Analisis Performa IndoBERT sebagai Feature Extractor dan SVM sebagai Classifier dalam Klasifikasi Multilabel Topik Hadits Terjemahan Bahasa Indonesia

1st Abiyyu Abdurrafi Ahmad
Department of Informatics
Telkom University
Bandung, Indonesia
abiyyuabdurrafi@student.telkomuniversity.ac.id

2nd Kemas Muslim Lhaksmana
Department of Informatics
Telkom University
Bandung, Indonesia
kemasmuslim@telkomuniversity.ac.id

***Abstrak* — Hadits sebagai sumber ajaran Islam memiliki kandungan makna yang beragam, seperti anjuran, larangan, dan informasi. Pada era digital, teks hadits terjemahan beredar luas tanpa anotasi topik yang jelas, sehingga menyulitkan pembaca dalam memahami pesan yang terkandung. Kondisi ini menimbulkan kebutuhan akan sistem klasifikasi otomatis yang mampu menangani lebih dari satu kategori makna dalam satu teks. Penelitian ini mengusulkan kombinasi IndoBERT sebagai ekstraktor fitur dan Support Vector Machine (SVM) sebagai pengklasifikasi untuk tugas klasifikasi multilabel topik hadits terjemahan Bahasa Indonesia. Pendekatan ini dipilih karena IndoBERT mampu menghasilkan representasi kontekstual yang kaya, sementara SVM efektif dalam memproses fitur berdimensi tinggi secara efisien. Tahapan penelitian mencakup praproses teks, ekstraksi embedding IndoBERT menggunakan strategi pooling CLS, mean, dan max, pelatihan model SVM, serta evaluasi menggunakan Hamming Loss, Micro F1, dan Macro F1. Hasil menunjukkan bahwa strategi mean pooling memberikan performa terbaik, dengan nilai Macro F1 mencapai 0.71649 dan mengungguli baseline TF-IDF + SVM. Temuan ini membuktikan bahwa pemanfaatan embedding IndoBERT dapat meningkatkan pemerataan performa antar label dan efektif digunakan dalam klasifikasi multilabel teks hadits.**

***Kata kunci*— klasifikasi multilabel, hadits, IndoBERT, SVM, ekstraksi fitur, NLP**

I. PENDAHULUAN

Hadits merupakan sumber ajaran Islam yang memiliki kedudukan penting setelah Al-Qur'an dan mengandung berbagai pesan keagamaan yang mencakup aspek hukum, akhlak, ibadah, serta penjelasan umum bagi umat Muslim. Karakteristik hadits yang kaya makna menyebabkan satu teks hadits sering kali memuat lebih dari satu pesan, seperti anjuran, larangan, dan informasi secara bersamaan. Seiring dengan perkembangan teknologi digital, teks hadits terjemahan Bahasa Indonesia semakin mudah diakses melalui berbagai platform daring. Namun, kemudahan akses ini juga menimbulkan tantangan, karena banyak hadits disebarkan tanpa konteks yang memadai, sehingga menyulitkan pembaca dalam memahami jenis pesan yang terkandung dan berpotensi menimbulkan kesalahpahaman, terutama ketika hadits disajikan secara terpotong.

Klasifikasi otomatis terhadap teks hadits menjadi solusi yang relevan untuk mendukung pemahaman yang lebih terstruktur. Tantangan utama dalam klasifikasi hadits terletak pada sifatnya yang multi-makna, sehingga pendekatan multilabel *classification* dinilai lebih sesuai dibandingkan pendekatan *single-label*. Pendekatan multilabel banyak digunakan pada teks yang memiliki lebih dari satu tujuan semantik, namun juga menghadapi tantangan berupa tumpang tindih antar label dan ketidakseimbangan distribusi data [1], [2].

Dalam konteks pemrosesan bahasa alami, berbagai penelitian menunjukkan bahwa model berbasis *deep learning* dan transformer memberikan performa yang lebih baik dibandingkan metode tradisional. Model BERT dan arsitektur transformer terbukti mampu menangkap konteks kata secara dua arah dan memahami hubungan semantik yang kompleks dalam teks [3], [4]. Untuk Bahasa Indonesia, IndoBERT dikembangkan sebagai model pralatih yang secara khusus menyesuaikan karakteristik linguistik lokal dan telah berhasil digunakan dalam berbagai tugas klasifikasi seperti analisis emosi dan sentimen [5], [6].

