Text Mining Repository Untuk Tren Tema Skripsi 2017-2020

Putry Wahyu Setyaningsih[1*], Arita Witanti[2]
[1] Sistem Informasi, Universitas Mercu Buana, Yogyakarta 55823, Indonesia
[2] Informatika, Universitas Mercu Buana, Yogyakarta 55823, Indonesia
*putryws@mercubuana-yogya.ac.id, arita@mercubuana-yogya.ac.id

Abstrak

Skripsi merupakan salah satu tolak ukur dari ketercapaian pembelajaran atau CPL dan sebuah kewajiban bagi mahasiswa di Perguruan Tinggi untuk menyelesaikan studi di tingkat strata 1. Apakah skripsi telah sesuai atau belum dengan target skema penelitian universitas atau target capaian lulusan setiap prodi dapat akan terlihat dari hasil publikasi dokumen. Publikasi dokumen skripsi yang lengkap dapat ditemui di halaman repository universitas. Gambaran ringkasnya terkhusus pada dokumen abstrak dari skripsi. Dari dokumen abstrak skripsi ini, peneliti dapat memotret tren tema skripsi di repository universitas yang nantinya hasilnya akan bermanfaat untuk manajemen. Tujuan dari penelitian ini adalah menemukan tren tema skripsi di repository dari 2017 sampai dengan 2020 awal. Tren ini dapat menjadi potret sebenarnya apakah jurusan tersebut update atau tidak untuk tema skripsinya serta untuk mengukur apakah sudah tercapai capaian pembelajaran lewat tema skripsi. Sejak repository mulai ada tahun 2017, setidaknya terdapat 800 dokumen repository skripsi dari 13 program studi. Semakin lama dokumen semakin bertambah seiring bertambahnya lulusan. Agar data dokumen ini bisa dimanfaakan maka kita memerlukan teknik tertentu untuk mengambilnya dianatarnya text mining. Text mining adalah salah satu teknik dalam data mining untuk mengambil 'intisari' dari sebuah text. Teknik clustering akan dipakai untuk mendapatkan intisari dari semua abtrak pada repository.
Kata kunci: Clustering, Text Mining, Repository

Abstract

Thesis is one of the benchmarks for learning achievement or CPL and an obligation for students in tertiary institutions to complete studies at the strata 1 level. Whether or not the thesis is in accordance with the target of the university research scheme or the achievement target of graduates of each study program can be seen from the publication results. document. The publication of complete thesis documents can be found on the university repository page. The brief description is especially in the abstract document from the thesis. From the abstract of this thesis, the researcher can capture the trend of thesis themes in the university repository, the results of which will be useful for management. The purpose of this study is to find trends in thesis themes in the repository from 2017 to early 2020. This trend can be a true portrait of whether the department is updating or not for the thesis theme and to measure whether learning outcomes have been achieved through the thesis theme. Since the repository started in 2017, there are at least 800 thesis repository documents from 13 study programs. The longer the document, the more the number of graduates increases. In order for this document data to be used, we need certain techniques to retrieve it, including text mining. Text mining is one of the techniques in data mining to extract the 'gist' of a text. Clustering technique will be used to get the gist of all abstracts in the repository.
Keywords: *Clustering, Text Mining, Repository*

1. PENDAHULUAN

Skripsi atau Tugas Akhir merupakan salah satu mata kuliah wajib yang harus diselesaikan oleh setiap mahasiswa. Skripsi merupakan persyaratan yang harus ditempuh untuk mendapatkan status sarjana (S1) di setiap Perguruan Tinggi Negeri (PTN) maupun Perguruan Tinggi Swasta (PTS) yang ada di Indonesia (Nurdin & Munthoha, 2017). Skripsi menghasilkan dokumen skripsi salah satunya adalah abstrak yang disimpan di repository universitas. Semakin lama semakin banyak dokumen repository di unggah semakin kaya sumber repository internal. Kekayaannya bisa diambil dan disarikan untuk tujuan rekomendasi pengelola universitas. Program studi memiliki capaian pembelajaran dan kompetensi lulusan dalam mewujudkan visi dan misi program studi yang selalu diperbaharui setiap 5 tahun sekali. Tema skripsi yang diambil mahasiswa harusnya masih relevan dengan capaian pembelajaran program studi untuk itu di-

