

**PERINGKASAN TEKS EKSTRAKTIF PADA ARTIKEL BERITA
BAHASA INDONESIA MENGGUNAKAN METODE
BERT DAN *COSINE SIMILARITY***

SKRIPSI

**Oleh :
NURMA ROMIHIM FADLILAH
NIM. 18650017**



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025**

**PERINGKASAN TEKS EKSTRAKTIF PADA ARTIKEL BERITA
BAHASA INDONESIA MENGGUNAKAN METODE
BERT DAN *COSINE SIMILARITY***

SKRIPSI

**Diajukan kepada :
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)**

**Oleh :
NURMA ROMIHIM FADLILAH
NIM. 18650017**

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025**

HALAMAN PERSETUJUAN

**PERINGKASAN TEKS EKSTRAKTIF PADA ARTIKEL BERITA
BAHASA INDONESIA MENGGUNAKAN METODE
BERT DAN COSINE SIMILARITY**

SKRIPSI

Oleh :
NURMA ROMIHIM FADLILAH
NIM. 18650017

Telah Diperiksa dan Disetujui untuk Diuji :
Tanggal : 17 Juni 2025

Pembimbing I

Pembimbing II



Dr. Zainal Abidin, M. Kom
NIP. 19760613 200501 1 004



Dr. M. Imamudin, Lc, MA
19740602 200901 1 010

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Fachrul Kurniawan, S.T., M.MT., IPU
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

PERINGKASAN TEKS EKSTRAKTIF PADA ARTIKEL BERITA BAHASA INDONESIA MENGGUNAKAN METODE *BERT* DAN *COSINE SIMILARITY*

SKRIPSI

Oleh :
NURMA ROMIHIM FADLILAH
NIM. 18650017

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 23 Juni 2025

Susunan Dewan Penguji

Ketua Penguji : Dr. Ir. Fachrul Kurniawan, S.T., M.MT., IPU
NIP. 19771020 200912 1 001

Anggota Penguji I : Shoffin Nahwa Utama, M.T
NIP. 19860703 202012 1 003

Anggota Penguji II : Dr. Zainal Abidin, M. Kom
NIP. 19760613 200501 1 0004

Anggota Penguji III : Dr. M. Imamudin, Lc, MA
19740602 200901 1 010

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi

Universitas Islam Negeri Maulana Malik Ibrahim Malang



Dr. Ir. Fachrul Kurniawan, S.T., M.MT., IPU
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya bertanda tangan dibawah ini :

Nama : Nurma Romihim Fadlilah
NIM : 18650017
Fakultas / Program Studi : Sains dan Teknologi / Teknik Informatika
Judul Skripsi : Peringkasan Teks Ekstraktif Pada Artikel Berita Bahasa Indonesia Menggunakan Metode *BERT* dan *Cosine Similarity*

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar- benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 17 Juni 2025

Yang membuat pernyataan,



Nurma Romihim Fadlilah
NIM. 18650017

HALAMAN MOTTO

“Setiap tetes keringat kedua orang tuaku, adalah jutaan langkahku untuk selalu bangkit dan maju”

(Penulis)

فَإِنَّ مَعَ الْعُسْرِ يُسْرًا ﴿٥﴾ إِنَّ مَعَ الْعُسْرِ يُسْرًا ﴿٦﴾

“Maka sesungguhnya beserta kesulitan ada kemudahan, Sesungguhnya beserta kesulitan ada kemudahan”

(QS. Al-Insyirah : 5-6)

HALAMAN PERSEMBAHAN

Puji Syukur atas kehadiran Allah Subhanahu wa ta'ala, karena berkat rahmat dan petunjuk-Nya penulis dapat menyelesaikan skripsi ini dengan keadaan sehat wal 'afiyat.

Penulis mempersembahkan karya ini kepada kedua orang tua penulis, dosen, teman, dan sahabat yang telah menemani masa perkuliahan serta seluruh pihak yang telah membantu dalam menyelesaikan skripsi ini.

KATA PENGANTAR

Puji syukur penulis ucapkan kepada Allah SWT yang telah memberikan rahmat, taufiq dan hidayat-Nya sehingga penulis dapat menyelesaikan skripsi dengan judul “Peringkasan Teks Ekstraktif Pada Artikel Berita Bahasa Indonesia Menggunakan Metode *BERT* dan *Cosine Similarity*”.

Sholawat serta salam semoga tetap tercurahkan kepada Rasulullah Muhammad SAW yang telah memberikan bimbingan dan petunjuk sehingga kita semua tetap dalam iman islam dan dapat melaksanakan kegiatan belajar dengan cahaya islam.

Penulisan dan penyusunan skripsi ini dimaksudkan untuk memenuhi salah satu persyaratan dalam menyelesaikan program studi Teknik Informatika di Universitas Islam Negeri Maulana Malik Ibrahim Malang. Penulis menyadari bahwa skripsi ini tidak dapat terwujud tanpa adanya bantuan, bimbingan dan arahan dari berbagai pihak. Oleh karena itu, dengan penuh ketulusan hati penulis mengucapkan terima kasih dikhususkan kepada :

1. Prof. Dr. M. Zainuddin, MA selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Sri Harini, M.Si., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Ir. Fachrul Kurniawan, S.T., M.MT., IPU selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.

4. Dr. Zainal Abidin, M.Kom selaku Dosen Pembimbing I yang telah memberikan arahan dan bimbingan dengan baik dalam penulisan dan program dalam penyelesaian skripsi ini.
5. Dr. M. Imamudin, Lc, MA selaku Dosen Pembimbing II yang telah memberikan bimbingan, arahan serta bantuan dalam skripsi ini.
6. Dr. Ir. Fachrul Kurniawan, S.T., M.MT., IPU selaku ketua penguji, Shoffin Nahwa Utama, M.T selaku penguji I, Dr. Zainal Abidin, M.Kom selaku penguji II, Dr. M. Imamudin, Lc, MA selaku penguji III yang telah meluangkan waktunya untuk menguji dan dengan sabar memberikan arahan serta saran dalam menyelesaikan skripsi ini.
7. Prof. Dr. Muhammad Faisal, M.T selaku Dosen Wali yang telah dengan sabar memberikan arahan dalam proses perkuliahan.
8. Segenap Dosen, Laboran, dan jajarannya pada Program Studi Teknik Informatika yang telah memberikan bimbingan dan bantuan selama perkuliahan.
9. Mbak Nia Faricha, S.Si selaku Admin Program Studi Teknik Informatika yang telah sabar dalam membantu dan memberikan arahan informasi terkait perkuliahan.
10. Kedua orang tua penulis, Ayah Zaenal dan Ibu Zuhrotul Mu'awanah, terimakasih tiada terhingga penulis sampaikan atas segala cinta kasih sayang, arahan, dukungan dan apapun yang telah diberikan. Terimakasih atas do'a yang selalu engkau panjatkan selama ini, terimakasih untuk semua hal apapun itu Ayah dan Ibuku tercinta.

11. Sahabatku di Tuban Utamimatul Ula yang senantiasa mendukung dalam hal positif apapun. Teman-temanku juga di Permata Ronggolawe kurnia putri, alfianita, ulum arofatin.
12. Teman-teman seperjuangan, luqyana, riska dwi, nisa kholifatul, diyah, naila, agil, abidzar, yusuf, ilham rafif, juga teman-teman lain yang saya sayangi.
13. Seluruh keluarga besar Saudara Teknik Informatika UIN Malang terkhusus angkatan 2018 “UFO (*Unity Of Informatics Force*)”, terimakasih sekali
14. Seluruh pihak yang telah terlibat secara langsung maupun tidak langsung yang tidak bisa saya sebutkan satu persatu dalam penyelesaian skripsi ini.
15. *And lastly*, kepada diri saya sendiri, Nurma Romihim Fadlilah. Terima kasih sudah bertahan sampai saat ini. Terima kasih sudah berusaha sampai saat ini, walaupun seringkali merasa putus asa dengan apa yang sudah diusahakan, namun tetap menjadi manusia yang mau berusaha dan bangkit dari keterpurukan. Terimakasih untuk tidak menyerah dalam proses penyusunan skripsi ini dan berusaha menyelesaikannya semaksimal mungkin.

Penulis menyadari dalam penulisan skripsi ini tidak luput dari kesalahan dan jauh dari kata sempurna. Oleh karena, itu, penulis mengharapkan kritik dan saran yang membangun sehingga skripsi ini bisa dikembangkan lagi.

Malang, 17 Juni 2025

Penulis

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PENGAJUAN	ii
HALAMAN PERSETUJUAN	iii
HALAMAN PENGESAHAN	iv
PERNYATAAN KEASLIAN TULISAN	v
HALAMAN MOTTO	vi
HALAMAN PERSEMBAHAN	vii
KATA PENGANTAR	viii
DAFTAR ISI	xi
DAFTAR GAMBAR	xiii
DAFTAR TABEL	xiv
ABSTRAK	xv
ABSTRACT	xvi
المُلخَص	xvii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Pernyataan Masalah	6
1.3 Tujuan Penelitian	6
1.4 Batasan Masalah.....	7
1.5 Manfaat Penelitian	7
BAB II STUDI PUSTAKA	8
2.1 Penelitian Terkait	8
2.2 <i>Natural Language Processing (NLP)</i>	17
2.2.1 Sistem Peringkasan Teks	18
2.2.2 <i>Text Preprocessing</i>	19
2.3 Berita	20
2.4 <i>BERT</i>	21
2.5 <i>Cosine Similarity</i>	28
2.6 <i>Recall-Oriented Understudy For Gisting Evaluation (ROUGE)</i>	29
BAB III DESAIN DAN IMPLEMENTASI	32
3.1 Desain Sistem.....	32
3.2 Pengumpulan Data	35
3.3 <i>Data Loading dan Parsing</i>	37
3.4 <i>Preprocessing</i>	39
3.4.1 <i>Text Cleaning</i> (Pembersihan Teks).....	39
3.4.2 <i>Case Folding</i>	40
3.5 Pembagian Data	42
3.6 <i>BERT Tokenization</i>	42
3.7 <i>Fine-Tuning Model BERT</i>	44
3.8 Evaluasi Model.....	45
3.9 <i>BERT Importance Score</i>	46
3.10 <i>BERT Embeddings dan Cosine Similarity</i>	47
3.11 <i>Combined Score Calculation</i>	49

3.12 <i>Sentence Selection</i>	49
3.13 Evaluasi Ringkasan (<i>ROUGE Evaluation</i>)	51
3.14 Contoh Hasil <i>Summary</i> (Contoh Hasil Ringkasan).....	53
BAB IV HASIL DAN PEMBAHASAN	55
4.1 Skenario Uji Coba.....	55
4.1.1 Dataset Uji Coba.....	55
4.1.2 Parameter Uji Coba.....	56
4.1.3 Metrik Evaluasi.....	57
4.2 Hasil Uji Coba.....	58
4.2.1 Hasil <i>Fine Tuning BERT</i>	58
4.2.2 Hasil Eksperimen Parameter <i>Alpha</i>	59
4.2.3 Hasil Evaluasi pada <i>Testing Data</i>	60
4.2.4 Hasil Berdasarkan Kategori	61
4.2.5 Hasil Eksperimen <i>Ratio</i>	62
4.2.6 Hasil Peringkasan	63
4.3 Pembahasan.....	68
4.3.1 Analisis Hasil <i>Fine Tuning</i>	69
4.3.2 Analisis Parameter <i>Alpha</i>	69
4.3.3 Analisis Performa Pada Kategori.....	70
4.3.4 Analisis Pemilihan <i>Ratio</i>	71
4.4 Gambar Implementasi Sistem	71
4.5 Integrasi Islam.....	72
4.5.1 <i>Mu'amalah Ma'a Allah</i>	73
4.5.2 <i>Mu'amalah Ma'a An-Nas</i>	75
BAB V KESIMPULAN DAN SARAN	77
5.1 Kesimpulan	77
5.2 Saran.....	78
DAFTAR PUSTAKA	

DAFTAR GAMBAR

Gambar 2. 1 Gambar <i>pre-training</i> dan <i>fine tuning</i> dalam <i>BERT</i>	24
Gambar 2. 2 Arsitektur umum metode <i>BERT</i>	25
Gambar 2. 3 Arsitektur <i>BERT</i> dalam <i>text summarization</i>	25
Gambar 3. 1 Desain sistem peringkasan teks ekstraktif menggunakan <i>BERT</i> dan <i>cosine similarity</i>	33
Gambar 3. 2 <i>Flowchart preprocessing</i>	39
Gambar 4. 1 Teks asli artikel kategori olahraga	63
Gambar 4. 2 Teks ringkasan sistem kategori olahraga	64
Gambar 4. 3 Teks ringkasan <i>indosum</i> kategori olahraga	64
Gambar 4. 4 Evaluasi <i>rouge</i> pada artikel kategori olahraga	64
Gambar 4. 5 Hasil skor ringkasan kalimat kategori olahraga	65
Gambar 4. 6 Teks asli artikel kategori hiburan	66
Gambar 4. 7 Teks ringkasan sistem kategori hiburan	66
Gambar 4. 8 Teks ringkasan <i>indosum</i> kategori hiburan.....	66
Gambar 4. 9 Evaluasi <i>rouge</i> pada artikel kategori hiburan.....	67
Gambar 4. 10 Hasil skor ringkasan kalimat kategori hiburan.....	67
Gambar 4. 11 Implementasi sistem	71
Gambar 4. 12 Implementasi sistem 2	72

DAFTAR TABEL

Tabel 2. 1 Penelitian terkait	14
Tabel 3. 1 Fitur- fitur <i>dataset</i>	36
Tabel 3. 2 Contoh <i>dataset</i>	37
Tabel 3. 3 Tabel contoh hasil <i>parsing</i>	38
Tabel 3. 4 Contoh proses <i>text cleaning</i>	40
Tabel 3. 5 Contoh proses <i>case folding</i>	41
Tabel 3. 6 Tabel contoh hasil <i>preprocessing</i>	41
Tabel 3. 7 Tabel contoh <i>tokenization BERT</i>	43
Tabel 3. 8 Hasil contoh <i>BERT importance score</i>	47
Tabel 3. 9 Contoh perhitungan <i>similarity matrix</i>	48
Tabel 3. 10 Tabel contoh skor hasil kombinasi <i>BERT</i> dan <i>cosine similarity</i>	49
Tabel 3. 11 Contoh hasil ringkasan.....	53
Tabel 3. 12 <i>ROUGE score</i> contoh hasil ringkasan	54
Tabel 4. 1 Pembagian <i>dataset</i>	55
Tabel 4. 2 Distribusi kategori pada <i>test set</i>	56
Tabel 4. 3 Spesifikasi lingkungan pelatihan	56
Tabel 4. 4 Skenario parameter uji coba.....	56
Tabel 4. 5 Parameter uji coba sistem	57
Tabel 4. 6 Hasil <i>training</i> per <i>epoch</i>	58
Tabel 4. 7 Eksperimen parameter <i>alpha</i>	59
Tabel 4. 8 Hasil evaluasi <i>testing</i> data.....	60
Tabel 4. 9 Performa per kategori artikel	61
Tabel 4. 10 Hasil eksperimen <i>ratio</i>	62

ABSTRAK

Fadlilah, Nurma Romihim. 2025. **Peringkasan Teks Ekstraktif Pada Artikel Berita Bahasa Indonesia Menggunakan Metode *BERT* dan *Cosine Similarity***. Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Zainal Abidin, M. Kom (II) Dr. M. Imamudin Lc, MA.

Kata kunci : Peringkasan Teks Ekstraktif, *BERT*, *Cosine Similarity*, *ROUGE*

Volume informasi berita online yang meningkat secara cepat menyebabkan pembaca kesulitan dalam menyaring informasi penting secara efisien. Penelitian ini bertujuan mengembangkan sistem peringkasan teks ekstraktif untuk artikel berita berbahasa Indonesia dengan mengombinasikan metode *Bidirectional Encoder Representations from Transformers (BERT)* dan *Cosine Similarity*. *BERT* digunakan untuk memahami konteks semantik kalimat secara *bidirectional*, sementara *Cosine Similarity* mengukur keterkaitan antar kalimat untuk mengurangi redundansi. Penelitian menggunakan dataset *IndoSum* pada file “*train.03.jsonl*” yang menggunakan 2000 artikel berita dengan pembagian 70% data training (1400 artikel), 20% data validasi (400 artikel), dan 10% data testing (200 artikel). Proses penelitian meliputi tahapan *preprocessing (text cleaning dan case folding)*, *fine-tuning* model *IndoBERT* selama 12 *epoch*, ekstraksi fitur *BERT*, perhitungan *Cosine Similarity*, dan kombinasi skor dengan parameter *alpha* optimal 0.9 (90% bobot *BERT*, 10% bobot *Cosine Similarity*). Evaluasi menggunakan metrik *ROUGE* menunjukkan hasil *ROUGE-1*: 0.537 (± 0.155), *ROUGE-2*: 0.456 (± 0.191), dan *ROUGE-L*: 0.499 (± 0.178) pada 200 artikel *test set*. Sistem mencapai *compression rate* rata-rata 72.7% dengan waktu eksekusi 0.8 detik per artikel. Hasil tersebut menunjukkan bahwa sistem dapat menghasilkan ringkasan yang cukup baik dalam konteks peringkasan teks ekstraktif.

ABSTRACT

Fadlilah, Nurma Romihim. 2025. **Extractive Text Summarization for Indonesian-Language News Articles Using the BERT dan Cosine Similarity Methods.** Thesis. Informatics Engineering Study Program. Faculty of Science and Technology Universitas Islam Negeri Maulana Malik Ibrahim Malang. Advisor: (I) Dr. Zainal Abidin, M. Kom (II) Dr. M. Imamudin Lc, MA.

Keywords: Extractive Text Summarization, BERT, Cosine Similarity, ROUGE

The rapid growth of online news has made it increasingly challenging for readers to efficiently filter and identify key information. The research aims to develop an extractive text summarization system for Indonesian-language news articles by combining the Bidirectional Encoder Representations from Transformers (BERT) method to capture the semantic meaning and Cosine Similarity method to measure sentence cohesion and minimize redundancy. The research uses file *train.03.jsonl* of the IndoSum dataset, consisting of 2,000 news articles split into 70% training data (1,400 articles), 20% validation data (400 articles), and 10% test data (200 articles). The process includes preprocessing (text cleaning and case folding), fine-tuning the IndoBERT model over 12 epochs, extracting BERT features, calculating Cosine Similarity, and combining scores using an optimal alpha parameter of 0.9—giving 90% weight to BERT and 10% to Cosine Similarity. Evaluation using ROUGE metrics produced the following results on 200 test set articles: ROUGE-1: 0.537 (± 0.155), ROUGE-2: 0.456 (± 0.191), and ROUGE-L: 0.499 (± 0.178). The system achieved an average compression rate of 72.7% and an execution time of 0.8 seconds per article. These results demonstrate that the system can generate concise and fairly accurate extractive text summaries.

الملخص

فاضيلة، نورما رومي حيم. ٢٠٢٥. تلخيص النصوص الاستخراجية لمقالات الأخبار باللغة الإندونيسية باستخدام طريقة BERT و Cosine Similarity. الرسالة الجامعية. برنامج دراسات تكنولوجيا المعلومات، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف: (١) الدكتور زين العابدين، ماجستير في علوم الحاسوب (٢) الدكتور محمد الإمام الدين، الماجستير.

الكلمات المفتاحية: تلخيص النصوص الاستخراجية، BERT، تشابه جيب التمام، ROUGE.

الزيادة السريعة في حجم المعلومات الإخبارية عبر الإنترنت تجعل من الصعب على القراء فرز المعلومات الهامة بكفاءة. تهدف هذه الدراسة إلى تطوير نظام تلخيص نصوص استخراجية لمقالات الأخبار باللغة الإندونيسية من خلال الجمع بين طريقة BERT (تمثيلات المبرمجين الثنائية الاتجاه من المحولات) وتشابه جيب التمام. يتم استخدام BERT لفهم السياق الدلالي للجمل بشكل ثنائي الاتجاه، بينما يقيس تشابه جيب التمام العلاقة بين الجمل لتقليل التكرار. تستخدم الدراسة مجموعة بيانات IndoSum في ملف "train.03.jsonl" الذي يحتوي على ٢٠٠٠ مقال إخباري، مع تقسيم ٧٠٪ للتدريب (١٤٠٠ مقال)، و ٢٠٪ للتحقق (٤٠٠ مقال)، و ١٠٪ للاختبار (٢٠٠ مقال). تشمل عملية البحث مراحل المعالجة المسبقة (تنظيف النص وطى الحالة)، وضبط النموذج IndoBERT لمدة ١٢ حقبة، واستخراج ميزات BERT، وحساب تشابه جيب التمام، ودمج النقاط مع المعلمة ألفا المثلى ٠.٩ (٩٠٪ وزن BERT، ١٠٪ وزن تشابه جيب التمام). تظهر التقييمات باستخدام مقياس ROUGE النتائج التالية: ROUGE (٠.٥٣٧، ±٠.١٥٥) (1: ROUGE-2: ٠.٤٥٦، ±٠.١٩١)، و ROUGE-L: ٠.٤٩٩، ±٠.١٧٨) على ٢٠٠ مقال اختبار. حقق النظام معدل ضغط متوسط قدره ٧٢.٧٪ مع وقت تنفيذ ٠.٨ ثانية لكل مقال. تشير النتائج إلى أن النظام يمكن أن ينتج ملخصات جيدة إلى حد ما في سياق تلخيص النصوص الاستخراجية.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pada era modernisasi saat ini, volume informasi yang tersedia secara *online* meningkat secara cepat. Salah satu sumber informasi yang paling menonjol yakni artikel berita. Menurut data dari Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) pada tahun 2024, 79.50% (221.563.479 jiwa) penduduk Indonesia menggunakan internet, dan konsumsi berita *online* mengalami peningkatan yang signifikan (APJII, 2024). Dengan demikian untuk menyaring serta meringkas konten menjadi semakin penting.

Artikel berita adalah informasi yang disampaikan melalui berbagai media seperti seperti surat kabar, televisi, radio, dan internet yang mengandung laporan mengenai peristiwa atau keadaan tertentu yang memiliki nilai baru, menarik, atau bermanfaat bagi pembaca, pendengar, atau penonton. Berita seringkali memiliki ciri khas yang mencakup kebaruan, relevansi, fakta- fakta yang akurat, objektivitas, serta penyajian yang jelas dan ringkas. Perkembangan teknologi informasi yang pesat telah membawa perubahan signifikan dalam berbagai bidang, termasuk dalam penyebaran informasi. Media berita, baik cetak maupun digital, kini menjadi salah satu sumber utama informasi bagi masyarakat yang mudah diakses dimana saja melalui internet dan dapat dibaca kapan saja.

Membaca salah salah satu kegiatan yang tidak terpisahkan dari kehidupan manusia, baik itu membaca buku, majalah, maupun artikel berita. Artikel berita memungkinkan individu untuk tetap mendapatkan informasi terkini tentang

peristiwa dan perkembangan terbaru di berbagai bidang seperti kesehatan, olahraga, teknologi, dan lain- lain. Tantangan muncul ketika dokumen ataupun teks yang akan dibaca memiliki panjang dan jumlah yang signifikan, karena hal ini memerlukan waktu yang cukup lama untuk memahami isi dokumen tersebut (Savanti *et al.*, 2018). Pembaca juga seringkali mengalami kesulitan dalam menyaring informasi yang relevan dan penting. Hal ini menjadi tantangan tersendiri di tengah rendahnya tingkat literasi di Indonesia.

