Jurnal Komputer Multidisipliner

Vol.7 No.3, Maret 2024
ISSN: 24559633

PENERAPAN TEXT MINING DALAM KLASTERISASI BUKU DI PERPUSTAKAAN DENGAN ALGORITMA K-MEANS CLUSTERING (STUDI KASUS : PERPUSTAKAAN UNIVERSITAS NAHDLATUL ULAMA KALIMANTAN BARAT)

Dea Nanda[1], **Aditya Pratama**[2], **Awanis Hidayati**[3]
Universitas Nahdlatul Ulama Kalimantan Barat
E-mail: deananda.edu@gmail.com[1], adityapratama@unukalbar.ac.id[2], awanishidayati@gmail.com[3]

Abstrak

Pendataan buku yang ada di perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat masih dilakukan secara manual karena belum terkomputerisasi untuk mengelompokkan setiap buku pada setiap kategori. Permasalahan lain dari mahasiswa yang mengalami kesulitan dalam mencari sumber referensi. Tujuan penelitian untuk menghasilkan sebuah klasterisasi dokumen dalam mengelompokkan buku. Penelitian ini menggunakan jenis penelitian kuantitatif dengan metode pendekatan komparatif. Objek pada penelitian ini adalah klasterisasi buku di perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat. Metode yang digunakan adalah pemodelan klasterisasi menggunakan algoritma K-Means. Metodologi penelitian ini merujuk pada model CRISP-DM yang terdiri dari enam fase, yaitu Business Understanding Phase, Data Understanding Phase, Data Preparation Phase, Modelling Phase, Evaluation Phase, dan Deployment Phase. Proses pengimplementasian text mining dengan model penelitian CRISP-DM pada klasterisasi buku menghasilkan klasterisasi dengan software RapidMiner adalah berupa sebuah knowledge atau pengetahuan terkait klasterisasi buku. Pada penelitian ini diperoleh hasil 4 klaster dengan pembagian setiap dokumen buku yang berbeda-beda. Dalam hal ini, buku-buku yang telah diklasterisasi pada proses clustering terdapat data yang tidak relevan berada pada klaster. Sehingga, data hasil klaster tersebut tidak dapat diterapkan pada perpustakaan, namun proses penelitian yang telah dilakukan dapat menjadi acuan dalam penelitian selanjutnya.

Kata Kunci — Text Mining, Klasterisasi, CRISP-DM, K-Means, Rapidminer

1. PENDAHULUAN

Buku menjadi salah satu cara yang sering digunakan untuk menambah pengetahuan yang luas. Salah satu manfaat dari buku yaitu menyangga alur proses akademik, baik itu di sekolah maupun perguruan tinggi [1]. Buku banyak ditemukan di perpustakaan dengan berbagai jenis tema yang ada. Buku yang terdapat di perpustakaan akan disusun pada setiap rak berdasarkan letak kategori yang telah di tandai dengan label kategori.

Perpustakaan berfungsi sebagai sarana untuk menyediakan informasi kepada setiap orang dengan tujuan meningkatkan kecerdasan masyarakat dan memudahkan penyebaran informasi [2]. Perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat telah banyak berperan dalam mendukung baik dosen maupun mahasiswa dalam proses perkuliahan. Khususnya bagi mahasiswa, perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat berperan besar sebagai tempat mencari referensi belajar maupun membantu dalam penyelesaian setiap tugas mahasiswa.

Buku yang tersedia tidak dapat langsung ditentukan dalam pengelompokan yang sama, dikarenakan tidak semua buku baru yang diperoleh oleh UPT Perpustakaan memiliki kata kunci yang relevan terkait dengan pelabelan kategori pada perpustakaan, maka dengan

begitu petugas perpustakaan dapat mengalami kekeliruan dalam proses pengelompokkan buku tersebut. Selain itu, dari kalangan mahasiswa yang mengalami kesulitan dalam mencari sumber referensi, karena letak buku yang tidak sesuai dengan label kategori yang ada diperpustakaan.

