JINTEKS (Jurnal Informatika Teknologi dan Sains) ISSN 2686-3359 (Online)
Vol. 4 No. 3, Agustus 2022, hlm. 255 – 259

TEXT MINIG CLUSTERING UNTUK PENGELOMPOKAN TOPIK DOKUMEN PENELITIAN MENGGUNAKAN ALGORITMA *K-MEANS* DENGAN *COSINE SIMILARITY*

Nengah Widya Utami[1*], I Gede Juliana Eka Putra[2]
[1]Sistem Informasi Akuntansi, STMIK Primakara
[2]Teknik Informatika, STMIK Primakara
email: widya@primakara.ac.id

Abstrak: Penelitian merupakan salah satu unsur yang wajib dilakukan baik oleh dosen maupun mahasiswa di perguruan tinggi. Dalam hal ini memungkinkan para peneliti mengambil topik yang sama atau hampir serupa. Melalui penelitian ini akan dilakukan analisis untuk mengelompokkan dokumen penelitian. Hasil dari pengelompokan dokumen penelitian ini akan memperlihatkan bagaimana pola kemiripan dan keterkaitan antar penelitian dalam bentuk *cluster*. Data yang digunakan dalam penelitian ini adalah judul penelitian dosen tahun 2019-2021 berumlah 52 data. Proses ekstraksi dokumen dilakukan dengan menggunakan proses *text mining*, sedangkan untuk proses pengelompokan dokumen dilakukan dengan menggunakan metode k-*means clustering* dengan *cosine similarity*. Pada tahap *text mining* dilakukan proses *preprocessing* diantaranya proses *tokenization, filtering dan stemming*. Algoritma K-Means mampu menghasilkan *cluster* optimal yang berjumlah 6 *cluster*. Trend topik penelitian yang dilakukan dosen di STMIK Primakara meliputi Pengembangan dan Evaluasi Sistem Informasi, *E-Government*, *Data Mining*, Teknologi Pendidikan, *Machine Learning/Artificial Intelligence*, serta Manajemen dan Bisnis.

Kata kunci: text mining, clustering, data mining, dokumen

Abstract: *Research is one of the elements that must be carried out by both lecturers and students in universities. In this case it allows researchers to take the same or almost similar topics. Through this research, analysis will be carried out to classify research documents. The results of grouping research documents will show how the patterns of similarity and interrelationships between studies are in the form of clusters. The data used in this study is the title of the lecturer's research for the years 2019-2021 totaling 52 data. The document extraction process is carried out using a text mining process, while the document grouping process is carried out using the k-means clustering method with cosine similarity. At the text mining stage, a preprocessing process is carried out including the tokenization, filtering and stemming processes. The K-Means algorithm is able to produce an optimal cluster of 6 clusters. Trends in research topics conducted by lecturers at STMIK Primakara include Information System Development and Evaluation, E-Government, Data Mining, Educational Technology, Machine Learning/Artificial Intelligence, and Management and Business.*

Keywords: text mining, clustering, data mining, document

PENDAHULUAN

STMIK Primakara merupakan salah satu Perguruan Tinggi Swasta yang berada di Provinsi Bali. STMIK Primakara memiliki Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) yang merupakan unit Lembaga yang memiliki tugas mengkoordinasikan, memonitor pelaksanaan kegiatan penelitian dan pengabdian, dan mengembangkan bidang penelitian yang dilaksanakan oleh dosen-dosen di STMIK Primakara. Saat ini LPPM belum memiliki sistem informasi dalam pengelolaan data penelitian dan pengabdian kepada masyarakat. Pengarsipan judul dan data penelitian dan pengabdian dosen dilakukan dalam bentuk *file excel*. LPPM tidak mengetahui secara pasti judul penelitian dosen dikarenakan dari pihak LPPM belum mengelompokkan judul penelitian yang dimiliki dosen belum berdasarkan kategori penelitian dan LPPM juga belum mengetahui dosen-dosen yang telah melakukan penelitian dengan judul yang sama. Data-data judul penelitian dosen yang telah ada dapat diidentifikasi kemiripan judul penelitiannya yang dihasilkan dari pengelompokan judul penelitian dosen publikasi LPPM.

