IMPLEMENTASI PERINGKAS DOKUMEN BERBAHASA INDONESIA MENGGUNAKAN METODE TEXT TO TEXT TRANSFER TRANSFORMER (T5)

I Nyoman Purnama[1)], **Ni Nengah Widya Utami**[2)]
Program Studi Sistem Informasi [1)], Sistem Informasi Akutansi[2)]
Universitas Primakara, Denpasar, Bali [1) 2)]
purnama@primakara.ac.id [(1)], widya@primakara.ac.id [2)]

ABSTRACT

With so much information contained in digital news, it makes it difficult for readers to know the essence of this collection of texts. For this reason, a system is needed that can automatically summarize digital news in Indonesian. Document summarization is the process of extracting text from a document, exploring and presenting important information to users or applications in the form of a short and concise summary. When we are faced with a language structure that is quite complex, such as in human language, then it captures the main idea and the meaning of the original text. This is where the Transformer model is used, which has a high-performance and a compact model. T5 is an example of an abstractive transformer model that rewrites sentences rather than just taking sentences directly from the original text. In this research, the process of summarizing news documents in Indonesian was carried out using the T5 transformer method. This research was done with three scenarios. The differences for each scenario is the preprocessing part. In scenario 1 stemming and stopwords removal are implemented, in scenario 2 stemming is implemented without stopwords removal, and in scenario 3 neither is implemented. The conclusion that can be drawn in this research, is that the best test obtained with scenario 2, by implementing stemming without stopwords removal with a ROUGE-1 evaluation value of 0.17568.

Keywords*: Summarization, T5, ROUGE, Document, Indonesian Language*

ABSTRAK

Dengan banyaknya informasi yang terdapat pada sebuah berita digital, membuat pembaca terkadang mengalami kesulitan untuk mengetahui intisari dari kumpulan teks ini. Untuk itu dibutuhkan sebuah sistem yang bisa meringkas berita digital berbahasa Indonesia secara otomatis. Peringkasan dokumen adalah proses mengambil teks dari sebuah dokumen, menggali dan menyajikan informasi penting bagi user atau aplikasi dalam bentuk rangkuman yang singkat dan padat. Ketika kita dihadapkan pada struktur bahasa yang cukup kompleks, seperti pada Bahasa manusia, kemudian menangkap ide dan makna utama teks asli. Disinilah digunakan model Transformer yang merupakan model peringkas yang berkinerja tinggi. T5 merupakan contoh model transformer abstraktif yang menulis ulang kembali kalimatnya daripada hanya mengambil kalimat langsung dari teks aslinya. Pada penelitian ini dilakukan proses peringkasan dokumen berita berbahasa Indonesia dengan metode transformer T5. Penelitian ini dikerjakan dengan tiga skenario. Bagian yang membedakan masing-masing skenario adalah pada bagian preprocessing katanya. Pada skenario 1 diimplementasikan *stemming* dan *stopwords removal*, pada skenario 2 diimplementasikan *stemming* tanpa *stopwords removal*, dan pada skenario 3 tidak diimplementasikan keduanya. Adapun kesimpulan yang dapat diambil pada penelitian ini adalah pengujian terbaik yang didapatkan adalah pengujian dengan skenario 2, yaitu dengan mengimplementasikan *stemming* tanpa *stopwords removal* dengan nilai evaluasi ROUGE-1 0.17568.

Kata Kunci: Peringkas, dokumen, T5, Bahasa Indonesia, ROUGE

PENDAHULUAN

Kebutuhan akan informasi pada era digital ini bisa diperoleh melalui berbagai macam media yang tidak terbatas sumbernya. Salah satu sumber informasi yang paling cepat dan akurat bisa diperoleh melalui media digital. Internet merupakan salah satu media digital yang paling popular saat ini. Dimana informasi di internet bisa diperoleh melalui berita, artikel ataupun status dari media social. Berita memiliki sumber informasi yang cukup padat dan beragam. Dengan banyaknya informasi yang terdapat pada berita, membuat pembaca terkadang mengalami kesulitan untuk mengetahui intisari dari kumpulan teks ini. Untuk itu dibutuhkan sebuah sistem yang bisa meringkas dokumen digital seacra otomatis. Ringkasan juga merupakan tugas penting pada teknologi pemrosesan Bahasa (Natural language Processing). Adapun tugas dari peringkas dokumen adalah untuk memberikan ringkasan dari sebuah teks Panjang[1]. Ringkasan bisa dijadikan pratinjau untuk pembaca sebelum mengetahui lebih detail isi dari sebuah artikel atau berita. Peringkasan dokumen adalah proses mengambil teks dari sebuah dokumen, menggali dan menyajikan informasi penting bagi user atau aplikasi dalam bentuk rangkuman yang singkat dan padat. Peringkasan dokumen dapat menjadi solusi bagi setiap orang yang tidak memiliki banyak waktu dan sedang membutuhkan informasi penting dalam tumpukan dokumen yang terus berkembang[2] Bahasa Indonesia digunakan dalam penelitian ini karena peringkasan Bahasa Indonesia masih belum terlalu berkembang dibandingkan dengan Bahasa Inggris,

