REKOMENDASI KATA KUNCI ARTIKEL ILMIAH BERBASIS CONTEXTUALIZED WORD EMBEDDING DAN AGGLOMERATIVE CLUSTERING

Sulthan Rafif, Faisal Zaidan Pradana
Program Studi Sains Data, Universitas Telkom, Kampus Surabaya
Jl. Ketintang No. 156, Kec. Gayungan, Surabaya, Jawa Timur
sulthanrafif@telkomuniversity.ac.id

ABSTRAK

Kata kunci adalah alat untuk memudahkan pembaca menemukan artikel ilmiah. Kata kunci yang ditentukan dengan tepat akan menarik perhatian pembaca untuk mengkaji dan mengutip artikel ilmiah. Kata kunci yang tidak ditentukan secara tepat dapat mengurangi kesempatan artikel ilmiah dijangkau oleh pembaca. Berdasarkan permasalahan tersebut perlu menerapkan penentuan kata kunci secara otomatis untuk mengurangi kesalahan dan mempercepat pekerjaan. Pada penelitian sebelumnya diterapkan pendekatan *Contextualized Word Embedding*, *Clustering*, dan *Part-of-Speech Tagging* dalam ekstraksi kata kunci. Kombinasi pendekatan *Contextualized Word Embedding* dan *Part-of-Speech* diterapkan pada penelitian sebelumnya dengan Model *PatternRank*. Model *PatternRank* dalam ekstraksi kata kunci hanya mempertimbangkan perspektif dari satu artikel ilmiah yang menyebabkan informasi yang diperoleh terbatas sehingga perlu menerapkan pendekatan *clustering*. Pendekatan *clustering* dalam ekstraksi kata kunci pada penelitian sebelumnya menerapkan TF-IDF dalam proses *embedding*. TF-IDF tidak memperhatikan urutan kata, sehingga menyebabkan nilai *embedding* yang dihasilkan dari TF-IDF tidak sesuai dengan konteks kalimat. Sehingga berdasarkan kelemahan dari pendekatan *Contextualized Word Embedding* dan *Clustering*, pada penelitian ini menerapkan kombinasi dari kedua pendekatan tersebut dalam ekstraksi kata kunci. Berdasarkan nilai indikator *precision*, *recall*, dan *f1-score* hasil dari evaluasi menunjukkan kombinasi dari Metode *Clustering* dan *Contextualized Word Embedding* mengungguli Metode *PatternRank*.

Kata kunci : *Ekstraksi Kata Kunci, Agglomerative Clustering, Contextualized Word Embedding*

1. PENDAHULUAN

Kata kunci merupakan frasa yang merepresentasikan pembahasan utama dari artikel ilmiah [1]. Kata kunci merupakan alat untuk memudahkan pembaca menemukan artikel ilmiah. Kata kunci yang ditentukan dengan tepat akan menarik perhatian pembaca untuk mengkaji dan mengutip artikel ilmiah [2] [3] [4]. Namun disisi lain, kata kunci adalah bagian dari artikel ilmiah yang sering diabaikan. Kata kunci seringkali disusun secara terburu-buru untuk memenuhi kriteria penyerahan artikel ilmiah, sehingga menimbulkan ketidaktepatan [5]. Kata kunci yang tidak ditentukan secara tepat dapat mengurangi kesempatan artikel ilmiah dijangkau oleh pembaca [6]. Berdasarkan permasalahan tersebut perlu menerapkan penentuan kata kunci secara otomatis untuk mengurangi kesalahan serta mempercepat pekerjaan. Penentuan kata kunci secara otomatis disebut dengan ekstraksi kata kunci [7].

Ekstraksi kata kunci adalah proses otomatis untuk memilih sekumpulan frasa paling relevan dari artikel ilmah. *Contextualzied Word Embedding*, *Part-of-Speech* (PoS) dan *Clustering* merupakan beberapa pendekatan yang digunakan dalam ekstraksi kata kunci. PoS diimplementasikan oleh Model PatternRank [8]. PatternRank menerapkan *PoS* untuk menyeleksi frasa kata kunci yang terkandung dalam artikel ilmiah.

Namun, PatternRank hanya mempertimbangkan perspektif dari satu artikel ilmiah yang menyebabkan informasi yang diperoleh terbatas [8]. Terbatasnya informasi yang diperoleh menyebabkan kata kunci yang diekstraksi menjadi kurang relevan. Berdasarkan pada permasalahan tersebut, ekstraksi kata kunci perlu mempertimbangkan perspektif dari kumpulan artikel ilmiah sehingga informasi yang diperleh bersifat lebih komprehensif.

