BAB I PENDAHULUAN

1.1 Latar Belakang

Dalam era digital saat ini, perkembangan penelitian ilmiah terjadi dengan sangat pesat. Setiap tahunnya, jutaan artikel dan paper akademik dipublikasikan di berbagai jurnal dan konferensi. Dengan begitu banyaknya publikasi, para peneliti, mahasiswa, dan praktisi akademik sering kesulitan untuk mengikuti perkembangan terkini secara efektif. Dibutuhkan waktu yang tidak sedikit untuk memahami, memilah, dan menemukan kontribusi utama dari masing-masing paper, terlebih lagi dalam bidang yang kompleks dan terus berkembang.

Di sisi lain, kontribusi dari sebuah penelitian merupakan salah satu aspek paling penting dalam sebuah artikel ilmiah. Bagian ini menjelaskan nilai tambah atau keunikan dari penelitian tersebut dibandingkan dengan karya-karya terdahulu. Identifikasi kontribusi ini sangat berguna dalam menentukan apakah sebuah artikel memiliki relevansi atau keunikan yang dibutuhkan oleh pembaca. Bagi seorang peneliti yang akan merujuk atau melanjutkan penelitian yang sudah ada, memahami kontribusi dari karya ilmiah sebelumnya adalah langkah awal yang krusial. Namun, upaya untuk memahami kontribusi ini secara manual bisa sangat melelahkan, apalagi mengingat format penulisan kontribusi dalam setiap paper bisa sangat bervariasi tergantung pada jurnal atau konferensi yang menerbitkannya.

Pemrosesan Bahasa Alami atau Natural Language Processing (NLP) telah muncul sebagai solusi yang menjanjikan untuk membantu pemahaman dan penanganan informasi dalam teks. Teknik NLP sangat penting untuk mengekstraksi informasi yang bermakna dari label teks dalam model grafis. Ini sangat penting untuk tugas-tugas seperti mengidentifikasi kata benda dan frasa kata kerja, yang penting untuk membentuk hubungan dalam model. Kemampuan untuk mengekstrak frasa ini secara akurat dalam kondisi dunia nyata menunjukkan kekokohan metode NLP (Danenas & Skersys, 2022). Berbagai model NLP telah dikembangkan untuk menyelesaikan berbagai permasalahan dalam analisis teks, termasuk dalam merangkum teks, menjawab pertanyaan, hingga melakukan klasifikasi informasi. Salah satu model yang dimaksud adalah T5 (Text-To-Text Transfer Transformer), yang dikembangkan oleh Google Research. T5 adalah model berbasis transformer yang dapat menyelesaikan berbagai tugas NLP dengan pendekatan yang sangat fleksibel, yaitu dengan mengonversi setiap masalah menjadi format input-output berbasis teks. Model T5 ini telah terbukti efektif dalam tugas-tugas seperti summarization (meringkas teks) , machine translation (penerjemahan bahasa), question answering (menjawab pertanyaan), dan information extraction (ekstraksi informasi). T5 juga menunjukkan kinerjanya yang unggul untuk analisis subjektivitas dan klasifikasi opini, dengan peningkatan ACC dan skor F1 sebesar 0,6% untuk tugas analisis subjektivitas dan peningkatan ACC sebesar 0,4% dan skor F1 sebesar 0,6% untuk tugas klasifikasi opini, jika dibandingkan dengan model yang lain (Zhao et al., 2025). Studi terbaru di bidang Pemrosesan Bahasa Alami (NLP) telah menunjukkan bahwa arsitektur Text-To-Text Transfer Transformer (T5) dapat mencapai kinerja canggih untuk berbagai tugas NLP (Mastropaolo et al., 2021).

Selain T5, untuk membangun sistem pembacaan kontribusi dan menganalisis makna teks secara kompleks dalam penelitian ini jg menggunakan BERT (Bidirectional

Encoder Representations from Transformers). Model BERT dapat digunakan untuk menghasilkan contextual embeddings dari setiap kalimat dalam makalah jurnal. Contextual embeddings memungkinkan representasi semantik yang kaya dengan mempertimbangkan konteks setiap kata dalam kalimat (Samosir et al., 2022).

Penelitian ini bertujuan untuk menerapkan Algoritma T5 dan BERT dalam mengekstraksi kontribusi utama dari paper akademik secara otomatis. Dengan menggunakan model T5 dan BERT, diharapkan proses ini dapat dilakukan secara efektif dan efisien, sehingga dapat membantu para peneliti atau pembaca untuk dengan cepat memahami kontribusi dari sebuah artikel tanpa harus membaca seluruh isi artikel tersebut secara mendalam. Penelitian ini juga bertujuan untuk mengidentifikasi tingkat akurasi dan efisiensi dari model T5 dan BERT dalam melakukan tugas ini, serta mengeksplorasi bagaimana model ini dapat ditingkatkan untuk aplikasi lebih lanjut. Selain itu, penelitian ini memiliki implikasi penting dalam meningkatkan aksesibilitas informasi ilmiah. Dengan penerapan algoritma ini, diharapkan tercipta sebuah sistem otomatis yang dapat membantu mengelola informasi akademik dalam skala besar, terutama di tengah kebutuhan akan literatur ilmiah yang terus bertambah. Sebuah sistem yang mampu memberikan ringkasan kontribusi ilmiah secara otomatis dapat menjadi alat yang sangat berguna, baik dalam manajemen literatur penelitian maupun dalam pengembangan perpustakaan digital.

Pada akhirnya, penelitian ini tidak hanya akan berkontribusi pada pengembangan teknologi NLP di bidang akademis tetapi juga membuka peluang untuk integrasi lebih lanjut dari model NLP dalam sistem manajemen literatur ilmiah selain itu Kemajuan dalam teknologi ringkasan dapat diterapkan dalam berbagai skenario dunia nyata, termasuk penelitian akademis Ini secara signifikan dapat meningkatkan aksesibilitas informasi dan pengalaman pengguna di platform digital (Dilawari et al., 2023). Melalui penelitian ini, diharapkan dapat dihasilkan suatu pendekatan yang memungkinkan untuk memproses informasi ilmiah secara lebih cepat dan akurat, sehingga mendukung kemajuan penelitian di berbagai bidang.