Peringkasan teks sendiri digolongkan menjadi peringkasan teks ekstraktif dan abstraktif[3]. Peringkasan teks secara ekstraktif merupakan peringkasan teks yang dilakukan dengan menampilkan kembali paragraf atau kalimat dari dokumen teks yang merupakan topik utama sebuah dokumen teks namun dengan bentuk yang sederhana. Pada metode ini peringkas tidak merubah kata dalam kalimat, melainkan dengan memilih kata yang penting. Beberapa metode ektraktif yang bisa digunakan yakni textRank, tf-idf dan Word2vec. Sedangkan peringkasan teks abstraktif merupakan sebuah interpretasi dari teks asli yang ada. Kalimat yang terdapat pada dokumen ditransformasikan kembali menjadi kalimat yang lebih singkat[4].Umumnya abstraksi dapat membentuk teks yang lebih kuat dibandingkan dnegan metode ekstraktif. Metode abstraktif membangkitkan hasil ringkasan sesuai dengan bagaimana manusia meringkas teks. Beberapa model yang bisa digunakan pada metode abstraktif yakni Seq2seq, BART, T5 dll.

Berdasarkan penelitian yang dilakukan Kulkarni and Apte[5] disebutkan bahwa pendekatan yang bagus untuk metode ekstraktif terdiri dari 4 langkah yakni pemrosesan teks, ektraksi fitur pada kata dan kalimat, pemilihan kalimat dan merakitnya, serta pembuatan ringkasan. Pemrosesan teks terdiri dari tokenisasi, penghapusan stop words dan stemming. Proses ini dilakukan untuk mengekstrak fitur yang ada pada dokumen kemudian berdasarkan bobotnya dan memberikan nilai 1 atau 0. Langkah berikutnya yaitu pemilihan dan perakitan kalimat, dimana kalimat disimpan dengan urutan sesuai dengan rangkingnya. Pada Langkah terakhir akan diperoleh ringkasan dari dokumen aslinya. Metode ini mengambil bebrapa bariis dari teks yang dipandang penting, tapi tidak menghasilkan kata yang tidak ada dalam teks aslinya. Hampir semua metode peringkasan yang ada, termasuk kedalam ekstraktif. Sedangkan pada metode abstraktif, ringkasan dibangun dengan menggunakan representasi semantic dari blok teks asli[6]. Hasil dari ringkasan berisi kalimat yang berbeda dengan blok teks aslinya. Peringkasan ini hampir sama dengan bagaimana manusia meringkas blok teks yang panjang. Metode abtraktif lebih kompleks dan membutuhkan proses yang lebih lama dibandingkan dengan metode ekstraktif. Namun hasil

ringkasan yang dihasilkan lebih baik dan lebih bermakna.

Ketika kita dihadapkan pada struktur bahasa yang cukup kompleks, seperti pada Bahasa manusia untuk memparafrasekan kalimat rumit menjadi frasa pendek Kemudian menangkap ide dan makna utama teks asli. Disinilah digunakan model Transformer yang merupakan model peringkas yang berkinerja tinggi. Model transformer menjadi referensi dalam hal NLP, karena strukturnya yang kompleks dan mekanisme perhatiannya yang memungkinkan mereka untuk memahami cara kerja bahasa lisan dan tulisan. Contoh model transformer yang bisa digunakan untuk meringkas teks berbahasa Indonesia yakni BERT dan T5. BERT merupakan kependekan dari Bidirectional Encoder Representations from Transformers. BERT [7] merupakan model yang dikembangkan oleh *Google* untuk keperluan *Natural Language Processing* seperti *text classification*, *question answering*, dan lain-lain. Sedangkan T5 merupakan model yang dikembangkan oleh perusahaan raksasa *Google*. Merupakan kependekan dari text to text transfer transformer. T5 merupakan algoritma peringkas abstraktif yang menulis ulang kembali kalimatnya daripada hanya mengambil kalimat langsung dari teks aslinya.

Penelitian mengenai peringkasan teks menggunakan beberapa metode telah dilakukan oleh Barbella dengan judul "A Comparison of Methods for the Evaluation ofText Summarization Techniques"[8]. Pada penelitiannya diuji bagaimana pengukuran ROUGE bisa digunakan untuk menguji kinerja metode abstraktif dan ekstraktif. Hasil pengujian dengan metode BERT memiliki hasil yang terbaik dibandingkan dengan metode lainnya. Penelitian lain mengenai metode peringkas dengan K-means clustering berhasil memberikan hasi ringkasan yang baik, berdasarkan 50 responden dengan nilai 87%. Dimana data yang digunakan berupa berita berbahasa Indonesia dari website CNN[9]. Metode ektraktif juga digunakan untuk meringkas dokumen berbahasa Bali[1]. Pada penelitian ini proses yang dilakukan yakni text preprocessing, ekstraksi fitur, skor kalimat dan penyusunan ringkasan. Hasil yang diperoleh diuji dengan metode ROUGE, dan diperoleh hasil sebesar 0.52. Penelitian mengenai penggunaan metode T5 telah dilakukan oleh Afeefa Farhath., dimana pada penelitiannya menggunakan ringkasan berita yang diperoleh dari dataset Kaggle. Data ini terlebih dahulu diproses dengan metode preprosesing, sebelum dimasukkan pada model T5. Pada penelitian ini digunakan metode ROUGE untuk menguji hasil ringkasannya[10]. Berdasarkan hasil penelitian diperoleh T5 merupakan metode yang terbaik untuk peringkasan abstraktif.