Penelitian terkait klasifikasi hadits juga telah dilakukan dengan berbagai pendekatan. Studi sebelumnya memanfaatkan metode Support Vector Machine (SVM) [7] serta Random Forest untuk klasifikasi multilabel topik hadits terjemahan Bahasa Indonesia [8]. Selain itu, pendekatan *deep learning* non-transformer seperti Recurrent Convolutional Neural Network (CRNN) juga telah diterapkan pada klasifikasi teks hadits dan menunjukkan performa yang cukup baik [9]. Meskipun demikian, pendekatan tersebut masih memiliki keterbatasan dalam menangkap konteks semantik secara menyeluruh dibandingkan model berbasis transformer.

Di sisi lain, metode ekstraksi fitur konvensional seperti TF-IDF masih sering digunakan sebagai baseline karena kesederhanaannya. Namun, studi komparatif menunjukkan bahwa metode berbasis embedding transformer secara signifikan mengungguli TF-IDF dalam klasifikasi teks modern [10], [11]. Salah satu pendekatan yang dinilai efisien

adalah menggunakan transformer sebagai *feature extractor* tanpa melakukan *fine-tuning* penuh, kemudian mengombinasikannya dengan algoritma klasifikasi yang lebih ringan seperti SVM. Beberapa penelitian menunjukkan bahwa kombinasi embedding transformer dan SVM mampu memberikan performa kompetitif dengan kebutuhan komputasi yang lebih rendah [12].

Meskipun berbagai penelitian telah dilakukan, hingga saat ini belum ditemukan penelitian yang secara khusus mengkaji kombinasi IndoBERT sebagai *feature extractor* dan SVM sebagai *classifier* untuk klasifikasi multilabel topik hadits terjemahan Bahasa Indonesia. Oleh karena itu, penelitian ini bertujuan untuk mengisi celah tersebut dengan mengevaluasi efektivitas kombinasi IndoBERT dan SVM serta menganalisis strategi pooling yang paling optimal berdasarkan metrik F1 macro, F1 micro, dan Hamming loss.

II. KAJIAN TEORI

A. Klasifikasi Teks

Klasifikasi teks merupakan proses pengkategorian suatu teks ke dalam kelas tertentu berdasarkan isi dari teks tersebut. Penerapannya dapat dijumpai di berbagai bidang seperti analisis sentimen, pendeteksian spam, pengkategorian konten sosial media, hingga analisis konten keagamaan.

Klasifikasi teks memiliki dua pendekatan utama, yaitu single-label dan multilabel. Single-label memiliki makna satu teks atau kalimat hanya dikategorikan ke satu kelas, sementara multilabel, satu teks atau kalimat dapat dikategorikan ke dalam lebih dari satu kelas secara bersamaan. Dalam klasifikasi teks hadits, lebih sesuai menggunakan klasifikasi multilabel, karena dalam praktiknya, satu hadits dapat mengandung lebih dari satu makna [8].

B. Representasi Fitur Teks

Representasi fitur merupakan tahapan penting dalam sistem klasifikasi teks. TF-IDF merupakan salah satu metode representasi yang paling umum digunakan dan telah terbukti efektif dalam berbagai penelitian [5], [11]. Namun, metode ini memiliki keterbatasan karena tidak dapat menangkap konteks semantik.

Sebaliknya, perkembangan model transformer seperti BERT, RoBERTa, dan varian lain memungkinkan representasi kata berbasis konteks yang jauh lebih kaya. BERT bekerja secara bidirectional, sehingga mampu memahami hubungan antar kata baik dari sisi kiri maupun kanan kalimat [10]. IndoBERT, sebagai model pralatih khusus Bahasa Indonesia, telah terbukti memberikan peningkatan signifikan pada berbagai tugas NLP Bahasa Indonesia [6]. Representasi embedding dari transformer juga dapat dipadukan dengan *classifier* tradisional seperti SVM, menghasilkan pendekatan hybrid yang efisien tetapi tetap berkinerja tinggi [11].

C. IndoBERT

IndoBERT merupakan model bahasa berbasis arsitektur transformer yang dikembangkan secara khusus untuk Bahasa Indonesia. Model ini dilatih pada korpus Bahasa Indonesia berskala besar sehingga mampu menangkap pola bahasa, relasi antar kata, dan konteks kalimat dengan lebih baik. IndoBERT juga dilaporkan efektif pada berbagai tugas NLP berbahasa Indonesia, seperti klasifikasi sentimen maupun klasifikasi teks pada domain tertentu [13]. Selain itu, terdapat studi yang menunjukkan bahwa model pralatih Bahasa Indonesia dapat diadaptasi lebih jauh untuk domain spesifik melalui modifikasi *vocabulary* maupun *pretraining* lanjutan [14].

