KEYPHRASE EXTRACTION PADA TEKS BERBAHASA INDONESIA MENGGUNAKAN METODE *TOPICRANK*

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika

Oleh :

Muhammad Raihan Almenata
NIM : 09021381924149

Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2023

LEMBAR PENGESAHAN SKRIPSI

KEYPHRASE EXTRACTION PADA TEKS BERBAHASA INDONESIA MENGGUNAKAN METODE *TOPICRANK*

Oleh :

Muhammad Raihan Almenata
NIM : 09021381924149

Palembang, 12 Juli 2023

Pembimbing I,	Pembimbing II,
Novi Yusliani, S.Kom., M.T.	M Naufal Rachmatullah M.T.
NIP. 198211082012122001	NIP. 199212012022031008

Mengetahui,

Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom.
NIP. 19781222200642003

TANDA LULUS UJIAN KOMPREHENSIF

Pada hari rabu tanggal 14 juni 2023 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya

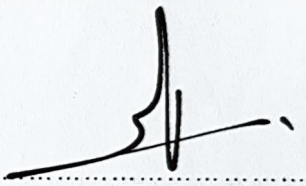
Nama : Muhammad Raihan Almenata
NIM : 09021381924149
Judul : Keyphrase Extraction pada Teks Berbahasa Indonesia Menggunakan Metode Topicrank

dan dinyatakan LULUS.

1. Ketua Penguji

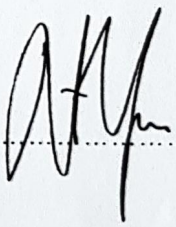
Alvi Syahrini Utami, M.Kom. .....
NIP. 19781222200642003

2. Penguji I



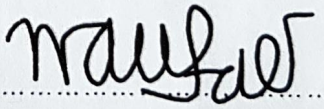
Dr. Abdiansah, S.Kom., M.Cs. .....
NIP. 198410012009121005

3. Pembimbing I



Novi Yusliani, S.Kom., M.T. .....
NIP. 198211082012122001

4. Pembimbing II



M Naufal Rachmatullah M.T. .....
NIP. 199212012022031008

Mengetahui,
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom.
NIP. 19781222200642003

HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini :

Nama : Muhammad Raihan Almenata

NIM : 09021381924149

Program Studi : Teknik Informatika

Judul Skripsi : *Keyphrase Extraction* pada Teks Berbahasa Indonesia Menggunakan Metode *Topicrank*

Hasil Pengecekan Software iThenticate/Turnitin : 12%

Menyatakan bahwa laporan skripsi saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan skripsi ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Palembang, 15 Juli 2023

Muhammad Raihan Almenata

NIM. 09021381924149

MOTTO DAN PERSEMBAHAN

"Whatever happens, do your best."

Skripsi ini kupersembahkan kepada :

- Allah Subhanahu wa Ta'ala
- Orang Tua, Saudari, serta Keluargaku
- Dosen dan Guru - Guruku
- Teman - Temanku
- Fakultas Ilmu Komputer Universitas Sriwijaya
- Universitas Sriwijaya

ABSTRACT

This study investigates the implementation as well as the performance of the application of TopicRank method on performing keyphrase extraction towards Indonesian text. The utilisation of keyphrase as one of the tools on managing the growing textual information drives the need for a relevant keyphrase extraction methodology. By developing the keyphrase extraction system, applying the extraction process towards the title and abstract of Indonesian scientific paper and utilising the comparison of the keyphrase obtained by the extraction against the keyphrase determined by the writer, performance measurement was done towards the extraction process. The matter then followed by performing comparison between the obtained performance of TopicRank extraction process that did and didn't utilise cosine similarity at the postprocessing stage. The obtained result shows that the extraction system utilising the TopicRank method managed to obtain the acquisition of performance metrics by the magnitude of value of 0.7 for accuracy, 0.08 for precision, 0.09 for recall, and 0.09 for f-score with the parameter configuration of 5 selected keyphrases which assessed to be optimal compared to 2 other parameter configuration. Furthermore, the extraction that applied the TopicRank methodology but didn't utilise cosine similarity at the postprocessing stage obtained relatively lower performance metrics values against the optimal potency simulated through the application of cosine similarity at the postprocessing stage. The keyphrase extraction utilising the TopicRank method performed is judged to still be improvable in order to maximise the potency of the extraction performance.