1.2 Kajian Pustaka

1.2.1 Pengertian Paper

Paper adalah karya tulis ilmiah yang memuat hasil penelitian, kajian, atau eksperimen yang dilakukan oleh seorang peneliti atau akademisi. Dalam sebuah paper, penulis biasanya menjelaskan tujuan penelitian, metode yang digunakan, hasil yang ditemukan, serta analisis terhadap temuan tersebut. Paper ini disusun dengan sistematis, dimulai dari judul yang menggambarkan topik penelitian, dilanjutkan dengan abstrak yang merangkum inti dari isi paper, lalu diikuti dengan pendahuluan yang memberikan latar belakang serta tujuan penelitian.

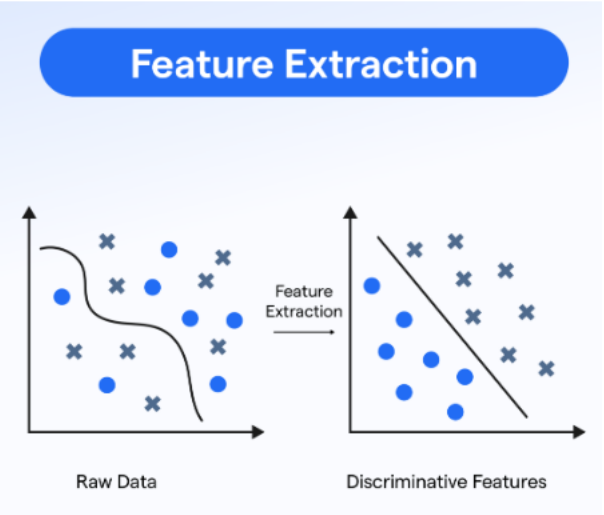
Bagian metodologi dalam paper menjelaskan secara rinci langkah-langkah yang diambil dalam penelitian, agar pembaca dapat memahami prosesnya. Selanjutnya, hasil penelitian disajikan dalam bentuk data atau temuan, yang kemudian dianalisis dan dibahas dalam bagian pembahasan. Bagian ini bertujuan untuk memberikan interpretasi dan makna dari data yang diperoleh. Paper diakhiri dengan kesimpulan yang

merangkum temuan utama, serta daftar referensi yang mencantumkan sumber-sumber yang mendukung penelitian.

Paper yang telah ditulis dengan baik biasanya dikirimkan ke jurnal ilmiah atau konferensi untuk dipublikasikan. Melalui publikasi ini, hasil penelitian tersebut dapat diakses oleh komunitas akademik dan berkontribusi pada perkembangan ilmu pengetahuan di bidang terkait.

1.2.2 Ektraksi Fitur

Ekstraksi fitur merupakan tahap yang penting pada sistem pengenalan karakter. Proses tersebut dilakukan untuk mendapatkan ciri khusus yang dimiliki dalam sistem pengenalan karakter.



Gambar 1. 1 Ekstraksi fitur

Gambar *Feature Extraction* menggambarkan bagaimana data mentah yang bentuknya masih acak dan sulit dipisahkan dapat diproses menjadi fitur yang lebih teratur sehingga mudah dikenali oleh model. Pada bagian kiri, data masih belum memiliki pola yang jelas. Setelah melalui proses ekstraksi fitur, data tersebut berubah menjadi representasi yang lebih rapi dan dapat dibedakan antar kelasnya, seperti terlihat pada bagian kanan gambar.

Dalam pemrosesan bahasa alami, tahap ini biasanya meliputi pemecahan teks menjadi token, pembersihan teks, serta pembentukan representasi vektor. Model seperti BERT mengekstraksi fitur berdasarkan konteks, sehingga setiap kata dipahami sesuai kalimatnya. Hasil ekstraksi inilah yang kemudian digunakan oleh model seperti T5 untuk menghasilkan ringkasan kontribusi yang lebih tepat. Dengan kata lain, ekstraksi fitur menjadi tahapan penting untuk membantu model memahami struktur dan isi teks secara lebih menyeluruh.

Abstrak:
"Penelitian ini mengusulkan sebuah algoritma pembelajaran mesin baru untuk meningkatkan akurasi prediksi penyakit pada data medis. Algoritma ini menggunakan pendekatan ensembel yang menggabungkan metode pohon keputusan dan regresi logistik. Hasil eksperimen menunjukkan peningkatan signifikan dibandingkan metode yang sudah ada."
Pendahuluan:
"Di era big data, analisis data medis menjadi semakin penting. Namun, masih terdapat tantangan dalam meningkatkan akurasi prediksi penyakit. Penelitian sebelumnya telah menggunakan berbagai metode seperti pohon keputusan, regresi logistik, dan jaringan saraf tiruan. Penelitian ini bertujuan untuk mengatasi tantangan tersebut dengan mengembangkan algoritma baru yang lebih akurat."
Kesimpulan:
"Hasil penelitian menunjukkan bahwa algoritma yang diusulkan berhasil meningkatkan akurasi prediksi hingga 15% dibandingkan metode sebelumnya. Pendekatan ensembel yang digunakan menjadi salah satu kontribusi utama dalam penelitian ini."

Fitur yang Dapat Diambil

1. Fitur Tekstual:
   Fitur yang berkaitan dengan karakteristik dasar teks, seperti panjang kata, pola huruf, atau urutan kata. Fitur ini tidak memperhatikan struktur tata bahasa atau makna.
   a. Frasa yang mengindikasikan kontribusi:
      "Penelitian ini mengusulkan sebuah algoritma pembelajaran mesin baru."
      "Pendekatan ensembel yang digunakan menjadi salah satu kontribusi utama."
   b. Kata kunci seperti "mengusulkan," "berhasil," "kontribusi," "baru," "meningkatkan."
1. Fitur Sintaksis:
   Fitur yang berkaitan dengan struktur tata bahasa, seperti peran kata (part-of-speech) atau hubungan antar kata dalam kalimat.
   a. Kalimat aktif dengan kata kerja seperti "mengusulkan," "berhasil," "menunjukkan."
   b. Contoh: *"Penelitian ini mengusulkan sebuah algoritma pembelajaran mesin baru."*
2. Fitur Semantik:
   Fitur yang menangkap makna dan konteks dari teks, sering menggunakan representasi seperti embeddings atau analisis entitas.
   Hubungan antara subjek, tindakan, dan objek:
   "Penelitian ini mengusulkan (tindakan) sebuah algoritma pembelajaran mesin baru (kontribusi)."
   "Hasil penelitian menunjukkan (tindakan) bahwa algoritma yang diusulkan berhasil meningkatkan akurasi (kontribusi dan hasil)."

