EVALUASI EFEKTIVITAS PRA-PEMROSESAN TERHADAP AKURASI MODEL DISTILBERT PADA KLASIFIKASI TOPIK PENELITIAN HUKUM

(Studi Kasus: Fakultas Hukum Universitas Tanjungpura Kota Pontianak)

KHALID IBNU NATSIR
H1101221046

SKRIPSI

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TANJUNGPURA
PONTIANAK
2026

Evaluasi Efektivitas Pra-Pemrosesan Terhadap Akurasi Model Distilbert Pada Klasifikasi Topik Penelitian Hukum (Studi Kasus: Fakultas Hukum Universitas Tanjungpura Kota Pontianak)

KHALID IBNU NATSIR
H1101221046

Skripsi
Sebagai salah satu syarat untuk memperoleh gelar
Sarjana Komputer pada Program Studi Sistem Informasi



PROGRAM STUDI SISTEM INFORMASI
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TANJUNGPURA
PONTIANAK
2026

LEMBAR PENGESAHAN

Judul Tugas Akhir	: Evaluasi Efektivitas Pra-Pemrosesan Terhadap Akurasi Model Distilbert Pada Klasifikasi Topik Penelitian Hukum. (Studi Kasus: Fakultas Hukum Universitas Tanjungpura Kota Pontianak)
Nama Mahasiswa	: Khalid Ibnu Natsir
NIM	: H1101221046
Jurusan/Program Studi	: Sistem Informasi / Sistem Informasi
Tanggal Lulus	: 27 Juli 2026
SK Pembimbing	: No. 4525/DST/UN22.8/DT.00.09/2026
SK Penguji	: No. 4767/DST/UN22.8/DT.00.09/2026

Dosen Pembimbing

Pembimbing I

Ibnur Rusi, S.Kom., M.M.
NIP 198907282019031008

Pembimbing II

Syahru Rahmayuda, S.Kom., M.Kom.
NIP 198306172025211055

Dosen Penguji

Ketua Penguji

Dian Prawira, S.T., M.Eng.
NIP 198411132015041001

Anggota Penguji

Nurul Mutiah, S.T., M.T.
NIP 198711182015042002

Pimpinan Sidang
(merangkap anggota penguji)

Ibnur Rusi, S.Kom., M.M.
NIP 198907282019031008

Sekretaris Sidang
(merangkap anggota penguji)

Syahru Rahmayuda, S.Kom., M.Kom.
NIP 198306172025211055

Mengesahkan
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Tanjungpura

Prof. Dr. Gusrizal, S.Si., M.Si.
NIP 197108022000031001

Scanned with TapScanner

LEMBAR PENGESAHAN

Judul Tugas Akhir	: Evaluasi Efektivitas Pra-Pemrosesan Terhadap Akurasi Model Distilbert Pada Klasifikasi Topik Penelitian Hukum. (Studi Kasus: Fakultas Hukum Universitas Tanjungpura Kota Pontianak)
Nama Mahasiswa	: Khalid Ibnu Natsir
NIM	: H1101221046
Jurusan/Program Studi	: Sistem Informasi / Sistem Informasi
Tanggal Lulus	:
SK Pembimbing	: No. 4525/DST/UN22.8/DT.00.09/2026
SK Penguji	: No.

Dosen Pembimbing

Pembimbing I	Pembimbing II
Ibnur Rusi, S.Kom., M.M.	Syahru Rahmayuda, S.Kom., M.Kom.
NIP 198907282019031008	NIP 198306172025211055

Dosen Penguji

Ketua Penguji	Anggota Penguji
Dian Prawira, S.T., M.Eng.	Nurul Mutiah, S.T. , M.T.
NIP 198411132015041001	NIP 198711182015042002
Pimpinan Sidang (merangkap anggota penguji)	Sekretaris Sidang (merangkap anggota penguji)
Ibnur Rusi, S.Kom., M.M.	Syahru Rahmayuda, S.Kom., M.Kom.
NIP 198907282019031008	NIP 198306172025211055

Mengesahkan
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Tanjungpura

Prof. Gusrizal, S.Si., M.Si.
NIP 197108022000031001

PERNYATAAN INTEGRITAS AKADEMIK

Saya yang bertanda tangan di bawah ini:

Nama	: Khalid Ibnu Natsir
NIM	: H1101221046
Program Studi/ Jurusan	: Sistem Informasi/ Sistem Informasi
Fakultas	: Matematika dan ilmu pengetahuan alam

dengan ini menyatakan bahwa dokumen ilmiah Tugas Akhir yang disajikan ini tidak mengandung unsur pelanggaran integritas akademik sesuai Peraturan Menteri Pendidikan, Kebudayaan, Riset, Dan Teknologi Republik Indonesia Nomor 39 Tahun 2021. Apabila di kemudian hari dokumen ilmiah Tugas Akhir ini mengandung unsur pelanggaran integritas akademik sesuai ketentuan perundangan tersebut, maka saya bersedia menerima sanksi akademik dan/atau sanksi hukum yang berlaku.