Keywords : Keyphrase Extraction, Graph-Based Ranking, TopicRank

ABSTRAK

Penelitian ini menyelidiki implementasi serta kinerja dari penerapan metode *TopicRank* dalam ekstraksi kata kunci pada teks berbahasa indonesia. Pemanfaatan frasa kunci sebagai salah satu alat dalam pengelolaan informasi tekstual yang bertumbuh mendorong kebutuhan akan metode ekstraksi frasa kunci yang relevan. Dengan mengembangkan sistem ekstraksi kata kunci, menerapkan proses ekstraksi frasa kunci terhadap judul serta abstrak makalah ilmiah berbahasa indonesia dan memanfaatkan perbandingan frasa kunci yang diperoleh dari hasil ekstraksi tersebut terhadap kata atau frasa kunci yang ditentukan oleh penulis, dilakukan pengukuran performa terhadap proses ekstraksi. Hal tersebut dilanjutkan dengan melakukan perbandingan perolehan performa proses ekstraksi *TopicRank* yang menggunakan dan tidak menggunakan *cosine similarity* pada tahapan *postprocessing*, untuk mensimulasikan potensi optimal yang dapat dicapai metode *TopicRank* dengan konfigurasi proses *preprocessing* dan dataset yang sama dengan tetap memberikan derajat error tertentu. Hasil yang diperoleh menunjukkan sistem ekstraksi menggunakan metode *TopicRank* berhasil memperoleh perolehan metrik performa dengan besaran nilai 0.7 untuk akurasi, 0.08 untuk presisi, 0.09 untuk *recall* dan 0.09 untuk *f-score* dengan menggunakan konfigurasi parameter 5 *selected keyphrases* yang dinilai optimal dibandingkan 2 konfigurasi parameter lainnya. Selanjutnya, ekstraksi dengan menggunakan metode *TopicRank* yang tidak menerapkan proses cosine similarity pada *postprocessing* masih memperoleh nilai metrik performa yang relatif lebih rendah terhadap potensi optimal yang disimulasikan melalui penerapan *cosine similarity* pada *postprocessing*. Ekstraksi frasa kunci menggunakan metode *TopicRank* yang dilakukan dinilai masih dapat ditingkatkan untuk memaksimalkan potensi performa ekstraksi yang dilakukan.

Kata Kunci : Ekstraksi Frasa Kunci, Pemeringkatan Berbasis Grafik, *TopicRank*

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Allah SWT atas berkat, rahmat, karunia dan hidayahNya sehingga penulisan karya tulis tugas akhir dengan judul "*Keyphrase Extraction* pada Teks Berbahasa Indonesia Menggunakan Metode *Topicrank*" dapat penulis selesaikan dengan baik. Penulisan karya tulis ini ditujukan salah satunya untuk memenuhi persyaratan dalam menyelesaikan studi Strata-I pada program studi Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya. Selanjutnya, penulis juga menyampaikan banyak ungkapan terima kasih kepada :

1. Allah SWT atas berkat, rahmat, karunia dan hidayahNya.
2. Mama dan Papa terkasih atas do'a, dukungan, dan kesabarannya.
3. Adik perempuanku tersayang, Mutia Adilah Almenata atas dukungan serta teladannya.
4. Alm. bapak Jaidan Jauhari, S.Pd., M.T atas jasa kepemimpinannya selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
5. Ibu Alvi Syahrini Utami, M.Kom. atas jasa dan bantuannya selaku Ketua Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
6. Ibu Novi Yusliani, S.Kom., M.T. atas bantuan, bimbingan, arahan, dan dukungannya selaku Dosen Pembimbing 1.
7. Bapak Muhammad Naufal Rachmatullah M.T. atas bimbingan, arahan serta dukungannya selaku Dosen Pembimbing 2.

8. Alm. Bapak Drs. Megah Mulya, M.T. atas dedikasi, inspirasi, teladan, serta jasanya.
9. Para Dosen Jurusan Teknik Informatika serta Fakultas Ilmu Komputer Universitas Sriwijaya atas jasa, bantuan serta kontribusinya dalam membantu penulisan karya tulis ini.
10. Staf dan Pegawai Jurusan Teknik Informatika serta Fakultas Ilmu Komputer Universitas Sriwijaya atas jasa, bantuan serta kontribusinya dalam membantu penulisan karya tulis ini.
11. Fadel, Bintang, Andre, Alvin dan Asyraf selaku sahabat penulis atas dukungan, bantuan, doa, inspirasi, perjuangan bersama serta persahabatannya selama ini dan Insyaallah berlanjut pada tahun-tahun kedepan.
12. Teman-teman FOVV, Nilam, Zafira, Shabrina, Rani, Reyhani, Aulia, Tarisa, Nurul, dan Fidyah atas dukungan, bantuan, dan perjuangan bersamanya.
13. Rekan-rekan *Development Student Club* Universitas Sriwijaya, tim IT Magang RSUD Siti Fatimah Program Kampus Merdeka 2021 dan Tim *Capstone Adha'ar* pada Bangkit 2022 atas dukungan, inspirasi, serta bantuannya.
14. Bu Diana, Kak Indra, Kak Emir dan Kak Nabilah atas teladan, inspirasi, dan bantuannya.
15. Teman-teman dan pihak-pihak lainnya yang tidak disebutkan satu persatu.

