KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
PROGRAM STUDI S1 TEKNOLOGI INFORMASI
Jalan Alumni No. 3 Gedung C, Kampus USU Padang Bulan, Medan 20155
Telepon/Fax: 061-8210077 | Email: tek.informasi@usu.ac.id | Laman: http://it.usu.ac.id

FORM PENGAJUAN JUDUL

Nama : Warida Hafni Hasibuan

NIM : 201402018

Judul diajukan oleh* : ☐ Dosen
☑ Mahasiswa

Bidang Ilmu (tulis dua bidang) : Machine Learning

Uji Kelayakan Judul** : ◯ Diterima ◯ Ditolak

Hasil Uji Kelayakan Judul :

Calon Dosen Pembimbing I: Dr. Muhammad Anggia Muchtar, S.T., MM.IT.
(Jika judul dari dosen maka dosen tersebut berhak menjadi pembimbing I)

Calon Dosen Pembimbing II: Ade Sarah Huzaifah, S.Kom., M.Kom

Paraf Calon Dosen Pembimbing I

Medan, Juli 2024

Ka. Laboratorium Penelitian,

** Centang salah satu atau keduanya*

*** Pilih salah satu*

(Fanindia Purnamasari, S.TI., M.IT)

NIP. 198908172019032023

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
PROGRAM STUDI S1 TEKNOLOGI INFORMASI
Jalan Alumni No. 3 Gedung C, Kampus USU Padang Bulan, Medan 20155
Telepon/Fax: 061-8210077 | Email: tek.informasi@usu.ac.id | Laman: http://it.usu.ac.id

RINGKASAN JUDUL YANG DIAJUKAN

*Semua kolom di bawah ini diisi oleh mahasiswa yang sudah mendapat judul

Judul / Topik Skripsi	**Perancangan Sistem Deteksi Kesamaan Dokumen Executive Summary dengan menggunakan Large Language Models pada Proses Pengajuan Skripsi di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara**
Latar Belakang dan Penelitian Terdahulu	**Latar Belakang** Di era digital yang terus berkembang saat ini, teknologi telah menjadi elemen fundamental dalam berbagai aspek kehidupan manusia, termasuk dalam dunia pendidikan. Transformasi digital telah membawa perubahan signifikan dalam cara kita mengakses, menyimpan, dan memproses informasi. Dalam konteks pendidikan, institusi pendidikan tinggi di seluruh dunia mengalami peningkatan volume dan keragaman dokumen akademik yang harus dikelola, terutama dalam proses pengajuan skripsi. Salah satu fenomena yang mencolok adalah tantangan dalam pemeriksaan kesamaan dokumen akademik seperti executive summary, yang menjadi prasyarat sebelum mahasiswa dapat mengajukan proposal skripsi. Proses ini penting untuk memastikan bahwa topik penelitian yang diajukan adalah orisinal dan tidak duplikasi dari penelitian sebelumnya. Di Fakultas Ilmu Komputer dan Teknologi Informasi, proses pengecekan kesamaan dokumen executive summary saat ini masih dilakukan secara manual oleh kepala bidang. Proses manual ini memiliki beberapa kelemahan utama. Pertama, membutuhkan waktu dan tenaga yang signifikan, mengingat jumlah proposal yang harus diperiksa semakin meningkat setiap tahunnya. Kedua, metode manual ini rentan terhadap kesalahan subjektif yang dapat mempengaruhi keakuratan penilaian. Kesalahan dalam penilaian ini tidak hanya dapat menyebabkan redundansi dalam topik penelitian tetapi juga dapat berdampak serius pada kualitas akademik institusi. Kesalahan subjektif dalam pengecekan manual dapat disebabkan oleh berbagai faktor, termasuk kelelahan, ketidakjelasan dalam pedoman penilaian, dan interpretasi pribadi. Selain itu, beban kerja yang tinggi pada staf akademik juga dapat mengurangi waktu yang tersedia untuk evaluasi yang teliti dan mendalam. Dengan demikian, ada kebutuhan mendesak untuk mengembangkan sistem yang lebih efisien dan akurat dalam mendeteksi kesamaan dokumen executive summary. Dalam konteks pendidikan tinggi saat ini, penelitian ini sangat relevan dan penting. Untuk membuat proses akademik lebih jelas dan berhasil, peningkatan efisiensi dan akurasi dalam proses pembuatan dokumen sangat penting. Alat teknologi canggih seperti *Large Language Models* (LLM) memiliki banyak peluang di bidang ini. Seperti BERT (*Bidirectional Encoder Representations from Transformers*), LLM mampu memahami dan membuat human language dengan sangat akurat. BERT dapat digunakan dalam sistem deteksi dokumen untuk mengurangi biaya tenaga kerja, meningkatkan

kualitas tulisan, dan menjamin bahwa hasil analisis adalah tepat dan sah. Dengan menggunakan teknologi ini, diharapkan bahwa implementasi BERT akan memungkinkan adopsi

