



UNIVERSITAS INDONESIA

**EKSTRAKSI KATAKUNCI DENGAN METODE SAINSMETRIKA
DAN KLASIFIKASI LABEL BANYAK**

RINGKASAN DISERTASI

Harry Tursulistyono Yani Achsan
NPM. 1706010445

Fakultas Ilmu Komputer
Program Doktor Ilmu Komputer
Depok
2023

ABSTRACT

Keywords and keyphrases are crucial for various purposes such as search engines, article annotation and classification, and image processing. Unfortunately, most downloadable article metadata from Google Scholar lacks keywords/keyphrases, making it difficult to extract the essence of the writing. Several methods have been developed over the years, but some have low accuracy or take too long to implement. This study seeks to improve keyword extractor performance by introducing a new method that uses multi-label classification in machine learning. Labels are obtained from popular vocabulary, the most popular keyword from computer science scientific articles. The new method was tested on popular datasets such as Semeval, Nguyen, and Krapivin2019 and compared with ten state-of-the-art keyword extraction methods. Results show that this new method consistently has the highest quality (Mean Reciprocal Rank value), and it is faster than other good-performing methods. This study also created a new dataset with 3,229 scientific article metadata, and a publicly accessible recommendation system for keyword extraction at <http://extract.my.id/>.

Keywords: *Keyword extraction, Recommender systems, Mean reciprocal rank, Popular vocabulary.*

ABSTRAK

Kata kunci dan frasa kunci sangat penting untuk berbagai tujuan seperti mesin telusur, anotasi dan klasifikasi artikel, dan pemrosesan gambar. Sayangnya, sebagian besar metadata artikel yang dapat diunduh dari Google Cendekia kekurangan kata kunci/keyphrases, sehingga sulit untuk mengekstrak inti tulisan. Beberapa metode telah dikembangkan selama bertahun-tahun, tetapi beberapa memiliki akurasi yang rendah atau terlalu lama untuk diimplementasikan. Penelitian ini berusaha untuk meningkatkan kinerja ekstraktor kata kunci dengan memperkenalkan metode baru yang menggunakan klasifikasi multi-label dalam pembelajaran mesin. Label diperoleh dari kosakata populer, kata kunci paling populer dari artikel ilmiah ilmu komputer. Metode baru diuji pada kumpulan data populer seperti Semeval, Nguyen, dan Krapivin2 019 dan dibandingkan dengan sepuluh metode ekstraksi kata kunci yang canggih. Hasil menunjukkan bahwa metode baru ini secara konsisten memiliki kualitas tertinggi (nilai Mean Reciprocal Rank), dan lebih cepat dibandingkan metode lain yang berkinerja baik. Studi ini juga membuat dataset baru dengan 3.229 metadata artikel ilmiah, dan sistem rekomendasi ekstraksi kata kunci yang dapat diakses publik di <http://extract.my.id/>.

Katakunci: Ekstraksi katakunci, Sistem rekomendasi, Mean reciprocal rank, Kosakata populer.

DAFTAR ISI

ABSTRACT	ii
ABSTRAK	iii
DAFTAR ISI	iv
1 PENDAHULUAN.....	7
1.1 Perumusan Masalah	8
1.2 Ruang Lingkup	9
1.3 Tujuan dan Manfaat	9
1.4 Signifikansi Penelitian	10
1.5 Penelitian Awal.....	10
2 TINJAUAN PUSTAKA.....	13
2.1 Metadata bahan pustaka.....	13
2.2 Kosakata populer	14
2.3 Kata/frasa kunci	14
2.4 Riwayat penelitian ekstraksi katakunci.....	15
3 METODOLOGI	18
3.1 Pengantar	18
3.2 Pendekatan	18
3.3 Klasifikasi multi-label.....	19
3.4 Langkah-Langkah	21
3.4.1 Peninjauan Pustaka.....	21
3.4.2 Penelitian Awal	21
3.4.3 Penambangan Web.....	22
3.4.4 Pengembangan Metoda Ekstraksi Katakunci	22
3.4.5 Pra-Proses.....	23
3.4.6 Pembuatan Dataset & Kosakata Populer.....	24
3.4.7 Pengembangan Aplikasi Proof of Concepts.....	24
3.4.8 Eksperimen dengan berbagai parameter	24

3.4.9	Analisis & Evaluasi.....	25
3.4.10	Penulisan Laporan Disertasi.....	26
4	PENGEMBANGAN METODA BARU.....	27
4.1	Pembuatan Daftar Kosakata populer.....	27
4.2	Dataset.....	27
4.2.1	Dataset pembentuk kosakata populer.....	27
4.2.2	Dataset umum.....	28
4.2.3	Dataset baru.....	28
4.3	Teknik Ekstraksi Kata/Frasa Kunci.....	29
4.3.1	Pembobotan.....	29
4.3.2	Frasa dalam satu kata.....	29
4.3.3	Jumlah katakunci.....	30
4.4	Eksperimen.....	31
4.4.1	Perangkat yang Digunakan.....	31
4.4.2	Dampak tiap peubah.....	31
4.4.3	Formula.....	33
4.5	Evaluasi.....	35
4.5.1	Evaluasi metode baru.....	35
4.5.2	Komparasi dengan 10 metode state-of-the-art.....	36
4.6	Proof of Concepts.....	39
4.6.1	Metoda.....	39
4.6.2	Antarmuka Pengguna.....	40
4.6.3	Cara Penggunaan.....	40
4.6.4	Akurasi.....	44
4.7	Keterbatasan.....	44
5	PENUTUP.....	45
5.1	Simpulan.....	45
5.2	Luaran.....	46
5.3	Saran.....	46

DAFTAR PUSTAKA.....	47
---------------------	----

1 PENDAHULUAN

Kata/frasa kunci sangat penting untuk membantu menyelesaikan berbagai permasalahan dan berguna juga untuk keperluan-keperluan lainnya. Salah satunya adalah dalam hal pengolahan citra (*image*), beberapa hal diantaranya: klasifikasi citra [1], anotasi citra [2, 3, 4, 5], keterangan citra (*image caption*) [6], temu kembali citra [7], pencarian citra [8], pencarian kemiripan citra merk dagang [9], pengenalan kata-gambar kuno [10], deteksi keberadaan suatu produk [11], dan temu kembali citra berdasarkan kemiripannya [12].

Dalam bidang paten, kata/frasa kunci juga mempunyai berbagai kegunaannya. Beberapa kegunaannya antara lain: anotasi semantik paten [13], menemukan peluang teknologi dengan analisis paten berbasis katakunci [14, 15], ekstraksi informasi semantik dari paten (Ding, Wang, & Zhu, 2019), ekstraksi pengetahuan dari paten [16], klasifikasi paten [17], dan temu kembali paten [18, 19].

Kata/frasa kunci juga dibutuhkan pada riset bidang-bidang lain selain yang tersebut di atas. Permasalahan muncul ketika mendapati di Internet bahwa salah satu pengindeks dokumen/artikel terbesar yaitu Google Scholar yang menjadi rujukan Sinta (Science and Technology Index) dan juga mengindeks dokumen lebih banyak dari Scopus ternyata tidak menyertakan katakunci dokumen/artikel dalam metadatanya. Jika dokumen yang tidak memiliki katakunci dibuatkan secara manual maka akan membutuhkan biaya sangat mahal karena jumlah artikel yang terindeks GS sudah sangat banyak. Pada tahun 2014 Google Scholar telah mengindeks 240 juta artikel ilmiah [20], saat ini tentu sudah jauh lebih banyak.

Untuk menyelesaikan permasalahan tersebut perlu adanya metode ekstraksi kata/frasa kunci. Berbagai metode ekstraksi telah dikembangkan sejak puluhan tahun yang lalu, antara lain: ekstraksi katakunci berdasarkan derajat relevansi kata [21], ekstraksi katakunci otomatis berbasis grafik dari dokumen teks [22], ekstraksi katakunci berdasarkan Word2Vec dan TextRank [23], ekstraksi katakunci dengan model jaringan saraf [24], analisis koleksi teks untuk ekstraksi katakunci [25], ekstraksi katakunci dalam bahasa Jerman: teori informasi vs *deep learning* [26], model LSTM-CRF dua arah untuk ekstraksi katakunci dalam berita olahraga Cina [27], YAKE! ekstraksi katakunci dari dokumen tunggal menggunakan banyak fitur lokal [28], ekstraksi

katakunci berbasis DOM dari halaman Web [29], ekstraksi katakunci otomatis menggunakan *Support Vector Machine* [30], dan indeks multi-sentralitas untuk ekstraksi katakunci berbasis grafik [31].

Adapun metode ekstraksi katakunci yang dikembangkan pada penelitian ini menggunakan metode yang diambil dari ilmu sainsmetrika (*scientometrics*) dan pembelajaran mesin. Dimana sainsmetrika adalah bidang studi yang mengevaluasi ciri-ciri dan perkembangan ilmu pengetahuan dan teknologi dengan menggunakan data bibliometrik serta metode pengukuran kuantitatif lainnya [32], dan menurut Narin et al. [33] sainsmetrika adalah bidang studi yang mengukur dan menganalisis kinerja sains dan teknologi dengan mempertimbangkan faktor-faktor seperti publikasi, sitasi, dan kolaborasi. Sedangkan pembelajaran mesin menurut Ethem Alpaydin [34] dalam bukunya yang berjudul "*Introduction to Machine Learning (3rd ed.)*" adalah bidang studi yang menggunakan algoritma untuk membuat sistem komputer yang dapat belajar dari data dan melakukan tugas tertentu tanpa dikode secara eksplisit. Defini lain dari pembelajaran mesin adalah bidang studi yang berkaitan dengan pengembangan algoritma yang memungkinkan komputer untuk membuat prediksi atau keputusan berdasarkan data [35].

Sainsmetrika dalam penelitian Ekstraksi Katakunci dalam disertasi ini adalah untuk menentukan atau membentuk daftar kosakata populer. Selanjutnya, kosakata populer ini dipakai sebagai dataset pada pembelajaran mesin untuk melakukan prediksi katakunci dari suatu artikel ilmiah. Pembelajaran mesin yang dipakai adalah berjenis tersupervisi (*supervised learning*). Pembelajaran mesin jenis ini memerlukan dataset untuk pembelajarannya. Dataset diperlukan untuk membuat model pembelajaran mesinnya.

1.1 PERUMUSAN MASALAH

1. Bagaimana menentukan/memilih kosakata agar dapat menghasilkan pengekstrak katakunci dengan kinerja yang baik?
2. Bagaimana penggunaan sistem pembelajaran mesin untuk ekstraksi katakunci, metoda apakah yang menghasilkan kinerja yang terbaik?
3. Bagaimana strategi implementasi dari metode ekstraksi katakunci ini agar dapat dimanfaatkan oleh para penulis artikel ilmiah?

1.2 RUANG LINGKUP

Penelitian ini difokuskan pada ekstraksi kata/frasa kunci dari bibliografi atau metadata bahan pustaka. Bibliografi atau metadata bahan Pustaka yang dipakai antara lain adalah: judul, katakunci dan abstrak artikel, jumlah sitasi, kualitas penerbit (jurnal/konferensi), dan tahun diterbitkannya. Dalam pencarian metadata bahan pustaka tidak menutup kemungkinan diperoleh juga data pribadi penulis dan terjadi gangguan pada server komputer penyedia data. Privasi, keamanan dan ketentuan hukum tidak dibahas dalam penelitian ini.

Penyelesaian semua pertanyaan penelitian tersebut hanya menggunakan konsep, model maupun algoritma dari ranah ilmu temu kembali informasi (*Information Retrieval*) dengan menitik-beratkan pada penggalian data (*data mining*) dan lebih sempit lagi pada penggalian web (*web mining*). Selain itu juga digunakan teori dari ranah ilmu perpustakaan dan informasi (*library and information science/LIS*). Dari ranah LIS tersebut diambil ilmu bidang sainsmetrika untuk membentuk kosakata populer sebagai dataset.

1.3 TUJUAN DAN MANFAAT

Penelitian ini dilakukan dengan tujuan untuk mendapatkan algoritma dan metode ekstraksi kata/frasa kunci dari suatu artikel ilmiah menggunakan pembelajaran mesin tersupervisi. Dataset yang digunakan adalah metadata artikel ilmiah dari pengindeks dokumen bereputasi. Tetapi karena metadata tersebut tidak bisa digunakan secara langsung maka diperlukan beberapa metoda dalam sainsmetrika untuk membuat kosakata populer yang menjadi bahan dasar dataset dalam penelitian ini.

Adapun manfaat penelitian ini apabila berhasil menjawab semua pertanyaan penelitian di atas antara lain adalah untuk:

1. Ekstraksi/pencarian frasa/katakunci dari suatu artikel ilmiah, yang dapat dimanfaatkan pada berbagai riset sebagaimana disebut dalam Bagian 1.1.
2. Diperolehnya metode dan algoritma ekstraksi katakunci baru.

1.4 SIGNIFIKANSI PENELITIAN

Dari banyaknya permasalahan yang harus diselesaikan pada penelitian ini, tentunya harus mempunyai nilai penting yang banyak juga. Beberapa kontribusi (*novelty*) yang diharapkan muncul dari penelitian ini antara lain adalah:

1. Metoda dan algoritma baru dalam ekstraksi kata/frasa kunci.
2. Model kosakata untuk ekstraksi katakunci.

1.5 PENELITIAN AWAL

Penelitian awal terkait dengan bidang web mining, bibliografi/metadana artikel ilmiah, dan sainsmetrika telah dilakukan dan hasilnya telah dipublikasikan baik dalam berbagai konferensi. Beberapa publikasi terindeks Scopus yang telah dilakukan dan berkaitan dengan penelitian ini yaitu:

1. Achsan, H.T.Y., Suhartanto, H., Wibowo, W.C., Dewi, D.A., Ismed, K. (2023). *Automatic Extraction of Indonesian Stopwords*. International Journal of Advanced Computer Science and Applications, 14(2), pp. 166–171.
2. Achsan, H.T.Y., Suhartanto, H., Wibowo, W.C., Putri, W.T.H (2020). *An Approach for Measuring Research Strength Map of an Institution*. Journal of Physics: Conference Series 1566 (1).
3. Achsan, H.T.Y., Suhartanto, H., Wibowo, W.C. (2019). *A Technique in Information Retrieval and Bibliometrics to Check the Reliability of an Article Indexing*. Proceedings of the 4th International Conference on Contemporary Computing and Informatics, IC3I 2019, Singapore. pp. 148-153
4. Achsan, H.T.Y. (2019). *The research trends of partnership based on scientific publications*. Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2019, Saint Petersburg, Russia. pp. 164-168.