Pengelompokkan artikel ilmiah dapat menerapkan pendekatan *clustering*. Pada penelitian sebelumnya pendekatan *clustering* diterapkan oleh CollabRank dengan mengelompokkan artikel ilmiah yang memiliki kesamaan topik pembahasan menjadi beberapa *cluster* [9]. Data yang dikelompokkan adalah bobot atau nilai *embedding* setiap frasa artikel ilmiah. Nilai *embedding* ditentukan dengan nilai TF-IDF. Nilai TF-IDF merepresentasikan frekuensi kemunculan frasa dalam artikel ilmiah. Namun TF-IDF tidak memperhatikan urutan frasa, mengingat perbedaan urutan frasa dapat mempengaruhi konteks dari artikel ilmiah [10]. Berdasarkan permasalahan tersebut perlu menerapkan pendekatan *Clustering Word Embedding* atau disebut dengan pembobotan frasa berdasarkan konteks.

Proses pembobotan frasa berdasarkan konteks dapat menerapkan Model *Pretrained* BERT [11]. Pada penelitian sebelumnya, Model *Pretrained* BERT *embedding* diterapkan oleh KeyBERT [12]. *Pretrained* BERT memperhatikan urutan frasa dalam artikel ilmiah [13], sehingga nilai *embedding* yang dihasilkan dari setiap frasa artikel ilmiah berdasarkan pada konteks.

Berdasarkan pada permasalahan penentuan kata kunci, pada penelitian ini menerapkan Pendekatan *Clustering* dan *Contextualized Word Embedding*

dalam ekstraksi kata kunci. Pendekatan *Clustering* dan *Contextualized Word Embedding* akan dibandingkan dengan Metode *Baseline* yaitu Metode *PatternRank*. Dengan menerapkan pendekatan *clustering* dan BERT *Contextualized Word Embedding* diharapkan setiap kata kunci yang dihasilkan relevan dengan pembahasan yang terkandung di dalam artikel ilmiah dengan tujuan dapat menarik pembaca untuk mengkaji dan mengutipnya.

2. TINJAUAN PUSTAKA

Pada bagian ini dideskripsikan penelitian terdahulu yang menerapkan Metode Ekstraksi Kata Kunci. *Contextualized Word Embedding* [12], *clustering* [9] [14], dan *PoS* [8] meraih *State-of-the-art* untuk Metode Ekstraksi Kata kunci. Pendekatan *clustering* mengelompokkan artikel ilmiah berdasarkan topik pembahasan [9]. Penelitian *CollabRank* menerapkan lima Algoritma *Clustering*, yaitu *K-Means Clustering*, *Agglomerative (Average Link) Clustering*, *Agglomerative (CompleteLink) Clustering*, *Divisive Clustering*, dan *Random Clustering*. Kelima algoritma tersebut diuji dan dibandingkan dengan Metode *Baseline* yaitu *TextRank* dan TF-IDF. Berdasarkan nilai *precision*, *recall*, dan *f-measure*, *Algoritma Agglomerative Complete Clustering* mengungguli Metode *TextRank* dan TF-IDF serta empat Algoritma *Clustering* yang lain.

Pada proses *clustering*, data yang dikelompokkan adalah bobot atau nilai *embedding* setiap frasa artikel ilmiah hasil *preprocessing*. Nilai *embedding* ditentukan dengan nilai TF-IDF. Nilai TF-IDF merepresentasikan frekuensi kemunculan frasa dalam artikel ilmiah. Frasa dengan nilai TF-IDF yang tinggi cenderung mengandung informasi yang lebih signifikan dalam suatu artikel ilmiah [15]. Namun TF-IDF tidak memperhatikan urutan kata dalam frasa mengingat perbedaan urutan kata dapat mempengaruhi konteks dari frasa [16]. Pada penelitian selanjutnya mengusulkan Metode KeyBERT yang menerapkan Pendekatan *Pretrained* BERT berbasis *Contextualized Word Embedding* untuk menentukan nilai *embedding* setiap frasa pada artikel ilmiah.

KeyBERT menerapkan Model *Pretrained Bidirectional Encoder Representation from Transformers* atau BERT dan menggunakan N-Gram untuk menyeleksi frasa kata kunci [12]. *Pretrained* BERT memperhatikan urutan kata dalam frasa, sehingga nilai *embedding* yang dihasilkan dari setiap kata berdasarkan pada konteks frasa [13]. N-Gram merepresentasikan jumlah kata pada frasa yang diekstraksi menjadi kata kunci. Jumlah kata pada frasa ditentukan secara manual. Dalam implementasinya, KeyBERT perlu menentukan jumlah kata yang tepat dalam menghasilkan frasa kata kunci yang sesuai. Penelitian selanjutnya mengusulkan Model PatternRank yang merupakan pengembangan dari KeyBERT.