Indonesia memiliki tingkat literasi yang rendah apabila dibandingkan dengan negara lain. Berdasarkan hasil survei yang dilakukan oleh *Program for International Student Assessment (PISA)* yang dirilis oleh *Organization for Economic Co-operation and Development (OECD)* pada tahun 2022, Indonesia berada di peringkat 10 terbawah dalam kategori literasi membaca. Indonesia kini menempati peringkat 70 dari 80 negara dengan skor literasi membaca sebesar 359 (OECD, 2023). Beberapa faktor yang menyebabkan rendahnya literasi di Indonesia yakni tidak membiasakan diri untuk membaca buku dari rumah, perkembangan teknologi yang semakin pesat, minimnya saran untuk membaca, kurangnya motivasi untuk membaca, dan sifat malas untuk mengembangkan ide.

Dari permasalahan yang telah disebutkan diatas, di dalam QS. Al-Asr surat ke 103 ayat 1-3 yang berbunyi :

وَالْعَصْرِ ﴿١﴾ إِنَّ الْإِنْسَانَ لَفِي خُسْرٍ ﴿٢﴾ إِلَّا الَّذِينَ آمَنُوا وَعَمِلُوا الصَّالِحَاتِ وَتَوَّصَوْا بِالْحَقِّ وَتَوَّصَوْا

بِالصَّبْرِ ﴿٣﴾

“ 1) Demi Masa, 2) Sungguh, manusia berada dalam kerugian, 3) Kecuali orang-orang yang beriman dan mengerjakan kebajikan serta saling menasihati untuk kebenaran dan saling menasihati untuk kesabaran” (QS. Al- Ashr/103: 1-3).

Dalam QS.Al-Ashr ayat 1-3(Online, n.d.-a) ini sebagai salah satu surat yang paling sempurna petunjuknya, dimana mengingatkan betapa pentingnya waktu, karena waktu merupakan modal utama manusia. Apabila waktu tidak di isi dengan baik maka ia akan berlalu begitu saja, ia akan hilang. Kemudian, jangankan keuntungan yang di dapat, modal saja pun juga tidak akan kembali (Bluto, 2020).

Berdasarkan surah tersebut, manusia dianjurkan untuk memanfaatkan waktu dengan sebaik mungkin guna meningkatkan kualitas ibadah lainnya. Sangat disayangkan apabila seseorang tidak menggunakan waktu dengan bijak, karena pada dasarnya, mereka yang tidak memanfaatkan waktu dengan baik termasuk golongan yang merugi. Sehingga artikel berita yang panjang, apabila di baca, maka pembaca akan cepat jenuh karena memakan waktu yang lama. Maka perlu adanya sistem peringkasan artikel berita secara otomatis agar pembaca lebih cepat mengetahui intisari dari suatu artikel berita yang di baca serta dapat mengefisiensi waktu yang ada. Dengan berbekal ilmu pengetahuan dan perkembangan teknologi saat ini maka sangat diperlukan suatu sistem yang dapat melakukan peringkasan informasi dalam teks artikel secara tepat, benar dan efisien.

Natural Language Processing (NLP) adalah kombinasi dari ilmu komputer dan kecerdasan buatan yang terkait dengan *linguistik*. NLP berfokus pada bagaimana mesin memahami bahasa manusia untuk dapat saling berinteraksi. Dengan adanya NLP, komputer dapat belajar dan memahami bahasa manusia, sehingga memungkinkan komputer untuk berkomunikasi dengan manusia. Bahasa

manusia merupakan sesuatu yang unik karena dirancang khusus untuk menyampaikan makna. Membuat komputer dapat mengerti bahasa manusia merupakan tugas yang menantang, karena bahasa manusia memiliki struktur yang kompleks. Selain itu bahasa memiliki keunikannya sendiri dan mungkin memiliki makna ganda (Prasetyo *et al.*, 2021). NLP (*Natural Language Processing*) dalam penerapannya dapat ditemukan pada mesin penerjemah, pendeteksi email spam, *chatbot*, dan peringkasan teks. Kemampuan NLP dalam menganalisis bahasa memungkinkan aplikasi pada sistem peringkasan otomatis.

Metode peringkasan teks ekstraktif merupakan salah satu teknik yang digunakan untuk mengekstraksi informasi penting dari teks yang panjang. Teknik ini bekerja dengan cara memilih kalimat- kalimat kunci dari teks asli dan menyusunnya kembali menjadi ringkasan yang lebih singkat. Dalam proses ini, peringkasan menerima input berupa teks, kemudian melakukan peringkasan, dan menghasilkan *output* berupa teks yang lebih singkat dibandingkan teks aslinya. Hasil peringkasan tersebut mengandung poin-poin penting atau informasi utama dari teks sumber (Khaqiqi *et al.*, 2022). Pada sistem peringkasan otomatis dapat menggunakan metode berbasis kueri, yang mana hasil ringkasan disesuaikan dengan kueri atau topik tertentu (Abdi *et al.*, 2018).

Model yang digunakan dalam penelitian ini adalah *BERT* (*Bidirectional Encoder Representations from Transformers*), yang dikembangkan oleh Google. *BERT* merupakan metode lanjutan/ perkembangan dari *Transformer* yang diperkenalkan pada tahun 2019. Sesuai dengan namanya, *BERT* hanya berfokus pada proses *encode* dan menghasilkan sebuah model bahasa. Oleh karena itu *BERT*

hanya memerlukan *encoder* (Pratama & Romadhony, 2020). *BERT* memiliki kemampuan untuk memahami konteks dua arah dalam teks, sehingga menghasilkan ringkasan yang lebih akurat dan relevan (Devlin, Chang, Lee, Google, et al., 2019), apabila dibandingkan dengan model yang hanya memproses teks dari kiri ke kanan atau sebaliknya. Tujuan dari metode *BERT* ini adalah untuk membantu komputer memahami makna bahasa yang ambigu di dalam sebuah teks artikel dengan menggunakan teks di sekitarnya guna menetapkan konteks.

Selain *BERT*, teknik *Cosine Similarity* juga di gunakan dalam proses peringkasan teks ekstraktif pada penelitian ini. *Cosine Similarity* mengukur kemiripan antara dua dokumen atau teks. Di dalam *Cosine Similarity* teks ataupun dokumen disebut sebagai *vector* (Sanjaya et al., 2023). Dengan menggabungkan *BERT* dan *Cosine Similarity*, diharapkan dapat menghasilkan sistem peringkasan teks yang tidak hanya akurat, tetapi juga relevan dengan konteks aslinya.

Penelitian pernah dilakukan oleh (Samosir et al., 2022) yakni memperkenalkan pendekatan inovatif untuk peringkasan teks ekstraktif menggunakan kombinasi *K-Means Clustering* dan arsitektur berbasis *BERT*, khususnya *Sentence-BERT*. Dengan mengelompokkan kalimat- kalimat yang tersemat dari jurnal ilmiah, penulis telah mengembangkan metode yang secara efektif mengidentifikasi kalimat- kalimat kunci untuk di ringkas. Hasilnya menunjukkan performa yang menjanjikan. Menunjukkan potensi model ini dalam menyempurnakan tugas peringkasan teks. Namun, ada beberapa kelauman model dalam penelitian ini, seperti ketidakmampuan untuk menangkap formula matematis dengan baik dan masih menyertakan bagian dari daftar pustaka dalam ringkasan.

Hal ini dikarenakan model yang menganggap judul literatur merupakan bagian dari isi makalah jurnal.

Pada penelitian ini menggunakan metode *BERT (Bidirectional Encoder Representations from Transformers)* dan *Cosine Similarity* untuk membuat sistem peringkasan teks ekstraktif pada artikel berita bahasa Indonesia. Metode *BERT* dipilih karena memiliki kemampuan untuk memahami konteks kata secara *bidirectional*. *Cosine Similarity* digunakan untuk mengukur seberapa mirip kalimat satu dengan kalimat yang lain. Hal ini memungkinkan untuk menemukan kalimat-kalimat yang paling mirip dalam artikel berita. Penelitian ini menggunakan dataset *Indosum*, yang merupakan kumpulan artikel berita berbahasa Indonesia beserta ringkasannya, dan metrik *ROUGE (Recall Oriented Understudy for Gisting Evaluation)* untuk mengevaluasi performa sistem. Penelitian ini diharapkan dapat mempermudah pembaca dalam memahami artikel berita bahasa Indonesia secara relevan dengan lebih cepat, tepat, dan efisien.

1.2 Pernyataan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah pada penelitian ini yaitu bagaimana implementasi pengembangan sistem peringkasan teks ekstraktif menggunakan *BERT* dan *Cosine Similarity* pada artikel berita berbahasa Indonesia?

1.3 Tujuan Penelitian

1. Mengembangkan metode peringkasan teks ekstraktif untuk artikel berita Indonesia menggunakan model *BERT* dan *Cosine Similarity*

2. Mengevaluasi performa metode yang diusulkan dengan menggunakan metrik evaluasi *ROUGE*.

1.4 Batasan Masalah

1. Pada penelitian ini hanya berfokus pada artikel berita berbahasa Indonesia
2. Metode yang digunakan dalam penelitian ini yakni metode peringkasan teks ekstraktif, bukan metode ringkasan teks abstraktif
3. Evaluasi efektivitas metode hanya dilakukan berdasarkan kriteria akurasi dan relevansi ringkasan teks

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis. Secara teoritis, penelitian ini diharapkan dapat memperkaya literature mengenai teknik peringkasan teks ekstraktif. Khususnya yang menggunakan *BERT* dan *Cosine Similarity* dalam konteks bahasa Indonesia. Penelitian ini diharapkan dapat membantu pembaca artikel berita dalam menyaring informasi yang penting dan relevan secara lebih cepat dan efisien. Selain itu, metode yang dikembangkan dalam penelitian ini juga diharapkan dapat di implementasikan oleh media berita dalam menyajikan ringkasan artikel yang lebih informatif dan mudah dipahami oleh pembaca.

BAB II

STUDI PUSTAKA

Pada bab ini mengulas penelitian terkait yang telah dilakukan sebelumnya yang relevan dengan penelitian ini. Selain itu, bab ini juga menjelaskan studi pustaka yang digunakan sebagai landasan teori dalam penelitian ini.

2.1 Penelitian Terkait

Terdapat beberapa penelitian terkait sebelumnya, salah satunya penelitian yang telah dilakukan oleh (Utomo *et al.*, 2022) yang berjudul “*Text Summarization Pada Artikel Berita Menggunakan Vector Space Model dan Cosine Similarity*”. Pada penelitian tersebut dengan model yang digunakan, pembaca dapat memahami isi artikel dengan lebih cepat tanpa mengubah substansi asal. Prosesnya yakni melibatkan beberapa langkah penting seperti *crawling*, *tokenization*, penghapusan tanda baca, *stopword*, dan *stemming* sebelum menerapkan algoritma. Data yang diujikan dalam penelitian tersebut berasal dari salah satu artikel publikasi dengan judul “Otomasi Peringkasan Dokumen Sebagai Pendukung Sistem Manajemen Surat” yang di tulis oleh Ahmad Najibullah dan Wang Mingyan dari Universitas Nanchang, Republik Rakyat Tiongkok. Hasil dari penelitian tersebut yakni hasil uji coba menunjukkan tingkat kompresi 53,67 %, dengan ringkasan yang dihasilkan mencerminkan 96,67% informasi utama dari dokumen asli. Proses tokenisasi, penghapusan *stopword*, dan *stemming* berkontribusi signifikan terhadap kualitas ringkasan, misalnya setelah proses *stopword* jumlah kata dalam judul artikel berkurang dari 8 kata menjadi 7 kata, yang menunjukkan penghapusan kata yang

tidak penting. Perbedaan penelitian terletak pada metode serta data yang digunakan, pada penelitian rujukan yakni menggunakan *Vector Space Model* sedangkan dalam penelitian ini menggunakan metode BERT.

Penelitian yang dilakukan oleh (Fatmalasari & Lumbanraja, 2022) tentang pengembangan sistem peringkasan teks otomatis untuk artikel ilmiah berbahasa Indonesia menggunakan metode pembobotan kalimat. Data yang digunakan dalam penelitian ini adalah berupa file yang berekstensi PDF dari Jurnal Komputasi yang diperoleh dari website Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung, dengan jumlah 100 dokumen. Tahapan dalam penelitian tersebut yaitu mengumpulkan data 100 jurnal ilmiah, ekstraksi file dari PDF ke format TXT, menghapus abstrak dokumen, parsing kalimat, pengambilan term, pembobotan kalimat menggunakan algoritma TF-IDF, pemilihan kalimat, pengujian sistem menggunakan *User Acceptance Testing* (UAT). Hasil penelitian yang diperoleh dari penelitian tersebut adalah penerimaan pengguna, sistem peringkasan teks otomatis mendapatkan respon positif dari pengguna, berdasarkan hasil pengujian 82.6% responden setuju dengan sistem, 80.2% setuju dengan efisiensi ringkasan yang dihasilkan, dan 83,7% menyatakan puas dengan penggunaan sistem. Pengujian dilakukan menggunakan *User Accepted Testing* (UAT) dengan melibatkan 41 responden yang terdiri dari mahasiswa dan siswa untuk memperoleh umpan balik yang representatif mengenai sistem. Perbedaan penelitian ada pada metode peringkasan yang digunakan, pada penelitian rujukan menggunakan metode pembobotan kalimat dan TF-IDF sebagai salah satu teknik utama yang digunakan dalam pembobotan kalimat, *Cosine Similarity* untuk

memeriksa suatu kesamaan dengan dokumen yang sudah ada. Sedangkan pada penelitian ini menggunakan metode *BERT* dan *Cosine Similarity*. Data yang digunakan dalam penelitian rujukan adalah dokumen- dokumen dari jurnal ilmiah khususnya dalam bidang komputasi, sedangkan dalam penelitian ini menggunakan artikel berita berbahasa Indonesia yang diambil dari IndoSum.

Penelitian di lakukan oleh (AL-Hafiidh *et al.*, 2022) tentang peringkasan teks otomatis pada portal berita olahraga menggunakan metode *Maximum Marginal Relevance (MMR)* serta metode *TF-IDF* untuk pembobotan kata. Data yang digunakan dalam penelitian ini adalah berita olahraga yang berasal dari tiga portal berita olahraga, yaitu sport.okezone.com, sport.detik.com, dan sport.sindonews.com. Data tersebut diambil menggunakan teknik *web scraping* pada portal berita olahraga tersebut. Adapaun tahapan dalam penelitian tersebut diantaranya mengidentifikasi masalah, pengumpulan data, *preprocessing* teks (*tokenization, filtering, stemming*), pembobotan kata, penerapan metode *MMR*, mengevaluasi hasil. Pengujian diambil dari data sampel berita *online* yakni 5 berita sindonews yang menggunakan lamda 0,7. Hasil pengujian sistem menunjukkan akurasi rata rata yakni *precision* : 57,7%, *recall* : 48,5%, *f-measure* : 50,3%. Pada penelitian rujukan, metode yang di gunakan untuk peringkasan teks yakni metode *Maximum Marginal Relevance (MMR)*, *TF-IDF* untuk memberikan bobot pada kata- kata dalam teks, dan *Cosine Similarity* untuk menghitung tingkat kesamaan antar dua objek. Sedangkan dalam penelitian ini menggunakan metode *BERT* untuk peringkasan teks dan *Cosine Similarity* untuk mengukur kemiripan antara dua dokumen/ teks juga. Perbedaan lainnya yakni data yang digunakan, pada penelitian

rujukan menggunakan artikel berita olahraga pada portal berita sport.okezone.com, sport.detik.com, dan sport.sindonews.com. Sedangkan dalam penelitian ini menggunakan dataset artikel berita berbahasa Indonesia yang diambil dari *IndoSum*.

Penelitian dilakukan oleh (Astuti *et al.*, 2024) yang berjudul *Indonesian News Text Summarization Using MBART Algorithm*. Penelitian tersebut membahas tentang efektivitas penggunaan algoritma *MBART* untuk melakukan peringkasan teks berita dalam bahasa Indonesia. Metode yang digunakan dalam penelitian tersebut yakni algoritma *MBART*. Adapun data yang digunakan dalam penelitian tersebut adalah dataset *XL-SUM* yang berisi 47.800 artikel berita berbahasa Indonesia. Pada saat pengujian dataset ini dibagi menjadi data pelatihan (38.200) kemudian data pengujian dan validasi (4.780), pembagian ini penting guna memastikan model dapat diuji secara efektif serta hasilnya pun dapat dievaluasi secara akurat. Adapun tahapan dalam penelitian tersebut yaitu pengumpulan *dataset*, *pre-processing*, *fine-tuning model*, pelatihan model, evaluasi, prediksi ringkasan. Hasil dari penelitian tersebut yaitu menghasilkan evaluasi menggunakan metrik *ROUGE-1* : 35.94, *ROUGE-2* : 16.43, *ROUGE-L* : 29.91 yang mencerminkan bahwa kinerja model dalam menghasilkan ringkasan yang relevan serta informatif. Perbedaan penelitian ada pada metode yang digunakan, pada penelitian rujukan menggunakan metode *MBART*, sedangkan pada penelitian ini menggunakan metode *BERT* dan *Cosine Similarity*. Perbedaan lain juga ada pada data penelitian yang digunakan, pada penelitian rujukan data yang digunakan yakni

berasal dari *dataset XL-SUM*, sedangkan pada penelitian ini menggunakan data dari dataset *Indosum*.

Penelitian juga di lakukan oleh (Halim *et al.*, 2022) yang berjudul “Ringkasan Ekstraktif Otomatis pada Berita Berbahasa Indonesia Menggunakan Metode *BERT*”. Pada penelitian tersebut membahas mengenai pembuatan ringkasan otomatis berita berbahasa Indonesia menggunakan metode *Bidirectional Encoder Representations from Transformers (BERT)*. Tujuan penelitiannya adalah mengembangkan dan mengevaluasi metode ringkasan otomatis untuk berita berbahasa Indonesia menggunakan model *BERT*. Penelitian ini membedakan antara dua metode ringkasan, yakni ekstraktif dan abstraktif. *Dataset* yang digunakan yakni dataset *IndoSum*, dataset ini menyediakan *label gold summarization* untuk peringkasan ekstraktif. Pada penelitian tersebut dilakukan pencarian konfigurasi yang tepat dalam model *BERT* kemudian hasilnya dibandingkan dengan referensi ekstraktif dan abstraktif. Di dapatkan *score* pengujiannya untuk referensi ekstraktif *ROUGE-1* : 84.46, *ROUGE-2* : 83.21, *ROUGE-L* : 83.40, dan untuk referensi abstraktifnya yakni *ROUGE-1* : 57.17, *ROUGE-2* : 51.27, *ROUGE-L* : 55.20, sedangkan abstraktif hasilnya *ROUGE-1* : 57.17, *ROUGE-2* : 51.27, *ROUGE-L* : 55.20. Perbedaan penelitian, pada penelitian rujukan tidak mencakup teknik *cosine similarity* di dalamnya, sedangkan dalam penelitian ini terdapat teknik *cosine similarity* untuk mengukur kesamaan/ kemiripan pada dokumen artikel.

Penelitian pernah dilakukan oleh (Wang *et al.*, 2019) berjudul “*A Text Abstraction Summary Model Based on BERT Word Embedding and Reinforcement Learning*”. Penelitian tersebut membahas terkait pengembangan model ringkasan

otomatis yang memanfaatkan model *BERT* dengan pendekatan *hybrid* (menggabungkan antara pendekatan ekstraktif dan abstraktif). *Dataset* yang digunakan dalam penelitian tersebut CNN/Daily Mail dan DUC2002. Hasil dari penelitian tersebut model *BERT* menunjukkan peningkatan apabila dibandingkan dengan model *baseline*, yakni peningkatan sebesar *ROUGE-1* : 1.07%, *ROUGE-2* : 2.46%, *ROUGE-L* : 0.95%, *R-AVG* : 1.46%. Perbedaan penelitian terletak pada pendekatannya, penelitian rujukan menggunakan pendekatan *Hybrid* (abstraktif dan ekstraktif), sedangkan pada penelitian ini sifatnya ekstraktif. Kemudian dataset yang digunakan penelitian rujukan berasal dari CNN/Daily Mail dan DUC2002, sedangkan dalam penelitian ini dari *IndoSum*.

Penelitian dilakukan oleh (Juna & Hayaty, 2023) yang berjudul “*The Observed Preprocessing Strategies For Doing Automatic Text Summarizing*”. Penelitian tersebut membahas tentang strategi *preprocessing* yang digunakan untuk meningkatkan kualitas ringkasan teks otomatis (ATS) menggunakan model *BERT*. Penelitian tersebut juga menggunakan 16 eksperimen yang berbeda untuk mengevaluasi dampak dari berbagai teknik *preprocessing*. Data yang digunakan dalam penelitian yakni data berita berbahasa Indonesia sebanyak 3762 artikel. Hasil penelitian didapatkan dengan menggabungkan proses *data cleaning* dan *case folding* menunjukkan bahwa *ROUGE-1* : 0.78, *ROUGE-2* : 0.60, dan *ROUGE-L* : 0.68. Perbedaan penelitian yakni penelitian rujukan berfokus pada perbandingan berbagai teknik *preprocessing* yang digunakan guna menghasilkan ringkasan teks menggunakan metode *BERT*, sedangkan pada penelitian ini yakni teknik *preprocessing* tetap digunakan tetapi tidak berfokus pada perbandingannya.

Model *BERT* digunakan dalam penelitian ini serta *cosine similarity* digunakan untuk menghitung kemiripan pada dokumen teks, pada penelitian rujukan untuk *cosine similarity* tidak diterapkan, akan tetapi pada penelitian ini *cosine similarity* di terapkan.