Dalam penelitian ini, IndoBERT digunakan sebagai metode embedding, di mana setiap teks hadits diubah menjadi representasi vektor berdimensi tinggi. Proses ini dimulai dengan tokenisasi menggunakan tokenizer IndoBERT, yang memecah kalimat menjadi token sub-word. Setelah itu, model menghasilkan embedding untuk masing-masing token. Dari keseluruhan output tersebut, digunakan representasi dari token khusus, yang secara konvensional dianggap merepresentasikan makna global dari seluruh kalimat.

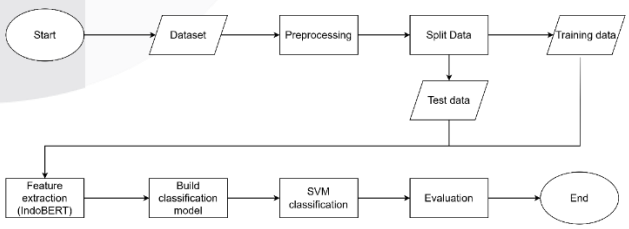
Hasil akhir dari proses ini adalah vektor numerik yang merepresentasikan isi semantik dari hadits. Vektor inilah yang kemudian digunakan sebagai input fitur dalam proses klasifikasi menggunakan SVM di tahap selanjutnya.

D. SVM

SVM merupakan algoritma machine learning yang populer untuk tugas klasifikasi, khususnya klasifikasi teks. Algoritma ini bekerja dengan cara mencari hyperplane terbaik yang memisahkan dua kelas atau lebih dengan margin terbesar. Keunggulan SVM terletak pada kemampuannya dalam menangani data berdimensi tinggi dan tetap mempertahankan performa yang stabil meski ukuran dataset yang digunakan berjumlah sedikit [15].

Dalam konteks klasifikasi hadits, penelitian yang dilakukan oleh Kusumaningrum et al. (2017) menunjukkan SVM mampu memberikan nilai akurasi yang lebih tinggi dibanding pendekatan rule-based sederhana, meskipun hasilnya masih terbatas. Selain itu, penelitian yang dilakukan oleh Surianto et al. (2023) membandingkan performa SVM dan Artificial Neural Network (ANN) dalam klasifikasi hadits, dan hasilnya menunjukkan bahwa SVM lebih unggul dari ANN dalam hal stabilitas dan akurasi, baik pada data yang telah dilakukan stemming maupun tanpa stemming. Hasil ini memperkuat posisi SVM sebagai baseline yang cukup kuat dalam klasifikasi teks hadits [16].

III. METODE



GAMBAR 1
ALUR PEMODELAN

Berdasarkan gambar 1, dapat dilihat rincian proses utama yang dilakukan, mulai dari pengumpulan dataset, *preprocessing*, ekstraksi fitur, pengembangan model klasifikasi, serta evaluasi model.

A. Dataset

Dataset yang digunakan merupakan dataset yang didapat dari Telkom University Dataverse, berisi kumpulan hadits terjemahan bahasa Indonesia berjumlah 7000 data yang sudah diberi label secara multi-topik. Satu hadits dapat dilabeli lebih dari satu kategori sehingga sesuai dengan pendekatan multilabel. Label yang digunakan dalam penelitian ini antara lain: anjuran, larangan, dan informasi. Jumlah data per kelas dapat dilihat pada tabel berikut:

TABEL 1
JUMLAH KEMUNCULAN LABEL

Label	Jumlah Data
Anjuran	1335
Larangan	844
Informasi	6634

Data diperoleh dari penelitian sebelumnya yang juga meneliti topik klasifikasi multilabel, sehingga layak digunakan dalam pengembangan dan pengujian model klasifikasi topik hadits.

B. Preprocessing Data

Tahap praproses dilakukan untuk membersihkan dan menormalisasi teks hadits sebelum masuk ke proses ekstraksi fitur. Langkah-langkah yang diterapkan mencakup konversi seluruh teks menjadi huruf kecil untuk menyamakan bentuk kata, normalisasi spasi untuk menghapus spasi ganda, serta pembersihan simbol, tanda baca, dan karakter non-alfanumerik menggunakan ekspresi reguler. Nilai kosong juga ditangani dengan menggantinya menjadi string kosong agar setiap baris dapat diproses tanpa error. Tahapan ini memastikan bahwa teks berada dalam format yang bersih dan terstandardisasi, sehingga memudahkan proses tokenisasi dan ekstraksi fitur pada langkah selanjutnya.