Demikian pernyataan ini untuk digunakan sebagaimana mestinya.

Pontianak, Juli 2026

Khalid Ibnu Natsir

H1101221046

Evaluasi Efektivitas Pra-Pemrosesan Terhadap Akurasi Model Distilbert Pada Klasifikasi Topik Penelitian Hukum (Studi Kasus: Fakultas Hukum Universitas Tanjungpura Kota Pontianak)

ABSTRAK

Pre-processing teks merupakan tahapan krusial dalam *pipeline* klasifikasi teks, tahapan tersebut juga relevan untuk model klasifikasi teks arsitektur transformer seperti DistilBERT, namun pengaruhnya pada dataset domain hukum berbahasa Indonesia yang kaya akan singkatan dan istilah teknis masih memerlukan kajian lebih lanjut. Penelitian ini bertujuan mengevaluasi efektivitas *pre-processing* dan augmentasi data *Back Translation* (BT) terhadap akurasi model DistilBERT dalam mengklasifikasikan abstrak skripsi hukum ke dalam lima topik: Hukum Internasional, Tata Negara, Ekonomi, Perdata, dan Pidana. Dataset 586 abstrak diproses melalui dua skenario: Skenario A (pembersihan dasar dan normalisasi singkatan) dan Skenario B (pembersihan dasar, *case folding*, tokenisasi, *stopword removal*, *stemming*). Masing-masing dilatih dengan dan tanpa *Back Translation* menggunakan MarianMT, menghasilkan empat model. Hasil pengujian pada 119 data uji menunjukkan Skenario A+BT mencapai akurasi tertinggi 79,66%, diikuti Skenario B+BT 78,81%, Skenario A 72,03%, dan Skenario B 69,49%. *Back Translation* meningkatkan akurasi +7,63% pada Skenario A dan +9,32% pada Skenario B. *Pre-processing* penuh (Skenario B) justru menurunkan akurasi 2,54% dibanding Skenario A, mengindikasikan *stemming* dan *stopword removal* menghilangkan informasi penting pada teks hukum. Penelitian menyimpulkan Skenario A+BT sebagai konfigurasi optimal, menunjukkan augmentasi *Back Translation* efektif meningkatkan performa, sementara *pre-processing* ekstensif berpotensi menurunkan akurasi model.

Kata Kunci: Klasifikasi Teks, DistilBERT, Pra-Pemrosesan, Back Translation, Teks Hukum Bahasa Indonesia

Evaluation of the Effectiveness of Pre-Processing on the Accuracy of the DistilBERT Model in Legal Research Topic Classification (Case Study: Faculty of Law, Tanjungpura University, Pontianak)

ABSTRACT

Text pre-processing is a crucial stage in the text classification pipeline, including for transformer-based models such as DistilBERT. However, its impact on Indonesian legal domain datasets—which are rich in abbreviations and technical terms—still requires further investigation. This study aims to evaluate the effectiveness of pre-processing and Back Translation (BT) data augmentation on the accuracy of the DistilBERT model in classifying legal thesis abstracts into five topics: International Law, Constitutional Law, Economics, Civil Law, and Criminal Law. A dataset of 586 abstracts was processed using two scenarios: Scenario A (basic cleaning and abbreviation normalization) and Scenario B (basic cleaning, case folding, tokenization, stopword removal, stemming). Each scenario was trained with and without Back Translation using MarianMT, resulting in four models. Evaluation on 119 test samples showed that Scenario A+BT achieved the highest accuracy of 79.66%, followed by Scenario B+BT at 78.81%, Scenario A at 72.03%, and Scenario B at 69.49%. Back Translation improved accuracy by +7.63% in Scenario A and +9.32% in Scenario B. Full pre-processing (Scenario B) actually decreased accuracy by 2.54% compared to Scenario A, indicating that stemming and stopword removal removed important information from legal texts. This study concludes that Scenario A+BT is the optimal configuration, demonstrating that Back Translation augmentation effectively improves performance, while extensive pre-processing has the potential to reduce model accuracy.