Penelitian ini sangat relevan dan penting dalam konteks pendidikan saat ini. Peningkatan efisiensi dan akurasi dalam proses pengecekan kesamaan dokumen sangat penting untuk mendukung proses akademik yang lebih baik dan transparan. Teknologi canggih seperti *Large Language Models* (LLM) menunjukkan potensi besar dalam bidang ini. LLM, seperti BERT (*Bidirectional Encoder Representations from Transformers*), dilatih dengan sejumlah besar data teks dan mampu memahami serta menghasilkan bahasa manusia dengan akurasi tinggi. Implementasi BERT dalam sistem deteksi kesamaan dokumen dapat mengurangi beban kerja manual, meningkatkan akurasi penilaian, dan memastikan bahwa topik penelitian yang diajukan benar-benar orisinal. Dengan memanfaatkan teknologi BERT, diharapkan dapat mengadopsi pendekatan yang lebih modern dan efektif dalam pengelolaan dokumen akademik. Penelitian ini juga berpotensi memberikan kontribusi signifikan pada pengembangan teknologi pemrosesan bahasa alami (NLP) di Indonesia. Penerapan BERT tidak hanya meningkatkan efisiensi proses penilaian tetapi juga mendukung integritas akademik dengan mencegah duplikasi penelitian.

Dalam penelitian ini, data yang diperlukan untuk melatih dan menguji model akan dikumpulkan menggunakan API dari website repository Universitas Sumatera Utara. Penggunaan API memungkinkan pengumpulan data secara otomatis dari website, termasuk metadata penting seperti judul, penulis, abstrak, metode, algoritma, rumusan masalah, diagram arsitektur, dan referensi. Data ini akan digunakan untuk membangun dan melatih model BERT sehingga dapat mendeteksi kesamaan teks dengan lebih akurat dan efisien. Proses pengumpulan data akan dilakukan dengan memperhatikan aspek legal dan etis, serta menggunakan API yang disediakan oleh repository USU.

Berberapa penelitian sebelumnya telah mengeksplorasi metode untuk deteksi kesamaan teks. Arnold Pramudita Tjiawi et al. (2018) menggunakan algoritma *Winnowing* yang efektif untuk mendeteksi kesamaan lokal namun rentan terhadap perubahan struktur kalimat. Berlin Ong et al. (2020) menerapkan metode *Latent Semantic Analysis* (LSA) dan cosine similarity untuk menganalisis struktur semantik teks dan menghitung kemiripan, meskipun akurasinya masih perlu ditingkatkan. Devlin et al. (2019) memperkenalkan model BERT, yang menunjukkan kinerja superior dalam berbagai tugas NLP, termasuk deteksi kesamaan teks. Berdasarkan kajian literatur tersebut, terlihat bahwa penggunaan *Large Language Models* (LLMs) memiliki potensi besar untuk meningkatkan akurasi dan efisiensi dalam deteksi kesamaan teks. Oleh karena itu, penelitian ini akan memanfaatkan BERT untuk mengembangkan sistem deteksi kesamaan executive summary yang lebih baik.

Meskipun berbagai metode untuk deteksi kesamaan teks telah dikembangkan, masing-masing memiliki keterbatasan. Algoritma *Winnowing* yang digunakan oleh Arnold Pramudita Tjiawi et al. (2018) efektif untuk mendeteksi kesamaan lokal namun rentan terhadap perubahan struktur kalimat. Metode LSA dan *cosine similarity* yang diterapkan oleh Berlin Ong et al. (2020) mampu mendeteksi kemiripan meskipun ada perubahan struktur kalimat, namun akurasinya masih perlu ditingkatkan. Model berbasis transformer seperti BERT yang diperkenalkan oleh Devlin et al. (2019) menunjukkan kinerja superior dalam berbagai tugas NLP, termasuk deteksi kesamaan teks, namun membutuhkan fine-tuning dan data yang besar untuk hasil optimal.

