Jurnal Sistem Informasi, Teknologi Informasi, dan Edukasi Sistem Informasi (JUST-SI)
Vol. 5, No. 1, Agustus 2024, hlm.34-41

IMPLEMENTASI METODE TEXTRANK DAN NAMED ENTITY RECOGNITION UNTUK EKSTRAKSI KATA KUNCI PADA MEDIA ONLINE BERITA

Muhammad Theofany Aulia Anwar*[1], Satrio Hadi Wijoyo [2], Widhy Hayuhardhika Nugraha Putra[3]

[1,2,3]Universitas Brawijaya, Kota Malang
Email: [1]theofany007@student.ub.ac.id, [2]satriohadi@ub.ac.id, [3]widhy@ub.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 18 April 2024, diterima untuk diterbitkan: 13 Agustus 2024)

Abstrak

Kata kunci adalah bagian penting untuk memahami isi berita secara singkat dan mendukung indeksasi serta pencarian, proses identifikasi kata kunci yang efisien dan akurat sering kali menjadi tantangan dalam pengelolaan konten digital. Penelitian ini bertujuan untuk meningkatkan proses identifikasi kata kunci yang relevan dalam artikel berita online dengan memanfaatkan metode TextRank dan *Named Entity Recognition* (NER). NER digunakan untuk mengenali dan mengklasifikasikan entitas penting dalam teks, sementara TextRank, yang merupakan algoritma berbasis graf, digunakan untuk menentukan pentingnya kata berdasarkan struktur jaringan mereka. Gabungan dari kedua metode ini diharapkan dapat meningkatkan akurasi ekstraksi kata kunci. Teknik NER yang diimplementasikan adalah model bahasa Indonesia pada spaCy, yang dilatih khusus untuk tujuan ini. Selain itu, TF-IDF (*Term Frequency-Inverse Document Frequency*) digunakan untuk pembobotan kata dalam penerapan algoritma TextRank. Pada penelitian sebelumnya telah dilakukan ekstraksi kata kunci menggunakan kombinasi TextRank dan NER dalam bahasa Inggris, penelitian ini mengarah pada penggunaan kedua metode tersebut untuk mengekstraksi kata kunci dalam bahasa Indonesia, menggunakan data berita *online* dari Times Indonesia. Dari penelitian ini dapat dibuktikan, kinerja gabungan metode TextRank dan NER dalam mengekstraksi kata kunci dari artikel berita lebih baik dibandingkan dengan penggunaan TextRank secara tunggal. Hal ini dapat dilihat dari nilai rata-rata *recall, precision, f-measure,* dan *accuracy* yang dihasilkan dari eksperimen dengan 300 artikel dan *weight multiplier* 2 dengan nilai masing-masing 0.652, 0.645, 0.648, 0.505. Secara kesimpulan, integrasi TextRank dan NER dapat secara signifikan meningkatkan kualitas ekstraksi kata kunci dari artikel berita *online*.

Kata kunci: *TextRank, Named Entity Recognition, ekstraksi kata kunci, berita online, spaCy*

IMPLEMENTATION OF TEXTRANK AND NAMED ENTITY RECOGNITION METHODS FOR KEYWORD EXTRACTION IN ONLINE NEWS MEDIA

Abstract

Keywords are an important part of understanding the content of news briefly and supporting indexation and search, an efficient and accurate keyword identification process is often a challenge in digital content management. This research aims to improve the process of identifying relevant keywords in online news articles by utilizing TextRank and Named Entity Recognition (NER) methods. NER is used to recognize and classify important entities in text,

while TextRank, which is a graph-based algorithm, is used to determine the importance of words based on their network structure. The combination of these two methods is expected to improve the accuracy of keyword extraction. The implemented NER technique is the Indonesian language model on spaCy, which is specially trained for this purpose. In addition, TF-IDF (Term Frequency-Inverse Document Frequency) is used for word weighting in the application of the TextRank algorithm. While previous research has conducted keyword extraction using a combination of TextRank and NER in English, this research leads to the use of both methods to extract keywords in Indonesian, using online news data from Times Indonesia. From this research, it can be proven that the combined performance of TextRank and NER methods in extracting keywords from news articles is better than the use of TextRank alone. This can be seen from the average values of recall, precision, f-measure, and accuracy generated from experiments with 300 articles and weight multiplier 2 with values of 0.652, 0.645, 0.648, 0.505, respectively. In conclusion, the integration of TextRank and NER can significantly improve the quality of keyword extraction from online news articles.