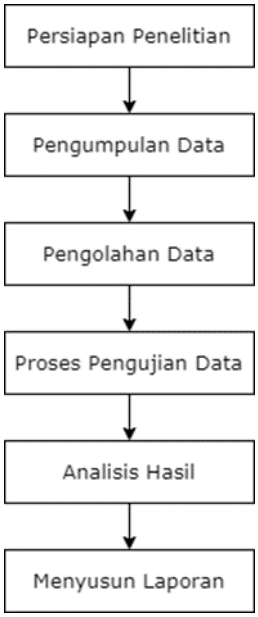
Pengelompokkan buku perpustakaan dapat mempermudah kalangan mahasiswa dalam mencari sumber referensi dengan melihat label kategori yang telah tersedia diperpustakaan. Hal ini dikarenakan dalam proses klasterisasi dapat mengumpulkan data berdasarkan kemiripan tekstual antar tema buku atau berdasarkan kata kunci pada setiap pembahasan yang ada di buku tersebut.

Penelitian terkait yang telah dilakukan oleh Sari dkk, dengan judul "Penerapan Algoritma K-Means Untuk Klastering Data Kemiskinan Provinsi Banten Menggunakan Rapidminer". Dalam penelitian tersebut menjelaskan tentang dimana Masyarakat termiskin tinggal berdasarkan data yang dikumpulkan. Oleh karena itu, studi ini memberikan hasil tiga klaster dengan skor analitis berbeda untuk setiap klaster, yang dapat digunakan sebagai acuan pemerintah dalam menentukan prioritas daerah dengan tingkat kemiskinan terendah [3].

Oleh karena itu, dengan menghasilkan sebuah klasterisasi dokumen dalam mengelompokkan buku yang ada di perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat, dapat memberikan informasi tentang buku sesuai dengan kategorinya. Pada klasterisasi buku memerlukan suatu proses yang dapat diimplementasikan dalam menganalisis data yaitu dengan Text Mining menggunakan algoritma K-Means. Penelitian ini menerapkan software Rapidminer Studio dalam proses klasterisasi buku, dimana akan menampilkan hasil dari pengelompokkan kategori berdasarkan dataset yang diinputkan berupa judul dan kata kunci dari setiap buku. Sehingga akan memudahkan petugas perpustakaan dalam tata letak buku yang sesuai dengan label kategorinya.

2. METODE PENELITIAN

Adapun tahapan yang dilakukan dalam penelitian adalah sebagai berikut:



Gambar 1 Alur Tahapan Penelitian

1) Persiapan Penelitian

Pada tahapan persiapan penelitian dilakukan analisa terhadap kebutuhan yang diperlukan dalam penelitian. Persiapan yang dilakukan adalah menyampaikan perizinan untuk pengambilan data buku di Perpustakaan Universitas Nahdlatul Ulama Kalimantan

Barat sebagai sumber data yang digunakan dalam penelitian. Persiapan alat bantu untuk mengumpulkan data pada penelitian menggunakan perangkat keras scanner dan aplikasi scanner. Perangkat keras yang digunakan adalah printer Epson L3110 dan aplikasi CamScanner.

2) Pengumpulan Data

a) Observasi

Pada tahapan ini dilakukan pengamatan dan pencatatan secara sistematis tentang hal-hal yang berhubungan dengan basis data dokumen buku-buku yang ada di perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat.

b) Studi Pustaka

Pada tahapan ini dilakukan scanning terhadap semua buku yang ada di perpustakaan dari bahan-bahan referensi, arsip, dan dokumen yang berhubungan dengan permasalahan dalam pengumpulan dataset pada penelitian ini.

3) Pengolahan Data

Penelitian ini membagi alur proses pengolahan data menjadi enam tahap, dengan hubungan dari setiap tahap ditunjukkan dengan panah. Metode CRISP-DM ditunjukkan pada Gambar 2.

Gambar 2. Tahapan CRISP-DM

Pada penelitian ini, metodologi yang digunakan adalah model CRISP-DM yang merupakan singkatan dari Cross Industry Standard Process for Data Mining. Penelitian ini terdiri dari enam tahap, yaitu fase pemahaman bisnis, fase pemahaman data, fase persiapan data, fase pemodelan, fase evaluasi, dan fase penyebaran [4].

