Identifikasi Divisi Pada Struktur Organisasi Pondok Pesantren Berdasarkan Standar Sekolah Berasrama Menggunakan Metode *Semantic Similarity*

Muhammad Ainul Yaqin[1,*], Munajatul Azizah[2], Syahiduz Zaman[3]

Teknik Informatika, Universitas Islam Negeri Maulana Malik Ibrahim, Indonesia

[1]yaqinov@ti.uin-malang.ac.id; [2]16650058@student.uin-malang.ac.id; [3]syahid@ti.uin-malang.ac.id

*corresponding author

INFO ARTIKEL

Sejarah Artikel
Diterima: 19 Desember 2022
Direvisi: 7 Januari 2023
Diterbitkan: 28 April 2023

Kata Kunci
K-Means
Pondok Pesantren
Semantic Similarity
Struktur Organisasi

ABSTRAK

Sebagai salah satu lembaga pendidikan, pondok pesantren memerlukan adanya pedoman atau standar sebagai acuan dalam pengembangan proses bisnisnya. *National Minimum Standards for Boarding Schools* merupakan salah satu standar nasional yang diperuntukkan bagi sekolah berasrama atau pondok pesantren. Standar ini berisikan 52 poin yang akan mengatur proses bisnis pada sebuah pondok pesantren. Sebagai sebuah organisasi pondok pesantren memerlukan struktur organisasi agar proses bisnis pada pondok pesantren dapat berjalan. Karena perannya dalam mengatur interaksi antar unit kerja atau divisi, serta dalam pembagian dan koordinasi tanggung jawab dan wewenang yang lebih efektif, struktur organisasi menjadi sangat penting. Pada penelitian ini, untuk menentukan divisi pada struktur organisasi dilakukan dengan menggunakan metode *semantic similarity*. Dilakukan tahap *text preprocessing* pada data standar. Kemudian dilakukan pembobotan untuk menentukan kata kunci pada setiap standar, menggunakan *Term Frequency - Inverse Document Frequency* (TF-IDF). Kata kunci yang diperoleh dibandingkan tingkat kemiripannya antar standar menggunakan metode *semantic similarity*. Tahap selanjutnya adalah proses clustering terhadap data matrik hasil *semantic similarity* menggunakan algoritma *k-means*. Untuk mengetahui berapa cluster yang optimal, digunakan perhitungan SSE. Setelah data cluster diperoleh, dilakukan identifikasi divisi yang akan dibentuk. Penelitian ini menggunakan dua skenario, yakni *clustering* menggunakan data kemiripan semantik antar kata dan menggunakan data kemiripan semantik antar standar. Skenario pertama, diperoleh selisih nilai SSE tertinggi yaitu 0,4245 pada percobaan ke-2 dengan K=6. Sehingga diperoleh 6 *cluster* yang kemudian menjadi divisi kesantrian, divisi keamanan, divisi pendidikan, divisi sarana dan fasilitas, divisi kepegawaian, dan divisi kesejahteraan. Skenario kedua menghasilkan selisih nilai SSE tertinggi yaitu 0,0011 pada percobaan ke-3 dengan K=7. Sehingga diperoleh 7 *cluster* yang kemudian menjadi divisi kesantrian, divisi keamanan, divisi kepegawaian, divisi sarana dan fasilitas, divisi kesejahteraan, divisi asrama, dan divisi humas.

PENDAHULUAN

Pondok pesantren terkenal sebagai salah satu lembaga pendidikan agama Islam di Indonesia. Tercatat sebanyak 26.975 pondok pesantren yang telah terdaftar di Kementerian Agama [1]. Pondok pesantren terkenal dengan pendidikan agama Islam. Umumnya, siswa yang belajar di sebuah pondok pesantren dikenal dengan nama santri dan tinggal atau bermukim di sebuah asrama atau pondok di bawah naungan seorang guru yang disebut

sebagai Kiai [2]. Saat ini pondok pesantren tidak hanya menyampaikan ilmu pengetahuan agama, namun mengajarkan ilmu pengetahuan umum lainnya mengikuti perkembangan zaman. Pondok pesantren tersebar di 34 provinsi dengan jumlah pondok pesantren paling banyak terdapat pada provinsi Jawa Barat dengan total 8.343 pondok pesantren [1].

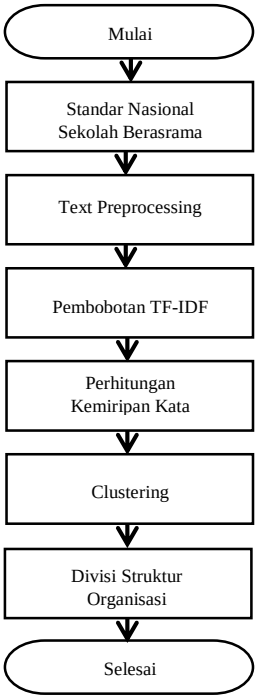
Suatu perencanaan tidak lepas dari aturan-aturan atau standar yang dijadikan sebagai acuan atau pedoman. Standar diartikan sebagai ukuran tertentu yang dijadikan patokan, dalam Kamus Besar Bahasa Indonesia. Dalam hal perencanaan sebuah pondok pesantren agar dapat mencapai tujuan atau memenuhi visi dan misi, diperlukan suatu standar pedoman dalam pengembangannya. *National Minimum Standards for Boarding Schools* merupakan salah satu standar nasional yang diperuntukkan bagi sekolah berasrama atau pondok. Standar ini diterbitkan oleh Menteri Kesehatan dan Layanan Sosial Welsh Assembly Government, di UK. *National Minimum Standards for Boarding Schools* mengatur segala aktivitas dan kebutuhan santri selama di pondok, mulai dari fasilitas tempat tidur hingga kegiatan di waktu luang [3].