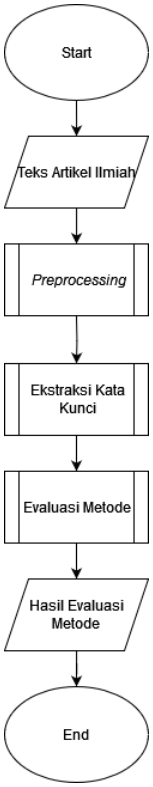
Berbeda dengan KeyBERT, PatternRank menyeleksi kata kunci berdasarkan pola *Part-of-Speech* atau PoS [8]. Selanjutnya PatternRank menerapkan *Pretrained* BERT *embedding* untuk menentukan nilai *embedding* setiap kata. Jumlah kata pada frasa diekstraksi berdasarkan hasil seleksi PoS dan nilai nilai *embedding* dari setiap kata.

Berdasarkan kajian penelitian terdahulu, pada penelitian ini menerapkan pendekatan *clustering* serta *Contextualized Word Embedding* dalam ekstraksi kata kunci. Algoritma *Clustering* yang digunakan adalah *Algoritma Agglomerative Clustering* serta menerapkan *Pretrained* BERT *embedding* untuk proses pembobotan kata.

3. METODE PENELITIAN

Pada penelitian ini menerapkan ekstraksi kata kunci terhadap dataset berisi sejumlah 224 artikel ilmiah Berbahasa Inggris yang terdiri dari bab abstrak dan pendahuluan. Pada penelitian ini menggunakan Dataset SemEval2010. SemEval2010 berisi sejumlah 244 artikel ilmiah, diambil dari *ACM Digital Library* dan diterbitkan pada Tahun 2010.

Terdapat tiga tahapan dalam ekstraksi kata kunci. Tahap pertama adalah melakukan *preprocessing*. Tahap kedua adalah menerapkan model ekstraksi kata kunci. Tahap ketiga adalah melakukan evaluasi model. Gambar 1 menampilkan *flowchart* alur dari tahapan ekstraksi kata kunci.



Gambar 1. Tahapan ekstraksi kata kunci

Pada Bagian Metode Penelitian, tahapan ekstraksi kata kunci diterapkan pada contoh kalimat abstrak. Contoh kalimat abstrak ditampilkan pada Tabel 1.

Tabel 1. Contoh kalimat abstrak

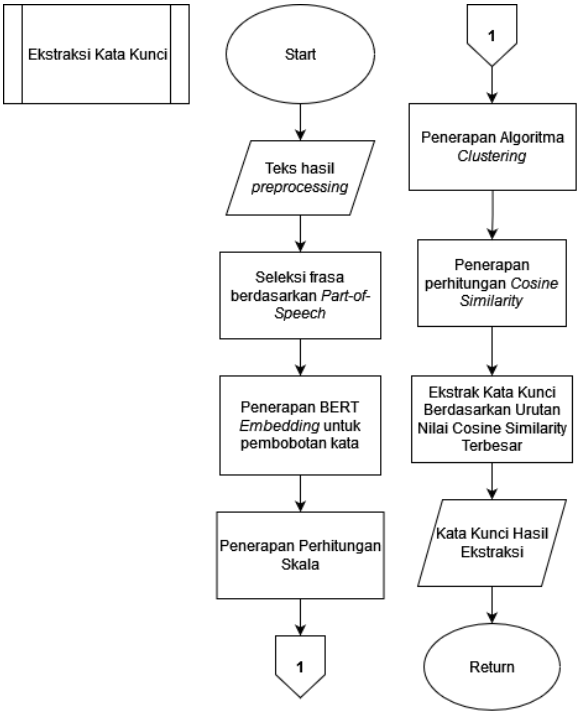
No	Kalimat Abstrak
1.	*Real-time services have been supported by and large on circuit-switched networks*
2.	*A relatively minor outage can disrupt and inconvenience a large number of users*

3.1. Preprocessing

Preprocessing diterapkan pada teks abstrak dan pendahuluan. Tahap *preprocessing* bertujuan untuk menyiapkan teks yang siap diseleksi menggunakan *Part-of-Speech.* Pada penelitian ini menerapkan dua tahap *preprocessing*, yaitu *case-folding* dan *stopword removal*.

