TELEMATIKA MKOM
Volume 15, Nomor 2, September, 2023
Halaman 86-92
ISSN: 2527-9033 (*online*), 2085-725X (cetak)

PENDEKATAN *HYBRID* PADA SISTEM PERINGKAS TEKS ARTIKEL BERITA BAHASA INGGRIS MENGGUNAKAN *NATURAL LANGUAGE PROCESSING*

Farah Raihanunnisa[1], Muhammad Arhami [2], Rahmad Hidayat [2*]

[1]Teknik Informatika, Teknologi Informasi dan Komputer, Politeknik Negeri Lhokseumawe
[2]Teknologi Informasi dan Komputer, Politeknik Negeri Lhokseumawe, Lhokseumawe
Jl. Banda Aceh-Medan Km. 280,3, Buketrata, Mesjid Punteut, Blang Mangat, Lhokseumawe, Aceh, Indonesia
e-mail koresponden: rahmad_hidayat@pnl.ac.id

(***received*****:** 11/08/2023, ***revised*****:** 22/08/2023, ***accepted*****:** 24/08/2023)

Abstrak

Kegiatan mengumpulkan informasi melalui sejumlah artikel yang dilakukan dalam kehidupan sehari hari baik oleh kalangan pelajar, peneliti, jurnalis, dan sebagainya, memakan waktu yang relatif lama. Hal ini menimbulkan masalah ketika seseorang harus mengumpulkan informasi yang cukup dalam waktu yang terbatas. Penelitian ini bertujuan untuk membuat sebuah sistem peringkas teks otomatis yang dapat menghasilkan ringkasan yang relevant dan informatif sehingga membantu penggali informasi untuk dapat menemukan informasi penting dalam sebuah artikel dengan waktu yang lebih sedikit dibandingkan dengan membaca keseluruhan artikel. Sistem peringkas teks otomatis yang diajukan menerapkan NLP *(Natural Language Processing)* dengan pendekatan *hybrid*. Pendekatan *hybrid* merupakan gabungan dari dua teknik, yaitu teknik peringkasan ekstraktif dan teknik peringkasan abstraktif. Peringkasan ekstraktif merupakan peringkasan yang dilakukan dengan mengekstrak kalimat dari dokumen asli, Sedangkan peringkasan abstraktif dilakukan dengan menghasilkan kalimat baru mendekati peringkasan yang dihasilkan oleh manusia. Peringkasan ekstraktif yang dilakukan menggunakan algoritma *Textrank*, sedangkan teknik peringkasan abstraktif dilakukan dengan menerapkan arsitektur *Transformer*. *Textrank* merupakan pendekatan berbasis *graph*, sedangkan *transformer* merupakan rangkaian algoritma berbasis *encoder decoder*. Pengujian model dilakukan dengan menerapkan teknik pengujian *ROUGE (Recall Oriented Understudy for Gisting Evaluation)*, dimana rouge melakukan pengujian berdasarkan n-gram kata. Hasil yang diperoleh pada penelitian ini menunjukkan nilai *F1-Score* 0.34 pada *ROUGE-1*, 0.15 pada *ROUGE-2*, dan 0.25 pada *ROUGE-L*.

Kata kunci: Arsitektur *Transformer*, Pendekatan *Hybrid* , Peringkas Teks Otomatis, *Recall Oriented Understudy for Gisting Evaluation, Textrank.*

Abstract

The activity of gathering information through a number of articles carried out in everyday life by students, researchers, journalists, and so on, takes a relatively long time. This creates a problem when one has to gather sufficient information in a limited time. This study aims to create an automatic text summary system that can produce relevant and informative summaries so as to help information seekers to be able to find important information in an article in less time than reading the entire article.The proposed automatic text summary system applies NLP (Natural Language Processing) with a hybrid approach. The hybrid approach is a combination of two techniques, namely extractive summary techniques and abstractive summary techniques. Extractive summarization is carried out using the Textrank algorithm, while abstractive summarization techniques are carried out by applying the Transformer architecture. Model testing is carried out by applying the ROUGE (Recall Oriented Understudy for Gisting Evaluation) testing technique. The result in this study shows F1-Score value 0.34 on ROUGE-1, 0.15 on ROUGE-2, and 0.25 on ROUGE-L.