Berdasarkan penelitian yang telah dilakukan sebelumnya, pada penelitian ini diimplementasikan metode peringkasan abstraktif. Metode ini akan diuji menggunakan model Transformer. Model transformer yang digunakan yakni T5 untuk metode abstraktif. Dataset yang digunakan yakni data berita Bahasa Indonesia, yang diperoleh dari dataset INDOSUM. Dimana INDOSUM merupakan dataset yang diklaim sebagai benchmark untuk kasus peringkas teks otomatis Bahasa Indonesia[11]. Dataset tersebut berisi 19 ribu pasangan artikel berita dan gold summaries. Artikel-artikel pada dataset tersebut merupakan artikel berita yang diambil dari portal berita online seperti CNN Indonesia Kedua model tadi selanjutnya akan diuji menggunakan metode ROUGE.

METODE PENELITIAN

A. Dataset

Data yag digunakan pada penelitian ini berasal dari dataset INDOSUM. Dataset INDOSUM merupakan dataset berbahasa Indonesia yang merupakan hasil penelitian yang dilakukan Kurniawan dan Louvan. Dataset ini sering dijadikan tolak ukur dalam penelitian mengenai ringkasan otomatis berbahasa Indonesia. Dataset ini berbentuk json yang terdiri dari 19.000 berita yang telah ditokenisasi dan bersasal

dari shortir.com[7]. Dataset dibagi menjadi 5 *folds cross-validation* dimana setiap fold terdiri dari training set, development set, dan test set. Dalam dataset ini terdapat kategori yang menunjukkan kategori berita, *gold label* yang merupakan label *gold summarization* untuk peringkasan ekstraktif, id unik tiap berita dari berita yang bersangkutan, *source* yang berisi sumber berita, *source url* yang berisi tautan dari berita, *paragraph* yang berisi paragraf teks dari berita, dan *summary* yang berisi ringkasan berita yang dibuat oleh manusia dan bersifat abstraktif[12]. Paragraf diberikan dalam bentuk daftar kalimat. Dimana kalimat merupakan kumpulan kata. Sebelum data digunakan ke dalam aplikasi yang dikembangkan dalam Bahasa Python, data akan diolah terlebih dahulu sehingga memudahkan dalam implementasinya. Terdapat beberapa hal yang akan dilakukan dalam pengolahan data. Pertama, mengambil bagian 'paragraphs', dan 'summary' dari dataset. Kemudian dilanjutkan dengan membuka array pada tiap bagian 'paragraphs', 'summary' untuk diambil isinya. Terakhir, isi pada masing-masing bagian akan digabungkan dengan delimiter '<q>'. Pada bagian 'paragraphs' dan 'summary', delimiter ini bertujuan untuk memisahkan tiap kalimat sedangkan pada 'gold_labels' bertujuan untuk memisahkan label dari tiap kalimat yang akan digunakan sebagai referensi ekstraktif. Berikut pada table 1 merupakan salah satu isi dari artikel berita yang digunakan pada penelitian ini, dimana flatten article dan flatten summary merupakan isi artikel asli dan ringkasan dari berita yang telah digabungkan dari kumpulan katanya.

Table 1. Contoh dataset yang digunakan

Kategori	Flatten_article	Flatten_summary
Tajuk Utama	Jakarta , CNN Indonesia - - Dokter Ryan Thamrin , yang terkenal lewat acara Dokter Oz Indonesia , meninggal dunia pada Jumat (4 / 8) dini hari ., Dokter Lula Kamal yang merupakan selebriti sekaligus rekan kerja Ryan menyebut kawannya itu sudah sakit sejak setahun yang lalu ., Lula menuturkan , sakit itu membuat Ryan mesti vakum dari semua kegiatannya , termasuk menjadi pembawa acara Dokter Oz Indonesia ., Kondisi itu membuat Ryan harus kembali ke kampung halamannya di Pekanbaru , Riau untuk menjalani istirahat ., " Setahu saya dia orangnya sehat , tapi tahun lalu saya dengar dia sakit ., (Karena) sakitnya , ia langsung pulang ke Pekanbaru , jadi kami yang mau jenguk juga susah ., Barangkali mau istirahat , ya betul juga , kalau di Jakarta susah isirahatnya , " kata Lula kepada CNNIndonesia.com , Jumat (4 / 8) ., Lula yang mengenal Ryan sejak sebelum aktif berkarier di televisi mengaku belum sempat membesuk Ryan lantaran lokasi yang jauh dst..	Dokter Lula Kamal yang merupakan selebriti sekaligus rekan kerja Ryan Thamrin menyebut kawannya itu sudah sakit sejak setahun yang lalu ., Lula menuturkan , sakit itu membuat Ryan mesti vakum dari semua kegiatannya , termasuk menjadi pembawa acara Dokter Oz Indonesia ., Kondisi itu membuat Ryan harus kembali ke kampung halamannya di Pekanbaru , Riau untuk menjalani istirahat