***Keywords**: Text Classification, DistilBERT, Pre-processing, Back Translation, Indonesian Legal Text*

PRAKATA

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Evaluasi Efektivitas Pra-Pemrosesan Terhadap Akurasi Model DistilBERT pada Klasifikasi Topik Penelitian Hukum (Studi Kasus: Fakultas Hukum Universitas Tanjungpura Kota Pontianak)" dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Sistem Informasi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Tanjungpura.

Dalam proses penyusunan skripsi ini, penulis menyadari bahwa banyak pihak yang telah memberikan bantuan, dukungan, serta bimbingan sehingga penelitian ini dapat terselesaikan. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih kepada:

1. Tuhan Yang Maha Esa yang telah memberikan kesehatan, kekuatan, dan kelancaran selama proses penyusunan skripsi ini.
2. Orang tua dan keluarga yang selalu memberikan doa, dukungan, serta motivasi selama menjalani masa perkuliahan hingga penyusunan skripsi ini.
3. Bapak Ibnur Rusi, S.Kom., MM. selaku Dosen Pembimbing Akademik, yang telah memberikan ilmu, perhatian, serta bimbingan dan saran-saran sejak awal perkuliahan hingga saat ini.
4. Bapak Syahru Rahmayuda, S.Kom., M.Kom. selaku Dosen Pembimbing kedua, yang telah memberikan saran serta masukkan selama proses penyusunan skripsi ini
5. Bapak Dian Prawira, ST., M.Eng. selaku penguji pertama, saya ucapkan terima kasih atas ilmu, saran-saran, serta bantuan dalam menggali dan mengembangkan kemampuan saya.
6. Ibu Nurul Mutiah, ST., MT. selaku penguji kedua, saya ucapkan terima kasih atas ilmu, saran-saran, serta bantuan dalam menggali dan mengembangkan kemampuan saya.

7. Seluruh dosen dan staf Program Studi Sistem Informasi Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Tanjungpura yang telah memberikan ilmu dan bantuan selama masa perkuliahan.
8. Pihak Fakultas Hukum Universitas Tanjungpura yang telah memberikan izin serta data yang diperlukan dalam pelaksanaan penelitian ini.
9. Teman-teman dan semua pihak yang tidak dapat disebutkan satu per satu yang telah memberikan dukungan serta bantuan dalam penyusunan skripsi ini.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan dan kekurangan. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun demi perbaikan di masa yang akan datang. Semoga skripsi ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya di bidang data mining dan klasifikasi topik dari text.

Pontianak, Juli 2026

Khalid Ibnu Natsir
H1101221046

DAFTAR ISI

LEMBAR PENGESAHAN ..... **ii**
PERNYATAAN INTEGRITAS AKADEMIK iii
ABSTRAK iv
ABSTRACT v
PRAKATA vi
DAFTAR ISI viii
DAFTAR GAMBAR xi
DAFTAR TABEL xii
DAFTAR SINTAKS KODE xiii
DAFTAR PERSAMAAN xiv
BAB I PENDAHULUAN 1
1.1. Latar Belakang 1
1.2. Rumusan Masalah 3
1.3. Tujuan Penelitian 3
1.4. Batasan Masalah 3
1.5. Manfaat Penelitian 4
1.6. Sistematika Penulisan 4
BAB II TINJAUAN PUSTAKA 7
2.1. Landasan Teori 7
2.1.1 Fakultas Hukum Universitas Tanjungpura 7
2.1.2 *Deep Learning* 8
2.1.3 *Natural Language Processing* (*NLP*) 8
2.1.4 *Text mining* 9
2.1.5 Klasifikasi Teks 9
2.1.6 Klasifikasi Teks Hukum 10
2.1.7 *Pre-processing* Dalam Data Teks 10
2.1.8 Skenario Dalam *Pre-processing* 10
2.1.9 Arsitektur *Transformer* 11
2.1.10 Model BERT 13
2.1.11 Model DistilBERT 15
2.1.12 DistilBERT *Multilingual* (distilbert-base-multilingual-cased) 16
2.1.13 Tokenisasi pada DistilBERT 17
2.1.14 *Fine-Tuning* Model *Transformer* 17
2.1.15 *Optimizer* Dan *Learning Rate Scheduler* 18
2.1.16 Ketidakseimbangan Data (*Imbalanced Data*) 18
2.1.17 *Back Translation* Untuk Menangani Ketidakseimbangan Data 19
2.1.18 *Confusion Matrix* Dan Metrik Evaluasi Klasifikasi 19
2.1.19 Hugging Face *Transformers* 21
2.1.20 Python Dan *Library* Pendukung 22
2.2. Penelitian Terdahulu 24
BAB III METODOLOGI PENELITIAN 26
3.1. Metodologi Penelitian 26
3.1.1 *Environment* 27
3.1.2 *IS Research* 27