Tabel 2. 1 Penelitian terkait

No.	Peneliti	Judul	Metode	Hasil	Perbedaan Penelitian
1.	Mardi Siswo Utomo, Jati Sasongko Wibowo, dan Eko Nur Wahyudi	<i>Text Summarization</i> Pada Artikel Berita Menggunakan <i>Vector Space Model</i> dan <i>Cosine Similarity</i>	Metode <i>Vector Space Model</i> dan <i>Cosine Similarity</i>	Hasilnya yakni hasil uji coba menunjukkan tingkat kompresi 53,67%, dengan ringkasan yang dihasilkan mencerminkan 96,67% informasi utama dari dokumen asli	1. Menggunakan metode <i>Vector Space</i> untuk memberikan bobot nilai pada semua kata pada artikel.
2.	Desi Fatmalasari dan Favorisen Rosyking Lumabanraja	Peringkasan Teks Artikel Ilmiah Berbahasa Indonesia Dengan Metode Pembobotan Kalimat	Metode pembobotan kalimat <i>Terms Frequency-Invers Document Frequency (TF-IDF)</i> dan <i>Cosine Similarity</i>	Berdasarkan hasil pengujian 82.6% responden setuju dengan sistem, 80.2% setuju dengan efisiensi ringkasan yang dihasilkan, dan 83,7% menyatakan puas dengan penggunaan sistem	1. Menggunakan metode pembobotan kalimat (<i>TF-IDF</i>) 2. Data yang digunakan adalah dokumen-dokumen dari jurnal ilmiah khususnya dalam bidang komputasi.
3.	Rahma Hayuning Astuti, Mujiono, dan Sutriawan	<i>Indonesian News Text Summarization Using MBART Algorithm</i>	Algoritma <i>MBART</i>	Hasil dari penelitian tersebut yaitu menghasilkan evaluasi menggunakan metrik <i>ROUGE-1</i> : 35.94, <i>ROUGE-2</i> : 16.43, <i>ROUGE-L</i> : 29.91 yang mencerminkan bahwa kinerja model dalam	1. Menggunakan metode <i>MBART</i> 2. Data yang digunakan adalah dataset <i>XL-SUM</i>

No.	Peneliti	Judul	Metode	Hasil	Perbedaan Penelitian
				menghasilkan ringkasan yang relevan serta informatif	
4.	Franky Halim, Liliana, Kartika Gunadi	Ringkasan Ekstraktif Otomatis pada Berita Berbahasa Indonesia Menggunakan Metode BERT	Metode <i>BERT</i> , metode peringkasan yang membedakan antara 2 metode yakni ekstraktif dan abstraktif	Pada referensi ekstraktif, nilai <i>F-1 Score</i> yang diperoleh adalah sebagai berikut : <i>ROUGE-1</i> : 84.46, <i>ROUGE-2</i> : 83.21, <i>ROUGE-L</i> : 83.40, sedangkan abstraktif hasilnya <i>ROUGE-1</i> : 57.17, <i>ROUGE-2</i> : 51.27, <i>ROUGE-L</i> : 55.20.	<ol style="list-style-type: none"> 1. Tidak menggunakan <i>Cosine Similarity</i> di dalam jurnal tersebut 2. Pendekatan <i>hybrid</i> (abstraktif dan ekstraktif)
5.	Qicai Wang, Peiyu Liu, Zhenfang Zhu, Hongxia Yin, Qiuyue Zhang, dan Lindong Zhang	A Text Abstraction Summary Model Based on BERT Word Embedding and Reinforcement Learning	Model BERT	Hasil akhir model BERT menunjukkan peningkatan apabila dibandingkan dengan model baseline, yakni meningkat <i>ROUGE-1</i> : 1.07%, <i>ROUGE-2</i> : 2.46%, <i>ROUGE-L</i> : 0.95%, <i>R-AVG</i> : 1.46%.	<ol style="list-style-type: none"> 1. Menggunakan pendekatan <i>hybrid</i> (abstraktif dan ekstraktif) 2. <i>Dataset</i> berasal dari CNN/Daily Mail dan DUC2002
6.	Muhammad Farhan Juna, Mardhiya Hayaty	The Observed Preprocessing Strategies For Doing Automatic Text Summarizing	Model BERT	Hasil penelitian menunjukkan bahwa <i>ROUGE-1</i> : 0.78, <i>ROUGE-2</i> : 0.60, dan <i>ROUGE-L</i> : 0.68	<ol style="list-style-type: none"> 1. Berfokus pada perbandingan berbagai teknik <i>preprocessing</i> untuk menghasilkan ringkasan teks. 2. Tidak menggunakan <i>cosine similarity</i>
-	Usulan Penelitian	<i>Extractive Text</i>	Metode BERT dan	Penelitian yang akan dilakukan	1. Menggunakan metode <i>BERT</i>

No.	Peneliti	Judul	Metode	Hasil	Perbedaan Penelitian
		<i>Summarization Of Indonesian News Articles Using BERT And Cosine Similarity</i>	<i>Cosine Similarity</i>		dan <i>Cosine Similarity</i> 2. Menggunakan dataset dari <i>IndoSum (Indonesian Text Summarization)</i>

Berdasarkan penelitian terkait pada tabel 2.1 diatas diketahui bahwa pada penelitian terdahulu menggunakan beberapa metode yakni *Vector Space Model*, metode *Term Frequency-Inverse Document Frequency (TF-IDF)*, metode *MBART*, metode *BERT*. Pada penelitian rujukan, algoritma *BERT* mampu menghasilkan akurasi tinggi saat digunakan dalam pengujian peringkasan teks ekstraktif, hal ini membuat peneliti tertarik untuk menggunakan metode *BERT* untuk peringkasan teks ekstraktif dan juga ditambah dengan *cosine similarity* untuk mengukur tingkat kesamaan kata pada dokumen teks artikel. Pada penelitian usulan untuk mengevaluasi hasil dari peringkasan teks, penulis menggunakan metrik *ROUGE (Recall-Oriented Underresearch for Gisting Evaluastion)* dimana metrik ini digunakan untuk membandingkan kualitas teks yang dihasilkan dengan teks referensi. Ada dua pembaruan yang terdapat pada penelitian ini, yang pertama yaitu penggunaan data berupa dataset dari *Indonesian Text Summarization (Indosum)* dan yang kedua menggunakan metode *Bidirectional Encoder Rrepresentations from Transformers (BERT)* untuk sistem peringkasan teks artikel berita berbahasa Indonesia serta penggunaan *cosine similarity* untuk menghitung kesamaan *cosine* antara vektor kalimat dengan keseluruhan artikel untuk menentukan kalimat mana yang paling representatif.

2.2 *Natural Language Processing (NLP)*

Natural *Language Processing* (NLP) merupakan cabang dari AI yang membantu komputer membaca dan memahami bahasa alami manusia, dengan tujuan meningkatkan komunikasi antara mesin dengan manusia (Apriliani *et al.*, 2023). Pemrosesan bahasa alami (NLP) telah mendapatkan perhatian yang signifikan dalam representasi dan analisis komputasi bahasa manusia. Penerapan konsep ini dapat digunakan di berbagai bidang termasuk ringkasan, terjemah mesin, deteksi spam email, ekstraksi informasi, dsb (Khurana *et al.*, 2023).

Dalam konteks peringkasan teks, NLP memainkan peranan penting dalam menganalisis struktur dan makna dari teks yang akan diringkas. Berbagai teknik NLP, seperti *tokenisasi*, *stemming*, *lemmatization*, *part-of-speech tagging*, dan *named entity recognition* dapat digunakan untuk memahami struktur serta makna pada suatu teks (Oktavia *et al.*, 2024). Salah satu aspek yang penting dalam pemrosesan bahasa alami (NLP) adalah kemampuan untuk melaksanakan analisis sintaksis dan semantik. Analisis sintaksis menitikberatkan pada struktur kalimat, mencakup tata bahasa dan hubungan antar kata, sementara analisis semantik berfokus pada makna yang terkandung dalam kalimat tersebut. Dengan memanfaatkan algoritma dan model statistik, sistem NLP dapat mengidentifikasi pola dalam tata bahasa yang besar, sehingga mampu menghasilkan ringkasan teks, terjemahan, bahkan menjawab pertanyaan berdasarkan teks yang disediakan.

Secara keseluruhan, proses pemrosesan bahasa alami (NLP) bukan hanya tentang memahami bahasa tetapi juga menawarkan solusi untuk masalah komunikasi di era digital ini. Dengan kemajuan teknologi, harapannya NLP akan

terus berkembang dan berkontribusi lebih banyak dalam efisiensi informasi di Indonesia.

2.2.1 Sistem Peringkasan Teks

Salah satu bidang yang semakin berkembang dalam pengolahan bahasa alami (NLP) adalah sistem peringkasan teks. Ringkasan adalah suatu proses yang bertujuan untuk memperpendek teks sumber tanpa menghilangkan informasi penting dan makna dari teks asli. Tujuannya adalah untuk menyampaikan kepada pembaca inti dari pemikiran utama yang terdapat dalam teks tersebut. Peringkasan teks otomatis (*automatic text summarization*) adalah proses pembuatan versi teks yang lebih ringkas dengan memanfaatkan aplikasi yang dijalankan melalui komputer (Ayu Syahfitri *et al.*, 2022).

Peringkasan teks bertujuan untuk menyajikan gambaran yang jelas dan ringkas mengenai isi dokumen, tanpa mengorbankan esensi informasi yang disampaikan. Melalui ringkasan yang efektif, pembaca dapat dengan cepat menangkap inti dari berita tanpa perlu membaca seluruh artikel. Pada peringkasan teks terdapat dua pendekatan yang pertama yakni pendekatan ekstraktif. Pendekatan ini menghasilkan ringkasan dengan cara mengambil kalimat- kalimat penting dari dokumen asli dan menyusunnya menjadi dokumen yang lebih pendek. Yang kedua adalah pendekatan abstraktif. Pendekatan ini menghasilkan ringkasan dimana kalimat yang tersusun tidak ada dalam dokumen aslinya (Widiastutik *et al.*, 2021).

Model peringkasan teks otomatis terbagi menjadi dua kategori. Pertama, ringkasan umum (*generic summary*), yang merupakan representasi dari teks asli

dengan tujuan untuk menyajikan semua fitur penting dari teks tersebut. Pendekatan yang digunakan dalam model ini yakni *bottom up (information retrieval)*. Kedua, ringkasan yang berpusat pada pemakai (*query driven*), dimana peringkasan ini didasarkan pada spesifikasi kebutuhan informasi pengguna, seperti topik atau *query* tertentu, dan mengikuti pendekatan *top down (information extraction)* (Khaqiqi *et al.*, 2022).

2.2.2 Text Preprocessing

Text Preprocessing atau pra-pemrosesan teks merupakan tahapan untuk mempersiapkan data tersebut agar siap untuk analisis lebih lanjut. Tahapan ini sangat penting sebelum memulai suatu penelitian, karena keberhasilan dan kelancaran suatu penelitian sangat bergantung pada minimnya kesalahan dalam proses *text preprocessing* tersebut (Albab *et al.*, 2023). Proses ini dirancang untuk menyiapkan data mentah sehingga dapat dimanfaatkan secara optimal oleh algoritma pemodelan.

Pembersihan teks (*text cleaning*) merupakan tahapan awal dalam *preprocessing* yang berguna untuk menghilangkan *noise* pada *dataset* artikel berita. Tujuan dari *text cleaning* adalah untuk menemukan, memperbaiki, atau menghapus data yang tidak valid maupun yang tidak berguna dari kumpulan data (Salam *et al.*, 2023). Proses dari *text cleaning* diantaranya menghapus karakter khusus serta tanda baca yang tidak relevan seperti (#,\$,%,@, dll), penghapusan digit atau angka yang tidak mempunyai kontribusi pada makna teks, menghapus *whitespace* yang berlebih dan juga baris kosong. Setelah *text cleaning* tahapan selanjutnya ada *case folding*. *Case folding* ini mengubah teks menjadi huruf kecil (*lowercase*) guna

menghindari kata kata yang ambigu serta mengurangi variasi kata. Setelah seluruh rangkaian *preprocessing* dilakukan, maka data siap untuk digunakan pada tahapan berikutnya. Serangkaian proses tersebut dilakukan dalam tahapan *preprocessing* teks dan akan mempengaruhi hasil analisis teks yang dilakukan pada model penelitian dengan sistem NLP.

2.3 Berita

Berita merupakan salah satu bentuk komunikasi yang menyampaikan informasi terbaru mengenai peristiwa, isu, dan perkembangan yang terjadi di tingkat lokal, nasional, maupun internasional. Secara umum, berita dapat diartikan sebagai laporan mengenai fakta atau ide terkini yang memiliki sifat menarik, akurat, atau penting bagi sebagian besar masyarakat . Di Indonesia, berita memainkan peran penting dalam membentuk opini publik serta memberikan pemahaman yang lebih mendalam tentang berbagai permasalahan yang dihadapi oleh masyarakat. Berita tersebut disampaikan melalui media massa, baik media elektronik maupun media cetak. Jenis informasi yang disajikan kepada publik oleh media massa beragam, dengan tujuan untuk menyampaikan informasi yang mendidik serta menghibur (Effendy *et al.*, 2023).

Dalam bidang jurnalistik, berita dikategorikan ke dalam beberapa jenis, termasuk *hard news*, *soft news*, dan *feature*. *Hardnews* umumnya menyajikan menyajikan informasi yang langsung dan faktual mengenai peristiwa terkini, seperti bencana alam atau kecelakaan. Di sisi lain, *soft news* lebih menekankan pada aspek hiburan dan *human interest*, seperti cerita inspiratif atau profil individu. Sementara itu, *feature* merupakan jenis berita yang lebih mendalam, seringkali

mengintegrasikan elemen naratif untuk memberikan konteks yang lebih komprehensif terhadap isu yang di bahas.

Dalam era *digital* saat ini, informasi tidak hanya disampaikan dalam bentuk teks, tetapi juga melalui video, audio, serta infografis. Media sosial telah menjadi salah satu platform utama untuk penyebaran dan konsumsi berita. Interaksi pengguna di media sosial memiliki dampak signifikan terhadap persepsi dan kepercayaan publik terhadap informasi yang disampaikan. Oleh karena itu, pentingnya verifikasi dan pemahaman yang baik terhadap sumber berita, karena tujuan utama dari penyampaian berita adalah memberikan informasi kepada pembaca dan pendengar mengenai peristiwa yang berlangsung di sekitar mereka atau dalam bidang tertentu.

2.4 BERT

Bidirectional Encoder Representation from Transformers (BERT) merupakan suatu teknik pembelajaran mesin yang berbasis pada arsitektur yang dikembangkan oleh *Google*. *BERT* dikembangkan dan diluncurkan pada tahun 2018 oleh *Jacob Devlin* dari *Google* bersama dengan timnya. Tahun 2019, *Google* mengumumkan bahwa mereka telah mulai mengimplementasikan *BERT* dalam mesin pencari dan berencana untuk menerapkan *BERT* pada hampir semua permintaan dalam bahasa Inggris pada akhir tahun 2020 (Putu *et al.*, 2023). Model ini telah menunjukkan performanya yang sangat baik dalam berbagai tugas NLP termasuk peringkasan teks. Bidang pemrosesan bahasa alami (*Natural Language Processing*) telah mengalami kemajuan yang signifikan dalam beberapa tahun terakhir, terutama dengan hadirnya model- model berbasis *transformer* seperti

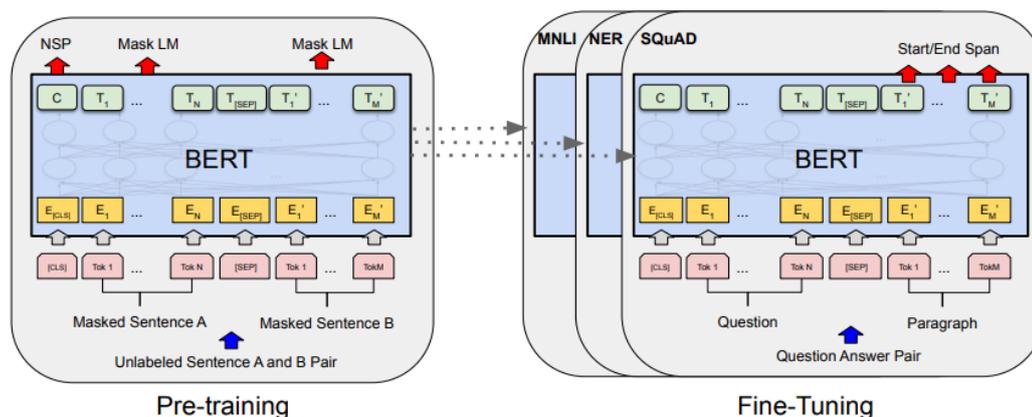
BERT. Model ini dirancang untuk memahami konteks dua arah dalam teks atau pemodelan bahasa *masked*, sehingga dapat menghasilkan representasi yang lebih kaya dan akurat dibandingkan model yang hanya memproses teks dari kiri ke kanan atau sebaliknya serta memiliki kemampuan untuk menangkap hubungan kontekstual antara kata- kata yang terpisah jauh. *BERT* menghasilkan representasi untuk setiap kata dalam kalimat sebagai *output*, serta dapat meningkatkan kinerja model dalam menyelesaikan tugas- tugas berurutan yang kompleks dalam bidang NLP (Vidya Chandradev *et al.*, 2023).

BERT menerapkan teknik *pre-training* dan *fine tuning* dalam proses pelatihannya. Pada tahap *pre- training*, model dilatih melalui dua tugas utama, yakni *Masked Language Model (MLM)* dan *Next Sentence Prediction (NSP)*. Dalam tugas *MLM*, beberapa kata dalam kalimat diacak, dan model ditugaskan untuk memprediksi kata- kata yang hilang tersebut atau melakukan mask terhadap beberapa token dari kalimat input dengan tujuan untuk memprediksi kata yang telah di mask berdasarkan konteks kalimat tersebut.

Pada model *BERT*, input berupa teks kalimat akan diubah menjadi *token sequence*. Terdapat token [CLS] yang berfungsi sebagai penanda awal dari suatu kalimat, serta token [SEP] yang digunakan untuk memisahkan kalimat- kalimat. Selain itu, token [PAD] juga digunakan untuk menambahkan *padding* atau token kosong ke dalam kalimat guna memenuhi *fixed length* dari kalimat yang di inputkan. *Fixed length* pada model *BERT* yakni 512 token. Sementara itu, dalam tugas NSP, model harus menentukan apakah satu kalimat mengikuti kalimat lainnya. Untuk melakukan pelatihan pada representasi *bidirectional*, dilakukan

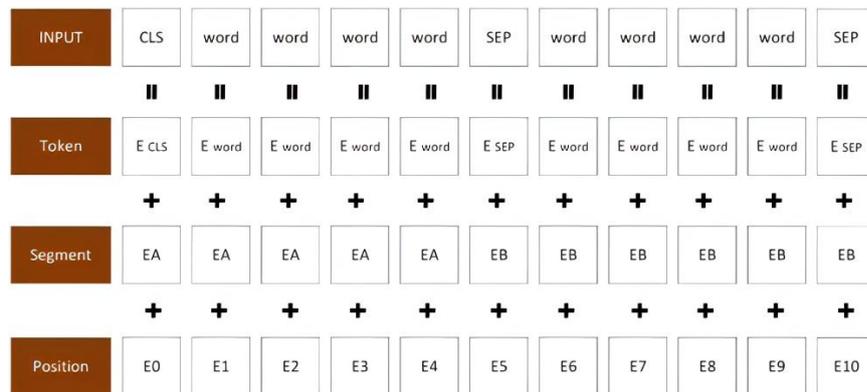
proses *masking* dengan presentase tertentu dari *input* token secara acak. Selanjutnya, dilakukan prediksi terhadap token yang di *mask*, yang dilatih melalui tugas *pre-training* menggunakan *Next Sentence Prediction* (NSP). Proses ini memungkinkan *BERT* untuk mengembangkan pemahaman yang mendalam mengenai struktur dan konteks bahasa (Devlin, Chang, Lee, & Toutanova, 2019).

Setelah menyelesaikan tahap *pre-training*, model dapat di *fine-tune* untuk berbagai tugas spesifik seperti halnya sistem peringkasan teks. Untuk melakukan *fine-tuning* pada model *BERT*, akan dilakukan inisialisasi parameter yang telah di latih sebelumnya. Dalam penerapannya untuk pembuatan ringkasan, terdapat beberapa modifikasi yang dilakukan terhadap model *BERT* asli. Modifikasi ini mencakup penambahan token [CLS] dan [SEP] yang disisipkan di setiap awal dan akhir kalimat. Selanjutnya, apabila panjang input dokumen melebihi batas tetap yang di tetapkan oleh *BERT*, akan dilakukan pemotongan terhadap kalimat berikutnya. Dengan memanfaatkan *BERT*, kita dapat melakukan evaluasi terhadap relevansi kalimat- kalimat dalam artikel berita Indonesia dengan cara yang lebih efisien. *BERT* memungkinkan kita untuk menghitung representasi vektor dari setiap kalimat (Zhang *et al.*, 2020).

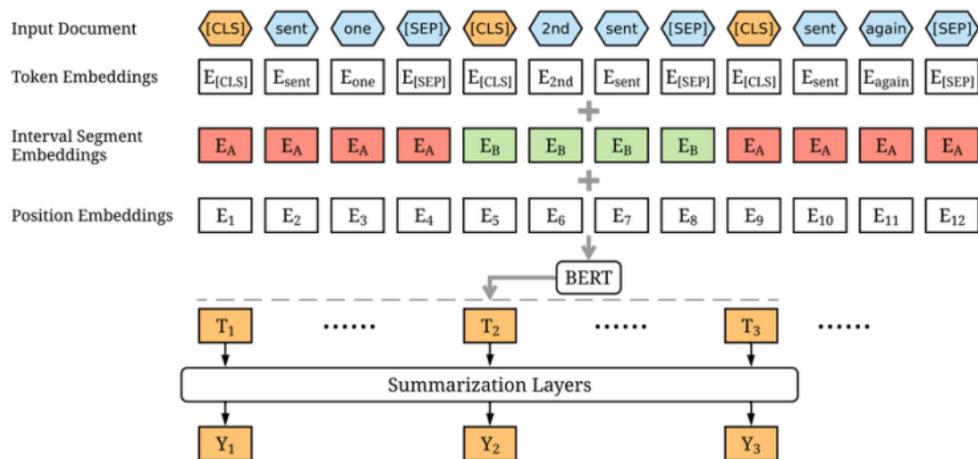


Gambar 2. 1 Gambar *pre-training* dan *fine tuning* dalam *BERT*
 Sumber : (Devlin, Chang, Lee, & Toutanova, 2019)

Input yang digunakan dalam model *BERT* adalah kalimat yang diubah menjadi vektor per kata, yang dikenal sebagai token. Pada model *BERT*, token memiliki representasi token *embeddings* yang berbeda dalam arsitektur *transformer*, yaitu dengan menggunakan *positional embeddings* untuk menunjukkan posisi token dalam urutan. Token pertama dari setiap urutan adalah klasifikasi khusus (CLS), diikuti oleh token dari teks pertama, kemudian token pemisah (SEP), dan diikuti oleh token dari teks kedua, hingga seluruh urutan selesai. Setelah token terbentuk, mereka dimasukkan ke dalam *encoder*. Dalam proses *encoding*, terdapat mekanisme seperti *multi-head attention* dan *add-norm*, serta *feed-forward* dan *add-norm*. *Multi-head attention* berfungsi untuk membantu *encoder* dalam memfokuskan perhatian pada kata tertentu dan memahami konteks keseluruhan dari input yang diberikan (Basbeth & Fudholi, 2024). Gambar 2.2 berikut merupakan arsitektur umum metode *BERT* :



Gambar 2. 2 Arsitektur umum metode *BERT*
 Sumber : (Despotovic *et al.*, 2024)



Gambar 2. 3 Arsitektur *BERT* dalam *text summarization*
 Sumber : (Harinatha *et al.*, 2021)

Pada Gambar 2.3 terdapat dua bagian, yakni pengkode *BERT* dan pengklasifikasi ringkasan. *Encoder* menghasilkan representasi vektor untuk setiap kalimat, yang selanjutnya digunakan oleh klasifikator ringkasan untuk menentukan label pada setiap kalimat. Label ini menunjukkan apakah kalimat tersebut akan dimasukkan ke dalam laporan akhir atau tidak. Dalam hal ini, ditambahkan token [CLS] sebelum setiap kalimat untuk memisahkan kalimat-kalimat tersebut dan mengumpulkan fitur dari kalimat sebelumnya. Setiap kalimat juga diberikan label

Ea jika memiliki jumlah genap, dan Eb jika jumlah kata tersebut ganjil. Selain itu, model ini memberi skor untuk setiap kalimat berdasarkan tingkat kepentingannya. Berdasarkan skor tersebut, ditentukan apakah kalimat- kalimat itu akan dimasukkan ke dalam ringkasan atau tidak.

