Extractive Text Summarization Terhadap Artikel Berita Indonesia Berbasis Machine Learning

1st Muhammad Raihan Hadwirianto
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
mraihanh@student.telkomuniversity.ac.id

2nd Faqih Hamami
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
faqihhamami@telkomuniversity.ac.id

3rd Oktariani Nurul Pratiwi
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
onurulp@telkomuniversity.ac.id

***Abstrak* — Terciptanya internet, jejaring sosial, forum, dan teknologi informasi yang tersebar secara cepat, menyebabkan interaksi terhadap informasi semakin sulit untuk dipahami, dibuat, dikembangkan, dan disimpan. Dengan luasnya informasi sehingga hampir tidak mungkin untuk seorang pun untuk memproses dan meringkas semua data informasi yang tersedia. Indonesia memiliki literasi yang sangat rendah dari negara lain dengan beberapa faktor seperti tidak membiasakan diri untuk membaca buku dari rumah, perkembangan teknologi yang semakin pesat, minimnya sarana untuk membaca, kurangnya motivasi untuk membaca, dan sifat malas untuk mengembangkan ide. *Automatic text summarization* adalah salah satu alternatif teknologi yang bisa digunakan untuk menyelesaikan masalah tersebut. *Automatic text summarization* merupakan bagian dari bidang *Natural Language Processing* (NLP) yang bertujuan untuk merepresentasikan dokumen teks yang panjang menjadi lebih ringkas, sehingga pengguna dapat dengan mudah memahami informasi dengan cepat. Berbagai metode telah dilakukan untuk mengatasi masalah peringkasan teks otomatis untuk objek berbahasa Indonesia, yaitu berbasis *extractive* dan *abstractive*. Untuk mengatasi masalah ini, pada penelitian digunakan *extractive text summarization* berbasis *machine learning*. Pada penelitian ini menggunakan *dataset* publik yang bisa digunakan untuk penelitian-penelitian selanjutnya. Metode yang digunakan untuk mendapatkan hasil *summarization* dengan menggunakan metode *Word2Vec* dengan penerapan model *Continous Bag-of-Word* (CBOW) dan *Skip-Gram*. Metode yang digunakan untuk evaluasi akurasi hasil ringkasan adalah *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE).**

Kata kunci*— *Automatic Text Summarization, Word2Vec, Continous Bag-of-Words, Skip-Gram, Recall-Oriented Understudy for Gisting Evaluation

I. PENDAHULUAN

Terciptanya internet, jejaring sosial, forum, dan teknologi informasi yang tersebar secara cepat, menyebabkan interaksi terhadap informasi semakin sulit untuk dipahami, dibuat, dikembangkan, dan disimpan. Dengan besarnya akan kebutuhan informasi dan pesatnya perkembangan internet menyebabkan dorongan pertumbuhan situs media *online* di Indonesia [1], [2]. Membaca merupakan salah satu kegiatan yang tidak dapat dipisahkan dari kehidupan manusia, baik membaca buku, majalah, atau sebuah artikel berita. Tetapi, permasalahan akan muncul ketika sebuah teks atau artikel yang akan dibaca memiliki isi yang banyak dan panjang karena membutuhkan waktu yang cukup lama untuk dapat memahami isi dari teks tersebut [3]. Hal ini menjadi sebuah tantangan di tengah rendahnya tingkat literasi di Indonesia.

Indonesia memiliki literasi yang sangat rendah dari negara lain. Berdasarkan hasil survei yang dilakukan oleh *Program for International Student Assessment* (PISA) yang diterbitkan *Organization for Economic Co-operation and Development* (OECD) pada tahun 2019, Indonesia berada di peringkat 62 dari 70 negara, dimana Indonesia berada di peringkat 10 terbawah. [4]. Beberapa faktor yang menyebabkan rendahnya literasi di Indonesia adalah tidak membiasakan diri untuk membaca buku dari rumah, perkembangan teknologi yang semakin pesat, minimnya sarana untuk membaca, kurangnya motivasi untuk membaca, dan sifat malas untuk mengembangkan ide. [5].

Automatic text summarization adalah salah satu alternatif teknologi yang bisa digunakan untuk menyelesaikan masalah tersebut [6]. *Automatic text summarization* merupakan bagian dari bidang *Natural Language Processing* (NLP) yang bertujuan untuk merepresentasikan dokumen teks yang panjang menjadi lebih ringkas, sehingga pengguna dapat dengan mudah memahami informasi dengan cepat [7].