Keywords: *Automatic Text Summarizer, Hybrid Approach, Recall Oriented Understudy for Gisting Evaluation, Textrank, Transformer Architecture*

TELEMATIKA MKOM
Volume 15, Nomor 2, September, 2023
Halaman 86-92
ISSN: 2527-9033 (*online*), 2085-725X (cetak)

1. Pendahuluan

Natural Language Processing atau Pemrosesan Bahasa Alami merupakan bidang dalam *Artificial Intelligence* yang mempelajari cara komunikasi antara manusia dan komputer. Bidang ini mendalami bagaimana suatu mesin seolah dapat memahami dan mengerti bahasa alami yang digunakan manusia [1]. Salah satu bagian dari NLP adalah *Automatic Text Summarization* atau Peringkas Teks Otomatis. Fokus terhadap bidang ini pertama kali dipelajari oleh Lun sekitar enam puluh tahun yang lalu dan telah menyita perhatian dalam beberapa tahun terakhir [2]. Peringkas teks otomatis merupakan sebuah teknik yang membangun ringkasan singkat dan akurat dari sebuah dokumen yang panjang [3]. Peringkas teks otomatis memberikan gambaran informasi keseluruhan dari dokumen hanya melalui beberapa kalimat yang cukup dibaca dalam waktu yang singkat. Hal ini akan membantu pekerjaan dalam mengumpulkan informasi penting melalui sejumlah dokumen tekstual yang menghabiskan waktu yang lama untuk dibaca secara keseluruhan. Alasan tersebut menjadikan peringkas teks otomatis sebagai salah satu teknik yang diminati dan mulai dipelajari oleh manusia. Tujuan utama dari peringkas otomatis yaitu untuk mendapatkan informasi penting yang menjadi inti dari sebuah dokumen.

Informasi merupakan kumpulan data dan fakta yang begitu penting dan banyak digunakan dalam seluruh aspek kegiatan manusia. Informasi dapat tersedia dalam berbagai jenis media baik cetak maupun digital. Artikel merupakan salah satu media sumber informasi yang cukup sering digunakan untuk berbagai keperluan. Seorang jurnalis harus membaca banyak artikel sebagai informasi tambahan dan referensi dalam membuat sebuah berita. Seorang pelajar, peneliti, penulis harus berhadapan dengan banyak artikel untuk mengumpulkan data dan informasi yang nantinya akan dijadikan landasan terhadap penelitian atau karya tulis yang akan diangkat. Ada beberapa permasalahan yang diperoleh para peneliti, pelajar, atau masyarakat umum ketika akan melakukan pengumpulan informasi diantaranya adalah : Membutuhkan waktu yang relative lama karena harus membaca keseluruhan dokumen yang panjang dan Perlunya perbaikan dan peningkatan relevansi dalam pengembangan peringkas teks otomatis

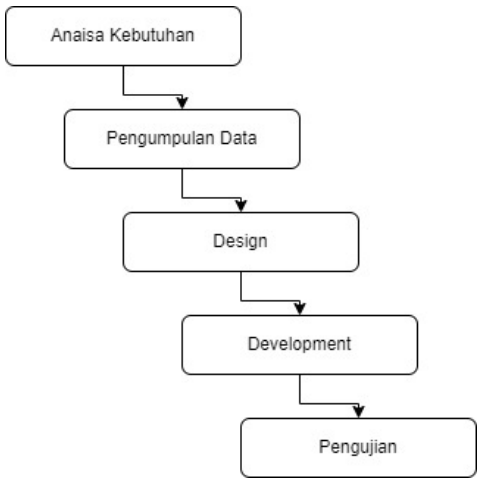
Penelitian sebelumnya telah menunjukkan berbagai metode peringkas teks otomatis dengan akurasi yang berbeda beda. Penelitian yang dilakukan oleh Ade Naufal, peringkas teks otomatis dengan menggunakan algoritma Binary Firefly dan menunjukkan nilai F1 score 0.46 pada ROUGE-1, 0.34 pada ROUGE-2, dan 0.42 pada ROUGE-L [4]. Penelitian yang dilakukan oleh Leni, peringkas teks otomatis dengan menggunakan algoritma *Textrank* pada dokumen berbahasa Indonesia, namun sayangnya Leni tidak menejelaskan hasil pengujian secara spesisik [5]. Penelitian menerapkan *deep learning* dengan algoritma RNN pada peringkas teks otomatis menunjukkan hasil nilai F1 score 0.43 pada ROUGE-1, 0.20 pada ROUGE-2 dan 0.39 pada ROUGE-L [6]. Penelitian menerapkan pendekatan Hybrid dengan menggabungkan algoritma LSA dan RBM pada peringkas teks otomatis dan menunjukkan penurunan nilai F1 *score* yang cukup drastis dibandingkan menggunakan kedua metode secara terpisah [7].