Dalam konteks berita bahasa Indonesia, dimana terdapat volume informasi yang besar dan bervariasi, *BERT* dapat berperan penting dalam membantu pembaca dalam menyaring informasi yang paling relevan. Contohnya, dalam artikel berita yang membahas mengenai fenomena alam, *BERT* dapat digunakan untuk mengekstrak kalimat- kalimat yang menjelaskan lokasi, waktu serta dampak dari fenomena alam tersebut. Pendekatan ini tidak hanya meningkatkan efisiensi dalam mengakses informasi, tetapi juga mendukung pengambilan keputusan yang lebih cepat dan akurat.

Penerapan model *BERT* dalam peringkasan teks ekstraktif, berikut rumusnya :

$$\text{TokenEmbedding}(t) = \text{WordEmbedding}(t) + \text{PositionEmbedding}(p) + \text{SegmentEmbedding}(s)$$

Keterangan :

WordEmbedding(*t*) : Embedding kata untuk token *t*
 PositionEmbedding(*p*) : Embedding posisi token dalam kalimat
 SegmentEmbedding(*s*) : Embedding yang membedakan segmen kalimat (A dan B)

Selanjutnya, untuk *self-attention* dalam model *BERT* dapat dihitung menggunakan rumus :

$$\text{Attention}(Q, K, V) = \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (2.1)$$

Keterangan :

Q : *Query matrix*

- K : *Key matrix*
 V : *Value matrix*
 d_k : *Dimensi dari key vectors*
 $\sqrt{d_k}$: *Scaling factor* untuk menghindari gradien yang terlalu kecil

Kemudian untuk representasi kalimat dari *output*, terdapat beberapa rumus, sebagai berikut :

CLS Token Representation :

$$S = BERT([CLS] + sentence + [SEP])_0 \quad (2.2)$$

Keterangan :

- S : Vektor representasi kalimat, yang diambil dari *output* posisi ke-0
 [CLS] : Token khusus di awal *input BERT* untuk klasifikasi
 [SEP] : Token pemisah yang menunjukkan akhir dari input kalimat

Selanjutnya *Mean Pooling* :

$$S = \frac{1}{n} \sum_{i=1}^n BERT(sentence)_i \quad (2.3)$$

Keterangan :

- S : Vektor representasi kalimat (hasil mean pooling)
 n : Jumlah token dalam kalimat
 $BERT(sentence)_i$: Vektor *output BERT* pada posisi token ke- i

Untuk menentukan probabilitas ekstraksi kalimat terpilih dalam ringkasan , berikut rumusnya :

$$P(y|s) = \sigma(W \cdot BERT(s) + b) \quad (2.4)$$

Keterangan :

- s : *Kalimat input*
 W : *Matrix bobot*
 b : *Bias*
 y : *Label (1 untuk kalimat yang dipilih, 0 untuk yang tidak)*

2.5 Cosine Similarity

Cosine similarity merupakan suatu ukuran yang digunakan untuk menentukan tingkat kesamaan antara dua vektor dalam ruang dimensi. Secara umum penghitungan menggunakan algoritma ini didasarkan pada ukuran kesamaan ruang vektor yang dinyatakan dalam dua vektor, dengan memanfaatkan kata kunci dari sebuah dokumen sebagai acuan. Keunggulan dari metode *cosine similarity* ini terletak pada kesederhanaannya, efisiensinya, serta mudah untuk dipahami. Selain itu, algoritma *cosine similarity* ini tidak terpengaruh dengan panjang atau pendeknya suatu dokumen, serta kemampuan untuk memberikan tingkat akurasi yang tinggi (Kurniawan & Arif, 2022). Ukuran dari *cosine similarity* ini diperoleh dari nilai *cosinus* sudut yang dihasilkan dari perkalian dua vektor yang sedang dibandingkan. Mengingat bahwa nilai *cosinus* untuk sudut 0° adalah 1, dan sudut lainnya nilainya kurang dari 1, maka dua vektor dianggap memiliki kesamaan yang tinggi apabila nilai *cosine similarity*-nya mencapai 1 (Vendyansyah & Pranoto, 2021). Berikut persamaan untuk perhitungan *cosine similarity* :

$$\text{Cos } \alpha = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (2.5)$$

Keterangan :

- A : Vektor A, yang akan dibandingkan kemiripannya
- B : Vektor B, yang akan dibandingkan kemiripannya
- $A \cdot B$: *Dot product* antara vektor A dan vektor B
- |A| : Panjang vektor A
- |B| : Panjang vektor B
- |A| |B| : *Cross product* antara |A| dan |B|
- A_i : Bobot istilah i pada dokumen A
- B_i : Bobot istilah i pada dokumen B

Dengan A dan B adalah nilai vektor A dan vektor B atau dua vektor kalimat yang akan diukur tingkat kesamaannya, sedangkan $A \cdot B$ menunjukkan *dot product* antara setiap komponen (bobot kata) dari vektor A dan Vektor B . Untuk $|A|$ adalah panjang vektor A , dan untuk $|B|$ adalah panjang vektor B (R. Kurniawan & Arif, 2022). Nilai *cosine* diperoleh dengan membagi antara (*dot product*) dengan (*cross product*) dari vektor.

2.6 Recall-Oriented Understudy For Gisting Evaluation (ROUGE)

ROUGE (*Recall Oriented Understudy for Gisting Evaluation*) merupakan suatu metode evaluasi yang dilakukan dengan menghitung jumlah *n-gram* yang identik antara ringkasan yang dihasilkan oleh sistem dan ringkasan yang dibuat secara manual (Widiastutik *et al.*, 2021). ROUGE digunakan untuk mengukur tingkat kesamaan antara ringkasan yang dihasilkan oleh sistem dan ringkasan referensi yang disusun oleh manusia. Dengan demikian, ROUGE memberikan gambaran yang komprehensif mengenai efektivitas kinerja sistem tersebut.

Terdapat beberapa varian dari metrik ROUGE, yang paling umum digunakan adalah ROUGE-N, ROUGE-L, dan ROUGE-W. ROUGE-N mengukur kesamaan berdasarkan *n-gram*, yakni urutan kata yang diambil dari teks. Sebagai contoh, ROUGE-1 mengukur kesamaan berdasarkan *unigram* (kata tunggal), sedangkan ROUGE-2 mengukur kesamaan berdasarkan *bigram* (dua kata berturut-turut). Disisi lain, ROUGE-L mengukur kesamaan berdasarkan panjang *sub sequence* terpanjang yang berurutan dalam teks. Pendekatan ini memberikan keunggulan dalam memahami konteks serta struktur kalimat dalam ringkasan (Lin & Rey, 2004).

Penelitian ini menggunakan *ROUGE-N*, dan *ROUGE L*, yang terdiri atas *precision*, *recall*, dan *f-measure* sebagai *scoring* untuk mengukur kinerja sistem.

Berikut rumusnya :

$$ROUGE - 1 recall = \frac{\text{Jumlah unigram kata yang sama}}{\text{Total kata di ringkasan manual}} \quad (2.6)$$

$$ROUGE - 1 precision = \frac{\text{Jumlah unigram kata yang sama}}{\text{Total kata di ringkasan sistem}} \quad (2.7)$$

$$ROUGE - 2 recall = \frac{\text{Jumlah bigram kata yang sama}}{\text{Total kata di ringkasan manual}} \quad (2.8)$$

$$ROUGE - 2 precision = \frac{\text{Jumlah bigram kata yang sama}}{\text{Total kata di ringkasan sistem}} \quad (2.9)$$

Persamaan di atas menunjukkan *ROUGE-N* yang membandingkan *n-gram* jumlah kata yang sesuai antara ringkasan sistem dan ringkasan manual. Jumlah *n-gram* yang digunakan bermacam-macam, pada umumnya yakni $n=\{1,\dots,4\}$. Akan tetapi, mayoritas yang digunakan dalam penelitian adalah *n-gram* dengan jumlah $n=1$ (*ROUGE-1*) dan $n=2$ (*ROUGE-2*).

Untuk *ROUGE-L* yakni mencocokkan seluruh susunan kata terpanjang yang sama yang dikenal dengan *longest common subsequence (LCS)* antara ringkasan sistem dan ringkasan manual. Berikut rumus untuk menghitung *ROUGE-L* :

$$ROUGE - L recall = \frac{LCS(\text{sistem},\text{manual})}{\text{Total kata di ringkasan manual}} \quad (2.10)$$

$$ROUGE - L precision = \frac{LCS(\text{sistem},\text{manual})}{\text{Total kata di ringkasan sistem}} \quad (2.11)$$

Sedangkan untuk *F-Measure* merupakan hasil kolaborasi antara nilai *recall* dan *precision* untuk menghitung akurasi suatu sistem. Komposisi yang berimbang

antara *precision* dan *recall* disebut dengan *F-1 Measure* (Halimah et al., 2022), berikut rumus untuk menghitungnya :

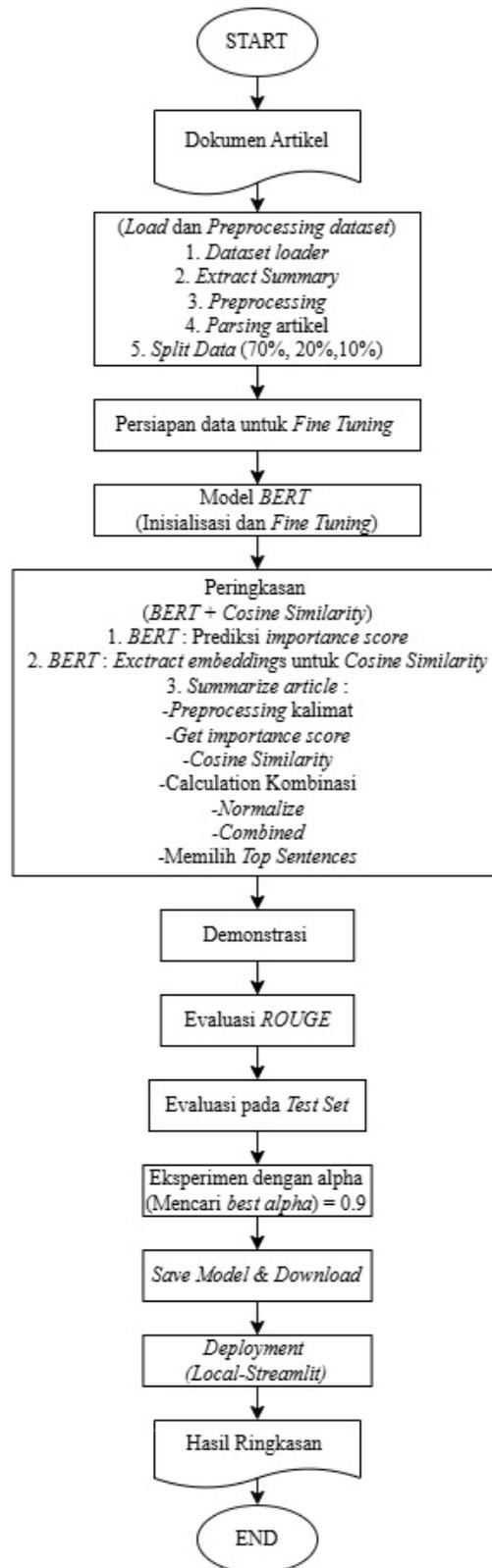
$$F1 - Measure = 2 * \frac{precision*recall}{precision+recall} \quad (2.12)$$

BAB III

DESAIN DAN IMPLEMENTASI

3.1 Desain Sistem

Pada tahapan ini merupakan alur dari sistem yang akan di buat. Dimulai dari *input* data hingga menghasilkan *output* yang sesuai dengan tujuan penelitian. Dalam melakukan peringkasan teks secara ekstraktif, gambaran akan keseluruhan sistem akan di jelaskan pada penelitian ini. Berikut desain sistem yang akan di implementasikan pada penelitian ini digambarkan pada gambar 3.1 di bawah ini :



Gambar 3. 1 Desain sistem peringkasan teks ekstraktif menggunakan *BERT* dan *cosine similarity*

Desain sistem ini menggambarkan alur kerja sistem peringkasan teks ekstraktif pada artikel berita berbahasa Indonesia menggunakan metode *BERT* dan *Cosine Similarity*. Tahapan awal dalam penelitian ini adalah adanya dataset dokumen artikel yakni dataset indosum. Kemudian tahap *load* serta *preprocessing* dataset. Pada tahapan ini, sistem melakukan pembacaan dataset untuk membaca file “*train.03.jsonl*” dari *Indosum*, kemudian untuk *extract summary* untuk mengekstrak ringkasan referensi, pada *preprocessing* yang dilakukan adalah *text cleaning* dan *case folding*, untuk parsing artikel dilakukan untuk memecah dokumen menjadi sebuah kalimat- kalimat, kemudian untuk *split data* atau pembagian data dilakukan dengan presentasi 70% *data training*, 20 % *data validation*, dan 10% adalah *data testing*.

Tahapan selanjutnya setelah data terpreparasi adalah tahapan persiapan data untuk *fine tuning*. Selanjutnya model *BERT* dilakukan inisialisasi kemudian dilakukan proses *fine-tuning* menggunakan *pre-trained IndoBERT*. *Fine tuning* bertujuan melatih model dalam mengklasifikasikan *importance score* pada tiap kalimat, dan menghasilkan model yang sudah disesuaikan dengan karakteristik artikel berita bahasa Indonesia. Selanjutnya tahapan inti yakni proses peringkasan yang menggunakan metode *BERT* dan *Cosine Similarity*. Dalam tahap ini, *BERT* bertugas untuk memprediksi *importance score* pada tiap kalimat berdasarkan pemahaman semantic, sementara itu untuk *cosine similarity* yakni menghitung kemiripan antara kalimat menggunakan *embeddings* untuk menjaga koherensi. Proses dimulai dengan *preprocessing* kalimat input, dilanjutkan dengan mengekstrak *importance score* dari *BERT* dan perhitungan *similarity* matrik. Kedua

skor ini kemudian dikombinasikan. Kemudian sistem melakukan normalisasi skor, menggabungkan kedua metode, serta memilih *top sentence* yang berdasarkan perankingan skor final. Hasil dari ringkasan kemudian di demonstrasikan, kemudian dilanjutkan dengan evaluasi metrik *ROUGE* (*ROUGE-1*, *ROUGE-2*, dan *ROUGE-L*) untuk mengukur kualitas ringkasan dibandingkan dengan referensi. Selanjutnya untuk evaluasi dilakukan pada *test set* yang terdiri dari 200 artikel guna mendapatkan performa keseluruhan sistem. Selanjutnya, untuk eksperimen dengan parameter *alpha* dilakukan untuk menemukan kombinasi optimal antara *BERT* dan *cosine similarity*. Setelah mendapatkan parameter yang optimal, model disimpan dan di download untuk di lakukan *deployment*.

Proses *deployment* dilakukan secara lokal menggunakan *Streamlit* untuk mengakses sistem secara *interface*. Sistem menghasilkan ringkasan ekstraktif dari artikel berita yang di input yang telah melalui seluruh *pipeline processing*. Kemudian *outputnya* berupa hasil ringkasan yang siap digunakan.

3.2 Pengumpulan Data

Penelitian yang berjudul “*Exctractive Text Summarization Of Indonesia News Articles Using BERT and Cosine Similarity*” ini menggunakan *dataset* dari *IndoSum (Indonesian Text Summarization) update* terbaru tahun 2022 (K. Kurniawan & Louvan, 2018). *IndoSum* merupakan dataset berita berbahasa Indonesia yang dihasilkan dari penelitian yang dilakukan oleh Kurniawan dan Louvan. *Dataset* ini dapat dipergunakan sebagai acuan dalam pengembangan sistem ringkasan otomatis berbahasa Indonesia. Format *dataset* ini yakni JSON

serta minimal mencakup sebanyak 18.774 artikel berita yang diambil dari berbagai portal berita, termasuk CNN Indonesia, kumparan, dan portal berita lainnya.

Data yang digunakan dalam penelitian ini sebanyak 2000 data artikel berita berbahasa Indonesia. Presentasi data yang digunakan adalah 70% data pelatihan (1400 artikel), 20 % data validasi (400 artikel), dan 10% data testing (200 artikel). Dalam dataset ini terdapat beberapa elemen penting, diantaranya kategori yang menunjukkan jenis berita, *gold label* yang berfungsi sebagai label ringkasan untuk peringkasan ekstraktif, ID unik untuk setiap berita, sumber berita, URL sumber, dan paragraf yang berisi teks dari berita (K. Kurniawan & Louvan, 2018). Ringkasan tersebut di kerjakan secara manual oleh dua orang penutur asli Bahasa Indonesia. Dataset *Indosum* sendiri terdiri dari *5 fold cross validation* dimana data dibagi menjadi 5 kumpulan data *training*, *development*, serta *testing*. Setiap kumpulan data yang ada terdiri dari 3 file *.jsonl* yakni *file train*, *dev*, dan *test*. Salah satu file pada *dataset* yaitu *train.03.jsonl* yang jumlah beritanya 14290 dan ada 7 fitur. Berikut fitur- fitur pada dataset *Indosum* :

Tabel 3. 1 Fitur- fitur *dataset*

Fitur	Penjelasan
<i>category</i>	Kategori artikel tersebut
<i>gold_labels</i>	Label ekstraktif kalimat- kalimat dalam artikel tersebut
<i>id</i>	Pengenal unik di setiap artikel
<i>paragraphs</i>	Daftar paragraf dari artikel asli
<i>source</i>	Sumber artikel
<i>source_url</i>	Url artikel yang asli
<i>summary</i>	Ringkasan manual pada artikel yang ada dalam daftar kalimat

Pada Tabel 3.1 ada 7 fitur yang ada pada dataset *IndoSum*, akan tetapi terdapat 2 fitur utama yang digunakan, yakni *paragraphs* dan *summary*. Berikut

Berikut contoh dari hasil *parsing* :

Tabel 3. 3 Tabel contoh hasil *parsing*

No	Kalimat	Label	Gold Label	Alasan
1	Kebakaran pagi ini terjadi di kawasan Pasir Koneng , Kedoya Utara , Jakarta Barat .	1	<i>True</i>	Info utama: lokasi kejadian
2	18 Unit Damkar dikerahkan untuk memadamkan api .	1	<i>True</i>	Info penting: respons penanganan
3	Informasi dari Sudin Damkar Jakarta Barat , Kamis (13 / 7) , kebakaran terjadi sekitar pukul 04.45 WIB .	0	<i>False</i>	Detail tambahan: sumber & waktu
4	Kebakaran ini melalap beberapa unit rumah .	1	<i>True</i>	Info penting: dampak kejadian
5	Hingga saat ini proses pemadaman masih berlangsung .	0	<i>False</i>	Info temporal yang bisa berubah
6	Belum diketahui kerugian , penyebab hingga korban jiwa dalam kebakaran tersebut .	1	<i>True</i>	Info penting: status korban
7	Warga sekitar diimbau untuk menjauhi lokasi kebakaran , hingga api berhasil dipadamkan .	0	<i>False</i>	Himbauan umum

Keterangan :

Paragraf 1:

- Kalimat 1: *True* → Label = 1 (Penting untuk ringkasan)
- Kalimat 2: *True* → Label = 1 (Penting untuk ringkasan)

Paragraf 2:

- Kalimat 3: *False* → Label = 0 (Tidak penting)
- Kalimat 4: *True* → Label = 1 (Penting untuk ringkasan)

Paragraf 3:

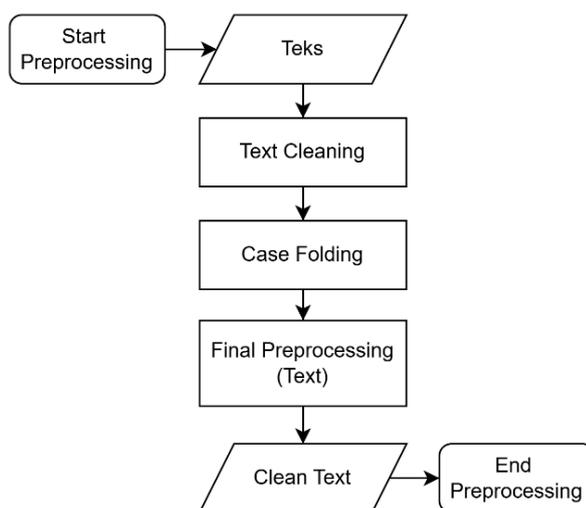
- Kalimat 5: *False* → Label = 0 (Tidak penting)
- Kalimat 6: *True* → Label = 1 (Penting untuk ringkasan)

Paragraf 4:

- Kalimat 7: *False* → Label = 0 (Tidak penting)

3.4 Preprocessing

Tahapan awal dalam pemrosesan ringkasan teks adalah *preprocessing*. Tahap *preprocessing* merupakan langkah yang penting untuk mempersiapkan teks agar siap digunakan dalam model yang menghasilkan *indeks term* dari dokumen teks yang akan digunakan pada tahap berikutnya. Berikut tahapan *preprocessing* yang akan di implementasikan pada penelitian ini :



Gambar 3. 2 Flowchart *preprocessing*

Pada gambar 3.2 diatas merupakan *flowchart* dari tahapan *preprocessing*, dimana dalam hal ini *preprocessing* yang digunakan terdiri dari *text cleaning* dan *case folding*.

3.4.1 Text Cleaning (Pembersihan Teks)

Text cleaning (pembersihan teks) adalah tahapan yang bertujuan untuk menghapus karakter khusus, tanda baca, serta elemen yang tidak relevan pada teks artikel berita.