3.1.3 *Knowledge Base* 28
3.2. Metode Penelitian 29
BAB IV HASIL DAN PEMBAHASAN 32
4.1. *Dataset* Penelitian 32
4.1.1 Deskripsi *Dataset* 32
4.1.2 Distribusi Label Topik Hukum 33
4.1.3 Pembagian Data Latih Dan Data Uji 34
4.2. Perancangan Skenario 34
4.2.1 Skenario A – *Minimal Pre-Processing* 34
4.2.2 Skenario B – *Full Pre-Processing* 35
4.2.3 Konfigurasi Parameter *Training* Model 35
4.3. Tahapan *Pre-Processing* 36
4.3.1 *Install* dan *Import Library* 36
4.3.2 Konfigurasi *File* Dan Inisialisasi Data 37
4.3.3 Kamus Normalisasi Istilah Hukum, *Stopword* Akademik, ...37
4.3.4 Fungsi *Pre-Processing Text Cleaning* 39
4.3.5 Fungsi *Pre-Processing Case Folding* Dan Normalisasi Istilah 39
4.3.6 Fungsi *Pre-Processing Stopword Removal*, Dan *Stemming* 40
4.3.7 Fungsi *Pre-Processing* Tokenisasi DistilBERT 41
4.3.8 Fungsi *Load Dataset* Dan Distribusi Topik 41
4.3.9 Fungsi Pipeline Skenario A (*Minimal Pre-processing*) 42
4.3.10 Fungsi Pipeline Skenario B (*Full Pre-processing*) 43
4.4. Tahapan *Training* Model 44
4.4.1 *Import Library* 44
4.4.2 Konfigurasi Parameter Training Model 45
4.4.3 Fungsi *Label Mapping* Dan Fungsi *Parsing Token* 46
4.4.4 Fungsi Load Dan Persiapan *Data Training* 47
4.4.5 Fungsi Pembagian *Data training* 47
4.4.6 Fungsi Kelas Dataset Untuk DataLoader 48
4.4.7 Fungsi Training Model Tanpa Back Translation 49
4.4.8 Fungsi Back Translation Untuk Augmentasi Data 51
4.4.9 Fungsi Persiapan Data *Training* Dengan *Back Translation* 52
4.4.10 Fungsi Evaluasi Dan *Confusion Matrix* 55
4.4.11 Fungsi Plot Hasil Perbandingan 57
4.4.12 Fungsi Pembuatan Wordcloud Per Label 59
4.4.13 Fungsi *Argument Parser* Dan Persiapan DataLoader 60
4.4.14 Fungsi *Main Block*: Eksekusi Training, Penyimpanan Model 61
4.5. Tahapan *Testing* Interaktif 64
4.5.1 *Import Library*, Konfigurasi, Dan Komponen *Pre-processing* 64
4.5.2 Fungsi *Pre-processing* Untuk *Testing* 66
4.5.3 Fungsi load_models 67
4.5.4 Fungsi Predict 68
4.5.5 *Web Interface* Dan *Main Block Testing* 69
4.6. Hasil Dan Analisis Penelitian 70
4.6.1 Hasil *Pre-Processing* Data Setiap Tahap 70
4.6.2 Hasil Pelatihan Model 72

4.6.3 Analisis *Confusion Matrix* 74
4.6.4 Analisis Kurva Loss 78
4.6.5 Analisis Wordcloud 80
4.6.6 Tampilan Testing Interaktif 85
4.6.7 Hasil *Accuracy*, *Precision,* Dan *F1-Score* 86
4.6.8 Analisis Komparatif Dan Temuan Penelitian 88
BAB V PENUTUP 90
5.1. Kesimpulan 90
5.2. Saran 91
DAFTAR PUSTAKA 92
LAMPIRAN 97

