkatan yang konsisten pada kinerja pengujian.

3.4. Analisis Kontribusi Metode

Tabel 3. Peran komponen model yang diusulkan

Komponen	Peran
Mean–max pooling	Merangkum informasi rata-rata dan aktivasi token dominan.
Metadata fusion	Menambahkan sinyal keyword, location, dan pola permukaan.
Multi-sample dropout	Meningkatkan regularisasi classification head.
OOF threshold	Menyeimbangkan precision–recall dari seluruh data latih.
3-fold averaging	Mengurangi variansi prediksi dengan biaya komputasi moderat.

Pencapaian model terbaik tidak berasal dari satu perubahan tunggal, melainkan kombinasi representasi kontekstual, pooling, metadata, regularisasi, dan validasi. Mean pooling menangkap kecenderungan umum seluruh token, sedangkan max pooling mempertahankan aktivasi paling kuat yang mungkin berkaitan dengan istilah bencana. Metadata fusion memperkaya representasi tanpa menggantikan informasi tekstual. Multi-sample dropout membantu mengurangi ketergantungan classifier pada unit tertentu. Terakhir, OOF threshold menyesuaikan batas keputusan terhadap metrik F1

Dari sisi efisiensi, 3-fold menjadi kompromi terbaik pada eksperimen ini. Model hanya dilatih tiga kali, tetapi memberikan skor pengujian eksternal lebih tinggi daripada dua konfigurasi 5-fold. Temuan ini penting untuk penelitian dengan sumber daya terbatas karena peningkatan kompleksitas pelatihan tidak selalu sebanding dengan peningkatan kinerja. Meskipun demikian, hasil masih dipengaruhi variasi data, random seed, dan keterbatasan data uji tersembunyi. Validasi pada dataset bencana lain diperlukan untuk menguji generalisasi lintas kejadian dan lintas waktu.

Tabel 4. Hasil studi ablation metadata

Ablation	F1	ΔF1 vs Full
FULL metadata	**0,83246**	-
− Text length	0,82229	**- 0,01016**
− Keyword	0,82289	**- 0,00956**
− Capitalization	0,82377	**- 0,00868**
− URL	0,82533	- 0,00712
− Mention	0,82577	- 0,00669
− Hashtag	0,82622	- 0,00623
− Digit	0,82661	- 0,00584
− Location	0,82851	- 0,00394
− Punctuation	0,83141	**- 0,00104**

Studi ablation menunjukkan bahwa seluruh kelompok metadata memberikan kontribusi positif terhadap F1 pada evaluasi terkontrol. Konfigurasi penuh menghasilkan F1 sebesar 0,83246. Penurunan terbesar terjadi ketika fitur panjang teks dihilangkan, yaitu sebesar 0,01016, diikuti keyword sebesar 0,00956 dan kapitalisasi sebesar 0,00868. Penghapusan URL, mention, hashtag, dan digit menurunkan F1 masing-masing sebesar 0,00712, 0,00669, 0,00623, dan 0,00584. Sementara itu, location dan tanda baca memberikan perubahan yang lebih kecil, masing-masing sebesar 0,00394 dan 0,00104. Hasil ini menunjukkan bahwa peningkatan metadata fusion tidak bergantung pada satu atribut tunggal, tetapi berasal dari kontribusi beberapa sinyal dengan pengaruh yang berbeda. Karena studi ablation dilakukan pada satu pembagian holdout dan satu seed, besarnya kontribusi setiap fitur diinterpretasikan sebagai karakteristik pada pengaturan eksperimen ini dan bukan sebagai urutan kepentingan yang bersifat universal.