5. Achsan, H.T.Y., Wibowo, W.C., Purnama, D.G., Achsan, M.M.B. (2019). *Mining of Russian bibliography*. Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2019, Saint Petersburg, Russia. pp. 169-173.
6. Achsan, H.T.Y., Wibowo, W.C., Achsan, M.M.B., Purnama, D.G., Lubis, K.B. (2019). *The quality of Indonesian scientific articles and its neighboring countries*. Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2019, Saint Petersburg, Russia. pp. 174-178.
7. Achsan, H.T.Y., Wibowo, W.C., Ganesha, H., Achsan, M.M.B. (2019). *The importance of computer science in industry 4.0*. 2018 International Conference on Advanced Computer Science and Information Systems, ICACISIS 2018 8618217, pp. 29-37.
8. Achsan, H.T.Y., Wibowo, W.C., Putri, W.T.H., Achsan, M.M.B., Barcah, Q.K.D. (2019). *Harvesting bibliography multi-thread, safe and ethical web crawling*. 2018 International Conference on Advanced Computer Science and Information Systems, ICACISIS 2018, pp. 355-360.
9. Achsan, H.T.Y., Wibowo, W.C., Putri, W.T.H., Achsan, M.M.B. (2018). *Visulization of Indonesian Bibliography : A Scientometrics Approach*. 2018 International Workshop on Big Data and Information Security, IW BIS 2018, pp. 75-80.
10. Achsan, Harry T.Y.; Wibowo, Wahyu Catur; Suhartanto, Heru. (2017). *Ontology Enrichment for Multi-Domain Knowledge and Expertise Representation*. 28th DAAAM International Symposium, Zadar, Croatia.
11. Achsan, Harry T.Y.; Wibowo, Wahyu Catur; Suhartanto, Heru. (2017). *The Quality and Research Trend on Subject Area "Automation & Manufacturing" Published by DAAAM*. DAAAM Scientific Book 2017, Vol.

16, ISSN 1726-9687, ISBN 978-3-902734-12-9, Vienna University of Technology, Vienna, Austria.

12. Achsan, Harry T.Y. & Wibowo, Wahyu Catur. (2013). A Fast Distributed Focused-web Crawling. *Procedia Engineering*, volume 69, ISSN 1877-7058 (pp. 492-499). Elsevier BV.

Keterangan: terindex Scopus, disitasi: 19 kali.

13. Achsan, H.T.Y. (2011). Web mining based on user profile and preferences. *Annals of DAAAM and Proceedings of the International DAAAM Symposium* pp. 821-822.

Publikasi nomor 8 s/d 13 adalah hasil studi tentang pemanenan bibliografi atau metadata artikel ilmiah lalu menerapkannya pada penelitian awal bidang sainsmetrika. Publikasi nomor 2 s/d 7 merupakan hasil studi metode sainsmetrika dan sebagiannya digabungkan dengan pembelajaran mesin untuk melakukan: a) prediksi kekuatan riset institusi, b) prediksi validitas pengindeks artikel ilmiah, c) prediksi kecenderungan kemitraan, d) menghitung kualitas artikel ilmiah dari peneliti Indonesia dan negara tetangga, e) prediksi kecenderungan (*trend*) teknologi yang akan berkembang dan digunakan pada era *Industry 4.0*. Dari berbagai metode prediksi dengan menggunakan ilmu bidang sainsmetrika dan pembelajaran mesin menghasilkan ide dasar prediksi katakunci dari artikel ilmiah menggunakan bidang-bidang ilmu tersebut.

Sedangkan artikel nomor 1 merupakan studi awal untuk melakukan pencarian kata henti (*stop words*) dari artikel berbahasa Indonesia. Publikasi ini sebenarnya merupakan kelanjutan dari penelitian tentang pencarian kata henti dari artikel berbahasa Inggris, namun karena sudah banyak artikel yang membahas pembentukan kata henti dari artikel berbahasa Inggris maka hasil penelitian ini tidak dipublikasikan. Kata henti dalam Bahasa Inggris diperlukan dalam penelitian ini dalam pra-proses, yaitu dalam penghapusan kata henti.

2 TINJAUAN PUSTAKA

Penelitian ini masuk dalam ranah ilmu *data mining* (DM), bagian dari bidang ilmu Temu Kembali Informasi (*Informational Retrieval*). Dimana Data Mining adalah proses menemukan pola atau informasi yang berguna dari data besar yang terstruktur, semi-terstruktur, atau tidak terstruktur dengan menggunakan algoritma pembelajaran mesin [36]. Sedangkan Temu Kembali Informasi adalah teknik untuk menemukan informasi yang relevan dari kumpulan data yang besar, dimana pembelajaran mesin digunakan dalam teknik ini untuk meningkatkan efektivitas pencarian [37].

Pada bab ini akan dijelaskan tentang Data Mining dan metadata bahan pustaka yang akan digunakan sebagai dasar pembuatan dataset untuk pembelajaran mesin pada penelitian ini. Namun sebelumnya akan dibahas tentang data, metadata dan metadata bahan pustaka terlebih dahulu supaya lebih mudah untuk dipahami. Setelah penjelasan tentang Data Mining dilanjutkan dengan pembahasan tentang Klasifikasi yang merupakan bagian dari ilmu bidang Pembelajaran Mesin. Lalu dilanjutkan dengan kajian tentang Kosakata Populer yang dijadikan dataset untuk pembelajaran mesin serta penjelasan tentang kata/frasa kunci yang merupakan luaran (*output*) dari pembelajaran mesin. Pada akhir bab ini dibebaskan riwayat penelitian ekstraksi katakunci yang dilakukan para peneliti sebelumnya.

2.1 METADATA BAHAN PUSTAKA

Bahan pustaka ada beraneka ragam formatnya, antara lain: buku, peta, musik, *game*, film, dan berkas komputer. Metadata bahan Pustaka disebut juga dengan bibliografi. Bibliografi sebagaimana disebut dalam kamus Merriam-Webster adalah sebuah daftar buku, majalah, artikel, dan lain-lain tentang suatu subyek [38], sedangkan dalam kamus Oxford dinyatakan sebagai sebuah daftar buku yang direferensikan dalam karya ilmiah biasanya tercetak sebagai lampiran [39]. Dari kamus-kamus yang lain artinya juga hampir sama yang pada intinya adalah daftar pustaka.

2.2 KOSAKATA POPULER

Kosakata populer adalah himpunan kata/frasa yang terorganisasi yang diambil berdasarkan popularitasnya dalam banyak dokumen. Daftar kosakata populer dibuat untuk mengekspresikan suatu ide dengan pilihan kata terbaik. Dimana salah satu definisi populer adalah gaya bahasa yang mengandung unsur-unsur tertentu seperti kejelasan dan kebersahajaan untuk menarik perhatian [40].

Dari artikel-artikel ilmiah populer diambil katakuncinya, dimana katakunci mencerminkan tema-tema inti dalam suatu artikel ilmiah, untuk dimasukkan dalam daftar kosakata populer. Tingkat kepopuleran dari artikel ilmiah dapat didasarkan pada jumlah sitasi pada artikel yang memuat kosakata/katakunci tersebut, dimana semakin tinggi sitasi dari suatu artikel ilmiah berarti artikel tersebut semakin populer karena banyak yang menyadurnya. Daftar kosakata ini dipakai sebagai dataset untuk pembuatan model pembelajaran mesin pada ekstraksi kata/frasa kunci dari suatu metadata artikel ilmiah dalam penelitian ini.

2.3 KATA/FRASA KUNCI

Kata/frasa kunci, untuk selanjutnya bisa disebut sebagai katakunci saja, adalah istilah yang menggambarkan topik-topik inti dari suatu artikel. Beberapa definisi yang lainnya adalah: kata atau frasa yang merupakan inti atau topik utama dalam teks, dokumen, atau konten tertentu [41], sedangkan dalam dunia riset definisi katakunci adalah istilah atau frase yang digunakan untuk mencari dan mengidentifikasi sumber informasi yang relevan dalam literatur ilmiah atau pangkalan data [42].

Banyak kegunaan dari katakunci, terutama pada temu kembali informasi dan *Natural Language Processing* (NLP). Beberapa contoh kegunaan katakunci antara lain adalah untuk:

1. Klasifikasi [43, 44, 45, 46],
2. Klustering [47, 48, 49, 50],
3. Investigasi *infodemic monikers* [51],
4. Reviu artikel ilmiah [52, 53, 54, 55, 56, 57, 58, 59, 60, 61],
5. Rekomendasi tindakan/ramuan pencegahan Covid-19 [62, 63],
6. Deteksi wabah coronavirus [64],
7. Prediksi kecenderungan (*trend*) penelitian [65, 66, 67],

8. Pencarian data spasial [68, 69],
9. Berbagi data sensitive [70, 71], dan
10. Temu kembali informasi lainnya [72, 73, 74, 75]

Banyaknya kegunaan katakunci memperlihatkan pentingnya suatu katakunci dari dokumen. Namun sayangnya banyak sekali dokumen yang tidak disertai dengan katakunci. Pencarian dengan katakunci a pada Google Scholar menghasilkan lebih dari sebelas juta artikel. Google Scholar menyediakan metadata dari setiap dokumen yang dia indeks, lihat **Error! Reference source not found.**. Atribut metadatanya antara lain adalah: Authors, Publikation date, Journal, Description, dan Total citations. Tetapi Google Scholar tidak menyediakan katakunci dalam metadatanya. Hal inilah yang menarik para peneliti untuk mengembangkan metode dan algoritma ekstraksi katakunci.

2.4 RIWAYAT PENELITIAN EKSTRAKSI KATAKUNCI

Pada Riwayat pengembangan metode ekstraksi katakunci ini dijelaskan metodologi penelitian yang telah dilakukan oleh peneliti sebelumnya. Hal ini dilakukan untuk membangun landasan teori dan konteks penelitian yang lebih kuat dan untuk menunjukkan bagaimana penelitian sebelumnya dapat membantu dalam memahami fenomena yang sedang diteliti.

Tabel 2-1. Riwayat pengembangan metode-metode ekstraksi katakunci tanpa supervisi.

Tahun	Pengembang Metode	Stat.	Stats Graph	Clust.	LDA	KG	C/N	Sem.	Lang
2003	Tomokiyo and Hurst (2003)								√
2004	Mihalcea and Tarau (2004)		√						
2008	Wan and Xiao (2008) - SingleRank		√						
	Wan and Xiao (2008) - ExpandRank		√				√		
2009	Liu et al. (2009)	√		√					
	El-Beltagy & Rafea (2009)	√							
2010	Liu et al. (2010)		√		√				
	Rose et al. (2010)		√						
2013	Bougouin et al. (2013)		√	√					
2014	Gollapalli & Caragea (2014)		√				√		
	Wang et al. (2014)		√					√	
2015	Sterckx et al. (2015a)		√		√				
	Sterckx et al. (2015b)		√		√				
	Danesh et al. (2015)		√						
	Wang et al. (2015)		√					√	
2017	Teneva and Cheng (2017)		√		√				
	Florescu and Caragea (2017b)		√						
	Shi et al. (2017)		√	√				√	
2018	Campos et al. (2018b)	√							
	Boudin (2018)		√	√					
	Bennani-Smires et al. (2018)							√	
	Papagiannopoulou and Tsoumakas (2018)	√						√	
	Mahata et al. (2018)		√					√	
	Yu and Ng (2018)		√				√	√	
2019	Won et al. (2019)	√							

[76]

Tabel 2-2. Metode ekstraksi katakunci tersupervisi dalam dua dekade terakhir: a) metode dihasilkan oleh para peneliti sebelumnya, b) hasil penelitian dalam disertasi ini.

a) metode-metode yang dihasilkan oleh para peneliti sebelumnya

Tahun	Pengembang	Algoritma PM	Stat.	Posit.	Ling.	Cont.	Stack.	Ext.
1999	Witten et al. (1999)	<i>Naive Bayes</i>	√	√				
2003	Hulth (2003)	<i>Rule Induction/Bagging</i>	√		√	√		
2007	Nguyen and Kan (2007)	<i>Naive Bayes</i>	√	√	√			
2008	Shi et al. (2008)	<i>Logistic Regression</i>	√	√	√			√
2009	Medelyan et al. (2009)	<i>Bagged Decision Trees</i>	√	√	√			√
	Jiang et al. (2009)	<i>SVM</i>	√	√	√			
2010	Nguyen and Luong (2010)	<i>Naive Bayes</i>	√	√	√			√
2014	Caragea et al. (2014)	<i>Naive Bayes</i>	√	√	√	√		
2016	Zhang et al. (2016)	<i>RNN</i>						√
	Bougouin et al. (2016)	<i>Graph-based Method</i>	√					√
2017	Wang and Li (2017)	<i>Ensemble (RF/SVM)</i>	√	√	√	√	√	√
	Meng et al. (2017)	<i>seq2seq Learning</i>				√		
	Gollapalli et al. (2017)	<i>CRFs</i>	√	√	√	√	√	√
	Zhang et al. (2017)	<i>Random-walk</i>	√	√	√	√		
2018	McIlraith and Weinberger (2018)	<i>Naive Bayes</i>	√			√		
	Chen et al. (2018)	<i>seq2seq Learning</i>				√		
	Ye and Wang (2018)	<i>Multi-task Learning (seq2seq Model)</i>				√		
	Basaldella et al. (2018)	<i>Bi-LSTM RNN Task-oriented LDA Model</i>	√					√
	Yang et al. (2018)	<i>Topic-based</i>				√		√
	Wang et al. (2018)	<i>Adversarial Neural Network</i>						
2019	Alzaidy et al. (2019)	<i>Bi-LSTM-CRF Sequence Labeling</i>				√		√

[76]

b) metode hasil penelitian dalam disertasi ini

Tahun	Pengembang	Algoritma PM	Stat.	Posit.	Ling.	Cont.	Stack.	Ext.
2023	Harry T Y Achsan	Sainsmetrika dan klasifikasi multilabel	√		√			√

3 METODOLOGI

3.1 PENGANTAR

Metodologi riset diperlukan dalam suatu penelitian karena merupakan dasar untuk perencanaan dan perancangan penelitian, agar setiap langkah dalam proses penelitian berjalan dengan baik dan benar. Dengan proses seperti itu diharapkan dapat memastikan tercapainya luaran dan tujuan riset, serta hasilnya adalah valid dan dapat dipercaya. Metodologi riset dibahas mulai dari kerangka teoretis, rancangan penelitiannya, hingga langkah-langkah yang dilakukan dalam melaksanakan penelitian ini.

3.2 PENDEKATAN

Ada tiga pendekatan dalam melakukan penelitian yaitu metode kuantitatif, metode kualitatif dan metode campuran (gabungan dari metode kuantitatif dan kualitatif) [42]. Masing-masing pendekatan dapat diterapkan pada penelitian-penelitian tertentu. Penelitian kausal/sebab-akibat atau penelitian yang menggunakan eksperimen akan lebih tepat menggunakan kuantitatif [77]. Sedangkan penelitian yang melihat konteks sosial di sekeliling responden atau yang belum diketahui secara pasti peubah-peubahnya akan lebih tepat jika menggunakan pendekatan metode kualitatif [77].