perlukan cara menggali informasi tren tema skripsi berdasarkan dokumen dokumen skripsi yang sudah ada, agar nantinya bila tren tema tidak sesuai dengan CPL bisa direkomendasikan tindakan peningkatan tema penelitian skripsi kepada pengelola prodi. Pemberian label tren topik diharapkan dapat membantu mahasiswa dalam mengetahui topik apa yang sedang tren di tahun sebelumnya tanpa harus membaca secara keseluruhan skripsi yang ada (Sulartopo, 2015).

Beberapa publikasi ilmiah yang berkaitan dengan topik penelitian ini, salah satunya adalah "Aplikasi Text Mining Untuk Automasi Penentuan Tren Skripsi Dengan Metode K-Means Clustering (Studi Kasus: Prodi Sistem Komputer)". Penelitian ini memanfaatkan teknologi text mining dan algoritma K-Means Clustering untuk mengetahui tren topik skripsi mahasiswa (Riyadhi, 2019).

Artikel dengan judul "Implementasi *Text Mining* Pengelompokan Dokumen Skripsi Menggunakan Metode *K-Means Clustering*". Tujuan dari penelitian ini adalah pengelompokkan 2 cluster, cluster pertama didominasi dengan metode klasifikasi, sedangkan cluster kedua didominasi dengan metode analisis pengendalian mutu dan matematika asuransi (Adhe et al., 2020).

Artikel dengan judul "Automasi Penentuan Tren Topik Skripsi Menggunakan Algoritma K-Means Clustering". Penelitian ini menghasilkan aplikasi yang dirancang dapat berjalan dengan baik dengan tingkat akurasi sebesar 84% dari 70 data uji (Fuadi et al., 2022).

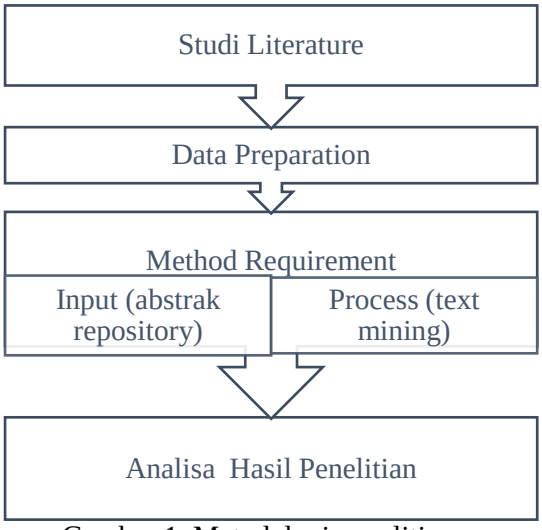
Salah satu cara untuk menggali informasi penting dari dokumen adalah menggunakan *text mining*, yaitu proses mensarikan dokumen dari pengolahan text. Text yang akan diolah berasal dari dokumen abstrak skripsi mahasiswa di repository yang sudah mencapai ratusan . Studi kasus diambil dari repository perpustakaan UMBY. Teknik yang akan dipakai untuk penelitian ini adalah clustering , yaitu penentuan cluster tren tema skripsi berdasarkan peminatan di setiap jurusan atau ranking rekomendasi tren tertinggi.

2. METODE

Alur penelitian dimulai dari *studi literature* yang berupa pemahaman akan masalah dengan dukungan pencarian referensi dan metodologi yang sesuai dengan penelitian terkait. Dengan mempelajari studi literatur dapat menemukan ide-ide penelitian dan mencari metode yang sesuai dengan penelitian terkait bahkan dapat juga memodifikasi penelitian sebelumnya.