Fase pemahaman bisnis berfokus pada pemahaman tujuan dan persyaratan proyek dari perspektif bisnis. Kemudian, terjemahkan tujuan Anda ke dalam masalah penambangan data dan kembangkan rencana awal untuk mencapainya. Data yang dikumpulkan terdiri dari judul buku, ringkasan, dan daftar isi untuk mengidentifikasi kata kunci pembahasan pada setiap buku.

Tahap persiapan data mempersiapkan data untuk digunakan pada tahap selanjutnya. Proses ini meliputi pemilihan data kasus dan parameter yang akan dianalisis, transformasi parameter tertentu, dan pembersihan data sehingga siap untuk tahap pemodelan. Dengan kata lain pengelompokan data buku dapat dipahami dari hasil pengelompokan yang dilakukan oleh pengguna.

Tahap pemodelan menentukan metode penambangan data, alat, dan algoritma yang tepat untuk digunakan. Pada tahap evaluasi, hasil pengolahan data yang dihasilkan pada proses pemodelan tahap sebelumnya diinterpretasikan dan dievaluasi dengan tujuan agar model yang diterapkan pada tahap sebelumnya dapat mencapai tujuan yang seharusnya

dicapai pada tahap sebelumnya.

Tahap terakhir adalah implementasi. Dengan kata lain membuat laporan hasil seluruh proses data mining atau evaluasi pada tahap sebelumnya. Seluruh tahapan penelitian dilakukan dalam framework CRISP-DM dengan konsep penemuan pengetahuan dalam database. Selain itu, tujuan awal survei adalah untuk mengolah data historis.

4) Proses Pengujian Data

Berikut metode yang akan diterapkan pada pengujian data antara lain :

a. Text Mining

Text Mining berusaha memecahkan masalah overload informasi text mining merupakan kelanjutan dari data mining atau penemuan pengetahuan dalam database (KDD) yang bertujuan untuk menemukan pola-pola menarik pada database berskala besar [5]. Langkah-langkah pada text mining yang umum adalah tokenisasi, pemfilteran, stemming, penandaan, dan analisis.

b. TF-IDF

Dalam perhitungan bobot menggunakan TF-IDF, hitung jumlah nilai TF kata dengan bobot masing-masing kata. Sedangkan nilai IDF di rumuskan pada persamaan berikut:

$$IDF\ (word) = log\frac{D}{df}$$

Keterangan:

IDF (word) : Nilai IDF dari setiap kata
D : Total dokumen
Df : Total kemunculan kata disemua dokumen

c. Algoritma K-Means Clustering

Secara umum algoritma K-Means Clustering bekerja seperti ini:

1) Tentukan k sebagai jumlah klaster yang diinginkan.
2) Tentukan nilai acak untuk centroid (klaster awal) hingga k.
3) Hitung jarak seluruh data masukan ke setiap centroid menggunakan rumus Euclidean Distance untuk mencari jarak terpendek seluruh data ke centroid. Persamaan rumusnya adalah:

Euclidean Distance:

$$d(ai, bj) = \sqrt{\sum(ai - bj)^2}$$

Keterangan :

ai : Data kriteria
bj : Centroid pada klaster ke-j

4) Mengklasifikasikan data berdasarkan proximity (jarak terpendek) terhadap centroid.
5) Perbarui nilai pusat massa. Melakukan perulangan dari langkah 2 hingga 5, sampai anggota tiap klaster tidak ada yang berubah .

d. RapidMiner

Rapidminer Studio melakukan proses dasar analisis klaster yang bertujuan untuk menemukan pola dalam kumpulan data sehingga dapat dibedakan berdasarkan kategori dengan cara:

1) Temukan klaster yang terpisah satu sama lain.
2) Memahami karakteristik utama klaster.

