Optimasi Prediksi Vonis Pidana Melalui Strategi Ekstraksi Fitur Hibrida

Rizky Fadhilah[1], Vincentius Jacob Gunawan[2], Josh Edward Sutanto[3]

LG01

***Abstract*— Ledakan volume dokumen putusan pengadilan digital dalam sistem hukum perdata di Indonesia menimbulkan tantangan besar bagi praktisi hukum dalam menganalisis tren vonis karena jumlah data yang masif dan inkonsistensi struktural. Penelitian ini mengusulkan model pembelajaran mesin berbasis pemrosesan Bahasa alami untuk memprediksi durasi vonis pidana (dalam bulan) dari teks putusan pengadilan yang tidak terstruktur. Menggunakan dataset dunia nyata dari kompetisi Dataquest 4.0 yang terdiri dari 23.238 dokumen hukum, kami mengembangkan strategi ekstraksi fitur hibrida yang inovatif. Pendekatan ini menggabungkan analisis N-gram pada level kata dan karakter dengan teknik TF-IDF Vectorizer dan CountVectorizer untuk mengatasi tingkat *noise* yang tinggi, kesalahan pengetikan, dan inkonsistensi data. Metodologi kami meliputi pra-pemrosesan yang cermat, ekstraksi fitur hibrida, dan pelatihan model menggunakan algoritma seperti XGBoost, CatBoost, LightGBM, TabNet, dan LSTM. Model XGBoost mencapai performa terbaik dengan RMSE 13,53, mengungguli model canggih seperti IndoBERT yang kesulitan menangani *noise* dan konteks spesifik hukum. Penelitian ini menunjukkan bahwa pendekatan hibrida yang disesuaikan tidak hanya meningkatkan akurasi prediksi, tetapi juga menawarkan efisiensi komputasi dan skalabilitas, menjadikannya alat praktis untuk mendukung keputusan peradilan yang konsisten dan berbasis data di sistem hukum Indonesia.**

***Keywords*— Pemrosesan Bahasa Alami, ekstraksi fitur hibrida, prediksi vonis pidana, analisis N-gram, TF-IDF Vectorizer, CountVectorizer, XGBoost, analisis teks hukum, keadilan berbasis data.**

I. PENDAHULUAN

Di era digital, ledakan informasi tidak hanya terjadi di sektor industri, tetapi juga merambah ke dalam sistem hukum. Sistem hukum di Indonesia yang menganut *civil law* yang menempatkan undang-undang tertulis sebagai acuan utama, dengan putusan hakim yang memegang peranan krusial dalam menjaga konsistensi dan kepastian hukum. Tantangan muncul ketika volume dokumen putusan pengadilan digital bertambah banyak. Dalam tiga bulan terakhir, tercatat sekitar 329 ribu putusan dan jumlah ini terus bertambah [1]. Besarnya volume data tersebut membuat analisis manual oleh praktisi hukum untuk memahami tren vonis pada kasus serupa menjadi pekerjaan yang kurang efisien karena memakan waktu dan rentan terhadap inkonsistensi.

Oleh karena itu, untuk menjawab tantangan tersebut dan memanfaatkan kemajuan teknologi dalam analisis data, kami mengembangkan sebuah model *machine learning* berbasis *Natural Language Processing* (NLP). Model ini dirancang untuk menganalisis dan memprediksi durasi vonis pidana (dalam bulan) secara otomatis, hanya berdasarkan narasi teks dari dokumen putusan tersebut. Tujuan utama dari model ini bukanlah untuk menggantikan peran dan kebijaksanaan hakim, melainkan untuk berfungsi sebagai alat bantu yang menyediakan referensi *data-driven*. Dengan adanya referensi ini, diharapkan dapat mendukung terciptanya keadilan yang lebih konsisten di seluruh Indonesia, sejalan dengan prinsip kepastian hukum.

Pada upaya pengembangan model ini, kami menggunakan dataset *real-world* dari kompetisi **AIRNOLOGY FTMM UNAIR DATAQUEST 4.0** cabang lomba **Objective Quest**, yang diselenggarakan oleh Himpunan Mahasiswa Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga. Dataset tersebut mereplikasi tantangan sesungguhnya dalam analisis teks hukum, terdiri dari ribuan dokumen putusan pengadilan dalam format teks beserta label spesifik mengenai durasi hukuman yang dijatuhkan kepada terdakwa.