3. Fitur Posisi:
   Fitur yang mempertimbangkan lokasi teks dalam dokumen, seperti apakah kalimat berada di awal, tengah, atau akhir.
   Frasa kontribusi utama sering ditemukan:
   a. Di awal atau akhir abstrak.
   b. Awal pendahuluan.
   c. Awal atau akhir kesimpulan.
4. Fitur Statistik:
   Fitur berbasis perhitungan kuantitatif yang menangkap distribusi kata atau frasa dalam teks. Frekuensi kemunculan kata kunci seperti "mengusulkan," "berhasil," "baru," "kontribusi."

1.2.3 Pengertian Kontribusi Secara Umum

Secara umum, kontribusi adalah bentuk sumbangan, peran, atau partisipasi seseorang atau suatu pihak terhadap pencapaian tujuan tertentu, baik dalam bentuk pemikiran, tenaga, tindakan, maupun hasil karya. Kontribusi dapat bersifat material (uang, barang) maupun non-material (ide, gagasan, waktu, keahlian). Dalam konteks sosial, kontribusi berarti keterlibatan individu dalam mendukung keberhasilan kelompok atau masyarakat. Sedangkan dalam konteks akademik, kontribusi merujuk pada nilai kebaruan (novelty), temuan, atau pengembangan ilmu pengetahuan yang diberikan oleh suatu penelitian.

Definisi Kontribusi Menurut Beberapa Sumber Teori

1. Kamus Besar Bahasa Indonesia (KBBI) Kontribusi diartikan sebagai *sumbangan*. Sumbangan ini dapat berupa uang, tenaga, pikiran, atau bentuk lainnya yang mendukung suatu kegiatan atau tujuan tertentu.
2. Soerjono Soekanto (Teori Peran dalam Sosiologi) Dalam teori peran, kontribusi berkaitan dengan pelaksanaan peran sosial individu sesuai dengan kedudukannya dalam masyarakat. Artinya, setiap individu memberikan kontribusi melalui peran yang dijalankannya.
3. Robert K. Merton (Teori Fungsionalisme Struktural) Merton menjelaskan bahwa setiap unsur dalam sistem sosial memiliki fungsi. Kontribusi dipahami sebagai fungsi atau dampak yang diberikan suatu tindakan atau struktur terhadap kestabilan dan perkembangan sistem.
4. John W. Creswell (Metodologi Penelitian) Dalam penelitian ilmiah, kontribusi diartikan sebagai bagian dari hasil penelitian yang memberikan tambahan pengetahuan baru, memperluas teori, atau memberikan solusi terhadap masalah tertentu.

Jenis-Jenis Kontribusi

1. Kontribusi Material : uang, barang, fasilitas.
2. Kontribusi Non-Material : ide, gagasan, tenaga, waktu.
3. Kontribusi Akademik : temuan baru, metode baru, pengembangan teori, aplikasi praktis.
4. Kontribusi Sosial dampak positif bagi masyarakat.

Kontribusi pada dasarnya adalah segala bentuk sumbangan atau peran yang memberikan dampak terhadap suatu tujuan atau sistem tertentu. Dalam bidang akademik, kontribusi berarti nilai tambah atau kebaruan ilmiah yang dihasilkan oleh suatu penelitian dibandingkan penelitian sebelumnya.

1.2.4 Kontribusi Paper

Kontribusi utama dari sebuah paper adalah menambah pengetahuan baru, wawasan, atau solusi terhadap suatu masalah dalam bidang tertentu. Paper yang diterbitkan memberikan kontribusi penting dalam mengembangkan ilmu pengetahuan, dengan menyajikan penemuan atau teori baru yang belum pernah dijelaskan sebelumnya. Paper ini juga sering bertujuan untuk menyelesaikan masalah spesifik, baik dengan menawarkan teknologi baru, strategi, atau metode yang dapat diterapkan dalam dunia nyata. Selain itu, sebuah paper berfungsi sebagai dasar penelitian lanjutan. Hasil dan data yang dipublikasikan dapat menjadi rujukan bagi peneliti lain yang ingin mengembangkan atau menguji lebih lanjut temuan yang ada. Dalam banyak kasus, hasil penelitian di dalam paper bisa diterapkan dalam industri atau bidang profesional, sehingga meningkatkan praktik dan teknologi yang ada.

Paper juga dapat mempengaruhi pembuatan kebijakan, terutama dalam bidang kesehatan, lingkungan, atau pendidikan. Berdasarkan data ilmiah yang ada, pemerintah atau organisasi dapat menyusun kebijakan yang lebih informatif dan berbasis bukti. Selain itu, paper juga membantu memecahkan pertanyaan ilmiah atau menguji hipotesis tertentu, sehingga memberikan bukti atau jawaban yang membantu menjelaskan masalah yang sedang dikaji. Secara keseluruhan, kontribusi sebuah paper sangat luas, tidak hanya dalam pengembangan ilmu pengetahuan dan akademik, tetapi juga memiliki dampak praktis pada masyarakat, teknologi, industri, dan kebijakan di berbagai bidang. Maka dari itu melakukan Peringkasan dokumen adalah salah satu solusi yang dapat dilaksanakan untuk menemukan kontribusi dalam paper pada yang sedang dikaji proses mengekstraksi teks dari suatu dokumen, mengeksplorasi dan menyajikan informasi penting kepada pengguna atau aplikasi dalam bentuk ringkasan yang singkat dan ringkas. Ketika kita dihadapkan pada struktur bahasa yang cukup kompleks, seperti dalam bahasa manusia, maka itu menangkap gagasan utama dan makna dari teks aslinya. Di sinilah model Transformer digunakan, yang memiliki model berkinerja tinggi dan (Purnama & Widya Utami, 2023). Penelitian NLPContributions menetapkan struktur kontribusi paper berdasarkan hasil analisis empiris melalui anotasi manual terhadap artikel ilmiah di bidang Natural Language Processing berbasis Machine Learning. Penentuan struktur ini dilakukan dengan mengidentifikasi kalimat-kalimat yang secara eksplisit merepresentasikan kontribusi penelitian, yang umumnya ditemukan pada bagian judul, abstrak, pendahuluan, hasil, dan kesimpulan. Selanjutnya, kalimat-kalimat kontribusi tersebut dikelompokkan ke dalam unit informasi semantik yang konsisten dan berulang antar-paper, sehingga membentuk sepuluh unit kontribusi inti seperti *ResearchProblem*, *Approach*, dan *Results*. Pendekatan ini bersifat bottom-up, yaitu struktur kontribusi tidak ditentukan secara teoritis, melainkan diturunkan langsung dari pola kontribusi yang ditemukan pada artikel ilmiah nyata (D'Souza, 2020).