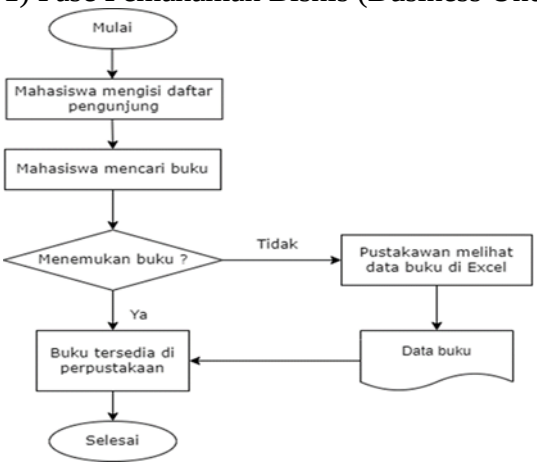
5) Menyusun Laporan

Laporan penelitian ini memuat keseluruhan proses dan tahapan penelitian sampai dengan hasil analisis data klasterisasi buku yang sesuai dengan tujuan penelitian hingga kesimpulan yang diperoleh sebagai acuan dalam pengelompokkan buku di Perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat.

3. HASIL DAN PEMBAHASAN

Hasil Pengujian Data

1) Fase Pemahaman Bisnis (Business Understanding Phase)



Gambar 3. Alur Proses Bisnis Perpustakaan

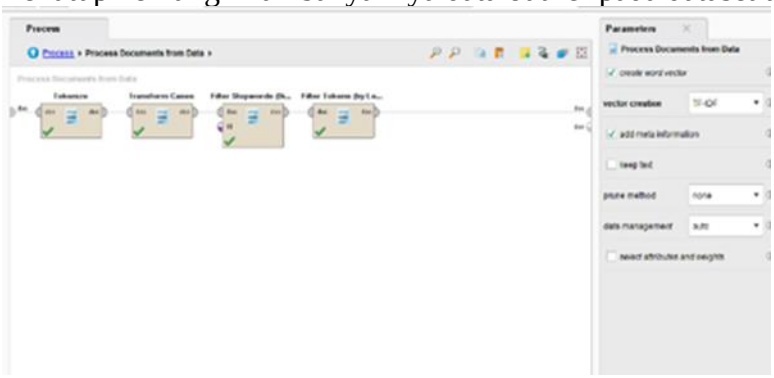
Pada gambar 3. flowchart diatas dapat diketahui alur dari proses bisnis yang dijalankan di Perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat masih secara manual dengan pendataan buku pada MS Excel. Tujuan dari proses bisnis yang dijalankan adalah memberikan layanan terkait dengan peran perpustakaan sebagai tempat dalam mendapatkan berbagai referensi yang di perlukan oleh mahasiswa.

2) Fase Pemahaman Data (Data Understanding Phase)

Fase pemahaman data ini didorong oleh pemahaman akan kebutuhan data terkait pencapaian tujuan bisnis. Data yang dikumpulkan dalam penelitian ini berasal dari koleksi buku, arsip dan dokumen yang bertempat di Perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat.

3) Fase Persiapan (Data Preparation Phase)

Pada persiapan data ini dataset akan dianalisis dengan proses preprocessing untuk mencari kesalahan atau missing eror pada data, dikarenakan sumber data yang beragam sehingga tidak menutup kemungkinan banyaknya data outlier pada dataset tersebut.

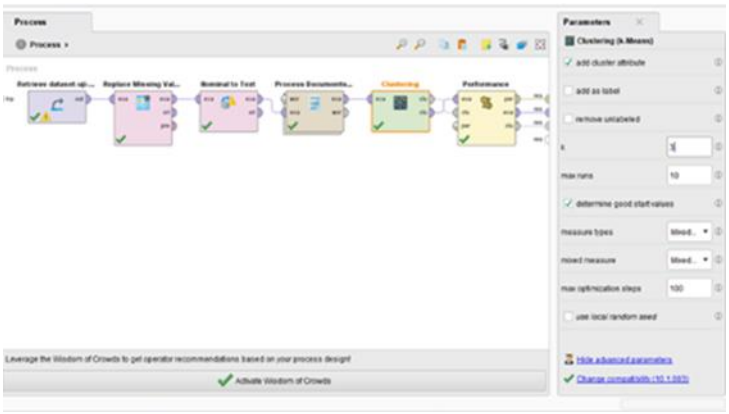


Gambar 4. Proses Preprocessing

Preprocessing adalah beberapa tahapan yang dilakukan pada pengolahan data agar menjadi lebih baik dan siap untuk digunakan saat proses pengujian data.

