BAB I
PENDAHULUAN

Bab ini membahas pendahuluan yang mencakup latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan skripsi mengenai prediksi lama hukuman tindak pidana berdasarkan dokumen riwayat putusan pengadilan Mahkamah Agung menggunakan model *word embedding* BERT (*Bidirectional Encoder Representations from Transformers*) Hierarchical dan model LSTM (*long short-term memory*).

1.1. Latar Belakang

Menurut Supriyadi (2015) tindak pidana merupakan perbuatan yang oleh peraturan perundang-undangan diancam dengan sanksi pidana dan/atau tindakan yang diatur pada pasal 53 KUHP ayat (1) yang menyatakan "Percobaan melakukan kejahatan dapat dikenakan pidana, jika niat untuk melakukannya telah terbukti dari adanya permulaan pelaksanaan, dan tidak terselesainya tindakan tersebut disebabkan oleh hal-hal di luar kehendaknya sendiri.". Untuk dinyatakan sebagai tindak pidana, suatu perbuatan yang diancam dengan sanksi pidana dan tindakan oleh peraturan perundang-undangan harus bersifat melawan hukum atau bertentangan dengan hukum yang hidup dalam masyarakat. Setiap tindak pidana selalu bersifat melawan hukum, kecuali ada alasan pembenar. Tindak pidana mencakup berbagai jenis kejahatan seperti pembunuhan, pencurian, penipuan, dan korupsi. Setiap tindak pidana ini memiliki karakteristik dan dampak yang berbeda-beda terhadap masyarakat dan negara.

Penegakan hukum terhadap tindak pidana umum memerlukan upaya yang terorganisir dan sistematis, salah satunya melalui pengadilan Mahkamah Agung (MA) Republik Indonesia. Pengadilan memainkan peran penting dalam menjatuhkan hukuman yang sesuai dan adil bagi pelaku kejahatan. Selain itu, pengadilan menjaga konsistensi penerapan hukum melalui putusan kasasi dan peninjauan kembali, serta memastikan bahwa semua hukum dan undang-undang di seluruh wilayah Republik Indonesia diterapkan secara adil, tepat, dan benar. Dokumen hasil putusan pengadilan, terutama hasil sidang MA Republik Indonesia, sangat penting sebagai

acuan hukum dan panduan bagi pengadilan tingkat bawah ketika menghadapi kasus-kasus yang melibatkan delik hukum, yaitu kondisi di mana terjadi kekosongan hukum, yakni saat tidak ada regulasi untuk suatu kasus tertentu. Dalam hal ini, keputusan hakim sebelumnya juga harus menjadi instrumen hukum untuk menjaga kepastian hukum (Simanjuntak, 2019). Dokumen keputusan MA ini mencerminkan bagaimana hukum diterapkan dan keadilan ditegakkan di Indonesia. Melalui analisis dokumen tersebut, masyarakat dapat memahami berbagai pertimbangan hukum yang digunakan oleh pengadilan MA dalam menjatuhkan putusan, termasuk faktor-faktor yang memperberat atau meringankan hukuman.

Analisis dokumen putusan pengadilan akan menjadi sulit jika jumlah dokumen sangat banyak dan setiap bagian dalam putusan sidang harus diperhatikan secara seksama untuk menentukan lama hukuman dalam suatu perkara pidana. Oleh karena itu, diperlukan pemanfaatan teknologi dengan regresi untuk menganalisis setiap dokumen hasil sidang perkara pidana pengadilan MA. Regresi adalah metode statistik yang digunakan untuk memahami hubungan antara variabel dependen (yang ingin diprediksi atau dijelaskan) dengan satu atau lebih variabel independen (yang digunakan untuk membuat prediksi) atau bagaimana variabel-variabel itu berhubungan atau dapat diramalkan. Pengertian regresi secara umum adalah sebuah alat statistik yang memberikan penjelasan tentang pola hubungan (model) antara dua variabel atau lebih (Suhandi dkk., 2018).

Penelitian mengenai regresi dokumen pengadilan untuk memprediksi lama hukuman, belum banyak dilakukan. Penelitian Nuranti dkk., (2022) memprediksi kategori dan lama hukuman untuk kasus hukum tertentu berdasarkan kasus serupa sebelumnya yang diarsipkan dalam dokumen keputusan MA Republik Indonesia dengan penggunaan metode Word2Vec dan FastText untuk memproses teks dokumen dan metode *deep learning*. Dari penelitian tersebut, diperoleh nilai *R2 score* untuk LSTM+*attention* sebesar 38,65% dan *R2 score* untuk BiLSTM+*attention* sebesar 41,54%. Nilai kesalahan yang cukup besar tersebut dapat disebabkan oleh kekurangan dalam mendalami konteks berdasarkan hasil *word embedding Word2Vec* yang tidak cocok dengan teks yang mempunyai banyak kemiripan, di mana dokumen hasil persidangan di Indonesia memiliki banyak frasa yang diulang-ulang (Nuranti dkk., 2022). Dengan demikian menyebabkan LSTM