Text Summarization terdiri dari dua metode, yaitu *abstractive text summarization* dan *extractive text summarization*. *Abstractive text summarization* memparafrasakan teks secara keseluruhan sehingga ringkasan tersebut memiliki kosa kata yang bervariasi. Sedangkan, *extractive text summarization* melibatkan pemilihan kalimat penting dari naskah asli dan menggabungkannya menjadi kalimat yang lebih pendek tanpa kehilangan informasi penting. [8]–[10].

Sudah banyak penelitian tentang *automatic text summarization* berbasis bahasa Indonesia terhadap sebuah artikel berita dan dokumen secara ekstraktif maupun abstraktif, tetapi beberapa peneliti menggunakan *dataset* yang dikumpulkan sendiri dan tidak dipublikasikan sehingga tidak mudah untuk digunakan sebagai *benchmark text summarization* berbahasa Indonesia. Dari permasalahan di

atas, peneliti ingin menyelidiki *extractive text summarization* berbahasa Indonesia berbasis *machine learning*. *Dataset* yang akan digunakan dalam penelitian ini adalah Indosum yang dibuat oleh Kurniawan & Louvan [11] yang tersedia secara publik. Dengan ini, peneliti mengambil judul: "*Extractive Text Summarization* Terhadap Artikel Berita Indonesia Berbasis *Machine Learning*".

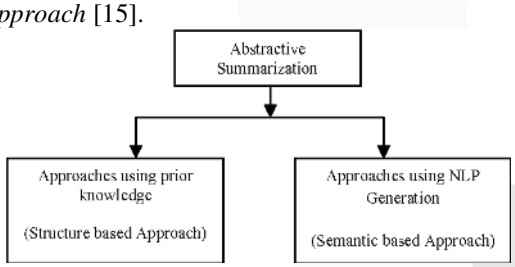
II. KAJIAN TEORI

A. *Text Summarization*

Text Summarization adalah sebuah proses mengubah teks yang panjang menjadi kelompok kalimat yang lebih kecil, akurat, dan mudah dimengerti dan dapat memberikan informasi yang diperlukan kepada pembaca dengan jumlah kata yang lebih ringkas [8]. *Text Summarization* bertujuan untuk menghasilkan bentuk teks yang lebih ringkas yang dapat menyampaikan informasi penting dari teks asli yang umumnya disampaikan dalam bentuk yang lebih panjang [12]. Dengan bertambahnya data, waktu, dan uang yang diperlukan untuk mengolah data, maka perlu adanya data yang perlu di rangkum dengan semua informasi penting dari data asli yang dapat di akses dengan mudah. Metode *Text Summarization* dibagi menjadi dua, yaitu *abstractive text summarization* dan *extractive text summarization* [9], [10].

B. *Abstractive Text Summarization*

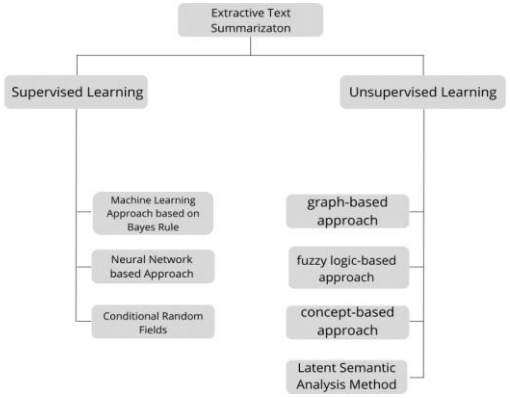
Abstractive Text Summarization mengandalkan *Natural Language Processing* (NLP) untuk menghasilkan ringkasan yang singkat dan padat, mendapatkan ide-ide penting dari sumber teks yang dapat berpotensi menghasilkan frasa dan kalimat baru yang kemungkinan tidak muncul dalam sumber teks [13], [14]. *Abstractive text summarization* menampilkan informasi yang sudah diringkas dalam bentuk koheren yang mudah dibaca dan benar secara tata Bahasa dan memiliki dua pendekatan, yaitu *structure-based approach* dan *semantic based approach* [15].