Penulis menyadari bahwa masih terdapat banyak kekurangan pada proses maupun hasil dari penulisan tugas akhir ini, maka dari itu kritik dan saran yang konstruktif sangat penulis apresiasi dan harapkan. Terakhir, penulis berharap karya tulis ini dapat bermanfaat dan berdampak kebaikan bagi banyak orang dan penulis.

Palembang, 30 Mei 2022

Muhammad Raihan Almenata

DAFTAR ISI

Halaman

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Pendahuluan

Pada bagian ini, diuraikan berbagai pokok-pokok substansi yang berperan sebagai landasan sekaligus pondasi dilakukannya penelitian ini. Pokok-pokok substansi yang dimaksud diantaranya adalah latar belakang masalah penelitian, perumusan masalah/permasalahan penelitian, tujuan penelitian dan manfaat penelitian.

1.2 Latar Belakang Masalah

Pesatnya perkembangan serta pemanfaatan teknologi informasi turut mendorong terjadinya peningkatan jumlah informasi digital secara signifikan. Dewasa ini, data dari berbagai sumber heterogen dihasilkan dalam jumlah yang besar dihasilkan sehari-hari dalam tingkatan yang belum pernah terjadi sebelumnya (Oussous et al., 2018). Fenomena ledakan pada jumlah informasi digital yang tersedia juga turut membawa tantangan serta kesempatan baru dalam proses pengolahan informasi. Misalnya, pertumbuhan penggunaan media sosial membuka kesempatan baru untuk menganalisa sejumlah aspek dan pola dalam komunikasi. Sedangkan, sangat pesatnya pertumbuhan tersebut juga turut menyebabkan sebuah peningkatan akumulasi data (Stieglitz et al., 2018). Contoh konkrit lainnya, yaitu dokumen ilmiah disertai dengan kompleksitasnya yang dihasilkan pada volume yang cukup besar menyebabkan diperlukannya waktu

serta upaya yang tidak layak untuk secara manual mengekstraksi kumpulan ringkas *keyphrases* (Asl & Banda, 2020).

Eksplorasi dan penerapan pada bidang *Natural Language Processing,* yaitu suatu bidang penelitian dan aplikasi yang menelusuri penggunaan komputer untuk memahami dan memanipulasi teks atau ucapan bahasa alami manusia untuk melakukan hal yang bermanfaat (Chowdhary, 2020) diharapkan dapat menjadi salah satu jawaban dalam menghadapi tantangan serta kesempatan yang muncul, terkhususnya terhadap informasi dalam bentuk tekstual. *Automatic Keyphrase Extraction* merupakan kegiatan mengidentifikasi kata dan frasa penting yang paling menggambarkan dokumen teks tertentu (Saxena et al., 2020). *Automatic Keyphrase Extraction* merupakan salah satu aplikasi dalam bidang *Natural Processing Language,* dan diharapkan dapat membantu membangun metode pengelolaan informasi yang lebih efektif dan efisien melalui pemanfaatan frasa kunci dokumen. Selanjutnya, *Automatic Keyphrase Extraction* dapat diterapkan melalui sejumlah metode yang berbeda.

TopicRank merupakan sebuah metodologi ekstraksi frasa kunci yang menerapkan pemeringkatan berbasis grafik dengan mengandalkan representasi topik pada sebuah dokumen (Bougouin dkk, 2013). Metode ini menawarkan karakteristik yang memanfaatkan penggambaran topik pada sebuah dokumen dalam melakukan ekstraksi kata kunci. Sehingga, melalui metode ini penulis berharap dapat memanfaatkan karakteristik yang ditawarkan dalam menerapkan *Automatic Keyphrase Extraction*.

1.3 Rumusan Masalah

Penulis memecah rumusan masalah pada penelitian ini kedalam 2 pertanyaan, yaitu :

1. Bagaimana mengembangkan perangkat lunak ekstraksi kata kunci menggunakan metode *TopicRank* pada teks berbahasa indonesia?
2. Bagaimana kinerja perangkat lunak dalam mengekstraksi kata kunci pada teks berbahasa indonesia menggunakan metode *TopicRank*?