Pondok pesantren sebagai sebuah organisasi memerlukan adanya struktur organisasi. Struktur organisasi dibentuk agar proses bisnis pada pondok pesantren dapat berjalan. Struktur organisasi sangat penting karena mengatur interaksi antara unit kerja serta membantu membagi dan mengkoordinasikan tanggung jawab dan wewenang dengan lebih efektif. Struktur organisasi membantu mempermudah pola koordinasi yang dijalankan untuk mewujudkan tujuan yang telah ditetapkan berdasarkan standar sekolah asrama. Oleh sebab itu diperlukan sebuah struktur organisasi yang sesuai dengan standar sekolah asrama [4]. Dalam pengembangan struktur organisasi perlu diketahui bagian-bagian atau unit kerja apa saja yang dibutuhkan agar standar sekolah berasrama dapat terpenuhi. TF-IDF merupakan metode yang digunakan untuk mengetahui frekuensi kemunculan kata dari sebuah dokumen. Dengan metode ini dapat diketahui kata apa saja yang sering muncul pada dokumen standar sekolah berasrama. Kata-kata yang muncul merepresentasikan seberapa penting kata tersebut dalam dokumen standar sekolah berasrama [5]. Data kata yang memiliki frekuensi kemunculan yang tinggi akan dilakukan pengelompokan atau *clustering*.

Data dengan karakteristik yang sama dikelompokkan ke dalam satu "wilayah", sedangkan data dengan karakteristik yang berbeda dikelompokkan ke dalam "wilayah" yang berbeda dengan menggunakan teknik analisis data yang disebut *clustering*[6]. Pengelompokan dilakukan berdasarkan pada tingkat kemiripan katanya. Proses perhitungan kemiripan kata dihitung dengan menggunakan metode kemiripan semantik atau *semantic similarity* [7]. Metode kemiripan semantik atau *semantic similarity* digunakan untuk menghitung kemiripan antar kata tidak hanya melihat dari susunan katanya namun juga akan mempertimbangkan arti atau makna dari kata tersebut[8]. Data matrik perbandingan kemiripan semantik akan dikelompokkan menggunakan metode *K-means clustering* untuk dapat diketahui divisi-divisi atau unit kerja apa saja yang dibutuhkan untuk membangun struktur organisasi.

METODE

Tahapan-tahapan yang dilakukan dapat dilihat dalam Gambar 1. Tahap awal yang dilakukan dalam penelitian ini adalah proses *text preprocessing*, yang kemudian dilanjutkan dengan pembobotan kata untuk mengetahui frekuensi kemunculan kata dalam dokumen standar. Sehingga diperoleh kata dengan nilai frekuensi tertinggi yang akan menjadi kata kunci pada masing-masing standar. Kata kunci yang diperoleh dihitung kemiripannya. Perhitungan kemiripan antar kata kunci dilakukan dengan metode kemiripan semantik. Data matrik kemiripan semantik yang diperoleh selanjutnya dilakukan proses *clustering*. Proses *clustering* bertujuan untuk menentukan banyaknya unit atau divisi dalam struktur organisasi.



Gambar 1. Alur Penelitian

Text Preprocessing

Text preprocessing adalah proses memilih data teks dan mengaturnya agar lebih mudah untuk diproses. Terdapat 4 (empat) tahap pada *text preprocessing* yakni tahapan *case folding*, *tokenizing*, *stopword removal/filtering* dan *stemming* [9]. 1) Tahap case folding, disebabkan data yang dimiliki tidak selalu menggunakan huruf kapital, maka dalam tahap *case folding* akan terjadi proses perubahan huruf kecil yang awalnya berupa huruf kapital. Hal ini bertujuan untuk menyamaratakan penggunaan huruf. Pada tahap ini pula dilakukan penghapusan angka maupun simbol khusus, contoh tanda baca dan lainnya [9]. 2) *Tokenizing*, adalah tahap memecah sebuah kalimat berdasarkan kata penyusunnya. Proses pemecahan kalimat menjadi kata disebut dengan token [9]. 3) Proses pemilihan kata-kata penting yang diperoleh dari hasil *tokenizing* disebut *filtering*. *Stopwords* adalah kata-kata umum yang sering digunakan tetapi tidak memiliki makna yang nyata. *Stopword* dapat berupa kata hubung misal, dan, setelah, yang, di dan lain sebagainya [9]. Dan 4) *Stemming* adalah tahap mengembalikan kata ke bentuk aslinya setelah menambahkan akhiran atau awalan. Prosedur ini digunakan untuk mengurangi jumlah berbagai indeks data. Pengelompokan kata lain yang memiliki akar kata dan arti yang sama tetapi bentuk yang berbeda juga dilakukan dalam tahap *stemming* [9].

Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency - Inverse Document Frequency adalah gabungan dari 2 (dua) metode yakni *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). Hubungan antara kata dan dokumen diberi bobot menggunakan TF-IDF. Metode ini menggabungkan frekuensi kemunculan suatu kata dalam dokumen tertentu dan frekuensi kebalikan dari dokumen yang mengandung kata tersebut untuk menentukan bobot [5]. Model pembobotan ini diperlukan untuk membentuk vektor dari data text dengan menggunakan Persamaan (1):

$$Wdt = tf\,dt * Id\,ft \qquad (1)$$

Keterangan:
Wdt = bobot ke -d dokumen terhadap ke -t data
tf dt = jumlah kata dicari dalam sebuah dokumen
Id ft = inverse document frequency (log N/df)
N = banyak dokumen
df = banyak dokumen yang memiliki kata yang dicari

Frekuensi sebuah kata muncul dalam dokumen memberi tahu seberapa penting kata tersebut bagi dokumen itu. Jumlah dokumen yang menggunakan kata mengungkapkan seberapa sering kata itu digunakan. Akibatnya, jika sebuah kata sering muncul dalam dokumen maka frekuensi kata tersebut besar, dan hubungan antara kata dan dokumen akan sangat signifikan [5].

Semantic Similarity

Kemiripan semantik atau *semantic similarity* merupakan teknik menghitung kemiripan antar kata. Dalam membandingkan dua kata, sangat penting untuk mempertimbangkan tingkat kesetaraan antara kata-kata daripada mengasumsikan bahwa mereka setara secara tulisan[8]. Karena kedua istilah tersebut adalah sinonim, bahkan kata-kata yang tampak berbeda mungkin memiliki arti yang sama., contoh kata "baik" dan "bagus". Metode kemiripan semantik digunakan untuk menghitung kemiripan antar kata tidak hanya melihat dari susunan katanya namun juga akan mempertimbangkan arti atau makna dari kata tersebut.

$$sim_{wup}(s1,s2) = \frac{2\ x\ depth(LCS)}{(depth(s1) + depth(s2))} \quad (2)$$

Wu palmer merupakan salah satu algoritma dalam perhitungan kemiripan semantik. Perhitungan dalam algoritma *wu palmer* menggunakan Persamaan (2). Untuk menghitung LCS, *Wu Palmer* terlebih dahulu menentukan jalur terpendek dari setiap *concept* dan kemudian menggabungkan jalur tersebut. Menemukan *Lowest Common Subsumer* (LCS) dengan mencari *sense* yang dihasilkan dari dua jalur yang terhubung[8]. Sinonim, hipernim, dan akronim dengan konotasi semantik serupa akan dicari dalam algoritma *Wu Palmer*.[10]. Skor yang dihasilkan berada dalam rentang 0 (nol) sampai 1 (satu) dengan skor error -1 (negatif satu). Semakin besar skor menandakan kata tersebut memiliki tingkat kemiripan yang baik[18].

Clustering

Data disusun menjadi beberapa *cluster* atau kelompok dengan menggunakan metode *clustering*. Sehingga data di dalam *cluster* akan menjadi yang paling mirip, sedangkan data antar *cluster* akan menjadi yang paling tidak mirip [11]. Data dengan karakteristik yang sama dikelompokkan ke dalam satu "wilayah", sedangkan data dengan karakteristik yang berbeda harus dikelompokkan ke dalam "wilayah" yang berbeda dengan menggunakan teknik analisis data yang disebut *clustering* [6]. *Hierarchical clustering* dan *non-hierarchical clustering* adalah dua jenis teknik pengelompokan data yang digunakan dalam *data mining*. Data yang dikelompokkan menggunakan *hierarchical clustering* akan membangun hierarki dendogram dengan data serupa ditempatkan dalam hierarki dekat dan data berbeda dalam hierarki jauh [6]. *Non-hierarchical clustering* juga dikenal sebagai *partitional clustering*, adalah teknik pengelompokan yang membagi data menjadi beberapa kelompok tanpa menggunakan kerangka hierarki untuk menghubungkan satu kluster ke kluster berikutnya.

Setiap *cluster* dalam pendekatan partisi ini memiliki *centroid* atau titik pusat *cluster*. Tujuan utama dari fungsi metode ini adalah untuk mengurangi jarak antara setiap *centroid* dan semua data [12].

K-means adalah teknik pengelompokan yang menggunakan *partitional clustering* atau *non-hierarchical clustering* [12]. Adapun tahapan-tahapan dalam proses *K-means* adalah berikut:

1. Berikan nilai K atau banyak *cluster*
2. Tentukan titik pusat cluster atau *centroid*.
3. Tentukan jarak antara setiap titik data dan setiap *centroid*. Gunakan rumus *Euclidean Distance* untuk menghitung jarak

$$\text{dist}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

dengan;
n = banyaknya data
i = indeks data

4. Setiap centroid akan mengambil data yang memiliki jarak terdekat dengannya
5. Perbaiki titik *centroid.* Hitung kembali nilai rata-rata dari seluruh data yang masuk dalam klasternya. Sehingga centroid akan berubah posisi mengikuti nilai barunya.
6. Hitung kembali jarak centroid dengan data-data terbaru yang masuk dalam klasternya.
7. Ulangi langkah 5 dan 6 sampai titik pusat cluster atau centroid tidak berpindah posisi atau memiliki perubahan yang sangat kecil.