Selain pembersihan dasar, tersedia opsi stemming menggunakan Sastrawi untuk mengembalikan kata ke bentuk dasarnya. Namun, proses stemming tidak diterapkan pada eksperimen utama karena model transformer seperti IndoBERT bekerja lebih optimal ketika struktur kalimat dipertahankan utuh. Stopword juga tidak dihapus karena dalam konteks hadits, kata-kata seperti "jangan", "tidak", dan "hendaklah" sering menjadi penanda penting kategori larangan atau anjuran. Model transformer sendiri mampu mengabaikan token tidak penting melalui mekanisme self-attention, sehingga penghapusan stopword tidak diperlukan dan justru dapat menghilangkan informasi semantik yang relevan.

C. Split Data

Split data merupakan proses membagi data menjadi 2 bagian, yaitu data latih dan data uji dengan perbandingan 80:20. Data latih berguna untuk melatih model yang akan digunakan, sedangkan data uji berguna untuk menguji performa model yang telah dilatih menggunakan data latih.

D. Ekstraksi Fitur

Ekstraksi fitur menggunakan IndoBERT merupakan tahap penting dalam penelitian ini karena model transformer mampu menghasilkan representasi teks yang bersifat kontekstual. Berbeda dengan metode tradisional seperti TF-IDF yang hanya mempertimbangkan frekuensi kata, IndoBERT memproses kalimat secara menyeluruh menggunakan mekanisme self-attention sehingga setiap kata direpresentasikan berdasarkan hubungan maknanya dengan kata lain dalam kalimat [3]. Pendekatan ini sangat relevan untuk teks hadits yang memiliki struktur bahasa kompleks dan makna yang padat. Dalam penelitian ini, embedding IndoBERT diperoleh dari last hidden state model, kemudian diubah menjadi vektor representasi dokumen melalui beberapa strategi pooling. Untuk menjaga fokus penelitian, tiga strategi pooling utama digunakan, yaitu CLS Token, Mean Pooling, dan Max Pooling. Berikut adalah penjelasan setiap strategi pooling:

a. CLS Token

Strategi CLS memanfaatkan token khusus [CLS] pada awal input sebagai representasi global dokumen setelah diproses oleh seluruh lapisan transformer. Pendekatan ini sederhana dan efisien, namun performanya bergantung pada optimalisasi token CLS oleh model pralatih, sehingga kurang stabil pada teks panjang.

b. Mean Pooling

Mean pooling menghitung rata-rata seluruh embedding token, sehingga setiap kata berkontribusi terhadap representasi akhir. Strategi ini menghasilkan embedding yang lebih stabil dan konsisten untuk teks panjang dan kompleks, serta sering menunjukkan performa yang baik pada tugas klasifikasi teks .

c. Max Pooling

Max pooling mengambil nilai maksimum pada setiap dimensi embedding untuk menekankan fitur paling dominan dalam teks. Pendekatan ini efektif dalam menangkap token bermuatan semantik kuat, seperti kata perintah atau larangan, namun rentan terhadap noise dari token yang tidak relevan.

E. SVM

Proses klasifikasi dilakukan setelah setiap teks hadits direpresentasikan dalam bentuk vektor menggunakan metode ekstraksi fitur. Pada penelitian ini, algoritma Support Vector Machine (SVM) digunakan sebagai model klasifikasi utama karena kemampuannya dalam menangani data berdimensi tinggi serta kestabilannya pada berbagai jenis representasi teks. SVM bekerja dengan mencari sebuah hyperplane yang memisahkan data dari dua kelas berbeda dengan margin terbesar. Untuk satu label biner, fungsi keputusan dapat dituliskan sebagai:

$$f(x) = w^T x + b \quad (1)$$

dimana x merupakan vektor fitur dokumen, w adalah bobot model, dan b adalah bias. Nilai f(x) menentukan apakah suatu dokumen masuk ke kelas positif atau negatif pada skema klasifikasi biner.

Klasifikasi multilabel diimplementasikan menggunakan algoritma SVM dengan pendekatan Binary Relevance melalui skema One-vs-Rest (OvR). Dalam skema ini, satu model SVM dilatih secara independen untuk setiap label (Anjuran, Larangan, dan Informasi). Optimasi performa dilakukan menggunakan Grid Search Cross-Validation guna mencari nilai parameter regularisasi (C) terbaik secara

spesifik bagi masing-masing model label. Penentuan prediksi akhir didasarkan pada decision boundary standar SVM (nilai fungsi keputusan positif), tanpa menerapkan penyesuaian ambang batas (threshold tuning) pada probabilitas output.