Kemudian dilanjutkan proses kedua, yaitu mempersiapkan data yang sesuai untuk inputan metodologi. Data yang digunakan dalam penelitian ini adalah data-data dokumen abstrak skripsi Fakultas Ilmu Komunikasi dan Fakultas Teknologi Informasi dari tahun 2017 hingga awal tahun 2020. Data yang dikumpulkan dari dua fakultas tersebut mendapatkan sampel sebanyak 5747 data abstrak skripsi.

Tahap selanjutnya adalah menginput data 5747 dokumen abstrak skripsi ke aplikasi *Orange*, lalu diproses dengan *text mining*. Setelah menginput data 5747 dokumen abstrak dan diproses *text mining* maka akan keluar hasil kata-kata yang sering digunakan dalam dokumen abstrak tersebut. Alur metode penelitian ditunjukkan seperti Gambar 1.



Gambar 1. Metodologi penelitian

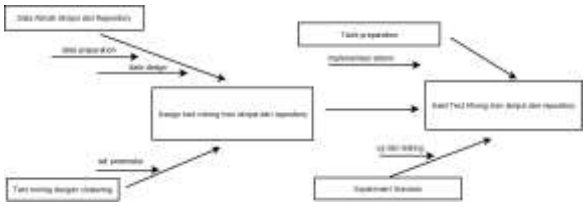
3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data primer sebanyak 800 dokumen abstrak skripsi. Periode transaksi data dari Januari 2017 sampai Februari 2020 yaitu sejak UMBY menerapkan online repository.

Data tersebut adalah data abstrak skripsi dari semua program studi di UMBY yang telah mengikuti proses yudisium. Selain data primer terdapat juga data sekunder yang mendukung yaitu dokumen kurikum UMBY tahun 2017.

Analisis yang dilakukan adalah mendapatkan kata paling banyak dari total skripsi setiap prodi berdarkan tahun upload abstrak skripsi. Kemudi-

an melakukan proses clustering terhadap hasil *text mining* abstrak tersebut. Hasilnya nantinya akan dipakai untuk rekomendasi ketercapaian CPL dan kompetensi lulusan berbasis pemintan program studi. Langkah-langkah dalam proses pengelompokkan data diperoleh sebagai berikut pada Gambar 2.



Gambar 2. Tahap data diperoleh

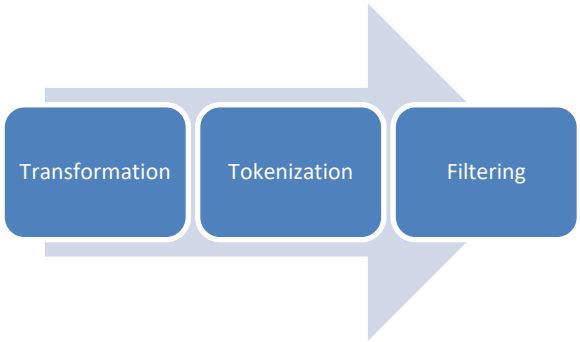
Penelitian ini telah sampai kepada tahap pengumpulan data dimana diperoleh data mentah berupa data text abstrack dari repository UMBY data tersebut terbagi menjadi 8 file XLS. Gambar 3 adalah contoh salah satu data dari repository.

Gambar 3. Gambaran data mentah

Ada beberapa tahap penyelesaian dalam mengolah data abstrak skripsi dengan studi kasus di repository UMBY. Tahap pertama adalah *preprocessing* data, tahap kedua pengolahan *text mining*, tahap ketiga analisis system.

Tahap pertama melakukan *preprocessing* data yaitu mengolah dokumen yang didapat agar lebih mudah untuk dilakukan proses *clustering*. Tahap *preprocessing* ini berfungsi untuk menghilangkan *noise* data, meningkatkan citra dan juga bisa menentukan bagian citra yang akan digunakan dalam tahap selanjutnya (Sutramiani et al., 2015).