Algoritma *K-means* dianggap mudah digunakan dan dijalankan, relatif lebih cepat, serba bisa, dan dapat digunakan di banyak kasus [17].

Untuk menentukan banyaknya K yang optimal, dilakukan perhitungan dengan metode *elbow*. Dengan membandingkan persentase hasil dari jumlah *cluster* yang akan membentuk siku pada suatu titik, metode *elbow* dapat digunakan untuk mendapatkan informasi tentang jumlah *cluster* yang ideal [14]. Grafik dapat digunakan sebagai sumber informasi untuk menampilkan hasil berbagai persentase dari setiap nilai *cluster*. Jumlah klaster paling baik jika perpotongan nilai klaster pertama dan nilai klaster kedua nilainya berkurang paling banyak[14]. Perhitungan perbandingan dilakukan dengan mencari nilai SSE atau *sum of squared error* terhadap setiap *cluster* di setiap percobaan.

$$SSE = \sum_{K=1}^{K} \sum_{x_i s S_k} \|X_i - C_k\|^2 \quad (4)$$

Keterangan;
K = banyak klaster
Xi = data ke-i
Ck = *centroid* klaster

Perhitungan SSE menggunakan Persamaan (4) dengan langkah perhitungan adalah sebagai berikut:

1. Tentukan *value* pada data Xi. Pada penelitian ini, *value* merupakan rata-rata *centroid* pada masing-masing *cluster* di setiap percobaan
2. Hitung rata-rata dari *value* yang telah diperoleh
3. Hitung selisih value dengan rata-ratanya
4. Kuadratkan hasil selisih
5. Jumlah hasil kuadrat selisih

Nilai SSE merupakan nilai akhir dari penjumlahan hasil kuadrat selisih data ke-i dengan rata-ratanya. Hasil dari perhitungan nilai SSE direpresentasikan ke dalam suatu grafik untuk melihat perubahan nilai SSE yang paling besar. Titik dimana terjadi perubahan nilai SSE paling besar menunjukkan nilai yang optimal untuk dijadikan sebagai nilai K pada proses *clustering*.

Struktur Organisasi

Pembagian kegiatan di antara berbagai sub-unit, alokasi tanggung jawab di antara pos-pos administrasi, pola tugas dan keterkaitan tugas yang telah ditetapkan, dan jaringan komunikasi formal membentuk struktur organisasi [15]. Struktur organisasi, yang sering diwakili oleh bagan organisasi, adalah pola formal pengelompokan orang dan pekerjaan, pola aktivitas formal, dan interaksi antara sub-unit organisasi yang berbeda [16]. Untuk menciptakan bagaimana suatu organisasi dapat berfungsi dan membantunya mencapai tujuan yang dinyatakan untuk masa depan, hierarki didefinisikan dalam organisasi menggunakan struktur organisasi [4]. Struktur organisasi fungsional akan dikembangkan pada penelitian ini. Struktur organisasi fungsional adalah struktur yang komponen atau unit kerjanya berbagi pengetahuan yang sama. Sehingga memiliki tugas dan beban kerja yang sama. Klasterisasi dilakukan dengan memasukkan nilai K yang diperoleh dari metode *elbow*. Proses *clustering* menghasilkan kelompok-kelompok atau *cluster* yang memiliki kemiripan. Bagian-bagian *cluster* yang dihasilkan akan menjadi bagian atau divisi dalam struktur organisasi.