3.2. Metode Ekstraksi Kata Kunci

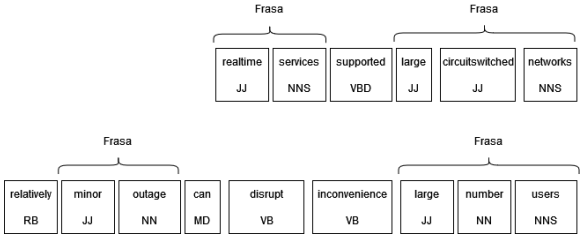
Penerapan Metode Ekstraksi Kata Kunci terdiri dari lima tahapan. Tahapan pertama adalah melakukan seleksi frasa berdasarkan *Part-of-Speech*. Tahap kedua adalah menerapkan pembobotan frasa dengan BERT *Embedding*. Tahap ketiga adalah menerapkan perhitungan skala. Tahap keempat adalah menerapkan Algoritma *Clustering*. Tahap kelima adalah menerapkan *Cosine Similarity* untuk menentukan frasa kata kunci yang diekstraksi. Gambar 2 menampilkan *flowchart* penerapan metode ekstraksi kata kunci.



Gambar 2. Metode ekstraksi kata kunci

Seleksi *Part-of-Speech* adalah tahapan menyeleksi frasa dari teks abstrak dan pendahuluan hasil *preprocessing*. Proses seleksi *Part-of-Speech* dilakukan dengan memberikan *tag* pada setiap kata pada kalimat. *Tag* yang diberikan berupa pengisi fungsi sintaksis. Fungsi sintaksis yang diseleksi adalah *nouns* dan *adjectives*. *Nouns* dan *Adjectives* memiliki pengisi yaitu berupa *tag* "JJ", "NN", "NNS", "NNP", "NNPS" [17].

Setiap kata *adjectives* dengan tag "JJ" dan kata *noun* dengan tag "NN", "NNS", "NNP", "NNPS" yang berdekatan digabung menjadi frasa. Gambar 3 menampilkan contoh penerapan *tag* pada salah satu kalimat abstrak.



Gambar 3. Contoh *part-of-speech*

Berdasarkan pada Gambar 3 kata "*realtime*" dan "*services*" berdekatan dan memiliki *tag* "JJ" dan "NNS" sehingga kedua kata tersebut digabung menjadi frasa. Penggabungan kata juga berlaku terhadap frasa "*large circuitswitched networks*", "*minor outage*", dan "*large number users*". Setiap frasa yang dihasilkan pada tahapan seleksi *Part-of-Speech* selanjutnya diterapkan pembobotan menggunakan BERT *embedding.* Setiap bobot memiliki nilai fitur. Bobot setiap frasa digunakan sebagai representasi data pada *cluster*.

Selanjutnya, nilai bobot setiap frasa dilakukan proses *feature scaling*. *Feature scaling* merupakan tahapan melakukan transformasi nilai fitur pada dataset berdasarkan skala yang sama [18]. *Feature scaling* penting diterapkan pada *dataset* dengan nilai fitur yang memiliki perbedaan skala. Penelitian ini menerapkan perhitungan *standar scaler*, *min max scaler*, dan *log scaler*.

Algoritma *clustering* berperan untuk mengelompokkan frasa dari seluruh artikel ilmiah menjadi beberapa *cluster*. Data yang dikelompokkan adalah bobot setiap frasa hasil dari proses *embedding* yang sudah diterapkan *feature scaling*. Pada penelitian ini menerapkan Algoritma *Agglomerative Clustering* untuk mengekstraksi kata kunci.

Pada penelitian sebelumnya, Algoritma *Agglomerative Clustering* mengungguli Metode *TextRank* dan TF-IDF [9]. *Agglomerative Clustering* merupakan metode hirarki dalam *clustering*. Metode hirarki mengelompokkan data dalam bentuk struktur pohon atau disebut dengan *tree* [19].

Hasil dari tahapan *clustering* adalah *cluster* dari kumpulan *frasa*. Frasa yang dimiliki oleh setiap *cluster* ditampilkan pada Tabel 2.

Tabel 2. Frasa setiap *cluster*

No	Cluster 1	Cluster 2
1	*realtime services, large circumswitched networks*	*large number users, minor outage*

Frasa yang dimiliki oleh setiap *cluster* diterapkan perhitungan *cosine similarity* untuk mengetahui tingkat kesamaan antara frasa dengan artikel ilmiah.

3.3. Penerapan Algoritma *Clustering* untuk Setiap Frasa Milik Setiap Artikel Ilmiah

Selain diterapkan untuk mengelompokkan frasa dari seluruh artikel ilmiah pada dataset, penerapan *Algoritma Clustering* juga digunakan untuk mengelompokkan frasa setiap artikel ilmiah, pendekatan ini disebut dengan *local clustering*. *Local clustering* adalah proses mengelompokkan frasa setiap artikel ilmiah dengan tujuan untuk menyeleksi frasa yang memiliki makna yang sama berdasarkan perhitungan *euclidean distance*. Frasa dengan makna yang sama dijadikan sebagai kandidat kata kunci untuk diseleksi menggunakan persamaan *Cosine Similarity*. Tabel 3 menampilkan daftar frasa dari dua dokumen abstrak yang dikelompokkan oleh *cluster*.