DAFTAR GAMBAR

Gambar 2.1 Fakultas Hukum Universitas Tanjungpura 7
Gambar 2.2 DistilBERT model architecture and components. 16
Gambar 3.1 Kerangka Kerja Penelitian 26
Gambar 3.2 Flowchart Metode Penelitian 29
Gambar 4.1 Distribusi Data Per Topik Hukum 33
Gambar 4.2 Perbandingan Akurasi Model 73
Gambar 4.3 Heatmap Confusion Matrix Skenario A 75
Gambar 4.4 Heatmap Confusion Matrix Skenario A + BT 76
Gambar 4.5 Heatmap Confusion Matrix Skenario B 77
Gambar 4.6 Heatmap Confusion Matrix Skenario B + BT 78
Gambar 4.7 Kurva Loss Model 80
Gambar 4.8 Wordcloud Skenario A Label Internasional 80
Gambar 4.9 Wordcloud Skenario A Label Tatanegara 81
Gambar 4.10 Wordcloud Skenario A Label Ekonomi 81
Gambar 4.11 Wordcloud Skenario A Label Perdata 82
Gambar 4.12 Wordcloud Skenario A Label Pidana 82
Gambar 4.13 Wordcloud Skenario B Label Internasional 83
Gambar 4.14 Wordcloud Skenario B Label Tatanegara 83
Gambar 4.15 Wordcloud Skenario B Label Ekonomi 84
Gambar 4.16 Wordcloud Skenario B Label Perdata 84
Gambar 4.17 Wordcloud Skenario B Label Pidana 85
Gambar 4.18 Tampilan Testing Interaktif 86
Gambar 4.19 Wawancara pengambilan data di tempat studi kasus 97

DAFTAR TABEL

Tabel 2.1 Confusion Matrix 20
Tabel 2.2 Penelitian Terdahulu 24
Tabel 3.1 Pembagian Dataset 30
Tabel 4.1 Jumlah Abstrak di setiap Topik 32
Tabel 4.2 Tahapan *Pre-Processing* Skenario A 35
Tabel 4.3 Tahapan *Pre-Processing* Skenario B 35
Tabel 4.4 Hasil Pre-Processing Data Setiap Tahap 71
Tabel 4.5 Hasil Akurasi Pelatihan Model 73
Tabel 4.6 Hasil Accuracy, Precision, Dan F1-Score 87

DAFTAR SINTAKS KODE

Sintaks Kode 4.1 Command Terminal install Library 36
Sintaks Kode 4.2 Import Library Pre-Processing 37
Sintaks Kode 4.4 Normalisasi Istilah Hukum, Stopword Akademik.....39
Sintaks Kode 4.5 Fungsi Text Cleaning.....39
Sintaks Kode 4.6 Fungsi Case Folding dan Normalisasi 40
Sintaks Kode 4.7 Fungsi Tokenisasi, Stopword Removal, dan Stemming.....40
Sintaks Kode 4.8 Fungsi Tokenisasi DistilBERT 41
Sintaks Kode 4.9 Load Dataset dan Distribusi Topik 42
Sintaks Kode 4.10 Pipeline Skenario A 42
Sintaks Kode 4.11 Pipeline Skenario B 43
Sintaks Kode 4.12 Import Library Training Model 45
Sintaks Kode 4.13 Parameter Training Model.....46
Sintaks Kode 4.14 Mapping Label dan Parsing Token.....46
Sintaks Kode 4.15 Load dan Persiapan Data Training 47
Sintaks Kode 4.16 Pembagian Data training.....48
Sintaks Kode 4.17 Kelas HukumDatasetFromCSV.....49
Sintaks Kode 4.18 Training Model 51
Sintaks Kode 4.19 Penerapan Back Translation 52
Sintaks Kode 4.20 Persiapan Data Training dengan Back Translation 55
Sintaks Kode 4.21 Evaluasi dan Confusion Matrix 57
Sintaks Kode 4.22 Plot Hasil Perbandingan 58
Sintaks Kode 4.23 Wordcloud per Label 60
Sintaks Kode 4.24 Argument Parser dan Persiapan DataLoader.....61
Sintaks Kode 4.25 Eksekusi Training, Penyimpanan Model, dan Plot.....64
Sintaks Kode 4.26 Import Library, Konfigurasi, dan Komponen Preprocessing ..66
Sintaks Kode 4.27 Preprocessing untuk Testing.....67
Sintaks Kode 4.28 Fungsi load_models 68
Sintaks Kode 4.29 Predict.....69
Sintaks Kode 4.30 Web Interface dan Main Block Testing.....70