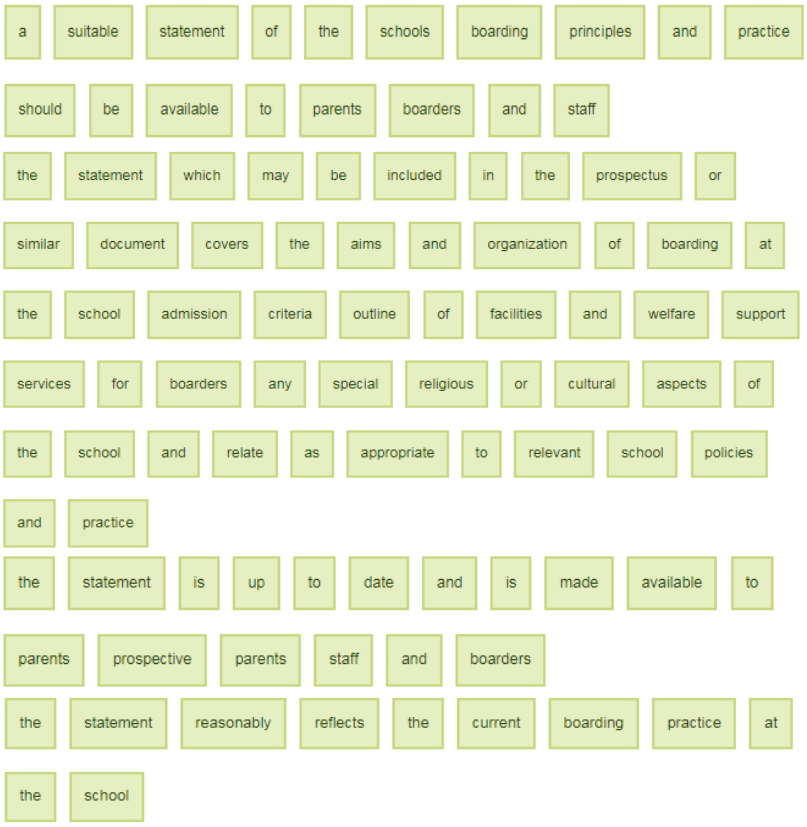
HASIL DAN PEMBAHASAN

Text Preprocessing

Tahap *text preprocessing* dilakukan terhadap dokumen *National Minimum Standards for Boarding Schools* yang berisikan 52 poin standar sekolah berasama. 52 poin standar melalui tahapan-tahapan yang ada dalam *text preprocessing*. Proses *case folding* menghasilkan data dalam Tabel 1. Proses *tokenizing* menghasilkan data token seperti pada Gambar 2. Data kata dalam token hasil proses *tokenizing* akan melewati tahap *filtering* sehingga dihasilkan data seperti pada Tabel 2. Data pada Tabel 2 diproses dalam tahap *stemming*, sehingga diperoleh data dalam Tabel 3.

Tabel 1. Hasil tahap *case folding* standar 1

No.	Hasil tahap *case folding*
1	a suitable statement of the schools boarding principles and practice should be available to parents boarders and staff
2	the statement which may be included in the prospectus or similar document covers the aims and organization of boarding at the school admission criteria outline of facilities and welfare support services for boarders any special religious or cultural aspects of the school and relate as appropriate to relevant school policies and practice
3	the statement is up to date and is made available to parents prospective parents staff and boarders
4	the statement reasonably reflects the current boarding practice at the school



Gambar 2. Hasil *tokenizing* standar 1

Tabel 2. Hasil tahap *stopword removal* standar 1

No.	Hasil tahap *stopword removal*
1	suitable statement schools boarding principles practice parents boarders staff
2	statement included prospectus document covers aims organization boarding school admission criteria outline facilities welfare support services boarders special religious cultural aspects school relate relevant school policies practice
3	statement parents prospective parents staff boarders
4	statement reflects current boarding practice school

Tabel 3. Hasil tahap *stemming* standar 1

No.	Hasil tahap *stemming*
1	suitable statement school boarding principle practice parent boarder staff
2	statement include prospectus document cover aim organization boarding school admission criteria outline facility welfare support service boarder special religious cultural aspect school relate relevant school policy practice
3	statement parent prospective parent staff boarder
4	statement reflect current boarding practice school

Pembobotan TF-IDF

Setiap standar yang telah melalui tahap *preprocessing* dilakukan perhitungan TF-IDF. Tahapan ini menghitung jumlah setiap kata yang ada pada masing-masing standar dengan dokumen standar tersebut. Perhitungan TF-IDF dilakukan dengan menggunakan Persamaan (1). Hasil yang didapat akan menunjukkan kata apa saja yang memiliki nilai TF-IDF terbesar. Tahap TF-IDF menghasilkan data dalam Tabel 4. Perhatikan dalam Tabel 4 bahwa setiap kata memiliki bobot atau frekuensi kemunculannya terhadap suatu dokumen. Data dalam Tabel 4 kemudian diurutkan berdasarkan nilai TF x IDF yang terbesar. Kata dengan nilai terbesar menunjukkan bahwa kata tersebut dinilai penting dalam standar tersebut. Kata-kata penting ini disebut dengan kata kunci atau *keyword*. Tahapan TF-IDF dilakukan terhadap setiap standar. Sehingga akan diperolah kata kunci pada setiap standar. Pada tahapan ini dihasilkan kata kunci seperti dalam Tabel 5.

Tabel 4. Hasil TF-IDF standar 1

No.	Token	TF	IDF	TF × IDF
1	suitable	0.0208	0.602	0.013
2	statement	0.0833	0.602	0.05
3	school	0.1042	0.602	0.063
4	boarding	0.0625	0.602	0.038
5	principle	0.0208	0.602	0.013
6	practice	0.0625	0.602	0.038
7	parent	0.0625	0.602	0.038
8	boarder	0.0625	0.602	0.038
9	staff	0.0417	0.602	0.025
10	include	0.0208	0.602	0.013
11	prospectus	0.0208	0.602	0.013
12	document	0.0208	0.602	0.013
13	cover	0.0208	0.602	0.013
14	aim	0.0208	0.602	0.013
15	organization	0.0208	0.602	0.013
16	admission	0.0208	0.602	0.013
17	criteria	0.0208	0.602	0.013
18	outline	0.0208	0.602	0.013
19	facility	0.0208	0.602	0.013
20	welfare	0.0208	0.602	0.013
21	support	0.0208	0.602	0.013
22	service	0.0208	0.602	0.013
23	special	0.0208	0.602	0.013
24	religious	0.0208	0.602	0.013
25	cultural	0.0208	0.602	0.013
26	aspect	0.0208	0.602	0.013

27	relate	0.0208	0.602	0.013
28	relevant	0.0208	0.602	0.013
29	policy	0.0208	0.602	0.013
30	prospective	0.0208	0.602	0.013
31	reflect	0.0208	0.602	0.013
32	current	0.0208	0.602	0.013

Tabel 5. *Keyword* standar 1 sampai 10

Standar ke-	Kata Kunci
1	*school, statement, boarding*
2	*bully, staff, school*
3	*child, school, protection*
4	*child, school, physical*
5	*complaint, parent, boarder*
6	*abuse, policy, substance*
7	*boarder, record, welfare*
8	*staff, welfare, provision*
9	*boarder, welfare, crisis*
10	*boarding, house, provision*

Perhitungan Kemiripan Semantik

Pada penelitian ini perhitungan kemiripan semantik dilakukan dengan membandingkan kata kunci pada Tabel 5 yang diperoleh dalam setiap standar. Setiap kata kunci akan dibandingkan dengan masing-masing kata kunci lainnya. Perhitungan *semantic similarity* dihitung menggunakan pendekatan *wu palmer* dengan tools WS4J. Hasil perbandingan antar kata kunci di setiap standar menghasilkan suatu matrik kemiripan semantik seperti pada Tabel 6.