Tujuan utama dari penelitian ini adalah untuk merancang dan mengimplementasikan sebuah strategi ekstraksi fitur hibrida yang inovatif guna meningkatkan akurasi prediksi durasi vonis secara signifikan. Melalui pendekatan multi-perspektif yang menggabungkan analisis N-gram pada level kata dan karakter serta teknik vektorisasi TF-IDF Vectorizer dan CountVectorizer, kami berupaya tidak hanya menghasilkan model dengan akurasi tinggi. Lebih dari itu, penelitian ini bertujuan untuk menunjukkan bagaimana metode NLP yang canggih dapat mentransformasi data teks hukum yang kualitatif dan tidak terstruktur menjadi insight kuantitatif yang berharga, yang pada akhirnya dapat berkontribusi pada efisiensi dan transparansi dalam sistem peradilan.

II. DATA DAN EXPLORATORY DATA ANALYSIS

A. *Pengumpulan Data*

Pada perlombaan ini dataset yang digunakan berupa 16.572 dokumen putusan pengadilan untuk data latihan model dan 6.666 dokumen untuk data testing. Setiap dokumen disimpan dalam format .txt merepresentasikan satu kasus hukum dan disertai dengan label berupa durasi vonis pidana yang ditulis dalam satuan bulan. Format data dan isi narasi hukum yang kompleks dan tidak teratur menjadikan dataset ini cukup sulit untuk digunakan.

B. *Exploratory Data Analysis (EDA)*

Tahap *Exploratory Data Analysis* atau Analisis Data Eksploratif dilakukan untuk memahami karakteristik dan kualitas dari dataset yang mencakup 23.238 dokumen putusan

hukum. Hasil analisis mengungkapkan bahwa dataset tersebut memiliki tingkat *noise* dan inkonsistensi yang sangat tinggi, sehingga menimbulkan tantangan dalam proses pemodelan.

B.1. Profil Kuantitatif Dataset

Analisis kuantitatif menunjukan bahwa dataset tidak hanya memiliki volume yang besar dengan rata-rata teks per dokumen mencapai lebih dari 78.000 karakter, tetapi juga sangat kotor. Setiap dokumen rata-rata mengandung ribuan karakter non-esensial, termasuk 2.137 karakter spesial, 4.522 huruf kapital, dan 1.956 tanda baca. Selain itu, keseluruhan dataset teridentifikasi memiliki masalah spasi berlebih (*whitespace issues*), yang mengindikasikan adanya inkonsistensi struktural.

B.2. Analisis Kualitatif Inkonsistensi Data

Analisis kualitatif mengungkapkan berbagai isu kualitas data yang lebih dalam. Teks dalam dokumen putusan ini bukanlah narasi yang bersih, sebaliknya, ia dipenuhi dengan berbagai kesalahan pengetikan.

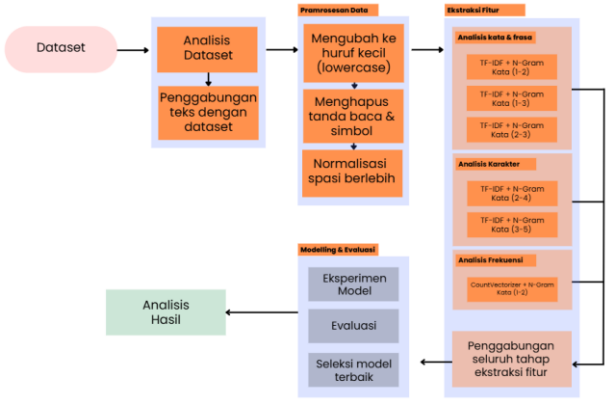
B.2.1. Inkonsistensi Kapitalisasi dan Spasi

Penulisan frasa yang sama bisa sangat bervariasi mulai dari “hkama”, “ahkamah Agung Repu”, “mah Agung Republik Indonesia”, dan beberapa jenis varian inkonsistensi lainnya, ditambah dengan spasi yang tidak beraturan, membuat proses ekstraksi fitur standar menjadi tidak efektif.