Text mining adalah salah satu teknik yang dapat dipergunakan untuk menggali informasi yang tersembunyi dari data yang bersifat *text*. Salah satu metode dalam *text mining* yang dapat dikombinasikan adalah metode *clustering*. Menurut Mushlihudin dan Zahrotun (2017) tujuan utama dari metode *cluster* adalah mengelompokkan sejumlah data/objek ke dalam *cluster* sehingga dalam setiap *cluster* berisi data yang semirip mungkin [1]. Salah satu cara untuk dapat mengetahui kemiripan judul-judul penelitian berdasarkan kategori telah banyak dilakukan pada penelitian sebelumnya. Dilakukan pengelompokan dengan menggunakan metode *K-Means* dengan *Cosine Similarity*.

K-Means *clustering* adalah metode untuk pengelompokkan data non-hierarki yang mempartisi data ke dalam bentuk dua atau lebih kelompok, sehingga data berkarakteristik sama dimasukkan dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan dalam kelompok yang lain [2]. *K-Means* juga merupakan salah satu metode pengelompokan data yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sedemikian rupa agar data yang berkarakteristik sama dimasukan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain [3]. Algoritma K-Means ini dikatakan algoritma

 ISSN 2686-3359 (Online)

yang dapat dengan mudah untuk diimplementasikan, diadaptasi, dan membutuhkan waktu pembelajaran yang relatif cepat [4].

Dengan demikian peneliti tertarik untuk melakukan penelitian dengan judul "*Text Mining Clustering* untuk Pengelompokan Topik Dokumen Penelitian Menggunakan Algoritma *K-Means* dengan *Cosine Similarity*" yang bertujuan untuk mengelompokkan dokumen penelitian sehingga dapat menjadi acuan untuk menentukan roadmap penelitian perguruan tinggi.

TINJAUAN PUSTAKA

2.1 Data Mining

Data mining adalah sekumpulan teknik untuk menemukan pengetahuan yang sebelumnya tidak diketahui dalam basis data yang besar. Pola yang ditemukan tersebut bisa digunakan untuk membantu pengambilan sebuah keputusan [2]. *Data mining* tidak hanya dapat digunakan dalam menemukan pengetahuan atau fenomena baru, melainkan dapat juga untuk meningkatkan pemahaman kita mengenai apa yang kita ketahui. *Data mining* juga dapat didefinisikan sebagai proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidenfikasi informasi yang bermanfaat serta pengetahuan yang terakit dari berbagai database besar atau data *warehouse* [5].

2.2 Text Mining

Text Mining adalah penerapan konsep dari teknik *data mining* untuk dapat mencari pola dalam teks yang memiliki tujuan mencari informasi yang bermanfaat dengan tujuan tertentu [6]. Pada proses *text mining* dilakukan proses *text preprocessing* yang memiliki tujuan untuk menghasilkan sebuah set *term index* yang bisa mewakili dokumen. Komponen dari *text preprocessing* diantaranya sebagai berikut.

a) Tokenization

Proses ini memisahkan setiap kata yang menyusun suatu dokumen. Selain itu juga dilakukan penghilangan angka, tanda baca dan karakter lain selain huruf alphabet [7].

b) Filtering

Filtering merupakan tahap pemilihan kata-kata penting dari hasil token, yaitu kata-kata yang bisa digunakan untuk mewakili isi dari sebuah dokumen. Proses filtering juga biasa disebut sebagai *stopword removal* [5].

c) Stemming

Stemming adalah salah satu proses penghilangan/pemotongan *prefiks* (awalan) dan *sufiks* (akhiran) dari kata dan istilah-istilah dokumen [6].

2.3 K-Means

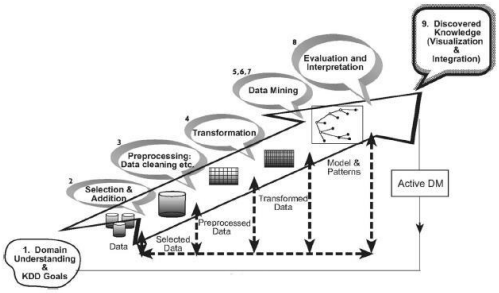
Algoritma K-Means yang digunakan dalam proses pengelompokan topik dokumen adalah sebagai berikut[8]:

a) Tahap diawali dengan menentukan nilai k
b) Selanjutnya Menentukan centroid awal secara acak.
c) Penghitungan jarak data ke *centroid* digunakan metode *Cosine Similarity*. Pengelompokan objek berdasarkan jarak terdekat objek dengan centroid.
d) Kembali Langkah b hingga *centroid* atau pusat cluster tidak mengalami perubahan.