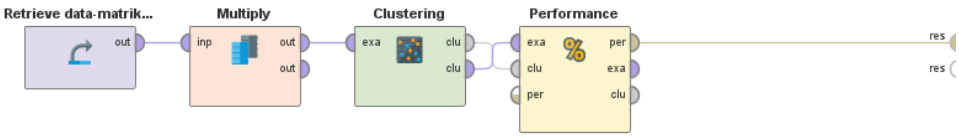
Tabel 6 Matrik hasil perhitungan kemiripan semantik kata pada standar 1 dan 2

	1-school#n#1	1-statement#n#1	1-boarding#n#1
2-bully#n#1	**0,3158**	**0,3750**	**0,2857**
2-staff#n#1	**0,7059**	**0,4286**	**0,6316**
2-school#n#1	**1,0000**	**0,4000**	**0,9000**

Clustering

Matrik kemiripan semantik yang diperoleh seperti dalam Tabel 6 diproses dengan 2 skenario klasterisasi. Pertama, klasterisasi dengan menggunakan nilai perbandingan setiap kata kunci dalam standar. Dan kedua, dengan menggunakan nilai rata-rata perbandingan

pada masing-masing standar. Masing-masing skenario dilakukan percobaan sebanyak 11 kali.



Gambar 3. Desain percobaan pada RapidMiner matrik antar kata

Proses *clustering* dilakukan dengan menggunakan *tools* RapidMiner. Desain percobaan dalam RapidMiner seperti pada Gambar 3. Proses *clustering* dikerjakan menggunakan pendekatan *K-Means*. Dalam prosesnya, *K-Means* menggunakan persamaan (2) untuk menentukan anggota kelompoknya. Dalam menentukan K yang paling baik, dilakukan perhitungan perbandingan dengan mencari nilai SSE atau *sum of squared error* terhadap setiap *cluster* di setiap percobaan dengan menggunakan Persamaan (3).

Identifikasi Divisi Struktur Organisasi

Divisi atau unit-unit kerja yang dibentuk didasarkan pada data hasil klasterisasi. Data klaster menunjukkan bagian-bagian yang dibutuhkan untuk mengembangkan sebuah struktur organisasi pada pondok pesantren dengan tujuan yang telah tertuang dalam standar sekolah berasrama. Pada penelitian ini, dilakukan 2 skenario. Skenario pertama, struktur organisasi yang dibentuk berdasarkan matrik kemiripan semantik antar kata. Dan kedua, struktur organisasi yang dibentuk berdasarkan matrik kemiripan semantik antar standar.

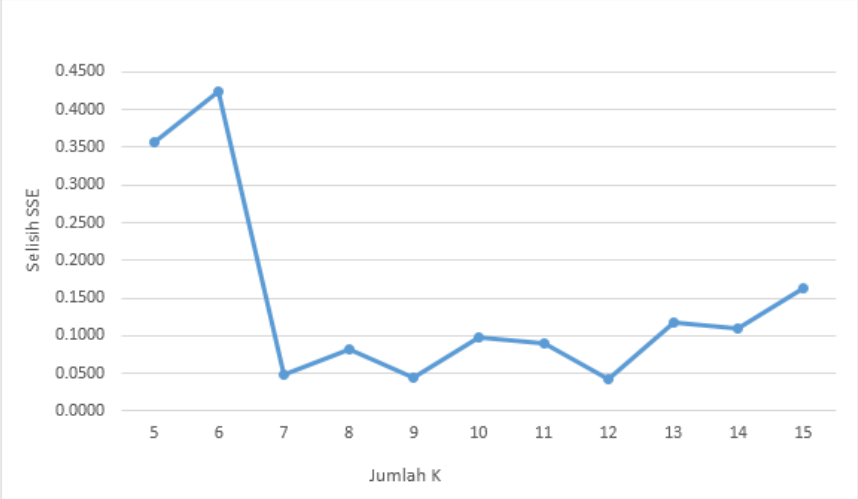
Skenario 1: *Clustering* Matrik *Semantic Similarity* Antar Kata

Proses *clustering* dilakukan sebanyak 11 kali dengan variasi jumlah K sehingga diperoleh data seperti pada Tabel 7. Dalam menentukan jumlah cluster yang terbaik, dilakukan perhitungan SSE menggunakan Persamaan (3). Nilai selisih SSE yang diperoleh dalam clustering (lihat Tabel 7) kemudian digambarkan ke dalam grafik di Gambar 4. Dalam Gambar 4 diketahui bahwa selisih SSE tertinggi ada pada percobaan ke-2 dengan jumlah K=6. Sehingga jumlah *cluster* terbaik adalah 6.