Keywords: *TextRank, Named Entity Recognition, keyword extraction, online news, spaCy*

1. PENDAHULUAN

Di era digitalisasi yang terus berkembang, akses terhadap informasi dan berita mengalami peningkatan yang signifikan. Media berita online, portal berita, dan platform media sosial menyajikan konten berita secara terus-menerus. Kondisi ini, meskipun menguntungkan dalam hal ketersediaan informasi, juga memunculkan tantangan dalam pengelolaan informasi yang berlimpah agar menjadi lebih relevan dan mudah diakses. Peran pengelolaan kata kunci menjadi krusial dalam menangani masalah ini, mengingat kata kunci merupakan elemen penting yang merefleksikan inti dari sebuah teks, memudahkan pengguna dalam memahami isi berita secara singkat, dan berperan penting dalam indeksasi serta pencarian berita, yang merupakan fundamental dari manajemen berita digital.

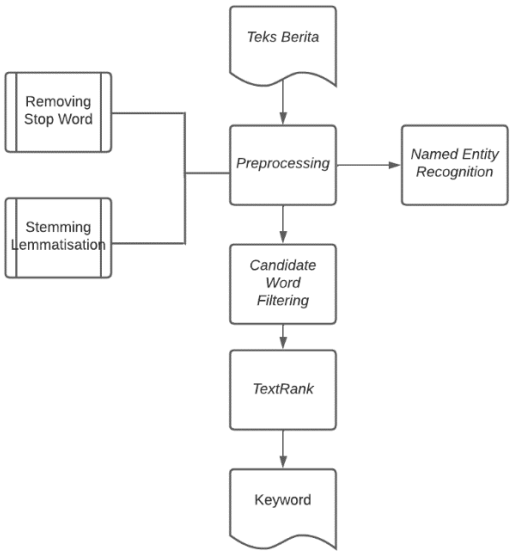
Proses ekstraksi kata kunci secara tradisional seringkali dilakukan secara manual, yang tidak hanya memakan waktu, tetapi juga bersifat subyektif. Pendekatan ini sering kali tidak efisien dan menghasilkan ketidakobjektifan serta inkonsistensi dalam hasilnya. Salah satu metode untuk mengatasi hal tersebut adalah ekstraksi menggunakan metode TextRank, Namun, dengan berkembangnya kebutuhan akan ekstraksi informasi yang lebih spesifik dan mendalam, penelitian ini mengintegrasikan metode *Named Entity Recognition* (NER). NER berperan dalam mengidentifikasi dan klasifikasi entitas bernama dalam teks, seperti nama orang, organisasi, lokasi, dan lainnya. Integrasi NER dalam proses ekstraksi kata kunci memungkinkan pemahaman yang lebih komprehensif terhadap konten, sekaligus meningkatkan relevansi dan akurasi informasi yang dihasilkan.

Sebelum melakukan penelitian terkait penulis menemukan beberapa penelitian terdahulu yang dapat membantu penulis menyelesaikan penelitian ini. Seperti penelitian menurut Lu Yao et al. (2019), penggabungan TF-IDF dan algoritma TextRank untuk mengekstrak kata kunci dari teks dengan membangun model grafik kata, menghitung frekuensi kata frekuensi kata dan frekuensi dokumen terbalik, dan mempertimbangkan bobot dari posisi berita utama. Penelitian ini menggunakan data berupa 1230 koleksi berita dari Sina news yang masing-masing telah ditandai dengan kata kunci. Penelitian ini menggunakan *Recall, Precision* dan *F-measure* sebagai kriteria untuk mengevaluasi hasil eksperimen dan berhasil menunjukkan bahwa kombinasi TextRank dan algoritma TF-IDF serta pemilihan bobot judul yang tepat yang tepat dapat secara efektif meningkatkan efisiensi ekstraksi kata kunci.