Tabel 3. 4 Contoh proses *text cleaning*

Sebelum <i>text cleaning</i>	[[["Kebakaran", "pagi", "ini", "terjadi", "di", "kawasan", "Pasir", "Koneng", ",", "Kedoya", "Utara", ",", "Jakarta", "Barat", "."], ["18", "Unit", "Damkar", "dikerahkan", "untuk", "memadamkan", "api", "."]], [{"Informasi", "dari", "Sudin", "Damkar", "Jakarta", "Barat", ",", "Kamis", "("}, {"13", "/", "7", ")"}, {"kebakaran", "terjadi", "sekitar", "pukul", "04.45", "WIB", "."}], [{"Kebakaran", "ini", "melalap", "beberapa", "unit", "rumah", "."}], [{"Hingga", "saat", "ini", "proses", "pemadaman", "masih", "berlangsung", "."}], [{"Belum", "diketahui", "kerugian", ",", "penyebab", "hingga", "korban", "jiwa", "dalam", "kebakaran", "tersebut", "."}], [{"Warga", "sekitar", "diimbau", "untuk", "menjauhi", "lokasi", "kebakaran", ",", "hingga", "api", "berhasil", "dipadamkan", "."}]]
Setelah <i>text cleaning</i>	[[["Kebakaran", "pagi", "ini", "terjadi", "di", "kawasan", "Pasir", "Koneng", "Kedoya", "Utara", "Jakarta", "Barat"], ["18", "Unit", "Damkar", "dikerahkan", "untuk", "memadamkan", "api"]], [["Informasi", "dari", "Sudin", "Damkar", "Jakarta", "Barat", "Kamis", "13", "7", "kebakaran", "terjadi", "sekitar", "pukul", "0445", "WIB"], [{"Kebakaran", "ini", "melalap", "beberapa", "unit", "rumah"}]], [["Hingga", "saat", "ini", "proses", "pemadaman", "masih", "berlangsung"], [{"Belum", "diketahui", "kerugian", "penyebab", "hingga", "korban", "jiwa", "dalam", "kebakaran", "tersebut"}]], [["Warga", "sekitar", "diimbau", "untuk", "menjauhi", "lokasi", "kebakaran", "hingga", "api", "berhasil", "dipadamkan"}]]

3.4.2 Case Folding

Tahapan setelah *text cleaning* (pembersihan teks) adalah *case folding*. *Case folding* merupakan tahapan dimana semua huruf dalam dokumen diubah menjadi huruf kecil. Proses ini bertujuan untuk menghindari perbedaan makna pada saat terkandung huruf kapital di dalam kata tersebut. Hasilnya yang berupa daftar kalimat dengan *list* token di setiap kalimat, akan diubah seluruh tokennya menjadi huruf kecil. Kemudian hasil dari tahapan ini yakni berupa *list* token yang telah menjadi huruf kecil semuanya. Berikut contoh proses dari *case folding* dapat dilihat pada tabel 3.3 :

Tabel 3. 5 Contoh proses *case folding*

Sebelum <i>case folding</i>	[[["Kebakaran", "pagi", "ini", "terjadi", "di", "kawasan", "Pasir", "Koneng", "Kedoya", "Utara", "Jakarta", "Barat"], ["18", "Unit", "Damkar", "dikerahkan", "untuk", "memadamkan", "api"]], [["Informasi", "dari", "Sudin", "Damkar", "Jakarta", "Barat", "Kamis", "13", "7", "kebakaran", "terjadi", "sekitar", "pukul", "0445", "WIB"], ["Kebakaran", "ini", "melalap", "beberapa", "unit", "rumah"]], [["Hingga", "saat", "ini", "proses", "pemadaman", "masih", "berlangsung"], ["Belum", "diketahui", "kerugian", "penyebab", "hingga", "korban", "jiwa", "dalam", "kebakaran", "tersebut"]], [["Warga", "sekitar", "diimbau", "untuk", "menjauhi", "lokasi", "kebakaran", "hingga", "api", "berhasil", "dipadamkan"]]]
Setelah <i>case folding</i>	[[["kebakaran", "pagi", "ini", "terjadi", "di", "kawasan", "pasir", "koneng", "kedoya", "utara", "jakarta", "barat"], ["18", "unit", "damkar", "dikerahkan", "untuk", "memadamkan", "api"]], [["informasi", "dari", "sudin", "damkar", "jakarta", "barat", "kamis", "13", "7", "kebakaran", "terjadi", "sekitar", "pukul", "0445", "wib"], ["kebakaran", "ini", "melalap", "beberapa", "unit", "rumah"]], [["hingga", "saat", "ini", "proses", "pemadaman", "masih", "berlangsung"], ["belum", "diketahui", "kerugian", "penyebab", "hingga", "korban", "jiwa", "dalam", "kebakaran", "tersebut"]], [["warga", "sekitar", "diimbau", "untuk", "menjauhi", "lokasi", "kebakaran", "hingga", "api", "berhasil", "dipadamkan"]]]

Berikut contoh hasil *preprocessing*-nya :

Tabel 3. 6 Tabel contoh hasil *preprocessing*

No	Kalimat Asli	Setelah <i>Cleaning</i>	Setelah <i>Case Folding</i>
1	Kebakaran pagi ini terjadi di kawasan Pasir Koneng , Kedoya Utara , Jakarta Barat .	Warga sekitar diimbau untuk menjauhi lokasi kebakaran , hingga api berhasil dipadamkan .	kebakaran pagi ini terjadi di kawasan pasir koneng , kedoya utara , jakarta barat .
2	18 Unit Damkar dikerahkan untuk memadamkan api .	18 Unit Damkar dikerahkan untuk memadamkan api .	18 unit damkar dikerahkan untuk memadamkan api .
3	Informasi dari Sudin Damkar Jakarta Barat , Kamis (13 / 7) , kebakaran terjadi sekitar pukul 04.45 WIB .	Informasi dari Sudin Damkar Jakarta Barat , Kamis (13 / 7) , kebakaran terjadi sekitar pukul 04.45 WIB .	informasi dari sudin damkar jakarta barat , kamis (13 / 7) , kebakaran terjadi sekitar pukul 04.45 wib .
4	Kebakaran ini melalap beberapa unit rumah .	Kebakaran ini melalap beberapa unit rumah .	kebakaran ini melalap beberapa unit rumah .
5	Hingga saat ini proses pemadaman masih berlangsung .	Hingga saat ini proses pemadaman masih berlangsung .	hingga saat ini proses pemadaman masih berlangsung .

No	Kalimat Asli	Setelah <i>Cleaning</i>	Setelah <i>Case Folding</i>
6	Belum diketahui kerugian , penyebab hingga korban jiwa dalam kebakaran tersebut .	Belum diketahui kerugian , penyebab hingga korban jiwa dalam kebakaran tersebut .	belum diketahui kerugian , penyebab hingga korban jiwa dalam kebakaran tersebut .
7	Warga sekitar diimbau untuk menjauhi lokasi kebakaran , hingga api berhasil dipadamkan .	Warga sekitar diimbau untuk menjauhi lokasi kebakaran , hingga api berhasil dipadamkan .	warga sekitar diimbau untuk menjauhi lokasi kebakaran , hingga api berhasil dipadamkan .

3.5 Pembagian Data

Setelah pelabelan kalimat selesai, *dataset* yang telah diberi *label* dapat digunakan untuk tahap berikutnya. Tahapan ini adalah pembagian data menjadi data latih, data validasi, dan data uji. Data latih (*training data*) digunakan untuk melatih evaluasi model peringkasan teks. Data validasi (*validation data*) digunakan untuk mengevaluasi dan menyetel model selama pelatihan. Dan data pengujian (*testing data*) digunakan setelah proses pelatihan selesai untuk mengevaluasi performa akhir model. Berikut contoh pembagian data menggunakan proporsi umum data pelatihan 70%, data validasi 20%, dan data pengujian 10%.

3.6 BERT *Tokenization*

Pada tahapan *BERT Tokenization* ini menggunakan *WordPiece Tokenization*. BERT menggunakan *WordPiece Tokenization* yang memecah kata menjadi sub-kata guna menangani kata-kata yang jarang muncul (*OOV- Out of Vocabulary*). Proses *tokenization* yaitu pertama menambahkan token khusus [CLS] diawal dan [SEP] diakhir, kedua memecah kata menjadi sub-kata dengan prefix ## untuk sub kata lanjutan, ketiga melakukan konversi token ID berdasarkan vocabulary, keempat *paddingnya* sampai *max length* 128. Berikut contoh *tokenization* pada semua kalimat :

Tabel 3. 7 Tabel contoh *tokenization BERT*

No	Kalimat	Tokens	Token IDs	Length
K1	kebakaran pagi ini terjadi di kawasan pasir koneng , kedoya utara , jakarta barat .	[CLS] ke ##bara ##kan pagi ini terjadi di kawasan pasir kon ##eng , ke ##doya utara , jakarta barat . [SEP]	[101, 1176, 2878, 1489, 6790, 1179, 4678, 1301, 5412, 7008, 3600, 1998, 117, 1176, 9457, 7994, 117, 4251, 3003, 119, 102]	21
K2	18 unit damkar dikerahkan untuk memadamkan api .	[CLS] 18 unit dam ##kar di ##kerah ##kan untuk me ##mada ##m ##kan api . [SEP]	[101, 2466, 3988, 6000, 7654, 1301, 12757, 1489, 1334, 1143, 23456, 131, 1489, 9630, 119, 102]	16
K3	informasi dari sudin damkar jakarta barat , Kamis (13 / 7) , kebakaran terjadi sekitar pukul 04.45 wib .	[CLS] informasi dari sud ##in dam ##kar jakarta barat , Kamis (13 / 7) , ke ##bara ##kan terjadi sekitar pukul 04 . 45 wi ##b . [SEP]	[101, 5869, 1330, 9554, 1199, 6000, 7654, 4251, 3003, 117, 10976, 113, 1492, 120, 128, 114, 117, 1176, 2878, 1489, 4678, 7310, 9040, 1827, 119, 3546, 7178, 1830, 119, 102]	30
K4	kebakaran ini melalap beberapa unit rumah	[CLS] ke ##bara ##kan ini me ##lala ##p beberapa unit rumah . [SEP]	[101, 1176, 2878, 1489, 1179, 1143, 14372, 1643, 5312, 3988, 7029, 119, 102]	13
K5	hingga saat ini proses pemadaman masih berlangsung .	[CLS] hingga saat ini proses pe ##mada ##man masih berlangsung . [SEP]	[101, 6021, 6416, 1179, 6599, 1609, 23456, 1399, 5023, 11213, 119, 102]	12
K6	belum diketahui kerugian , penyebab hingga korban jiwa dalam kebakaran tersebut .	[CLS] belum di ##ketahui ke ##rugi ##an , penyebab hingga korban jiwa dalam ke ##bara ##kan tersebut . [SEP]	[101, 4222, 1301, 23100, 1176, 21719, 1389, 117, 15187, 6021, 10843, 9386, 1633, 1176, 2878, 1489, 6680, 119, 102]	19
K7	warga sekitar diimbau untuk menjauhi lokasi kebakaran , hingga api berhasil dipadamkan	[CLS] warga sekitar di ##imba ##u untuk men ##jau ##hi lokasi ke ##bara ##kan , hingga api berhasil di ##pada ##m ##kan . [SEP]	[101, 10148, 7310, 1301, 23654, 131, 1334, 1386, 28901, 3766, 11934, 1176, 2878, 1489, 117, 6021, 9630, 8938, 1301, 9851, 131, 1489, 119, 102]	24

Padding dan Attention Mask :

Berikut contoh pada Kalimat K5, dengan *max length* 12

Input IDs = [101, 6021, 6416, 1179, 6599, 1609, 23456, 1399, 5023, 11213, 119, 102] + [0] * 116

Attention mask = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1] + [0] * 116

3.7 Fine-Tuning Model BERT

Tahapan setelah dilakukan pemasangan *input-label* adalah *fine-tuning* model BERT. Dalam hal ini *training configuration*-nya *epoch* yakni 12, *batch size* 32, *learning rate* 3e-5, *max length* 128. *Fine-tuning* dimulai dengan inisialisasi menggunakan *pretrained IndoBERT weights*, dilanjutkan dengan *loading* data menggunakan *stratified sampling* untuk menjaga proporsi *label*. Tiap iterasi *training* melakukan *forward pass* dari *input* melalui BERT hingga *classification head* guna menghasilkan *logits*, kemudian menghitung *weighted BCE loss*, dan melakukan *backward pass* untuk kalkulasi gradien kemudian melakukan update parameter menggunakan *optimizer*. Untuk akhir per *epoch*, dilakukan evaluasi pada *validation set* untuk melihat progress performa model. Selanjutnya, untuk menghindari *overfitting*, dilakukan *early stopping* untuk menghentikan *training* jika *validation loss* tidak menurun selama 3 *epoch* berturut-turut, *dropout rate* 0.1 pada BERT layers, *gradient clipping* dengan *max norm* 1.0, dan *learning rate decay* yang menurunkan *learning rate* secara bertahap selama proses *training*.

Berikut rumus untuk *loss function-Binary Cross Entropy* :

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (3.1)$$

Keterangan :

- L : Nilai total *loss* dari seluruh sampel
- N : Jumlah total sampel (kalimat) dalam batch
- y_i : Label sebenarnya dari sampel ke- i , bernilai 0 atau 1
- \hat{y}_i : Probabilitas prediksi dari model bahwa sampel ke- i bernilai 1
- Log : Logaritma natural (basis e)

Binary Cross Entropy digunakan untuk mengukur seberapa baik prediksi probabilitas model terhadap label biner.

3.8 Evaluasi Model

Evaluasi model digunakan untuk mengukur kemampuan model dalam mengklasifikasi kalimat penting (1) atau tidak penting (0) untuk selanjutnya dimasukkan dalam ringkasan. Evaluasi dilakukan pada validation set di setiap *epoch* yang berguna untuk memantau *progress training* serta menentukan kapan model mencapai performa yang optimal. Dasar dari evaluasi yakni *confusion matrix* 2 x 2 yang terdiri dari *True Positif* (TP) yaitu kalimat penting yang di prediksi sebagai kalimat penting, *True Negative* (TN) yakni kalimat tidak penting yang di prediksi tidak penting, *False Positive* (FP) yakni kalimat tidak penting yang diprediksi penting, dan *False Negative* (FN) yakni kalimat penting yang diprediksi tidak penting.

Metrik evaluasi yang digunakan meliputi *accuracy* yang mengukur prediksi benar dari total prediksi $(TP + TN) / (TP + TN + FP + FN)$. Untuk *precision* yakni mengukur ketepatan prediksi kelas positif $TP / (TP + FP)$. *Recall* yakni mengukur kelengkapan deteksi kelas positif dengan rumus $TP / (TP + FN)$. *F1-Score* sebagai *harmonic mean* dari *precision* dan *recall* dengan rumus $2 * (precision * recall) / (precision + recall)$. Setelah model *BERT* sudah selesai dan optimal maka di *save*. *Checkpoint* model disimpan setiap kali *validation F1-score* meningkat, kemudian model terbaik digunakan untuk tahapan selanjutnya dalam peringkasan.

3.9 BERT Importance Score

BERT *Importance Score* merupakan probabilitas yang berarti suatu kalimat diklasifikasikan sebagai kalimat “penting” (kelas 1) dan dimasukkan dalam ringkasan. *Score* ini diperoleh dari *fine-tuning* BERT yang telah dilakukan *training* untuk *binary classification*.

Rumus *Linear Transformation* :

$$z = Wx + b \quad (3.2)$$

Keterangan :

z = logits (z0, z1)
 W = weight matrix
 x = BERT output vector
 b = bias vector

Rumus *Softmax Function* :

$$P(y = k|x) = \frac{e^{z_k}}{\sum_{j=0}^1 e^{z_j}} = \frac{e^{z_k}}{e^{z_0} + e^{z_1}} \quad (3.3)$$

Keterangan :

$P(y = k|x)$: Probabilitas kelas k diberikan input x
 z_k : Skor logit (*output linear*) untuk kelas k
 e : Bilangan Euler

Rumus *Importance Score* :

$$Importance\ Score = P(y = 1|x) = \frac{e^{z_k}}{e^{z_0} + e^{z_1}} \quad (3.4)$$

Keterangan :

Importance Score: Probabilitas bahwa $y=1$ diberikan input x
 z_0, z_1 : logits atau skor *output* model sebelum fungsi *softmax*
 e : Bilangan Euler

Berikut dibawah ini merupakan contoh perhitungan *importance score* setiap kalimat :

Contoh Kalimat 1 : "kebakaran pagi ini terjadi di kawasan pasir koneng , kedoya utara , jakarta barat ."

Tabel 3. 8 Hasil contoh BERT *importance score*

Kalimat	Logits	Softmax	Importance Score
K1	[-1.234, 1.567]	[0.057, 0.943]	0.943
K2	[-1.789, 2.123]	[0.028, 0.972]	0.972
K3	[0.234, -0.456]	[0.639, 0.361]	0.361
K4	[-0.987, 1.234]	[0.123, 0.877]	0.877
K5	[0.567, -0.234]	[0.703, 0.297]	0.297
K6	[-0.654, 0.876]	[0.213, 0.787]	0.787
K7	[0.789, -0.345]	[0.762, 0.238]	0.238

3.10 BERT Embeddings dan Cosine Similarity

Tahapan ini mengekstrak CLS *Token Embeddings* (768 dimensi), akan tetapi untuk simplifikasi digunakan 5 dimensi. *Output* dari BERT adalah *tensor* dengan *shape* [*batch size*, *sequence length*, 768]. Dalam tahapan ini digunakan contoh untuk 5 dimensi :

Dalam representasinya 768-D ke 5-D :

K1 *full* = [0.234, -0.156, 0.789, -0.345, 0.567, 0.123, -0.456, ..., 0.891]

768 values

K1 *simplified* = [0.234, -0.156, 0.789, -0.345, 0.567] 5 value

Berikut *embeddings* semua kalimat (5 dimensi)

K1: [0.234, -0.156, 0.789, -0.345, 0.567]

K2: [0.345, -0.234, 0.678, -0.456, 0.789]

K3: [0.456, -0.345, 0.567, -0.567, 0.234]

K4: [0.567, -0.456, 0.456, -0.678, 0.345]

K5: [0.678, -0.567, 0.345, -0.789, 0.456]

K6: [0.789, -0.678, 0.234, -0.234, 0.567]

K7: [0.234, -0.789, 0.123, -0.345, 0.678]

Rumus *Cosine Similarity* :

$$\text{cosine}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^d A_i B_i}{\sqrt{\sum_{i=1}^d A_i^2} \cdot \sqrt{\sum_{i=1}^d B_i^2}} \quad (3.5)$$

Berikut perhitungan *similarity matrix* nya :

Tabel 3. 9 Contoh perhitungan *similarity matrix*

	K1	K2	K3	K4	K5	K6	K7
K1	1.000	0.972	0.856	0.923	0.678	0.812	0.734
K2	0.972	1.000	0.823	0.891	0.712	0.845	0.756
K3	0.856	0.823	1.000	0.934	0.789	0.867	0.823
K4	0.923	0.891	0.934	1.000	0.834	0.912	0.867
K5	0.678	0.712	0.789	0.834	1.000	0.756	0.812
K6	0.812	0.845	0.867	0.912	0.756	1.000	0.889
K7	0.734	0.756	0.823	0.867	0.812	0.889	1.000

Kemudian untuk *averagenya* berikut hasilnya pada per kalimat :

$$\text{K1: } (0.972+0.856+0.923+0.678+0.812+0.734)/6 = 0.829$$

$$\text{K2: } (0.972+0.823+0.891+0.712+0.845+0.756)/6 = 0.833$$

$$\text{K3: } (0.856+0.823+0.934+0.789+0.867+0.823)/6 = 0.849$$

$$\text{K4: } (0.923+0.891+0.934+0.834+0.912+0.867)/6 = 0.894$$

$$\text{K5: } (0.678+0.712+0.789+0.834+0.756+0.812)/6 = 0.764$$

$$\text{K6: } (0.812+0.845+0.867+0.912+0.756+0.889)/6 = 0.847$$

$$\text{K7: } (0.734+0.756+0.823+0.867+0.812+0.889)/6 = 0.814$$

Berikut perhitungan *normalization* :

$$\text{min} = 0.764, \text{max} = 0.894, \text{range} = 0.130$$

$$\text{K1}_{\text{norm}} = (0.829 - 0.764) / 0.130 = 0.500$$

$$\text{K2}_{\text{norm}} = (0.833 - 0.764) / 0.130 = 0.531$$

$$\text{K3}_{\text{norm}} = (0.849 - 0.764) / 0.130 = 0.654$$

$$K4_norm = (0.894 - 0.764) / 0.130 = 1.000$$

$$K5_norm = (0.764 - 0.764) / 0.130 = 0.000$$

$$K6_norm = (0.847 - 0.764) / 0.130 = 0.638$$

$$K7_norm = (0.814 - 0.764) / 0.130 = 0.385$$

3.11 Combined Score Calculation

Selanjutnya yakni menghitung kombinasi skor pada *BERT* dan *Cosine Similarity* nya :

$$Score_{final} = \alpha . Score_{BERT} + (1 - \alpha) . Score_{Cosine} \quad (3.6)$$

Pada perhitungan *combined score*, dilakukan dengan $\alpha = 0.9$, dan berikut tabel hasil perhitungannya :

Tabel 3. 10 Tabel contoh skor hasil kombinasi *BERT* dan *cosine similarity*

Kalimat	<i>BERT Score</i>	<i>Cosine Norm</i>	<i>Calculation</i>	<i>Final Score</i>
K1	0.943	0.500	$(0.9 \times 0.943) + (0.1 \times 0.500)$	$0.849 + 0.050 = 0.899$
K2	0.972	0.531	$(0.9 \times 0.972) + (0.1 \times 0.531)$	$0.875 + 0.053 = 0.928$
K3	0.361	0.654	$(0.9 \times 0.361) + (0.1 \times 0.654)$	$0.325 + 0.065 = 0.390$
K4	0.877	1.000	$(0.9 \times 0.877) + (0.1 \times 1.000)$	$0.789 + 0.100 = 0.889$
K5	0.297	0.000	$0.9 \times 0.297 + 0.1 \times 0.000$	$0.267 + 0.000 = 0.267$
K6	0.787	0.638	$0.9 \times 0.787 + 0.1 \times 0.638$	$0.708 + 0.064 = 0.772$
K7	0.238	0.385	$0.9 \times 0.238 + 0.1 \times 0.385$	$0.214 + 0.039 = 0.253$

3.12 Sentence Selection

Kemudian setelah menghitung hasil dari kombinasi *BERT* dan *Cosine Similarity*, selanjutnya adalah *sentence selection*. Berikut perhitungannya :

Summary ratio = 0.3

Sentences to select = $\max(2, \text{int}(7 \times 0.3))$
 = $\max(2, 2) = 2$

Perankingan *Final Score* :

1. K2 (0.928): "18 unit damkar dikerahkan untuk memadamkan api ."
2. K1 (0.899): "kebakaran pagi ini terjadi di kawasan pasir koneng , kedoya utara , jakarta barat ."
3. K4 (0.889): "kebakaran ini melalap beberapa unit rumah ."
4. K6 (0.772): "belum diketahui kerugian , penyebab hingga korban jiwa dalam kebakaran tersebut ."
5. K3 (0.390): "informasi dari sudin damkar jakarta barat , kamis (13 / 7) , kebakaran terjadi sekitar pukul 04.45 wib ."
6. K5 (0.267): "hingga saat ini proses pemadaman masih berlangsung ."
7. K7 (0.253): "warga sekitar diimbau untuk menjauhi lokasi kebakaran , hingga api berhasil dipadamkan ."

Seleksi kalimat : ada 2 kalimat yang dipilih

1. K1: "Kebakaran pagi ini terjadi di kawasan Pasir Koneng , Kedoya Utara , Jakarta Barat ."
2. K2: "18 Unit Damkar dikerahkan untuk memadamkan api ."

Generate summary :

"Kebakaran pagi ini terjadi di kawasan Pasir Koneng , Kedoya Utara , Jakarta Barat . 18 Unit Damkar dikerahkan untuk memadamkan api ."

3.13 Evaluasi Ringkasan (*ROUGE Evaluation*)

Pada tahapan ini melakukan evaluasi ringkasan menggunakan metrik *ROUGE*, berikut contoh perhitungannya :

Reference Summary : "Kebakaran pagi ini terjadi di kawasan Pasir Koneng , Kedoya Utara , Jakarta Barat . 18 Unit Damkar dikerahkan untuk memadamkan api . Kebakaran terjadi sekitar pukul 04.45 WIB . Kebakaran ini melalap beberapa unit rumah . Belum diketahui kerugian , penyebab hingga korban jiwa dalam kebakaran tersebut ."