3.5. Analisis Kesalahan dan Keterbatasan

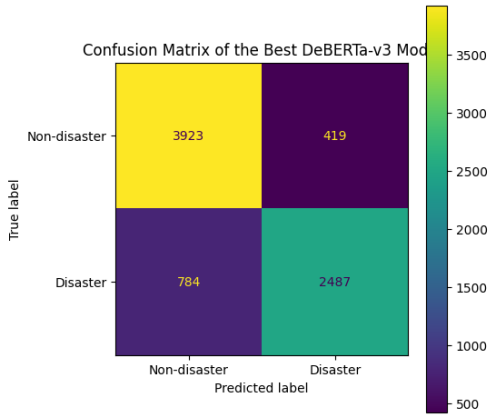
Evaluasi diagnostik terhadap prediksi out-of-fold Konfigurasi nomor 13 pada 7.613 data berlabel menghasilkan accuracy sebesar 0,84198, precision sebesar 0,85582, recall sebesar 0,76032, F1-score sebesar 0,80525, dan ROC-AUC sebesar 0,88944. ROC-AUC digunakan sebagai metrik diagnostik tambahan untuk menilai kemampuan model membedakan kelas berdasarkan probabilitas prediksi tanpa bergantung pada satu threshold tertentu, sedangkan F1 pada data uji tetap digunakan sebagai metrik utama perbandingan antarkonfigurasi.

Tabel 5. Hasil Percobaan konfigurasi No. 13 OOF

Metrik	Hasil
Accuracy	0,84198
Precision	0,85582
Recall	0,76032
F1-score	0,80525
ROC-AUC	0,88944

Confusion matrix menunjukkan bahwa model menghasilkan 3.923 true negative dan 2.487 true positive, dengan 419 false positive dan 784 false negative. Jumlah false negative yang lebih tinggi daripada false positive sejalan dengan nilai precision yang lebih tinggi daripada recall. Dengan kata lain, ketika model menetapkan suatu tweet sebagai bencana, prediksinya relatif presisi, tetapi model masih melewatkan sebagian tweet yang sebenarnya berkaitan dengan bencana. Pola ini menunjukkan bahwa kesalahan utama model lebih banyak berasal dari kegagalan

mendeteksi kelas positif daripada pemberian label bencana secara berlebihan.



Gambar 2. Confusion matrix

Analisis terhadap pola prediksi menunjukkan bahwa kesalahan terutama berpotensi muncul pada tweet dengan kata berkonotasi bencana tetapi digunakan secara metaforis. Ungkapan seperti on fire, blown away, atau disaster dapat memiliki aktivasi token yang kuat, padahal tidak selalu merepresentasikan kejadian aktual. Mean pooling membantu mempertimbangkan konteks keseluruhan, sementara max pooling mempertahankan sinyal token dominan. Kombinasi keduanya mengurangi ketergantungan pada satu jenis agregasi, tetapi belum sepenuhnya mengatasi ironi, humor, kutipan, dan konteks yang berada di luar teks. Kesalahan lain dapat terjadi ketika tweet sangat singkat, hanya berisi tautan, atau menyebut lokasi tanpa menjelaskan peristiwa.

Tabel 6. Contoh kesalahan prediksi OOF konfigurasi 13

Jenis	Keyword	Cuplikan tweet	Aktual	Prediksi	Prob.
FP	hazardous	"Hazardous Weather Outlook ..."	0	1	0,98828
FP	chemical emergency	"Emergency units simulate a chemical ..."	0	1	0,98828
FP	nuclear disaster	"Over half of poll respondents worry nuclear disaster ..."	0	1	0,98584
FN	blaze	"Love living on my own. I can blaze inside my apartment ..."	1	0	0,05167
FN	crashed	"My iPod crashed ..."	1	0	0,05319
FN	sinking	"Do you feel like you are sinking in low self-..."	1	0	0,05310

Contoh kesalahan menunjukkan bahwa false positive terutama muncul ketika tweet mengandung istilah yang sangat dekat dengan domain bencana tetapi tidak secara langsung melaporkan kejadian aktual. Tweet mengenai Hazardous Weather Outlook, simulasi chemical emergency, dan kekhawatiran terhadap nuclear disaster memperoleh probabilitas kelas bencana yang sangat tinggi meskipun berlabel nonbencana. Hal ini menunjukkan bahwa istilah bencana, keyword, dan konteks pendukung dapat menghasilkan sinyal kuat meskipun tweet bersifat peringatan, simulasi, atau diskusi.