DAFTAR PERSAMAAN

Persamaan 2.1 ReLU.....12
Persamaan 2.2 Residual Connection.....12
Persamaan 2.3 Layer normalization 12
Persamaan 2.4 Scaled Dot-Product Attention.....13
Persamaan 2.5 Precision.....20
Persamaan 2.6 Recall 20
Persamaan 2.7 F1-Score.....21
Persamaan 2.8 Akurasi.....21

BAB I
PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi informasi dan komunikasi berdampak besar di berbagai bidang, termasuk didalamnya dunia akademik. Salah satu dampaknya adalah meningkatnya jumlah penelitian dan publikasi ilmiah, termasuk skripsi yang ditulis oleh mahasiswa. Skripsi sebagai karya ilmiah memegang peran penting dalam pengembangan ilmu pengetahuan, khususnya di bidang hukum. Tetapi, dengan semakin banyaknya skripsi yang dihasilkan, proses pengelolaan dan pengelompokan topik penelitian menjadi tantangan tersendiri. Dalam konteks ini, text mining dapat digunakan untuk mengekstraksi informasi dan pengetahuan dari data teks. Salah satu metode yang digunakan dalam text mining adalah klasifikasi teks, yaitu proses pengelompokan dokumen teks ke dalam kelas-kelas tertentu secara otomatis (Farisi et al., 2019).

Fakultas Hukum Universitas Tanjungpura Pontianak adalah salah satu institusi pendidikan tinggi yang menghasilkan banyak skripsi setiap tahun. Ketiadaan klasifikasi artikel ilmiah yang terstruktur menghambat institusi dalam mengekstraksi wawasan berharga (Wolff et al., 2024). Hal ini membuat kesulitan dalam menelusuri dan menganalisis topik-topik penelitian yang sudah dilakukan, menganalisis tren penelitian, mengidentifikasi topik yang kurang diminati, serta mengembangkan penelitian lanjutan.

Dalam beberapa tahun terakhir, model berbasis *Transformer*, seperti BERT (*Bidirectional Encoder Representations from Transformers*), telah menunjukkan bahwa model ini sebagai *State of the art* (SOTA) dalam berbagai tugas Klasifikasi (Bello et al., 2023). DistilBERT, sebagai versi yang lebih ringan dan cepat dari BERT, menawarkan keseimbangan ideal antara akurasi tinggi dan efisiensi komputasi (Fajri et al., 2022).

Kesuksesan model seperti DistilBERT tidak hanya bergantung pada arsitektur model itu sendiri, tapi juga dipengaruhi dengan kualitas data yang digunakan. Kualitas data ditentukan dari tahapan pra-pemrosesan data dan komposisi dataset

(Schröer et al., 2021). Masalah lain yang sering muncul dalam klasifikasi dokumen di dunia nyata adalah ketidakseimbangan data (*imbalanced data*), yaitu ketimpangan jumlah sampel antar kelas. Pada konteks penelitian ini, tidak semua topik hukum memiliki jumlah skripsi yang sama, topik seperti "Hukum Perdata" dapat memiliki jauh lebih banyak dokumen dibandingkan "Hukum Internasional". Jika tidak ditangani, model akan cenderung lebih memihak kelas mayoritas, yang mengakibatkan akurasi tinggi secara semu namun performa (F1-Score) buruk pada kelas minoritas. Untuk mengatasi ketidakseimbangan kelas pada *dataset* dalam penelitian ini, teknik augmentasi data seperti *back translation* dapat digunakan. *Back translation* bekerja dengan menerjemahkan teks dari bahasa sumber ke bahasa lain, lalu diterjemahkan kembali ke bahasa asal sehingga menghasilkan variasi parafrase yang alami tanpa mengubah makna inti. Hasil penelitian Rajagopalan Nair et al. (2024) menunjukkan bahwa augmentasi data teks menggunakan *back translation* mampu meningkatkan performa klasifikasi pada model DistilBERT secara signifikan.