Pada penelitian ini dilakukan pendekatan serupa namun dengan beberapa penyesuaian dan perluasan. Beberapa perluasan yang dilakukan adalah dengan menggunakan dataset bahasa yang berbeda pada berita dan menyelaraskan lebih lanjut algoritma TextRank dengan konteks data yang spesifik. Selain itu terdapat tambahan metrik evaluasi *accuracy* agar terlihat lebih jelas perbedaan antara hasil ekstraksi kata kunci dengan menggunakan NER dan yang tidak menggunakan NER. Penyesuaian ini dibuat untuk memastikan bahwa metode yang diusulkan dapat menangani perbedaan pada data.

2. METODE PENELITIAN

Hasil dari penelitian ini bertujuan untuk mengimplementasikan dan mengembangkan suatu konsep, metode, atau teknologi menjadi suatu produk atau sistem yang dapat digunakan. Ekstraksi kata kunci secara garis besar dibagi menjadi tiga langkah: *preprocessing* teks, *candidate word filtering* dan ekstraksi kata kunci.



Gambar 1. Flowchart Ekstraksi Kata Kunci Menggunakan Kombinasi TextRank dan TF-IDF

Preprocessing teks dimulai dengan melakukan tokenisasi, yang melibatkan pemisahan teks menjadi token-token individu atau kata-kata untuk memudahkan analisis. Selanjutnya, dilakukan pengubahan semua huruf dalam teks menjadi huruf kecil untuk membentuk kekonsistenan data atau disebut *case folding*. Selanjutnya dilakukan *text cleaning* yang meliputi penghapusan karakter khusus yang tidak relevan serta kata-kata yang tidak diperlukan seperti nama, Alamat dan lain-lainnya. Langkah terakhir adalah menghilangkan *stopwords*, yaitu kata-kata umum yang tidak memberikan informasi penting seperti “dan”, “adalah”, “akan”, “bagi”, dan kata-kata umum lain pada teks Bahasa Indonesia. *Preprocessing* ini memastikan bahwa teks bersih dan siap untuk analisis.

Sebelum melakukan ekstraksi kata kunci dari data hasil *preprocessing*, kata yang akan diekstrak perlu disaring agar proses ekstraksi kata kunci tidak memakan banyak waktu. *Candidate Word Filtering* berasal dari kata-kata hasil NER yang sudah diidentifikasi dan juga berasal dari kata yang sudah mengalami pembersihan.

Ekstraksi kata kunci menggunakan metode TextRank yang menggunakan TF-IDF untuk pembobotan kata dalam penerapan algoritma TextRank. Ekstraksi kata kunci didapat dari kombinasi metode TextRank dan NER. Metode TextRank ini merupakan salah satu pendekatan berbasis graf yang digunakan untuk mengekstrak kata kunci dari sebuah teks. TextRank bekerja dengan menganalisis hubungan antar kata dalam teks dan memberikan skor

berdasarkan tingkat kepentingan kata tersebut. Sedangkan NER berfungsi untuk menemukan dan menentukan jenis *named entity* pada teks

2.1 TF-IDF

Algoritma TF-IDF ini merupakan algoritma dengan metode statistik berbobot yang biasa digunakan dalam pencarian informasi dan penggalian data. Metode ini digunakan untuk mengevaluasi tingkat kepentingan kata (Lu Yao et al., 2019).

2.1.1 *Term Frequency (tf)*

Pada persamaan (1) *Term Frequency* menggambarkan frekuensi munculnya kata dalam teks. Berapa kali kata tertentu muncul dalam teks yang dimaksud dan akhirnya menyebabkan masalah (Lu Yao et al., 2019).

$$tf_{ij} = \frac{n_{ij}}{\Sigma_k n_{kj}} \quad (1)$$

Keterangan:
n_{ij} = Jumlah kemunculan kata kunci **i** dalam dokumen **j**
tf_{ij} = *Term Frequency* (TF) untuk kata tertentu (kata kunci) **i** dalam dokumen **j**.
n_{kj} = Jumlah kemunculan kata **k** dalam dokumen **j**

2.1.2 *Document Frequency (df)*

Document Frequency berfungsi untuk mengukur seberapa sering kata tertentu muncul dalam seluruh dokumen. *Document frequency* nantinya dibutuhkan untuk mendapatkan nilai *inverse document frequency* (Lu Yao et al., 2019).