Penelitian untuk disertasi ini pada awalnya dilakukan dengan pendekatan metode kualitatif. Pendekatan ini diambil karena pada awal dilakukan penelitian tidak diketahui peubah-peubah apa saja yang berpengaruh. Untuk mengetahui peubah-peubahnya maka dilakukan observasi, yaitu dengan melakukan pengamatan secara langsung dan detil terhadap berbagai sumberdata untuk mendapatkan informasinya dan menentukan peubah-peubah yang berpengaruh pada penelitian ini. Hasil dari observasi diperoleh data/peubah yang bersifat kualitatif karena datanya berupa teks/tekstual seperti judul dan abstrak suatu artikel ilmiah.

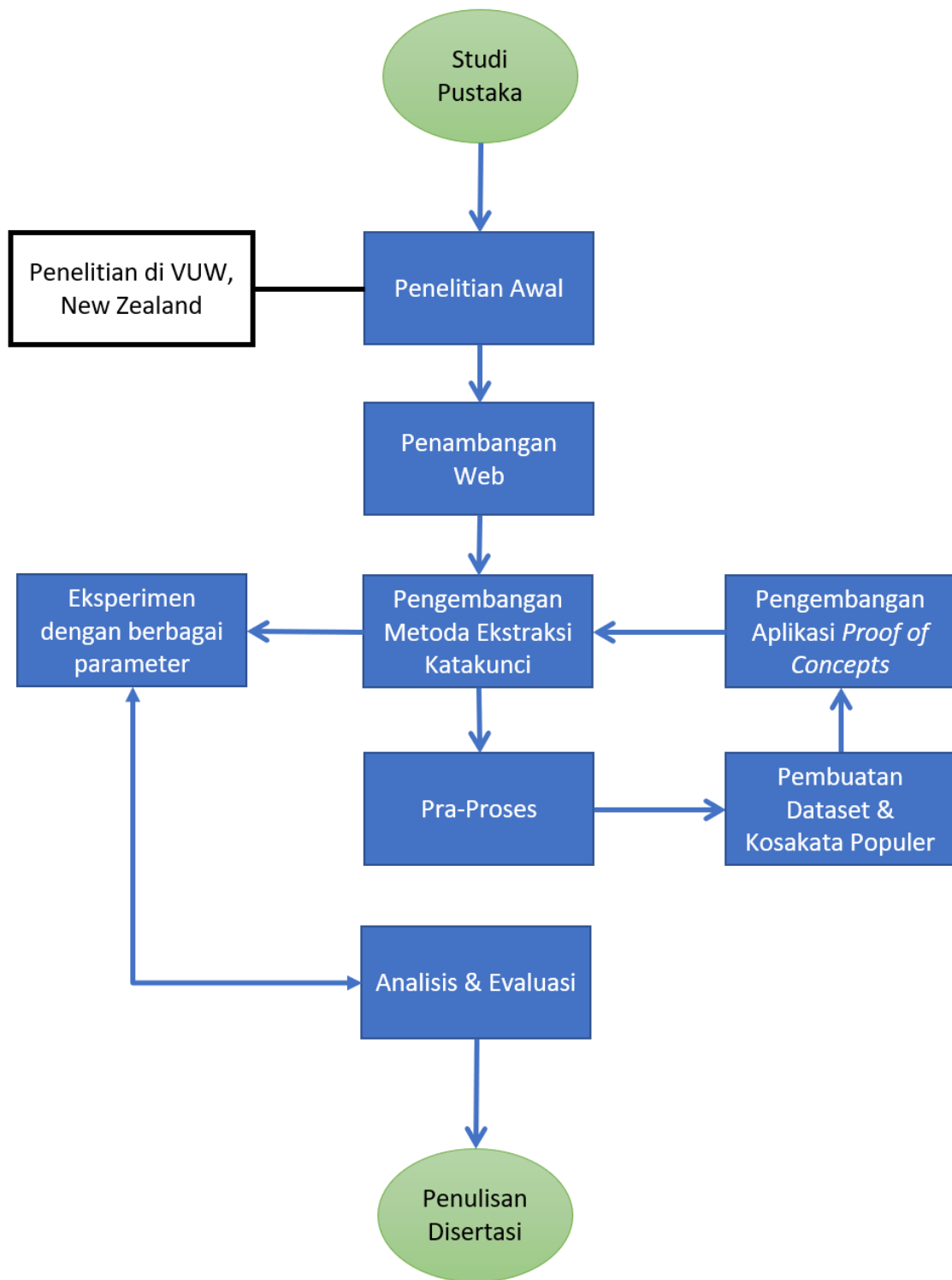
Jumlah item dari data yang diperoleh dari salah satu sumber data ternyata banyak, ada 50.000 item. Dengan data sebesar itu tentu harus dibantu perangkat komputasi. Untuk itu, data diubah dari kualitatif, dalam penelitian ini adalah data tekstual, menjadi data kuantitatif yang berupa angka-angka. Guna mengolah data numerik tersebut digunakan pendekatan dengan metode kuantitatif.

Karena menggunakan pendekatan kombinasi antara metode kualitatif dan kuantitatif, maka penelitian ini lebih tepat disebut menggunakan pendekatan metode campuran (*mixed method*). Ada tiga jenis pendekatan metode campuran, yaitu konvergen, eksplanatori dan eksploratori [78]. Pendekatan metode campuran konvergen adalah apabila peneliti mengumpulkan dan menganalisis data kuantitatif dan kualitatif secara bersamaan. Pendekatan metode campuran sekuensial eksplanatori adalah apabila peneliti melakukan penelitian awal dengan pendekatan metode kuantitatif lalu dilakukan penelitian dengan lebih detil menggunakan pendekatan metode kualitatif. Sedangkan pendekatan metode campuran sekuensial eksploratori adalah kebalikan dari eksplanatori, yaitu peneliti melakukan penelitian dengan pendekatan metode kualitatif sehingga diperoleh data kualitatif, lalu diubah menjadi data kuantitatif/numerik/angka dan dianalisis dengan menggunakan pendekatan metode kuantitatif. Dari tiga jenis metode campuran tersebut, penelitian ini sesuai dengan pendekatan metode campuran sekuensial eksploratori.

3.3 KLASIFIKASI MULTI-LABEL

Klasifikasi multi-label atau klasifikasi multi-output adalah varian dari masalah klasifikasi dalam pembelajaran mesin, dimana beberapa label non-eksklusif dapat ditetapkan ke setiap objek. Dalam klasifikasi multi-label, label tidak eksklusif dan tidak ada batasan pada berapa banyak kelas/label yang dapat ditempelkan ke setiap objek.

Dalam penelitian ini, klasifikasi multi-label diterapkan pada objek input/masukan berupa metadata artikel ilmiah. Adapun luarannya adalah label-label yang diambil dari daftar kosakata populer.



Gambar 3.1. Alur langkah-langkah dalam penelitian ini.

3.4 LANGKAH-LANGKAH

Beberapa langkah-langkah yang dilakukan dalam penelitian ini adalah: peninjauan pustaka, penambangan web, klasifikasi, pengembangan beberapa konsep, dan pengembangan sistem. Beberapa publikasi ilmiah dihasilkan dalam pelaksanaan langkah-langkah tersebut sebagaimana disebutkan dalam sub-bab 1.2. Publikasi tidak hanya dilakukan penulis sendirian tetapi berkolaborasi dengan peneliti lain maupun pembimbing.

Masukan utama diperoleh dari pembimbing penelitian ini. Penulis juga melakukan kolaborasi penelitian dengan pihak lain yang tertuang dalam tulisan ilmiah. Selain itu, penulis juga melakukan penelitian di luar negeri, yaitu di School of Engineering and Computer Science, Victoria University of Wellington, New Zealand. Selama tiga bulan di New Zealand, penulis dibimbing oleh beberapa guru besar dari New Zealand dan Amerika.

3.4.1 Peninjauan Pustaka

Peninjauan pustaka adalah salah satu tahapan dalam penelitian yang bertujuan untuk mengumpulkan dan mengevaluasi sumber-sumber literatur yang relevan dengan topik penelitian atau karya ilmiah yang akan dilakukan. Peninjauan pustaka dilakukan untuk memperoleh pemahaman yang lebih baik tentang topik, mengidentifikasi gap pengetahuan, menemukan teori-teori atau model-model yang relevan, dan menghindari duplikasi penelitian yang sudah dilakukan sebelumnya.

Pada langkah ini penulis mengumpulkan berbagai bahan pustaka dari beberapa perpustakaan, basisdata web, dan dari berbagai laman web. Peninjauan dititik-beratkan pada teori-teori pendukung pelaksanaan penelitian ini dan melakukan evaluasi penelitian orang lain yang terkait dengan penelitian ini. Penelitian orang lain diambil dari prosiding konferensi ilmiah maupun dari jurnal ilmiah.

3.4.2 Penelitian Awal

Penelitian awal (*preliminary research*) adalah tahap awal dalam sebuah penelitian yang bertujuan untuk mengumpulkan informasi dan data dasar tentang topik yang akan diteliti. Penelitian awal dilakukan untuk memperoleh pemahaman yang lebih baik tentang topik dan masalah yang akan diteliti, serta untuk membantu peneliti dalam

merancang penelitian yang lebih efektif dan efisien. Dalam penelitian ilmiah, penelitian awal merupakan tahapan penting yang harus dilakukan sebelum melakukan penelitian utama atau eksperimen. Dengan melakukan penelitian awal yang cermat, peneliti dapat menghindari kesalahan dalam merancang penelitian dan meningkatkan validitas dan kepercayaan hasil penelitian yang akan dilakukan.

Berbagai penelitian telah dilakukan pada fase ini untuk memahami dan memperdalam berbagai metoda ekstraksi katakunci. Luaran dari penelitian awal ini berupa belasan publikasi, lihat bagian 1.5. Penelitian awal difokuskan pada bidang ilmu sainsmetrika dan pembelajaran mesin. Dataset yang diperlukan pada penelitian awal ini dibentuk dari hasil penambangan web.

3.4.3 Penambangan Web

Penambangan web dilakukan untuk mendapatkan dataset yang akan digunakan pada klasifikasi bibliografi. Penambangan dilakukan pada situs WorldCat dengan menggunakan Focused Web Crawler terdistribusi [79], yaitu menambang web menggunakan satu sistem terdistribusi. Dengan sistem terdistribusi pada ratusan komputer maka penambangan dapat dilakukan jauh lebih cepat. Penambangan web juga dilakukan pada beberapa pengindeks dokumen, dimana salah satunya adalah Scopus. Dari Scopus diperoleh metadata artikel ilmiah atau bibliografi dengan fitur cukup lengkap. Berdasarkan bibliografi serta pemahaman tentang sainsmetrika dan pembelajaran mesin yang diperoleh pada penelitian awal maka dikembangkanlah metoda ekstraksi katakunci.

3.4.4 Pengembangan Metoda Ekstraksi Katakunci

Metoda yang dikembangkan didasarkan pada ilmu bidang sainsmetrika dan pembelajaran mesin. Dari sainsmetrika diperoleh ide penggunaan bibliografi untuk mengembangkan dataset yang akan dipakai pada pembelajaran mesin. Pembelajaran mesin yang digunakan adalah klasifikasi untuk memprediksi label atau katakunci dari suatu artikel ilmiah.

Algoritma klasifikasi semi otomatis telah dikembangkan guna melakukan pengelompokan bibliografi berdasar nomor panggilnya (*call number*) atau kode katalog. Kode katalog masih penting meskipun pencarian bahan pustaka sudah dapat dilakukan

dengan mudah menggunakan mesin pencari yang dimiliki perpustakaan. Kode katalog mampu meningkatkan akurasi mesin pencari, selain itu kode katalog juga dipakai pada penyusunan dokumen/buku dalam rak sehingga pengunjung perpustakaan maupun pustakawan dapat mencari dokumen dengan mudah dan cepat.

Penelitian ekstraksi katakunci ini juga menggunakan salah satu metode klasifikasi. Klasifikasi secara otomatis merupakan salah satu bagian dari pembelajaran mesin (*Machine Learning/ML*) yang dipakai untuk memprediksi label atau kelas suatu objek dalam hal ini adalah dokumen/artikel ilmiah. Klasifikasi dilakukan terhadap katakuncinya. Dengan adanya katakunci hasil klasifikasi ini, sebagaimana katalog yang dibahas di atas, maka pencarian dokumen oleh mesin pencari akan lebih cepat dan tepat. Karena dalam satu dokumen/artikel biasanya terdapat lebih dari satu katakunci maka klasifikasi pada ekstraksi katakunci termasuk dalam *Multi-label Classification*.

Multi-label classification (klasifikasi multi-label) adalah sebuah teknik dalam data mining dan machine learning yang digunakan untuk memprediksi beberapa label atau kategori untuk sebuah data atau objek. Dalam multi-label classification, setiap data atau objek dapat memiliki lebih dari satu label atau kategori. Objek yang diklasifikasi pada penelitian ini adalah dokumen atau artikel ilmiah, sedangkan labelnya adalah katakunci.

Dataset diperlukan pada multi-label classification. Dataset dibentuk dari bibliografi hasil penambangan web. Namun sebelum bisa dijadikan dataset maka data mentah hasil penambangan web harus melewati pra-proses terlebih dahulu.

3.4.5 Pra-Proses

Pra-proses atau *preprocessing* adalah salah satu tahapan penting dalam analisis data yang bertujuan untuk membersihkan, menyiapkan, dan mentransformasi data mentah menjadi bentuk yang dapat digunakan untuk analisis lebih lanjut. Pra-proses melibatkan beberapa kegiatan seperti pemilihan data, penghapusan data yang tidak relevan atau duplikat, pengisian data kosong atau hilang, serta transformasi data seperti normalisasi dan standarisasi.

Tahapan pra-proses sangat penting karena kualitas data yang buruk dapat menyebabkan hasil analisis yang salah atau tidak akurat. Dengan melakukan pra-proses

yang baik, data akan menjadi lebih terstruktur dan bersih, sehingga dapat meningkatkan kualitas analisis dan memperoleh wawasan yang lebih baik dari data.

Beberapa teknik yang digunakan dalam pra-proses antara lain adalah: a) pembersihan data (*data cleaning*): mengidentifikasi dan menghapus data yang tidak relevan atau duplikat, serta mengatasi data yang hilang atau kosong, b) transformasi data: mengubah bentuk atau format data menjadi bentuk yang lebih mudah diolah atau dipahami, contohnya adalah normalisasi, standardisasi, dan pengubahan skala, c) seleksi fitur (*feature selection*): memilih fitur atau variabel yang paling relevan dan berpengaruh dalam analisis, d) penghapusan kata henti, e) tokenisasi, dan f) stemming.