4) Fase Pemodelan (Modelling Phase)

Pada penelitian ini menerapkan algoritma K-Means Clustering, maka diperlukannya validasi data hasil klaster karena terdapat banyak jenis atribut pada dataset. Proses klasterisasi dengan algoritma K-Means menggunakan parameter indeks jarak Euclidean Distance dengan performance Indeks Davies Bouldin yang secara default tersedia di software Rapidminer.



Gambar 5. Proses Clustering

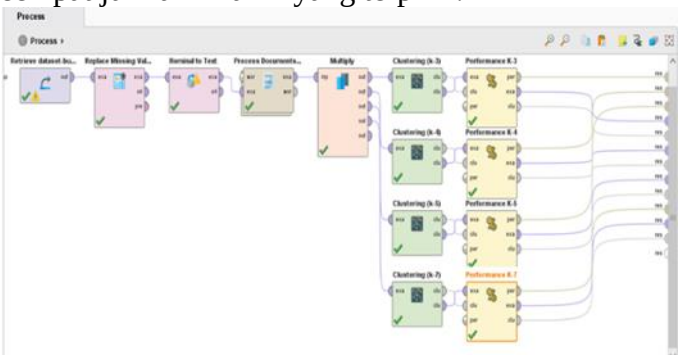
Proses clustering dengan menerapkan operator k-means dan Index Davies Bouldin pada software Rapidiner. Berikut tabel data klaster hasil uji klasterisasi pada algoritma K-Means Clustering dengan performance Davies Bouldin Index menggunakan klaster model sebagai berikut :

Tabel 1. Hasil Uji Clustering

Jumlah (k)	Klaster Model	DBI	Rata-Rata *Centroid*
2	Klaster 0 = 584 anggota Klaster 1 = 584 anggota	-0,501	-28556,670
3	Klaster 0 = 390 anggota Klaster 1 = 388 anggota Klaster 2 = 390 anggota	-0.500	-12690,818
4	Klaster 0 = 290 anggota Klaster 1 = 293 anggota Klaster 2 = 293 anggota Klaster 3 = 292 anggota	-0.500	-7126.052
5	Klaster 0 = 232 anggota Klaster 1 = 235 anggota Klaster 2 = 235 anggota Klaster 3 = 234 anggota Klaster 4 = 232 anggota	-0.500	-4565,737
6	Klaster 0 = 193 anggota Klaster 1 = 195 anggota Klaster 2 = 197 anggota Klaster 3 = 194 anggota Klaster 4 = 192 anggota Klaster 5 = 197 anggota	-0.501	-3170,926
7	Klaster 0 = 166 anggota Klaster 1 = 168 anggota Klaster 2 = 166 anggota Klaster 3 = 168 anggota Klaster 4 = 163 anggota Klaster 5 = 169 anggota Klaster 6 = 170 anggota	-0.500	-2327,846
8	Klaster 0 = 147 anggota Klaster 1 = 147 anggota Klaster 2 = 146 anggota Klaster 3 = 143 anggota Klaster 4 = 146 anggota Klaster 5 = 147 anggota	-0.501	-1781,259

	Klaster 6 = 148 anggota Klaster 7 = 144 anggota		
9	Klaster 0 = 135 anggota Klaster 1 = 128 anggota Klaster 2 = 125 anggota Klaster 3 = 132 anggota Klaster 4 = 127 anggota Klaster 5 = 131 anggota Klaster 6 = 125 anggota Klaster 7 = 136 anggota Klaster 8 = 129 anggota	-0.501	-1412.430
10	Klaster 0 = 118 anggota Klaster 1 = 119 anggota Klaster 2 = 114 anggota Klaster 3 = 118 anggota Klaster 4 = 118 anggota Klaster 5 = 121 anggota Klaster 6 = 116 anggota Klaster 7 = 115 anggota Klaster 8 = 114 anggota Klaster 9 = 115 anggota	-0.502	-1145.005