Tabel 7. Nilai dan selisih SSE pada proses *clustering* matrik kemiripan semantik antar kata

Percobaan ke-	Jumlah K	Nilai SSE	Selisih SSE
1	5	0,3571	0,3571
2	6	0,7816	0,4245
3	7	0,7339	0,0477
4	8	0,6531	0,0808
5	9	0,5651	0,0439
6	10	0,6022	0,0971
7	11	0,5261	0,0902
8	12	0,5609	0,0421
9	13	0,4476	0,1179

10	14	0,3064	0,1089
11	15	0,3664	0,1631



Gambar 4. Grafik selisih SSE *clustering* matrik kemiripan antar kata

Dilakukan analisis terhadap data klaster yang diperoleh pada clustering data matrik kemiripan antar kata dengan K=6, untuk kemudian dijadikan sebagai suatu divisi atau unit kerja dalam struktur organisasi. Sehingga didapatkan hasil sebagai berikut: *cluster* 0 menjadi divisi kesantrian; *cluster* 1 menjadi divisi keamanan; *cluster* 2 menjadi divisi pendidikan; *cluster* 3 menjadi divisi sarana dan fasilitas; *cluster* 4 menjadi divisi kepegawaian; dan *cluster* 5 menjadi divisi kesejahteraan (lihat Gambar 5).

Gambar 5. Divisi berdasarkan matrik kemiripan antar kata

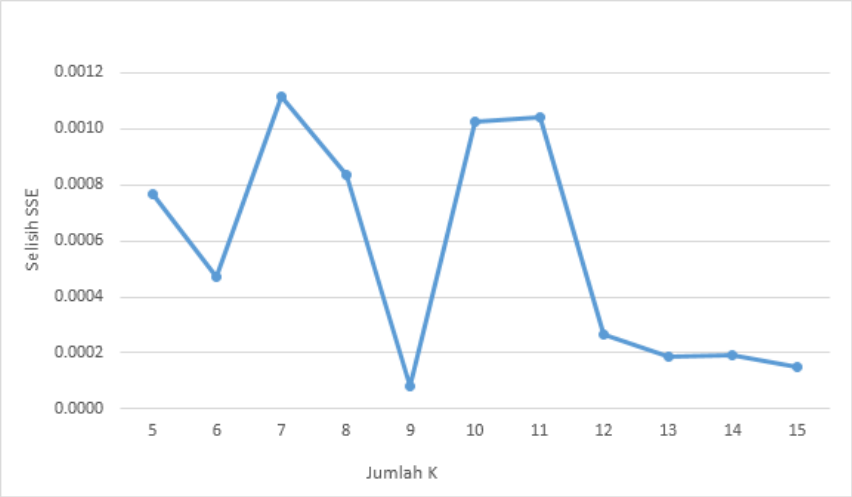
Skenario 2: *Clustering* Matrik *Semantic Similarity* Antar Standar

Clustering dilakukan dengan 11 kali percobaan dengan jumlah K yang bervariasi. Penentuan banyaknya cluster yang optimal dengan memperhatikan perubahan nilai SSE terbesar. Proses *clustering* pada skenario kedua menghasilkan data pada Tabel 8. Data selisih SSE digambarkan ke dalam bentuk grafik di Gambar 6. Dalam Gambar 6 diketahui bahwa selisih SSE terbesar ada pada percobaan ke-3 dengan K=7. Sehingga dalam skenario kedua, *cluster* terbaik yang dapat dibentuk adalah 7.

Tabel 8. Nilai dan selisih SSE pada proses *clustering* matrik kemiripan semantik antar standar

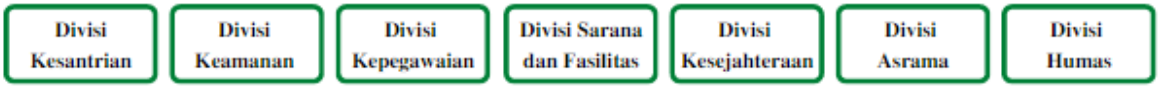
Percobaan ke-	Jumlah K	Nilai SSE	Selisih SSE
1	5	0,0007	0,0008
2	6	0,0001	0,0005
3	7	0,0015	0,0011
4	8	0,0022	0,0008

5	9	0,0023	0,0001
6	10	0,0033	0,0010
7	11	0,0023	0,0010
8	12	0,0025	0,0003
9	13	0,0023	0,0002
10	14	0,0025	0,0002
11	15	0,0027	0,0002



Gambar 6. Grafik selisih SSE *clustering* matrik kemiripan antar standar

Analisis dilakukan terhadap data klaster yang diperoleh pada clustering data matrik kemiripan antar standar dengan K=7, menghasilkan data sebagai berikut: *cluster* 0 menjadi divisi kesantrian; *cluster* 1 menjadi divisi keamanan; *cluster* 2 menjadi divisi kepegawaian; *cluster* 3 menjadi divisi sarana dan fasilitas; *cluster* 4 menjadi divisi kesejahteraan; *cluster* 5 menjadi divisi asrama; dan *cluster* 6 menjadi divisi humas (lihat Gambar 7.