Perhitungan jarak pada algoritma K-means ini dengan *Cosine Similarity*. *Cosine Similarity* adalah sebuah metode pengukuran kemiripan kalimat dengan berdasarkan sudut dua vektor. Kalimat disini dianggap vektor, dengan nilai cosinus sudut antara dua vektor tersebut sebagai parameter jarak. Parameter jarak sebagai parameter kemiripan dua vektor. Semakin dekat jarak antara dua vektor maka semakin mirip dua vektor tersebut[8].

METODE

Metode yang dipilih dan digunakan dalam penelitian ini adalah *Knowledge Discovery in Database* (KDD). KDD adalah sebuah proses komputasi yang didalamnya terdapat penggunaan algoritma-algoritma matematika yang dapat berfungsi untuk mengekstraksi data dan melakukan perhitungan probabilitas kemungkinan tindakan di masa yang akan datang. Hasil dari KDD dapat berupa pengetahuan yang sebelumnya tidak diketahui, potensial, dan bermanfaat [4]. Terdapat 9 tahapan dalam proses KDD seperti yang dapat dilihat pada Gambar 1 berikut.



Gambar 1. Metode Knowledge Discovery in Datbase

Penjelasan dari setiap tahapan dalam proses KDD adalah sebagai berikut:

1) *Domain Understanding* and KDD Goals

Pada tahapan ini dilakukan dengan pemahaman mengenai apa yang akan dilakukan dalam proses pemodelan data mining dan penentuan tujuan dilakukan data mining hingga implementasi ke lingkungan dimana proses penemuan pengetahuan akan berlangsung. Tahapan ini dapat direvisi ketika hasil pemodelan data mining tidak sesuai dengan tujuannya.

 ISSN 2686-3359 (Online)

2) *Selection and Addition*

Tahapan selanjutnya yaitu dilakukan pengumpulan data. Data yang telah dikumpulkan kemudian dilakukan seleksi atribut dan hasil dari seleksi tersebut diintegrasikan menjadi sebuah dataset. Proses pembangunan dataset merupakan suatu proses yang penting karena proses pembelajaran data mining dan penemuan pola baru didasarkan pada dataset yang telah dibentuk.

3) *Preprocessing and Data Cleaning*

Pada tahapan ini dilakukan pembersihan data untuk meningkatkan keandalan data. Pembersihan data dilakukan dengan menangani nilai kosong, menangani baris data yang tidak relevan, dan menghilangkan noise atau outlier. Proses persiapan awal ini dapat melibatkan metode statistik yang kompleks atau menggunakan algoritma data mining yang spesifik.

4) *Transformation*

Tahap selanjutnya, dilakukan pengembangan data sehingga data dapat dipersiapkan dengan lebih baik dan siap untuk dilakukan pemodelan data mining. Hal yang dilakukan untuk mempersiapkan data menjadi lebih baik adalah melakukan reduksi dimensi seperti pemilihan fitur dan ekstrasi sampel data. Selain itu, dapat juga dilakukan transformasi atribut seperti mengubah atribut numerik menjadi atribut diskrit dan transformasi fungsional.

5) *Data Mining*

Tahapan data mining terdiri dari tiga (3) tahapan, yaitu yang pertama pemilihan model data mining, kedua pemilihan algoritma data mining, dan yang ketiga adalah penggunaan data mining. Model data mining yang dipilih adalah model yang sesuai dengan kebutuhan, yaitu klasifikasi, regresi, atau pengelompokan. Algoritma yang dipilih disesuaikan dengan model data mining yang dipilih dengan tetap mempertimbangkan kelebihan dan kekurangan dari algoritma tersebut. Setelah dilakukan pemilihan model dan algoritma data mining, selanjutnya data mining digunakan untuk menemukan pola atau aturan yang baru. Algoritma data mining dimungkinkan untuk digunakan berulang kali hingga memperoleh hasil yang sesuai. Pada penggunaan data mining juga dilakukan pengaturan parameter kontrol algoritma.

6) *Evaluation and Interpretation*

Pada tahapan ini dilakukan evaluasi dan penafsiran pengetahuan yang berupa pola hasil penggunaan data mining. Selain itu, dilakukan pendokumentasian pengetahuan yang ditemukan untuk penggunaan lebih lanjut.

7) *Discovered Knowledge*

Pada tahapan ini, pengetahuan yang ditemukan telah siap untuk diimplementasikan ke dalam sebuah sistem. Tahapan ini menentukan keaktifan keseluruhan proses.