GAMBAR 1
Abstractive Text Summarization Approach [15]

C. *Extractive Text Summarization*

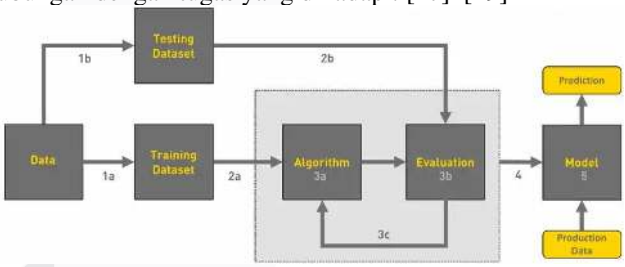
Extractive Text Summarization didasarkan metode pembobotan kalimat dengan memperoleh kata dan frasa dalam teks dengan frekuensinya [16]. Setelah itu, akan dilakukan pemilihan kalimat dengan nilai tertinggi dari dokumen dan akan menghapus sisa-sisa kalimat yang tidak berguna. Metode *extractive text summarization* dapat diklasifikasikan menjadi dua, yaitu *Unsupervised Learning* dan *Supervised Learning*. *Unsupervised Learning* memiliki metode pendekatan seperti *graph-based approach*, *fuzzy logic-based approach*, *concept-based approach*, dan *Latent Semantic Analysis Method*, Sedangkan Supervised Learning memiliki pendekatan seperti *Machine Learning Approach based on Bayes Rule*, *Neural Network based Approach*, dan *Conditional Random Fields* [10].



GAMBAR 2
Extractive Text Summarization Approaches [10]

D. *Machine Learning*

Machine Learning adalah sub disiplin kecerdasan buatan (AI) yang berupaya meniru bagaimana otak manusia memahami dan berinteraksi dengan dunia dan hubungan antara objek dan proses di dunia. *Machine Learning* berkisar pada konsep membuat program komputer yang meningkatkan kinerja dengan secara otomatis belajar dan beradaptasi dengan pengalaman. Ini pada dasarnya bekerja pada banyak hipotesis dan kemudian menemukan yang terbaik yang sesuai dengan data yang diamati. Tujuan dari setiap *Machine Learning*, bagaimanapun, dapat dirumuskan mengingat pengamatan (mungkin, parsial) dari keadaan sistem dan model parametrik yang menghasilkan prediksi berdasarkan keadaan sistem yang diamati, temukan parameter terbaik yang ditetapkan untuk model yang diberikan untuk memaksimalkan akurasi prediksi sehubungan dengan tugas yang dihadapi. [17]–[19]



GAMBAR 3
Ilustrasi *Workflow Machine Learning* [20]

E. *K-Means Clustering*

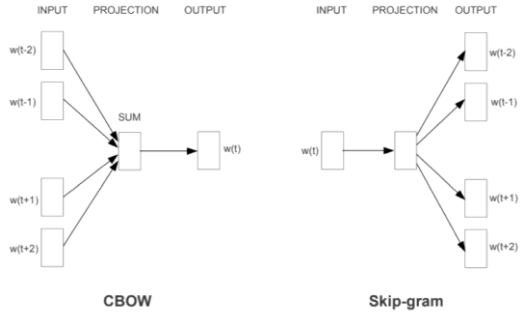
K-Means Clustering adalah sebuah algoritma *unsupervised machine learning* yang digunakan untuk mengelompokkan data. [21]. *K-Means* bertujuan untuk mengelompokkan data dengan memaksimalkan kemiripan karakteristik setiap data di dalam sebuah kluster dan juga memaksimalkan perbedaannya dengan kluster lain. [22]. Namun, *K-Means* memiliki kekurangan berupa ketergantungannya pada pengelompokan awal. Jika pemilihan *centroid* dalam *clustering* tidak tepat, hasil dari *clustering* akan menjadi optimal secara lokal, sehingga memungkinkan *centroid* awal yang sangat dibutuhkan [23].

F. *Word2Vec*

Word2Vec adalah algoritma *word embedding* yang memetakan setiap kata dalam teks menjadi sebuah vektor. [24]. Metode ini diperkenalkan oleh Tomas Mikolov dan rekan-rekannya dari Google pada tahun 2013 [25]. Cara kerja Word2Vec adalah algoritma representasi vektor kata yang dapat mengelompokkan kata-kata yang serupa ke dalam

vektor yang sama. Keuntungan dari *Word2Vec* adalah dapat merepresentasikan kemiripan kontekstual dari dua kata dalam vektor yang dihasilkan. [26]. Meskipun *Word2Vec* tidak secara langsung terkait dengan *text summarization*, representasi vektor kata yang dihasilkannya sering digunakan sebagai fitur dalam tugas-tugas seperti *text summarization*. *Word2Vec* bisa masuk dalam kategori *supervised learning* ataupun *unsupervised learning* [27].