3.4.6 Pembuatan Dataset & Kosakata Populer

Pembuatan dataset adalah proses pengumpulan data dari berbagai sumber untuk digunakan dalam analisis data. Dataset dapat diperoleh dari berbagai sumber seperti survei, sensor, platform media sosial, dan lain sebagainya. Pentingnya dataset dalam analisis data adalah untuk memberikan representasi yang akurat dari fenomena atau objek yang ingin diteliti.

Pada penelitian ini dataset dibentuk dari bibliografi melalui rangkaian proses tertentu. Sedangkan kosakata populer merupakan daftar katakunci yang diambil dari artikel ilmiah bersitasi tinggi.

3.4.7 Pengembangan Aplikasi Proof of Concepts

Pengembangan sistem dilakukan untuk membuktikan bahwa konsep dan algoritma yang telah dikembangkan adalah benar dapat diterapkan. Sistem ini diharapkan dapat merepresentasikan pengetahuan sesuai konsep yang telah dikembangkan. Sistem juga harus dapat: membuat dan menyimpan profil pengguna, mendeteksi latar belakang pengetahuan pengguna, serta dapat mengidentifikasi kepakaran pengguna.

3.4.8 Eksperimen dengan berbagai parameter

Eksperimen dengan berbagai parameter adalah proses menguji berbagai konfigurasi atau pengaturan parameter dalam model atau algoritma pembelajaran mesin untuk memperoleh kinerja yang optimal pada tugas tertentu. Konfigurasi atau

pengaturan parameter yang tepat dapat meningkatkan kinerja model atau algoritma dalam memprediksi atau mengklasifikasikan data.

Parameter yang dapat disesuaikan dalam pembelajaran mesin termasuk parameter pembelajaran, parameter regularisasi, parameter arsitektur model, dan parameter optimisasi. Parameter pembelajaran dapat disesuaikan untuk mengatur kecepatan pembelajaran model, sedangkan parameter regularisasi digunakan untuk mengurangi overfitting dalam model. Parameter arsitektur model meliputi formula yang digunakan dalam model. Sedangkan, parameter optimisasi digunakan untuk mengatur strategi optimisasi untuk menemukan nilai optimal dari parameter model.

Proses eksperimen dengan berbagai parameter melibatkan pemilihan sejumlah parameter dan nilai-nilainya yang akan diuji dalam model atau algoritma klasifikasi. Setelah itu, model atau algoritma tersebut akan dievaluasi pada data uji menggunakan metrik yang sesuai. Proses ini kemudian diulang dengan variasi parameter yang berbeda untuk memperoleh kombinasi parameter yang optimal. Eksperimen dengan berbagai parameter dapat membantu meningkatkan kinerja model atau algoritma machine learning dalam memprediksi atau mengklasifikasikan data.

3.4.9 Analisis & Evaluasi

Analisis dan evaluasi adalah tahap akhir dari proses pengembangan model atau algoritma machine learning. Pada tahap ini, hasil dari model atau algoritma akan dievaluasi untuk memastikan bahwa kinerjanya optimal dan sesuai dengan tujuan yang diinginkan. Evaluasi model atau algoritma biasanya dilakukan dengan menggunakan metrik yang relevan untuk tugas yang diinginkan, seperti akurasi, presisi, recall, dan F1 score.

Analisis dilakukan dengan memeriksa hasil prediksi model atau algoritma dan menganalisis kesalahan yang terjadi. Dalam analisis, biasanya dilakukan visualisasi data untuk memahami karakteristik data dan hasil prediksi model atau algoritma. Hal ini dapat membantu mengidentifikasi faktor-faktor yang memengaruhi kinerja model atau algoritma, serta menemukan cara untuk meningkatkan kinerja model atau algoritma.

Evaluasi dilakukan dengan menghitung metrik evaluasi yang relevan untuk menentukan kinerja model atau algoritma. Metrik evaluasi yang umum digunakan termasuk akurasi, presisi, recall, dan F1 score. Akurasi mengukur seberapa banyak

prediksi model atau algoritma yang benar, sedangkan presisi mengukur seberapa banyak prediksi positif yang benar dari total prediksi positif. Recall mengukur seberapa banyak data positif yang terdeteksi dari total data positif, sedangkan F1 score adalah rata-rata harmonik dari presisi dan recall. Detil evaluasi yang digunakan pada penelitian ini dapat dilihat pada Bab 2 bagian Evaluasi Hasil Ekstraksi.

3.4.10 Penulisan Laporan Disertasi

Langkah ini perlu dilakukan agar hasil penelitian terdokumentasi. Dokumentasi hasil penelitian ini penting karena dapat dipakai peneliti lain dalam penelitiannya. Di dalamnya juga memuat temuan/kebaharuan (*novelty*) hasil penelitian

4 PENGEMBANGAN METODA BARU

4.1 PEMBUATAN DAFTAR KOSAKATA POPULER

Pembuatan daftar kosakata populer dimulai dengan melakukan pra-proses terhadap metadata hasil pemanenan yang memiliki atribut cukup banyak. Meskipun atributnya cukup banyak, tetapi untuk membuat kosakata populer cukup atribut *Author Keywords* saja. Kata/frasa kunci pilihan penulis artikel tersebut dikumpulkan untuk kemudian menjalani pra-proses.

4.2 DATASET

Dataset dalam penelitian ini ada tiga jenis yaitu: a) dataset untuk membentuk kosakata populer, b) dataset umum yang dipakai untuk menguji metode ekstraksi katakunci, dan c) dataset baru hasil dari penelitian ini.

4.2.1 Dataset pembentuk kosakata populer

Dataset untuk ekstraktor katakunci adalah kosakata populer yang sudah diberi frekwensi kemunculannya. Dari 50 ribu metadata artikel ilmiah yang dipanen dari Scopus dihitung setiap kata/frasa kunci yang dibuat oleh penulis artikel. Misalnya ditemukan 181 frasa: *Adaptive control*, *adaptive control*, *Adaptive controller*, *Adaptive controls*, dan *Adaptive Control*, maka frekwensi untuk token *adapt control* adalah 181. Dari 50 ribu metadata artikel diperoleh 56.727 baris dataset. Karena dataset dibuat pada tahun 2019 dengan data tahun 2018 dan sebelumnya, maka tidak ada katakunci baru yang tercatat di dalam dataset. Di dataset tidak ada katakunci Covid-19, karena istilah tersebut baru muncul tahun 2020. Sebagian dari dataset yang diperoleh dapat dilihat pada Tabel 4.2.

4.2.2 Dataset umum

Maksud dataset umum di sini adalah himpunan artikel beserta katakuncinya yang sudah umum dijadikan sebagai dataset pengujian dari berbagai metode ekstraksi katakunci. Berbagai dataset ini sudah dibahas pada bagian **Error! Reference source not found.** Dari belasan dataset tersebut, ada 3 dataset yang telah dipilih untuk digunakan pada penelitian ini yaitu dataset Krapivin2019 (2.304 item), dataset Semeval (243 item) dan dataset Nguyen (209 item).

Dataset umum ini digunakan dengan tujuan utama untuk membandingkan secara objektif metode ekstraksi katakunci *novelty* dari penelitian ini dengan metode-metode *State-of-the-Art* yang sudah dipaparkan pada bagian **Error! Reference source not found.** Perbandingan metode akan bagus bila menggunakan dataset yang sudah umum dipakai para peneliti lain dalam menghasilkan metode baru mereka. Hal ini karena bisa dipakai untuk membandingkan metode secara *apple-to-apple* dengan dataset yang sama.

4.2.3 Dataset baru

Selain menggunakan dataset umum di atas, penulis juga menggunakan dataset buatan sendiri yang dapat diunduh di alamat situs web <https://github.com/harry-achsan/achsan2022>. Dataset ini terdiri dari metadata artikel ilmiah yang diunduh dari situs web pengindeks dokumen Scopus. Dibuatnya dataset ini karena dataset umum yang ada berisi metadata artikel-artikel yang rata-rata diterbitkan belasan tahun lalu sehingga kurang mewakili artikel-artikel baru seperti artikel tentang Covid19 (*Corona Virus Disease* 2019). Selain itu jumlah fiturnya terbatas/sedikit.

Dataset ini telah didaftarkan identifikasinya ke International DOI Foundation dengan kode DOI: 10.17605/OSF.IO/9GHWM. Digital Object Identifier (DOI) merupakan standar ISO (ISO 26324) untuk sistem DOI yang menyediakan infrastruktur teknis dan sosial untuk pendaftaran dan penggunaan pengidentifikasi interoperabilitas persisten untuk digunakan pada jaringan digital [80]. Dengan adanya kode DOI ini maka semua objek dapat ditelusuri keberadaannya secara digital melalui situs <https://www.doi.org/> dengan memasukkan kodenya.

4.3 TEKNIK EKSTRAKSI KATA/FRASA KUNCI

Algoritma dan metode serta teknik ekstraksi kata/frasa kunci yang dibahas di sini adalah novelty dari penulis. Untuk melihat kinerja metode baru ini, novelty dari penelitian ini, maka dilakukan evaluasi untuk membandingkan kinerjanya dengan 10 metode ekstraksi kata kunci lain yang merupakan state-of-the-art. Adapun 10 metoda lain sebagai pembandingnya adalah: KPMiner, MultipartiteRank, PositionRank, RAKE, SingleRank, TextRank, TF-IDF, TopicalPageRank, TopicRank dan YAKE, yang sudah dibahas pada Bab 2. Meskipun ini novelty dari penulis, tetapi penulis tidak mengklaim bahwa keseluruhannya adalah hasil pemikiran penulis. Penulis hanya mencoba memperbaiki teknik, metoda dan algoritma yang sudah ada, sebagaimana yang dilakukan para peneliti pada umumnya.

4.3.1 Pembobotan

Ada 5 teknik pembobotan yang dilakukan pada kata/frasa kunci. Pembobotan ini tergantung jumlah kata dalam frase atau biasa disebut dengan istilah n-gram, dimana n menyatakan jumlah kata. Untuk setiap jumlah kata ada istilahnya masing-masing, uni-gram untuk 1-gram atau jumlah kata dalam frase adalah satu, bi-gram untuk jumlah kata 2, tri-gram untuk 3 kata, dan four-gram untuk 4 kata. Semakin banyak jumlah kata dalam suatu frase kunci menandakan bahwa frase kunci tersebut semakin spesifik, misalnya *network* akan kalah spesifik dengan *sensor network*, tetapi *sensor network* yang berupa bi-gram akan kalah spesifik dengan *wireless sensor network* yang tri-gram. Semakin spesifik suatu frase kunci maka bobotnya akan semakin tinggi. Pembobotan untuk *wireless sensor network* akan lebih tinggi dibandingkan dengan *sensor network*, dan pembobotan untuk *sensor network* lebih tinggi daripada *network*.

4.3.2 Frasa dalam satu kata

Pembobotan tersebut terlihat sederhana, tetapi ternyata ada kendala yang muncul. Kendala tersebut adalah ketidak mampuan untuk mendeteksi suatu frase yang tergabung dalam satu kata. Contohnya untuk itu adalah *genet* yang merupakan token untuk *genetic* (yang berarti asal-usul) akan mempunyai bobot yang sama dengan *epigenet* yang merupakan token dari *epigenetic* dan juga berbobot sama dengan *phyloenet* yang merupakan token dari *phylogenetic*, padahal *epigenetic* merupakan

paduan kata *epi* dari bahasa Yunani kuno atau Latin yang berarti “di atas” dan kata *genetic*, begitu juga dengan *phylogenetic* yang merupakan paduan dari *phylo* (Bahasa Latin yang berarti ras, suku atau marga) dan *genetic*.

Untuk mengetahui bahwa *phylogenetic* dan *epigenetic* itu adalah kata yang lebih spesifik dari *genetic* adalah mudah bagi ahli biologi tetapi tidak mudah bagi komputer. Misalnya untuk kata *phylogenetic*, komputer bisa mengira itu gabungan *phy* (bahasa Latin yang berarti ekspresi menjijikkan) + *lo* (bahasa Latin yang merupakan mantra untuk menyembuhkan dari gigitan anjing gila) + *genetic* (kata *genetic* tidak ada dalam kamus bahasa Latin, yang ada sebenarnya adalah *genetica*, *geneticae*, *geneticus*, *geneticum* dimana mempunyai definisi yang sama dengan *genetic* dalam Bahasa Inggris), tetapi bisa pula dianggap gabungan dari *phylo* + *genetic*. Gabungan-gabungan ini perlu suatu penelitian mendalam secara terpisah karena seperti *phylogenetic* itu ternyata gabungan antara *phylo* yang merupakan kata dalam Bahasa Latin dan *genetic* yang merupakan kata dalam Bahasa Inggris.

4.3.3 Jumlah katakunci

Katakunci dibutuhkan oleh mesin pencari agar suatu dokumen/artikel/objek mudah ditemukan. Pemilihan katakunci penting karena menentukan visibilitas artikel yang dapat meningkatkan sitasi atau dampak dari artikel tersebut. Semakin banyak jumlah katakunci akan memperbesar kemungkinan artikel ditemukan oleh mesin pencari. Namun terlalu banyak katakunci akan mengurangi relevansi katakunci tersebut dengan topik artikelnya.

Jumlah katakunci yang akan diekstrak secara otomatis dari setiap artikel adalah 20. Penentuan jumlah ini jauh lebih banyak dari kisaran maksimum jumlah katakunci yang diminta penerbit jurnal ilmiah. Beberapa penerbit besar dunia menetapkan jumlah katakunci untuk artikel berbeda-beda: Elsevier maksimum 5, Taylor and Francis hanya 5 atau 6 saja, Frontiers minimum 5 maksimum 8, dan Wiley minimal 5 katakunci. Sedangkan RistekDikti menentukan maksimal 5 katakunci dalam pengajuan proposal hibah penelitian. Penentuan jumlah katakunci yang diekstrak secara otomatis lebih didasarkan pada metode evaluasi katakunci yang umum dipakai para peneliti sesuai **Error! Reference source not found.**, dimana evaluasinya menggunakan MAP@20 yang mengharuskan ada 20 katakunci.

4.4 EKSPERIMEN

Percobaan dilakukan berulang kali dengan formula berbeda-beda dan pada banyak artikel. Pada tahap penentuan formula, evaluasi hasil setiap percobaan dilakukan sendiri oleh penulis. Penentuan formula dilakukan secara empiris, dimana formula akan diurutkan, yang hasil ekstraksi katakuncinya menurut penulis paling bagus akan ditempatkan paling atas, kemudian diikuti dengan formula yang hasilnya tidak lebih bagus dari formula di atasnya.