Dari seluruh penelitian yang pernah dilakukan belum pernah ada penelitian yang melakukan peringkasan dengan pendekatan *Hybrid* yang menggabungkan algoritma *Textrank* dan *Transformer Architecture*. Oleh karena itu, penelitian ini mengajukan sistem peringkas teks otomatis dengan menerapkan pendekatan *Hybrid* yang mengggabungkan peringkasan ekstraktif dan abstraktif dengan menggunakan algoritma *Textrank* dan *Transformer Architecture* yang diharapkan memiliki nilai keakuratan yang lebih baik.

2. Metode Penelitian

Secara sederhana alur tahapan pembuatan sistem untuk penelitian terdapat pada Gambar 1.

TELEMATIKA MKOM
Volume 15, Nomor 2, September, 2023
Halaman 86-92
ISSN: 2527-9033 (*online*), 2085-725X (cetak)



Gambar 1. Alur Tahapan Penelitian

Berikut ini adalah penjelasan tahapan alur penelitian yang terdapat pada Gambar 1:

a. Analisa Kebutuhan, pada tahapan ini akan dilakukan peninjauan lebih lanjut terkait dengan kebutuhan system yang dibangun, apakah system ini dibutuhkan, dan seperti apa detailnya system yang dibutuhkan
b. Pengumpulan Data, penelitian ini hanya melibatkan data sekunder, yaitu data yang akan digunakan untuk membangun system peringkas teks otomatis, data yang digunakan merupakan kumpulan dataset CNN/Daily Mail yang berisi artikel berita dan ringkasan yang dibuat secara manual.
c. Rancang dan Bangun Sistem, pada tahapan ini dilakukan perancangan sistem termasuk design UML, design *User Interface*, dan dilanjutkan dengan pembangunan sistem peringkas teks otomatis
d. Pengujian, pada tahapan ini dilakukan untuk menunjukkan performa dari pendekatan yang digunakan dalam peringkas teks otomatis, hasil dari tahapan ini yang menunjukkan bahwa metodologi yang diajukan mampu menghasilkan ringkasan yang relevan atau tidak.

2.2. Metodologi Penelitian

Penelitian ini mengajukan Teknik *Hybrid* dalam membangun ringkasan. Teknik *hybrid* yang dimaksud merupakan kombinasi dari teknik peringkasan ekstraktif dan abstraktif. Secara garis besar proses peringkasan terbagi kedalam tiga tahapan sebagaimana yang terlihat pada Gambar 2.

Gambar 2. Alur Peringkasan

Gambar 2. Merupakan tampilan dari alur peringkasan secara keseluruhan, berikut merupakan tahapan peringkasan:

a. *Preprocessing*
   Tahap *preprocessing* merupakan tahapan awal yang bertujuan untuk menyiapkan teks sedemikian rupa sehingga siap untuk diringkas. Tahapan ini meliputi proses segmentasi, tokenisasi, menghapus *stopwords* dan tanda baca, dan lemmatisasi.
b. Peringkasan Ekstraktif
   Peringkasan ekstraktif bertujuan untuk ekstraksi kalimat penting dalam keseluruhan dokumen. Adapun metode yang digunakan adalah *TextRank*. *TextRank* merupakan model peringkatan berbasis grafik dalam pemrosesan teks yang diajukan oleh Mihalchea dan Paul. Ada dua pembelajaran