Pada false negative ditemukan pola yang berbeda. Beberapa data berlabel bencana menggunakan istilah secara nonliteral atau dalam konteks yang secara semantik tampak nonbencana, seperti *blaze* dalam konteks aktivitas pribadi, *crashed* untuk perangkat elektronik, dan *sinking* sebagai ungkapan mengenai kondisi personal. Model memberikan probabilitas kelas bencana yang sangat rendah pada contoh tersebut. Temuan ini menunjukkan bahwa sebagian kesalahan tidak hanya berkaitan dengan keterbatasan model dalam memahami konteks, tetapi juga dapat dipengaruhi oleh ambiguitas anotasi pada dataset. Oleh karena itu, hasil analisis kesalahan perlu diinterpretasikan dengan mempertimbangkan kualitas dan konsistensi label.

Metadata juga perlu ditafsirkan secara hati-hati. Keberadaan location tidak selalu menunjukkan lokasi kejadian karena kolom tersebut dapat berisi domisili pengguna, frasa kreatif, atau nilai yang tidak dapat dipetakan secara geografis. Keyword dapat membantu menunjukkan topik, tetapi sebagian nilainya bersifat umum dan muncul pada kelas bencana maupun nonbencana. Oleh sebab itu, metadata dalam penelitian ini tidak digunakan sebagai aturan keputusan langsung, melainkan diproyeksikan ke ruang fitur kecil dan digabungkan dengan representasi kontekstual. Desain tersebut membatasi dominasi metadata sekaligus mempertahankan kontribusinya sebagai sinyal komplementer.

Nilai F1 utama pada Tabel 2 berasal dari evaluasi pada data uji terpisah, sehingga perbedaan kecil antarmodel tetap perlu diinterpretasikan secara hati-hati. Analisis statistik pada subset berlabel yang sama menunjukkan bahwa perbedaan antara metadata fusion holdout dan 3-fold OOF belum signifikan pada taraf 0,05, sebagaimana ditunjukkan oleh interval kepercayaan bootstrap yang mencakup nol dan exact McNemar test dengan $p = 0{,}07224$. Namun, pola yang lebih besar, misalnya penurunan pada threshold 0,50 atau 5-fold murni, memberikan indikasi bahwa kalibrasi keputusan dan stabilitas fold berpengaruh terhadap hasil. Repeated stratified cross-validation dapat digunakan pada penelitian lanjutan untuk mengukur variasi antarpelatihan secara lebih sistematis.

Evaluasi beberapa konfigurasi pada sistem pengujian eksternal juga berpotensi menimbulkan model-selection bias karena skor pengujian digunakan dalam menentukan konfigurasi terbaik yang dilaporkan. Oleh karena itu, hasil 0,84063 diinterpretasikan sebagai kinerja terbaik pada benchmark dan pengaturan eksperimen yang digunakan, bukan sebagai estimasi generalisasi yang sepenuhnya independen. Penggunaan held-out test set independen atau nested cross-validation dapat dipertimbangkan pada penelitian selanjutnya.

Penelitian ini juga memiliki keterbatasan domain. Dataset berbahasa Inggris dan berasal dari konteks media sosial tertentu, sehingga model belum tentu langsung berlaku pada bahasa Indonesia, platform lain, atau jenis bencana yang distribusinya berbeda. Perubahan kosakata dari waktu ke waktu dapat menyebabkan concept drift, sedangkan kejadian baru dapat menghadirkan nama tempat, singkatan, atau pola penyebaran informasi yang belum terdapat pada data pelatihan. Evaluasi lintas dataset dan pengujian temporal diperlukan untuk mengetahui apakah metadata fusion tetap bermanfaat ketika distribusi data berubah.

Dari sisi operasional, model DeBERTa-v3 membutuhkan sumber daya komputasi lebih besar daripada TF-IDF atau GloVe–SVM. Oleh karena itu, penerapan waktu nyata dapat menggunakan strategi bertingkat: model ringan menyaring sebagian besar pesan, sedangkan DeBERTa-v3 memproses kasus yang ambigu. Pendekatan distilasi, kuantisasi, atau pruning juga dapat dipertimbangkan agar model lebih mudah diterapkan pada layanan pemantauan dengan keterbatasan latensi dan memori.