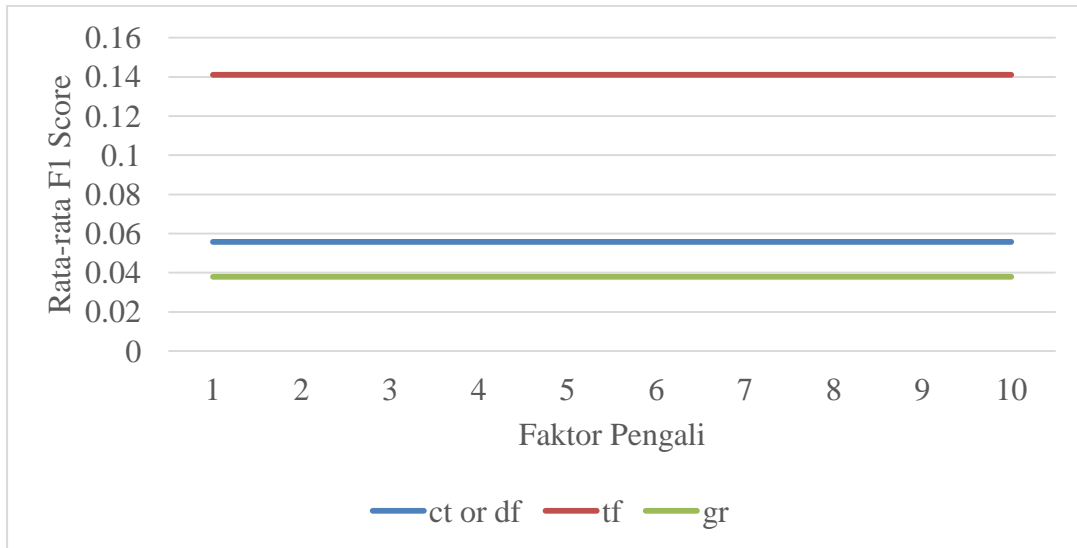
4.4.1 Perangkat yang Digunakan

Piranti keras dalam penelitian ini adalah komputer jenis laptop dengan prosesor Intel i7 berkecepatan 4,5 GHz dan dengan jumlah core 6 serta memori terpasang 24 GB. Agar proses input/output dapat berjalan cepat maka dua buah penyimpanannya dipilih yang berjenis SSD (*Solid State Drive*). Laptop juga dilengkapi dengan dua buah GPU (*Graphics Processing Unit*) yaitu Intel UHD Graphics dan Nvidia GeForce GTX 1650 untuk mempercepat proses yang dilakukan secara parallel, dimana Nvidia GeForce GTX 1650 memiliki jumlah Cuda Core 896.

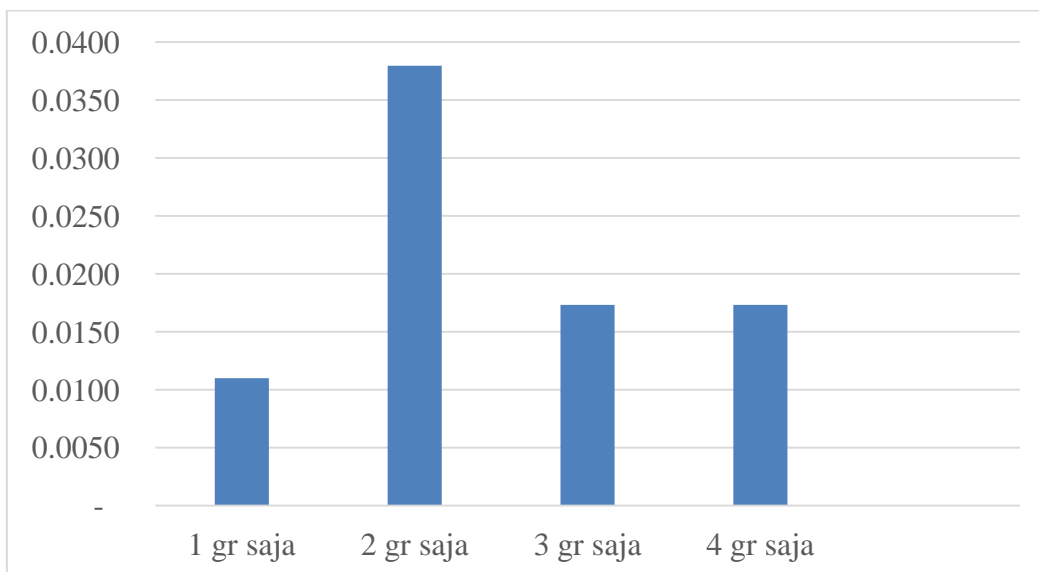
4.4.2 Dampak tiap peubah

Ada 4 peubah dalam penelitian ini yang berpengaruh pada hasil dari penerapan metode ekstraksi katakunci. Ke-empat peubah tersebut adalah: Jumlah sitasi (ct), Jumlah dokumen (df), Jumlah gram katakunci (gr) dan Jumlah kemunculan katakunci dalam dokumen (tf). Data 3 buah peubah pertama (ct, df dan gr) diambil dari dataset kosakata populer, sedangkan tf diambil dari korpus/dataset yang berisi metadata artikel ilmiah sebagaimana telah dibahas pada bagian 4.2.2 dan 4.2.3.

Untuk mengetahui dampak tiap peubah maka dilakukan eksperimen dengan satu peubah yang diisolasi dari peubah-peubah lainnya. Peubah tersebut kemudian dilihat kinerjanya ketika dipakai dalam metode baru untuk ekstraksi katakunci. Adapun dataset yang dipakai untuk ini adalah Nguyen.



Gambar 4.1. Dampak tiap peubah terhadap F1 Score hasil ekstraksi katakunci dengan menggunakan metode baru.



Gambar 4.2. Pengaruh/dampak dari jumlah kata (gram) dalam katakunci terhadap F1 Score.

Peubah-peubah terisolasi tersebut kemudian dikalikan dengan suatu konstanta atau factor pengali mulai dari 1 hingga 10. Hasil eksperimen terhadap peubah-peubah terisolasi tersebut dapat dilihat pada Gambar 4.1. Pada tersebut terlihat bahwa peubah ct dan df mempunyai dampak yang sama nilainya terhadap F1 Score yaitu 0.0557. Dampak peubah gr adalah paling kecil, sedangkan terbesar adalah dampak dari tf yaitu 0.1411.

Gambar 4.2 memperlihatkan dampak jumlah kata (gram) dalam katakunci terhadap nilai F1 Score hasil evaluasi ekstraksi katakunci menggunakan metode baru. Terlihat bahwa katakunci mono-gram (terdiri dari satu kata saja) ternyata mempunyai dampak paling kecil yang artinya bahwa dalam korpus Nguyen hanya sedikit artikel yang menggunakan katakunci monogram. Hal ini bagus karena katakunci monogram terlalu general/umum sehingga kurang mampu memberikan informasi yang lebih spesifik.

Dari gambar ini juga nampak bahwa bi-gram adalah katakunci yang paling banyak dipakai oleh para penulis, sedangkan tri-gram maupun four-gram meskipun lebih spesifik tetapi tidak sepopuler bi-gram. Di sini nampak bahwa model Distribusi Luhn (*Luhn Distribution*) berlaku, dimana mono-gram yang sangat umum dan tri-gram/four-gram yang sangat spesifik lemah dalam penentuan topik sedangkan bi-gram sangat kuat dalam menentukan topik. Penjelasan tentang Distribusi Luhn dapat dilihat pada sub-bab **Error! Reference source not found.**

4.4.3 Formula

Dari berbagai eksperimen terhadap peubah terisolasi tersebut di atas diperoleh diketahui bahwa tf mempunyai pengaruh paling besar terhadap F1 Score. Sedangkan gram dari katakunci paling populer adalah bi-gram. Dari temuan-temuan tersebut maka dibuatlah beberapa formula:

- a) Bobot = c

Pembobotan dilakukan dengan suatu konstanta, dimana uni-gram bobotnya 250, bi-gram bobotnya 1000, tri-gram bobotnya 2000 dan four-gram bobotnya 4000. Adapun persamaan untuk formula ini adalah:

$$B(x_i) = \sum_{j=1}^N (c_j) \quad (4-1)$$

dimana B adalah bobot dari kata/frasa kunci x_i , sedangkan N adalah jumlah atau frekwensi ditemukannya x_i dalam metadata artikel dan c_j adalah konstanta sesuai jumlah gram dari x_i .

Contoh: dalam sebuah metadata ditemukan 21 kata *network*, 5 frasa *sensor network*, dan 4 frasa *wireless sensor network*. Jika ditemukan tiga frasa

kunci tersebut maka perhitungan bobot untuk setiap frasa adalah sebagai berikut:

$$B(network) = \sum_{j=1}^{21}(250) = 5250 \quad (4-2)$$

$$B(sensor\ network) = \sum_{j=1}^5(1000) = 5000 \quad (4-3)$$

$$B(wireless\ sensor\ network) = \sum_{j=1}^4(2000) = 8000 \quad (4-4)$$

Dari pembobotan tersebut maka urutan akan dijadikan sebagai hasil dari metode baru adalah: *wireless sensor network*, diikuti *network*, dan *sensor network*. Frasa kunci tersebut akan tergeser apabila ternyata dalam metadata artikel ditemukan frasa kunci lain yang bobotnya lebih tinggi.

b) Bobot = f + c

Simbol f di atas adalah frekwensi dari suatu katakunci. Formula terbaik yang ditemukan saat percobaan adalah f untuk uni-gram, f+100 untuk bi-gram, f+200 untuk tri-gram dan f+400 untuk four-gram. Apabila dalam proses pencarian katakunci ditemukan frasa *neural network*, *genetic algorithm*, dan *wireless sensor network*, maka berdasarkan **Error! Reference source not found.** diperoleh bobot untuk masing-masing frasa tersebut adalah: *neural network* 575 (475+100), *genetic algorithm* 420 (320+100), dan *wireless sensor network* 520 (320+200). Disini terlihat bahwa *wireless sensor network* yang tri-gram dapat dikalahkan oleh *neural network* yang bi-gram karena frekwensi kemunculan *neural network* dalam artikel-artikel yang diunduh dari Scopus lebih tinggi daripada *wireless sensor network* yang tri-gram dengan selisih di atas 100.

Dengan f_i sebagai frekwensi frasa x_i di dalam dataset, maka persamaan untuk formula ini adalah:

$$B(x_i) = \sum_{j=1}^N(f_i + c_i) \quad (4-5)$$

c) Bobot = cf

Formula pembobotan ini adalah frekwensi dikalikan suatu konstanta. Dalam berbagai percobaan ditemukan bahwa ada dua varian pembobotan yang cukup bagus hasilnya yaitu:

1. uni-gram = f, bi-gram = 10f, tri-gram = 20f, dan four-gram = 40f

2. uni-gram = 10f, bi-gram = 20f, tri-gram = 40f, dan four-gram = 80f

Persamaan untuk formula ini adalah:

$$B(x_i) = \sum_{j=1}^N (c_j f_j) \quad (4-6)$$

d) Frasa Judul ditambah f+c

Frasa judul diperoleh dengan cara memotong judul menggunakan stop-words. Misalkan diperoleh suatu artikel berjudul “*Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set*” dengan stop-words yang bergaris bawah, maka akan hasil prediksi frasa kuncinya adalah: *Efficiency*, *ab-initio total energy calculations*, *metals*, *semiconductors*, dan *plane-wave basis set*. Jika luaran dari ekstraktor katakunci dibatasi hanya lima kata/frasa kunci maka frasa judul tersebut sudah memenuhi. Tetapi jika judul artikelnya adalah “*Deep residual learning for image recognition*” maka didapatkan frasa kunci: *Deep residual learning* dan *image recognition*. Karena jumlah frasa kunci yang diperoleh baru dua maka yang tiga lagi dicari menggunakan formula f+c (sudah dibahas di formula b di atas).

4.5 EVALUASI

Evaluasi secara kuantitatif terhadap klasifikasi multi-label diperlukan untuk menilai metode baru. Metode evaluasi yang digunakan adalah F1@5, F1@10, F1@20, MAP@5, MAP@10, MAP@20 dan MRR. Penjelasan tentang metode evaluasi tersebut sudah dibahas pada bagian 3.5.5.

4.5.1 Evaluasi metode baru

Formula yang dibentuk pada penelitian ini harus dievaluasi terlebih dahulu. Hasil eksperimen yang telah dilakukan terhadap 3 korpus Nguyen, Krapivin2019 dan Semeval dapat dilihat peta panasnya (*Heat Map*) pada **Error! Reference source not found.** Terlihat bahwa kinerja dari setiap formula berbeda untuk setiap korpus. Formula cv4 selalu menjadi yang terbaik sedangkan cv2 adalah yang terburuk.

Eksperimen ini dilakukan berulang-ulang dengan mengubah bobot dari tiap peubah sehingga menghasilkan formula terbaik.

Untuk membuktikan bahwa formula yang dibentuk sudah bagus maka harus dibandingkan dengan metode lain. **Error! Reference source not found.** memperlihatkan peta panas hasil perbandingan formula-formula metode baru dengan metode ekstraksi katakunci RAKE, TF-IDF, TextRank dan YAKE. Terlihat bahwa RAKE adalah metoda ekstraksi katakunci tercepat tetapi sekaligus terburuk. Jika menghilangkan RAKE maka metode baru dari penulis adalah yang paling cepat karena untuk menjalankan 5 formula hanya dibutuhkan waktu 37 detik saja. Formula cv4 dan cv1 dari metode baru terlihat mendominasi warna hijau yang mengindikasikan bahwa formula-formula tersebut kinerjanya lebih bagus dari metoda lainnya.

4.5.2 Komparasi dengan 10 metode state-of-the-art

Jika pada evaluasi metode baru hanya dibandingkan dengan 4 metoda lain maka di sini akan dibandingkan dengan sepuluhg metoda lainnya. Adapun 10 metoda lain sebagai pembandingnya adalah: KPMiner, MultipartiteRank, PositionRank, RAKE, SingleRank, TextRank, TF-IDF, TopicalPageRank, TopicRank dan YAKE, yang sudah dibahas pada Bab 2. Adapun korpus yang dipakai adalah Krapivin2019, Nguyen, Semeval dan dataset baru.