B.2.2. Kesalahan pengetikan dan kata yang menyatu

Kami menemukan banyak sekali kesalahan pengetikan, singkatan yang tidak umum, dan kata-kata yang memiliki makna penting yang menyatu tanpa spasi. Contohnya termasuk “hikama” (kesalahan pengetikan dari 'mahkamah'), kepaniteraanmahka, dan umurtanggal. Terdapat pula teks berulang seperti “...tahan tahan than...” yang merupakan *noise* dan tidak membawa makna hukum yang substantif. Diperlukan sebuah tahap *preprocessing* yang baik, namun juga harus berhati-hati. Tantangannya adalah bagaimana menemukan keseimbangan antara membersihkan *noise* (seperti kesalahan pengetikan dan spasi berlebih), dengan tidak menghilangkan informasi penting. *Preprocessing* yang terlalu agresif, misalnya menghapus *stopwords* atau melakukan *stemming* yang ekstrem, berisiko mengubah makna penting dari narasi hukum [2]. Oleh karena itu, tahap pembersihan data yang cermat menjadi pilar utama dalam penelitian kami.

III. METODOLOGI



Gambar 1. Alur Eksperimen

Gambar 1 menunjukkan metodologi yang diterapkan dalam tim kami. Metodologi tersebut dirancang secara sistematis untuk mengubah data teks pengadilan di Indonesia yang tidak terstruktur. Pendekatan yang kami lakukan berpusat pada strategi kami dalam melakukan *feature extraction* yang menggabungkan beberapa teknik N-gram untuk menangkap pola linguistik yang relevan dari data teks dalam berbagai tingkatan [3].

A. Pra-pemrosesan Data

Sebelum melakukan *feature extraction*, setiap dokumen melewati beberapa tahap *pre-processing* untuk membersihkan dan menstandarisasi teks:

A.1. Pengelompokan data

Pada tahap ini, seluruh dokumen teks dipetakan sesuai dengan pembagiannya pada file train dan test dalam dataset, sehingga terbentuk dataframe terstruktur yang memudahkan proses analisis selanjutnya.

A.2. Pembersihan Teks

Pada tahap ini terdapat beberapa proses yang dilakukan untuk menjaga konsistensi dan menstadarisasi dataset:
- Mengubah ke huruf kecil: Semua teks pada dataset kami ubah menjadi huruf kecil untuk memastikan konsistensi.
- Menghapus simbol dan tanda baca: Seluruh karakter non-alfabet (seperti @, !, #) dihilangkan untuk memfokuskan *feature extraction* hanya pada konten tekstual.
- Normalisasi spasi: Spasi yang tidak beraturan pada dokumen kami hapus untuk menjaga kerapihan dataset.

A.3. Tokenisasi dan Stemming

Proses tokenisasi, atau pemecahan teks menjadi unit-unit kata, merupakan langkah fundamental yang terjadi secara implisit di dalam TF-IDF Vectorizer dan CountVectorizer. Vectorizer ini secara otomatis menangani tokenisasi sebagai bagian dari proses pembentukan fitur N-gram, sehingga tidak memerlukan langkah tokenisasi eksplisit yang terpisah. Namun, kami memutuskan untuk tidak menerapkan proses stemming maupun lemmatization pada tahap pra-pemrosesan ini. Keputusan ini didasarkan pada hipotesis bahwa dalam konteks teks hukum, variasi bentuk kata seringkali membawa makna yang krusial. Sebagai contoh, kata "putusan" dan "memutuskan" dapat memberikan hasil yang berbeda bagi model. Proses stemming yang agresif berisiko menyamakan kedua kata ini dan menghilangkan informasi yang berharga. Oleh karena itu, pendekatan kami adalah dengan langsung menerapkan ekstraksi fitur N-gram pada teks yang hanya telah dibersihkan (diubah ke huruf kecil dan dihapus dari simbol). Strategi ini memungkinkan model untuk mempelajari pola kata dan frasa secara langsung dari data asli tanpa mereduksi variasi leksikalnya secara drastis.