1.2.5 Algorithma T5 (Text-To-Text Transformer)

Text-To-Text Transfer Transformer (T5) adalah sebuah algoritma dalam bidang pemrosesan bahasa alami (Natural Language Processing, NLP) yang mengubah semua tugas berbasis teks menjadi format teks-ke-teks. Ini berarti bahwa baik input maupun output dari model selalu berupa teks, terlepas dari jenis tugas yang sedang dikerjakan, seperti klasifikasi, terjemahan, atau ringkasan. Pendekatan ini memberikan fleksibilitas yang lebih besar dibandingkan dengan model-model sebelumnya yang biasanya dirancang untuk tugas tertentu saja (Purnama & Widya Utami, 2023)**.** T5 adalah model bahasa berbasis transformer yang dirancang untuk tugas-tugas bahasa alami dalam format teks-ke-teks, yang semakin populer karena kerangka kerjanya yang terpadu, mengubah semua masalah berbasis teks menjadi format teks-ke-teks. Model ini juga mudah untuk diskalakan dengan jumlah parameter yang bervariasi, mulai dari 60 juta hingga 11 miliar parameter, melalui metode pemrosesan paralel pada model (Guo et al., 2022).

Model T5, atau Text-to-Text Transfer Transformer, adalah model bahasa berbasis transformer yang dikembangkan oleh Raffel et al. (2019) dengan tujuan untuk menyatukan berbagai tugas pemrosesan bahasa alami (Natural Language Processing/NLP) dalam satu kerangka kerja yang konsisten, yakni format "teks-ke-teks". Dalam kerangka ini, setiap tugas NLP, seperti penerjemahan, klasifikasi, dan peringkasan, dikonversi ke format input-output berbasis teks. Misalnya, dalam kasus tugas klasifikasi, model T5 menerima teks sebagai input dan menghasilkan teks berupa kategori sebagai output. T5 menggunakan arsitektur transformer berbasis self-attention penuh dengan encoder-decoder, sehingga memungkinkan model ini memproses data sekuensial dengan efisien. Keuntungan utama T5 adalah fleksibilitas dalam menangani berbagai tugas NLP hanya dengan perubahan pada format teks input. Hal ini juga memungkinkan model T5 untuk memanfaatkan pendekatan *transfer learning* atau pembelajaran transfer, yang membuatnya sangat efektif pada berbagai tugas dengan melakukan fine-tuning pada data spesifik (Alexandria, 2023)**.** Model T5 didasarkan pada arsitektur Transformer, yang terdiri dari dua komponen utama: encoder dan decoder. Enkoder memproses teks input untuk membuat representasi keadaan tersembunyi, sementara decoder menghasilkan teks keluaran dari representasi ini (Balakrishnan et al., 2024). Transformer Transfer Teks-ke-Teks (T5) membingkai ulang beragam tugas pemrosesan bahasa alami sebagai masalah masuk/keluar teks terpadu: model selalu menerima teks sebagai input dan mengembalikan teks sebagai output, baik yang diterapkan pada klasifikasi, terjemahan mesin, ringkasan, atau tugas serupa. Formulasi terpadu ini menawarkan lebih banyak fleksibilitas daripada pendekatan sebelumnya yang direkayasa untuk tujuan khusus tugas tunggal (Saptadi, 2025). T5 adalah model bahasa berbasis transformator yang dirancang untuk tugas bahasa alami dalam format teks-ke-teks, yang semakin populer karena kerangka kerjanya yang terpadu, mengubah semua masalah berbasis teks menjadi format teks-ke-teks. Model ini juga mudah

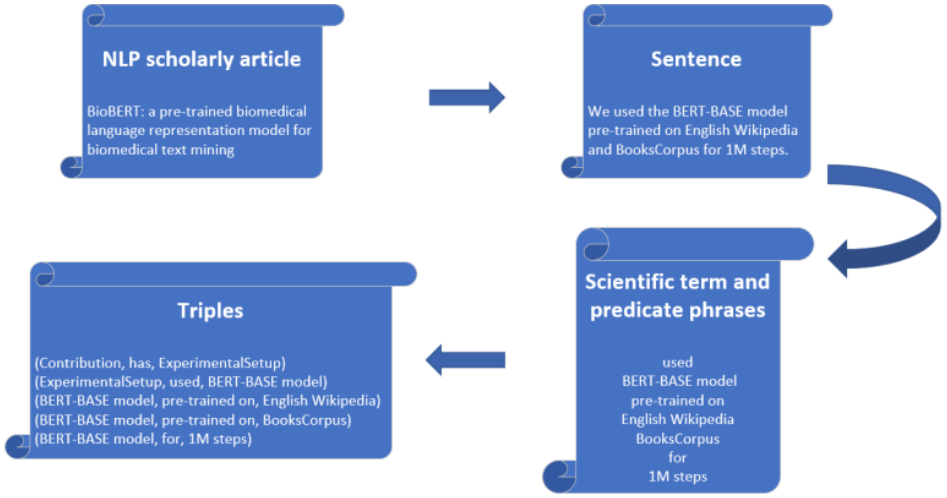
diskalakan dengan jumlah parameter yang bervariasi, mulai dari 60 juta hingga 11 miliar parameter, melalui metode pemrosesan paralel pada model (Guo, 2022).

1.2.6 Algoritma BERT (Bidirectional Encoder Representation from Transformers)

BERT (Bidirectional Encoder Representations from Transformers**)** adalah model deep learning berbasis transformer yang dikembangkan oleh Google. Model ini menjadi tonggak baru dalam bidang Natural Language Processing (NLP) karena mampu memahami konteks suatu kata dalam sebuah kalimat secara bidirectional**,** yaitu dengan membaca teks dari kedua arah (kiri ke kanan dan kanan ke kiri). Hal ini memberikan BERT kemampuan untuk memahami makna kata dalam konteks yang lebih baik dibandingkan pendekatan tradisional yang hanya bersifat unidirectional (Ma et al., 2021). Berikut rincian fitur dan fungsionalitas utamanya:

1. Konteks Dua Arah: Tidak seperti model tradisional yang membaca teks secara berurutan (kiri-ke-kanan atau kanan-ke-kiri), BERT memproses teks di kedua arah secara bersamaan. Hal ini memungkinkannya untuk memahami konteks kata berdasarkan semua kata di sekitarnya, yang mengarah pada pemahaman yang lebih baik tentang makna dalam kalimat.
2. Arsitektur Transformer: BERT dibangun di atas arsitektur transformator, yang menggunakan mekanisme perhatian diri untuk menimbang pentingnya kata-kata yang berbeda dalam sebuah kalimat. Hal ini memungkinkan model untuk menangkap hubungan antara kata-kata terlepas dari posisinya dalam teks.