1.4 Tujuan Penelitian

Adapun tujuan penelitian ini dirumuskan sebagai berikut:

1. Menghasilkan perangkat lunak ekstraksi kata kunci menggunakan metode *TopicRank* pada teks berbahasa indonesia.
2. Mengetahui kinerja metode *TopicRank* dalam mengekstraksi kata kunci pada teks berbahasa indonesia.

1.5 Manfaat Penelitian

Adapun manfaat yang diharapkan diperoleh melalui penelitian dijabarkan sebagai berikut:

1. Hasil penelitian ini dapat dijadikan referensi pada bidang penelitian terkait.
2. Perangkat lunak yang dihasilkan dapat digunakan untuk melakukan operasi *keyphrase extraction* menggunakan metode *TopicRank*.

1.6 Batasan Masalah

Adapun batasan permasalahan yang diangkat pada penelitian ini dijabarkan sebagai berikut :

1. Jenis data yang digunakan pada penelitian ini merupakan judul, abstrak, dan kata kunci makalah ilmiah berbahasa indonesia dengan predikat sinta 2 atau 3.

1.7 Sistematika Penelitian

Berikut penjabaran sistematika penulisan yang diterapkan dalam penulisan penelitian ini :

BAB I. PENDAHULUAN

Bab ini menguraikan latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah/ruang lingkup dan sistematika penulisan.

BAB II. KAJIAN LITERATUR

Bab ini menjabarkan dasar-dasar teori yang digunakan dalam penelitian, seperti *Automatic Keyphrase Extraction, TopicRank,* dan *Confusion Matrix.*

BAB III. METODOLOGI PENELITIAN

Bab ini membahas komponen-komponen pada metodologi penelitian yang terdiri dari pengumpulan data, tahapan penelitian, metode pengembangan perangkat lunak, dan manajemen proyek perangkat lunak.

1.8 Kesimpulan

Pada bab ini telah dijabarkan latar belakang, rumusan masalah, tujuan dan manfaat, batasan masalah, serta sistematika penulisan pada penelitian ini.

DAFTAR PUSTAKA

Asl, J. R., & Banda, J. M. (2020). Gleake: Global and local embedding automatic keyphrase extraction. arXiv preprint arXiv:2005.09740.

Basil, V. R., & Turner, A. J. (1975). Iterative enhancement: A practical technique for software development. IEEE Transactions on Software Engineering, (4), 390-396.

Boudin, F. (2018). Unsupervised keyphrase extraction with multipartite graphs. arXiv preprint arXiv:1803.08721.

Bougouin, A., Boudin, F., & Daille, B. (2013, October). Topicrank: Graph-based topic ranking for keyphrase extraction. In International joint conference on natural language processing (IJCNLP) (pp. 543-551).

Bowes, D., Hall, T., & Gray, D. (2012, September). Comparing the performance of fault prediction models which report multiple performance measures: recomputing the confusion matrix. In Proceedings of the 8th international conference on predictive models in software engineering (pp. 109-118).

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-117.

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. Annals of Mathematics and Artificial Intelligence, 81(3), 429-450.

Chen, J., Zhang, X., Wu, Y., Yan, Z., & Li, Z. (2018). Keyphrase generation with correlation constraints. arXiv preprint arXiv:1808.07185.

Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. Fundamentals of artificial intelligence, 603-649.

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).

Hasan, K. S., & Ng, V. (2014, June). Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1262-1273).

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 216-223).

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. International Journal of Computer Science & Communication Networks, 5(1), 7-16.

Liao, S., Yang, Z., Liao, Q., & zheng, Z. (2023). TopicLPRank: a keyphrase extraction method based on improved TopicRank. The Journal of Supercomputing, 1-20.

Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).

Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University-Computer and Information Sciences, 30(4), 431-448.

Papagiannopoulou, E., & Tsoumakas, G. (2020). A review of keyphrase extraction. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(2), e1339.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., & Mononen, J. (2018). Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. Behavioural processes, 148, 56-62.

Sasirekha, K., & Baby, P. (2013). Agglomerative hierarchical clustering algorithm-a. International Journal of Scientific and Research Publications, 83(3), 83.

Saxena, A., Mangal, M., & Jain, G. (2020, December). Keygames: A game theoretic approach to automatic keyphrase extraction. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 2037-2048).

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. International journal of information management, 39, 156-168.