1. *ROUGE-1 (Unigram Overlap)* :

$$ROUGE - 1 = \frac{Count_{match}(unigram)}{Count_{reference}(unigram)} \quad (3.7)$$

Reference unigrams:

[kebakaran, pagi, ini, terjadi, di, kawasan, pasir, koneng, kedoya, utara, jakarta, barat, 18, unit, damkar, dikerahkan, untuk, memadamkan, api, kebakaran, terjadi, sekitar, pukul, 04.45, wib, kebakaran, ini, melalap, beberapa, unit, rumah, belum, diketahui, kerugian, penyebab, hingga, korban, jiwa, dalam, kebakaran, tersebut]

System unigrams:

[kebakaran, pagi, ini, terjadi, di, kawasan, pasir, koneng, kedoya, utara, jakarta, barat, 18, unit, damkar, dikerahkan, untuk, memadamkan, api]

Unique reference = 35 tokens

Unique system = 19 tokens

Overlap = 19 tokens

ROUGE-1 *Recall* = $19/35 = 0.543$

ROUGE-1 *Precision* = $19/19 = 1.000$

ROUGE-1 F1 = $2 \times (0.543 \times 1.000) / (0.543 + 1.000) = 0.704$

2. ROUGE-2 (Bigram Overlap) :

$$ROUGE - 2 = \frac{Count_{match}(bigram)}{Count_{reference}(bigram)} \quad (3.8)$$

Reference bigrams = 40 bigrams

System bigrams = 18 bigrams

Overlap = 17 bigrams

Reference bigrams (40 total) :

[kebakaran pagi], [pagi ini], [ini terjadi], [terjadi di], [di kawasan], [kawasan pasir], [pasir koneng], [koneng kedoya], [kedoya utara], [utara jakarta], [jakarta barat], [18 unit], [unit damkar], [damkar dikerahkan], [dikerahkan untuk], [untuk memadamkan], [memadamkan api], [kebakaran terjadi], [terjadi sekitar], [sekitar pukul], [pukul 04.45], [04.45 wib], [kebakaran ini], [ini melalap], [melalap beberapa], [beberapa unit], [unit rumah], [belum diketahui], [diketahui kerugian], [kerugian penyebab], [penyebab hingga], [hingga korban], [korban jiwa], [jiwa dalam], [dalam kebakaran], [kebakaran tersebut]

System bigrams (18 total) :

[kebakaran pagi], [pagi ini], [ini terjadi], [terjadi di], [di kawasan], [kawasan pasir], [pasir koneng], [koneng kedoya], [kedoya utara], [utara jakarta], [jakarta barat], [18

unit], [unit damkar], [damkar dikerahkan], [dikerahkan untuk], [untuk memadamkan], [memadamkan api]

Matching bigrams = 17 (semua kecuali bigram terakhir)

ROUGE-2 Recall = $17/40 = 0.425$

ROUGE-2 Precision = $17/18 = 0.944$

ROUGE-2 F1 = $2 \times (0.425 \times 0.944) / (0.425 + 0.944) = 0.586$

3. ROUGE-L (Longest Common Subsequence)

$$ROUGE - L = \frac{LCS(Reference, System)}{Lenght(Reference)} \quad (3.9)$$

Reference tokens = 41 tokens

System tokens = 19 tokens

LCS length = 19

ROUGE-L Recall = $19/41 = 0.463$

ROUGE-L Precision = $19/19 = 1.000$

ROUGE-L F1 = $2 \times (0.463 \times 1.000) / (0.463 + 1.000) = 0.633$

3.14 Contoh Hasil *Summary* (Contoh Hasil Ringkasan)

Pada sub bab ini yakni menjelaskan tentang contoh hasil ringkasan yang di hasilkan oleh sistem peringkasan teks ekstraktif menggunakan metode *BERT* dan *Cosine Similarity*, berikut contoh hasil ringkasannya :

Tabel 3. 11 Contoh hasil ringkasan

Jenis	Kalimat	Jumlah Kata
Artikel Asli	Kebakaran pagi ini terjadi di kawasan Pasir Koneng , Kedoya Utara , Jakarta Barat . 18 Unit Damkar dikerahkan untuk memadamkan api . Informasi dari Sudin Damkar Jakarta Barat , Kamis (13 / 7) , kebakaran terjadi sekitar pukul	84

Jenis	Kalimat	Jumlah Kata
	04.45 WIB . Kebakaran ini melalap beberapa unit rumah . Hingga saat ini proses pemadaman masih berlangsung . Belum diketahui kerugian , penyebab hingga korban jiwa dalam kebakaran tersebut . Warga sekitar diimbau untuk menjauhi lokasi kebakaran , hingga api berhasil dipadamkan .	
Ringkasan Sistem	Kebakaran pagi ini terjadi di kawasan Pasir Koneng , Kedoya Utara , Jakarta Barat . 18 Unit Damkar dikerahkan untuk memadamkan api .	22
Ringkasan Referensi	Kebakaran pagi ini terjadi di kawasan Pasir Koneng , Kedoya Utara , Jakarta Barat . 18 Unit Damkar dikerahkan untuk memadamkan api . Kebakaran terjadi sekitar pukul 04.45 WIB . Kebakaran ini melalap beberapa unit rumah . Belum diketahui kerugian , penyebab hingga korban jiwa dalam kebakaran tersebut .	48

Compression Rate :

$$\text{Sistem} = (84 - 22) / 84 \times 100\% = 73.8\%$$

$$\text{Referensi} = (84 - 48) / 84 \times 100\% = 42.9\%$$

Tabel 3. 12 *ROUGE score* contoh hasil ringkasan

Metrik	<i>Recall</i>	<i>Precision</i>	<i>F1-Score</i>
ROUGE-1	0.543	1.000	0.704
ROUGE-2	0.425	0.944	0.586
ROUGE-L	0.463	1.000	0.633

Analisis hasil pada contoh hasil ringkasan diatas yakni sistem berhasil mengidentifikasi 2 kalimat paling penting, ringkasan mencakup informasi kunci yakni lokasi kejadian dan respons penanganan, kemudian didapatkan compression rate sebesar 73.8 %. Performa model *score* dominan 90 % memberikan hasil yang baik, dalam hal ini *cosine similarity* memperkuat keterkaitan antar kalimatnya, model juga dapat menunjukkan kemampuan yang membedakan antara kalimat penting atau tambahannya. Untuk evaluasi *ROUGE*, *ROUGE-1 F1*: 0.704 (baik untuk overlap kata), *ROUGE-2 F1*: 0.586 (cukup baik untuk *overlap* frasa), *ROUGE-L F1*: 0.633 (struktur kalimat terpreservasi).

BAB IV

HASIL DAN PEMBAHASAN

4.1 Skenario Uji Coba

Skenario uji coba pada penelitian ini yakni untuk mengevaluasi performa sistem peringkasan teks ekstraktif. Pengujian yang dilakukan melalui beberapa tahapan untuk mengidentifikasi konfigurasi optimal dan memahami karakteristik sistem. Skenario ini mencakup pengujian terhadap parameter-parameter yang mempengaruhi kualitas pada ringkasan, performa berdasarkan kategori artikel, dan evaluasi kombinasi metode terhadap hasil akhir.

4.1.1 Dataset Uji Coba

Penelitian ini menggunakan dataset *Indosum*. *Indosum* merupakan dataset peringkasan teks berita berbahasa Indonesia. *Dataset* yang digunakan dalam penelitian ini yakni 2000 artikel berita yang diambil dari file “*train.03.jsonl*”. *Split data* atau pembagian data dilakukan dengan presentasi 70% *data training* (1400 artikel), 20% *data validation* (400 artikel), dan 10% *data testing* (200 artikel). Total seluruh dataset yakni 35.928 kalimat, yakni distribusi 25.207 kalimat pada *training*, 7.139 kalimat pada *data validation*, serta 3.582 kalimat pada *data test*.

Tabel 4. 1 Pembagian *dataset*

Pembagian Data	Jumlah Artikel	Presentase	Jumlah Kalimat
<i>Training</i>	1400	70%	25.207
<i>Validation</i>	400	20%	7.139
<i>Test</i>	200	10%	3.582
Total	2000	100%	35.928

Tabel 4. 2 Distribusi kategori pada *test set*

Kategori	Jumlah Artikel	Presentase
Tajuk Utama	70	39.5%
Olahraga	52	26.0%
Showbiz	29	14.5%
Teknologi	23	11.5%
Hiburan	17	8.5%

Distribusi kategori pada *test set* menunjukkan bahwa variasi yang cukup beragam dengan tajuk utama yang mendominasi 79 artikel (39.5%), kemudian olahraga dengan 52 artikel (26.0%), selanjutnya showbiz yakni 29 artikel (14.5%), kemudian untuk teknologi 23 artikel (11.5%), serta hiburan 17 artikel (8.5%).

Uji coba dilakukan dalam dua lingkungan berbeda untuk memastikan sistem dapat berjalan dengan baik. Tahapan pelatihan model dilakukan menggunakan *Google Colaboratory* yang menyediakan akses GPU NVIDIA A100-SXM-40 GB dengan memory sebesar 42.5 GB. Berikut spesifikasi lengkap lingkungan pelatihan:

Tabel 4. 3 Spesifikasi lingkungan pelatihan

Komponen	Spesifikasi
<i>Platform</i>	<i>Google Colaboratory</i>
GPU	NVIDIA A100-SXM4-40GB
GPU Memory	42.5 GB
<i>Python Version</i>	3.11
<i>PyTorch Version</i>	2.6.0 + cu124
<i>Transformers</i>	4.52.4
CUDA Support	Yes

4.1.2 Parameter Uji Coba

Pada parameter uji coba seperti pada tabel berikut :

Tabel 4. 4 Skenario parameter uji coba

Skenario	<i>Alpha</i>	<i>Ratio</i>	Keterangan
S1	1.0	0.3	<i>BERT</i>
S2	0.0	0.3	<i>Cosine</i>
S3	0.5	0.3	<i>Balanced</i>
S4	0.7	0.3	Dominan <i>BERT</i> (70:30)
S5	0.9	0.3	Dominan <i>BERT</i> (90:10)
S6	0.9	0.2	Ringkasan Pendek
S7	0.9	0.4	Ringkasan Panjang

Uji coba dilakukan dengan tujuh skenario parameter yang berbeda untuk mengevaluasi pengaruh nilai *alpha* (α) dan *ratio* terhadap performa sistem. pada skenario pertama menggunakan $\alpha=1.0$ yang menunjukkan hanya menggunakan *BERT* saja. Selanjutnya, skenario kedua dengan $\alpha=0.0$ yakni untuk menguji performa *Cosine Similarity* saja. Skenario 3 hingga 5 yakni untuk menguji berbagai kombinasi dengan $\alpha=0.5$ (*balanced*), $\alpha=0.7$ (dominan *BERT* 70:30), dan $\alpha=0.9$ (dominan *BERT* 90:10). Seluruh skenario ini menggunakan *ratio* 0.3 yang berarti memilih 30% kalimat teratas sebagai ringkasan. Kemudian untuk skenario 6 dan 7 secara khusus menguji pengaruh *ratio* dengan menggunakan α optimal (0.9) namun dengan *ratio* berbeda yaitu 0.2 untuk ringkasan lebih pendek dan 0.4 untuk ringkasan lebih panjang.

Tabel 4. 5 Parameter uji coba sistem

Parameter	Nilai	Deskripsi
<i>Epochs</i>	12	Jumlah iterasi pelatihan
<i>Batch Size</i>	32	Ukuran batch untuk training
<i>Learning Rate</i>	3e-5	Laju pembelajaran
<i>Max Length</i>	128	Panjang maksimal token
<i>Summary Ratio</i>	0.1 - 0.5	Persentase kalimat yang dipilih
<i>Alpha (α)</i>	0.5 - 0.9	Bobot <i>BERT</i> vs <i>Cosine Similarity</i>
<i>Min Sentences</i>	2	Minimal kalimat dalam ringkasan
<i>Max Sentences</i>	7	Maksimal kalimat dalam ringkasan

4.1.3 Metrik Evaluasi

Evaluasi performa sistem menggunakan metrik *ROUGE* (*Recall-Oriented Understudy for Gisting Evaluation*) yang merupakan standart evaluasi untuk peringkasan teks. Pada *ROUGE-1* yakni digunakan untuk mengukur *overlap unigram* antara ringkasan referensi dengan ringkasan sistem, serta memberikan gambaran tentang seberapa baik sistem menangkap kata-kata penting. *ROUGE-2*

mengukur *overlap bigram* yang menunjukkan kemampuan sistem dalam mempertahankan frasa dan urutan kata yang tepat. Untuk *ROUGE-L* menggunakan *longest common subsequence* untuk mengevaluasi *fluency* serta koherensi secara keseluruhan.

4.2 Hasil Uji Coba

Pada hasil uji coba menunjukkan temuan- temuan dari semua tahapan yang dilakukan dalam penelitian ini. Dalam hasil uji coba di hasilkan yakni hasil dari *fine-tuning* model *BERT*, performa sistem dengan berbagai konfigurasi parameter, evaluasi pada dataset, serta analisis kategori artikel berita. Untuk memudahkan interpretasi, maka setiap hasil akan di sertakan data berupa tabel. Hasil uji coba akan menjadi dasar pada bagian pembahasan tentang peringkasan teks ekstraktif menggunakan metode *BERT* dan *Cosine Similarity*.

4.2.1 Hasil *Fine Tuning BERT*

Proses *fine tuning* model *IndoBERT* merupakan tahapan yang sangat penting dalam membangun suatu sistem peringkasan teks ekstraktif. Model *base IndoBERT* yang telah di latih pada *corpus* bahasa Indonesia akan diadaptasi secara spesifik untuk dilakukan klasifikasi *importance* kalimat dalam artikel berita. Hasil uji coba yakni pada proses *fine-tuning* dapat diselesaikan dalam waktu 8.8 menit dengan hasil sebagai berikut :

Tabel 4. 6 Hasil *training* per *epoch*

<i>Epoch</i>	<i>Training Loss</i>	<i>Validation Loss</i>	<i>Accuracy</i>	<i>F1-Score</i>	<i>Precision</i>	<i>Recall</i>
1	0.4232	0.3942	0.8403	0.4031	0.7347	0.2778
2	0.3773	0.3994	0.8349	0.4725	0.6219	0.3810
3	0.3556	0.4251	0.8078	0.5110	0.5049	0.5173

<i>Epoch</i>	<i>Training Loss</i>	<i>Validation Loss</i>	<i>Accuracy</i>	<i>F1-Score</i>	<i>Precision</i>	<i>Recall</i>
4	0.2433	0.5550	0.7608	0.4831	0.4161	0.5758
5	0.1458	0.6203	0.7897	0.4765	0.4612	0.4928
6	0.0841	0.8283	0.7918	0.4678	0.4644	0.4711
7	0.0563	1.0055	0.8046	0.4769	0.4965	0.4589
8	0.0534	1.1828	0.7605	0.4595	0.4089	0.5245
9	0.0292	1.1478	0.7833	0.4630	0.4462	0.4812
10	0.0318	1.3492	0.7714	0.4796	0.4297	0.5426
11	0.0109	1.4073	0.7938	0.4758	0.4698	0.4820
12	0.0146	1.3801	0.7937	0.4793	0.4699	0.4892

Berdasarkan tabel 4.3, *training loss* menurun konsisten dari 0.42 menuju ke 0,01. Sementara itu, *validation loss* awalnya stabil sampai *epoch* 3 yakni 0.42. Kemudian secara drastis sampai 1.38 pada *epoch* 12 yang berarti *overfitting*, dalam hal ini titik optimal berada di *epoch* 3. Berdasarkan hasil training, performa optimal di capai pada *epoch* 3 dengan *F1-score* 0.5110. Setelah *epoch* 3, terjadi *overfitting* yang ditandai dengan *training loss* terus turun (0.0146) sementara itu *validatin loss* naik secara drastis (1.3801). Setelah *epoch* 3 tidak ada peningkatan yang signifikan, *epoch* 3 memberikan keseimbangan yang optimal antara *precision* (0.5049) dan *recall* (0.5173).

4.2.2 Hasil Eksperimen Parameter *Alpha*

Setelah dilakukan *fine-tuning* model *BERT*, kemudian dilakukan eksperimen parameter *alpha*. Berikut hasil eksperimen pada parameter *alpha*, berikut hasilnya :

Tabel 4. 7 Eksperimen parameter *alpha*

<i>Alpha (a)</i>	<i>BERT Weight</i>	<i>Cosine Weight</i>	<i>ROUGE-1</i>	<i>ROUGE-2</i>	<i>ROUGE-L</i>
0.0	0%	100%	0.486	0.387	0.441
0.5	50%	50%	0.571	0.482	0.532
0.6	60%	40%	0.574	0.487	0.536
0.7	70%	30%	0.565	0.478	0.527
0.8	80%	20%	0.564	0.476	0.525

<i>Alpha (α)</i>	<i>BERT Weight</i>	<i>Cosine Weight</i>	<i>ROUGE-1</i>	<i>ROUGE-2</i>	<i>ROUGE-L</i>
0.9	90%	10%	0.577	0.491	0.540
1.0	100%	0%	0.523	0.432	0.482

Pada tabel 4.7, eksperimen terhadap nilai *alpha* menunjukkan hasil yang baik tentang kontribusi masing-masing komponen. Ketika sistem hanya menggunakan *Cosine Similarity* ($\alpha=0.0$), *ROUGE-1 score* yang dicapai senilai 0.486. Performa meningkat seiring dengan peningkatan kontribusi *BERT*, mencapai puncaknya pada $\alpha=0.9$ dengan *ROUGE-1 score* 0.577. Akan tetapi, saat sistem sepenuhnya pada *BERT* ($\alpha=1.0$), performa justru menurun menjadi 0.523. Temuan ini membuktikan bahwa kombinasi kedua metode memberikan hasil yang baik dibandingkan penggunaan salah satu metode saja. Nilai $\alpha=0.9$ menghasilkan performa terbaik tidak hanya pada *ROUGE-1*, tetapi juga pada *ROUGE-2* (0.491) dan *ROUGE-L* (0.540).

4.2.3 Hasil Evaluasi pada *Testing Data*

Tahapan selanjutnya, setelah dilakukan eksperimen pada parameter *alpha* dan di hasilkan *alpha* yang optimal. Tahapan selanjutnya yakni mengevaluasi data *testing*. Berikut hasil evaluasi pada data *testing* :

Tabel 4. 8 Hasil evaluasi *testing data*

Metrik	Mean	Standar Deviasi	Min	Max
<i>ROUGE-1</i>	0.537	0.155	0.182	0.923
<i>ROUGE-2</i>	0.456	0.191	0.095	0.897
<i>ROUGE-L</i>	0.499	0.178	0.156	0.912

Evaluasi yang dilakukan pada 200 artikel *test set* dengan menggunakan parameter yang optimal ($\alpha=0.9$, *ratio*=0.3) menghasilkan performa yang konsisten

dan baik serta menjanjikan. Sistem mencapai rata-rata *ROUGE-1 score* 0.537 dengan *standar deviasi* 0.155, menunjukkan performa yang relatif stabil meskipun ada variasi antar artikel. Kemudian untuk *ROUGE-2* mencapai rata-rata 0.456 (std 0.191) serta *ROUGE-L* 0.499 (std 0.178). Pada distribusi skor menunjukkan pola normal dengan *slight right skew*, di mana mayoritas artikel mendapat skor *ROUGE-1* antara 0.4 sampai 0.7. Nilai minimum *ROUGE-1* yakni 0.182 sementara untuk maksimum mencapai 0.923, menunjukan bahwa performa sistem sangat bergantung pada karakteristik artikel.

4.2.4 Hasil Berdasarkan Kategori

Setelah tahapan sebelumnya melakukan evaluasi pada *data testing*, maka tahapan berikutnya yakni melakukan analisis performa artikel pada kategori per artikel. Berikut hasilnya berdasarkan kategori :

Tabel 4. 9 Performa per kategori artikel

Kategori	Jumlah	<i>ROUGE-1</i>	<i>ROUGE-2</i>	<i>ROUGE-L</i>
Olahraga	52	0.574 (± 0.137)	0.511	0.550
Tajuk Utama	79	0.561 (± 0.165)	0.479	0.524
Showbiz	29	0.497 (± 0.148)	0.419	0.462
Teknologi	23	0.470 (± 0.153)	0.385	0.422
Hiburan	17	0.467 (± 0.127)	0.338	0.400

Analisis performa berdasarkan kategori menunjukkan perbedaan yang signifikan antar domain berita. Kategori olahraga menunjukkan performa terbaik dengan *ROUGE-1 score* 0.574 (± 0.137), selanjutnya oleh tajuk utama dengan 0.561 (± 0.165). Kategori showbiz mencapai 0.497 (± 0.148), sementara teknologi dan hiburan menunjukkan performa terendah dengan masing-masing 0.470 (± 0.153) dan 0.467 (± 0.127). Perbedaan performa ini dapat dijelaskan oleh karakteristik

masing-masing kategori, di mana artikel olahraga cenderung memiliki struktur yang lebih faktual dan kronologis sehingga memudahkan identifikasi kalimat penting.

4.2.5 Hasil Eksperimen *Ratio*

Tahapan sebelumnya yakni analisis performa artikel yang dilakukan per kategori artikel, tahapan selanjutnya yakni eksperimen terhadap *ratio*, dan berikut hasil dari eksperimen berdasarkan *ratio* :

Tabel 4. 10 Hasil eksperimen *ratio*

<i>Ratio</i>	<i>Avg Sentences</i>	<i>ROUGE-1</i>	<i>ROUGE2</i>	<i>ROUGE-L</i>	<i>Compression</i>
0.2	2.5	0.498	0.412	0.459	80%
0.3	3.8	0.537	0.456	0.499	70%
0.4	5.0	0.542	0.461	0.503	60%
0.5	6.3	0.531	0.448	0.491	50%

Pada tabel diatas eksperimen terhadap nilai *ratio* menunjukkan bahwa pemilihan *ratio* 0.3 memberikan *balance* optimal antara kelengkapan informasi dan keringkasan. Dengan *ratio* 0.2, sistem rata-rata memilih 2.5 kalimat per artikel dengan *ROUGE-1 score* 0.498, mengindikasikan hilangnya beberapa informasi penting. *Ratio* 0.3 menghasilkan rata-rata 3.8 kalimat dengan performa terbaik (*ROUGE-1*: 0.537). Peningkatan *ratio* menjadi 0.4 dan 0.5 tidak memberikan improvement signifikan pada performa, bahkan cenderung menurun karena meningkatnya redundansi informasi. *Compression rate* 70% (*ratio* 0.3) terbukti ideal untuk menjaga keseimbangan antara *informativeness* dan *conciseness*.

4.2.6 Hasil Peringkasan

Tahapan selanjutnya unntuk memberikan gambaran tentang *output* sistem, berikut beberapa hasil ringkasan dari beberapa kategori dengan tingkat performa yang berbeda. Berikut hasilnya :

a. Kategori Olahraga

Rimanews - Tim balap sepeda yang dipersiapkan untuk SEA Games 2017 Malaysia , pada 19 - 31 Agustus bakal dikirim ke Belanda dan Belgia guna memaksimalkan kemampuan sebelum berlaga pada kejuaraan multi event terbesar di Asia Tenggara itu . Ketua Umum Pengurus Besar Ikatan Sepeda Sport Indonesia (PB ISSI) Raja Sapta Oktohari , mengatakan , dari dua negara yang bakal dituju baru Belanda yang sudah dipastikan menjadi tuan rumah training center (TC) pebalap - pebalap Indonesia itu . " Yang di Belanda khusus untuk road race dan track putri saja . Sedangkan yang putra di Belgia . Namun , untuk yang Belgia masih dalam tahap pembicaraan , " kata Okto di Jakarta , Jumat (7 / 4 / 2017) . Menurut dia , TC bagi pebalap yang dipersiapkan untuk SEA Games 2017 memang harus secepatnya dilakukan mengingat pelaksanaan kejuaraan dua tahunan itu sudah tidak lama lagi . Hal ini dilakukan agar target untuk meraih hasil maksimal bisa tercapai . Sebenarnya , ada negara lain yang sebelumnya diputuskan untuk menjadi lokasi TC . Ada dua tempat yang siap untuk menjadi tuan rumah yaitu Hongkong dan China . Hanya saja , lokasi tersebut berpeluang basa diubah sesuai dengan kebutuhan program latihan . Selama ini , timnas Indonesia untuk nomor road race berlatih di Yogyakarta , sedangkan untuk nomor track menjalani pemusatan latihan di velodrome Manahan , Solo , Jawa Tengah . Rencana menggelar TC di Belanda juga dibenarkan oleh Direktur Performa Tinggi Satlak Prima Deny Gumulya . Menurut dia ada beberapa pertimbangan sehingga TC harus dilakukan di luar negeri yang salah satunya untuk mengejar kemampuan pebalap . Selama TC di Solo dan Yogyakarta , kata dia , memang terdapat kendala terutama masalah cuaca sehingga kondisi tersebut berpengaruh kepada fisik atlet . Selain itu , terutama di Yogyakarta , pebalap terkendala dengan debu saat menjalani latihan . " Akibatnya banyak yang terserang flu . Makanya kami akan segera mengirim dokter Satlak Prima untuk melihat secara langsung kondisi atlet termasuk memberikan vaksin , " katanya . Sedangkan untuk kondisi di Solo , mantan pebalap nasional itu menjelaskan jika velodrome Manahan tidak sesuai dengan standart internasional sehingga ia cukup kesulitan untuk meningkatkan catatan waktu . Deny menegaskan , idealnya pebalap diberangkat ke TC di Eropa pada akhir Mei . Hal tersebut disesuaikan dengan program yang ada serta waktu pelaksanaan SEA Games 2017 Malaysia . Apalagi Indonesia dituntut meningkatkan prestasi dibandingkan hasil SEA Games 2015 .

