Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 7, No. 7, Juli 2023, hlm. 3498-3506 http://j-ptiik.ub.ac.id

Analisis Clustering Topik *Survey* menggunakan Algoritme K-Means (Studi Kasus: Kudata)

Muhammad Arienal Haq[1], Welly Purnomo[2], Nanang Yudi Setiawan[3]

Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]muharienal@student.ub.ac.id, [2]wepe@ub.ac.id, [3]nanang@ub.ac.id

Abstrak

Kudata merupakan salah satu *platform* yang menyediakan layanan untuk menghubungkan pembuat survey yang ingin mencari responden. Pada permasalahan perusahaan ini tidak dapat diketahui untuk topik *survey*, karena *platform* yang digunakan Kudata, yakni *Google Forms*, yang menimbulkan data tidak terstruktur dalam *internal database* Sehingga, peneliti melakukan penelitian terkait permasalahan Kudata dalam melakukan beberapa pengembangan bisnis, seperti melakukan kategori *survey*, membuat *template survey*, dan mengetahui topik *survey* yang banyak digunakan oleh pengguna, serta tren waktu ke waktu pada topik *survey*. Metode pada penelitian ini menggunakan *scraping* untuk mengumpulkan data instrumen *survey*, meliputi deskripsi, pertanyaan dan kombinasi keduanya pada setiap formulir di *Google Forms*, serta hasilnya mendapatkan 1913 data URL dan kemudian dilakukan penerapan skenario pengujian, yang dilakukan dengan membagi *dataset* utama menjadi 3 rentang waktu (setiap 6 bulan) dan membagi kembali *dataset* tersebut menjadi 3 instrumen penting *survey* (deskripsi, pertanyaan dan kombinasi keduanya). Selain itu, penelitian ini menggunakan *text representation* dengan metode TF-IDF *(Term Frequency-Inverse Document Frequency)* dan reduksi dimensi menggunakan PCA *(Principal Component Analysis)*, serta menggunakan *Silhouette Score* untuk menghasilkan *cluster* secara optimal dalam algoritme *K-means*. Sehingga, penelitian ini menghasilkan rekomendasi dan tren topik *survey*. Terdapat 16 rekomendasi topik yang sering digunakan dan 4 topik unik yang diidentifikasi dalam seluruh skenario pengujian.

Kata kunci: *kudata, survey, scraping, k-means, cluster, topik*

Abstract

Kudata is a platform that provides services to connect survey makers who want to find respondents. In this company's problem, it is not possible to know the survey topics, because the platform used by Kudata, namely Google Forms, which creates unstructured data in the internal database So, researchers conducted research related to Kudata's problems in conducting several business developments, such as categorizing surveys, creating survey templates, and knowing survey topics that are widely used by users, as well as trends in survey topics over time. This research used the scraping method to collect survey instrument data, including descriptions, questions and a combination of both on each form in Google Forms, and the results obtained 1913 URL data and then implemented a test scenario, which was carried out by dividing the main dataset into 3 time spans (every 6 months) and re-dividing the dataset into 3 important survey instruments (descriptions, questions and a combination of both). In addition, this research uses text representation with TF-IDF (Term Frequency-Inverse Document Frequency) method and dimensionality reduction using PCA (Principal Component Analysis), and uses Silhouette Score to generate optimal clusters in K-means algorithm. Thus, this research produces recommendations and trends in survey topics. There are 16 recommended frequently used topics and 4 unique topics identified in all test scenarios.

***Keywords**: kudata, survey, scraping, k-means, cluster, topics*

1. PENDAHULUAN

Pesatnya perkembangan penggunaan komputer dan teknologi internet telah menyebabkan sejumlah besar data tidak terstruktur dihasilkan oleh berbagai perangkat dan *platform*. Menurut (Maryanto, 2017) kumpulan data tekstual dalam format apa pun atau tanpa struktur yang melekat disebut dengan data tidak terstruktur. Hampir setiap organisasi

di seluruh dunia menyimpan data tidak terstruktur dalam *database*, sebagian besar dalam bentuk teks, dan pertumbuhan itu konstan pada tingkat eksponensial dari waktu ke waktu.

Penelitian ini menggunakan data format teks, maka metode yang dapat digunakan adalah *text mining* untuk melakukan *text preprocessing* dan *text representation*. Sebelum masuk ke dalam metode tersebut, perlu adanya proses *collecting data*, salah satunya menggunakan metode *scraping* yang mengambil dokumen semi-terstruktur dari internet (Setiawan, Tristiyanto, & Hijriani, 2020).