Berdasarkan informasi tersebut, maka didapatkan informasi mengenai jumlah k dengan performance terbaik yaitu K-3, K-4, K-5, dan K-7 dengan performance Indeks Davies Bouldin -0,500. Oleh karena itu dilakukan uji ulang kembali dalam menentukan performance dari keempat jumlah nilai K yang terpilih.



Gambar 6. Proses Uji Perbandingan Clustering

Berikut tabel data klaster hasil uji perbandingan performance clustering pada algoritma K-Means dengan performance Davies Bouldin Index menggunakan klaster model sebagai berikut :

Tabel 2. Hasil Uji Perbandingan Clustering

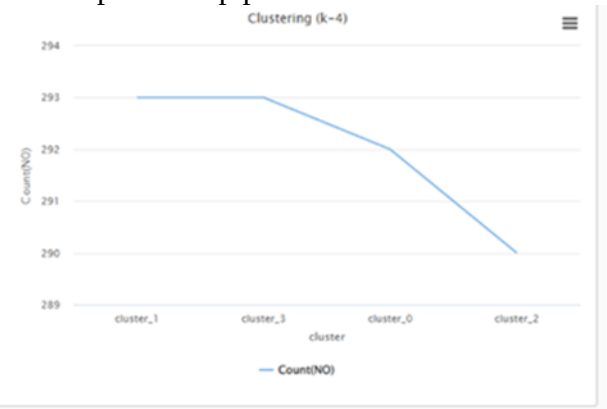
Jumlah (k)	Klaster Model	DBI	Rata-Rata *Centroid*
3	Klaster 0 = 390 anggota Klaster 1 = 390 anggota Klaster 2 = 388 anggota	-0,500	-12690,810
4	Klaster 0 = 292 anggota Klaster 1 = 293 anggota Klaster 2 = 290 anggota Klaster 2 = 293 anggota	-0.500	-7126.043
5	Klaster 0 = 233 anggota Klaster 1 = 234 anggota Klaster 2 = 234 anggota	-0.500	-4565,263

	Klaster 3 = 235 anggota Klaster 3 = 232 anggota		
7	Klaster 0 = 168 anggota Klaster 1 = 166 anggota Klaster 2 = 169 anggota Klaster 3 = 165 anggota Klaster 4 = 166 anggota Klaster 4 = 168 anggota Klaster 4 = 166 anggota	-0.500	-2327,549

Berdasarkan hasil uji perbandingan performance pada tabel diatas, maka dapat ditentukan jumlah k terbaik pada clustering dari perolehan selisih perbandingan paling tinggi yaitu pada k-4 dengan selisih dari k-3 adalah -5564,767.

5) Fase Evaluasi (Evaluation Phase)

Berdasarkan hasil pengujian yang telah dilakukan dengan menggunakan algoritma K-Means Clustering pada software RapidMiner, maka diperoleh pemodelan data berupa klasterisasi yang menampilkan hasil analisis berupa visualisasi data klaster dan grafik yang menunjukkan jumlah data pada setiap klaster, sehingga sudah memenuhi tujuan penelitian seperti yang telah ditentukan pada tahap pemahaman bisnis.



Gambar 7. Grafik Klaster

Pada gambar grafik klaster diatas, dapat dilihat pengemlompokkan jumlah setiap klaster tidak sama, dikarenakan tingkat kemiripan dan jarak terdekat antar atribut kata kunci yang telah di proses pada saat klasterisasi.

6) Fase Penyebaran (Deployment Phase)

Fase penyebaran merupakan tahap pembuatan laporan hasil penelitian, pembuatan laporan hasil dapat dilakukan setelah selesai fase evaluasi model pengelompokan (clustering) dan mendapatkan hasil yang sesuai dengan tujuan pada proses pemahaman bisnis.