Oleh karena itu, diperlukan penelitian yang lebih mendalam dan inovatif untuk mengembangkan sistem deteksi kesamaan teks yang lebih akurat dan efisien, khususnya untuk executive summary dalam pengajuan skripsi di Fakultas Ilmu Komputer dan Teknologi Informasi. Implementasi BERT berpotensi untuk memberikan solusi yang lebih baik dengan akurasi tinggi dan kemampuan menangani variasi dalam struktur kalimat dan konteks yang kompleks.

Penelitian ini bertujuan untuk mengembangkan dan mengimplementasikan sistem deteksi kesamaan executive summary menggunakan BERT di Fakultas Ilmu Komputer dan Teknologi Informasi. Sistem ini diharapkan dapat membantu dalam proses pengajuan skripsi, memberikan kemudahan bagi Kepala bidang dalam mengevaluasi orisinalitas karya ilmiah mahasiswa, dan meningkatkan standar akademik di fakultas ini.

Dengan latar belakang yang komprehensif ini, diharapkan penelitian ini dapat memberikan kontribusi signifikan dalam bidang deteksi kesamaan teks, khususnya dalam konteks pengajuan skripsi di Fakultas Ilmu Komputer dan Teknologi Informasi. Implementasi sistem deteksi kesamaan berbasis BERT diharapkan dapat meningkatkan efisiensi, akurasi, dan integritas proses penilaian akademik, serta mendukung upaya peningkatan kualitas pendidikan tinggi di Indonesia.Penelitian ini berjudul "**Perancangan Sistem Deteksi Kesamaan Dokumen Executive Summary dengan menggunakan Large Language Models pada Proses Pengajuan Skripsi di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara**".