2.1.3 *Inverse Document Frequency (idf)*

Pada persamaan (2) *Inverse Document Frequency* berfungsi untuk mengukur signifikansi universal sebuah kata . Ini memberikan bobot lebih tinggi pada kata-kata yang lebih jarang muncul, yang cenderung lebih informatif (Lu Yao et al., 2019).

$$i\,df_i = log\frac{|D|}{|\{jt_i \in d_j\}|} \quad (2)$$

Keterangan:
$i\,df_i$= Bobot *Inverse Document Frequency* untuk kata kunci t_i.
$|D|$ = Jumlah total dokumen dalam koleksi dokumen.
$|\{jt_i \in d_j\}|$ = Jumlah kemunculan kata **k** dalam dokumen **j**

Untuk menghasilkan bobot yang lebih tinggi pada kata kunci yang muncul sering dalam teks diperlukan TF-IDF. Mengacu pada penelitian yang dilakukan Lu Yao, et al.(2019) TF-IDF mencapai persamaan (3)

$$TF - IDF_{ij} = tf_{ij} \times i\,df_i \quad (3)$$

Keterangan:

$TF - IDF_{ij}$= Bobot *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk kata kunci t_i dalam dokumen.
tf_{ij} = *Term Frequency* (TF) untuk kata tertentu (kata kunci) **i** dalam dokumen **j**.
$i\,df_i$= Bobot *Inverse Document Frequency* untuk kata kunci t_i.

2.2 *Named Entity Recognition*

Named Entity Recognition adalah sebuah pendekatan praktis untuk mengidentifikasi entitas bernama secara otomatis dalam teks dan data. NER bertujuan untuk menemukan dan menentukan jenis named entity pada teks. NER dapat digunakan untuk mengetahui relasi antar *named entity* dan *question answering system*. Tugas utama NER adalah untuk mencari *named entiy* dan menentukan *tipe named entity* (Wulandari, J., et al., 2018).

2.3 *Confussion Matrix*

Confussion Matrix adalah tabel yang menunjukkan jumlah prediksi yang benar dan salah yang dibuat oleh model klasifikasi. *Confussion matrix* membantu dalam menghitung dan memvisualisasikan hasil prediksi dari model. Hasil dari *Confussion matrix* berfungsi untuk menghitung berbagai metrik evaluasi seperti *recall*, *precision*, *accuracy*, dan *f-measure*.

2.3.1 *Recall*

Metrik *Recall* atau *True Positive Rate* berfungsi untuk mengukur sejauh mana model dapat mengidentifikasi semua instance positif yang ada. Mengacu pada penelitian milik (Boxley et al., 2023) metrik *Recall* mencapai persamaan (1) berikut:

$$R = \frac{TP}{TP+FN} \quad (1)$$

Keterangan:
$TP =$ Jumlah kata kunci yang diekstraksi dengan benar.
$FN =$ Jumlah kata kunci yang sebenarnya termasuk tetapi tidak berhasil diekstraksi.

2.3.2 *Precision*

Metrik *Precision* digunakan untuk mencari tahu bagaimana sebuah sistem dapat menggunakan data dari kumpulan data untuk mendeteksi contoh yang relevan secara akurat. Precision mengukur sejauh mana prediksi positif model itu benar. Mengacu pada penelitian milik (Boxley et al., 2023) metrik *Precision* mencapai persamaan (2) berikut:

$$P = \frac{TP}{TP+FP} \quad (2)$$

Keterangan:
$TP =$ Jumlah kata kunci yang diekstraksi dengan benar.
$FN =$ Jumlah kata kunci yang diekstraksi tetapi sebenarnya tidak termasuk kata kunci.
sebenarnya.