3.6. Reproduksibilitas dan Efisiensi Komputasi

Seluruh konfigurasi eksperimen dirancang agar dapat direplikasi pada lingkungan yang sama. Random seed ditetapkan sebelum pembagian data, inisialisasi model, dan pembuatan data loader. Parameter model dipertahankan dalam FP32, sedangkan automatic mixed precision digunakan untuk mengurangi penggunaan memori. Optimizer dibuat setelah model berada pada tipe data yang benar. Setiap fold menggunakan scaler metadata yang hanya dipelajari dari data latihnya, checkpoint dipilih berdasarkan F1 validasi, dan probabilitas OOF disimpan sesuai indeks asal. Prosedur ini penting karena kesalahan indeks atau penggunaan scaler global dapat menghasilkan evaluasi terlalu optimistis.

Untuk memisahkan pengaruh setiap perubahan, eksperimen diberi nomor dan file keluaran berbeda. Setiap konfigurasi menyimpan riwayat epoch, probabilitas validasi, probabilitas data uji, threshold, dan submission biner. Penyimpanan probabilitas memungkinkan evaluasi threshold dan ensemble tanpa pelatihan ulang. Praktik ini mendukung audit eksperimen karena perubahan skor dapat ditelusuri ke arsitektur, skema validasi, atau batas keputusan. AdamW, dropout, dan F1 diterapkan konsisten agar perbandingan tetap berfokus pada representasi dan strategi validasi.

Pada lingkungan NVIDIA Tesla T4, pelatihan konfigurasi 3-fold OOF memerlukan waktu total sekitar 25,25 menit dengan penggunaan memori GPU maksimum 3,90 GB. Pada tahap inferensi, ensemble tiga model memproses 3.263 tweet dalam 25,49 detik atau sekitar 128 tweet per detik, dengan rata-rata waktu inferensi teramortisasi sebesar 7,813 ms per tweet pada batch size 32. Penggunaan memori GPU maksimum selama inferensi mencapai sekitar 3,25 GB. Hasil ini menunjukkan bahwa konfigurasi yang diusulkan masih dapat digunakan untuk pemrosesan tweet secara batch pada lingkungan GPU yang digunakan, meskipun pengujian pada sistem produksi tetap diperlukan untuk mengevaluasi latensi individual dan skalabilitas layanan.

4. KESIMPULAN

Penelitian ini mengembangkan model klasifikasi tweet bencana berbasis DeBERTa-v3 dengan integrasi keyword, location, dan metadata numerik. Representasi token diringkas menggunakan masked mean–max pooling, kemudian digabungkan dengan proyeksi metadata dan diproses melalui multi-sample dropout. Dari 15 konfigurasi yang diuji, pendekatan metadata fusion dengan stratified 3-fold out-of-fold ensemble menghasilkan skor F1 tertinggi sebesar 0,84063. Pada dataset dan pengaturan eksperimen yang digunakan, konfigurasi tersebut memperoleh skor F1 lebih tinggi dibandingkan pendekatan TF-IDF, GloVe, DistilBERT, BERTweet, DeBERTa-v3 tanpa metadata, weighted ensemble, dan dua konfigurasi 5-fold.

Hasil penelitian menunjukkan tiga temuan utama. Pertama, representasi kontekstual transformer memberikan peningkatan nyata dibandingkan embedding statis dan fitur sparse. Kedua, metadata yang sederhana tetap memberikan kontribusi ketika digabungkan secara terkontrol dengan representasi teks. Ketiga, optimasi threshold OOF dan skema 3-fold menghasilkan skor pengujian tertinggi pada eksperimen ini, meskipun evaluasi OOF menunjukkan adanya trade-off antar-metrik dan peningkatan jumlah fold tidak memberikan keuntungan yang konsisten. Model yang diusulkan dapat mendukung lembaga penanggulangan bencana dalam menyaring informasi media sosial yang relevan selama tanggap darurat. Penelitian selanjutnya dapat mengevaluasi kalibrasi probabilitas, repeated cross-validation, adversarial training, pseudo-labeling yang terkontrol, serta pengujian lintas dataset agar ketahanan model terhadap pergeseran domain dapat diukur. Dengan demikian, hasil penelitian menunjukkan efektivitas pendekatan yang diusulkan pada benchmark yang digunakan, tetapi belum dapat digeneralisasikan sebagai keunggulan universal pada dataset atau konteks bencana lainnya. Secara keseluruhan, kontribusi utama penelitian ini terletak pada integrasi DeBERTa-v3, masked mean–max pooling, metadata fusion, dan optimasi threshold berbasis out-of-fold dalam satu kerangka klasifikasi tweet bencana.