Salah satu keuntungan menggunakan BERT adalah kemampuannya untuk menyoroti fitur-fitur penting dalam data input, yang membantu dalam memahami proses pengambilan keputusan model. Ini sangat berguna dalam aplikasi seperti klasifikasi berita palsu, di mana mengidentifikasi isyarat linguistik dapat membantu dalam mengenali informasi yang salah (Oad et al., 2024). Berikut adalah proses yang menggambarkan proses ekstraksi informasi dari artikel ilmiah dalam konteks pemrosesan bahasa alami (NLP).



Gambar 1. 2 Arsitektur BERT

Berikut adalah penjelasan setiap elemen dalam gambar tersebut:

1. NLP Scholarly Article:
   Ini adalah sumber utama, yaitu artikel ilmiah yang berkaitan dengan NLP. Artikel ini berisi informasi dan data yang relevan yang akan diproses untuk ekstraksi informasi.
2. Sentence:
   Dari artikel ilmiah tersebut, kalimat-kalimat diambil. Model yang digunakan untuk ekstraksi kalimat ini adalah BERT-BASE**,** yang telah dilatih sebelumnya pada dataset dari Wikipedia Bahasa Inggris dan BooksCorpus. Ini menunjukkan bahwa model ini telah mendapatkan pemahaman yang baik tentang bahasa dan konteks sebelum digunakan pada artikel ilmiah.
3. Triples:
   a. Setelah kalimat diambil, informasi lebih lanjut diekstraksi dalam bentuk triplet. Triplet ini biasanya terdiri dari subjek, predikat, dan objek. Misalnya, dalam konteks ilmiah, triplet dapat berupa informasi yang menjelaskan hubungan antara entitas yang berbeda dalam artikel.
   b. Di sini**,** ExperimentalSetup**,** Contribution**,** dan BERT-BASE model adalah komponen yang menandakan informasi spesifik yang diekstrak dari kalimat.
4. Scientific Term and Predicate Phrases:
   Langkah terakhir menunjukkan bahwa dari kalimat dan triplet yang telah diekstraksi, frasa ilmiah dan istilah yang relevan juga diidentifikasi. Ini membantu dalam memahami konteks dan makna yang lebih dalam dari artikel ilmiah yang sedang dianalisis.

1.3 Rumusan Masalah

1. Bagaimana algoritma T5 dan BERT dapat dilatih untuk mengenali dan mengekstraksi kontribusi utama dalam berbagai jenis paper akademik?
2. Bagaimana tingkat akurasi dan performa model T5 dan BERT dalam mengekstraksi kontribusi paper akademik ?

1.4 Tujuan Penelitian

1. Mengembangkan model berbasis algoritma T5 dan BERT yang mampu mengekstraksi kontribusi utama dari paper akademik secara otomatis
2. Menganalisis kinerja model T5 dan BERT+T5 dalam menghasilkan ekstraksi kontribusi berdasarkan hasil evaluasi ROUGE dan BERTScore.
3. Menguji pengaruh integrasi BERT dengan T5 terhadap peningkatan kualitas dan keakuratan hasil ekstraksi dibandingkan penggunaan T5 saja.

1.5 Manfaat Penelitian

1. Mempermudah Peneliti dalam Mengidentifikasi Kontribusi Utama
2. Meningkatkan Efisiensi Manajemen Literatur Ilmiah

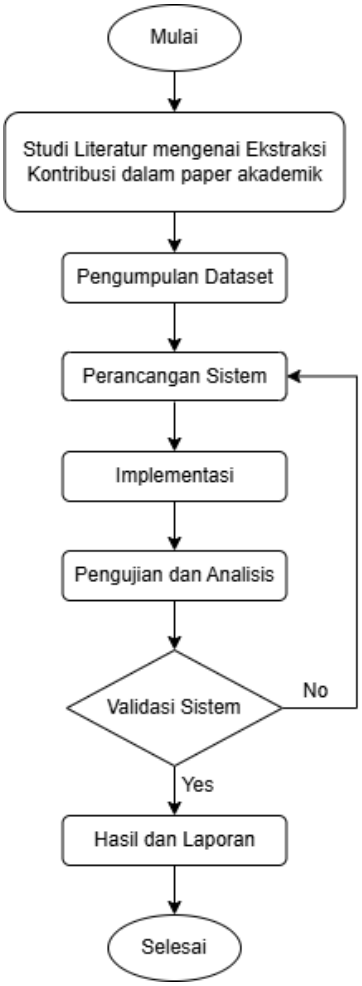
1.6 Batasan Masalah

1. Penelitian ini menggunakan paper akademik yang teksnya diambil dari bagian abstrak, pendahuluan, dan kesimpulan .
2. Paper yang digunakan berbahasa Inggris dan pada topik tertentu, seperti ilmu komputer atau teknik.
3. Penelitian hanya fokus pada mengambil kontribusi utama dari paper, seperti hasil utama atau inovasi, tanpa tugas lain.

BAB II METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini adalah penelitian berjenis eksperimental. Data yang dikumpulkan didapatkan melalui studi literatur dari berbagai jurnal dan artikel ilmiah, sehingga dapat ditemukan cakupan masalah dalam penelitian yang dilakukan. Berikut ini adalah tahapan penelitian yang dilakukan :



Gambar 2. 1 Tahapan Penelitian

2.1.1 Studi Literatur

Tahapan awal yang dilakukan adalah studi literatur. Tahapan studi literatur merupakan tahapan pengumpulan informasi dan refensi- referensi yang terkait dengan penelitian yang akan dilakukan yaitu Ekstraksi Kontribusi Dalam Paper Akademik. Pada tahap ini juga peneliti menentukan metode yang akan digunakan yaitu Metode T5 (Text-

To-Text Transfer Transformer) dan BERT (Bidirectional Encoder Representations from Transformers).