Word2Vec memiliki dua arsitektur, yaitu *Continuous Bag-Of-Word* (CBOW) dan *Skip-Gram*. *CBOW* adalah model yang menggunakan konteks untuk memprediksi suatu target kata, sedangkan *Skip-Gram* adalah model yang menggunakan suatu kata untuk memprediksi target konteks [24].



GAMBAR 4
Model CBOW dan Skip-Gram [28]

G. *Recall-Oriented Understudy for Gisting Evaluation (ROUGE)*

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) adalah sebuah metode penilaian yang digunakan untuk mengevaluasi sistem peringkasan secara otomatis dengan membandingkan ringkasan yang dihasilkan dengan ringkasan yang dibuat oleh referensi atau ringkasan yang ditulis oleh manusia [29].

Nilai *ROUGE* dilaporkan dalam bentuk *precision, recall,* dan *f-measure*. *Precision* mengukur nilai pembagian antara *n-grams* dalam ringkasan yang dihasilkan juga muncul di referensi. Nilai *precision* dapat dilihat dari rumus berikut [30].

$$\frac{jumlah\ kata\ tumpang\ tindih}{jumlah\ kata\ di\ ringkasan\ otomatis}$$

Recall mengukur nilai pembagian antara *n-grams* dalam referensi yang termasuk dalam ringkasan yang dihasilkan. Nilai *recall* didapatkan dari rumus berikut [30].

$$\frac{jumlah\ kata\ tumpang\ tindih}{jumlah\ kata\ di\ ringkasan\ manusia}$$

f-measure adalah rata – rata dari *precision* dan *recall* yang dapat memberikan hasil evaluasi yang ideal [31]. Nilai *f-measure* didapatkan dari rumus berikut [30].

$$\frac{2 \times Precision \times Recall}{(Precision \times Recall)}$$

ROUGE menggunakan beberapa *metric* yang digunakan untuk menjadi standar bagus atau tidaknya hasil suatu ringkasan yaitu, ROUGE-N, dan ROUGE-L.

ROUGE-N digunakan untuk membandingkan *n-grams* antara ringkasan yang dihasilkan dan ringkasan referensi. ROUGE-1 digunakan untuk membandingkan *unigram* (setiap kata) antara ringkasan yang dihasilkan dan ringkasan manual. Sedangkan ROUGE-2 digunakan untuk membandingkan *bigrams* (dua kata yang berurutan atau kombinasi dua kata) antara ringkasan yang dihasilkan dan ringkasan manual. Rumus untuk menghitung ROUGE-N dapat dilihat pada persamaan berikut [29].

$$\frac{\sum_{S\in\{referencesummaries\}gram_n}\sum Count_{match}(Gram_n)}{\sum_{S\in\{referencesummaries\}gram_n}\sum Count(Gram_n)}$$

ROUGE-L digunakan untuk mengukur perbandingan LCS (*Longest Common Subsequence*) antara ringkasan yang dihasilkan dan ringkasan referensi. *Metric* ini digunakan untuk evaluasi seberapa baiknya hasil ringkasan yang dihasilkan mempertahankan urutan dan struktur ringkasan referensi[6]. Rumus untuk menghitung ROUGE-L *precision*, *recall*, *f-measure* dapat dilihat pada persamaan berikut [29].

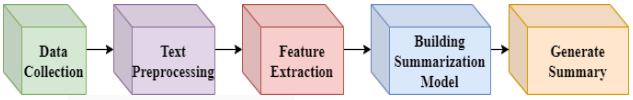
$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

III. METODE

Untuk sistematika penyelesaian, tahap-tahapannya bisa dilihat pada Gambar di bawah ini.



GAMBAR 5
Metode Penelitian

A. *Data Collection*

Data collection adalah sebuah proses pengumpulan informasi tentang variabel – variabel yang diinginkan, dengan metode sistematis yang memungkinkan seseorang untuk menjawab berbagai pertanyaan yang ditetapkan, menguji hipotesis dan mengevaluasi hasil [32]. Data yang akan digunakan dalam penelitian ini berupa sebuah *dataset* IndoSum yang dibuat oleh Kurniawan & Louvan [11] sebagai *benchmark text summarization* bahasa Indonesia. *Dataset* ini terdiri dari 20 ribu artikel berita Bahasa Indonesia dalam bentuk format *file JSON*. *Dataset* ini berisi tentang isi berita, ringkasan, kategori, sumber, dan sumber *url*..