No.	Metode Ekstraksi Katakunci	Metode Evaluasi							Waktu Proses (detik)
		F1@5	F1@10	F1@20	MAP@5	MAP@10	MAP@20	MRR	
1	cv1	0.1274	0.1368	0.1152	0.1317	0.1049	0.0729	0.2283	610
2	cv2	0.0656	0.0753	0.0765	0.0678	0.0578	0.0484	0.1550	610
3	cv3	0.1630	0.1531	0.1230	0.1685	0.1174	0.0779	0.2978	610
4	cv4	0.1679	0.1498	0.1116	0.1736	0.1148	0.0707	0.3170	610
5	cv5	0.1693	0.1523	0.1136	0.1750	0.1168	0.0719	0.3187	610
6	KPMiner	0.1587	0.1546	0.1127	0.1640	0.1186	0.0714	0.3004	6,143
7	MultipartiteRank	0.1551	0.1583	0.1322	0.1603	0.1214	0.0837	0.2705	6,197
8	PositionRank	0.1033	0.1255	0.1208	0.1068	0.0963	0.0765	0.1948	4,617
9	RAKE	0.0007	0.0015	0.0027	0.0007	0.0011	0.0017	0.0032	135
10	SingleRank	0.0366	0.0544	0.0653	0.0379	0.0417	0.0414	0.0819	4,373
11	TextRank	0.0737	0.0873	0.0759	0.0762	0.0670	0.0481	0.1087	2,720
12	TF-IDF	0.0751	0.0880	0.0869	0.0776	0.0674	0.0550	0.1520	5,350
13	TopicalPageRank	0.0393	0.0559	0.0671	0.0406	0.0429	0.0425	0.0856	6,592
14	TopicRank	0.0906	0.0936	0.0805	0.0936	0.0717	0.0509	0.1855	10,610
15	YAKE	0.0441	0.0650	0.1162	0.0456	0.0498	0.0736	0.1070	1,678

(a) Dataset Krapivin2019

No.	Metode Ekstraksi Katakunci	Metode Evaluasi							Waktu Proses (detik)
		F1@5	F1@10	F1@20	MAP@5	MAP@10	MAP@20	MRR	
1	cv1	0.0988	0.1309	0.1388	0.2032	0.1674	0.1234	0.2430	50
2	cv2	0.0499	0.0669	0.0816	0.1028	0.0855	0.0726	0.1975	50
3	cv3	0.1311	0.1596	0.1566	0.2699	0.2041	0.1393	0.3535	50
4	cv4	0.1215	0.1332	0.1272	0.2502	0.1703	0.1131	0.3755	50
5	cv5	0.1244	0.1309	0.1263	0.2559	0.1674	0.1123	0.3861	50
6	KPMiner	0.1427	0.1863	0.1802	0.2938	0.2382	0.1602	0.2891	562
7	MultipartiteRank	0.1444	0.1972	0.2135	0.2971	0.2522	0.1899	0.2618	573
8	PositionRank	0.0927	0.1361	0.1626	0.1909	0.1740	0.1446	0.3108	388
9	RAKE	0.0044	0.0054	0.0060	0.0090	0.0069	0.0053	0.0113	11
10	SingleRank	0.0484	0.0820	0.1108	0.0995	0.1049	0.0985	0.1406	378
11	TextRank	0.1036	0.1412	0.1515	0.2131	0.1806	0.1347	0.2206	269
12	TF-IDF	0.0920	0.1267	0.1510	0.1893	0.1621	0.1343	0.2270	476
13	TopicalPageRank	0.0492	0.0807	0.1135	0.1012	0.1032	0.1010	0.1492	658
14	TopicRank	0.1120	0.1448	0.1517	0.2304	0.1851	0.1349	0.2715	1465
15	YAKE	0.0548	0.1042	0.2058	0.1127	0.1333	0.1831	0.1167	151

(b) Dataset Semeval

No.	Metode Ekstraksi Katakunci	Metode Evaluasi							Waktu Proses (detik)
		F1@5	F1@10	F1@20	MAP@5	MAP@10	MAP@20	MRR	
1	cv1	0.1745	0.1975	0.1896	0.2966	0.2172	0.1516	0.2765	33
2	cv2	0.0782	0.1052	0.1172	0.1330	0.1157	0.0937	0.1903	33
3	cv3	0.2072	0.2318	0.2069	0.3521	0.2550	0.1655	0.3515	33
4	cv4	0.1739	0.1796	0.1480	0.2956	0.1976	0.1184	0.3824	33
5	cv5	0.1751	0.1783	0.1498	0.2976	0.1961	0.1198	0.3970	33
6	KPMiner	0.2432	0.2867	0.2649	0.4133	0.3153	0.2119	0.3559	418
7	MultipartiteRank	0.2393	0.2871	0.2949	0.4066	0.3157	0.2358	0.3382	421
8	PositionRank	0.1587	0.2057	0.2279	0.2698	0.2263	0.1822	0.2764	305
9	RAKE	0.0011	0.0043	0.0035	0.0019	0.0047	0.0028	0.0119	7
10	SingleRank	0.0929	0.1400	0.1812	0.1578	0.1540	0.1449	0.1789	292
11	TextRank	0.1413	0.1888	0.1887	0.2401	0.2076	0.1509	0.1984	148
12	TF-IDF	0.1509	0.2031	0.2165	0.2564	0.2234	0.1732	0.2473	371
13	TopicalPageRank	0.0906	0.1431	0.1854	0.1540	0.1574	0.1483	0.1922	538
14	TopicRank	0.1475	0.1862	0.1752	0.2507	0.2047	0.1401	0.2907	830
15	YAKE	0.0872	0.1274	0.2640	0.1483	0.1401	0.2112	0.1236	104

(c) Dataset Nguyen

Gambar 4.3. Hasil eksperimen berbagai metode ekstraksi katakunci pada 3 dataset berbeda. Angka dengan arsiran hijau menyatakan tiga peringkat terbaik, sedangkan yang tercetak tebal adalah yang paling bagus.

No.	Metode Ekstraksi Katakunci	Metode Evaluasi							Waktu Proses (detik)
		F1@5	F1@10	F1@20	MAP@5	MAP@10	MAP@20	MRR	
1	cv1	0.1630	0.1439	0.1070	0.1590	0.1062	0.0662	0.2994	283
2	cv2	0.1432	0.1297	0.0988	0.1397	0.0957	0.0611	0.2766	283
3	cv3	0.1469	0.1383	0.1107	0.1433	0.1020	0.0685	0.3010	283
4	cv4	0.1449	0.1371	0.1105	0.1413	0.1011	0.0683	0.3001	283
5	cv5	0.1633	0.1439	0.1126	0.1593	0.1062	0.0697	0.3226	283
6	KPMiner	0.0706	0.0494	0.0296	0.0689	0.0364	0.0183	0.2092	1386
7	MultipartiteRank	0.1486	0.1209	0.0770	0.1449	0.0892	0.0476	0.3217	1384
8	PositionRank	0.1628	0.1668	0.1316	0.1588	0.1231	0.0814	0.2730	1300
9	RAKE	0.0393	0.0730	0.1042	0.0383	0.0538	0.0644	0.0800	17
10	SingleRank	0.1178	0.1398	0.1240	0.1149	0.1031	0.0767	0.1817	1310
11	TextRank	0.0710	0.0637	0.0420	0.0693	0.0470	0.0260	0.1423	38
12	TF-IDF	0.1587	0.1798	0.1682	0.1549	0.1326	0.1041	0.2372	1359
13	TopicalPageRank	0.1201	0.1391	0.1230	0.1172	0.1026	0.0761	0.1875	3209
14	TopicRank	0.1262	0.1177	0.0871	0.1231	0.0868	0.0539	0.2531	1318
15	YAKE	0.0535	0.0781	0.1273	0.0522	0.0576	0.0787	0.1067	316

Gambar 4.4. Hasil eksperimen dengan menggunakan dataset hasil penelitian ini dengan keseluruhan data sebanyak 3.309 item metadata artikel ilmiah bidang Ilmu Komputer. Terlihat metode *novelty* dari penelitian ini (cv5) adalah yang terbaik untuk evaluasi dengan metode F1@5, MAP@5 dan MRR.

Hasil evaluasi tersebut dapat dilihat pada Gambar 4.9 dan Gambar 4.10. Dari Gambar 4.11 dapat dilihat kenyataan bahwa metode MultipartiteRank mendominasi penilaian dengan metode F1 Score dan MAP. Formula-formula metode baru meskipun tidak menempati urutan teratas dalam F1 Score maupun MAP tetapi banyak yang masuk dalam tiga besar, ditandai dengan warna hijau pada gambar tersebut. Namun dari sisi kualitas (MRR) cv5 yang merupakan salah satu formula dari metode baru selalu menjadi yang terbaik pada semua dataset tersebut.

Formula cv5 juga kualitasnya paling bagus pada dataset baru hasil penelitian ini, selain itu juga paling bagus nilainya pada F1@5 dan MAP@5. Pada dataset ini MultipartiteRank yang dominan pada Gambar 4.12 ternyata kinerjanya rendah, hanya MRR-nya saja yang masuk tiga teratas namun bukan teratas. Justru TF-IDF yang merupakan salah satu metoda tertua dalam ML menampakkan hasil bagus dengan menjadi teratas pada F1@10, F1@20, MAP@10 dan MAP@20. Karena jarang sekali penulis artikel membuat katakunci dengan jumlah 10 atau bahkan 20, maka metode ini kurang bagus untuk system rekomendasi ekstraksi katakunci.

Dari itu maka formula cv5 menjadi yang terbaik karena selain menempati peringkat tertinggi pada F1@5 dan MAP@5 juga kualitasnya paling bagus dilihat dari peringkat MRR-nya. Selain itu, diantara metoda-metoda yang menempati peringkat 3

besar maka metode baru adalah yang paling cepat, 10 kali lebih cepat dari MultipartiteRank.

4.6 PROOF OF CONCEPTS

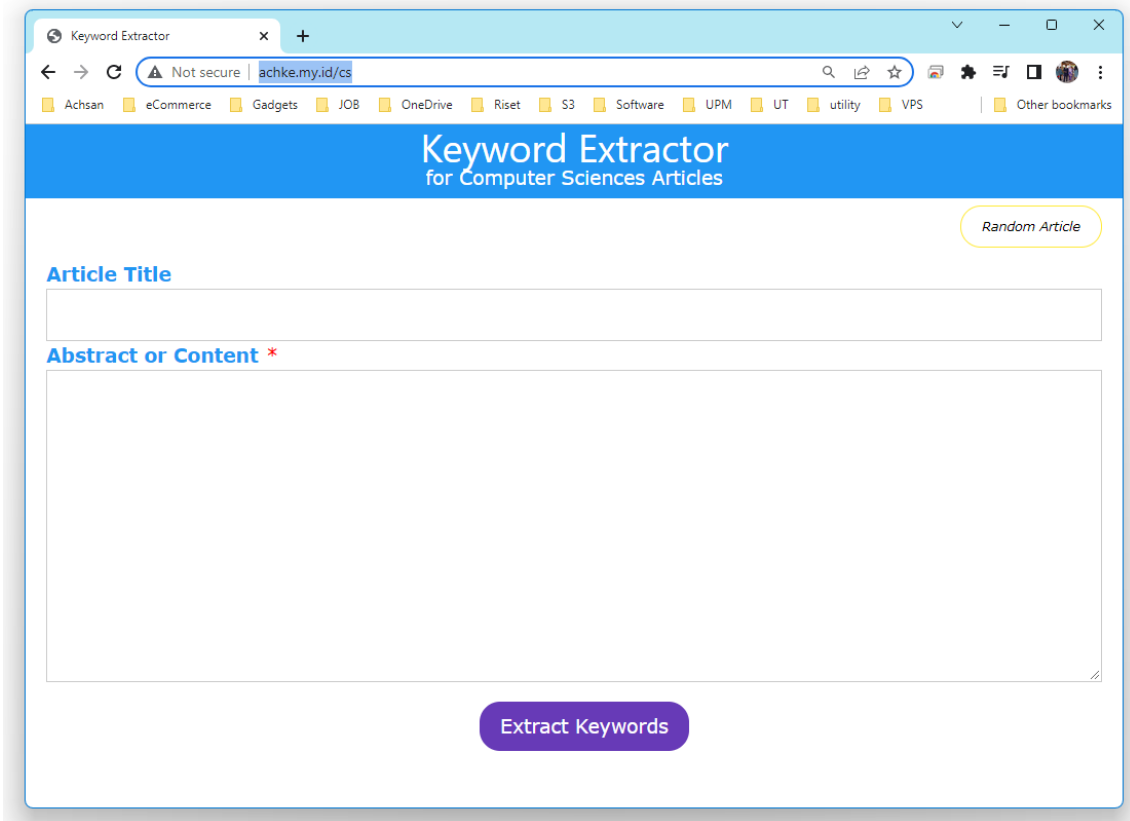
Untuk membuktikan kemampuan dan kinerja dari metode baru maka dibuatlah piranti lunak berbasis yang bisa diakses oleh siapa saja pada alamat web <http://extract.my.id/>. Siapapun dapat memanfaatkan piranti lunak ekstraksi katakunci tersebut tanpa harus mendaftar. Piranti lunak ini dapat dikatakan sebagai sebuah sistem rekomendasi.

4.6.1 Metoda

Metoda yang dipakai untuk pengembangan piranti lunak *proof of concepts* yang dapat dipakai sebagai sistem rekomendasi ini adalah dengan menerapkan metoda baru hasil penelitian ini sebagaimana telah dijelaskan pada bagian-bagian sebelumnya dalam bab ini. Ada 5 parameter yang ditampilkan pada aplikasi ini karena tiap parameter memiliki keunggulannya tersendiri.

Adapun metoda perbandingan yang berupa 10 metoda terbaru atau *state of the art* juga diterapkan pada aplikasi ini. Penerapan metoda-metoda ini relatif mudah karena pustakanya dapat ditelusuri di Internet. Di Internet juga tersedia kode sumber dari berbagai metoda tersebut. Kode sumber ini kemudian digabungkan dengan kode sumber hasil penelitian ini.

Karena banyak penelitian bidang pembelajaran mesin menggunakan Bahasa pemrograman Python maka aplikasi ini juga dibuat dengan Bahasa pemrograman Python versi 3.x. Perlu diketahui bahwa Python versi 3.x (x menyatakan sub-versinya) jauh berbeda dengan versi 2.x sehingga aplikasi ini tidak bisa dijalankan pada Python versi 2.x. Namun jika aplikasi dibuat menggunakan Python 2.x, maka dia juga tidak dapat dijalankan pada lingkungan Python 3.x. Bahasa pemrograman Python juga mendukung pembuatan aplikasi berbasis web sehingga antarmuka penggunaannya bisa langsung dibuat dalam Python tanpa harus menggunakan Bahasa pemrograman lain seperti PHP.



Gambar 4.5. Antarmuka pengguna dari piranti lunak pengestraksi katakunci.

Adapun dataset yang disertakan pada aplikasi ini adalah dataset hasil dari penelitian ini. Dataset ini menggunakan metadata artikel ilmiah terbaru dan jumlahnya paling banyak diantara semua dataset ekstraksi katakunci yang ada saat ini. Dataset ini juga paling lengkap fiturnya.