2.1.2 Pengumpulan Data

Dataset pada sistem dikumpulkan dari artikel akademik bidang *computer science* yang diperoleh melalui repositori publik (seperti arXiv dan jurnal *open access*). Teks dari setiap artikel diekstraksi, kemudian dipisah per kalimat dan dianotasi secara manual untuk menentukan kalimat yang mengandung kontribusi penelitian. Anotasi dibagi ke dalam tiga kategori kontribusi, yaitu *Problem Statement*, *Method/Approach*, dan *Main Findings/Results*. Dataset akhir terdiri dari 500 artikel akademik (± 8.500 kalimat dan 210.000 token), kemudian dibagi menjadi 70% data pelatihan, 15% validasi, dan 15% pengujian untuk keperluan pelatihan dan evaluasi model T5 serta BERT+T5.

2.1.3 Perancangan Sistem

Pada tahap perancangan sistem, ditentukan alur pemrosesan teks untuk ekstraksi kontribusi mulai dari *preprocessing*, pembentukan input model, hingga keluaran berupa kalimat kontribusi. Selain itu, ditetapkan komponen inti yang digunakan dalam sistem, yaitu model T5 sebagai *baseline* dan kombinasi BERT + T5 sebagai model hibrida, sehingga proses implementasi dapat berjalan terarah.

2.1.4 Implementasi

Tahap implementasi adalah proses membangun sistem sesuai dengan rancangan sebelumnya. Pada tahap ini, model T5 digunakan terlebih dahulu sebagai model dasar untuk mengekstraksi kalimat kontribusi secara langsung dari teks. Setelah itu, digunakan model gabungan BERT + T5 sebagai alternatif yang diharapkan memberi hasil lebih baik. Pada pendekatan ini, BERT berfungsi memahami makna dan konteks kalimat, lalu hasilnya diteruskan ke T5 agar T5 dapat menghasilkan kalimat kontribusi dengan lebih akurat. Seluruh proses implementasi dilakukan menggunakan *framework* PyTorch dan HuggingFace Transformers. PyTorch digunakan sebagai platform pelatihan model, sedangkan HuggingFace Transformers dimanfaatkan untuk memanggil dan mengelola model T5 dan BERT secara efisien.

2.1.5 Pengujian dan analisis

Setelah model selesai dibangun, dilakukan pengujian dan analisis**.** Model dievaluasi menggunakan data pengujian dengan memanfaatkan metrik berbasis kesamaan teks (ROUGE) dan kesamaan semantik (BERTScore). Analisis dilakukan untuk melihat tingkat ketepatan model dalam mengenali kontribusi, membandingkan kinerja kedua pendekatan (T5 saja dan BERT + T5), serta menilai konsistensi hasil pada berbagai skenario pengujian.

2.1.6 Validasi Sistem

Tahap berikutnya adalah validasi sistem, yang bertujuan menilai apakah performa yang dicapai telah memenuhi target penelitian. Jika kinerja model masih berada di bawah batas yang ditetapkan, dilakukan perbaikan dengan mengulang tahapan perancangan,

implementasi, atau parameter pelatihan hingga menghasilkan performa yang memadai. Bila sistem telah divalidasi, penelitian dapat dilanjutkan ke tahap akhir.

2.1.7 Penyusunan Hasil dan pelaporan

Penelitian ditutup dengan penyusunan hasil dan pelaporan, yang menguraikan keseluruhan proses mulai dari latar belakang hingga kesimpulan. Bagian ini juga memuat pembahasan performa sistem, manfaat penelitian, serta masukan untuk penelitian lanjutan.

2.2 Tempat dan waktu penelitian

Penelitian ini dimulai pada bulan Februari – November 2025. Penelitian dilaksanakan di Laboratorium Computer Based System (CBS) , Departemen Teknik Informatika Universitas Hasanuddin, Jalan Poros Malino, Kelurahan Borongloe, Kecamatan Bontomarannu, Kabupaten Gowa, Provinsi Sulawesi Selatan.

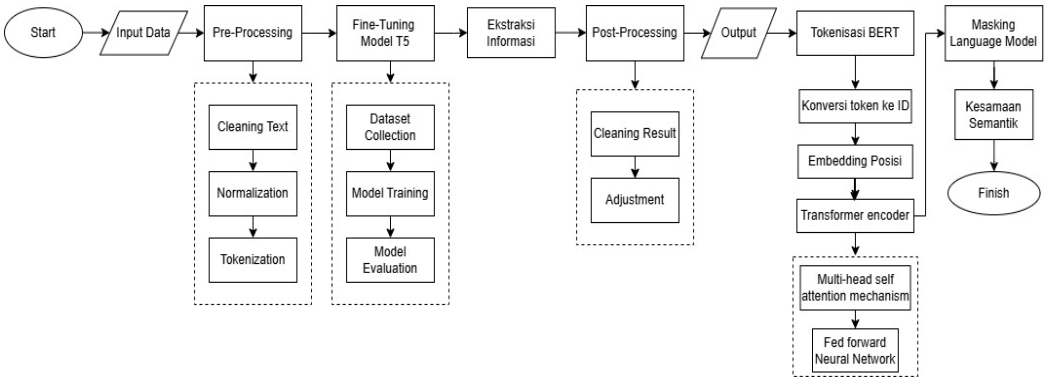
2.3 Perangkat Penelitian

Adapun perangkat-perangkat yang digunakan pada penelitian ini adalah sebagai berikut :

1. Perangkat Keras:
   a. Laptop
   b. Smartphone
2. Perangkat Lunak :
   a. Sistem Operasi Windows 11-64 bit
   b. Visual Studio code
   c. Microsoft Word

2.4 Rancangan Sistem

Rancangan Sistem Alur perancangan sistem pada penelitian ini ditunjukkan pada gambar 3.2 berikut:



Gambar 2. 2 Alur perancangan system

2.4.1 Input Data

Data yang di input berupa jurnal yang di upload, sistem akan membaca teks yang termasuk Abstrak, Pendahuluan, dan Kesimpulan. Pada tahap ini adalah proses untuk menyiapkan data yang akan di uji.

Contoh Teks :

"Penelitian ini mengusulkan penerapan algoritma pembelajaran mesin untuk meningkatkan akurasi prediksi dalam diagnosa penyakit jantung. Pendekatan yang digunakan adalah dengan memanfaatkan data historis pasien dan membangun model klasifikasi untuk memprediksi kemungkinan serangan jantung. Hasil eksperimen menunjukkan bahwa algoritma X memiliki performa yang lebih baik dibandingkan algoritma tradisional yang digunakan dalam penelitian sebelumnya."