TELEMATIKA MKOM
Volume 15, Nomor 2, September, 2023
Halaman 86-92
ISSN: 2527-9033 (*online*), 2085-725X (cetak)

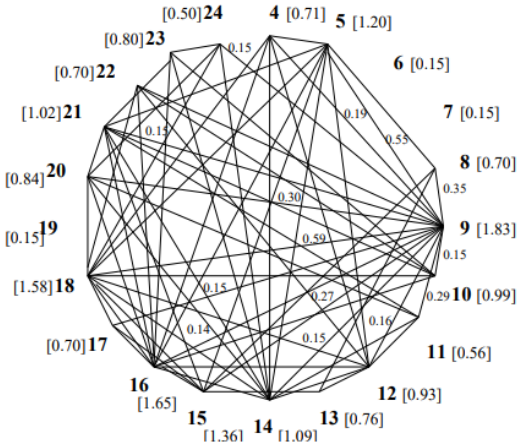
unsupervised yang diterapkan yakni, model untuk ekstraksi *keyword* dan model untuk ekstraksi kalimat [8]. *TextRank* sendiri merupakan perkembangan dari *PageRank*, dan graph yang dihasilkan pada model ini merupkan *graph* tidak berarah [9].

TextRank merupakan *method* berbasis *graph* yang dapat digunakan untuk meringkas dokumen tunggal. Pada metode ini kalimat kalimat akan direpresentasikan sebagai *node* yang dihubungkan oleh *edge*, dimana *edge* merupakan representasi *similarity* atau kemiripan antar kalimat. Adapun perhitungan kemiripan menggunakan rumus *cosine similarity*. *Cosine similarity* digunakan untuk menghitung jarak antara dua vektor. Sebagai contoh x dan y adalah dua vektor yang akan digunakan. Formula dari cosine similarity terdapat pada persamaan 1.

$$\cos(x, y) = \frac{x \,.\, y}{||x||.||y||} \quad (1)$$

dengan:
x = vektor x
y = vektor y
||x|| = √ (x1)^2 + (x2)^2 + (xn)^2
||y|| = √ (y1)^2 + (y2)^2 + (yn)^2
Perbedaan dua vektor diberikan oleh :
dis(x, y) = 1 - cos(x, y)

Gambar 3. merupakan gambaran pembobotan pada algoritma *TextRank* dimana 24 kalimat saling terhubung dengan edge yang merupakan perhitungan *similarity* terhadap kalimat lainnya.



Gambar 3. Pembobotan TextRank [5]

Adapun tahap peringkasan ekstarktif terdiri dari :

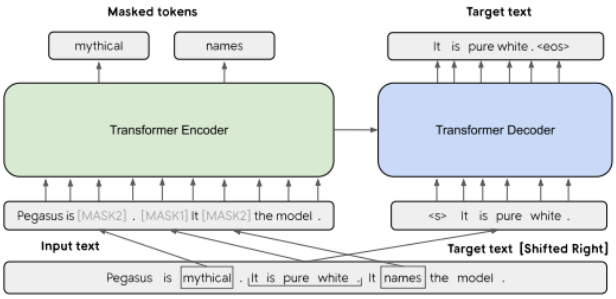
1. Mengubah kalimat menjadi vektor dengan menggunakan *sentence embedding*.
2. Menghitung kemiripan antar kalimat melalui vektor yang telah dibangun menggunakan *cosine similarity*.
3. Merepresentasikan *graph* dimana kalimat akan dijadikan *node* dan dihubungkan oleh nilai kemiripan kalimat tersebut dengan kalimat lainnya.
4. Setelah *graph* serta bobot direpresentasikan selanjutnya adalah proses perangkingan, dimana kalimat dengan nilai *top rank* akan dimasukkan kedalam ringkasan. Pemilihan dimulai dari kalimat dengan skor tertinggi dan seterusnya.

c. Peringkasan Abstraktif

Transformer architecture diimplementasikan dalam tahapan peringkasan abstraktif. Secara garis besar struktur dalam arsitektur *transformer* terbagi menjadi dua yaitu *Encoder* dan *decoder*. *Encoder* berfungsi untuk memproses kalimat *input* dan *decoder* berfungsi untuk membangun kalimat *output*.