2.3.3 *Accuracy*

Accuracy berfungsi untuk mengukur sejauh mana model benar dalam memprediksi semua kelas. Mengacu pada penelitian milik (Boxley et al., 2019) metrik *F-Measure* mencapai persamaan (3) berikut:

$$P = \frac{TP+TN}{Total\ Data\ Point} \quad (3)$$

Keterangan:
$TP =$ Jumlah kata kunci yang diekstraksi dengan benar.
$TN =$ Jumlah kata kunci yang tidak diekstraksi dan benar tidak seharusnya terkekstraksi.
$Total\ Data\ Point =$ Jumlah semua kata kunci (TP+FP+FN+TN).

2.3.4 *F-Measure*

F-Measure atau biasa dikenal sebagai *F1 score* merupakan penggabungan recall dan precision menjadi satu angka tunggal yang merepresentasikan kualitas keseluruhan sistem. Nilai *F-measure* berkisar dari 0 hingga 1, dengan nilai 1 menunjukkan kinerja sempurna dan nilai 0 menunjukkan kinerja yang sangat buruk. Mengacu pada penelitian milik (Boxley et al., 2019) metrik *F-Measure* mencapai persamaan 2.8 berikut:

$$F = \frac{2 \times P \times R}{P + R} \quad (4)$$

Keterangan:

$P =$ *Precision.*
$R =$ *Recall.*

3. HASIL

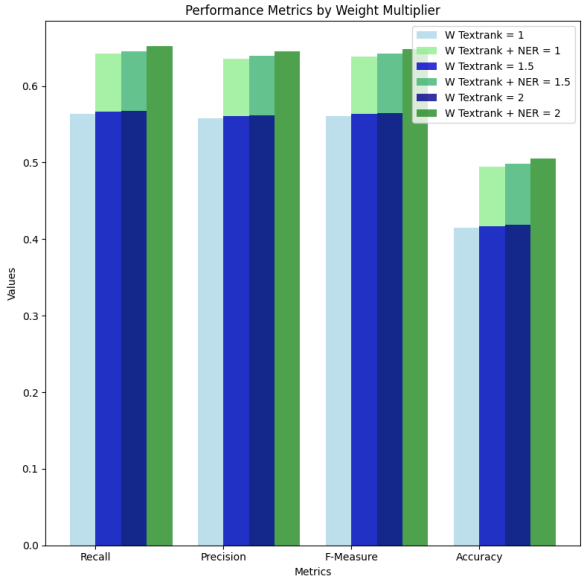
Analisis hasil dilakukan dengan membandingkan hasil ekstraksi kata kunci berlabel oleh pihak Times Indonesia dengan hasil ekstraksi kata kunci dengan metode TextRank dan integrasi Textrank dan NER. Analisis hasil dilakukan menggunakan analisis confussion matrix terhadap jumlah kata kunci yang berhasil diesktraksi serta menetapkan bobot kata kunci menjadi 1, 1,5 dan 2 kali dari teks. Tingkat kualitas yang didapatkan dihitung sesuai perhitungan kaidah *confussion matrix* seperti *recall, precision, f-measure* dan *accuracy*.

3.1 Hasil Pengujian Ekstraksi kata kunci

Hasil analisis ekstraksi kata kunci menggunakan TextRank memiliki nilai *recall, precision, f-measure* dan *accuracy* paling tinggi pada weight multiplier 2 dengan nilai masing-masing 0.567, 0.562, 0.564, 0.418. Sedangkan ekstraksi kata kunci menggunakan kombinasi TextRank dan NER memiliki nilai *recall, precision, f-measure* dan *accuracy* paling tinggi pada *weight multiplier* 2 dengan nilai masing-masing 0.652, 0.645, 0.648, 0.505. Hasil analisis dapat dilihat pada Tabel 1 dan Gambar 2.