B. Text processing

Sebelum dimasukkan ke masing-masing model, maka teks akan diprsoses terlebih dahulu menggunakan beberapa metode pemrosesan teks, sehingga diperoleh teks berita yang bersih dan bebas *typo*. Beberapa proses yang dilakukan yakni menghapus judul artikel, menghapus spasi dan tanda baca. Mengganti beberapa spasi dengan 1 spasi. Terakhir dilakukan proses penghapusan titik singkatan. Pembersihan selanjutnya adalah menghilangkan unicode whitespace characters seperti \t\n\r\f\v. Semua proses ini dilakukan dengan menggunakan *library Python.* Adapun proses yang dilakukan pada pemrosesan teks yakni :

Stemming

Stemming merupakan proses mengubah kata-kata yang memiliki imbuhan menjadi kata baku sehingga kata-kata yang diproses sistem adalah kata-kata dasar dalam bahasa indonesia[13]. *Stemming* di kesempatan ini menggunakan modul sastrawi yang tersedia di python.

Tokenizing dan Filtering

Tokenizing dan *Filtering* menghilangkan tanda baca yang tidak

diperlukan dan dilakukan dengan memisahkan kalimat menjadi kata-kata [14]. Kalimat dipisah menjadi kata-kata dengan pemisah karakter spasi. Setiap kata-kata di kalimat kemudian dibandingkan apakah tidak termasuk pada kumpulan tanda baca. Jika tidak termasuk pada kumpulan tanda baca, maka kata-kata masuk dalam hasil *preprocessing*, Tanda baca diambil dari modul standar library python *string punctuation*. Berikutnya mengubah semua token ke bentuk huruf kecil (*lower case*).

Proses tokenisasi melibatkan library *SentencePiece* pada Python. *SentencePiece* adalah sebuah *library* yang digunakan untuk tokenisasi teks dalam berbagai bahasa. Ini dapat digunakan untuk membagi teks menjadi unit-unit yang lebih kecil[15], seperti kata-kata atau subword, yang berguna dalam berbagai tugas pemrosesan bahasa alami seperti pemodelan bahasa, penerjemahan mesin, dan tugas lainnya.

Stopwords and numbers Removal

Stopwords removal merupakan proses penghilangan kata tidak penting pada deskripsi melalui pengecekan kata-kata hasil parsing deskripsi apakah termasuk di dalam daftar kata tidak penting (*stoplist*) atau tidak[12]. Angka dalam bentuk kata tidak dimasukkan dalam perhitungan karena dapat menimbulkan multitafsir. Kamus *stopword* bahasa Indonesia diperoleh dari modul python bernama NLTK.

C. Proses Utama

Setelah dilakukan pemrosesan teks, proses berikutnya yakni pembuatan ringkasan dengan menggunakan model T5. Model T5 merupakan model *embedding*, dimana *embedding* terdiri dari sekumpulan angka. *Word embedding* merupakan *embedding* untuk token atau kata. Penyematan kalimat adalah untuk kalimat. Penyematan teks adalah untuk teks. Karena pada penelitian ini akan melakukan peringkasan teks berita maka *embedding* yang digunakan yakni *embedding* kalimat T5 menggunakan arsitektur *Transformer* dan pendekatan *text-to-text transfer learning* tanpa metode *embedding* yang eksplisit.

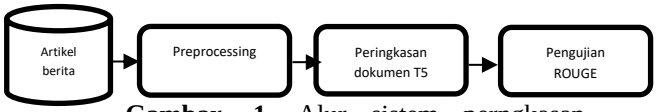
Sistem T5 (*Text-to-Text Transfer Transformer*) adalah model yang menggunakan arsitektur Transformer untuk melakukan berbagai tugas pemrosesan bahasa alami dalam format yang seragam[16]. Berikut adalah ringkasan desain sistem T5:

1. Arsitektur Transformer: T5 menggunakan arsitektur Transformer, yang terdiri dari beberapa lapisan yang saling terhubung. Setiap lapisan Transformer terdiri dari mekanisme self-attention dan lapisan feed-forward untuk memproses teks input.
2. Text-to-Text Format: T5 mengadopsi pendekatan "teks ke teks" yang seragam untuk semua tugas. Artinya, semua masukan dan keluaran dijelaskan dalam format teks, termasuk tugas-tugas seperti terjemahan, ringkasan, pertanyaan-jawaban, dan lainnya. Ini memungkinkan T5 untuk dilatih dan digunakan secara seragam dalam berbagai tugas pemrosesan bahasa alami.
3. Pretraining dengan Dataset Raksasa: T5 menjalani tahap pretraining yang besar dengan menggunakan dataset teks yang sangat besar, seperti Wikipedia. Selama pretraining, model belajar untuk melakukan berbagai tugas pemrosesan bahasa alami yang berbeda dalam format teks ke teks.
4. *Fine-tuning* untuk Tugas Tertentu: Setelah tahap pretraining, T5 dapat disesuaikan atau "fine-tuned" untuk tugas tertentu. Model T5 yang telah dilatih secara umum diubah dengan menambahkan lapisan klasifikasi di atasnya dan melatihnya menggunakan dataset yang spesifik untuk tugas tersebut.