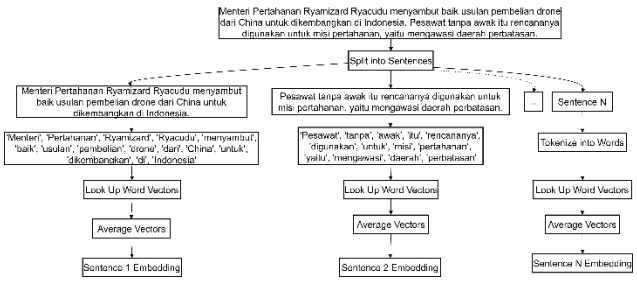
B. *Text Preprocessing*

Text Preprocessing adalah sebuah metode yang digunakan untuk mempersiapkan teks yang tidak terstruktur menjadi teks atau data yang siap digunakan atau diolah [9]. Dengan melakukan *text preprocessing*, data mentah yang tersedia dalam *dataset* dapat diolah. *Text Preprocessing* yang dilakukan memiliki beberapa tahapan yaitu, *case folding, remove punctuation,* dan *tokenizing*.

C. *Feature Extraction*

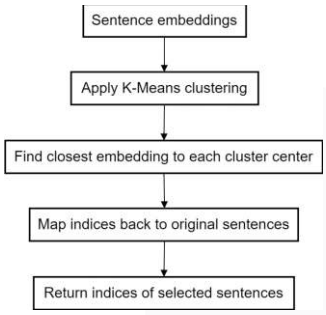
Feature Extraction dilakukan setelah tahap pre-processing yang dimana *feature extraction* adalah tahap pemilihan fitur yang akan digunakan untuk proses sumarisasi teks. *Feature extraction* yang dilakukan oleh peneliti ada 2, yaitu *Sentence Embedding* dan *Clustering*. *Sentence Embedding* dilakukan untuk membuat *word embeddings* dari kalimat yang diberikan. Lalu, merepresentasikan makna dari

sebuah kalimat sebagai vektor. Tahap ini dilakukan dengan menggunakan model *trained word2vec*.



GAMBAR *6*
Ilustrasi Sentence Embedding

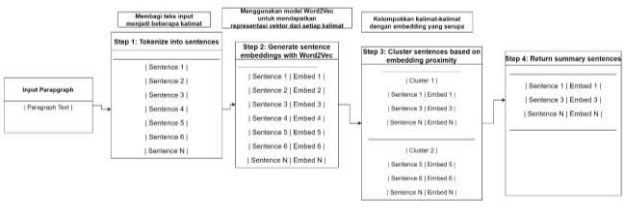
Clustering dilakukan untuk mengelompokkan kalimat-kalimat berdasarkan kata-kata yang terkandung di dalamnya. Hal ini dilakukan dengan mengambil serangkaian kalimat, di mana setiap kalimat diwakili oleh satu kumpulan angka. Lalu, angka-angka ini menggambarkan kata-kata dalam kalimat dengan cara yang dapat dimengerti oleh algoritma yang digunakan. *Clustering* ini menggunakan algoritma *K-Means* untuk melakukan pengelompokkan teks. Algoritma ini bekerja dengan memilih sejumlah kluster, yang menentukan berapa banyak kluster yang akan dibagi ke dalam kalimat-kalimat tersebut.



GAMBAR *7*
Ilustrasi Clstering

D. Bulding Word2Vec Summarization

Pada tahap ini, hasil dari *clustering* akan menghasilkan kluster yang menunjukkan hubungan antara kalimat-kalimat dalam teks. Setiap kluster terdiri dari kalimat yang memiliki kesamaan berdasarkan hasil *clustering*. Dalam proses tahap *summarization* ini, langkah selanjutnya yaitu memilih kalimat yang paling representatif dari setiap kluster sebagai bagian dari ringkasan. Proses ini dilakukan dengan memilih kalimat yang paling dekat dengan *centroid* (nilai pusat) dari setiap kluster. Dengan memilih kalimat yang paling dekat dengan nilai pusat (centroid), diharapkan kalimat tersebut merepresentasikan topik dari kluster tersebut.



GAMBAR *8*
Ilustrasi Tahap Summarization Secara Keseluruhan

Dengan mengambil kalimat dari setiap kluster sebagai ringkasan, proses ini secara efektif dapat mengekstrak kalimat-kalimat yang penting yang merepresentasikan makna keseluruhan dari teks asli. Panjang hasil ringkasan ini ditentukan oleh jumlah kluster yang digunakan. Semakin banyak jumlah kluster, semakin banyak kalimat yang dipilih untuk ringkasan, sehingga ringkasan akan lebih panjang dan lebih rinci. Rasio jumlah kluster dengan jumlah total kalimat asli menentukan persentase kalimat asli yang diekstrak ke dalam ringkasan. Dalam proses *summarization* ini akan menggunakan rasio 20% dan 40% dari *input* teks untuk menghasilkan sebuah ringkasan. Rasio ini digunakan untuk mengatur berapa persentase kalimat yang akan diambil dari teks asli sebagai bagian dari ringkasan. Rasio ini menentukan berapa banyak kluster yang akan digunakan dalam proses *summarization*.