Tabel 1. Hasil Ekstraksi Kata Kunci

Hasil Ekstraksi Kata Kunci				
Weight multiplier		**1**	**1.5**	**2**
Textrank	***Recall***	0.563	0.566	0.567
	Precision	0.558	0.560	0.562
	F-measure	0.560	0.563	0.564
	Accuracy	0.414	0.416	0.428
Textrank + NER	***Recall***	0.642	0.645	0.652
	Precision	0.636	0.639	0.645
	F-measure	0.638	0.641	0.648
	Accuracy	0.494	0.498	0.505



Gambar 2. Diagram 2 Ekstraksi Kata Kunci

4. KESIMPULAN DAN SARAN

Kombinasi metode TextRank dan *Named Entity Recognition* (NER) memberikan performa yang lebih baik dari analisis *recall, precision, f-measure,* dan *accuracy* dibandingkan dengan penggunaan metode TextRank saja. Kinerja gabungan metode TextRank dan NER dalam mengekstraksi kata kunci dari artikel berita lebih baik dibandingkan dengan penggunaan TextRank secara tunggal. Hal ini dapat dilihat dari nilai rata-rata *recall, precision, f-measure*, dan *accuracy* yang dihasilkan dari eksperimen dengan 300 artikel dan *weight multiplier* 2 dengan nilai masing-masing 0.652, 0.645, 0.648, 0.505.

Hal yang dapat dilakukan pada penelitian selanjutnya adalah menguji metode ekstraksi kata kunci ini pada berbagai jenis dataset, termasuk yang berbeda dalam topik dan bahasa. Hal ini akan memberikan wawasan yang lebih luas mengenai efektivitas metode dalam berbagai kondisi dan konteks, serta membuka peluang untuk penyesuaian dan perbaikan metode sesuai kebutuhan spesifik.

DAFTAR PUSTAKA

Ao, X. 2020. News keywords extraction algorithm based on TextRank and classified TF-IDF 1.

Azizi, M., Hayuhardhika, W., Putra, N., & Arwani, I. 2023. Ekstraksi Informasi pada Data Logbook KKN Mahasiswa Fakultas Ilmu Komputer Universitas Brawijaya Malang menggunakan Metode NER (Named Entity Recognition). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *7*(6), 2895–2903.

Boxley, C., Fujimoto, M., Ratwani, R. M., & Fong, A. 2023. A text mining approach to categorize patient safety event reports by medication error type. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-45152-w

Gunawan, D., Fanindia Purnamasari, Ranti Ramadhiana, & Romi Fadillah Rahmat. 2020. *Keyword Extraction from Scientific Articles in Bahasa Indonesia using TextRank Algorithm*. https://doi.org/10.1109/elticom50775.2020.9230514

Pirge, G. 2023, March 20. The Expert's Guide to Keyword Extraction from Texts with Python and Spark NLP. John Snow Labs. https://www.johnsnowlabs.com/the-experts-guide-to-keyword-extraction-from-texts-with-spark-nlp-and-python/#:~:text=Keywords%20extraction%20is%20the%20NLP

Rizal Maulana, A., Hadi Wijoyo, S., Mursityo, Y., Brawijaya, U., & Korespondensi, P. 2023. Analisis Sentimen Kebijakan Penerapan Kurikulum Merdeka Sekolah Dasar Dan Sekolah Menengah Pada Media Sosial Twitter Dengan Menggunakan Metode Word Embedding Dan Long Short-Term Memory Networks (Lstm) Sentiment Analysis Of Implementation Independent Curriculum Policy Elementary And Secondary School On Twitter Social Media Using Word Embedding And Long-Term Memory Networks (LSTM) Methods. Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK), 10(3), 523–530. https://doi.org/10.25126/jtiik.2023106977

Wang, P., Si, N., & Tong, H. 2022. A Named Entity Recognition Model Based on Entity Trigger Reinforcement Learning. 2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI), Beijing, China, 2022, Pp., 43-48. https://doi.org/10.1109/ccai55564.2022.9807747

Wulandari, D., Adikara, P., & Adinugroho, S. 2018. Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 2(11), 4555–4563.

Wongchaisuwat, P. (2019). *Automatic Keyword Extraction Using TextRank*.

Yao, L., Pengzhou, Z., & Chi, Z. 2019. Research on News Keyword Extraction Technology Based on TF-IDF and TextRank. 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS). https://doi.org/10.1109/icis46139.2019.8940293