Pada tahap *preprocessing* ada 3 tahapan *Transformation*, *Tokenization* dan *Filtering* yang ada pada Gambar 4.



Gambar 4. Tahap *preprocessing*

Tahap yang pertama adalah *Transformation*, tahap ini adalah membentuk proses yang diharapkan dengan cara mengubah kata-kata kedalam bentuk dasar dan pengurangan kata-kata yang tidak digunakan, semisal menghilangkan tanda titik koma dan huruf besar dalam sebuah kata. Tahap kedua adalah *Tokenization*, yaitu tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Tahap terakhir adalah *Filtering*, yaitu tahap mengambil kata-kata yang penting dari *tokenization* menggunakan stopworddina.txt yang ada pada aplikasi Orange.

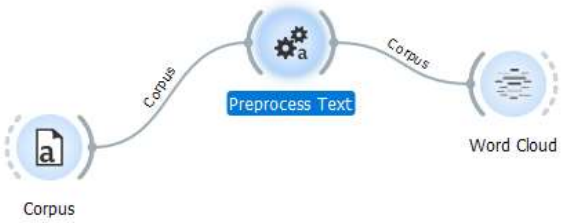
Hasil dari *preprocessing* data seperti Gambar 5.

Gambar 5. *Preprocessing* data

Tahapan yang kedua adalah pengolahan *text mining*. *Text mining* merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dimana teks mining adalah variasi dari data mining yang mencoba menemukan pola menarik dari kumpulan data tekstual.

Pengolahan *text mining* meliputi proses *loading corpus*, *preprocessing text* dan mencari *word cloud* untuk menemukan kata kata terbanyak. Disini dipakai dokumen *eprint* sebanyak 5747 dokumen abstrak dalam skripsi mahasiswa Uni-

versitas Mercu Buana Yogyakarta, seperti Gambar 6.



Gambar 6. Proses *text mining*

Tahap terakhir adalah analisis system. Tahap analisis ini mengecek kesesuaian *cluster* hasil data abstrak dengan *cluster* kamus tema skripsi di setiap program studi. Serta melakukan analisis kesesuaiannya. Tabel kurikulum 2017 pada Fakultas Psikologi dan Fakultas Teknologi Informasi ada pada Gambar 7 dan Gambar 8.

Words & weights

Weight	Word
5	pendidikan
5	masyarakat
4	pengabdian
3	menyelenggarakan
3	psikologi
3	bidang
3	pengabdian masyarakat
3	teknologi
3	informasi
3	teknologi informasi
2	bermutu
2	pengetahuan
2	nasional
2	menyelenggarakan pendidikan
2	ilmu
2	melaksanakan
2	mengembangkan
2	mewujudkan
2	cita

Gambar 7. Hasil banyaknya word cloud kurikulum 2017

Gambar 8. Hasil *word cloud* kurikulum 2017

Hasil yang dicapai adalah *word cloud* untuk data skripsi di Universitas Mercu Buana Yogyakarta sebanyak 5747 sample data seperti Gambar 9 dan Gambar 10.

Words & weights

Weight	Word
3872	data
3772	hasil
2761	analisis
2471	metode
2289	kunci
2142	perusahaan
2104	yogyakarta
1962	uji
1934	pengaruh
1872	variabel
1853	2
1771	berpengaruh
1727	kerja
1718	to
1714	in
1698	nilai
1630	hubungan
1590	bertujuan
1544	kinerja
1495	karyawan