Piasta & Kotas (2025) dalam studi komparatifnya menggunakan model DistilBERT untuk klasifikasi artikel berita menemukan bahwa perbedaan hasil antara teks yang melalui pra-pemrosesan lengkap (lemmatization, Punctuation removal) dengan akurasi 91% dan teks yang hanya melalui pra-pemrosesan minimal (penghapusan diakritik dan URL) dengan akurasi 90% hanya berbeda tipis. Temuan ini yang menjadi acuan utama dalam penelitian ini untuk melihat pengaruh pra-pemrosesan terhadap model DistilBERT ketika diterapkan pada *dataset* hukum berbahasa Indonesia yang kaya akan singkatan dan istilah teknis.

Oleh karena itu, penelitian ini bertujuan untuk menggunakan model DistilBERT, juga menganalisis pengaruh skenario pra-pemrosesan data dapat mempengaruhi akurasi dan keandalan model DistilBERT dalam mengklasifikasikan topik penelitian hukum di FH UNTAN.

1.2. Rumusan Masalah

Rumusan masalah dari penelitian "Evaluasi Efektivitas Pra-Pemrosesan Terhadap Akurasi Model Distilbert Pada Klasifikasi Topik Penelitian Hukum (Studi Kasus: Fakultas Hukum Universitas Tanjungpura Kota Pontianak)" yaitu:

1. Bagaimana cara mengelompokkan skripsi mahasiswa berdasarkan topik penelitian hukum?
2. Bagaimana pengaruh pra-pemrosesan data terhadap akurasi model klasifikasi?

1.3. Tujuan Penelitian

Adapun tujuan dari adanya penelitian ini antara lain:

1. Melatih model klasifikasi menggunakan arsitektur DistilBERT untuk mengelompokkan abstrak skripsi mahasiswa.
2. Menganalisis dampak dan performa model DistilBERT yang dilatih menggunakan data dengan tahapan *pre-processing minimal* dan *pre-processing full*.

1.4. Batasan Masalah

Adapun batasan masalah yang terdapat dalam penelitian ini antara lain:

1. Penelitian ini hanya berfokus pada klasifikasi topik penelitian hukum berdasarkan abstrak skripsi mahasiswa di Fakultas Hukum Universitas Tanjungpura Pontianak.
2. Data yang digunakan adalah abstrak skripsi mahasiswa Sarjana Fakultas Hukum Universitas Tanjungpura Pontianak.
3. Model klasifikasi yang digunakan terbatas pada DistilBERT (distilbert-base-multilingual-cased atau model sejenis yang relevan untuk Bahasa Indonesia).
4. Penelitian ini juga melihat pengaruh pra-pemrosesan dan teknik augmentasi data (Back Translation), bukan pada hyperparameter tuning model DistilBERT secara ekstensif.
5. Skenario pra-pemrosesan akan dibatasi pada beberapa kombinasi umum (misal: *minimal pre-processing*, dan *full pre-processing*).

6. Artefak evaluasi yang digunakan adalah Confusion Matrix.
7. Penelitian ini tidak membangun sistem informasi atau antarmuka pengguna (UI) berbasis web, melainkan berfokus pada eksperimen dan analisis model.

1.5. Manfaat Penelitian

Adapun manfaat dengan adanya penelitian ini antara lain:

1. Bagi Peneliti

   Diharapkan dapat meningkatkan pemahaman tentang implementasi model Transformer (DistilBERT) dan memberikan wawasan praktis mengenai pentingnya pra-pemrosesan serta penanganan data tidak seimbang dalam tugas klasifikasi teks.
2. Bagi Tempat Studi Kasus

   Diharapkan dapat memberikan blueprint atau model dasar untuk pengembangan sistem klasifikasi skripsi otomatis di masa depan. Ini dapat membantu administrasi, perpustakaan, dan pimpinan program studi dalam memetakan tren penelitian mahasiswa.
3. Bagi Jurusan

   Diharapkan dapat menjadi referensi ilmiah mengenai perbandingan efek pra-pemrosesan dan Back Translation pada model berbasis Transformer untuk klasifikasi teks berbahasa Indonesia di domain spesifik (hukum).

1.6. Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memberikan gambaran yang jelas mengenai alur pembahasan dalam penelitian, sehingga memudahkan pembaca dalam memahami isi setiap bab. Adapun sistematika penulisan skripsi ini adalah sebagai berikut:

1. **BAB I PENDAHULUAN**

   Bab ini berisi uraian mengenai latar belakang penelitian yang menjelaskan mengapa proses *pre-processing* akan diteliti untuk hasil akurasi model. Selain itu, bab ini juga memuat rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, serta sistematika penulisan skripsi.