B. Ekstraksi Fitur Hibrida

Ini adalah inti dari metodologi kami. Alih-alih menggunakan satu pendekatan, kami menggabungkan beberapa teknik vektorisasi untuk menangkap sinyal prediktif dari berbagai perspektif:

B.1 Analisis Kata & Frasa (Word N-grams)

TF-IDF Vectorizer (1,2)-grams: Menganalisis kata tunggal dan frasa dua kata (bigram) untuk menangkap konteks langsung, seperti "saksi ahli".
TF-IDF Vectorizer (1,3)-grams: Memperluas analisis hingga frasa tiga kata (trigram) untuk menangkap konteks yang lebih panjang dan spesifik.
TF-IDF Vectorizer (2,3)-grams: Fokus pada frasa dua dan tiga kata untuk mengisolasi hubungan antar kata yang lebih kompleks.

B.2 Analisis Karakter (Character N-grams)

TF-IDF Vectorizer (2,4)-grams Karakter: Menganalisis pola karakter untuk memberikan ketahanan terhadap kesalahan ketik, singkatan, dan variasi ejaan yang umum ditemukan dalam dokumen hukum.
TF-IDF Vectorizer (3,5)-grams Karakter: Menggunakan rentang yang sedikit lebih panjang untuk menangkap pola morfologis yang lebih detail.

B.3 Analisis Frekuensi (Frequency Analysis)

CountVectorizer (1,2)-grams: Selain bobot TF-IDF Vectorizer, kami juga menghitung frekuensi kemunculan kata dan frasa dua kata. Ini memberikan sinyal komplementer yang berfokus pada seberapa sering suatu istilah muncul, bukan hanya seberapa unik istilah tersebut.

Setelah keenam proses ekstraksi ini selesai, semua fitur yang dihasilkan digabungkan menjadi satu "Super Vektor" tunggal. Vektor gabungan ini akan berisikan ribuan fitur atau kolom yang merepresentasikan setiap dokumen dari berbagai sudut pandang, menciptakan sebuah set fitur yang sangat kaya dan siap untuk dimodelkan.

C. Pelatihan dan Pengujian Model

Untuk memprediksi durasi hukuman dari "Super Vektor" yang telah dibuat, kami melakukan perbandingan beberapa model *machine learning* dan *deep learning*. Metrik evaluasi utama yang digunakan adalah *Root Mean Squared Error* (RMSE), yang memberikan penalti lebih besar pada prediksi dengan selisih error yang tinggi.

Model-model yang diuji meliputi:

- LSTM (Long Short-Term Memory): Model *deep learning* berbasis jaringan saraf berulang (RNN) yang dirancang untuk memproses data sekuensial, seperti teks. LSTM mengatasi masalah *vanishing gradient* dengan mekanisme *gate* (input, forget, output) untuk menyimpan dan memperbarui informasi jangka panjang, cocok untuk menangkap pola dalam narasi hukum yang panjang [4].
- CatBoost: Model *gradient boosting* yang dioptimalkan untuk data tabular, terutama dengan fitur kategorikal. CatBoost menawarkan performa tinggi dengan pengaturan parameter otomatis [5].
- XGBoost: Model gradient boosting yang sangat efisien dan efektif untuk data tabular yang bersifat sparse seperti fitur hasil TF-IDF Vectorizer. Menggunakan pendekatan *level-wise tree growth* dan regularisasi untuk mencegah *overfitting*, XGBoost unggul dalam akurasi dan kecepatan pada dataset besar [6].
- LightGBM: Model *gradient boosting* yang dirancang untuk kecepatan dan efisiensi *training* yang sangat tinggi, terutama pada dataset berskala besar. Menggunakan pendekatan *leaf-wise tree growth* dan *histogram-based learning*, LightGBM mengurangi penggunaan memori dan mempercepat pelatihan, ideal untuk data berdimensi tinggi [7], [8].
- TabNet: Model *deep learning* yang menggunakan *attention mechanism* untuk memilih fitur pada data tabular dengan arsitekturnya yang menggunakan *sequential decision steps* [9].