Gambar 9. Hasil banyaknya jumlah per *word cloud*

Gambar 10. Hasil *word cloud*

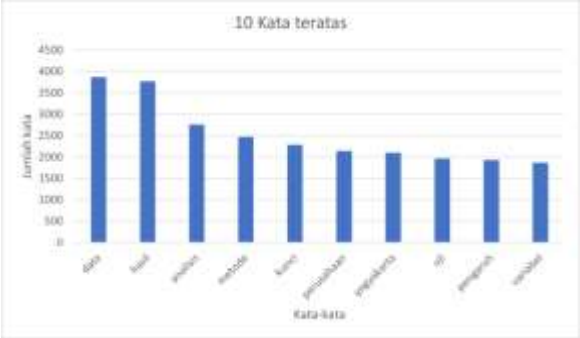
Dari hasil kesesuaian kurikulum 2017 pada Fakultas Psikologi dan Fakultas Teknologi Informasi didapatkan beberapa kesesuaian kata dalam Tabel 1 di bawah ini.

Tabel 1. Kesesuaian hasil abstrak

Kurikulum 2017	Skripsi
Tingkat	Tingkat
Menghasilkan	Hasil
Kerjasama	Kerja

4. KESIMPULAN

Kesimpulan dari penelitian dengan sampel data 5747 menghasilkan *word cloud* tren skripsi dari mahasiswa Universitas Mercu Buana Yogyakarta telah memenuhi tema. Dari 5747 abstrak skripsi, paling banyak mahasiswa menggunakan kata "data" dengan jumlah kata 3872 kata. Peringkat kedua, kata yang sering digunakan adalah "hasil" dengan jumlah 3772 kata. Peringkat ketiga, kata yang sering digunakan adalah "analisis" dengan jumlah kata 2761.



Gambar 11. Grafik perolehan 10 kata teratas

DAFTAR PUSTAKA

Adhe, D., Rachman, C., Goejantoro, R., & Tisna, D. (2020). Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering. *Jurnal EKSPONENSIAL*, *11*(2), 167–174.

Agham, V., & Shandilya, V. K. (2021). A Survey Paper on Extractive and Abstractive Techniques in Automatic Text Summarization. *International Journal of Research Publication and Reviews*, *2*(4), 619–625. www.ijrpr.com

Alami, N., Meknassi, M., & Rais, N. (2015). Automatic Texts Summarization: Current State of the Art. *Journal of Asian Scientific Research*, *5*(1), 1–15. https://doi.org/10.18488/journal.2/2015.5.1/2.1.1.15

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. http://arxiv.org/abs/1707.02919

Aristoteles. (2013). Penerapan Algoritma Genetika pada Peringkasan Teks Dokumen Bahasa Indonesia. *Semirata FMIPA Universitas Lampung*, 29–33. http://jurnal.fmipa.unila.ac.id/index.php/semirata/article/download/703/523

Aulia, T. M. P., Jamaludin, A., & ... (2021). Extractive Text Summerization Pada Berita Berbahasa Indonesia Menggunakan Algoritma Support Vector Machine. *J-SAKTI (Jurnal Sains …*, *5*(September), 727–735. http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/371

Burhan Ul Haq, H., Asif, M., & Bin Ahmad, M. (2021). Video Summarization Techniques: A Review Article in. *International Journal of Scientific & Technology Research*, *9*(11), 146–153. www.ijstr.org

Cartwright, R. (2010). Book Reviews: Book Reviews. *Perspectives in Public Health*, *130*(5), 239–239. https://doi.org/10.1177/1757913910379

198

Fauzan, I. (2020). Artificial Intelligence (AI) Pada Proses Pengawasan dan Pengendalian Kepegawaian - Sebuah Eksplorasi Konsep Setelah Masa Pandemi Berakhir. *Jurnal Civil Service*, *14*(1), 31–42.

Fuadi, W., Razi, A., & Fariadi, D. (2022). *Automasi Penentuan Tren Topik Skripsi Menggunakan Algoritma K-Means Clustering*. *VII*(2), 3072–3077.