Tabel 3. *Cluster* Frasa pada satu dokumen

No Frasa	Teks Abstrak	Frasa	Cluster
1	Real-time services have been supported by and large on circuit-switched networks	realtime services, large circumswitched networks	*Cluster 1*
2	A relativelyminor outage can disrupt and inconvenience a large number of users	large number users, minor outage	*Cluster 2*

Berdasarkan pada Tabel 3, setiap dokumen abstrak memiliki *cluster* masing-masing. Setiap *cluster* terdiri dari kumpulan frasa yang memiliki makna yang sama berdasarkan perhitungan *euclidean distance*. Kumpulan frasa pada *cluster* menjadi kandidat kata kunci yang akan diseleksi menggunakan perhitungan *cosine similarity*.

3.4. *Cosine Similarity*

Persamaan *Cosine Similarity* berfungsi untuk mengukur tingkat kesamaan antara bobot frasa dengan bobot teks artikel ilmiah. Semakin tinggi nilai *cosine similarity* maka semakin tinggi tingkat kesamaan antara frasa dengan artikel ilmiah. Nilai *cosine similarity* digunakan sebagai dasar untuk menentukan frasa yang diekstraksi menjadi kata kunci. Sebelum dilakukan perhitungan menggunakan persamaan *cosine similarity*, teks abstrak dan pendahuluan terlebih dahulu dikonversi menjadi bobot menggunakan *Pretrained BERT*.

Cosine similarity menghitung tingkat kesamaan berdasarkan derajat sudut antar vektor [20]. Vektor adalah bobot yang terdiri dari kumpulan nilai fitur. Contoh hasil perhitungan *Cosine Similarity* untuk setiap frasa hasil dari Algoritma *Agglomerative Clustering* terhadap teks abstrak ditampilkan pada Tabel 4.

Tabel 4. Hasil *cosine similarity*

No	Kalimat Abstrak	Frasa Hasil Ekstraksi	Nilai Cosine Similarity
1	Realtime services supported large circumswitched networks	realtime services; large circumswitched networks	0,577; 0,483
2	A relativel minor outage can disrupt and inconvenience a larger number of users	minor outage; large number users	0,984; 0,887

Berdasarkan pada Tabel 4, pada teks abstrak kedua, frasa *minor outage* memiliki nilai *cosine similarity* yang tinggi jika dibandingkan dengan *large number users*. Untuk teks abstrak pertama, untuk frasa *realtime services* memiliki nilai *cosine similarity* yang lebih besar jika dibandingkan dengan frasa *circumswitched networks*. Selanjutnya frasa diekstraksi berdasarkan nilai *cosine similarity* pada Tabel 4 yang diurutkan secara *descending*. Jumlah frasa yang diekstraksi menjadi kata kunci ditentukan secara manual. Pada contoh perhitungan diterapkan jumlah frasa kata kunci = 2. Frasa kata kunci yang diekstraksi untuk setiap teks abstrak ditampilkan pada Tabel 5.

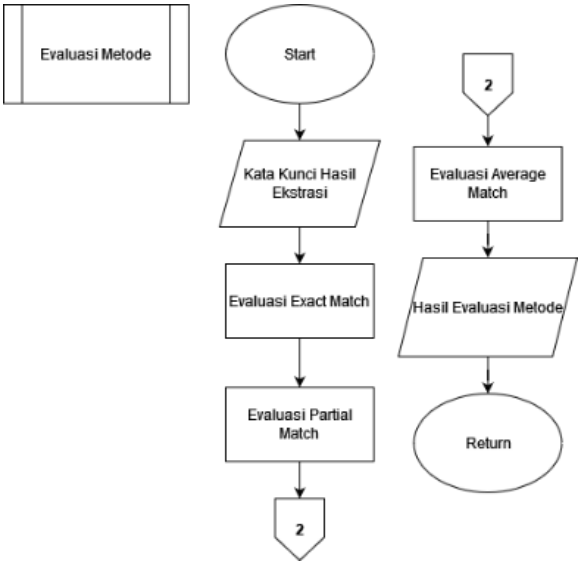
Tabel 5. Frasa kata kunci

No Frasa	Teks Abstrak	Frasa
1	Realtime services supported large circumswitched networks	realtime services, large circumswitched networks
2	A relatively minor outage can disrupt and inconvenience a larger number of users	minor outage, large number users