Gambar 7. Divisi berdasarkan matrik kemiripan antar standar

KESIMPULAN

Dari 2 skenario yang dilakukan dihasilkan kesimpulan seperti berikut: Pertama, skenario pertama adalah dengan melakukan *clustering* terhadap data matrik kemiripan semantik antar kata, diperoleh selisih nilai SSE terbesar yakni 0,4245 pada percobaan ke-2 dengan jumlah K=6. Sehingga cluster terbaik adalah 6 cluster dimana cluster-cluster tersebut menjadi 6 divisi yakni, 1) divisi kesantrian; 2) divisi keamanan; 3) divisi pendidikan; 4) divisi sarana dan fasilitas; 5) divisi kepegawaian; dan 6) divisi kesejahteraan. Kesimpulan kedua, Skenario kedua dengan melakukan clustering terhadap data matrik kemiripan semantik antar standar, diperoleh selisih nilai SSE terbesar yaitu 0,0011 pada percobaan ke-3 dengan jumlah

K=7. Sehingga cluster terbaik adalah 6 cluster dimana cluster-cluster tersebut menjadi 7 divisi yakni, 1) divisi kesantrian; 2) divisi keamanan; 3) divisi kepegawaian; 4) divisi sarana dan fasilitas; 5) divisi kesejahteraan; 6) divisi asrama; dan 7) divisi humas.

REFERENSI

[1] Kementerian Agama, "Pangkalan Data Pondok Pesantren", *ditpdpontren.kemenag.go.id*, 2019. [Online]. Tersedia: https://ditpdpontren.kemenag.go.id/pdpp/statistik [Diakses: 29 September 2022]
[2] H. Purnomo, "*Manajemen Pendidikan Pondok Pesantren*", Yogyakarta: Bilndung Pustaka Utama, 2017
[3] Minister for Health and Social Services, "*National Minimum Standards of Boarding School*", United Kingdom: Welsh Assembly Government, 2003.
[4] Ahmad, "Pengertian Struktur Organisasi: Fungsi, Jenis, dan Contoh", *Gramedia*, 2015. [Online]. Tersedia: https://www.gramedia.com/literasi/struktur-organisasi/ [Diakses: 28 September 2022].
[5] M. Nurjannah, Hamdani, and I. F. Astuti, "Penerapan Algoritma Term Frequency - Inverse Document Frequency (TF-IDF) Untuk Text Mining*", Jurnal Informatika Mulawarman*, vol. 8, no. 3, pp. 110-113, 2013.
[6] Y. Agusta, "Clustering", *Wordpress*, 23 Agustus 2014. [Online]. Tersedia: https://yudiagusta.wordpress.com/clustering/ [Diakses: 2 September 2022].
[7] Y. Caterina, M. A. Yaqin, S. Zaman, "Pengukuran Kemiripan Makna Kalimat Dalam Bahasa Indonesia Menggunakan Metode Path", *Fountain of Informatics Journal*, vol. 6, no. 2, pp. 45-50, 2021.
[8] A. Maulana, M. A. Bijaksana, and M. S. Mubarok, "Perancangan Semantic Similarity Based On Word Thesaurus Menggunakan Pengukuran Omiotis Untuk Pencarian Aplikasi pada I-GRACIAS", *E-proceeding of Engineering*, vol. 3, no. 2, pp. 3689-3699, 2016.
[9] R. Tineges, "Tahapan Text Preprocessing dalam Teknik Pengolahan Data", *Dqlab*, 17 Juni 2021. [Online]. Tersedia: https://www.dqlab.id/tahapan-text-preprocessing-dalam-teknik-pengolahan-data [Diakses: 20 Agustus 2022].
[10] A. Millah and S. Nurazizah, "Perbandingan Penggunaan Algoritma Cosinus dan Wu Palmer Untuk Mencari Kemiripan Kata Dalam Plagiarism Checker". *Jurnal Ilmu Komputer dan Desain Komunikasi Visual*, vol. 2, no. 1, pp. 15-25, 2017.
[11] B. Santosa, "*Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis*", Yogyakarta: Graha Ilmu, 2007.
[12] S. Santoso, "*Statistik Multivariat*", Jakarta: Elex Media Komputindo, 2010.
[13] X. Wu and V. Kumar, *The Top 10 Algorithms In Data Mining*, London: CRC Press Taylor & Francis Group, 2009.
[14] A. Larasati, R. Maren, and R. Wulandari, "Utilizing Elbow Method for Text Clustering Optimization in Analyzing Social Media Marketing Content of Indonesian e-Commerce", *Jurnal Teknik Industri*, vol. 23, no. 2, pp. 111-119, 2021.
[15] S. Haris, "*Politik Organisasi Perspektif Mikro Diagnosa Psikologis*", Yogyakarta: Pustaka Pelajar, 2006.
[16] V. Rivai, "*Kepemimpinan Dan perilaku Organisasi*", Jakarta: Grafindo Persada, 2008.
[17] P. Reyvan M., "Belajar Clustering dengan Kursus Data Scientist", *Dqlab*, 25 Desember 2020. [Online]. Tersedia: https://dqlab.id/belajar-clustering-dengan-kursus-data-scientist [Diakses: 2 September 2022]
[18] Ulta, A. Gandhis dan Yaqin, M. Ainul, "Implementasi Metode Semantic Similarity untuk Pengukuran Kemiripan Makna Antar Kalimat", *Journal of Computer Science and Applied Informatics*, vol. 1, no. 2, pp. 47-57, 2019.