IV. HASIL DAN PEMBAHASAN

Setelah membuat *summarization*, maka dapat dilakukan *generate summary* atau menghasilkan ringkasan terhadap artikel berita Indonesia. Proses *summarization* ini dilakukan dengan menggunakan model *trained Word2Vec* yang telah dilatih dengan menggunakan *corpus* Wikipedia Indonesia [33]. Dalam proses *summary* ini, ada 2 skenario yang dilakukan saat proses *summarization*, yaitu Skenario 1 dengan Penerapan model *trained* Word2Vec CBOW (Continous Bag-of-Words), dan Skenario 2 dengan penerapan model *trained* Word2Vec Skip-Gram.

A. Evaluasi ROUGE

Untuk metode evaluasi yang dilakukan adalah dengan menggunakan metode *ROUGE. Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) adalah sebuah metode penilaian yang digunakan untuk mengevaluasi sistem peringkasan secara otomatis dengan membandingkan ringkasan yang dihasilkan dengan ringkasan yang dibuat oleh referensi atau ringkasan yang ditulis oleh manusia [29]. Untuk penelitian ini, *metric* ROUGE yang akan digunakan oleh peneliti adalah ROUGE-N dan ROUGE-L. Kisaran hasil keluaran dari *ROUGE* sendiri adalah antara 0 dan 1. Semakin dekat hasil evaluasi dengan angka 1, maka hasil rangkuman semakin mirip dengan hasil rangkuman buatan manusia.

TABEL *1*
Hasil Penilaian ROUGE Model CBOW Dengan Rasio 40%

Scores	Rouge-1	Rouge-2	Rouge-l
Recall	0.670854	0.514328	0.654920
Precision	0.398481	0.279508	0.388795
F-measure	0.489486	0.352876	0.477789

TABEL *2*
Hasil Penilaian ROUGE Model CBOW Dengan Rasio 20%

Scores	Rouge-1	Rouge-2	Rouge-l
Recall	0.551787	0.407851	0.534674
Precision	0.482986	0.350126	0.467811
F-measure	0.505234	0.367277	0.489521

TABEL *3*
Hasil Penilaian ROUGE Model Skip-Gram Dengan Rasio 40%

Scores	Rouge-1	Rouge-2	Rouge-l
Recall	0.677995	0.519042	0.660808
Precision	0.400748	0.282117	0.389839
F-measure	0.492264	0.354803	0.479248

TABEL 4
Hasil Penilaian ROUGE Model Skip-Gram Dengan Rasio 20%

Scores	Rouge-1	Rouge-2	Rouge-l
Recall	0.536606	0.391615	0.519518
Precision	0.479207	0.343729	0.463698
F-measure	0.496307	0.356373	0.480395

Dari hasil penilaian *ROUGE* di atas, dapat dilihat bahwa perbedaan nilai antara dua skenario di atas memiliki perbedaan yang tidak terlalu signifikan berdasarkan rasio yang digunakan. Nilai *recall* dengan rasio 40% pada kedua model memiliki yang sedikit lebih baik dari rasio 20%, dimana menunjukkan kemampuan yang lebih baik dalam mencakup kalimat-kalimat penting dari teks asli. Tetapi, pada nilai *precision*, dengan rasio 20% pada kedua model memiliki hasil yang lebih baik dari pada dengan rasio 40%, yang dimana lebih selektif dalam menyertakan hanya kalimat – kalimat yang lebih relevan dan ada di teks asli. Untuk perbedaan nilai *f-measure* dengan model Skip-Gram tidak terlalu besar, tetapi nilai *f-measure* pada model CBOW dengan rasio 20% sedikit lebih tinggi daripada dengan rasio 40%. Sehingga hasil kualitas ringkasan bisa dikatakan kurang lebih bisa berbeda ataupun sama tergantung dengan penerapan model yang digunakan.