TELEMATIKA MKOM
Volume 15, Nomor 2, September, 2023
Halaman 86-92
ISSN: 2527-9033 (*online*), 2085-725X (cetak)

Masing – masing *encoder* dan *decoder* mengandung *multiple layers* dari *self attention* dan *feed forward neural networks* [10]. Implementasi *transformer* architecture dalam penelitian ini dilakukan dengan memanfaatkan Pegasus model. Pegasus merupakan sebuah *pretrained* model yang dikembangkan diatas *transformer* architecture [11]. Pegasus dilatih pada sekitar 1,5 juta dataset dari berbagai kumpulan dataset *summarization* seperti Xsum, CNN, Wikihow, Pubmed, Billsum, dan lain sebagainya. Arsitektur Pegasus dapat dilihat pada gambar 4.



Gambar 4. Arsitektur Pegasus Model [8]

Pegasus mengimplementasikan standard transformer encoder – decoder dengan mengaplikasikan GSG dan MLM.

1. *Gap Sentences Generation*
   Selama masa *pre-training* dokumen ditransformasikan kedalam *gap sentences. Gap sentences* dilakukan dengan menghapus sebagian kalimat dari dokumen dan menggunakan dokumen sisa untuk memprediksi sebagian kalimat yang telah dihapus.
2. *Masked Language Model*
   Dalam MLM dipilih token tertentu secara acak dari suatu dokumen untuk ditutup atau disebut dengan istilah *mask.* Dan selanjutnya dengan dokumen yang tersisa model akan dilatih untuk memprediksi token yang telah ditutup tersebut.

Model pegasus menggunakan kedua teknik GSG dan MLM selama masa pretraining yang bertujuan untuk memahami makna, pola bahasa dan relasi antar setiap kata. GSG berfokus untuk memahami struktur dan konteks dokumen, sedangkan MLM berfokus untuk mempelajari representasi token[8].

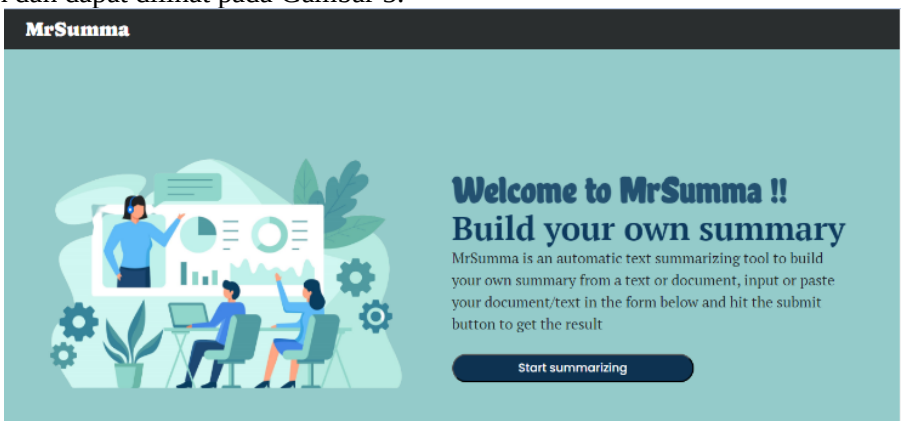
3. Hasil dan Pembahasan

Pada bagian ini berisi analisis, hasil implementasi ataupun pengujian serta pembahasan dari topik penelitian, yang bisa dibuat terlebih dahulu metodologi penelitian. Bagian ini juga merepresentasikan penjelasan yang berupa penjelasan, gambar, tabel dan lainnya.

3.1. *User Interface*

a. Halaman Depan

Tampilan *user interface* halaman depan pada sistem peringkas teks otomatis hanya terdiri dari satu halaman dan dapat dilihat pada Gambar 5.