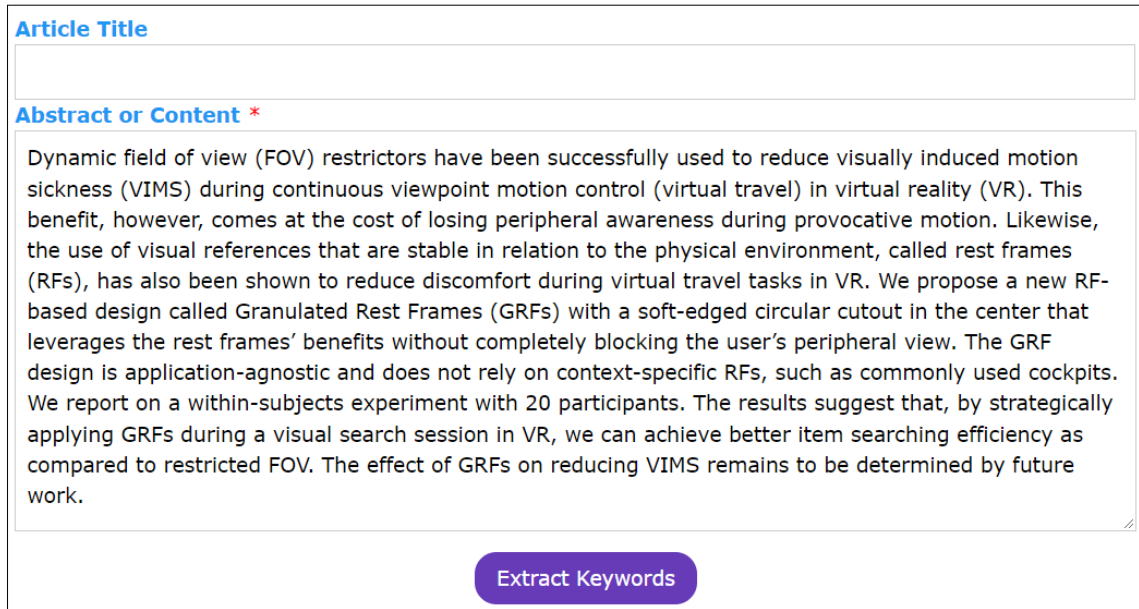
4.6.2 Antarmuka Pengguna

Antarmuka pengguna dibuat sesederhana mungkin, karena semakin sederhana maka biasanya akan semakin mempermudah pengguna. Terlihat pada Gambar 4.5 bahwa antarmuka pengguna hanya berupa satu formulir saja. Formulir tersebut hanya berisi dua kotak teks dan dua tombol.

4.6.3 Cara Penggunaan

Penggunaan aplikasi ini relatif mudah. Pengguna hanya harus memasukkan (bisa dengan copy-paste) abstrak atau keseluruhan isi dari sebuah artikel ilmiah bidang Ilmu Komputer, lebih bagus jika judul artikel juga diisi. Kemudian diikuti dengan klik

tombol “Extract Keywords”. Setelah itu tunggu beberapa detik maka akan muncul 15 deretan pilihan katakunci hasil ekstraksi dari berbagai metode.



Article Title

Abstract or Content *

Dynamic field of view (FOV) restrictors have been successfully used to reduce visually induced motion sickness (VIMS) during continuous viewpoint motion control (virtual travel) in virtual reality (VR). This benefit, however, comes at the cost of losing peripheral awareness during provocative motion. Likewise, the use of visual references that are stable in relation to the physical environment, called rest frames (RFs), has also been shown to reduce discomfort during virtual travel tasks in VR. We propose a new RF-based design called Granulated Rest Frames (GRFs) with a soft-edged circular cutout in the center that leverages the rest frames' benefits without completely blocking the user's peripheral view. The GRF design is application-agnostic and does not rely on context-specific RFs, such as commonly used cockpits. We report on a within-subjects experiment with 20 participants. The results suggest that, by strategically applying GRFs during a visual search session in VR, we can achieve better item searching efficiency as compared to restricted FOV. The effect of GRFs on reducing VIMS remains to be determined by future work.

Extract Keywords

Gambar 4.6. Contoh pengisian abstrak pada aplikasi.

Sebagai contoh pengisian abstrak pada aplikasi dapat dilihat pada Gambar 4.6 di atas. Setelah terisi, pengguna tinggal klik satu-satunya tombol di bawahnya. Hasilnya dapat dilihat pada Gambar 4.7. Terlihat hasilnya berupa 15 himpunan katakunci, dimana nomor 1-5 adalah formula dari metode baru, sedangkan 10 lainnya adalah hasil ekstraksi katakunci dari berbagai metode.

Terlihat bahwa KPMiner tidak dapat memberikan rekomendasi katakunci. Nampak pula bahwa formula CV5 ada tanda koma pada posisi paling depan yang menandakan bahwa pengguna tidak memasukkan judul artikel. Sebagaimana dibahas sebelumnya bahwa cv5 menggunakan judul sebagai salah satu sumber utama dari katakuncinya.

The Results:		
No.	Method	Extracted Keywords
1	ACHKE CV1	Visual search, Virtual reality, Subjective experiments, Search sessions, peripheral awareness
2	ACHKE CV2	Design, Visualization, Virtualization, Virtual reality, Search
3	ACHKE CV3	Visualization, Virtualization, REST, REDUCE, Motion
4	ACHKE CV4	Visual search, Virtual reality, Subjective experiments, Search sessions, peripheral awareness
5	ACHKE CV5	, Visual search, Virtual reality, Subjective experiments, Search sessions
6	KPMiner	
7	MultipartiteRank	grfs, rest frames, virtual travel, view, fov
8	PositionRank	dynamic field, peripheral view, virtual travel tasks, granulated rest frames, motion sickness
9	RAKE	rest frames ' benefits without completely blocking, based design called granulated rest frames, achieve better item searching efficiency, reduce visually induced motion sickness, continuous viewpoint motion control
10	SingleRank	continuous viewpoint motion control, granulated rest frames, rest frames, virtual travel tasks, new rf-based design
11	TF-IDF	motion, rest, rest frames, frames, grfs
12	TextRank	dynamic field, future, work, peripheral, fov
13	TopicRank	grfs, rest frames, virtual travel, view, motion sickness
14	TopicalPageRank	continuous viewpoint motion control, granulated rest frames, virtual travel tasks, visual search session, rest frames
15	YAKE	virtual travel tasks, Dynamic, reduce visually, called rest frames, virtual travel

Gambar 4.7. Hasil ekstraksi katakunci.

Article Title
Granulated Rest Frames Outperform Field of View Restrictors on Visual Search Performance
Abstract or Content *
Dynamic field of view (FOV) restrictors have been successfully used to reduce visually induced motion sickness (VIMS) during continuous viewpoint motion control (virtual travel) in virtual reality (VR). This benefit, however, comes at the cost of losing peripheral awareness during provocative motion. Likewise, the use of visual references that are stable in relation to the physical environment, called rest frames (RFs), has also been shown to reduce discomfort during virtual travel tasks in VR. We propose a new RF-based design called Granulated Rest Frames (GRFs) with a soft-edged circular cutout in the center that leverages the rest frames' benefits without completely blocking the user's peripheral view. The GRF design is application-agnostic and does not rely on context-specific RFs, such as commonly used cockpits. We report on a within-subjects experiment with 20 participants. The results suggest that, by strategically applying GRFs during a visual search session in VR, we can achieve better item searching efficiency as compared to restricted FOV. The effect of GRFs on reducing VIMS remains to be determined by future work.
Article's Attributes
[YEAR] 2021 [SUBJECT] LCC: QA75.5-76.95 - Electronic computers. Computer science [URL] https://www.frontiersin.org/articles/10.3389/frvir.2021.604889/full [JOURNAL] Frontiers in Virtual Reality vol. 2 [FILENAME] achsan2021/docsutf8/681.txt
Author Keywords
human performance, visual search, rest frames, virtual reality, HCI

Gambar 4.8. Contoh salah satu metadata artikel hasil penelitian ini.

The Results:

No.	Method	Extracted Keywords	Accuration
1	ACHKE CV1	Visual search, Field of view, Virtual reality, Subjective experiments, Search sessions	0.4000
2	ACHKE CV2	Visualization, Design, Virtualization, Performance, Search	0.0000
3	ACHKE CV3	Visual search, Visualization, REST, Field of view, Virtualization	0.2000
4	ACHKE CV4	Visual search, Field of view, Virtual reality, Subjective experiments, Search sessions	0.4000
5	ACHKE CV5	Granulated Rest Frames Outperform Field, View Restrictors, Visual Search Performance, Visual search, Field of view	0.2000
6	KPMiner		0.0000
7	MultipartiteRank	granulated rest frames outperform field, view restrictors, visual search performance, view, grfs	0.0000
8	PositionRank	granulated rest frames, rest frames, visual search performance, view restrictors, visual search session	0.2000
9	RAKE	rest frames ' benefits without completely blocking, based design called granulated rest frames, achieve better item searching efficiency, granulated rest frames outperform field, reduce visually induced motion sickness	0.0000
10	SingleRank	granulated rest frames outperform field, granulated rest frames, continuous viewpoint motion control, rest frames, visual search performance	0.2000
11	TF-IDF	rest, rest frames, frames, motion, grfs	0.2000
12	TextRank	motion, search, searching, frames, visual	0.0000
13	TopicRank	granulated rest frames outperform field, grfs, virtual travel, view, motion sickness	0.0000
14	TopicalPageRank	granulated rest frames outperform field, granulated rest frames, visual search session, visual search performance, rest frames	0.2000
15	YAKE	Field, View, Visual Search, Dynamic field, Outperform	0.2000

Gambar 4.9. Hasil ekstraksi katakunci beserta akurasinya.

4.6.4 Akurasi

Aplikasi ini juga dilengkapi dengan dataset baru hasil penelitian ini. Dataset buatan penulis ini terdiri dari 3.233 metadata artikel ilmiah bidang ilmu komputer dalam Bahasa Inggris. Tujuan adanya dataset ini adalah untuk melihat akurasi hasil ekstraksi katakunci.

Untuk itu pengguna hanya diminta klik tombol “Random Article” sehingga muncul metadata artikel yang dipilih secara acak dari dataset. Contoh hasilnya dapat dilihat pada Gambar 4.8. Pada gambar tersebut terlihat beberapa fitur dalam metadata artikel tersebut yaitu: Judul, Abstrak, Tahun Terbit, LCC, URL dari artikel, Nama Jurnal, nama file dari artikel, dan katakunci dari penulis.

Gambar 4.9 menunjukkan hasil ekstraksi katakunci beserta akurasinya. Akurasi dihitung berdasarkan kesamaannya dengan deretan katakunci dari penulis. Karena katakunci yang dihasilkan setiap metode adalah 5, maka jika ada satu katakunci hasil ekstraksi yang juga ada dalam katakunci buatan penulis maka akurasinya adalah $1/5$ atau 0.2000.

4.7 KETERBATASAN

Meskipun hasil dari metode ekstraksi baru yang dikembangkan penulis kualitasnya adalah yang paling bagus diantara 10 metode state of the art, tetapi metode baru ini memiliki keterbatasan. Keterbatasan pertama adalah pada kosakata yang dipakai dalam menentukan katakunci. Apabila kosakata tersebut tidak selalu diperbaharui maka ada istilah-istilah baru yang tidak bisa diekstrak sebagai katakunci. Misalnya katakunci dibuat pada awal tahun 2019, maka tidak akan ada/muncul hasil ekstraksi dengan katakunci Covid-19.

Keterbatasan kedua dari metode baru ini adalah bahwa ekstraksi akan optimal jika diterapkan pada naskah atau artikel ilmiah bidang Ilmu Komputer. Ini terjadi karena dataset pembentuk kosakata yang dipakai pada ekstraksi katakunci ini adalah dari bidang Ilmu Komputer. Namun keterbatasan ini dapat diatasi dengan membentuk dataset dari domain lain atau untuk seluruh domain.

5 PENUTUP

5.1 SIMPULAN

Berdasar hasil riset yang telah dilakukan penulis maka diperoleh solusi atau jawaban pertanyaan penelitian ini yaitu:

1. Cara menentukan/memilih kosakata populer agar dapat menghasilkan pengestrak katakunci dengan kinerja yang baik adalah dengan mengambil kosakata dari katakunci artikel bersitasi tinggi yang terbit 25 tahun terakhir.
2. Penggunaan sistem pembelajaran mesin untuk ekstraksi katakunci yang menghasilkan kinerja yang terbaik adalah dengan metoda klasifikasi, lebih tepatnya adalah *multi-label classification*.
3. Strategi implementasi dari metode ekstraksi katakunci ini agar dapat dimanfaatkan oleh para penulis artikel ilmiah adalah dengan menerapkan metoda temuan (*novelty*) hasil penelitian ini pada aplikasi berbasis web yang dapat diakses publik. Alamat aplikasi ada di <http://extract.my.id/>.

Dari hasil penelitian ini dan dari paparan di atas juga didapat beberapa simpulan yang bermanfaat bagi pengembangan ilmu pengetahuan, yaitu:

1. Metode novelty dari penelitian ini adalah paling cepat dalam waktu ekstraksi katakunci dibanding metode lain yang hasilnya bagus.
2. Metode ekstraksi katakunci baru secara kualitas (hasil evaluasi dengan metode MRR) adalah selalu paling bagus dibanding metode-metode *state-of-the-art*.
3. Karena kualitasnya paling bagus maka aplikasi hasil disertasi ini dapat dipakai untuk sistem rekomendasi. Rekomendasi dalam penentuan katakunci dari suatu artikel ilmiah.

5.2 LUARAN

Beberapa luaran yang juga novelty dari penelitian ini yang merupakan pengembangan ilmu pengetahuan dan dapat dipakai oleh publik adalah:

1. Metode ekstraksi katakunci baru. Metode ini dibandingkan dengan 10 metode-metode *state-of-the-art* adalah yang paling berkualitas. Diantara tiga metode terbaik, metode baru ini adalah yang tercepat, dimana kecepatannya bisa 10 kali lipatnya.
2. Aplikasi/piranti lunak hasil disertasi ini yang dapat diakses pada situs web <http://extract.my.id/>. Aplikasi ini merupakan system rekomendasi untuk ekstraksi katakunci yang dapat digunakan oleh para penulis artikel ilmiah bidang Ilmu Komputer.

Dalam penelitian ini juga dihasilkan dataset baru untuk ekstraksi katakunci. Dataset baru untuk ekstraksi katakunci tersebut berisi metadata artikel ilmiah lebih baru dan dengan fitur lebih lengkap dibanding dataset yang sudah ada. Dataset ini telah didaftarkan identifikasinya ke International DOI Foundation dengan DOI: 10.17605/OSF.IO/9GHWM, sehingga dataset ini dapat ditelusuri keberadaannya secara digital melalui situs <https://www.doi.org/> dengan memasukkan kodenya.

5.3 SARAN

Untuk para peneliti yang ingin melanjutkan riset ini, penulis memberikan beberapa saran:

1. Buatlah penelitian yang dapat menyelesaikan pembobotan frasa dalam satu kata dimana permasalahannya dijabarkan pada bagian 4.3.2.
2. Penelitian ini masih masih single-domain, untuk penelitian selanjutnya usahakan multi-domain, dimana metode ekstraksi kata/frasa kuncinya dapat berlaku pada semua ranah ilmu pengetahuan.