3.5. Metode Evaluasi

Pada tahapan ini dilakukan evaluasi terhadap frasa kata kunci hasil ekstraksi. Gambar 4 menampilkan *flowchart* alur dari evaluasi. Proses evaluasi dilakukan dengan memeriksa tingkat kesesuaian frasa kata kunci hasil ekstraksi dengan kata kunci pada dataset. Proses evaluasi dilakukan dengan perhitungan indikator *precision*, *recall*, dan *f1-score*. Evaluasi dengan indikator *precision*, *recall*, dan *f1-score* diterapkan untuk setiap artikel ilmiah berdasarkan tingkat kesesuaian frasa kata kunci hasil ekstraksi dengan kata kunci sebenarnya milik setiap artikel ilmiah. Nilai *precision*, *recall*, dan *f1-score* milik seluruh artikel ilmiah selanjutnya dilakukan perhitungan rata-rata untuk mengetahui performa dari model. Proses evaluasi menerapkan tiga pendekatan

evaluasi yaitu *exact match*, *partial match*, dan *average match*. Setiap pendekatan evaluasi dilakukan perhitungan untuk indikator *precision*, *recall*, dan *f1-score*. Ketiga indikator tersebut digunakan untuk mengetahui perbandingan hasil evaluasi antara model yang diusulkan dengan Model *PatternRank* sebagai metode *baseline.*



Gambar 4. Alur evaluasi

Pada pendekatan *exact match*, perhitungan *true positive* dilakukan jika kumpulan frasa hasil ekstraksi sesuai dengan kumpulan kata kunci pada salah satu artikel ilmiah [21].Contoh hasil evaluasi *exact match* ditampilkan pada Tabel 6.

Tabel 6. Hasil evaluasi *exact match*

No	Kata Kunci Hasil Ekstraksi	Kata Kunci Dataset	Hasil Evaluasi Exact Match
1	realtime services, large circumswitched networks	realtime services, large networks	realtime services (TRUE) large networks (FALSE)
2	minor outage, large number users	minor outage, large number users	minor outage (TRUE) large number users (TRUE)

Berdasarkan pada Tabel 6, *exact match* tidak menghiraukan jika hanya terdapat salah satu suku kata pada frasa yang sesuai dengan artikel ilmiah, hal tersebut menyebabkan pendekatan evaluasi ini memberikan sanksi pada kata kunci hasil ekstraksi yang secara sintaksis berbeda, namun sama secara semantik [22].

Pendekatan *partial match* menghitung *true positive* jika terdapat minimal salah satu suku kata pada frasa hasil ekstraksi sesuai dengan artikel ilmiah [23]. *Partial match* dilakukan dengan mengkonversi frasa ke *unigram*. *Partial match* memenuhi syarat jika terdapat frasa yang memiliki kesamaan dari segi semantik terhadap dataset [23]. Contoh hasil evaluasi *partial match* ditampilkan pada Tabel 7.

Tabel 7. Hasil evaluasi *partial match*

No	Kata Kunci Hasil Ekstraksi	Kata Kunci Dataset	Hasil Evaluasi Exact Match
1	realtime services, large circumswitched networks	realtime services, large networks	realtime services (TRUE) large networks (TRUE)
2	minor outage, large number users	minor outage, large number users	Minor outage (TRUE) large number users (TRUE)

Evaluasi *average match* diterapkan dengan menghitung rata-rata dari evaluasi *exact* dan *partial match*. Selanjutnya proses evaluasi menerapkan kombinasi nilai dari *hyperparameter*. Nilai *hyperparameter* yang diuji adalah *cluster*, jumlah fitur pada *embedding*, dan algoritma *clustering*. Kombinasi dari lima *hyperparameter* diterapkan dalam ekstraksi kata kunci. Setiap kombinasi nilai *hyperparameter* dievaluasi dengan pendekatan *exact match*, *partial match*, dan *average match*.

4. HASIL DAN PEMBAHASAN

Pada bagian ini ditampilkan hasil evaluasi dari kata kunci hasil ekstraksi berdasarkan kombinasi nilai *hyperparameter* yang diterapkan terhadap metode yang diusulkan dan metode *baseline*. Usulan metode yang diterapkan adalah pendekatan *clustering* dan *local clustering*.

4.1. Perbandingan Metode *Clustering*, *Local Clustering* dan *Baseline*

Berdasarkan hasil dari penerapan setiap *hyperparameter* tehadap Metode *Clustering*, *Local Clustering*, dan *Baseline*, didapatkan kombinasi *hyperparameter* terbaik. Kombinasi *hyperparameter* yang terbaik yaitu menerapkan nilai jumlah *cluster* = 2, nilai jumlah *embedding* = 384 dengan Model *Sentence BERT*, menerapkan Algoritma *SingleLink* untuk *Agglomerative Clustering* dan menerapkan *Log Scaler* sebagai perhitungan skala. Pada Tabel 8, 9, dan 10 ditampilkan hasil evaluasi dari penerapan kombinasi *hyperparameter* terbaik untuk Metode *Clustering*, *Local Clustering*, dan *Baseline*