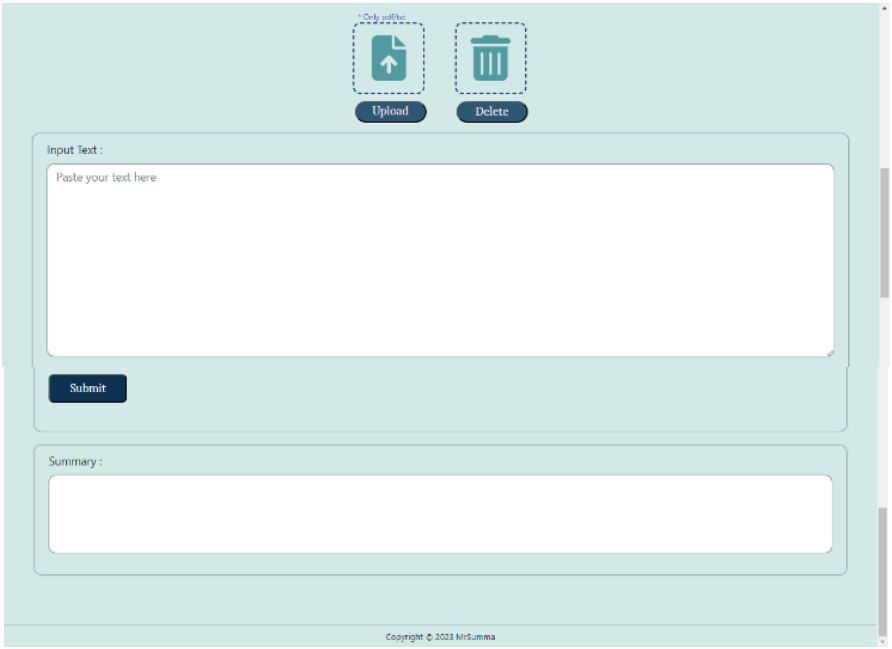


TELEMATIKA MKOM
Volume 15, Nomor 2, September, 2023
Halaman 86-92
ISSN: 2527-9033 (*online*), 2085-725X (cetak)

Gambar 5. Halaman Depan Sistem Peringkas Teks Otomatis

b. Form Peringkas Teks Otomatis

Pada *user interface* terdapat sebuah form untuk peringkas teks otomatis yang dapat digunakan untuk memasukkan dokumen atau teks yang ingin diringkas. Proses *input* dapat dilakukan melalui paste teks pada *textarea* ataupun melalui tombol *upload* dokumen. Selanjutnya untuk mendapatkan hasil ringkasan maka dapat digunakan tombol submit. Dan hasil ringkasana akan tampil pada *form output*. Jika ingin memasukkan teks baru maka *user* dapat menggunakan tombol delete untuk membersihan teks area.Tampilan form peringkas terdapat pada Gambar 6.



Gambar 6. Form Peringkas Teks Otomatis

3.2. Pengujian

Pada bagian ini merupakan hasil ringkasan dari proses pengujian menggunakan matriks *ROUGE (Recall Oriented Understudy for Gisting Evaluation)*. *ROUGE* merupakan standar pengujian yang selama ini digunakan dalam peringkasan. Peringkasan dilakukan pada 100 artikel CNN/Daily Mail yang dilengkapi dengan ringkasan manual sebagai hasil aktual atau sebagai referensi yang digunakan sebagai sampel pengujian.

Tabel 1. Sampel Hasil Pengujian 100 Artikel

Nama Dokumen	F1-Score ROUGE-1	F1-Score ROUGE-2	F1-Score ROUGE-L
doc0	0.518519	0.379747	0.469136
doc1	0.46	0.244898	0.32
doc2	0.142857	0	0.122449
doc3	0.408163	0.170213	0.367347
doc4	0.307692	0.105263	0.205128
doc5	0.273684	0.064516	0.189474
doc6	0.336634	0.141414	0.217822
doc7	0.690476	0.536585	0.547619
doc8	0.47619	0.196721	0.31746
…	…	…	…
doc95	0.242424	0	0.151515
doc96	0.36	0.183673	0.3
doc97	0.347826	0.059701	0.144928
doc98	0.314607	0.068966	0.202247

TELEMATIKA MKOM
Volume 15, Nomor 2, September, 2023
Halaman 86-92
ISSN: 2527-9033 (*online*), 2085-725X (cetak)

Adapun hasil rata–rata pengujian dari 100 artikel yang diperoleh terdapat pada Tabel 2.