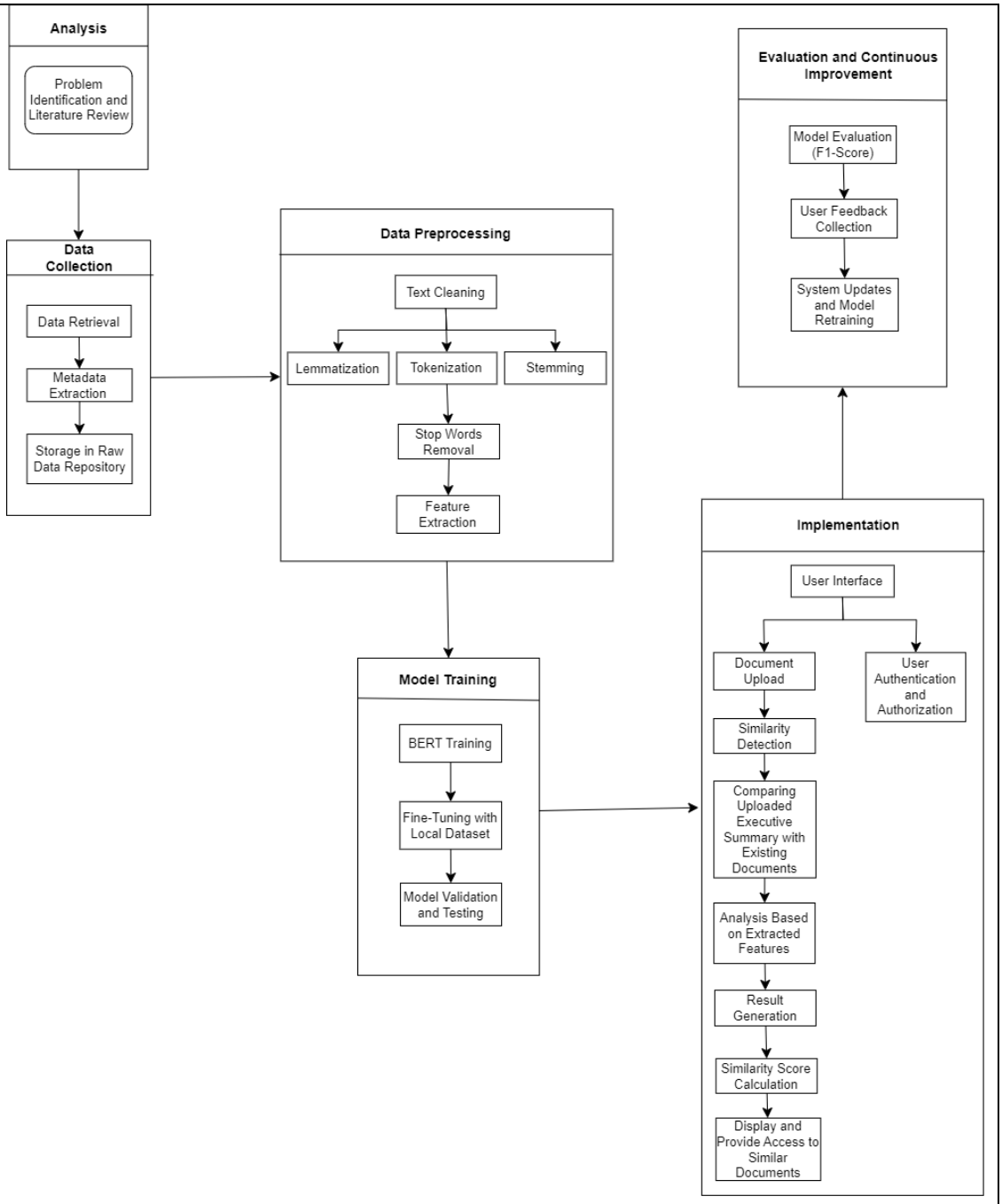
Penelitian Terdahulu

No.	Penulis	Judul	Tahun
1.	Berlin Ong, Dali S. Naga, Viny Christanti M.	Perancangan Aplikasi Pendeteksi Kemiripan Teks Dengan Menggunakan Metode Latent Semantic Analysis	2020
2.	Arnold Pramudita Tjiawi, Dyah E. Herwindiati, Lely Hiryanto	Perancangan Aplikasi Pendeteksi Tingkat Kesamaan Antar Dokumen Dengan Algoritma Winnowing	2018
3.	Denghui Yang, Dengyun Zhu, Hailong Gai, Fucheng Wan	Semantic Similarity Calculating Based on BERT	2024
4.	Kamal Fauzan Navaro, Septiyawan Rosetya Wardhana, Rinci Kembang Hapsari	Deteksi Plagiarisme Artikel Jurnal menggunakan Latent Semantic Analysis (LSA)	2023
5.	Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, Georg Rehm	Aspect-based Document Similarity for Research Papers	2020

	6.	Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanov	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	2019

Rumusan Masalah	Proses pengecekan executive summary secara manual dalam mendeteksi kesamaan dokumen pada pengajuan skripsi di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara tidak efisien dan kurang akurat. Metode ini memerlukan banyak waktu dan tenaga serta rentan terhadap kesalahan subjektif yang dapat menyebabkan duplikasi dalam topik penelitian. Kesalahan tersebut dapat terjadi akibat kelelahan, ketidakjelasan pedoman penilaian, dan interpretasi pribadi. Selain itu, beban kerja yang tinggi pada staf akademik mengurangi waktu yang tersedia untuk evaluasi yang teliti dan mendalam. Kepala bidang harus memeriksa berbagai elemen dalam executive summary seperti judul, metode, rumusan masalah, dan arsitektur umum. Untuk mengatasi masalah ini, penelitian ini mengusulkan pengembangan sistem otomatis yang lebih efisien dan akurat dalam mendeteksi kesamaan dokumen. Penelitian ini juga akan mengumpulkan data yang relevan untuk melatih model bahasa, dengan tujuan mengurangi beban kerja manual pada staf akademik, meningkatkan akurasi dan efisiensi penilaian, serta memastikan orisinalitas topik penelitian yang diajukan.

Metodologi



Tahapan-Tahapan Penelitian:

1. Analysis

   Pada tahap pengenalan masalah dan kajian literatur, fokus utama adalah memahami permasalahan dalam proses pengecekan kesamaan dokumen yang saat ini dilakukan secara manual di Fakultas Ilmu Komputer dan Teknologi Informasi. Langkah ini melibatkan identifikasi masalah utama yang dihadapi, seperti efisiensi waktu, akurasi, dan sumber daya yang dibutuhkan untuk

melakukan pengecekan manual. Selain itu, dilakukan kajian literatur untuk memahami solusi yang telah ada dan teknologi yang dapat digunakan, termasuk Large Language Models (LLM) seperti BERT, serta metode machine learning lainnya yang relevan.

2. Data Collection

   Pada tahap pengumpulan data, data yang diperlukan untuk melatih dan menguji model dikumpulkan. Proses ini melibatkan penggunaan API untuk mengumpulkan data dari repository Universitas Sumatera Utara. Data yang dikumpulkan mencakup metadata penting seperti judul, penulis, abstrak, metode, algoritma, rumusan masalah, diagram arsitektur, dan referensi. Data yang telah dikumpulkan kemudian disimpan dalam repositori data mentah untuk digunakan dalam tahap berikutnya.