HASIL DAN PEMBAHASAN

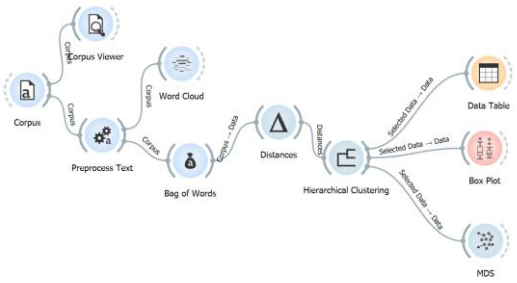
Sumber data yang digunakan dalam penelitian ini meliputi data dokumen penelitian dosen tahun 2019-2022 sejumlah 52 dokumen, dimana atribut yang digunakan adalah topik, judul, dan abstrak. Berikut adalah contoh dataset dalam penelitian ini.

Tabel 1. Dataset Penelitian

No	Judul
1	Rancang Bangun Sistem Informasi Penyedia Jasa Asisten Rumah Tangga di Yayasan Kasih Keluarga
2	Pengembangan Buku Cerita Bergambar untuk Siswa Sekolah Dasar Kelas Rendah
3	Analisis Value Equity, Relationship Equity, dan Brand Equity pada Strategi Pemasaran Objek Wisata Danau Buyan di Provinsi Bali Berbasis Social Media Marketing
4	Pengaruh Mekanisme Good Corporate Governance Terhadap Kinerja Keuangan Perusahaan di Bursa Efek Indonesia
5	Klasifikasi Teks Menggunakan Deep Learning dan Web Scraping
6	Evaluasi Sistem Informasi Kampus (Siska) dengan Teknik Heuristic Evaluation
7	Sistem Pendukung Keputusan Penilaian Kinerja Dosen Menggunakan Metode IT Balanced Scorecard
8	Sistem Informasi Marketplace Penyewaan Kendaraan Berbasis Website di Nusa Penida
9	Analisis Pengaruh Penerimaan Pengguna Sistem E-Learning
10	Dampak Pembelajaran Berbasis Proyek tentang Keterlibatan dan keterampilan Berbicara Siswa
11	Analisa Tingkat Kesiapan Penerapan Keamanan Teknologi Informasi Dalam Pelaksanaan E-Government Studi Kasus Pemerintah Kota Kediri
12	Analisis pengaruh Electronic Word of mouth marketing terhadap minat kunjungan wisatawan ke tempat wisata true bali experience
13	Literature Review Metode Valuasi Bisnis Startup Digital
14	Sentiment analisis menggunakan Deep Learning
15	Analisis Usability pada Sistem Informasi tanaman upacara Hindu Bali
	…
52	Deteksi Jenis Masker Menggunakan Algoritma Convolutional Neural Network

Selanjutnya data tersebut diolah menggunakan tools orange *data mining*. Berikut adalah hasil model yang telah dirancang di *orange data mining*.

 ISSN 2686-3359 (Online)



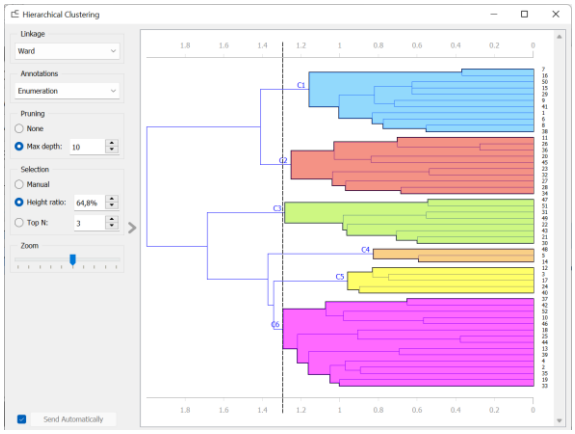
Gambar 2. Model Implementasi Text Mining Clustering pada Orange Data Mining

Pada proses processing dilakukan proses *tokenization, filtering dan stemming.* Hasil dari proses preprocessing seperti yang ditunjukkan pada tabel berikut.