V. KESIMPULAN

Tujuan utama dari penelitian ini adalah untuk mengatasi masalah rendahnya tingkat literasi di Indonesia dengan menyediakan ringkasan artikel berita yang ringkas dan mudah dipahami oleh pengguna. Penelitian ini menggunakan metode *Word2Vec* dengan penerapan model *CBOW* dan *Skip-Gram* untuk melakukan representasi vektor kata dalam kalimat-kalimat artikel berita. Penggunaan metode *clustering*, yaitu algoritma *KMeans*, digunakan untuk mengelompokkan kalimat-kalimat serupa dan mencari kalimat yang paling mewakili konten dari teks masukan. Hasil dari *Extractive Text Summarization* ini digunakan untuk menghasilkan ringkasan yang koheren dan efektif. Selain itu, metode evaluasi yang digunakan oleh peneliti adalah *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE).

Hasil akurasi yang dihasilkan baik dengan model CBOW maupun Skip-Gram menunjukkan potensi yang baik dalam menghasilkan sebuah ringkasan yang relevan. Penerapan dengan menggunakan model CBOW dan Skip-Gram pada *extractive text summarization* mampu menghasilkan ringkasan yang bisa mengangkat dan mencerminkan informasi penting dari artikel berita yang di analisa. Meskipun masih terdapat tantangan saat menghasilkan ringkasan, penelitian ini diharapkan dapat memberikan manfaat dalam meningkatkan literasi di Indonesia dengan menyediakan ringkasan artikel berita yang mudah diakses dan dimengerti.

REFERENSI

[1] Rupal. Bhargava and Yashvardhan. Sharma, "Deep Extractive Text Summarization," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 138–146. doi: 10.1016/j.procs.2020.03.191.

[2] B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Candidate sentence selection for extractive text summarization," *Inf Process Manag*, vol. 57, no. 6, Nov. 2020, doi: 10.1016/j.ipm.2020.102359.

[3] N. Savanti Widya Gotami, Indriati, and R. Kartika Dewi, "Peringkasan Teks Otomatis Secara Ekstraktif Pada Artikel Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Latent Semantic Analysis," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[4] B. U. Ilham, "Harbuknas 2022 : Literasi Indonesia Peringkat Ke-62 Dari 70 Negara," 2022. https://bisniskumkm.com/harbuknas-2022-literasi-indonesia-peringkat-ke-62-dari-70-negara/ (accessed Jan. 06, 2023).

[5] Jessica, "5 Penyebab Rendahnya Budaya Literasi di Indonesia," Jul. 10, 2017. https://www.educenter.id/5-penyebab-rendahnya-budaya-literasi-di-indonesia/ (accessed Jan. 06, 2023).

[6] M. M. Mubarok, "INDONESIAN ABSTRACTIVE NEWS SUMMARIZATION BERBASIS DEEP LEARNING DENGAN METODE SEQUENCE-TO-SEQUENCE LONG SHORT-TERM MEMORY," 2021.

[7] A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders," *Expert Syst Appl*, vol. 129, pp. 200–215, Sep. 2019, doi: 10.1016/J.ESWA.2019.03.045.

[8] S. Singh, A. Singh, S. Majumder, A. Sawhney, D. Krishnan, and S. Deshmukh, "Extractive Text Summarization Techniques of News Articles: Issues, Challenges and Approaches," in *Proceedings - International Conference on Vision Towards Emerging Trends in Communication and Networking, ViTECoN 2019*, 2019. doi: 10.1109/ViTECoN.2019.8899706.

[9] 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.

[10] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *International Conference on Computer, Communication, and Signal Processing: Special Focus on IoT, ICCCSP 2017*, 2017. doi: 10.1109/ICCCSP.2017.7944061.

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

[12] I. R. Musyaffanto, G. Budi Herwanto, and M. Riasetiawan, "Automatic extractive text summarization for indonesian news articles using maximal marginal relevance and non-negative matrix factorization," in *Proceedings - 2019 5th International Conference on Science and Technology, ICST 2019*, 2019. doi: 10.1109/ICST47872.2019.9166376.

[13] L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li, "Generative Adversarial Network for Abstractive

Text Summarization *," 2017, Accessed: Jan. 06, 2023. [Online]. Available: www.aaai.org

[14] H. T. Le and T. M. Le, "An approach to abstractive text summarization," in *2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, IEEE, Dec. 2013, pp. 371–376. doi: 10.1109/SOCPAR.2013.7054161.

[15] N. Moratanch and S. Chitrakala, "A survey on abstractive text summarization," in *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, IEEE, Mar. 2016, pp. 1–7. doi: 10.1109/ICCPCT.2016.7530193.