Gambar 4. 1 Teks asli artikel kategori olahraga

Rimanews - Tim balap sepeda yang dipersiapkan untuk SEA Games 2017 Malaysia , pada 19 - 31 Agustus bakal dikirim ke Belanda dan Belgia guna memaksimalkan kemampuan sebelum berlaga pada kejuaraan multi event terbesar di Asia Tenggara itu. Ketua Umum Pengurus Besar Ikatan Sepeda Sport Indonesia (PB ISSI) Raja Sapta Oktohari , mengatakan , dari dua negara yang bakal dituju baru Belanda yang sudah dipastikan menjadi tuan rumah training center (TC) pebalap - pebalap Indonesia itu. Menurut dia , TC bagi pebalap yang dipersiapkan untuk SEA Games 2017 memang harus secepatnya dilakukan mengingat pelaksanaan kejuaraan dua tahunan itu sudah tidak lama lagi. Hal ini dilakukan agar target untuk meraih hasil maksimal bisa tercapai. Hanya saja , lokasi tersebut berpeluang basa diubah sesuai dengan kebutuhan program latihan. Hal tersebut disesuaikan dengan program yang ada serta waktu pelaksanaan SEA Games 2017 Malaysia.

Gambar 4. 2 Teks ringkasan sistem kategori olahraga

Tim balap sepeda yang dipersiapkan untuk SEA Games 2017 Malaysia , pada 19 - 31 Agustus bakal dikirim ke Belanda dan Belgia guna memaksimalkan kemampuan sebelum berlaga pada kejuaraan multi event terbesar di Asia Tenggara itu . Ketua Umum Pengurus Besar Ikatan Sepeda Sport Indonesia (PB ISSI) Raja Sapta Oktohari , mengatakan , dari dua negara yang bakal dituju baru Belanda yang sudah dipastikan menjadi tuan rumah training center (TC) pebalap - pebalap Indonesia itu .

Gambar 4. 3 Teks ringkasan *indosum* kategori olahraga



Gambar 4. 4 Evaluasi *rouge* pada artikel kategori olahraga

Index	Sentence	Score	Selected
0	0 Rimanews - Tim balap sepeda yang dipersiapkan untuk SEA Games 2017 Malaysia , pa...	0.898	<input checked="" type="checkbox"/>
1	1 Ketua Umum Pengurus Besar Ikatan Sepeda Sport Indonesia (PB ISSI) Raja Sapta O...	0.435	<input checked="" type="checkbox"/>
2	2 " Yang di Belanda khusus untuk road race dan track putri saja.	0.227	<input type="checkbox"/>
3	3 Sedangkan yang putra di Belgia.	0.206	<input type="checkbox"/>
4	4 Namun , untuk yang Belgia masih dalam tahap pembicaraan , " kata Okto di Jakarta...	0.187	<input type="checkbox"/>
5	5 Menurut dia , TC bagi pebalap yang dipersiapkan untuk SEA Games 2017 memang haru...	0.401	<input checked="" type="checkbox"/>
6	6 Hal ini dilakukan agar target untuk meraih hasil maksimal bisa tercapai.	0.366	<input checked="" type="checkbox"/>
7	7 Sebenarnya , ada negara lain yang sebelumnya diputuskan untuk menjadi lokasi TC.	0.198	<input type="checkbox"/>
8	8 Ada dua tempat yang siap untuk menjadi tuan rumah yaitu Hongkong dan China.	0.323	<input type="checkbox"/>
9	9 Hanya saja , lokasi tersebut berpeluang basa diubah sesuai dengan kebutuhan prog...	0.365	<input checked="" type="checkbox"/>
10	10 Selama ini , timnas Indonesia untuk nomor road race berlatih di Yogyakarta , sed...	0.246	<input type="checkbox"/>
11	11 Rencana menggelar TC di Belanda juga dibenarkan oleh Direktur Performa Tinggi Sa...	0.247	<input type="checkbox"/>
12	12 Menurut dia ada beberapa pertimbangan sehingga TC harus dilakukan di luar negeri...	0.28	<input type="checkbox"/>
13	13 Selama TC di Solo dan Yogyakarta , kata dia , memang terdapat kendala terutama m...	0.203	<input type="checkbox"/>
14	14 Selain itu , terutama di Yogyakarta , pebalap terkendala dengan debu saat menjal...	0.304	<input type="checkbox"/>
15	15 " Akibatnya banyak yang terserang flu.	0.229	<input type="checkbox"/>
16	16 Makanya kami akan segera mengirim dokter Satlak Prima untuk melihat secara langs...	0.099	<input type="checkbox"/>
17	17 Sedangkan untuk kondisi di Solo , mantan pebalap nasional itu menjelaskan jika v...	0.185	<input type="checkbox"/>
18	18 Deny menegaskan , idealnya pebalap diberangkat ke TC di Eropa pada akhir Mei.	0.296	<input type="checkbox"/>
19	19 Hal tersebut disesuaikan dengan program yang ada serta waktu pelaksanaan SEA Gam...	0.428	<input checked="" type="checkbox"/>
20	20 Apalagi Indonesia dituntut meningkatkan prestasi dibandingkan hasil SEA Games 20...	0.233	<input type="checkbox"/>

Gambar 4. 5 Hasil skor ringkasan kalimat kategori olahraga

Hasilnya ringkasan pada hasil uji coba pada artikel kategori olahraga diatas , artikel asli dari 21 kalimat diringkas menjadi 6 kalimat dengan *compression rate* sebesar 64.5 % mengurangi 262 kata dari teks asli. Evaluasi menggunakan metrik *ROUGE* dan menghasilkan skor *ROUGE-1* senilai 0.687, *ROUGE-2* senilai 0.683, dan *ROUGE-L* senilai 0.687. Dalam hal ini sistem berhasil mengidentifikasi informasi kunci seperti “pengiriman tim ke Belanda dan Belgia, tujuan pembentukan training center, serta persiapan menghadapi *SEA Games 2017*”. Keberhasilan ini menunjukkan bahwa kategori olahraga menghasilkan performa terbaik dalam sistem peringkasan.

b. Kategori Hiburan

Kukis , camilan yang paling sering dikonsumsi oleh banyak orang ini dianggap mampu membuat pikiran menjadi lebih baik . Terutama kukis rasa coklat yang mana kandungan coklat mempunyai sifat menenangkan . Tapi , penelitian terbaru menunjukkan bahwa memakan kukis dan aneka kue lainnya malah dapat merusak suasana hati seseorang . Para peneliti dari Amerika menemukan bahwa asam lemak trans atau lemak jahat yang biasanya ditemukan di makanan olahan seperti biskuit dan kue dapat mengubah mood seseorang menjadi lebih buruk . Yahoo Health telah meneliti 5.000 orang yang gemar memakan kukis dan menemukan fakta bahwa mereka yang memakan makanan yang mengandung asam lemak jahat lebih banyak tidak dapat mengendalikan perasaan mereka . Seorang penulis penelitian dari San Diego State University , Megan Holt , mengungkapkan bahwa hubungan antara asam lemak trans dengan mental kesehatan seseorang sangat sulit untuk dipahami . Namun , Holt beranggapan jika penelitiannya harus bisa meningkatkan kesadaran masyarakat tentang bagaimana sebuah makanan dapat mempengaruhi perasaan seseorang . " Kita harus memikirkan bahwa nutrisi yang buruk juga bisa dijadikan implikasi sebagai penyebab seseorang tidak dapat mengendalikan perasaannya " tambah Megan Holt saat diwawancarai oleh Yahoo Health seperti dilansir Daily Mail . Lemak trans ini juga ditemukan di beberapa makanan yang berasal dari bahan pangan alami seperti daging dan produk susu . Dokter telah menyarankan agar tidak melakukan diet dengan mengonsumsi makanan yang mengandung tipe lemak ini karena dapat mengakibatkan kolesterol dan serangan jantung serta stroke . Penelitian sebelumnya juga menemukan orang-orang yang mengonsumsi makanan yang berlemak banyak cenderung lebih agresif . Menghindari mengonsumsi makanan yang mengandung asam lemak jahat dapat menjadi awal yang baik bagi kesehatan .

Gambar 4. 6 Teks asli artikel kategori hiburan

Para peneliti dari Amerika menemukan bahwa asam lemak trans atau lemak jahat yang biasanya ditemukan di makanan olahan seperti biskuit dan kue dapat mengubah mood seseorang menjadi lebih buruk. Seorang penulis penelitian dari San Diego State University , Megan Holt , mengungkapkan bahwa hubungan antara asam lemak trans dengan mental kesehatan seseorang sangat sulit untuk dipahami. Dokter telah menyarankan agar tidak melakukan diet dengan mengonsumsi makanan yang mengandung tipe lemak ini karena dapat mengakibatkan kolesterol dan serangan jantung serta stroke.

Gambar 4. 7 Teks ringkasan sistem kategori hiburan

Kukis , camilan yang paling sering dikonsumsi oleh banyak orang ini dianggap mampu membuat pikiran menjadi lebih baik . Penelitian terbaru menunjukkan bahwa memakan kukis malah dapat merusak suasana hati seseorang . Para peneliti dari Amerika menemukan bahwa asam lemak trans atau lemak jahat yang biasanya ditemukan di makanan olahan seperti biskuit dan kue dapat mengubah mood seseorang menjadi lebih buruk .

Gambar 4. 8 Teks ringkasan indosum kategori hiburan

Gambar 4. 9 Evaluasi *rouge* pada artikel kategori hiburan

	Index	Sentence	Score	Selected
0	0	Kukis , camilan yang paling sering dikonsumsi oleh banyak orang ini dianggap mam...	0.295	✗
1	1	Terutama kukis rasa coklat yang mana kandungan coklat mempunyai sifat menenangk...	0.147	✗
2	2	Tapi , penelitian terbaru menunjukkan bahwa memakan kukis dan aneka kue lainnya ...	0.345	✗
3	3	Para peneliti dari Amerika menemukan bahwa asam lemak trans atau lemak jahat yan...	0.412	☑
4	4	Yahoo Health telah meneliti 5.	0.204	✗
5	5	000 orang yang gemar memakan kukis dan menemukan fakta bahwa mereka yang memakan...	0.184	✗
6	6	Seorang penulis penelitian dari San Diego State University , Megan Holt , mengun...	0.411	☑
7	7	Namun , Holt beranggapan jika penelitiannya harus bisa meningkatkan kesadaran ma...	0.314	✗
8	8	" Kita harus memikirkan bahwa nutrisi yang buruk juga bisa dijadikan implikasi s...	0.163	✗
9	9	Lemak trans ini juga ditemukan di beberapa makanan yang berasal dari bahan panga...	0.339	✗
10	10	Dokter telah menyarankan agar tidak melakukan diet dengan mengonsumsi makanan ya...	0.397	☑
11	11	Penelitian sebelumnya juga menemukan orang-orang yang mengonsumsi makanan yang b...	0.328	✗
12	12	Menghindari mengonsumsi makanan yang mengandung asam lemak jahat dapat menjadi a...	0.355	✗

Gambar 4. 10 Hasil skor ringkasan kalimat kategori hiburan

Berdasarkan hasil ringkasan salah satu artikel pada kategori hiburan diatas, dari 13 kalimat artikel asli yang berisi 268 kata, sistem memilih 3 kalimat kunci *compression rate* 70.1%, dan mengurangi 188 kata dari teks asli. Evaluasi *ROUGE* pada Gambar 4.12 menunjukkan *ROUGE-1* senilai 0.515, *ROUGE-2* senilai 0.418, *ROUGE-L* senilai 0.426. Karakteristik artikel kategori hiburan pada artikel tersebut cenderung memberikan argumentasi secara bertahap, meskipun hasilnya agak rendah, sistem tetap berhasil mengidentifikasi informasi kunci terkait penelitian ilmiah dan rekomendasi medis.

Sistem peringkasan ekstraktif berfungsi dengan memilih kalimat-kalimat signifikan dari teks asli tanpa perubahan, sementara ringkasan referensi dalam dataset *IndoSum* bersifat abstraktif yang ditulis kembali dengan parafrase. Perbedaan ini membuat evaluasi *ROUGE* tidak bisa mencapai nilai tertinggi karena

istilah yang digunakan dalam ringkasan sistem berbeda dari ringkasan referensi, meskipun informasi yang disampaikan bisa jadi serupa.

Peringkasan teks ekstraktif memiliki limitasi tidak dapat melakukan parafrase seperti abstraktif, namun keunggulannya yakni *grammatically correct* dan *factual consistent* karena menggunakan kalimat asli.

Hasil ringkasan mengindikasikan bahwa sistem cenderung memilih kalimat-kalimat yang terdapat di bagian awal artikel. Karakteristik penulisan berita ini dapat dijelaskan dengan struktur piramida terbalik, di mana informasi utama diletakkan di bagian awal. Model *BERT* yang telah di-*fine-tune* berhasil mengidentifikasi pola ini, terbukti dari distribusi nilai yang menunjukkan bahwa kalimat-kalimat awal secara konsisten memperoleh skor lebih tinggi.

Kombinasi *BERT* dan *Cosine Similarity* terbukti berhasil mengatasi berbagai keterbatasan teknik ekstraktif konvensional. *BERT* memberikan pemahaman semantik yang mendalam tentang pentingnya kalimat berdasarkan konteks keseluruhan artikel, sedangkan *Cosine Similarity* membantu menangkap hubungan antar kalimat untuk meminimalkan redundansi. Walaupun kontribusi *Cosine Similarity* tergolong kecil (10% dengan α 0.9), keberadaannya tetap memperbaiki kinerja jika dibandingkan hanya dengan *BERT*.

4.3 Pembahasan

Setelah dilakukan skenario uji coba serta di dapatkan hasil evaluasi pada 200 artikel *test set Indosum*, berikut pembahasan terkait performa sistem peringkasan teks ekstraktif. Analisis ini mencakup hasil *fine tuning*, evaluasi *ROUGE*, optimasi parameter, berikut adalah penjelasannya :

4.3.1 Analisis Hasil *Fine Tuning*

Berdasarkan tabel pada hasil uji coba, model mempunyai performa yang optimal yakni pada *epoch* ke-3, dengan *F1-score* 0.511 yang menunjukkan keseimbangan terbaik antara *precision* dan *recall* dalam mengidentifikasi kalimat penting, kemudian setelah *epoch* ke-3 terjadilah *overfitting*. *Overfitting* yang terjadi setelah *epoch* ke-3 di tandai dengan meningkatnya *gap* yang terjadi di antara *training loss* dan *validation loss*. *Training loss* yang terus menurun hingga mendekati nol (0.0146), sementara itu untuk *validation loss* meningkat sampai pada 1.3801 menunjukkan bahwa model telah kehilangan kemampuan generalisasi dan hanya menghafal pola spesifik dari data *training*. Dalam hal ini, model *BERT* memiliki kecenderungan *overfitting* pada dataset dengan ukuran terbatas, terutama ketika *fine-tuning* dilakukan terlalu lama.

4.3.2 Analisis Parameter *Alpha*

Pada eksperimen parameter *alpha* memberikan peran penting terkait kontribusi *BERT* dan *Cosine Similarity* dalam sistem peringkasan teks ini. *BERT* ($\alpha=1.0$) yang hanya fokus pada *semantic importance* menghasilkan *ROUGE-1 score* 0.523, sementara itu *Cosine* ($\alpha=0.0$) yang hanya mempertimbangkan *similarity* antar kalimat menghasilkan *score* lebih rendah yaitu 0.486. Hasilnya, kombinasi optimal ditemukan pada $\alpha=0.9$ dengan *ROUGE-1 score* 0.577, menunjukkan peningkatan 10.3% dibanding *BERT* saja dan 18.7% dibanding *Cosine* saja. Hasil ini membuktikan bahwa metode tersebut dapat mengatasi kelemahan masing-masing metode. *BERT* yang kuat dalam *semantic understanding* namun lemah dalam koherensi, dikomplemen dengan baik oleh *Cosine Similarity*

yang memastikan kalimat terpilih memiliki keterkaitan yang baik. Dalam proporsi 90:10 menunjukkan bahwa *semantic importance* tetap menjadi faktor dominan, namun kontribusi kecil dari *coherence* (10%) memberikan *improvement* yang baik pada kualitas ringkasan.

4.3.3 Analisis Performa Pada Kategori

Untuk analisis variasi performa diantara kategori memberikan hal yang menarik tentang karakteristik domain yang mempengaruhi efektivitas sistem peringkasan. Kategori olahraga dengan *ROUGE-1 score* tertinggi (0.574) memiliki karakteristik artikel yang faktual, kronologis, dan terstruktur. Kalimat penting dalam artikel olahraga relatif lebih mudah untuk diidentifikasi karena biasanya mengandung informasi nyata seperti skor pertandingan, nama pemain, serta hasil kompetisi.

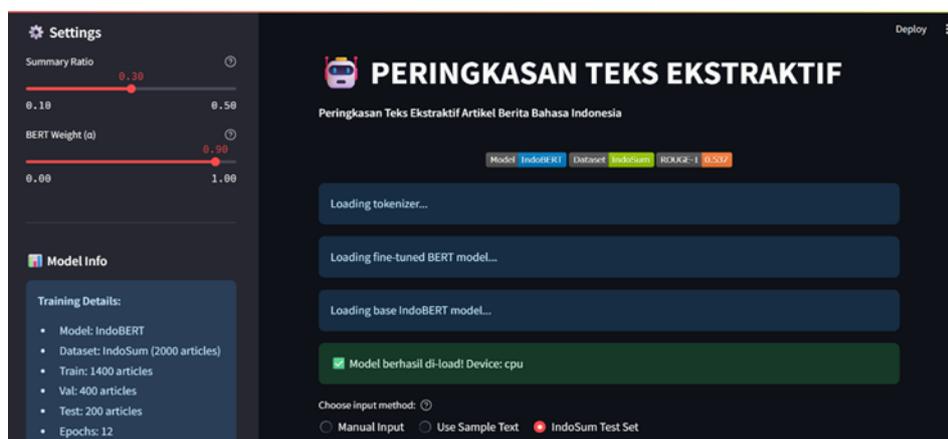
Namun sebaliknya, pada kategori hiburan dengan *score* terendah (0.467) cenderung memiliki artikel dengan gaya bahasa yang lebih deskriptif, subjektif, serta naratif. Untuk artikel hiburan sering mengandung opini serta deskripsi yang membuat penentuan *importance* menjadi lebih sulit. Kategori tajuk utama yang mencakup berita politik dan ekonomi menunjukkan performa baik (0.561) karena umumnya memiliki struktur piramida terbalik di mana informasi penting diletakkan di awal artikel. Hal ini menunjukkan bahwa sistem perlu penyesuaian domain spesifik untuk mengoptimalkan performa pada setiap kategori.

4.3.4 Analisis Pemilihan *Ratio*

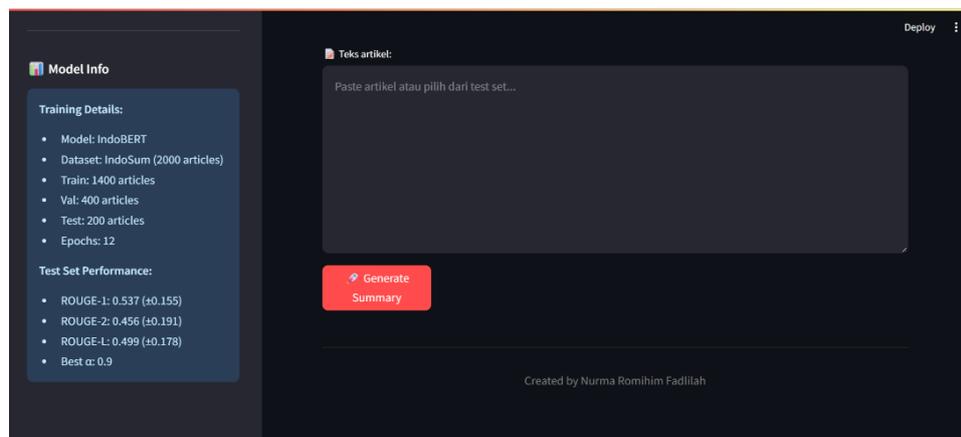
Pemilihan *ratio* 0.3 sebagai parameter optimal. *Ratio* yang terlalu kecil (0.2) menghasilkan ringkasan yang terlalu singkat dengan rata-rata kalimat hanya 2.5 kalimat, hal ini menyebabkan kehilangan informasi penting meskipun *compression rate* mencapai 80%. Di sisi lain, *ratio* yang terlalu besar (0.4-0.5) tidak memberikan *improvement* signifikan pada *ROUGE scores*, bahkan cenderung menurun karena meningkatnya *redundansi* dan berkurangnya fokus pada isi dari informasi tersebut. *Ratio* 0.3 dengan rata-rata 3.8 kalimat terpilih memberikan hasil yang mempertahankan informasi kunci sambil tetap menghasilkan ringkasan yang baik.

4.4 Gambar Implementasi Sistem

Berikut merupakan gambar implementasi dari sistem peringkasan teks ekstraktif menggunakan metode *BERT* dan *Cosine Similarity* :



Gambar 4. 11 Implementasi sistem



Gambar 4. 12 Implementasi sistem 2

Pada gambar implementasi sistem peringkasan teks menggunakan metode *BERT* dan *Cosine Similarity* diatas, user bisa melakukan *input* teks pada kolom teks artikel, kemudian melakukan peringkasan pada “*generate summary*” nantinya hasilnya akan muncul hasil ringkasan serta skor *rouge* pada artikel tersebut.

4.5 Integrasi Islam

Sistem peringkasan teks ekstraktif ini berkaitan dengan dua konsep *mu'amalah*, yakni *mu'amalah* kepada Allah (*mu'amalah ma'a Allah*) dan *mu'amalah* kepada sesama manusia (*mu'amalah ma'a an-nas*). *Mu'amalah* kepada Allah adalah segala sesuatu yang berhubungan diantara makhluk hidup dan Allah. Sedangkan *mu'amalah ma'annas* adalah hal- hal yang berhubungan dengan sesama manusia. Dalam Islam, *mu'amalah* digambarkan dengan suatu aktivitas atau kegiatan yang dilakukan seseorang dengan satu ataupun banyak orang demi memenuhi kebutuhan masing- masing. Konsep *mu'amalah ma'a Allah* dalam penelitian ini terdapat pada aspek penyampaian informasi, dalam menyajikan informasi penting untuk menyajikan informasi yang jelas, ringkas serta mudah di

pahami. Konsep *mu'amalah ma'an an-nas* terdapat pada aspek interaksi sosial, saling membantu, tolong menolong serta berbuat baik kepada sesama.