Setiap model dilatih pada 80% data latih dan dievaluasi pada 20% sisa data sebagai set validasi untuk melihat nilai RMSE pada data testing. Berdasarkan hasil perbandingan, XGBoost menunjukkan performa terbaik dalam nilai RMSE paling rendah dari model lain.

IV. HASIL DAN ANALISIS

Pada bagian ini, kami menyajikan hasil eksperimen dari berbagai model yang diuji, menganalisis faktor-faktor yang mendorong keunggulan arsitektur yang diusulkan yaitu kombinasi berbagai rentang N-gram (kata dan karakter) melalui TF-IDF Vectorizer dan CountVectorizer, serta membedah performa model *state-of-the-art* seperti IndoBERT jika digunakan sebagai *feature extractor* pada kasus ini.

A. Perbandingan Model dengan Pendekatan Hibrida

Setelah menggabungkan seluruh fitur hasil *extraction* dari TF-IDF Vectorizer. Kami menggunakan beberapa model *machine learning* dan *deep learning* untuk digunakan sebagai prediktor lama hukuman vonis tersangka. Tujuannya adalah untuk mengidentifikasi arsitektur model yang mampu menghasilkan prediksi yang akurat menggunakan hasil *feature extraction* kami.

Tabel 1. Hasil *Modelling* dengan Pendekatan Hibrida

Model	**Skor RMSE**
TabNet	13.74
CatBoost	13.57
LSTM	14.59
LightGBM	13.61
XGBoost	**13.53**

Berdasarkan Tabel 1, hasil evaluasi menunjukkan bahwa model **XGBoost** memberikan performa yang terunggul dibandingkan pendekatan lainnya. Model XGBoost muncul sebagai model dengan skor RMSE terendah **(13.53)**, membuktikan performanya dalam menangani data tabular berdimensi tinggi dan bersifat *sparse*, serta kemampuannya dalam memodelkan fitur yang kompleks.

Persaingan terketat datang dari sesama model *gradient boosting*. **CatBoost (13.57)** menunjukkan performa yang sangat baik. Akan tetapi, performa terbaik CatBoost adalah saat menangani fitur kategorikal, sedangkan pada dataset ini yang didominasi oleh fitur numerik *sparse*, keunggulan tidak sepenuhnya termanfaatkan.

Sementara itu, **LightGBM (13.61)**, yang dikenal karena performa pelatihan yang cepat, juga mencatatkan hasil yang sangat solid. LightGBM mencapai efisiensinya melalui strategi pertumbuhan pohon secara *leaf-wise* (vertikal), yang sangat cepat [10]. Namun, untuk dataset dengan ukuran seperti ini, pendekatan *level-wise* (horizontal) yang lebih menyeluruh dari XGBoost terbukti mampu menangkap interaksi fitur dengan sedikit lebih baik, memberikannya keunggulan tipis dalam hal akurasi.

Sebaliknya, perbandingan dengan model *deep learning* seperti TabNet (13.74) dan LSTM (14.59) menunjukkan performa yang lebih rendah karena kompleksitas model dan ketidakcocokan arsitektur LSTM dengan data tabular non-sekuensial. Hal ini semakin menunjukkan bahwa untuk tipe dataset dan target prediksi, arsitektur *gradient boosting* dan spesifik, model **XGBoost** yang seimbang antara akurasi dan efisiensi merupakan pendekatan terbaik.

B. Menganalisis Performa IndoBERT sebagai Feature Extractor

Kami menemukan dalam penelitian ini bahwa performa model IndoBERT sebagai *feature extractor* yang menggantikan TF-IDF Vectorizer dan digabungkan dengan model-model yang sama memiliki rata-rata RMSE lebih rendah. Seperti yang di tunjukkan pada tabel di bawah ini.