Tabel 2. Hasil Keseluruhan Pengujian

	Average	Low	High
F1-Score ROUGE-1	0.34	0.2	0.69
F1-Score ROUGE-2	0.15	0.	0.49
F1-Score ROUGE-L	0.25	0.02	0.66

Hasil menunjukkan nilai *F1-Score* tertinggi 0.69 pada *ROUGE-1*, 0.49 pada *ROUGE-2*, dan 0.66 pada *ROUGE-L*, rata–rata untuk *F1-Score* 0.34 pada *ROUGE-1*, 0.15 pada *ROUGE-2*, dan 0.25 pada *ROUGE-L*.

4. Kesimpulan

Penelitian ini menyimpulkan bahwa implementasi hybrid pada peringkas teks otomatis menunjukkan hasil *F1-Score* 0.34 pada *ROUGE-1*, 0.15 pada *ROUGE-2*, dan 0.25 pada *ROUGE-L*. Nilai ini menunjukkan penurunan jika dibandingkan dengan implementasi teknik abstraktif dengan Pegasus model sebelum digabungkan dengan TextRank. Disisi lain proses peringkasan memakan waktu yang lebih sedikit dibandingkan dengan penerapan teknik abstraktif dasar. Sehingga, dapat disimpulkan bahwa mengaplikasikan pendekatan hybrid dengan kombinasi TextRank dan *Transformer Architecture* memiliki keakuratan lebih kecil dari penerapan *Transformer Architecture* secara murni, namun mengungguli dari segi kecepatan

Daftar Pustaka

[1] H. Bryan *et al.*, "Pemanfaatan *Text Summarization Dengan Support Vector Machine Dan K-Nearest Neighbor* Pada Analisis Sentimen Untuk Mempermudah Pengguna Membaca Review Game Steam," *J. Infra*, vol. 10, no. 1, pp. 31–36, 2022.
[2] Y. J. Kumar, O. S. Goh, H. Basiron, N. H. Choon, and P. C. Suppiah, "A review on automatic text summarization approaches," *J. Comput. Sci.*, vol. 12, no. 4, pp. 178–190, 2016, doi: 10.3844/jcssp.2016.178.190.
[3] A. Raj, S. E. M, and D. P. S, "A Systematic Survey on Multi-document Text Summarization," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 6, pp. 3148–3153, 2021, doi: 10.30534/ijatcse/2021/111062021.
[4] A. N. Ammar and S. Suyanto, "Peringkasan Teks Ekstraktif Menggunakan Binary Firefly Algorithm," *Indo-JC*, vol. 5, no. September, pp. 31–42, 2020, doi: 10.21108/indojc.2020.5.2.440.
[5] L. Pertiwi, "Penerapan Algoritma Text Mining, Steaming Dan Texrank Dalam Peringkasan Bahasa Inggris," *Bimasati*, vol. 1, no. 3, pp. 100–104, 2022.
[6] D. Suleiman and A. Awajan, "Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges," *Math. Probl. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/9365340.
[7] R. Varma, "SJSU ScholarWorks A Hybrid Approach for Multi-document Text Summarization," San Jose State University, 2019.
[8] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," *Proc. 2004 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2004 - A Meet. SIGDAT, a Spec. Interes. Gr. ACL held conjunction with ACL 2004*, vol. 85, pp. 404–411, 2004.
[9] J. Pragantha, Eris, and V. C. M, "Penerapan Algoritma TextRank untuk Automatic Summarization pada Dokumen Berbahasa Indonesia," *Ilmu Tek. dan Komuter*, vol. 1, no. 1, pp. 71–78, 2017.
[10] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.
[11] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization," *37th Int. Conf. Mach. Learn. ICML 2020*, vol. PartF16814, pp. 11265–11276, 2020.