Selain itu, parameter class_weight='balanced' diaktifkan untuk menangani ketidakseimbangan data dengan memberikan bobot penalti proporsional pada kesalahan klasifikasi kelas minoritas. Kernel Linear dipilih sebagai fungsi pemetaan karena karakteristik fitur teks hasil ekstraksi IndoBERT yang berdimensi tinggi cenderung bersifat linearly separable. Kombinasi kernel linear dan penanganan imbalance ini memastikan model dapat bekerja secara efisien sekaligus efektif dalam memisahkan setiap kategori topik hadits.

F. Evaluasi

Tahap evaluasi dilakukan untuk menilai performa model klasifikasi multilabel dalam memprediksi kategori topik hadits. Evaluasi ini bertujuan mengukur kemampuan sistem dalam mengenali lebih dari satu label pada setiap teks hadits secara akurat. Tiga metrik evaluasi digunakan dalam penelitian ini, yaitu Hamming Loss, Micro F1-Score, dan Macro F1-Score.

Hamming Loss digunakan untuk mengukur proporsi kesalahan prediksi label terhadap keseluruhan label yang mungkin, di mana nilai yang lebih rendah menunjukkan performa model yang lebih baik. Sementara itu, F1-Score digunakan untuk menyeimbangkan antara precision dan recall dalam konteks klasifikasi multilabel. Micro F1-Score menghitung performa secara global dengan mempertimbangkan seluruh label dan instance, sehingga cenderung merefleksikan kinerja model pada kelas mayoritas. Sebaliknya, Macro F1-Score menghitung performa pada setiap label secara terpisah dan kemudian dirata-ratakan, sehingga mampu memberikan gambaran yang lebih adil terhadap kinerja model pada kelas minoritas.

Perbedaan nilai antara Micro dan Macro F1-Score digunakan sebagai indikator kemampuan model dalam menangani ketidakseimbangan distribusi data. Dalam penelitian ini, Macro F1-Score dijadikan tolok ukur utama keberhasilan model karena tujuan sistem adalah mengklasifikasikan seluruh topik hadits secara merata, bukan hanya pada kategori yang dominan. Evaluasi dilakukan menggunakan data uji yang terpisah dari data pelatihan untuk menjaga objektivitas hasil.

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

Rangkaian pengujian dilakukan untuk menjawab dua pertanyaan utama penelitian: strategi representasi fitur (pooling) mana yang paling optimal pada IndoBERT, dan seberapa efektif metode yang diusulkan dibandingkan dengan metode konvensional.

Pengujian pertama difokuskan pada *Ablation Study* untuk membandingkan tiga strategi *pooling* dasar: CLS Token, *Mean Pooling*, dan *Max Pooling*. Eksperimen ini bertujuan melihat pengaruh metode perangkuman vektor kata (*token embeddings*) terhadap akurasi klasifikasi akhir. Hasil perbandingan performa ketiga strategi tersebut disajikan secara lengkap pada Tabel 2:

TABEL 2
HASIL *ABLATION STUDY*

Pooling	Macro F1	Micro F1	Hamming Loss	Waktu (s)
Mean	0.7165	0.8400	0.1407	381.66
CLS	0.6883	0.8098	0.1667	445.32
Max	0.6527	0.7834	0.1890	385.51

Berdasarkan tabel tersebut, strategi Mean Pooling memberikan nilai Macro F1 tertinggi, yaitu 0.7165, dengan Micro F1 sebesar 0.8400, serta Hamming Loss terendah sebesar 0.1407. Perbandingan antar strategi menunjukkan bahwa Mean Pooling menghasilkan representasi yang lebih stabil dan menyeluruh karena mempertimbangkan seluruh token dalam kalimat secara merata. Sementara CLS Token hanya mengandalkan satu token (token awal), dan Max Pooling cenderung menyoroti token-token dengan aktivasi paling kuat, yang tidak selalu mencerminkan konteks teks hadits secara utuh. Dengan demikian, Mean Pooling dipilih sebagai strategi embedding utama pada model usulan.

Setelah strategi pooling ditetapkan, model diuji lebih lanjut untuk melihat performa prediksinya terhadap ketiga label multilabel: anjuran, larangan, dan informasi. Hasil prediksi diperlihatkan melalui tabel confusion matrix yang ditampilkan pada tabel 3.