Tabel 2. Hasil *Modelling* IndoBERT sebagai Feature Extractor

Model	**RMSE**
TabNet	22.67
CatBoost	22.35
LSTM	22.52
LightGBM	22.22
XGBoost	**22.15**

Ketika dibandingkan dengan Tabel 1, performa model pada Tabel 2 menunjukkan hasil yang lebih buruk. Fenomena tersebut merupakan contoh nyata dari "state-of-the-art fallacy" yang merupakan sebuah keyakinan keliru bahwa model yang dianggap paling baik oleh komunitas riset secara otomatis akan memberikan hasil terbaik di semua kasus [11]. Terdapat beberapa faktor yang mungkin menjadi alasan mengapa IndoBERT pada kasus ini kurang baik jika dibandingkan dengan arsitektur usulan:

- **Ketidaksesuaian data latih**: IndoBERT dilatih pada korpus data Bahasa Indonesia yang umum, bersih, dan terstruktur. Arsitektur IndoBERT dioptimalkan untuk memahami semantik dan konteks dari bahasa Indonesia yang umum digunakan. Ketika dihadapkan pada teks berbasis hukum yang penuh dengan kata spesifik, singkatan non-standar, kesalahan ketik masif, dan struktur kalimat yang repetitif, kemampuan pemahaman semantiknya justru menjadi kelemahan. Model ini "bingung" karena data input tidak sesuai dengan pola linguistik yang telah dipelajarinya.
- **Banyaknya *noise* pada data latih**: Model transformer seperti BERT sangat sensitif terhadap *noise*. Kesalahan pengetikan atau spasi yang tidak biasa dapat mengubah sebuah kata menjadi token yang sama sekali berbeda (*out-of-vocabulary*), sehingga informasi penting hilang. Sebaliknya, pendekatan N-gram karakter kami justru unggul dalam kondisi ini.

Temuan ini menegaskan bahwa pemilihan *feature extractor* tidak seharusnya hanya didasarkan pada tren atau popularitas, melainkan harus didasarkan pada pemahaman mendalam terhadap karakteristik unik dari dataset yang dihadapi.

C. Menganalisis Performa IndoBERT sebagai Feature Extractor

Di luar aspek akurasi prediksi model, arsitektur yang kami usulkan menawarkan keunggulan signifikan dari segi efisiensi dan skalabilitas. Kebutuhan sumber daya komputasi dengan proses ekstraksi fitur menggunakan TF-IDF Vectorizer dan CountVectorizer, diikuti dengan pelatihan model XGBoost, secara komputasi jauh lebih ringan dibandingkan dengan fine-tuning atau bahkan sekadar menjalankan inferensi pada model transformer besar seperti IndoBERT. Pendekatan kami tidak memerlukan perangkat keras khusus seperti GPU dan dapat dijalankan secara efisien pada CPU standar.

Selain itu, dalam skenario penerapan di mana sistem perlu memproses ribuan dokumen baru, kecepatan inferensi menjadi sangat penting. Studi oleh Zhang menemukan bahwa pendekatan hibrida yang menggabungkan metode klasik dengan panduan dari model modern mampu mencapai 95.2% akurasi BERT dengan hanya 1/50 biaya inferensi [12]. Hal ini menunjukkan bahwa arsitektur seperti yang kami usulkan tidak hanya lebih akurat untuk kasus spesifik ini, tetapi juga secara eksponensial lebih hemat biaya dan lebih cepat untuk dioperasikan dalam skala besar.

Keunggulan ini menjadikan pendekatan kami merupakan pendekatan yang layak dan praktis untuk dikembangkan lebih lanjut dalam sistem hukum di Indonesia, di mana ketersediaan sumber daya komputasi dapat menjadi kendala.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Dapat disimpulkan, pendekatan kami berhasil mengembangkan arsitektur model *machine learning* yang mampu memprediksi durasi vonis pidana dari teks putusan pengadilan dengan tingkat akurasi yang tinggi, mencapai skor **RMSE 13.53**. Kunci keberhasilan model ini terletak pada strategi ekstraksi fitur hibrida yang inovatif, yang menggabungkan analisis N-gram pada level kata dan karakter untuk secara efektif mengatasi tantangan data teks hukum yang masif, tidak terstruktur, dan penuh dengan inkonsistensi.