Goralski, M. A., & Tan, T. K. (2020). Artificial intelligence and sustainable development. *International Journal of Management Education*, *18*(1). https://doi.org/10.1016/j.ijme.2019.100330

Hai-Jew, S. (2019). *Applying Qualitative Matrix Coding Queries and Qualitative Crosstab Matrices for Explorations of Online Survey Data*. 181–204. https://doi.org/10.4018/978-1-5225-8563-3.ch008

Mahdiyah, E., & Andriyani, Y. (2013). Analisa Algoritma Pemahaman Kalimat Pada ALICE ChatBot Dengan Menggunakan Artificial Intelligence Markup Language (AIML). *Prosiding SEMIRATA 2013*, *1*(1), 193–201.

Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Edik Informatika*, *2*(2), 213–219. https://doi.org/10.22202/ei.2016.v2i2.1465

Moratanch, N., & Chitrakala, S. (2017). A survey on extractive text summarization. *International Conference on Computer, Communication, and Signal Processing: Special Focus on IoT, ICCCSP 2017*. https://doi.org/10.1109/ICCCSP.2017.7944061

Muttaqin, F. A., & Bachtiar, A. M. (2016). Implementasi Teks Mining Pada Aplikasi Pengawasan Penggunaan Internet Anak "Dodo Kids Browser." *Jurnal Ilmiah Komputer Dan Informatika*, 1–8.

Nurdin, N., & Munthoha, A. (2017). Sistem Pendeteksian Kemiripan Judul Skripsi Menggunakan Algoritma Winnowing. *InfoTekJar(Jurnal Nasional Informatika Dan Teknologi Jaringan)*, *2*(1), 90–97.

Pradnyana, G. A., & Mogi, I. K. A. (2014). Implementasi Automated Text Summarization Untuk Dokumen Tunggal Berbahasa Indonesia Dengan Menggunakan Graph-Based. *Jurnal Ilmiah NERO*, *1*(2), 33–46.

Riyadhi, M. F. (2019). *Aplikasi Text Mining Untuk Automasi Penentuan Tren Topik Skripsi Dengan Metode K-Means Clustering (Studi Kasus: Prodi Sistem Komputer)*. *2*(1), 1–6.

Romadhony, A., Z.R, F., Yusliani, N., & Abednego, L. (2017). Text Summarization untuk Dokumen Berita Berbahasa Indonesia. *Konferensi Nasional ICT-M Politeknik Telkom*, 408–414. //journals.telkomuniversity.ac.id/knip/article/view/586

Sah, S., Kulhare, S., Gray, A., Venugopalan, S., Prud'hommeaux, E., & Ptucha, R. (2017). Semantic text summarization of long videos. *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, *June 2020*, 989–997. https://doi.org/10.1109/WACV.2017.115

Setiawan, A., Kurniawan, E., & Handiwidjojo, W. (2013). Implementasi Stop Word Removal Untuk Pembangunan Aplikasi Alkitab Berbasis Windows 8. *Jurnal EKSIS*, *6*(2), 1–11.

Sulartopo, S. (2015). Pengkategorian Topik Skripsi Dengan Metode NBC. *E-Bisnis*, *8*(1), 49–53.

Suputra, I. P. G. H. (2018). Peringkasan Teks Otomatis Untuk Dokumen Bahasa

Bali Berbasis Metode Ektraktif. *Jurnal Ilmu Komputer*, *10*(1), 33–38.

Sutramiani, N. P., Darmaputra, Ik. G., & Sudarma, M. (2015). Local Adaptive Thresholding Pada Preprocessing Citra Lontar Aksara Bali. *Majalah Ilmiah Teknologi Elektro*, *14*(1), 27–30. https://doi.org/10.24843/mite.2015.v14i01p06

Wijayanto, I. R., Cholissodin, I., & Sari, Y. A. (2021). *Pengaruh Metode Word Embedding dalam Vector Space Model pada Pemerolehan Informasi Materi IPA Siswa SMP*. *5*(3), 950–959.

Yuliska, Y., & Syaliman, K. U. (2020). Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia. *IT Journal Research and Development*, *5*(1), 19–31. https://doi.org/10.25299/itjrd.2020.vol5(1).4688