Tabel 2. Hasil *Preprocessing Text*

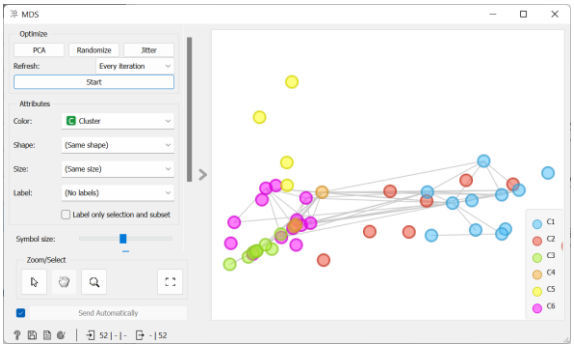
literatur	review	valuasi	bisnis	startup	digital			
ngembang	buku	cerita	gambar	siswa	sekolah	dasar	kelas	nengah
klasifikasi	text	deep	learning	web	scraping			
evaluasi	sistem	informasi	kampus	heuristic				
sentimen	analisis	deep	learning					
…								
deteksi	jenis	masker	convolutional	neural	network			

Berdasarkan model pada Gambar 2 diperoleh hasil cluster sejumlah 6 cluster, yaitu C1, C2, C3, C4, C5, dan C6 seperti yang ditunjukkan pada Gambar berikut.



Gambar 3. Hasil Cluster

Berdasarkan gambar di atas dapat disimpulkan bahwa trend topik penelitian yang dilakukan dosen di STMIK Primakara meliputi Pengembangan dan Evaluasi Sistem Informasi, *E-Government*, *Data Mining*, Teknologi Pendidikan, *Machine Learning/Artificial Intelligence*, serta Manajemen dan Bisnis.Persebaran data tiap cluster ditunjukkan pada gambar berikut.



Gambar 4. Persebaran Cluster

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian pengelompokan tema/topik dokumen Jurusan STMIK Primakara, jumlah kelompok optimal pada k=6. Hasil pengelompokan dapat dijadikan acuan untuk pelabelan kelompok topik. Topik penelitian yang dilakukan dosen di STMIK Primakara meliputi Pengembangan dan Evaluasi Sistem Informasi, *E-Government*, *Data Mining*, Teknologi Pendidikan, *Machine Learning/Artificial Intelligence*, serta Manajemen dan Bisnis. Untuk penelitian selanjutnya dapat dilakukan dengan menggunakan data dan atribut yang lebih beragam, serta melakukan uji evaluasi untuk mendapatkan hasil cluster yang lebih baik.

DAFTAR PUSTAKA

[1] L. Zahrotun, P. Studi, T. Informatika, F. T. Industri, and U. A. Dahlan, “Perancangan Text Mining Pengelompokkan Penelitian Dosen Menggunakan Metode Shared Nearest Neighbor Dengan,” pp. 849–855, 2017.

[2] A. A. I. I. P. Nengah Widya Utami, “Penerapan Data Mining Untuk Mengetahui Pola Pemilihan Program Studi Di Stmik Primakara Menggunakan Algoritma K-Means …,” *J. Teknol. Inf. dan …*, vol. 3, pp. 456–463, 2021, [Online]. Available: http://jurnal.undhirabali.ac.id/index.php/jutik/article/view/1540.

[3] H. Haviluddin, S. J. Patandianan, G. M. Putra, N. Puspitasari, and H. S. Pakpahan, “Implementasi Metode K-Means Untuk Pengelompokkan Rekomendasi Tugas Akhir,” *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 16, no. 1, p. 13, 2021, doi: 10.30872/jim.v16i1.5182.

[4] M. R. Muttaqin and M. Defriani, “Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa,” *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 121–129, 2020, doi: 10.33096/ilkom.v12i2.542.121-129.

[5] M. F. Riyadhi, “Aplikasi Text Mining Untuk Automasi Penentuan Tren Topik Skripsi Dengan Metode K-Means Clustering (Studi

 ISSN 2686-3359 (Online)

Kasus: Prodi Sistem Komputer),” vol. 2, no. 1, pp. 1–6, 2019.

[6] U. A. Dahlan, N. Anggraini, L. Zahrotun, R. Selatan, and U. A. Dahlan, “Pengelompokan Judul Penelitian Dosen Menggunakan Metode K-Means Dengan Cosine Similarity.”

[7] M. Sholehhudin, M. Fauzi Ali, and S. Adinugroho, “Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus : Universitas Brawijaya),” vol. 2, no. 11, pp. 5518–5524, 2018.

[8] M. Kurniana, I. R., Muhima, R.R., Wardana, S., Hakimah, “Penerapan Algoritma K-Means Untuk Pengelompokan Topik Dokumen Studi Kasus:Dokumen Abstrak Skripsi Jurusan Teknik Informatika ITATS Kurniana,” no. 1, pp. 219–224, 2021.