tidak begitu efektif dengan Word2Vec dan FastText (Van Houdt dkk., 2020). Menurut Vashisth dkk., (2020) hasil Word2Vec menghasilkan *word embedding* yang sama untuk satu kata meskipun kata tersebut muncul pada lokasi yang berbeda, sedangkan BERT menghasilkan *word embedding* yang berbeda untuk satu kata yang muncul pada lokasi yang berbeda karena pada kata sebelum dan sesudah juga dihitung *vector word embedding*-nya. Oleh karena itu, apabila BERT digabungkan dengan LSTM, maka dapat memberikan hasil yang lebih optimal.

Oleh karena kekurangan metode *Word2Vec* dan *FastText* dengan LSTM, penelitian ini mencoba menggunakan BERT. Hal ini dikarenakan kemampuan BERT untuk menangani berbagai variasi dalam teks dengan lebih efektif. BERT memiliki keunggulan dalam memprediksi hubungan antara kalimat dengan menganalisisnya secara keseluruhan. Hal tersebut sangat berguna untuk memahami konteks yang lebih luas dalam teks, memungkinkan model untuk melakukan tugas-tugas pada tingkat token seperti pengenalan entitas kata (Devlin dkk., 2018). BERT dapat menghasilkan representasi yang lebih kaya dan mendalam dari teks yang diolahnya, memberikan dasar yang kuat untuk analisis lanjutan. LSTM dapat dilatih lebih optimal dengan memanfaatkan representasi kontekstual yang dihasilkan oleh BERT untuk memperhitungkan urutan dan dinamika dalam teks dengan lebih baik. Kombinasi ini memungkinkan model untuk memahami dan memprediksi hubungan antara kalimat secara lebih mendalam, serta mengatasi tugas-tugas pada tingkat token dengan akurasi yang lebih tinggi.

Penelitian ini mengusulkan untuk melakukan regresi pada dokumen riwayat putusan pengadilan MA Republik Indonesia. Data yang digunakan berasal dari *dataset* penelitian yang disusun oleh Nuranti dkk., (2022) yang tersedia pada tautan GitHub https://github.com/ir-nlp-csui/indo-law. *Dataset* ini berisi dokumen putusan pengadilan Indonesia untuk kasus pidana umum dan pidana khusus, dengan jumlah total dokumen mencapai 22.630. Kemudian melakukan prediksi lama hukuman (regresi) pada suatu tindak pidana berdasarkan dokumen riwayat putusan pengadilan Mahkamah Agung. Untuk mencapai tujuan tersebut, penelitian menggunakan pendekatan yang menggabungkan model BERT untuk memproses teks dengan *word embedding Hierarchical* BERT dan model LSTM untuk regresi.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dirumuskan permasalahan dari penelitian skripsi ini adalah bagaimana kinerja regresi dari model *word embedding Hierarchical* BERT dan model LSTM dalam memprediksi lama hukuman tindak pidana berdasarkan dokumen riwayat putusan pengadilan MA. Dari permasalahan tersebut, dapat diturunkan ke dalam permasalahan khusus yaitu:

1. Bagaimana pengaruh akurasi beberapa model *word embedding* BERT dan model prapemrosesan data terhadap kinerja regresi pada dokumen putusan pengadilan MA Republik Indonesia?
2. Bagaimana pengaruh berbagai *hyperparameter* pada model LSTM terhadap kinerja hasil regresi?
3. Seberapa efektifkah metode LSTM dengan mekanisme attention dalam memprediksi lama hukuman berdasarkan dokumen putusan pengadilan di Indonesia?

1.3. Tujuan Penelitian Dan Manfaat

Tujuan secara umum dalam penelitian ini adalah sebagai berikut:

1. Mengembangkan model untuk memprediksi lama hukuman menggunakan *word embedding hierarchical* BERT dengan kombinasi pelatihan regresi LSTM.
2. Mengevaluasi pengaruh hasil regresi dari model *pretraining* BERT, model prapemrosesan, dan hyperparameter LSTM terhadap kinerja model prediksi lama hukuman, dengan menggunakan metode LSTM dan mekanisme attention berdasarkan dokumen putusan pengadilan.

Penelitian ini akan memberikan manfaat dalam memperdalam pemahaman tentang penerapan regresi pada dokumen putusan pengadilan, khususnya dalam menganalisis pengaruh model *word embedding* BERT dan model prapemrosesan data terhadap kinerja regresi, serta mengevaluasi metode LSTM dengan mekanisme *attention.* Bagi Universitas Diponegoro (UNDIP), penelitian ini akan menaikan kontribusi akademik dan bidang terkait untuk perkuliahan pemrosesan bahasa alami, khususnya dalam kasus menyelesaikan analisis dokumen hukum.