Temuan utama dari penelitian ini dapat dirangkum dalam tiga poin berikut:

- **Pendekatan hibrida yang superior:** Kombinasi berbagai rentang N-gram (kata dan karakter) melalui TF-IDF Vectorizer dan CountVectorizer, yang dipadukan dengan model prediktor XGBoost, terbukti menjadi arsitektur yang paling akurat dan tangguh untuk dataset spesifik ini.
- ***State-of-the-art fallacy*:** Pada kasus data di sini menunjukkan bahwa model SOTA seperti IndoBERT sebagai *feature extractor* tidak selalu menjadi solusi terbaik. Dari kelima model yang diuji, XGBoost konsisten menunjukkan performa terbaik. Ketika diuji dengan metode *feature extraction* yang berbeda, XGBoost dengan pendekatan klasik menghasilkan nilai RMSE sebesar 13.53, sedangkan XGBoost dengan *feature extractor* IndoBERT menghasilkan nilai RMSE sebesar 22.15. Temuan ini menegaskan pentingnya keselarasan antara metode dan karakteristik data, bukan sekadar mengadopsi teknologi terbaru.
- **Pendekatan yang efisien:** Selain unggul dalam akurasi, model yang kami usulkan juga jauh lebih efisien dari segi komputasi dan lebih *scalable* dibandingkan model berbasis transformer. Aspek efisiensi menjadikan model ini kandidat yang kuat untuk aplikasi dunia nyata dalam sistem peradilan, di mana kecepatan dan biaya operasional adalah pertimbangan utama.

B. Saran

Berdasarkan temuan yang diperoleh, terdapat beberapa saran untuk penelitian di masa depan. Pertama, pengembangan *domain-specific language model* untuk melihat apakah keterbatasan IndoBERT yang dilatih pada data dengan spektrum umum dapat diatasi. Penelitian selanjutnya dapat berfokus pada proses pre-training atau fine-tuning sebuah model transformer (seperti BERT) secara khusus pada korpus besar yang berisi dokumen hukum Indonesia. Ini berpotensi menciptakan model yang memiliki pemahaman semantik mendalam tentang istilah-istilah dalam hukum sekaligus lebih mampu menangkap maksud dari pengetikan teks yang variatif.

Kedua, penggabungan fitur hibrida dengan konteks dapat menjadi sebuah pendekatan yang menjanjikan. Dengan menggabungkan fitur N-gram yang telah terbukti efektif dengan *embedding* dari model transformer yang telah dilatih pada data hukum dapat menggabungkan ketahanan fitur klasik dan pemahaman kontekstual dari model modern yang berpotensi menghasilkan model yang lebih akurat dan efisien.

DAFTAR PUSTAKA

[1] Putusan Mahkamah Agung, “Mahkamah Agung Republik Indonesia.” Accessed: Sep. 19, 2025. [Online]. Available: https://putusan3.mahkamahagung.go.id

[2] A. K. Uysal and S. Gunal, “The impact of preprocessing on text classification,” *Inf Process Manag*, vol. 50, no. 1, pp. 104–112, 2014, doi: 10.1016/j.ipm.2013.08.006.

[3] T. Hasan and A. Matin, “Extract Sentiment from Customer Reviews: A Better Approach of TF-IDF and BOW-Based Text Classification Using N-Gram Technique,” 2021, pp. 231–244. doi: 10.1007/978-981-16-0586-4_19.

[4] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[5] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” Jan. 2019, [Online]. Available: http://arxiv.org/abs/1706.09516

[6] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[7] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” 2017. [Online]. Available: https://github.com/Microsoft/LightGBM.

[8] P. Florek and A. Zagdański, “Benchmarking state-of-the-art gradient boosting algorithms for classification,” May 2023, [Online]. Available: http://arxiv.org/abs/2305.17094

[9] S. O. Arik and T. Pfister, “TabNet: Attentive Interpretable Tabular Learning,” Dec. 2020, [Online]. Available: http://arxiv.org/abs/1908.07442

[10] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” 2017. [Online]. Available: https://github.com/Microsoft/LightGBM.

[11] A. Alvero, R. Dong, K. Kanopka, and D. Lang, “Algorithmic Tradeoffs, Applied NLP, and the State-of-the-Art Fallacy,” Sep. 2025, [Online]. Available: http://arxiv.org/abs/2509.08199

[12] L. Zhang, “Features extraction based on Naive Bayes algorithm and TF-IDF for news classification,” *PLoS One*, vol. 20, no. 7 July, Jul. 2025, doi: 10.1371/journal.pone.0327347.

.