DAFTAR PUSTAKA

- [1] S. Puri and S. Singh, "A fuzzy matching based image classification system for printed and handwritten text documents," *Journal of Information Technology Research*, vol. 13, no. 2, pp. 155-194, 2020.
- [2] K. Rajesh and B. Lalitha, "Automatic Image Annotation: A Review of Recent Advances and Literature," 2020.
- [3] H. Marin-Castro, J. Hernandez-Resendiz, H. Escalante-Balderas, L. Pellegrin and E. Tello-Leal, "Chained ensemble classifier for image annotation," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26263-26285, 2019.
- [4] S. Renuse and N. Bogiri, "Multi label learning and multi feature extraction for automatic image annotation," 2018.
- [5] P. Budikova, M. Batko and P. Zezula, "ConceptRank for search-based image annotation," *Multimedia Tools and Applications*, vol. 77, no. 7, pp. 8847-8882, 2018.
- [6] A. Abisha, S. Jeevitha, M. Madhurambigai and M. Hemalatha, "A semantic driven CNN - LSTM architecture for personalised image caption generation," 2019.
- [7] J. Hasoon and R. Hassan, "Face Image Retrieval Based on Fireworks Algorithm," 2019.
- [8] A. Priadana and M. Habibi, "Face detection using haar cascades to filter selfie face image on instagram," 2019.
- [9] G. Showkatramani, S. Nareddi, C. Doninger, G. Gabel and A. Krishna, "Trademark image similarity search," 2018.
- [10] S. Chanda, E. Okafor, S. Hamel, D. Stutzmann and L. Schomaker, "Deep learning for classification and as tapped-feature generator in medieval word-image recognition," 2018.
- [11] X. Chen, J. Zhao and R. Yang, "Product appearance detection based on visual keywords matching," *Microprocessors and Microsystems*, vol. 76, pp. -, 2020.
- [12] N. Varma and A. Riyaz, "Content Retrieval using Hybrid Feature extraction from Query Image," 2018.
- [13] X. Chen, W. Zong, N. Deng, S. Liu and Y. Li, "Incremental Patent Semantic Annotation Based on Keyword Extraction and List Extraction," 2020.
- [14] L. Feng, Y. Niu, Z. Liu, J. Wang and K. Zhang, "Discovering technology opportunity by keyword-based patent analysis: A hybrid approach of morphology analysis and USIT,"

Sustainability (Switzerland), vol. 12, no. 1, pp. -, 2020.

- [15] C. de, H. Bridi, P. von and M. Nemitz, "Hypericum species: An analysis on the patent technologies," *Fitoterapia*, vol. 139, pp. -, 2019.
- [16] M. Fomenkova, D. Korobkin, A. Kravets and S. Fomenkov, "Extraction of Knowledge and Processing of the Patent Array," 2019.
- [17] L. Xiao, G. Wang and Y. Liu, "Patent Text Classification Based on Naive Bayesian Method," 2018.
- [18] J. Rossi and E. Kanoulas, "Query generation for patent retrieval with keyword extraction based on syntactic features," 2018.
- [19] L. Helmers, F. Horn, F. Biegler, T. Oppermann and K.-R. Müller, "Automating the search for a patent's prior art with a full text similarity search," *PLoS ONE*, vol. 14, no. 3, pp. -, 2019.
- [20] A. W. Harzing, "A longitudinal study of Google Scholar coverage between 2012 and 2013," *Scientometrics*, pp. 565-575, 2014.
- [21] C. Chen, B. Yang and C. Zhao, "Keywords Extraction Based on Word Relevance Degrees," 2020.
- [22] M. Nayak, N. Das and U. Mohapatra, "Graph based automatic odia keyword extraction from text document," *Test Engineering and Management*, vol. 83, pp. 6389-6396, 2020.
- [23] Y. Zhang, F. Chen, W. Zhang, H. Zuo and F. Yu, "Keywords Extraction Based on Word2Vec and TextRank," 2020.
- [24] Y. Zhang, M. Tuo, Q. Yin, L. Qi, X. Wang and T. Liu, "Keywords extraction with deep neural network model," *Neurocomputing*, vol. 383, pp. 113-121, 2020.
- [25] A. Vanyushkin and L. Graschenko, "Analysis of text collections for the purposes of keyword extraction task," *Journal of Information and Organizational Sciences*, vol. 44, no. 1, pp. 171-184, 2020.
- [26] M. Kölbl, Y. Kyogoku, J. Philipp, M. Richter, C. Rietdorf and T. Yousef, "Keyword extraction in German: Information-theory vs. Deep learning," 2020.
- [27] Y. Jiang, T. Zhao, Y. Chai and P. Gao, "Bidirectional LSTM-CRF models for keyword extraction in Chinese sport news," in *0277786X*, PSISD, 2020.
- [28] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257-289, 2020.
- [29] H. Shah, M. Rezaei and P. Fränti, "DOM-based keyword extraction from Web pages," 2019.
- [30] B. Armouty and S. Tedmori, "Automated keyword extraction using support vector machine

from Arabic news documents," 2019.

- [31] D. Vega-Oliveros, P. Gomes, M. E. and L. Berton, "A multi-centrality index for graph-based keyword extraction," *Information Processing and Management*, vol. 56, no. 6, pp. -, 2019.
- [32] C. R. Sugimoto, V. Larivière, C. Ni and Y. Gingras, "Measuring "The Scientific" in Scientific Literacy.," in *Handbook of Scientometrics*, Springer, Cham., 2018, pp. 229-246.
- [33] F. Narin, K. Stevens and E. S. Whitlow, *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*, Washington, DC: National Science Foundation., 1976.
- [34] E. Alpaydin, *Introduction to Machine Learning* (3rd ed.), Cambridge: MIT Press, 2020.
- [35] Y. Zhang and C. Zhang, "A Review of Machine Learning for Air Quality Prediction," *Atmosphere*, 2020.
- [36] M. A. Khalid, Z. Ali, M. Aslam and H. Imran, "A Review of Data Mining Techniques for Educational Data Analytics," *IEEE Access*, pp. 124949-124969, 2021.
- [37] W. Liu, H. Liu, J. Zhang and X. Wu., "Research on Information Retrieval of E-commerce Website Based on Machine Learning," in *7th International Conference on Industrial Engineering and Applications (ICIEA)*, 2019.
- [38] Merriam-Webster, "Definition of Bibliography," 2 1 2016. [Online]. Available: <http://www.merriam-webster.com/dictionary/bibliography>.
- [39] Oxford, "Definition of bibliography in English," 2 1 2016. [Online]. Available: <http://www.oxforddictionaries.com/definition/english/bibliography>.
- [40] R. J. Sternberg and K. Sternberg, *Cognition* (6th ed.), Wadsworth: Cengage Learning, 2011.
- [41] C. D. Manning and P. R. H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
- [42] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.), SAGE Publications, 2018.
- [43] Y. Zheng and S. Liu, *Bibliometric analysis for talent identification by the subject–author–citation three-dimensional evaluation model in the discipline of physical education*, :, 2020, pp. -.
- [44] S. Song, Z. Wang, S. Xu, S. Ni and J. Xiao, "A novel text classification approach based on Word2vec and textrank keyword extraction," 2019.
- [45] Y. Samyuktha, V. Karthick and S. Magesh, *Effective analysis and classification of news with progressive rating system*, vol. 82, :, 2020, pp. 6653-6655.
- [46] S. Venkatraman, B. Surendiran and R. Arun, *Spam e-mail classification for the Internet of Things environment using semantic similarity approach*, vol. 76, :, 2020, pp. 756-776.

- [47] Y. Song, B. Liu, X. Chen and J. Liu, *Atmospheric pollution mapping of the yangtze river basin: An AQI-based weighted co-word analysis*, vol. 17, ;, 2020, pp. -.
- [48] J. Garrido-Cardenas, B. Esteban-García, A. Agüera, J. Sánchez-Pérez and F. Manzano-Agugliaro, *Wastewater treatment by advanced oxidation process and their worldwide research trends*, vol. 17, ;, 2020, pp. -.
- [49] M. Fakhar, M. Pellegrini, G. Marzi and M. Dabic, *Knowledge Management in the Fourth Industrial Revolution: Mapping the Literature and Scoping Future Avenues*, ;, 2020, pp. -.
- [50] X. Zhang, K. Mueller and S. Yoo, "Keyword extraction for document clustering using submodular optimization," 2017.
- [51] A. Rovetta and A. Bhagavathula, *COVID-19-related web search behaviors and infodemic attitudes in Italy: Infodemiological study*, vol. 22, ;, 2020, pp. -.
- [52] G. Marzi, A. Caputo, E. Garces and M. Dabić, *A Three Decade Mixed-Method Bibliometric Investigation of the IEEE Transactions on Engineering Management*, vol. 67, ;, 2020, pp. 4-17.
- [53] Q. Li, R. Long, H. Chen, F. Chen and J. Wang, *Visualized analysis of global green buildings: Development, barriers and future directions*, vol. 245, ;, 2020, pp. -.
- [54] X. Peng and J. Dai, *A bibliometric analysis of neutrosophic set: two decades review from 1998 to 2017*, vol. 53, ;, 2020, pp. 199-255.
- [55] M. Hirschmann, A. Hart, J. Henckel, P. Sadoghi, R. Seil and C. Mouton, *COVID-19 coronavirus: recommended personal protective equipment for the orthopaedic and trauma surgeon*, vol. 28, ;, 2020, pp. 1690-1698.
- [56] R. Vaishya, M. Javaid, I. Khan and A. Haleem, *Artificial Intelligence (AI) applications for COVID-19 pandemic*, vol. 14, ;, 2020, pp. 337-339.
- [57] N. Donthu, S. Kumar and D. Pattnaik, *Forty-five years of Journal of Business Research: A bibliometric analysis*, vol. 109, ;, 2020, pp. 1-14.
- [58] J. Lou, S.-J. Tian, S.-M. Niu, X.-Q. Kang, H.-X. Lian, L.-X. Zhang and J.-J. Zhang, *Coronavirus disease 2019: A bibliometric analysis and review*, vol. 24, ;, 2020, pp. 3411-3421.
- [59] G. Doğan and S. Kayır, *Global Scientific Outputs of Brain Death Publications and Evaluation According to the Religions of Countries*, vol. 59, ;, 2020, pp. 96-112.
- [60] G.-S. Huertas, P. Jones and O. Llanos-Contrera, *An overview of sport entrepreneurship field: a bibliometric analysis of the articles published in the Web of Science*, vol. 23, ;, 2020, pp. 296-314.
- [61] A. Haleem and M. Javaid, *3D printed medical parts with different materials using additive manufacturing*, vol. 8, ;, 2020, pp. 215-223.
- [62] Y.-H. Lin, C.-H. Liu and Y.-C. Chiu, *Google searches for the keywords of "wash hands"*

- predict the speed of national spread of COVID-19 outbreak among 21 countries*, vol. 87, ;, 2020, pp. 30-32.
- [63] Y. Li, X. Liu, L. Guo, J. Li, D. Zhong, Y. Zhang, M. Clarke and R. Jin, *Traditional Chinese herbal medicine for treating novel coronavirus (COVID-19) pneumonia: Protocol for a systematic review and meta-Analysis*, vol. 9, ;, 2020, pp. -.
- [64] D. Bonilla-Aldana, Y. Holguin-Rivera, I. Cortes-Bonilla, M. Cardona-Trujillo, A. García-Barco, H. Bedoya-Arias, A. Rabaan, R. Sah and A. Rodriguez-Morales, *Coronavirus infections reported by ProMED, February 2000–January 2020*, vol. 35, ;, 2020, pp. -.
- [65] E. Abad-Segura, M.-D. González-Zamar, J. Infante-Moro and G. García, *Sustainable management of digital transformation in higher education: Global research trends*, vol. 12, ;, 2020, pp. -.
- [66] E. Abad-Segura and M.-D. González-Zamar, *Global research trends in financial transactions*, vol. 8, ;, 2020, pp. -.
- [67] S. Sedita, A. Caloffi and L. Lazzeretti, *The invisible college of cluster research: a bibliometric core–periphery analysis of the literature*, vol. 27, ;, 2020, pp. 562-584.
- [68] L. Chen, S. Shang, C. Yang and J. Li, *Spatial keyword search: a survey*, vol. 24, ;, 2020, pp. 85-106.
- [69] B. Zheng, K. Zheng, C. Jensen, N. Hung, H. Su, G. Li and X. Zhou, *Answering Why-Not Group Spatial Keyword Queries*, vol. 32, ;, 2020, pp. 26-39.
- [70] Y. Yang, X. Liu, R. Deng and Y. Li, *Lightweight Sharable and Traceable Secure Mobile Health System*, vol. 17, ;, 2020, pp. 78-91.
- [71] Y. Lu, J. Li and Y. Zhang, *Privacy-Preserving and Pairing-Free Multirecipient Certificateless Encryption With Keyword Search for Cloud-Assisted IIoT*, vol. 7, ;, 2020, pp. 2553-2562.
- [72] R. Marcilly, L. Douze, C. Bousquet and S. Pelayo, *Modeling keyword search strategy: Analysis of pharmacovigilance specialists' search of MedDRA Terms*, vol. 257, ;, 2019, pp. 298-302.
- [73] R. Jayawardena, P. Sooriyaarachchi, M. Chourdakis, C. Jeewandara and P. Ranasinghe, *Enhancing immunity in viral infections, with special emphasis on COVID-19: A review*, vol. 14, ;, 2020, pp. 367-382.
- [74] P. Jiang, F. Guo, K. Liang, J. Lai and Q. Wen, *Searchain: Blockchain-based private keyword search in decentralized storage*, vol. 107, ;, 2020, pp. 781-792.
- [75] Z. Guan, X. Liu, L. Wu, J. Wu, R. Xu, J. Zhang and Y. Li, *Cross-lingual multi-keyword rank search with semantic extension over encrypted data*, vol. 514, ;, 2020, pp. 523-540.
- [76] E. Papagiannopoulou and G. Tsoumakas, "A Review of Keyphrase Extraction," *arXiv*, p. (preprinted article), 2019.

- [77] P. Liamputtong, Handbook of, Singapore: Springer Nature, 2019.
- [78] J. W. Creswell and V. L. P. Clark, Designing and Conducting Mixed Methods Research, SAGE Publications, Inc, 2017.
- [79] H. T. Y. Achsan and W. C. Wibowo, "A Fast Distributed Focused-Web Crawling," *Procedia Engineering vol. 69*, pp. 492-499, 2014.
- [80] DOI, "Digital Object Identifier," 30 12 2022. [Online]. Available: <https://www.doi.org/>.