1.4. Ruang Lingkup Penelitian

Penelitian dalam skripsi ini memerlukan adanya ruang lingkup agar penelitian yang dilakukan lebih terarah dan mencapai sasaran yang diharapkan. Dalam menerapkan penelitian ini untuk regresi dari model *word embedding Hierarchical* BERT dan model LSTM dalam memprediksi lama hukuman, seperti berikut:

1. Penelitian ini berfokus pada analisis dan pemrosesan *dataset* dokumen putusan pengadilan Indonesia yang mencakup kasus perkara pidana umum dan pidana khusus. *Dataset* ini terdiri dari 22.630 dokumen yang telah dianotasi berdasarkan bagian-bagian spesifik dari dokumennya dan hanya menggunakan 4 fitur berupa riwayat tuntutan, fakta, fakta hukum, dan pertimbangan hukum. Pemilihan fitur-fitur ini didasarkan pada penelitian yang dilakukan oleh Nuranti dkk., (2022)
2. *Dataset* tersebut dibagi menjadi tiga subset utama dengan proporsi 80:10:10, di mana 80% digunakan untuk pelatihan (*training*), 10% untuk pengujian model selama pelatihan (*validation*), dan 10% sisanya untuk evaluasi akhir hasil (*testing*).

1.5. Sistematika Penulisan

Sistematika penulisan yang digunakan dalam skripsi ini terbagi dalam beberapa pokok bahasan, yaitu:

BAB I PENDAHULUAN

Bab ini membahas mengenai latar belakang, rumusan masalah, tujuan dan manfaat, serta ruang lingkup pelaksanaan skripsi yang berfokus pada penerapan model BERT dan LSTM dalam analisis dokumen putusan pengadilan Indonesia untuk prediksi regresi lama hukuman berdasarkan dokumen hasil persidangan MA Republik Indonesia.

BAB II TINJAUAN PUSTAKA

Bab ini membahas kajian pustaka yang berhubungan dengan skripsi sebagai landasan untuk merumuskan dan menganalisis permasalahan. Kajian pustaka meliputi penelitian terdahulu terkait prediksi lama hukuman dari dokumen hukum, teori dasar mengenai model BERT dan

LSTM, dasar mengenai prapemrosesan, teori dasar mengenai pemrosesan bahasa alami (Natural Language Processing), regresi linear, bagian data yang digunakan, dan rumus evaluasi yang digunakan.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan penelitian skripsi yang meliputi pengumpulan data dokumen hasil persidangan Mahkamah Agung Republik Indonesia yang diambil dari Indo-Law Dataset dan dikonversi menjadi format CSV; proses anotasi dan prapemrosesan data yang mencakup penghapusan nilai kosong, normalisasi teks, penghapusan spasi berlebih, perubahan teks menjadi huruf kecil, tokenisasi, penghapusan stopwords, serta stemming menggunakan pustaka yang sesuai; ekstraksi fitur menggunakan model BERT yang melibatkan token embedding, segment embedding, dan positional embedding, serta penerapan metode Hierarchical BERT untuk mengatasi batasan panjang input; pembagian dataset menjadi data latih (80%), data uji (10%), dan data validasi (10%) untuk mengevaluasi kinerja model; pelatihan model LSTM dengan input representasi BERT dan berbagai kombinasi hyperparameter seperti batch size, learning rate, hidden size, num layers, dan epochs melalui grid search untuk menentukan konfigurasi terbaik; serta evaluasi model dengan menghitung rata-rata hasil pelatihan dari setiap kombinasi hyperparameter menggunakan metode post-hoc yang memanfaatkan fungsi idxmin dan idxmax dari pustaka Pandas untuk mencari nilai MSE dan SMAPE terendah serta R2 score tertinggi, yang membantu menentukan performa terbaik model dalam prediksi regresi lama hukuman.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini menguraikan hasil skenario dan analisis eksperimen yang dimulai dari teknis pengumpulan data sampai hasil dan analisis dari setiap eksperimen yang dilakukan. Termasuk di dalamnya adalah deskripsi *dataset*, hasil pengujian model BERT dan LSTM, serta interpretasi temuan berdasarkan analisis yang telah dilakukan.

BAB V PENUTUP

Bab ini menjelaskan mengenai kesimpulan dari uraian yang telah dijabarkan pada bab-bab sebelumnya dan saran untuk pengembangan penelitian lebih lanjut. Kesimpulan mencakup ringkasan temuan utama dan implikasinya, sedangkan saran difokuskan pada peningkatan metodologi dan eksplorasi lebih lanjut dalam konteks prediksi lama hukuman dari dokumen hasil persidangan di Indonesia.