[16] F. Horasan and B. Bilen, "Extractive Text Summarization System for News Texts," *International Journal of Applied Mathematics Electronics and Computers*, 2020, doi: 10.18100/ijamec.800905.

[17] R. Bonetto and V. Latzko, "Machine learning," *Computing in Communication Networks: From Theory to Practice*, pp. 135–167, Jan. 2020, doi: 10.1016/B978-0-12-820488-7.00021-9.

[18] K. Jain, M. Chawla, A. Gadhwal, R. Jain, and P. Nagrath, "Age and Gender Prediction Using Convolutional Neural Network," in *Lecture Notes in Networks and Systems*, 2020. doi: 10.1007/978-981-15-3369-3_19.

[19] C. E. Lawson *et al.*, "Machine learning for metabolic engineering: A review," *Metabolic Engineering*, vol. 63. 2021. doi: 10.1016/j.ymben.2020.10.005.

[20] A. Pant, "Workflow of a Machine Learning Project," 2019. https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94 (accessed Jan. 07, 2023).

[21] K. Shetty and J. S. Kallimani, "Automatic extractive text summarization using K-means clustering," in *International Conference on Electrical, Electronics, Communication Computer Technologies and Optimization Techniques, ICEECCOT 2017*, 2018. doi: 10.1109/ICEECCOT.2017.8284627.

[22] 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.

[23] A. Firdaus, N. Yusliani, and D. Rodiah, "Text Summarization using K-Means Algorithm," *Sriwijaya Journal of Informatic and Applications*, vol. 2, no. 2, pp. 16–22, 2021, [Online]. Available: http://sjia.ejournal.unsri.ac.id

[24] A. Nurdin, B. Anggo, S. Aji, A. Bustamin, and Z. Abidin, "PERBANDINGAN KINERJA WORD EMBEDDING WORD2VEC, GLOVE, DAN FASTTEXT PADA KLASIFIKASI TEKS," *Jurnal Tekno Kompak*, vol. 14, no. 2, pp. 74–79, Aug. 2020, doi: 10.33365/JTK.V14I2.732.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, Jan. 2013, Accessed: Jul. 20, 2023. [Online]. Available: https://arxiv.org/abs/1301.3781v3

[26] M. S. Asramanggala, "DATA TRAIN YANG OPTIMAL DALAM PENDETEKSIAN BERITA HOAX BAHASA INDONESIA MENGGUNAKAN SVM DAN WORD2VEC," 2023, Accessed: Aug. 29, 2023. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/198540/slug/data-train-yang-optimal-dalam-pendeteksian-berita-hoax-bahasa-indonesia-menggunakan-svm-dan-word2vec.html

[27] M. G. Adrian, "Efektifitas Word Embedding GloVe dan Word2Vec dalam Pendeteksian Berita Hoax Bahasa Indonesia Menggunakan LSTM," 2023, Accessed: Jul. 28, 2023. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/198502/slug/efektifitas-word-embedding-glove-dan-word2vec-dalam-pendeteksian-berita-hoax-bahasa-indonesia-menggunakan-lstm.html

[28] T. Mikolov and Q. V Le, "Exploiting Similarities among Languages for Machine Translation", Accessed: Jul. 28, 2023. [Online]. Available: https://code.google.com/p/word2vec/

[29] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," 2004.

[30] A. N. Ammar and S. Suyanto, "Peringkasan Teks Ekstraktif Menggunakan Binary Firefly Algorithm," *Indonesia Journal on Computing (Indo-JC)*, vol. 5, no. 2, pp. 31–42, Oct. 2020, doi: 10.34818/INDOJC.2020.5.2.440.

[31] M. A. Saputra, "Peringkas Teks Otomatis Bahasa Indonesia secara Abstraktif Menggunakan Metode Long Short-Term Memory," 2021, Accessed: Jan. 06, 2023. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/167769/slug/peringkas-teks-otomatis-bahasa-indonesia-secara-abstraktif-menggunakan-metode-long-short-term-memory.html

[32] S. Muhammad and S. Kabir, "METHODS OF DATA COLLECTION Article View project," 2016. [Online]. Available: https://www.researchgate.net/publication/325846997

[33] D. Nugraha K, "Membuat Model Word2Vec Bahasa Indonesia dari Wikipedia Menggunakan Gensim," Jun. 02, 2018. https://medium.com/@diekanugraha/membuat-model-word2vec-bahasa-indonesia-dari-wikipedia-menggunakan-gensim-e5745b98714d (accessed Jul. 28, 2023).