Tabel 8. Hasil evaluasi untuk setiap metode ekstraksi kata kunci untuk *exact match*

No	Variasi Perhitungan Skala	*Exact Match*		
		Precision	***Recall***	***F1 Score***
1	*Cluster*	0,36	0,12	0,16
2	*Local Cluster*	**0,75**	**0,14**	**0,23**
3	*Baseline (PatternRank)*	0,75	0,13	0,22

Tabel 9. Hasil evaluasi untuk setiap metode ekstraksi kata kunci untuk *partial match*

No	Variasi Perhitungan Skala	*Partial Match*		
		Precision	***Recall***	***F1 Score***
1	*Cluster*	0,95	**0,65**	0,74
2	*Local Cluster*	**0,99**	**0,65**	**0,76**
3	*Baseline (PatternRank)*	**0,99**	0,63	0,75

Tabel 10. Hasil evaluasi untuk setiap metode ekstraksi kata kunci untuk *average match*

No	Variasi Perhitungan Skala	*Average Match*		
		Precision	***Recall***	***F1 Score***
1	*Cluster*	**0,87**	**0,39**	**0,49**
2	*Local Cluster*	**0,87**	**0,39**	**0,49**
3	*Baseline (PatternRank)*	**0,87**	0,38	0,48

Berasarkan pada Tabel 12, 13, dan 14 Metode *Local Clustering* memiliki hasil yang unggul jika dibandingkan dengan Metode *Clustering* dan *Baseline* untuk pendekatan *exact match*, *partial match*, dan *average match*. Penerapan *clustering* terhadap bobot frasa untuk setiap artikel ilmiah dapat mengelompokkan frasa berdasarkan kesamaan konteks sebelum dilakukan perhitungan *cosine similarity*. Kandidat kata kunci yang dihasilkan dari proses *clustering* merepresentasikan pokok bahasan dari artikel ilmiah.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil evaluasi *exact match*, *partial match*, dan *average match*. Metode *local clustering* mengungguli Metode *Clustering* dan *PatternRank* dalam ekstraksi kata kunci. Kombinasi *hyperparameter* dengan menerapkan nilai jumlah *hyperparameter* dengan menerapkan nilai jumlah *cluster* = 2, nilai jumlah *embedding* = 384 dengan Model *Sentence BERT*, serta menerapkan Algoritma *Agglomerative Single Link* dengan *Log Scaler* dalam proses *clustering* untuk menyeleksi frasa yang dimiliki setiap dokumen dapat meningkatkan *precision*, *recall*, dan *f1-score* dalam ekstraksi kata kunci. Dalam Metode *Clustering* untuk pendekatan *exact match*, *partial match*, dan *average match* nilai dari *precision*, *recall* dan f1-score masih berada dibawah Metode *PatternRank*, hal tersebut dikarenakan terdapat beberapa *cluster* yang berisi frasa yang berbeda topik, jika jumlah *cluster* yang diterapkan dua. Sehingga berdasarkan hasil penerapan metode *clustering* pada ekstraksi kata kunci, pada penelitian selanjutnya perlu menerapkan suatu metode pengelompokkan frasa yang tepat berdasarkan topik pembahasan.

DAFTAR PUSTAKA

[1] G. de Abreu, "Keywords in a Research Paper: The Importance of the Right Choice," 2023. https://mindthegraph.com/blog/keywords-in-a-research-paper/ (accessed Oct. 18, 2023).

[2] L. Corrin, K. Thompson, G. J. Hwang, and J. M. Lodge, "The importance of choosing the right keywords for educational technology publications," *Australas. J. Educ. Technol.*, vol. 38, no. 2, pp. 1–8, 2022, doi: 10.14742/ajet.8087.

[3] L. J. Zunan Setiawan, Hildawati, Henny Sanulita, Dedy Afrizal, Siti Misaroh Ibrahim, Arif Susanto, Titi Indahyani, Saputra Adiwijaya, Laurensius Laka, Muhammad Ansor, Soetji Andari, Mohammad Fajri Mekka Putra, Asteria Permata Martawijaya, *METODOLOGI DAN TEKNIK PENULISAN ILMIAH*. Sonpedia Publishing Indonesia, 2024.

[4] Elsevier, "How to choose keywords for a manuscript?" https://scientific-publishing.webshop.elsevier.com/manuscript-preparation/how-choose-keywords-manuscript/ (accessed Oct. 30, 2023).

[5] A. E. Abusaada, Hisham, "Beyond Keywords : Effective Strategies for Building Consistent Reference Lists in Scientific Research," *Publications*, vol. 12(3), 25, 2024, doi: https://doi.org/10.3390/publications12030025.