TABEL 3
CONFUSION MATRIX MODEL USULAN

Label	TP	FP	TN	FN
Anjuran	197	196	937	70
Larangan	133	133	1099	35
Informasi	1221	51	22	106

Dari Tabel 3, terlihat bahwa model memiliki kecenderungan yang berbeda antar label. Pada label Larangan, jumlah False Negative sangat rendah (35 data), yang berarti model sangat jarang melewatkan ayat larangan. Namun, pada label Anjuran, tingkat False Positive cukup tinggi (196 data), mengindikasikan bahwa model terkadang "terlalu agresif" menganggap sebuah kalimat sebagai anjuran padahal bukan. Ringkasan metrik evaluasi per label ditampilkan pada Tabel 4 berikutnya, sebagai pelengkap interpretasi visual dari confusion matrix.

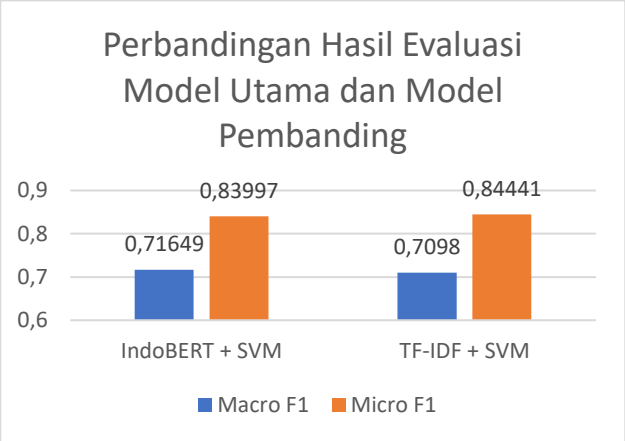
TABEL 4
HASIL EVALUASI MODEL USULAN

Label	Metrik Evaluasi			
	F1 Score	Precision	Recall	Accuracy
Anjuran	0.5970	0.5013	0.7378	0.8100
Larangan	0.6129	0.5000	0.7917	0.8800
Informasi	0.9396	0.9599	0.9201	0.8879

Dari tabel tersebut terlihat bahwa label informasi memperoleh nilai F1 tertinggi sebesar 0.9396, dengan precision tinggi (0.9599) dan recall tinggi (0.9201), menunjukkan bahwa model mampu mengidentifikasi pola informasi dengan sangat baik. Label larangan menunjukkan hasil moderat dengan F1 sebesar 0.6129, sementara label

anjuran memperoleh nilai F1 terendah yaitu 0.5970. Perbedaan performa antar label ini berkaitan erat dengan distribusi data dan karakteristik linguistik tiap kategori.

Untuk mengetahui posisi model usulan dibandingkan model baseline, dilakukan perbandingan kinerja yang ditampilkan pada Gambar 2. Grafik ini menampilkan nilai Macro F1 dan Micro F1 dari kedua model.



GAMBAR 2
PERBANDINGAN MODEL USULAN DAN MODEL PEMBANDING

Gambar 2 menunjukkan bahwa model SVM + IndoBERT unggul tipis pada Macro F1 (0.71649) dibandingkan baseline TF-IDF (0.70980). Keunggulan ini menandakan bahwa representasi semantik IndoBERT lebih efektif dalam mendeteksi kelas minoritas (Anjuran dan Larangan) yang pemahamannya bergantung pada konteks kalimat, bukan sekadar kata kunci. Sebaliknya, baseline TF-IDF memimpin pada Micro F1 (0.84441), serta jauh lebih efisien dari sisi biaya komputasi.

Analisis kompleksitas komputasi menunjukkan perbedaan yang sangat signifikan pada tahap ekstraksi fitur. Metode IndoBERT membutuhkan waktu 79,87 detik untuk memproses teks menjadi vektor embedding (termasuk waktu loading model), sedangkan metode TF-IDF hanya memerlukan waktu 0,76 detik. Kesenjangan drastis ini terjadi karena IndoBERT melakukan komputasi matriks yang kompleks pada setiap lapisan transformer, berbeda dengan TF-IDF yang hanya berbasis perhitungan statistik frekuensi kata sederhana. Implikasi dari kompleksitas fitur ini berlanjut ke tahap klasifikasi, di mana pelatihan SVM pada fitur IndoBERT memakan waktu 301,79 detik, jauh lebih lama dibandingkan pelatihan pada fitur TF-IDF yang selesai dalam 69,13 detik.

Meskipun baseline lebih unggul dalam efisiensi dan akurasi global, penggunaan IndoBERT tetap dinilai lebih strategis. Dalam domain hadits, kemampuan mendeteksi label hukum (Anjuran/Larangan) secara akurat dianggap memiliki nilai informasi yang lebih tinggi dibandingkan sekadar memprediksi kelas mayoritas (Informasi). Oleh karena itu, kompleksitas komputasi yang tinggi pada IndoBERT dianggap sebagai investasi yang sepadan untuk mendapatkan pemerataan performa antar kelas yang lebih baik, sebagaimana tercermin dari keunggulan nilai Macro F1.