Analisis Hasil

Hasil akhir dari proses klasterisasi dengan bantuan RapidMiner Studio ini adalah berupa sebuah knowledge atau pengetahuan terkait klasterisasi buku di perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat. Perpustakaan memiliki 4 klaster dengan label kategori yang dapat ditentukan oleh pakar ilmu atau ditentukan oleh kepala UPT Perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat.

Hasil klasterisasi dari penelitian dengan menerapkan algoritma K-Means Clustering telah mencapai tujuan dari penelitian yang dilakukan terhadap buku yang ada di Perpustakaan Universitas Nahdlatul Ulama Kalimantan Barat. Namun, dari hasil klasterisasi tersebut belum dinyatakan berhasil untuk diterapkan pada tata letak pengelompokkan buku, dikarenakan ada data buku yang masih menyebar pada klaster yang kata kuncinya tidak memiliki kemiripan atau kesamaan sehingga data yang tidak relevan masih terdapat dalam klaster.

Berdasarkan analisis tersebut, maka penelitian ini sudah sesuai dengan tujuan dari penelitian yang dilakukan, dengan hasil akhir yang belum bisa untuk diterapkan pada perpustakaan. Namun, dalam hal ini dapat menjadi acuan dalam penelitian selanjutnya agar lebih memperhatikan terhadap sumber data yang digunakan dalam jumlah yang besar.

4. KESIMPULAN

Pada penelitian ini diperoleh hasil 4 klaster dengan pembagian setiap dokumen buku yang berbeda-beda. Pengelompokkan buku ini bertujuan dalam menangani kekeliruan yang terjadi di Perputakaan Univitas Nahdlatul Ulama Kalimantan Barat dalam meletakkan buku yang kurang tepat pada rak yang telah dilabeli oleh petugas perpustakaan. Namun, dari hasil analisis terhadap proses klasterisasi yang dilakukan dengan metode Text Mining menerapkan algoritma K-Means Clustering pada software RapidMiner masih belum dinyatakan berhasil.

Dalam hal ini, buku-buku yang telah diklasterisasi pada proses clustering terdapat data yang tidak relevan berada pada klaster. Sehingga, data klaster tersebut tidak dapat diterapkan pada perpustakaan, namun proses penelitian yang telah dilakukan dapat menjadi acuan dalam penelitian selanjutnya agar mendapatkan hasil yang lebih baik.

DAFTAR PUSTAKA

[1] Setiawan, A., Astuti, I. F., & Kridalaksana, A. H. (2016). "Klasifikasi dan pencarian buku referensi akademik menggunakan metode naïve bayes classifier (nbc) (studi kasus: perpustakaan daerah provinsi kalimantan timur)". Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer, 10(1), 1-10.

[2] Riyanto, A. A., Daryanto, D., & Abdurrahman, G. (2022). Text Mining Untuk Clustering Buku Di Perpustakaan Menggunakan Metode K-Means. National Multidisciplinary Sciences, 1(6), 835-8.

[3] Sari, Y. R., Sudewa, A., Lestari, D. A., & Jaya, T. I. (2020). Penerapan Algoritma K-Means Untuk Clustering Data Kemiskinan Provinsi Banten Menggunakan Rapidminer. CESS (Journal of Computer Engineering, System and Science), 5(2), 192-198.

[4] Damayanti, S. E., & Kuswayati, S. (2006). Analisis Dan Implementasi Framework Crisp-Dm (Cross Industry Standard Process for Data Mining) Untuk Clustering Perguruan Tinggi Swasta. Jurnal STT Bandung.

[5] Februariyanti, H., & Santoso, D. B. (2017). Hierarchical agglomerative clustering untuk pengelompokan skripsi mahasiswa.

[6] Astuti, S. (2020). Algoritma K-Means Dalam Menentukan Penerima Beasiswa UPZ (Unit Pengumpulan Zakat) Pada Mahasiswa UIN Sumatera Uara Medan (Doctoral dissertation, Universitas Islam Negeri Sumatera Utara Medan).