[6] P. Pottier *et al.*, "Title , abstract and keywords : a practical guide to maximize the visibility and impact of academic papers," *R. Soc.*, vol. 1, 2024, doi: https://doi.org/10.6084/m9.figshare.c.7358189.

[7] P. Espada, "Expert system for extracting keywords in educational texts and textbooks based on transformers models," *Expert Syst. Appl.*, vol. 282, no. April, 2025, doi: 10.1016/j.eswa.2025.127735.

[8] T. Schopf, S. Klimek, and F. Matthes, "PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction," *Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag. IC3K - Proc.*, vol. 1, pp. 243–248, 2022, doi: 10.5220/0011546600003335.

[9] X. Wan and J. Xiao, "CollabRank: Towards a collaborative approach to single-document keyphrase extraction," *Coling 2008 - 22nd Int. Conf. Comput. Linguist. Proc. Conf.*, vol. 1, no. August, pp. 969–976, 2008.

[10] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, "Heliyon Investigating response behavior through TF-IDF and Word2vec text

analysis : A case study of PISA 2012 problem-solving process data," *Heliyon*, vol. 10, no. 16, p. e35945, 2024, doi: 10.1016/j.heliyon.2024.e35945.

[11] I. Beltagy, K. Lo, and A. Cohan, "SCIBERT: A pretrained language model for scientific text," *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process.*, vol. 9, pp. 3615–3620, 2019, doi: 10.18653/v1/d19-1371.

[12] Maarten Grootendorst, "KeyBERT: Minimal keyword extraction with bert," *Zenodo*, vol. 0.3.0, 2020, doi: https://doi.org/10.5281/zenodo.4461265.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[14] R. R. Cusumano L, Olsson N, Granath M, Jockwer R, "Clustering techniques and keyword extraction with large language models for knowledge discovery in building defects data," *Constr. Innov. Inf. Process Manag.*, vol. 25 No. 7 p, no. January 2026, 2025, doi: https://doi.org/10.1108/CI-04-2024-0123.

[15] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks," *SINTESIA J. Sist. dan Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 81–88, 2022.

[16] S. Rustad, G. F. Shidik, and I. Nlp, "Performance Evaluation of Text Embedding Models for Ambiguity Classification in Indonesian News Corpus : A Comparative Study of TF-IDF , Word2Vec , FastText BERT , and GPT," *Ingénierie des Systèmes d ' Inf.*, vol. 30, no. 6, pp. 1469–1482, 2025, doi: https://doi.org/10.18280/isi.300606 Received:

[17] H. Kunmei, "Modelling Lexical Characteristics of the Healthy Aging Population : A Corpus-Based Study," *Proc. Interspeech*, vol. 1090–1094, no. September, pp. 1090–1094, 2024, doi: 10.21437/Interspeech.2024-1871.

[18] R. M. O. C. Lucas B.V. de Amorim, George D.C. Cavalcanti, "The choice of scaling technique matters for classification performance," *Appl. Soft Comput.*, vol. 133, no. 2023, pp. 1–37, 2023, doi: https://doi.org/10.1016/j.asoc.2022.109924.

[19] Z. Michalewicz and H. Kwas, "Hierarchical data generator based on tree-structured stick breaking process for benchmarking clustering methods," *Inf. Sci. (Ny).*, vol. 554, no. 2020, pp. 99–119, 2020, doi: 10.1016/j.ins.2020.12.020.

[20] D. N. Quoc and N. T. Thanh, "Enhancing Accuracy for Classification Using the CNN Model and Hyperparameter Optimization Algorithm," *Indones. J. Electr. Eng. Informatics*, vol. 12, no. 3, pp. 575–582, 2024, doi: 10.52549/ijeei.v12i3.5545.

[21] M. Kölbl, Y. Kyogoku, J. N. Philipp, M. Richter, and C. Rietdorf, "Beyond the Failure of Direct-Matching in Keyword Evaluation : A Sketch of a Graph Based Solution," *Front. Artif. Intell.*, vol. 5:801564, no. March, pp. 1–13, 2022, doi: 10.3389/frai.2022.801564.

[22] I. Okpala and G. R. Rodriguez, "A Semantic Approach to Negation Detection and Word Disambiguation with Natural Language Processing," *Proc. 2022 6th Int. Conf. Nat. Lang. Process. Inf. Retr.*, vol. 6, no. January, pp. 36–43, 2023, doi: 10.1145/3582768.3582789.

[23] K. Ricci, H. C. Purujit, and G. Andrew, "Unsupervised Partial Sentence Matching for Cited Text Identification," *Proc. ofthe Third Work. Sch. Doc. Process.*, vol. 3, pp. 95–104, 2022, [Online]. Available: https://aclanthology.org/2022.sdp-1.11/.