5. DAFTAR PUSTAKA

[1] R. I. Ogie, S. James, A. Moore, T. Dilworth, M. Amirghasemi, and J. Whittaker, “Social media use in disaster recovery: A systematic literature review,” *International Journal of Disaster Risk Reduction*, vol. 70, p. 102783, Feb. 2022, doi: 10.1016/J.IJDRR.2022.102783.

[2] T. Acikara, B. Xia, T. Yigitcanlar, and C. Hon, “Contribution of Social Media Analytics to Disaster Response Effectiveness: A Systematic Review of the Literature,” *Sustainability*, vol. 15, no. 11, p. 8860, May 2023,

doi:10.3390/ SU15118860.

[3] N. S. N. Lam *et al.*, "Improving social media use for disaster resilience: challenges and strategies, " *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 3023–3044, Oct. 2023, doi:10.1080/17538947.2023.2239768.

[4] M. Ginzarly, J. Teller, and S. Dujardin, "Social media use in disaster response: empowering community resilience," *Int. J. Digit. Earth*, vol. 18, no. 1, Dec. 2025, doi:10.1080/17538947.2025.2521791.

[5] L. S. Gopal, R. Prabha, H. Thirugnanam, M. V. Ramesh, and B. D. Malamud, "Review article: Social media for managing disasters triggered by natural hazards: a critical review of data collection strategies and actionable insights," *Natural Hazards and Earth System Sciences*, vol. 26, no. 1, pp. 215–250, Jan. 2026, doi: 10.5194/NHESS-26-215-2026.

[6] T. Vishwanath, R. D. Shirwaikar, W. M. Jaiswal, and M. Yashaswini, "Social media data extraction for disaster management aid using deep learning techniques," *Remote Sens. Appl.*, vol. 30, p. 100961, Apr. 2023, doi: 10.1016/J.RSASE.2023.100961.

[7] R. Franceschini, A. Rosi, F. Catani, and N. Casagli, "Detecting information from Twitter on landslide hazards in Italy using deep learning models," *Geoenvironmental Disasters*, vol. 11, no. 1, Jul. 2024, doi: 10.1186/S40677-024-00279-4.

[8] C. He and D. Hu, "Social Media Analytics for Disaster Response: Classification and Geospatial Visualization Framework," *Applied Sciences*, vol. 15, no. 8, p. 4330, Apr. 2025, doi: 10.3390/APP15084330.

[9] K. Maswadi, A. Alhazmi, F. Alshanketi, and C. I. Eke, "The empirical study of tweet classification system for disaster response using shallow and deep learning models," *J. Ambient Intell. Humaniz. Comput.*, vol. 15, no. 9, pp. 3303–3316, May 2024, doi:10.1007/S12652-024-04807-W.

[10] E. Blomeier, S. Schmidt, and B. Resch, "Drowning in the Information Flood: Machine-Learning-Based Relevance Classification of Flood-Related Tweets for Disaster Management ," *Information*, vol. 15, no. 3, p. 149, Mar. 2024, doi: 10.3390/INFO15030149.

[11] W. Khallouli, J. Li, J. Huang, G. Rabadi, and S. Kovacic, "An integrated machine learning approach for identifying emergency rescue messages on social media during natural disasters," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, May 2025, doi: 10.1007/S13278-025-01462-7.

[12] S. Deb and A. K. Chanda, "Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data," *Machine Learning with Applications*, vol. 7, p. 100253, Mar. 2022, doi:10.1016/J .MLWA.2022.100253.

[13] M. L. Jamil, S. Pais, and J. Cordeiro, "Detection of dangerous events on social media: a critical review," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, Oct. 2022, doi: 10.1007/S13278-022-00980-Y.