Misalnya, jika digunakan untuk terjemahan, T5 akan dilatih menggunakan pasangan kalimat sumber dan target.

5. Pengkodean Teks Input: Sebelum memproses teks input, T5 menggunakan teknik "pengkodean" untuk menunjukkan tugas apa yang harus dilakukan model. Misalnya, jika tugasnya adalah terjemahan, teks input akan diawali dengan token "translate:" untuk memberi tahu model bahwa itu adalah tugas terjemahan.
6. Generasi Teks Output: Setelah memproses teks input, T5 dapat menghasilkan teks output yang sesuai dengan tugas yang ditentukan. Model dilatih untuk menghasilkan urutan teks yang benar dalam konteks tugas yang diberikan, seperti terjemahan kalimat, ringkasan artikel, atau jawaban pertanyaan.

Secara garis besar proses yang dilakukan pada penelitian ini dapat digambarkan pada gambar 1 dibawah ini :



Gambar. 1. Alur sistem perngkasan dokumen

Semua proses pembuatan model dilakukan menggunakan *library* yang ada pada python. Karena dalam proses kedua model ini memerlukan ukuran data yang besar dan sumberdaya GPU yang tinggi, maka pada penelitian ini digunakan Google colabs dengan lingkungan berbasis *cloud* dan memory besar serta memiliki GPU yang mendukung CUDA.

Berbeda dengan metode berbasis *embedding*, dimana pendekatan tradisional digunakan untuk mewakili kata-kata dalam bentuk numerik dan menghasilkan representasi numerik untuk setiap kata dalam korpus teks. Pada metode berbasis transformer merupakan arsitektur model yang inovatif dan mengandalkan mekanisme perhatian untuk membangun representasi kontekstual. Transformer menunjukan hasil yang sangat baik dalam tugas-tugas Natural Language Processing(NLP)[17].

Model T5 menggunakan Pustaka Python yaitu *Hugging Face Transformers*. *Hugging Face Transformers* adalah sebuah pustaka yang menyediakan implementasi dan antarmuka yang mudah digunakan untuk berbagai model pemrosesan bahasa alami, termasuk T5. Dengan menggunakan *Hugging Face Transformers*, dapat melakukan beberapa tugas pemrosesan bahasa dengan T5, termasuk peringkasan, terjemahan, generasi teks, dan lainnya. Pustaka ini menyediakan pre-trained model T5 yang dapat digunakan secara langsung, atau Anda dapat melatih model T5 Anda sendiri dengan menggunakan pustaka ini.

Pelatihan dan Pengujian

ROUGE merupakan salah satu cara untuk melakukan evaluasi kualitas dari suatu ringkasan yang dibuat secara otomatis. Ringkasan akan dibandingkan dengan gold summary yang ada pada dataset INDOSUM. *Gold summary* merupakan ringkasan dibuat oleh manusia dan menjadi standar dalam pengujian bagi teks yang akan diringkas.

Pemilihan ROUGE dalam evaluasi ringkasan dikarenakan sudah menjadi standar evaluasi dan telah banyak digunakan oleh penelitian lain. Model ROUGE yang digunakan adalah model ROUGE-1, dimana model ini menggunakan unigram untuk melihat kemiripan ringkasan hasil model dan ringkasan asli (*gold summaries*). ROUGE - 1 menilai kata-kata yang dihasilkan dari proses peringkasan dilihat dari kata perkata, bukan berdasarkan urutan.

Pelatihan model T5 membutuhkan waktu dan sumber daya komputasi yang signifikan. Oleh karena itu, mungkin perlu untuk menggunakan perangkat keras yang kuat atau platform komputasi yang disesuaikan seperti GPU atau TPU untuk

melatih model dengan efisiensi yang baik. Pada penelitian ini menggunakan sumber daya dari Google Colabs untuk melakukan semua proses dalam peringkasan dokumen.

Skenario Penelitian

Pada penelitian ini akan di tiga skenario. Pada skenario pertama, penulis mengimplementasikan stemming dan stopwords removal pada dataset saat melakukan preprocessing. Skenario kedua penulis hanya mengimplementasikan stemmingdan tanpa mengimplementasikan stopwordsremoval. Lalu pada skenario ketiga, penulis tidak mengimplementasikan stemmingdan stopwordsremoval. Mekanisme yang digunakan untuk melakukan stemmingdan stopwords removal adalah menggunakan library Sastrawi Python.