B. Analisis Hasil Pengujian

Hasil evaluasi menunjukkan bahwa pendekatan SVM + IndoBERT memberikan performa yang kompetitif dalam tugas klasifikasi multilabel hadits, ditandai dengan nilai Macro F1 sebesar 0.71649 yang lebih tinggi dibanding baseline. Keunggulan ini berasal dari kemampuan embedding IndoBERT dalam menangkap konteks semantik yang kompleks pada teks hadits sehingga model dapat menghasilkan prediksi yang lebih seimbang pada seluruh label. Namun, nilai Micro F1 model utama sedikit lebih rendah daripada baseline, menunjukkan bahwa akurasi keseluruhan masih dipengaruhi oleh dominasi label informasi dalam dataset.

Ketidakseimbangan data menjadi faktor paling berpengaruh terhadap performa model. Label informasi yang jumlahnya jauh lebih besar membuat model lebih mudah mempelajari pola-pola naratif khas kategori ini, sehingga menghasilkan F1-score yang tinggi baik pada model utama maupun baseline. Sebaliknya, label anjuran dan larangan memiliki jumlah sampel lebih sedikit dan struktur bahasa yang lebih beragam, sehingga model kesulitan mengenali pola-pola khas kedua label ini. Kondisi ini tercermin pada confusion matrix, di mana jumlah *False Positive* (FP) dan *False Negative* (FN) pada kedua label minoritas cukup tinggi.

Contoh kesalahan prediksi memperjelas pola tersebut. Pada label anjuran, model memberikan FP pada kalimat seperti “ia tidak boleh memakai baju imamah surban …”, yang seharusnya merupakan larangan. Kata-kata bernuansa normatif membuat model keliru mengasosiasikannya dengan anjuran. Sebaliknya, FN terjadi pada kalimat seperti “Hendaklah kalian berbuat baik kepada tetangga.”, yang diprediksi sebagai informasi karena strukturnya menyerupai pola naratif umum. Kesalahan serupa terjadi pada label larangan, misalnya pada kalimat “siapa yang menuduh budaknya berzina …” yang diprediksi sebagai larangan karena keberadaan kata “berzina”, padahal konteksnya informatif. Sementara itu, label informasi diprediksi dengan sangat baik (TP = 1221) karena pola bahasanya dominan, stabil, dan paling sering ditemukan dalam dataset.

Secara keseluruhan, model utama menunjukkan performa yang baik dari sisi pemerataan prediksi antar label, namun belum optimal untuk label minoritas yang membutuhkan pemahaman semantik lebih dalam. Kesalahan model sebagian besar muncul akibat ketidakseimbangan data dan tumpang tindih makna antar label, terutama pada teks yang mengandung ekspresi normatif namun tidak merujuk pada perintah atau larangan eksplisit.

V. KESIMPULAN

Penelitian ini mengevaluasi efektivitas IndoBERT sebagai ekstraktor fitur yang dikombinasikan dengan SVM dalam tugas klasifikasi multilabel topik hadits terjemahan Bahasa Indonesia. Berdasarkan hasil ablation study, strategi Mean Pooling memberikan performa terbaik dalam menghasilkan representasi fitur, dan model utama yang menggunakan embedding IndoBERT menunjukkan pemerataan performa yang lebih baik antar label dibanding baseline TF-IDF + SVM, tercermin dari nilai Macro F1 sebesar 0.71649. Namun, performa model masih dipengaruhi oleh ketidakseimbangan label, terutama pada kategori

anjuran dan larangan yang memiliki jumlah sampel lebih sedikit dan variasi bahasa lebih kompleks.

Meskipun baseline TF-IDF unggul pada Micro F1 dan Hamming Loss karena dominasi label informasi, hasil keseluruhan menunjukkan bahwa IndoBERT memiliki kemampuan yang lebih kuat dalam memahami konteks semantik yang kaya pada teks hadits. Analisis confusion matrix memperlihatkan bahwa model sangat baik dalam mengenali label informasi, namun masih menghadapi tantangan dalam membedakan pola anjuran dan larangan, yang menghasilkan nilai False Positive dan False Negative lebih tinggi. Hal ini menegaskan pentingnya pemilihan strategi representasi fitur dan penanganan ketidakseimbangan data dalam meningkatkan kualitas klasifikasi multilabel.