[14] V. Balakrishnan *et al.*, "A Comprehensive Analysis of Transformer-Deep Neural Network Models in Twitter Disaster Detection," *Mathematics*, vol. 10, no. 24, p. 4664, Dec. 2022, doi: 10.3390/MATH10244664.

[15] S. Wang, R. Li, H. Wu, J. Li, and Y. Shen, "Fine-grained flood disaster information extraction incorporating multiple semantic features," *Int. J. Digit. Earth*, vol. 18, no. 1, Dec. 2025, doi: 10.1080/17538947.2024.2448221.

[16] M. A. R. Noori, B. Sharma, and R. Mehra, "Feature selection from disaster tweets using Spark-based parallel meta-heuristic optimizers," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, Jul. 2022, doi: 10.1007/S13278-022-00930-8.

[17] T. D. N. Pavani and S. J. Malla, "A review of deep learning techniques for disaster management in social media: trends and challenges," *Eur. Phys. J. Spec. Top.*, vol. 234, no. 9, pp. 2803–2825, May 2024, doi:10.1140/ EPJS/S11734-024-01172-9.

[18] M. A. Islam, F. Rabbi, and N. U. I. Hossain, "Performance evaluation of NLP and CNN models for disaster detection using social media data," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, Nov. 2024, doi: 10.1007/S13278-024-01374-Y.

[19] P. B. Yandrapati and R. Eswari, "Classifying informative tweets using feature enhanced pre-trained language model," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, Feb. 2024, doi: 10.1007/S13278-024-01204-1.

[20] D. Hanny, S. Schmidt, S. Gandhi, M. Granitzer, and B. Resch, "A multimodal GeoAI approach to combining text with spatiotemporal features for enhanced relevance classification of social media posts in disaster response," *Big Earth Data*, vol. 10, no. 1, pp. 258–302, Jan. 2026, doi: 10.1080/20964471.2025.2572140.

[21] F. Sufi, "A Sustainable Way Forward: Systematic Review of Transformer Technology in Social-Media-Based Disaster Analytics," *Sustainability*, vol. 16, no. 7, p. 2742, Mar.2024 ,doi: 10.3390/SU16072742.

[22] M. S. I. Malik, M. Z. Younas, M. M. Jamjoom, and D. I. Ignatov, "Categorization of tweets for damages: infrastructure and human damage assessment using fine-tuned BERT model," *PeerJ Comput. Sci.*, vol. 10, p. e1859, Feb. 2024, doi: 10.7717/PEERJ-CS.1859.

[23] R. Koshy and S. Elango, "Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model," *Neural Comput. Appl.*, vol. 35, no. 2, pp. 1607–1627, Oct. 2022, doi:10.1007/S0052 1-022-07790-5.

[24] R. Prasad, A. U. Udeme, S. Misra, and H. Bisallah, "Identification and classification of transportation disaster tweets using improved

bidirectional encoder representations from transformers," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100154, Apr. 2023, doi: 10.1016/J.JJ IMEI.2023.100154.

[25] S. M. Jeba, T. T. Aurpa, and M. R. S. Adib, "From facebook posts to news headlines: using transformer models to predict post-disaster impact on mass media content," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, Oct. 2024, doi: 10.1007/S13278-024-01363-1.

[26] D. Dharrao *et al.*, "An efficient method for disaster tweets classification using gradient-based optimized convolutional neural networks with BERT embeddings," *MethodsX*, vol. 13, p. 102843, Dec. 2024, doi:10.1016/J.MEX.2024.102843.

[27] S. Madichetty, S. M, and S. Madisetty, "A RoBERTa based model for identifying the multi-modal informative tweets during disaster," *Multimed. Tools Appl.*, vol. 82, no. 24, pp. 37615–37633, Mar. 2023, doi:10.1007/S11042-023-14780-9.

[28] I. Sîrbu, R. A. Popovici, T. Rebedea, and Ștefan Trăușan-Matu, "Multihead Average Pseudo-Margin Learning for Disaster Tweet Classifica tion," *Information*, vol. 16, no. 6, p. 434, May 2025, doi: 10.3390/INFO16060434.