Penulis mendefinisikan tiga skenario ini adalah untuk melihat bagaimana pengaruh dari stemming dan stopwords removal pada sistem peringkasan ttrack computer teks ini. Jika terdapat pengaruh yang cukup signifikan ataupun tidak terdapat pengaruh, maka diharapkan hasil perbandingan ini dapat menjadi acuan bagi penelitian selanjutnya untuk mempertimbangkan apakah akan menggunakan stemming dan stopwords removal atau tidak sama sekali.

HASIL DAN PEMBAHASAN

Pada bagian ini terdiri dari implementasi dan hasil pengujian. Uji coba dilakukan untuk memperoleh nilai relevansi dari sistem yang telah dibangun dengan metode ROUGE.

Implementasi

Pertama kali dilakukan proses pengumpulan data rangkuman yang berasal dari dataset INDOSUM. Setelah dilakukan proses pengumpulan data berita dan rangkumannya, selanjutnya dilakukan beberapa seknario yakni skenario pertama ada proses stemming dan stopwords removal, Skenario kedua hanya mengimplementasikan stemming Lalu pada skenario ketiga, tidak mengimplementasikan stemming dan stopwords removal.

Pengujian sistem ini akan dilakukan untuk mencari tahu konfigurasi yang tepat dengan melakukan percobaan training. Selain itu, karena adanya keterbatasan resource yang dimiliki oleh peneliti maka untuk penelitian ini akan digunakan *resource* yang berasal dari Google Colab untuk melakukan proses training dan testing dari model. Kemudian, untuk pengujian lain selain pengujian awal dan akhir hanya akan diimplementasikan pada fold yang pertama saja untuk mempersingkat waktu pengujian pada penelitian ini.

Hasil pengujian

Bagian ini terdiri dari hasil pengumpulan dokumen, hasil pengolahan rangkuman referensi dan hasil pengujian rangkuman sistem. Semua pengujian dilakukan dengan menggunakan google Colabs. Pengumpulan dokumen dilakukan dengan mengunduh dataset INDOSUM di alamat https://www.kaggle.com/datasets/linkgish/indosum. Data ini terdiri dari 19.000 artikel berita ynag sudah ditokenisasi. Dataset ini dibagi menjadi 5 cross validation fold, yang didalamnya sudah rerdapat bagian training, development and bagian test. Pada penelitian ini digunakan data bagian training pada fold 1.

Data awal yang berupa JSON dirubah terlebih dahulu dalam bentuk flat paragraf. Ada 2 bagian pada dataset yang digunakan yaitu paragraf dan summary. Dikarenakan ukuran fold yang cukup besar, untuk mengurangi proses pelatihan maka digunakan hanya 500 data paragraph pertama saja. Data summary juga digunakan sesuai dengan jumlah data paragraph.

Proses *stemming* dilakukan dengan menggunakan library sastrawi untuk dataset berbahasa Indonesia. Sedangkan proses penghapusan *stopwords* berbahasa Indonesia mengunakan library python NLTK. Proses penghapusan *stopwords* otomatis akan menghilangkan beberapa

kata yang ada pada dataset sumber. Hal inilah yang akan dicoba dianalisa sejauh mana pengaruh penghapusan *stopwords* dan *stemming* pada proses pembuatan peringkasan berbahasa Indonesia. Setelah itu akan dilakukan proses peringkasan dokumen dnegan metode T5. Berikut hasil pengujian dari beberapa scenario pengujian yang dilakukan

A. Hasil pengujian skenario 1

Pada pengujian scenario pertama diimplementasikan proses *stemming* dan penghapusan *stop words* dari bagian paragraph. Hasil pengujian skenario 1 diperlihatkan pada table 1 dibawah :

Table 2. Hasil pengujian skenario 1

Skenario 1	NILAI ROUGE-1 (Average)	NILAI ROUGE-1 (Maximum)
F1-score	0.07612	0.5497
Precission	0.08459	0.4971
Recall	0.08891	0.6472

Berdasarkan hasil diatas diperoleh bahwa untuk skenario pertama ini mendapatkan hasil rata-rata F1-score untuk semua data uji sebesar 0.07612 dan nilai maksimum F1-score sebesar 0.5497. Nilai tersebut dapat dikatakan sangat rendah. Hal tersebut disebabkan karena dengan diimplementasikannya *stopwords removal*, maka panjang kata dalam suatu data (artikel berita) menjadi berkurang. Stopwords memiliki peran penting dalam suatu kalimat. Maka dari itu hasil peringkasan dengan skenario ini sangat rendah.

B. Hasil pengujian skenario 2

Pada pengujian skenario kedua diimplementasikan proses stemming saja dari bagian paragraf. Hasil pengujian skenario 2 diperlihatkan pada table 2 dibawah :

Table 3. Hasil pengujian skenario 2

Skenario 2	NILAI ROUGE-1 (Average)	NILAI ROUGE-1 (Maximum)
F1-score	0.17866	0.5297
Precission	0.16067	0.5172
Recall	0.16032	0.5472

Berdasarkan hasil di atas, diketahui bahwa untuk skenario kedua ini mendapatkan hasil rata- rata F1-score untuk semua data uji sebesar 0.17866 dan nilai maksimum F1-score sebesar 0.5297. Hasil pada skenario 2 ini cukup baik, hasil pada skenario 2 ini meningkat sangat signifikan. Maka dari itu dapat dinilai bahwa *stopwords removal* sangat mempengaruhi performansi dari sistem.