4.5.1 *Mu'amalah Ma'a Allah*

Sistem peringkasan teks dapat membantu mempersingkat waktu untuk mendapatkan informasi dari suatu artikel teks. Peringkasan teks juga bertujuan untuk menyajikan informasi yang jelas dan mudah untuk dipahami. Sehingga pembaca dapat memahami inti dari suatu teks tanpa harus membaca keseluruhan, hal ini sejalan dengan QS. Ali Imran ayat 138(Online, n.d.-b). Berikut firman Allah SWT QS. Ali Imran ayat 138 di dalam Al-Qur'an yang berbunyi :

هَذَا بَيَانٌ لِّلنَّاسِ وَهُدًى وَمَوْعِظَةٌ لِّلْمُتَّقِينَ ﴿١٣٨﴾

“(Al-Quran) ini adalah penerangan bagi seluruh manusia, dan petunjuk serta pelajaran bagi orang-orang yang bertakwa.”(QS. Ali Imron/3:138)

Dalam tafsir Al-Jalalain, yang ditulis oleh Imam Jalaluddin Al-Mahalli dan Imam Jalaluddin As-Suyuti, memberikan penjelasan ayat ini menunjukkan bahwa Al-Qur'an memberikan penjelasan yang jelas bagi orang-orang, Ini juga menekankan betapa pentingnya memahami ajaran Allah SWT dengan baik dan berbagi pengetahuan yang bermanfaat. Relevansi dengan sistem peringkasan yakni dalam konteks sistem peringkasan teks ekstraktif dapat membantu menyampaikan inti dari ajaran Islam dengan cara yang ringkas dan mudah di pahami oleh pembaca. Penekanan pada *“Bayaanun Linnaasi”* yang berarti penjelasan bagi manusia. Dalam sistem peringkasan teks ini menunjukkan bahwa informasi yang disajikan harus jelas dan mudah untuk dipahami.

Menyampaikan sesuatu secara ringkas serta jelas juga terdapat dalam hadis yang diriwayatkan oleh Imam Bukhori tentang menyampaikan ilmu, berikut bunyinya :

بَلِّغُوا عَنِّي وَلَوْ آيَةً

"Sampaikanlah dariku walaupun hanya satu ayat." (HR. Bukhori)

Dalam hadis tersebut Rasulullah SAW menyampaikan tentang pentingnya menyebarkan informasi dan pengetahuan yang bermanfaat. Bagian dari tanggung jawab setiap muslim dalam konteks *mu'amalah ma'a Allah* yakni menyampaikan ajaran Allah SWT dengan cara yang jelas dan ringkas. Hal ini pun juga sejalan dengan tujuan pada tafsir Al-Jalalain untuk memberikan penjelasan yang jelas bagi umat manusia. Hadis ini menunjukkan betapa pentingnya menyampaikan ilmu dengan cara yang ringkas. Integrasi antara hadits tersebut tentang menyampaikan ilmu dengan sistem peringkasan teks menunjukkan bahwa penyampaian informasi yang jelas, ringkas, sekaligus bermanfaat adalah kunci meningkatkan pemahaman dan pengetahuan umat Islam. Menyampaikan ilmu termasuk bentuk ibadah jika dilakukan dengan niat yang sungguh sungguh untuk mendapatkan ridha dari Allah SWT. Menyampaikan ilmu ini menjadi bagian dari *mu'amalah ma'a Allah* karena berusaha guna memenuhi perintah-Nya dalam berbuat baik kepada orang lain.

Sistem peringkasan teks ini bertujuan untuk memberikan informasi yang relevan dan penting dengan cara yang ringkas, oleh karena itu, sistem ini sejalan dengan ajaran hadits dalam memberikan informasi yang bermanfaat kepada orang lain.

4.5.2 *Mu'amalah Ma'a An-Nas*

Mu'amalah Ma'a An-Nas yakni hubungan serta interaksi antara manusia dengan sesama manusia, dalam hal ini yakni memberikan bantuan kepada mereka yang membutuhkan. Sistem peringkasan teks ini dapat membantu pembaca agar dapat mengetahui secara ringkas informasi yang ada dalam artikel secara efisien, dimana hal tersebut ada dalam potongan QS. Al-Ma'idah ayat 2 berikut :

وَتَعَاوَنُوا عَلَى الْبِرِّ وَالتَّقْوَىٰ وَلَا تَعَاوَنُوا عَلَى الْإِثْمِ وَالْعُدْوَانِ

“Tolong-menolonglah kamu dalam (mengerjakan) kebajikan dan takwa, dan jangan tolong-menolong dalam berbuat dosa dan permusuhan.” (QS.Al-Ma'idah 5:2)

Imam Ibnu Katsir membahas potongan QS. Al-Ma'idah ayat 2 (NU, n.d.) diatas yang berarti bahwa Allah SWT memerintahkan hamba-Nya yang beriman untuk selalu berbuat tolong menolong yakni dalam perbuatan yang baik, yang di dalam ayat tersebut terdapat pada kata “*al-birri*” atau “*birru*” yang artinya kebajikan, serta meninggalkan perilaku yang munkar. Dan Allah juga melarang untuk tolong menolong dalam perbuatan dosa. Sangat jelas dalam pesan Al-Qur'an diatas bahwa Allah SWT memerintahkan hambanya untuk saling membantu , dan membantu sesama manusia dengan cara yang baik. Selain itu, Allah melarang hambanya untuk membantu dalam hal yang tidak baik. Dalam *mu'amalah ma'a annas*, tolong menolong merupakan prinsip dasar yang harus diterapkan. Dalam sistem peringkasan teks ini adalah bentuk dari tolong menolong, dimana hal ini dapat membantu orang lain secara lebih efektif agar mendapatkan informasi yang berguna dan bermanfaat (Rulli Hastuti, 2022). Selain itu, dengan adanya sistem peringkasan ini dapat membantu pembaca untuk mendapatkan informasi yang

ringkas dan relevan tanpa membaca keseluruhan teks sehingga dapat menghemat waktu yang ada.

Membantu, tolong menolong juga dianjurkan oleh Rasulullah SAW, Rasulullah SAW bersabda :

عَنْ أَبِي هُرَيْرَةَ رَضِيَ اللَّهُ عَنْهُ قَالَ, قَالَ رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ قَالَ: مَنْ نَفَّسَ عَن مُؤْمِنٍ كُرْبَةً مِنْ كُرْبِ الدُّنْيَا نَفَّسَ اللَّهُ عَنْهُ كُرْبَةً مِنْ كُرْبِ الْيَوْمِ الْقِيَامَةِ, وَمَنْ يَسَّرَ عَلَى مُعَسِّرٍ يَسَّرَ اللَّهُ عَلَيْهِ فِي الدُّنْيَا وَالْآخِرَةِ, وَمَنْ سَتَرَ مُسْلِمًا سَتَرَهُ اللَّهُ فِي الدُّنْيَا وَالْآخِرَةِ, وَاللَّهُ فِي عَوْنِ الْعَبْدِ مَا كَانَ الْعَبْدُ فِي عَوْنِ أَخِيهِ... (رواه مسلم)

“Dari Abu Hurairah RA dia berkata: Rasulullah sallallahu ‘alaihi wa sallam bersabda: “Barangsiapa yang menghilangkan satu kesulitan seorang mukmin yang lain dari kesulitannya di dunia, niscaya Allah akan menghilangkan darinya satu kesulitan pada hari kiamat. Barangsiapa yang meringankan orang yang kesusahan, niscaya Allah akan meringankan baginya (urusannya) di dunia dan akhirat. Barangsiapa yang menutupi aib seorang muslim, niscaya Allah akan menutupi aibnya di dunia dan akhirat. Dan Allah akan senantiasa menolong hamba-Nya, selama hamba tersebut mau menolong saudaranya...” (HR. Muslim).

Dalam hadis tersebut yakni tentang tolong menolong dalam mengatasi kesulitan. Hadis ini menyatakan bahwa membantu menghilangkan kesulitan orang lain akan mendapatkan balasan baik dari Allah SWT. Dalam konteks sistem peringkasan teks, yakni membantu pembaca untuk mengatasi kesulitan dalam memahami informasi yang panjang, dengan adanya sistem peringkasan teks ini maka pembaca dengan mudah akan mendapatkan informasi secara ringkas dan jelas, serta memberikan kemudahan bagi pembaca untuk mendapatkan inti dari informasi tanpa harus membaca keseluruhan teks, hal ini sejalan dengan prinsip *muamalah ma’a annas* dimana saling membantu dan tolong menolong satu sama lain. Adanya sistem peringkasan teks ini sebagai perwujudan prinsip tolong menolong dan saling membantu antar sesama manusia, sesuai dengan ajaran Islam.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan pengujian peringkasan tesk ekstraktif artikel berita berbahasa Indonesia menggunakan *BERT* dan *Cosine Similarity* dapat disimpulkan bahwa penelitian ini berhasil mengimplementasikan sistem peringkasan teks ekstraktif yakni *BERT* yang telah di *fine-tuning* mampu mengidentifikasi tingkat kepentingan kalimat dengan akurasi 79.37%, *cosine similarity* dapat melengkapi dengan mengukur keterkaitan semantik antar kalimat, dalam hal ini kombinasi optimal dengan bobot $\alpha = 0.9$ (90% *BERT*, 10% *Cosine Similarity*) menghasilkan performa terbaik. Pada pemrosesan dataset *Indosum*, sistem berhasil memproses 2000 artikel dari dataset *IndoSum* dengan pembagian data yang proporsional: 70% *training* (1400 artikel), 20% *validation* (400 artikel), dan 10% *testing* (200 artikel) dan total 35,911 kalimat yang diproses dengan distribusi label tidak seimbang (18.9% kalimat penting). Dapat melakukan *parsing* yang efektif dari format token menjadi kalimat utuh dengan mempertahankan struktur dan makna.

Untuk proses *fine-tuning* yang optimal *fine-tuning* model *IndoBERT* selama 12 *epoch*, dan hasilnya terdapat pada *epoch* 3 yang menunjukkan *epoch* yang optimal memberikan hasil *training time* yang efisien (8.84 menit) dengan memanfaatkan GPU. Evaluasi pada 200 artikel *test set* menunjukkan *ROUGE-1*: 0.537 (± 0.155) menunjukkan *overlap* kata yang baik, *ROUGE-2*: 0.456 (± 0.191) preservasi frasa yang cukup baik, dan *ROUGE-L*: 0.499 (± 0.178) yang berarti struktur kalimat terpreservasi dengan baik. Sistem berhasil menghasilkan ringkasan

dengan *compression rate* 70-80% dengan mempertahankan informasi esensial, koherensinya tinggi karena mempertahankan kalimat asli (ekstraktif), mencakup informasi yang representatif dari artikel sumber. Keberhasilan implementasi sistem yang telah dikembangkan menjadi bentuk pengamalan dalam QS. Ali Imron ayat 138 dan QS. Al-Ma'idah ayat 2.

5.2 Saran

Peneliti menyadari terdapat beberapa kekurangan di dalam penelitian ini, untuk mendapatkan hasil yang lebih baik dalam pembuatan sistem peringkasan teks ekstraktif artikel berita berbahasa Indonesia dapat menerapkan beberapa saran. Berikut saran untuk peneliti selanjutnya :

1. Mengeksplorasi model *BERT* yang lebih mendalam lagi
2. Melakukan eksplorasi pendekatan abstraktif *summarization* untuk bahasa Indonesia, mengembangkan sistem yang dapat menghasilkan kalimat baru dan bukan hanya ekstraksi
3. Melakukan *fine-tuning* dengan *dataset* yang lebih besar serta beragam

DAFTAR PUSTAKA

- Abdi, A., Mariyam, S., & Aliguliyev, R. M. (2018). QMOS : Query-based multi-documents opinion-oriented summarization. *Information Processing and Management An International Journal*, 54(December 2017), 318–338.
- AL-Hafiidh, D. F., Rozi, I. F., & Putri, I. K. (2022). Peringkasan Teks Otomatis pada Portal Berita Olahraga menggunakan metode Maximum Marginal Relevance. *Jurnal Teknik Informatika Dan Sistem Informasi*, 8(3), 21–30. <https://doi.org/10.33795/jip.v8i3.519>
- Albab, M. U., Karuniawati P, Y., & Fawaiq, M. N. (2023). Optimization of the Stemming Technique on Text preprocessing President 3 Periods Topic. *Jurnal TRANSFORMATIKA*, 20(2), 1–10. <https://journals.usm.ac.id/index.php/transformatika/page1>
- APJII. (2024). Internet Indonesia. *Laporan Survei Internet APJII 2024. Asosiasi Penyelenggara Jasa Internet Indonesia*, 1–73. <https://survei.apjii.or.id/survei/group/9>
- Apriliani, D., Handayani, S. F., & Saputra, I. T. (2023). Implementasi Natural Language Processing (NLP) Dalam Pengembangan Aplikasi Chatbot Pada SMK YPE Nusantara Slawi. *Techno.Com*, 22(4), 1037–1047. <https://doi.org/10.33633/tc.v22i4.9155>
- Astuti, R. H., Muljono, M., & Sutriawan, S. (2024). Indonesian News Text Summarization Using MBART Algorithm. *Scientific Journal of Informatics*, 11(1), 155–164. <https://doi.org/10.15294/sji.v11i1.49224>
- Ayu Syahfitri, R., Kurniawan, A., & Irsan Humaidy, M. (2022). Bulletin of Data Science Penerapan Algoritma Maximum Marginal Relevance Dalam Peringkasan Teks Secara Otomatis. *Media Online*, 1(2), 49–56. <https://ejurnal.seminar-id.com/index.php/bulletinds>
- Basbeth, F., & Fudholi, D. H. (2024). Klasifikasi Emosi Pada Data Text Bahasa Indonesia Menggunakan Algoritma BERT, RoBERTa, dan Distil-BERT. *Jurnal Media Informatika Budidarma*, 8(2), 1160. <https://doi.org/10.30865/mib.v8i2.7472>
- Bluto, K. (2020). *KONSEP SUKSES DALAM PERSPEKTIF AL- QUR ' AN SURAH AL -ASR AYAT 1-3 Diajukan kepada : Lembaga Penelitian dan Pengabdian Masyarakat (LP2M)*.
- Despotovic, M., Stumpe, E., Brunauer, W. A., Lehner, M., & Zeppelzauer, M. (2024). Predictive Assessment of the Interior Quality of Apartments Based on Multimodal Data with Variable Visual Input. *SN Computer Science*, 5(6). <https://doi.org/10.1007/s42979-024-03028-4>
- Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2019). BERT:

- Pre-training of Deep Bidirectional Transformers for Language Understanding. *Naacl-Hlt 2019, Mlm*, 4171–4186. <https://aclanthology.org/N19-1423.pdf>
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics, Mlm*. [https://www.semanticscholar.org/venue?name=North American Chapter of the Association for Computational Linguistics](https://www.semanticscholar.org/venue?name=North%20American%20Chapter%20of%20the%20Association%20for%20Computational%20Linguistics)
- Effendy, E., Zakaria, Azlisa, & Anggarana. (2023). Dasar Dasar Penulisan Berita. *Jurnal Pendidikan Dan Konseling*, 5(2), 4042–4044. <https://journal.universitaspahlawan.ac.id/index.php/jpdk/article/view/13888>
- Fatmalasari, D., & Lumbanraja, F. R. (2022). Peringkasan Teks Artikel Ilmiah Berbahasa Indonesia dengan Metode Pembobotan Kalimat. *Jurnal Pepadun*, 3(3), 314–322. <https://doi.org/10.23960/pepadun.v3i3.127>
- Halim, F., Liliana, & Gunadi, K. (2022). Ringkasan Ekstraktif Otomatis pada Berita Berbahasa Indonesia Menggunakan Metode BERT. *Jurnal Infra*, 10(1), 162–168.
- Halimah, Surya Agustian, & Siti Ramadhani. (2022). Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexrank. *Jurnal CoSciTech (Computer Science and Information Technology)*, 3(3), 371–381. <https://doi.org/10.37859/coscitech.v3i3.4300>
- Harinatha, S. R. K., Tasara, B. T., & Qomariyah, N. N. (2021). Evaluating Extractive Summarization Techniques on News Articles. *Proceedings - 2021 International Seminar on Intelligent Technology and Its Application: Intelligent Systems for the New Normal Era, ISITIA 2021, July*, 88–94. <https://doi.org/10.1109/ISITIA52817.2021.9502230>
- Juna, M. F., & Hayaty, M. (2023). The observed preprocessing strategies for doing automatic text summarizing. *Computer Science and Information Technologies*, 4(2), 119–126. <https://doi.org/10.11591/csit.v4i2.pp119-126>
- Khaqiqi, R. F., Qomaruddin, M., & Wp, B. S. (2022). Rancang Bangun Aplikasi Baca Berita Online dengan Peringkas Teks Otomatis Menggunakan Algoritma Textrank. *Prosiding Seminar Nasional Konstelasi Ilmiah Mahasiswa UNISSULA 7 (KIMU 7)*, 7(Kimu 7), 21–34.
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3), 3713–3744. <https://doi.org/10.1007/s11042-022-13428-4>
- Kurniawan, K., & Louvan, S. (2018). IndoSum : A New Benchmark Dataset for Indonesian Text Summarization. *International Conference on Asian Language Processing, February*. <https://doi.org/10.48550/arXiv.1810.05334>

- Kurniawan, R., & Arif, M. F. (2022). Implementasi Text Mining Menggunakan Metode Cosine Similarity Untuk Klasifikasi Konten Berita Di Postingan Grup Facebook Info Lantas Dan Kriminal Pasuruan. *JAMI: Jurnal Ahli Muda Indonesia*, 3(1), 9–17. <https://doi.org/10.46510/jami.v3i1.41>
- Lin, C., & Rey, M. (2004). R OUGE : A Package for Automatic Evaluation of Summaries. *Annual Meeting of the Association for Computational Linguistics*.
- NU, O. Q. (n.d.). *Surat Al-Ma'idah Ayat 2: Arab, Latin, Terjemah dan Tafsir Lengkap*. Quran NU Online. Retrieved May 15, 2025, from <https://quran.nu.or.id/al-ma'idah/2>
- OECD. (2023). Equity in education in PISA 2022. In *PISA 2022 Results (Volume I): The State of Learning and Equity in education* (Vol. 1). https://www.oecd-ilibrary.org/education/pisa-2022-results-volume-i_03c74bdd-en
- Oktavia, A. N., Iqbal, M., Saputra, R. W., Zulfikar, M. I., & Saifudin, A. (2024). Implementasi Metode Natural Language Processing Dalam Studi Analisis Semantik Dan Emosi Buzzer Pada Tweet Di Aplikasi X. *BIIKMA : Buletin Ilmiah Ilmu Komputer Dan Multimedia*, 2(1), 154–159.
- Online, Q. N. (n.d.-a). *Surat Al-Ashr Ayat 1-3: Arab, Latin, Terjemah dan Tafsir Lengkap*. Quran NU Online. Retrieved October 10, 2024, from <https://quran.nu.or.id/al-ashr>
- Online, Q. N. (n.d.-b). *Surat Ali-Imran Ayat 138: Arab, Latin, Terjemah dan Tafsir Lengkap*. Quran NU Online. Retrieved May 15, 2025, from <https://quran.nu.or.id/ali-imran/138>
- Prasetyo, V. R., Benarkah, N., & Chrisintha, V. J. (2021). Implementasi Natural Language Processing Dalam Pembuatan Chatbot Pada Program Information Technology Universitas Surabaya. *Teknika*, 10(2), 114–121. <https://doi.org/10.34148/teknika.v10i2.370>
- Pratama, F. A., & Romadhony, A. (2020). Identifikasi Komentar Toksik Dengan BERT. *E-Proceeding of Engineering*, 7(2), 7941–7949.
- Putu, N., Saraswati, V. D., Yudistira, N., & Adikara, P. P. (2023). Analisis Sentimen terhadap Perundungan Siber pada Twitter menggunakan Algoritma Bidirectional Encoder Representations from Transformer (BERT). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 7(2), 909–916. <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/12345>
- Rulli Hastuti, U. (2022). Konsep Layanan Perpustakaan : Analisis Tafsir Surat Al-Maidah Ayat (2). *THE LIGHT : Journal of Librarianship and Information Science*, 2(2), 88–93. <https://doi.org/10.20414/light.v2i2.6182>
- Salam, R. R., Jamil, M. F., Ibrahim, Y., Rahmaddeni, R., Soni, S., & Herianto, H. (2023). Analisis Sentimen Terhadap Bantuan Langsung Tunai (BLT) Bahan Bakar Minyak (BBM) Menggunakan Support Vector Machine. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 3(1), 27–35.

<https://doi.org/10.57152/malcom.v3i1.590>

- Samosir, F. V. P., Toba, H., & Ayub, M. (2022). BESKlus : BERT Extractive Summarization with K-Means Clustering in Scientific Paper. *Jurnal Teknik Informatika Dan Sistem Informasi*, 8(1). <https://doi.org/10.28932/jutisi.v8i1.4474>
- Sanjaya, A., Setiawan, A. B., Mahdiyah, U., Farida, I. N., & Prasetyo, A. R. (2023). Pengukuran Kemiripan Makna Menggunakan Cosine Similarity dan Basis Data Sinonim Kata. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 10(4), 747–752. <https://doi.org/10.25126/jtiik.20241046864>
- Savanti, N., Gotami, W., & Dewi, R. K. (2018). Peringkasan Teks Otomatis Secara Ekstraktif Pada Artikel Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Latent Semantic Analysis. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIHK) Universitas Brawijaya*, 2(9), 2821–2828.
- Utomo, M. S., Wibowo, J. S., & Wahyudi, E. N. (2022). Text Summarization Pada Artikel Berita Menggunakan Vector Space Model Dan Cosine Similarity. *Jurnal Dinamika Informatika*, 14(1), 11–24. <https://doi.org/10.35315/informatika.v14i1.9163>
- Vendyansyah, N., & Pranoto, Y. A. (2021). Perancangan dan Pembuatan Aplikasi untuk Mendeteksi Kemiripan Jawaban Menggunakan Cosine Similarity. *Jurnal Teknika (Jurnal Fakultas Teknik Universitas Islam Lamongan)*, 13(1), 23–28.
- Vidya Chandradev, I Made Agus Dwi Suarjaya, & I Putu Agung Bayupati. (2023). Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT. *Jurnal Buana Informatika*, 14(02), 107–116. <https://doi.org/10.24002/jbi.v14i02.7244>
- Wang, Q., Liu, P., Zhu, Z., Yin, H., Zhang, Q., & Zhang, L. (2019). A text abstraction summary model based on BERT word embedding and reinforcement learning. *Applied Sciences (Switzerland)*, 9(21). <https://doi.org/10.3390/app9214701>
- Widiastutik, R., Zaman, L., & Santoso, J. (2021). Peringkasan Teks Ekstraktif pada Dokumen Tunggal Menggunakan Metode Restricted Boltzmann Machine. *INSYST: Journal of Intelligent System and Computation*, 58–64.
- Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS : Pre-training with Extracted Gap-sentences for Abstractive Summarization. *International Conference on Machine Learning*.