[29] L. Zou, Z. He, C. Zhou, and W. Zhu, "Multi-class multi-label classification of social media texts for typhoon damage assessment: a two-stage model fully integrating the outputs of the hidden layers of BERT," *Int. J. Digit. Earth*, vol. 17, no. 1, Dec. 2024, doi:10.108 0/17538947.2024.2348668.

[30] Z. He, C. Zhou, L. Zou, S. Zhou, and X. Zhao, "Siamese text classification network (SiamTCN) for multi-class multi-label information extraction of typhoon disasters from social media data," *Int. J. Digit. Earth*, vol. 18, no. 1, Dec. 2025, doi:10.1080/17538947.2025.2517790.

[31] D. S. Krishna, G. Srinivas, and P. V. G. D. Prasad Reddy, "A Deep Parallel Hybrid Fusion Model for disaster tweet classification on Twitter data," *Decision Analytics Journal*, vol. 11, p. 100453, Jun. 2024, doi:10.1016/J.DAJOUR.2024.100453.

[32] N. P. Shetty, Y. Bijalwan, P. Chaudhari, J. Shetty, and B. Muniyal, "Disaster assessment from social media using multimodal deep learning," *Multimed. Tools Appl.*, vol. 84, no. 18, pp. 18829–18854, Jul. 2024, doi: 10.1007/S11042-024-19818-0.

[33] S. Gite *et al.*, "Analysis of Multimodal Social Media Data Utilizing VIT Base 16 and GPT-2 for Disaster Response," *Arab. J. Sci. Eng.*, vol. 50, no. 23, pp. 19805–19823, May 2025, doi: 10.1007/S13369-025-10314-7.

[34] J. Lv, X. Wang, and C. Shao, "AMAE: Adversarial multimodal auto-encoder for crisis-related tweet analysis," *Computing*, vol. 105, no. 1, pp. 13–28, Aug. 2022, doi: 10.1007/S00607-022-01098-X.

[35] M. Rezk, N. Elmadany, R. K. Hamad, and E. F. Badran, "Categorizing Crises from Social Media Feeds via Multimodal Channel Attention," *IEEE Access*, vol. 11, pp. 72037–72049, 2023, doi: 10.1109/ACCESS.2023.3294474.

[36] A. Khattar and S. M. K. Quadri, "CAMM: Cross-Attention Multimodal Classification of Disaster-Related Tweets," *IEEE Access*, vol. 10, pp. 92889–92902, 2022, doi: 10.1109/ACCESS.2022.3202976.

[37] D. Adwaith, A. K. Abishake, S. V. Raghul, and E. Sivasankar, "Enhancing multimodal disaster tweet classification using state-of-the-art deep learning networks," *Multimed. Tools Appl.*, vol. 81, no. 13, pp. 18483–18501, Mar. 2022, doi: 10.1007/S11042-022-12217-3.

[38] S. Schmidt, M. Friedemann, D. Hanny, B. Resch, T. Riedlinger, and M. Mühlbauer, "Enhancing satellite-based emergency mapping: Identifying wildfires through geo-social media analysis," *Big Earth Data*, vol. 9, no. 3, pp. 389–411, Jul. 2025, doi:10.1080/20964471.2025.2454526.

[39] D. Hanny, K. Ghosh Dastidar, M. Wieland, M. Granitzer, and B. Resch, "Towards multimodal geospatial reasoning: a foundation model approach for disaster detection from social media, news, and weather data," *Natural Hazards*, vol. 122, no. 11, May 2026, doi:10.1007/S11069-026-08191-W.

[40] D. Effrosynidis, G. Sylaios, and A. Arampatzis, "The Effect of Training Data Size on Disaster Classification from Twitter," *Information*, vol. 15, no. 7, p. 393, Jul. 2024, doi:10.3390/I NFO15070393.

[41] Addison Howard, Devrishi, Phil Culliton, and Yufeng Guo, "Natural Language Processing with Disaster Tweets." Accessed: Aug. 18, 2026. [Online]. Available: https://www.kaggle.com/competitions/nlpgetting-started/overview