Berdasarkan temuan tersebut, penelitian selanjutnya disarankan untuk mengeksplorasi strategi pooling yang lebih kaya, seperti kombinasi CLS–mean atau agregasi multilayer. Selain itu, pendekatan *domain adaptation* melalui pelatihan lanjutan IndoBERT pada korpus hadits dapat dilakukan agar model lebih memahami karakteristik bahasa keagamaan yang khas.

REFERENSI

[1] W. Hu, Q. Fan, H. Yan, X. Xu, S. Huang, dan K. Zhang, "A Survey of Multi-Label Text Classification Under Few-Shot Scenarios," *Appl. Sci.*, vol. 15, no. 16, hlm. 8872, Agu 2025, doi: 10.3390/app15168872.

[2] S. P. K. Veeranki, A. Abdulnazar, D. Kramer, M. Kreuzthaler, dan D. B. Lumenta, "Multi-label text classification via secondary use of large clinical real-world data sets," *Sci. Rep.*, vol. 14, no. 1, hlm. 26972, Nov 2024, doi: 10.1038/s41598-024-76424-8.

[3] J. Devlin, M.-W. Chang, K. Lee, dan K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," dipresentasikan pada Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, hlm. 4171–4186.

[4] A. Vaswani *dkk.*, "Attention Is All You Need," dipresentasikan pada Neural Information Processing Systems, Long Beach, CA, USA, 2017, hlm. 5998–6008.

[5] H. Ahmadian, T. F. Abidin, H. Riza, dan K. Muchtar, "Transformer-Based Indonesian Language Model for Emotion Classification and Sentiment Analysis," dalam *2023 International Conference on Information Technology and Computing (ICITCOM)*, Yogyakarta, Indonesia: IEEE, Des 2023, hlm. 209–214. doi: 10.1109/ICITCOM60176.2023.10442970.

[6] F. Koto, A. Rahimi, J. H. Lau, dan T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020, *arXiv*. doi: 10.48550/ARXIV.2011.00677.

[7] A. Kusumaningrum, S. A. Faraby, dan A. Adiwijaya, "Klasifikasi Informasi, Anjuran dan Larangan pada Hadits Shahih Bukhari menggunakan Metode Support Vector Machine," *E-Proceeding Eng.*, vol. 4, hlm. 5014–5023, Des 2017.

[8] A. Wiraguna dan S. A. Faraby, "Klasifikasi Topik Multi Label pada Hadis Bukhari dalam Terjemahan Bahasa Indonesia Menggunakan Random Forest," *E-Proceeding Eng.*, vol. 6, hlm. 2144–2153, Apr 2019.

[9] M. Y. A. Bakar dan A. Adiwijaya, "Klasifikasi Teks Hadis Bukhari Terjemahan Indonesia Menggunakan Recurrent Convolutional Neural Network (CRNN)," *J. Teknol. Inf. Dan Ilmu Komput.*, vol. 8, no. 5, hlm. 907, Okt 2021, doi: 10.25126/jtiik.2021853750.

[10] A. Maiti, A. Abarda, dan M. Hanini, "The impact of feature extraction techniques on the performance of text data classification models," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 35, no. 2, hlm. 1041, Agu 2024, doi: 10.11591/ijeecs.v35.i2.pp1041-1052.

[11] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, dan J. Gao, "Deep Learning--based Text Classification: A Comprehensive Review," *ACM Comput. Surv.*, vol. 54, no. 3, hlm. 1–40, Apr 2022, doi: 10.1145/3439726.

[12] P. Aggarwal dan R. Mahajan, "Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification," *J. Inf. Syst. Inform.*, vol. 6, no. 2, hlm. 607–623, Jun 2024, doi: 10.51519/journalisi.v6i2.692.

[13] N. K. Nissa dan E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *Int. J. Electr. Comput. Eng. IJECE*, vol. 13, no. 5, hlm. 5641, Okt 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[14] F. Koto, J. H. Lau, dan T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," 10 September 2021, *arXiv*: arXiv:2109.04607. doi: 10.48550/arXiv.2109.04607.

[15] C. Cortes dan V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, hlm. 273–297, Sep 1995, doi: 10.1007/BF00994018.

[16] D. F. Surianto, M. Fajar B, M. R. Mulia, dan I. Indanasufya, "Comparative Analysis of the Performance of Hadith Text Classification Methods: A Case Study with ANN and SVM," *J. Embed. Syst. Secur. Intell. Syst.*, hlm. 89–98, Mar 2024, doi: 10.59562/jessi.v5i1.2942.