C. Hasil pengujian skenario 3

Pada pengujian skenario ketiga tidak dilakukan proses *stemming* dan penghapusan *stop words* dari bagian paragraph. Hasil pengujian skenario 3 diperlihatkan pada table 3 dibawah :

Table 4. Hasil pengujian skenario 3

Skenario 3	NILAI ROUGE-1 (Average)	NILAI ROUGE-1 (Maximum)
F1-score	0.17568	0.5397
Precission	0.16467	0.5273
Recall	0.16533	0.5273

Berdasarkan hasil di atas, diketahui bahwa untuk skenario ketiga menggunakan metode T5 ini mendapatkan hasil rata- rata F1-score untuk semua data uji sebesar 0.17568 dan nilai maksimum F1-score sebesar 0.5397. Jika dilihat berdasarkan hasil rata-rata f1-score nya, maka pada skenario ini juga mendapatkan hasil yang cukup baik. Jika dibandingkan pada skenario 2, maka pada skenario ke 3 ini mendapatkan hasil yang sedikit lebih rendah daripada skenario ke 2. Seperti dijelaskan sebelumnya bahwa pada skenario ke 2 mengimplementasikan *stemming*, sedangkan skenario 3 tidak mengimplementasikan *stemming* dan *stopwords removal*.

Disini dapat disimpulkan bahwa *stemming* memungkinkan untuk mempengaruhi performansi sistem, tetapi tidak signifikan. Hal tersebut disebabkan karena stemming tidak mengurangi kata

sedikitpun, hanya mengubah bentuk kata kembali ke kata dasar. Sehingga pengubahan kata ke dalam kata dasar di dalam suatu kalimat tidak mengubah banyak makna dari kalimat tersebut.

Perbandingan hasil dari ketiga skenario pengujian dilakukan berdasarkan nilai f1-score. ROUGE F1- score dipilih karena dinilai lebih baik dan lebih representatif untuk mengevaluasi suatu model karena melibatkan nilai *recall* dan *precision* sekaligus. Berikut hasil perbandingan yang didapatkan.

Table 5. Perbandingan F1-score

Skenario	F1 Score ROUGE-1 (Average)
Skenario 1	0.07612
Skenario 2	0.17866
Skenario 3	0.17568

Berdasarkan tabel 5 di atas, terlihat bahwa skenario yang mendapatkan hasil terbaik adalah skenario 2. Jika diamati berdasarkan hasil pengujian ROUGE dengan metode T5 terlihat bahwa nilai evaluasi ROUGE-1 pada skenario ke-2 dan ke-3 tidak berbeda secara signifikan, sedangkan jika dibandingkan dengan skenario ke-1 nilainya berbeda cukup jauh. Artinya penghapusan *stopwords* memberikan pengaruh kepada hasil pengujian, sedangkan untuk stemming memberikan pengaruh yang tidak terlalu jauh. *Stopwords removal* memberikan peranan sangat besar pada pengujian karena dengan *stopwords removal* akan memberikan masukan yang cukup berbeda bagi keseluruhan proses peringkasan dengan metode T5 ini.

Peringkasan dengan menggunakan model *Transformer* T5, dimana proses peringkasan dilakukan secara abstraktif dengan beberapa scenario. Pada skenario kedua dan ketiga proses *stemming* pada pengujian peringkasan otomatis ini memberikan sedikit pengaruh, dimana nilai F1-score yang didapatkan tidak jauh berbeda. Hal ini berarti proses *stemming* memberikan pengaruh yang baik, artinya dengan mengimplementasikan *stemming* maka akan mendapatkan performansi lebih baik. Hal ini disebabkan karena aturan Bahasa Indonesia yang memungkinkan satu kata dasar memiliki beberapa kata bentukan dengan penambahan beberapa imbuhan yang membuat kata-kata tersebut dapat berbeda makna meskipun kata dasarnya sama.

Sedangkan pada skenario 3 yang tidak mengimplementasikan *stemming* semua bentukan kata berimbuhan dari kata dasar dianggap berbeda. Namun, meskipun secara nilai evaluasi skenario 2 mendapatkan hasil yang sedikit lebih baik, hasil ringkasan pada skenario 2 terkesan tidak natural karena mengubah bentuk kata ke kata dasar. Sedangkan pada skenario ke-3 hasil yang didapatkan lebih natural mendekati hasil ringkasan alami oleh manusia.