2.4.2 Preprocessing

a. Cleaning : Membersihkan teks dari simbol atau format yang tidak relevan (stopwords).

   Proses ini bertujuan untuk membersihkan teks dari elemen yang tidak relevan atau tidak diperlukan untuk analisis lebih lanjut.

   Contoh elemen yang dibersihkan:

   1. Simbol: Seperti tanda baca (., !, ?), tanda khusus (#, $, %, @).
   2. Format tidak relevan: URL, angka, atau kode HTML (seperti atau).
   3. Stopwords: Kata-kata umum seperti *dan*, *atau*, *tetapi* yang biasanya tidak memiliki nilai informasi tinggi dalam analisis teks.
   4. Tujuan: Mengurangi "noise" dalam data agar lebih relevan dan mudah diproses.

      Contoh Teks Asli :

      "Penelitian ini mengusulkan penerapan algoritma pembelajaran mesin untuk meningkatkan akurasi prediksi dalam diagnosa penyakit jantung. Pendekatan yang digunakan adalah dengan memanfaatkan data historis pasien dan membangun model klasifikasi untuk memprediksi kemungkinan serangan jantung. Hasil eksperimen menunjukkan bahwa algoritma X memiliki performa yang lebih baik dibandingkan algoritma tradisional yang digunakan dalam penelitian sebelumnya."

      Teks setelah Cleaning:

      "Penelitian mengusulkan penerapan algoritma pembelajaran mesin meningkatkan akurasi prediksi diagnosa penyakit jantung Pendekatan digunakan memanfaatkan data historis pasien membangun model klasifikasi memprediksi kemungkinan serangan jantung Hasil eksperimen menunjukkan algoritma X performa lebih baik algoritma tradisional digunakan penelitian sebelumnya"

b. Normalization : Mengubah semua teks menjadi huruf kecil

   Proses ini digunakan untuk menyamakan format teks agar konsisten. Salah satu langkah yang umum dilakukan adalah mengubah seluruh teks menjadi huruf kecil (lowercase).

Tujuan: Menghindari perbedaan penanganan antara kata yang sama hanya karena perbedaan kapitalisasi. Misalnya, "Data", "data", dan "DATA" diperlakukan sebagai satu kata yang sama.

Teks setelah Normalisasi:

"penelitian mengusulkan penerapan algoritma pembelajaran mesin meningkatkan akurasi prediksi diagnosa penyakit jantung pendekatan digunakan memanfaatkan data historis pasien membangun model klasifikasi memprediksi kemungkinan serangan jantung hasil eksperimen menunjukkan algoritma x performa lebih baik algoritma tradisional digunakan penelitian sebelumnya"

c. Tokenisasi :

Proses memecah teks menjadi unit kecil, yang disebut *token*. *Token* dapat berupa kata, kalimat, atau paragraf, tergantung pada kebutuhan analisis.

Tujuan: Mempermudah analisis dengan memisahkan teks menjadi elemen yang dapat diproses oleh algoritma.

Tokenisasi kata: Setiap kata menjadi satu token.

Tokenisasi kalimat: Teks dipecah berdasarkan tanda titik (.), tanda seru (!), atau tanda tanya (?).

Tokenisasi per Kata:

["penelitian", "mengusulkan", "penerapan", "algoritma", "pembelajaran", "mesin", "meningkatkan", "akurasi", "prediksi", "diagnosa", "penyakit", "jantung", "pendekatan", "digunakan", "memanfaatkan", "data", "historis", "pasien", "membangun", "model", "klasifikasi", "memprediksi", "kemungkinan", "serangan", "jantung", "hasil", "eksperimen", "menunjukkan", "algoritma", "x", "performa", "lebih", "baik", "algoritma", "tradisional", "digunakan", "penelitian", "sebelumnya"]

Tokenisasi per Kalimat:

["penelitian mengusulkan penerapan algoritma pembelajaran mesin meningkatkan akurasi prediksi diagnosa penyakit jantung", "pendekatan digunakan memanfaatkan data historis pasien membangun model klasifikasi memprediksi kemungkinan serangan jantung", "hasil eksperimen menunjukkan algoritma x performa lebih baik algoritma tradisional digunakan penelitian sebelumnya"]

2.4.3 Fine-Tuning Model T5

Pada tahap ini melatih model T5 untuk mengenali pola kontribusi dalam jurnal dengan cara :

Dataset Collection : Menggunakan dataset yang sudah ada dari jurnal yang sudah dianotasi pada bagiam kontribusi , misalnya :

a. "Penelitian ini mengusulkan penerapan algoritma pembelajaran mesin untuk meningkatkan akurasi prediksi dalam diagnosa penyakit jantung."
b. "Hasil eksperimen menunjukkan bahwa algoritma X lebih baik dibandingkan dengan algoritma Y."

Training Model : Melakukan pelatihan dataset mnendapatkan pola kontribusi

Evaluation : Melakukan pengukuran / perangkingan

2.4.4 Ekstraksi Informasi

Pada tahap ini inputan akan diuji pada model T5 setelah ditunning, hasilnya ada jumlah kalimat yang sudah diatur sebagai output dengan nilai yang paling relevan dengan "kontribusi"

Kalimat Kontribusi yang Ditemukan:

"Penelitian ini mengusulkan penerapan algoritma pembelajaran mesin untuk meningkatkan akurasi prediksi dalam diagnosa penyakit jantung."

"Hasil eksperimen menunjukkan bahwa algoritma X memiliki performa yang lebih baik dibandingkan algoritma tradisional yang digunakan dalam penelitian sebelumnya."

2.4.5 Postprocessing

Cleaning : Menghapus atau mengurangi teks yang tidak relevan dari hasil ekstraksi

Teks Setelah Postprocessing Cleaning:

"Penelitian ini mengusulkan penerapan algoritma pembelajaran mesin untuk meningkatkan akurasi prediksi diagnosa penyakit jantung."
"Hasil eksperimen menunjukkan bahwa algoritma X memiliki performa yang lebih baik dibandingkan algoritma tradisional."

Adjustment : Menyusun result menjadi sebuah kalimat atau paragraf yang utuh

2.4.6 Output :

Menampilkan output berupa ringkasan dari kontribusi jurnal

Ringkasan Kontribusi Jurnal:

"Penelitian ini mengusulkan penerapan algoritma pembelajaran mesin untuk meningkatkan akurasi prediksi diagnosa penyakit jantung. Hasil eksperimen menunjukkan bahwa algoritma X memiliki performa yang lebih baik dibandingkan algoritma tradisional."