3. Data Prepocessing

   Pada tahap preprocessing data, data yang telah dikumpulkan dipersiapkan agar dapat digunakan untuk melatih model. Proses ini meliputi pembersihan teks dari karakter khusus dan elemen tidak relevan, tokenisasi teks menjadi kata-kata individual, lemmatization untuk mengubah kata ke bentuk dasarnya, dan stemming untuk menghilangkan imbuhan. Selain itu, kata-kata umum yang tidak bermakna penting dihapus (stop words removal), dan fitur-fitur penting dari dokumen seperti judul, metode, algoritma, rumusan masalah, dan arsitektur umum diekstraksi untuk digunakan dalam pelatihan model.

4. Model Training

   Tahap pelatihan model melibatkan penggunaan data yang telah dipreproses untuk melatih model machine learning. Model BERT dilatih untuk memahami konteks dalam kalimat menggunakan dataset yang telah dipreproses. Setelah pelatihan awal, dilakukan fine-tuning dengan dataset lokal untuk meningkatkan relevansi dan akurasi model dalam konteks spesifik. Proses ini diakhiri dengan

validasi dan pengujian model untuk memastikan performa yang optimal dan keakuratan hasil yang dihasilkan oleh model.

5. Implementation

   Pada tahap implementasi sistem, sistem yang akan digunakan oleh pengguna akhir dikembangkan. Ini mencakup pengembangan antarmuka pengguna untuk unggah dokumen dan otentikasi pengguna, serta fitur untuk membandingkan executive summary yang diunggah dengan dokumen yang ada dalam database berdasarkan fitur yang diekstraksi. Hasil analisis kesamaan kemudian dihitung dan ditampilkan dalam bentuk skor kesamaan, serta memberikan akses langsung ke file skripsi yang memiliki kesamaan.

6. Evaluation and Continuous Improvement

   Pada tahap terakhir ini,adalah tahapan evaluasi sistem, di mana kinerja sistem yang telah dibangun dievaluasi dan dilakukan perbaikan berkelanjutan. Model dievaluasi menggunakan metrik F1-score untuk mengukur keseimbangan antara precision dan recall. Umpan balik dari pengguna dikumpulkan untuk memahami kelemahan sistem dan area yang perlu diperbaiki. Berdasarkan hasil evaluasi dan umpan balik pengguna, sistem diperbarui dan model dilatih ulang untuk memastikan sistem tetap efektif dan relevan dengan kebutuhan pengguna.

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
PROGRAM STUDI S1 TEKNOLOGI INFORMASI
Jalan Alumni No. 3 Gedung C, Kampus USU Padang Bulan, Medan 20155
Telepon/Fax: 061-8210077 | Email: tek.informasi@usu.ac.id | Laman: http://it.usu.ac.id

Referensi

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Fauzan Navaro, K., Rosetya Wardhana, S., & Hapsari, R. K. (2023). Deteksi Plagiarisme Artikel Jurnal menggunakan Latent Semantic Analysis (LSA). *Seminar Nasional Sains Dan Teknologi Terapan XI*, 1–8.

Karo Karo, B. O., Naga, D. S., & Mawardi, V. C. (2020). Perancangan Aplikasi Pendeteksi Kemiripan Teks Dengan Menggunakan Metode Latent Semantic Analysis. *Computatio : Journal of Computer Science and Information Systems*, *4*(1), 1. https://doi.org/10.24912/computatio.v4i1.7191

Ostendorff, M., Ruas, T., Blume, T., Gipp, B., & Rehm, G. (2020). Aspect-based Document Similarity for Research Papers. *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, *2019*, 6194–6206. https://doi.org/10.18653/v1/2020.coling-main.545

Tjiawi, A. P., Herwindiati, D. E., & Hiryanto, L. (2018). Perancangan Aplikasi Pendeteksi Tingkat Kesamaan Antar Dokumen Dengan Algoritma Winnowing. *Computatio : Journal of Computer Science and Information Systems*, *2*(1), 36. https://doi.org/10.24912/computatio.v2i1.1918

Wahyuni, R. T., Prastiyanto, D., & Supraptono, E. (2017). Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi. *Jurnal Teknik Elektro Universitas Negeri Semarang*, *9*(1), 18–23. https://journal.unnes.ac.id/nju/index.php/jte/article/download/10955/6659

Yang, D., Zhu, D., Gai, H., & Wan, F. (2024). *Semantic Similarity Caculating based on BERT*. *2*, 73–79.

Medan, 5 Juli 2024

Mahasiswa yang mengajukan,

(Warida Hafni Hasibuan)
NIM.201402018