SIMPULAN

Pada penelitian ini dilakukan proses peringkasan dokumen berita berbahasa Indonesia dengan metode transformer T5. Penelitian ini dikerjakan dengan tiga skenario. Bagian yang membedakan masing-masing skenario adalah pada bagian preprocessing katanya. Pada skenario 1 diimplementasikan *stemming* dan *stopwords removal*, pada skenario 2 diimplementasikan *stemming* tanpa *stopwords removal*, dan pada skenario 3 tidak diimplementasikan keduanya. Semua skenario tersebut dilatih dengan metode yang sama, dan kondisi yang sama.

Dengan mengimplementasikan *stemming* tanpa *stopwords removal* dengan nilai evaluasi ROUGE-1 0.17568. Tetapi hasil tersebut hanya berbeda sedikit daripada skenario 3 yang tidak mengimplementasikan *stemming* dan *stopwords removal*, dimana nilai ROUGE ang diperoleh yakni 0.17568. *Stemming* memberikan pengaruh yang tidak terlalu signifikan dibandingkan dengan proses penghapusan *stop wods.*

DAFTAR PUSTAKA

[1] P. Studi, T. Informatika, J. I. Komputer, F. Matematika, D. Ilmu, and P. Alam, "PERINGKASAN TEKS OTOMATIS UNTUK DOKUMEN BAHASA BALI BERBASIS METODE EKTRAKTIF I Putu Gede Hendra Suputra," 2017.

[2] W. Budiyono and F. Solihin, "APLIKASI PERINGKAS BERITA ONLINE OTOMATIS MENGGUNAKAN METODE ORDINARY WEIGHTING PADA SITUS PENGUMPUL BERITA," 2014.

[3] M. A. Zamzam, "SISTEM AUTOMATIC TEXT SUMMARIZATION MENGGUNAKAN ALGORITMA TEXTRANK," *MATICS*, vol. 12, no. 2, pp. 111–116, Sep. 2020, doi: 10.18860/mat.v12i2.8372.

[4] R. Adelia, S. Suyanto, and U. N. Wisesty, "Indonesian abstractive text summarization using bidirectional gated recurrent unit," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 581–588. doi: 10.1016/j.procs.2019.09.017.

[5] Kulkarini and Apte, "A DOMAIN-SPECIFIC AUTOMATIC TEXT SUMMARIZATION USING FUZZY LOGIC," *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY*, vol. 4, no. 4, pp. 1–13, 2013, [Online]. Available: www.iaeme.com/ijcet.asp

[6] A. Y. Setiawan, I. Gede, M. Darmawiguna, and G. A. Pradnyana, "SENTIMENT SUMMARIZATION EVALUASI PEMBELAJARAN MENGGUNAKAN ALGORITMA LSTM (LONG SHORT TERM MEMORY)," *Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI)*, vol. 11, no. 2, 2022.

[7] F. V. P. Samosir, H. Toba, and M. Ayub, "BESKlus : BERT Extractive Summarization with K-Means Clustering in Scientific Paper," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 1, Apr. 2022, doi: 10.28932/jutisi.v8i1.4474.

[8] M. Barbella, M. Risi, and G. Tortora, "A comparison of methods for the evaluation of text summarization techniques," in *Proceedings of the 10th International Conference on Data Science, Technology and Applications, DATA 2021*, SciTePress, 2021, pp. 200–207. doi: 10.5220/0010523002000207.

[9] A. Firdaus, N. Yusliani, and D. Rodiah, "Text Summarization using K-Means Algorithm," 2021. [Online]. Available: http://sjia.ejournal.unsri.ac.id

[10] A. Farhath, "ABSTRACTIVE TEXT SUMMARIZATION," *International Research Journal of Modernization in Engineering*

Technology and Science www.irjmets.com @International Research Journal of Modernization in Engineering, pp. 2582–5208, 1967, doi: 10.56726/IRJMETS38320.

[11] K. Kurniawan and S. Louvan, "IndoSum: A New Benchmark Dataset for Indonesian Text Summarization," in *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, Institute of Electrical and Electronics Engineers Inc., Jan. 2019, pp. 215–220. doi: 10.1109/IALP.2018.8629109.

[12] F. Halim and K. G. Liliana, "Ringkasan Ekstraktif Otomatis pada Berita Berbahasa Indonesia Menggunakan Metode BERT."

[13] H. Dwiharyono and S. Suyanto, "Stemming for Better Indonesian Text-to-Phoneme," *Ampersand*, vol. 9, 2022, doi: 10.1016/j.amper.2022.100083.

[14] A. N. Laili, P. P. Adikara, and S. Adinugroho, "Rekomendasi Film Berdasarkan Sinopsis Menggunakan Metode Word2Vec," vol. 3, no. 6, pp. 6035–6043, 2019.

[15] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," *31st International Conference on Machine Learning, ICML 2014*, vol. 4, pp. 2931–2939, 2014.

[16] A. G. Etemad, A. I. Abidi, and M. Chhabra, "Fine-Tuned T5 for Abstractive Summarization," *International Journal of Performability Engineering*, vol. 17, no. 10, pp. 900–906, Oct. 2021, doi: 10.23940/ijpe.21.10.p8.900906.

[17] A. Gupta, D. Chugh, and R. Katarya, "Automated News Summarization Using Transformers."