2.4.7 Tokenisasi BERT :

Teks hasil dari T5 mejadi array kata, setiap kata di kembalikan ke kata asalnya (tanpa imbuhan), akan ditambahkan teks awal [CLS] dan akhiran [SEP] sebagai penanda range token kata

Contoh :

Teks : "Penelitian ini mengusulkan penerapan algoritma pembelajaran mesin untuk meningkatkan akurasi prediksi diagnosa penyakit jantung. Hasil eksperimen menunjukkan bahwa algoritma X memiliki performa yang lebih baik dibandingkan algoritma tradisional."

Tokens: ["[CLS]", "penelitian", "ini", "mengusulkan", "penerapan", "algoritma", "pembelajaran", "mesin", "untuk", "meningkatkan", "akurasi", "prediksi", "diagnosa", "penyakit", "jantung", ".", "hasil", "eksperimen", "menunjukkan", "bahwa", "algoritma", "x", "memiliki", "performa", "yang", "lebih", "baik", "dibandingkan", "algoritma", "tradisional", ".", "[SEP]"]

1. Konversi Token ke ID

   Setiap kata di konversi ke token dari pustaka BERT (berisi 30.000 token unik)

Contoh :

Token IDs: [2, 24744, 83, 1767, 13658, 14332, 21901, 2469, 129, 10263, 1423, 21419, 24945, 1656, 16807, 19, 671, 19839, 2922, 159, 14332, 95, 469, 13794, 40, 186, 296, 13569, 14332, 14966, 19, 3]

2. Penambahan Embedding Posisi

Setiap kata akan diurutkan sesuai posisi array nya dimana semua dimulai dari angka 0 dimana [CLS] masuk di posisi 0, “penelitian” masuk posisi 1 dan seterusnya.

3. Proses Transformer Encoder

Multi-Head Self-Attention Mechanism

Proses ini untuk menilai keterkaitan kata dari setiap token menggunakan rumus Attention :

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Feed-Forward Neural Network

Proses ini untuk proses non-linear agar menghasilkan representasi kontekstual berdasarkan fitur yang digunakan :

a. Proses Masking Language Model

Selama pretraining, 15% token diganti dengan [MASK].

Contoh :

"Penelitian ini mengusulkan penerapan [MASK1] pembelajaran mesin untuk meningkatkan akurasi [MASK2] diagnosa penyakit jantung. Hasil eksperimen menunjukkan bahwa :

[MASK3] X memiliki performa yang lebih baik dibandingkan [MASK4] tradisional."

BERT akan memprediksi kata yang hilang:

Prediksi pertama [MASK1]: "metode". -> sebelumnya di T5 “algoritma”

Prediksi kedua [MASK2]: "klasifikasi". -> sebelumnya di T5 “akurasi”

Prediksi ketiga [MASK3]: "model". -> sebelumnya di T5 “algoritma”

Prediksi keempa [MASK4]: "sistem". -> sebelumnya di T5 “algoritma” jika semua prediksi pada mask lebih baik maka bert akan menggunakan prediksi sebagai kata pelengkap dari kalimat tersebut

b. Kesamaan Semantik

Pada proses ini input dan output akan dibanding kan dengan aturan 0.1 tidak relevan dan 0.9 relevan

Hasil t5 :

"Penelitian ini mengusulkan penerapan algoritma pembelajaran mesin untuk meningkatkan akurasi prediksi diagnosa penyakit jantung. Hasil eksperimen menunjukkan bahwa algoritma X memiliki performa yang lebih baik dibandingkan algoritma tradisional."

Hasil Bert :

"Penelitian ini mengusulkan penerapan [metode] pembelajaran mesin untuk meningkatkan akurasi [klasifikasi] diagnosa penyakit jantung. Hasil eksperimen menunjukkan bahwa [model] X memiliki performa yang lebih baik dibandingkan [sistem] tradisional."

BERT akan memprediksi kata yang hilang:
Prediksi pertama **[MASK1]**: "metode". -> sebelumnya di T5 “algoritma” (relevan 0.9)
Prediksi kedua **[MASK2]**: "klasifikasi". -> sebelumnya di T5 “akurasi” (konteks sama 0.9)
Prediksi ketiga **[MASK3]**: "model". -> sebelumnya di T5 “algoritma” (relevan 0.9)
Prediksi keempa **[MASK4]**: "sistem". -> sebelumnya di T5 “algoritma” (sesuai konteks 0.9)
Jika hasil semantik prediksi masking dari BERT mendekati/relevan (0.9) maka output akan di gunakan, sedangkan jika hasil dari semantik tidak relevan (0.1), maka kalimat hasil dari T5 digunakan (dalam artian bahwa hasil T5 sudah baik)
Kesimpulan :
T5 : digunakan untuk generate kontribusi
Bert : digunakan untuk validasi hasil kontribusi

2.5 Evaluasi Sistem

Untuk melihat hasil kinerja dari sistem Ekstraksi Kontribusi Dalam Paper akademik yang dirancang pada masing-masing skenario maka digunakan teknik evaluasi confusion matrix yang menggunakan 4 variabel yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Skenario yang memiliki akurasi tertinggi merupakan skenario dengan algoritma terbaik untuk melakukan Ekstraksi Kontribusi Dalam Paper akademik. Penelitian ini melakukan pengujian akurasi untuk Ekstraksi Kontribusi Dalam Paper Akademik dan divalidasi menggunakan nilai akurasi, precision, recall. Pada pengujian Ekstraksi Kontribusi Dalam Paper Akademik, variabel pada confusion matrix didefinisikan sebagai berikut:

1. TP adalah hasil pendeteksian Kontribusi Dalam Paper akademik yang berhasil terdeteksi;
2. FN adalah sejumlah objek Kontribusi Dalam Paper akademik yang tidak berhasil terdeteksi;
3. FP adalah sebuah kondisi dimana sistem keliru dalam melakukan pendeteksian Kontribusi Dalam Paper akademik;
4. TN adalah ketika sistem tidak mendeteksi objek Kontribusi Dalam Paper akademik yang merupakan sampel negatif.

Penentuan tabel confusion matrix dapat dilihat pada gambar 3.9 di dawah:

		Predicted	
		True	False
Actual	True	*TP*	*FN*
	False	*FN*	*TN*

Gambar 2. 3 Evaluasi Sistem