Text mining adalah teknik penambangan data yang sumber datanya berupa teks diambil dari dokumen, untuk menemukan wakil isi dokumen berupa kata untuk melakukan analisis hubungan antar dokumen. Beberapa aspek spesifik dari *text mining* meliputi klasifikasi teks dan pengelompokan teks (Putri & Setiadi, 2014). Menurut (Rosell, 2009) *text clustering* merupakan proses membagi sekumpulan teks menjadi beberapa *cluster*, sehingga teks-teks tersebut berada di dalam masing-masing cluster serupa dalam konten. Algoritme K-means berbasis jarak untuk mengelompokkan data. Algoritme ini hanya bekerja dengan atribut numerik (Witten , Frank, & Hall, 2011). Pada proses *clustering* penentuan jumlah *cluster* (*k*) penting dilakukan karena mempengaruhi kualitas dan interpretasi hasil *clustering*. Metrik yang digunakan untuk menentukan jumlah *cluster* pada penelitian ini adalah *Silhouette Score* atau *Silhouette Coefficient*, adapun keuntungannya adalah nilai yang dihasilkan dapat digunakan untuk menentukan jumlah *cluster* alami dalam kumpulan data. Metrik ini merupakan kombinasi dari metode pemisahan dan kohesi (Kodinariya & Makwana, 2013).

Kudata merupakan salah satu platform penyedia jasa *survey online* terpusat berdiri sejak 2021 yang menyediakan layanan untuk menghubungkan *maker* atau pembuat *survey* yang ingin mencari responden. Terdapat berbagai macam *survey* yang ada pada Kudata, namun tidak dapat diketahui untuk topik *survey*, karena *platform* pengisian *survey* pada Kudata menggunakan pihak ketiga, yakni *Google Forms*. Hal tersebut seharusnya dapat dimanfaatkan untuk menjawab permasalahan Kudata dalam melakukan beberapa pengembangan bisnis, seperti melakukan kategori *survey*, membuat template *survey*, dan mengetahui topik *survey* yang banyak digunakan oleh pengguna, serta tren topik *survey* dari waktu ke waktu.

Berdasarkan permasalahan diatas, peneliti menawarkan solusi dengan melakukan pengelompokan data untuk menghasilkan rekomendasi dan tren topik *survey*, sehingga dapat memberikan rekomendasi topik *survey* berupa analisis deskriptif dengan memanfaatkan data yang dimiliki Kudata dalam rentang berdasarkan beberapa skenario pengujian dengan membagi *dataset* utama menjadi 3 rentang waktu (setiap 6 bulan) dan membagi kembali *dataset* tersebut menjadi 3 instrumen penting *survey* (deskripsi, pertanyaan dan kombinasi keduanya) untuk menghasilkan analisis yang komprehensif berdasarkan term atau istilah yang kuat pada tiap *cluster*. Pada akhirnya, peneliti tertarik untuk membahasnya dalam skripsi dengan judul "Analisis *Clustering* Topik *Survey* Menggunakan Algoritme *K-means* (Studi Kasus: Kudata)".

2. LANDASAN KEPUSATAKAAN

2.1 *Scraping*

Menurut (Turland, 2010), scraping melibatkan penggalian dokumen pada internet untuk semi-terstruktur dalam bentuk halaman web, menggunakan bahasa markup seperti HTML atau XHTML, serta melakukan analisis dokumen untuk mengambil informasi spesifik dari situs dan menggunakannya untuk tujuan tertentu.

Scraping melibatkan beberapa langkah, sebagai berikut: 1) Membuat template: *Programmer* mempelajari dokumen HTML dari halaman web dari mana informasi akan diambil untuk *tag* HTML yang mengelilingi informasi yang akan diambil, 2) Jelajahi navigasi situs: *Programmer* yang mempelajari teknik penjelajahan web akan memulihkan informasi mimik dalam program *scraping* yang akan dijalankan, 3) Navigasi dan Ekstrak: Berdasarkan informasi yang diperoleh pada langkah 1 dan 2 di atas, dibuatlah program pengumpulan data untuk mengumpulkan informasi secara otomatis dari situs web yang ditentukan dan 4) Mengekstraksi data paket dan riwayat: Informasi yang diperoleh pada langkah 3 disimpan dalam database (Josi, Abdillah, dan Suryayusra, 2022).

2.2 *Text Preprocessing*

Text preprocessing merupakan proses pertama pada *text mining*, proses ini meliputi

persiapan data teks yang akan digunakan untuk kemungkinan pengolahan pada tahap selanjutnya. *Text preprocessing* digunakan untuk menyiapkan teks sebelum menggunakannya dalam pengujian atau pelatihan untuk tujuan mengurangi *noise* dalam data (Indraloka & Santosa, 2017). Proses yang dilakukan sebagai berikut:

1. *Case folding* mengonversi semua kalimat yang memiliki huruf menjadi huruf kecil dan menghapus tidak valid pada karakter, termasuk angka, tanda baca, dan URL *(Uniform Resource Locators)*. Selain itu, proses penghilangan angka dan simbol khusus yang tidak begitu penting seperti tanda seru (!), koma (,), garis miring (/), lebih besar (>), lebih kecil (<) dan lain-lain.
2. *Tokenizing* melibatkan pemotongan kalimat menurut kata-kata yang menyusunnya. Tokenisasi memberikan gambaran proses pembagian teks menjadi kata-kata dengan menggunakan spasi sebagai pembatas dengan tujuan agar berdiri sendiri untuk setiap kata tanpa adanya hubungan dengan kata lain.
3. *Filtering* disebut juga dengan menghilangkan stopwords, yaitu proses kata-kata yang dihilangkan dan dianggap tidak relevan atau tidak menggambarkan makna isi kalimat. Dalam sebuah kalimat seringkali terdapat makna yang tidak lagi memiliki kaitan pada, seperti “ini”, “itu”, dll. Oleh karena itu, sering muncul kata-kata tersebut tetapi tidak berdampak besar jika dihilangkan pada tahap ini.
4. *Stemming* melibatkan konversi kata-kata dengan imbuhan yang berbeda ke kata dasarnya, langkah ini sering dilakukan untuk teks berbahasa Inggris, karena struktur afiks pada bahasa inggris cenderung stabil. *Stemming* merupakan kata yang diproses menjadi bentuk dasarnya dengan menghapus imbuhan yang melekat pada kata tersebut, seperti “in-”, “-nya” dan lain-lain.

2.3 *TF-IDF* (Term Frequency-Inverse Document Frequency)

TF *(Term Frequency)* adalah frekuensi kemunculan suatu kata dalam setiap dokumen. Dari TF kita mendapatkan DF *(Document Frequency)*, yaitu jumlah kata yang terkandung dalam dokumen tersebut. TF-IDF adalah nilai untuk menghitung dokumen dengan bobot kata yang telah ditemukan. TF-IDF diperoleh dengan mengalikan TF dan IDF, dimana IDF adalah kebalikan dari DF (Sammut & Webb, 2011). Perhitungannya dapat ditulis pada persamaan 2.1 dan 2.2 sebagai berikut:

$$IDF(w) = \log\left(\frac{n}{DF(w)}\right) \quad (2.1)$$

Keterangan:
IDF(w) : keseluruhan bobot kata dalam dokumen
w : kata
n : jumlah seluruh dokumen
DF(w) : jumlah dokumen yang terdapat kata w

$$TF - IDF(w,d) = TF(w,d) \times IDF(w) \quad (2.2)$$

Keterangan:
IDF(w) : invers DF dari kata w
TF(w,d) : frekuensi kemunculan kata w pada dokumen d

Pada perhitungan IDF persamaan 2.1, jika n = DF(w) maka hasilnya adalah 0 (nol) dan untuk menyiasatinya, dapat menambahkan nilai 1 ke sisi IDF dan perhitungan TF(w,d) menjadi:

$$TF - IDF(w,d) = TF(w,d) \times \left(log\left(\frac{n}{DF(w)}\right) + 1\right) \quad (2.3)$$

Kemudian, untuk menormalkan nilai TF-IDF pada rentang normalisasi dari 0 hingga 1, persamaan 2.3 dinormalkan menurut persamaan 2.4 sebagai berikut:

$$TF - IDF(w,d) = \frac{TF-IDF(w,d)}{\sqrt{\sum_{w=1}^{n} TF-IDF(w,d)^2}} \quad (2.4)$$

2.4 PCA *(Principal Component Analysis)*

PCA *(Principal Component Analysis)* merupakan metode analisis *multivariate* yang menyederhanakan digunakan pada data dengan cara mentransformasikan variabel-variabel awal sehingga jumlah variabelnya menjadi lebih sedikit, namun tetap mampu merepresentasikan sebagian besar variasi dari variabel asli. Reduksi (penyederhanaan) dimensi dilakukan berdasarkan kriteria persentase variasi data yang dijelaskan oleh komponen utama. Apabila beberapa komponen utama pertama menjelaskan lebih dari 85% hingga 95% variasi data asli, maka informasi dalam komponen utama ini

sudah mencukupi (Kodinariya & Makwana, 2013). Berikut pesamaan 2.5 mengenai analisis komponen utama:

$$PC_k = a_{12}X_1 + a_{22}X_2 + \cdots + a_{pk}X_p + \varepsilon_1 \quad (2.5)$$

2.5 *Silhouette Score*

Menurut (Handoyo, Rumami M, & Nasution, 2014) *silhouette score* atau yang biasa disebut *silhouette coefficient* merupakan sebuah metode untuk melakukan pengukuran terhadap kualitas dan kekuatan *cluster*. Metode ini menggabungkan konsep dari *cohesion* yang mengukur sebuah *cluster* yang memiliki hubungan antar objek dan *separation* yang mengukur jarak antara *cluster* yang berbeda. Berikut merupakan persamaan 2.6 mengenai perhitungan *silhouette score*:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (2.6)$$

2.6 *K-Means*

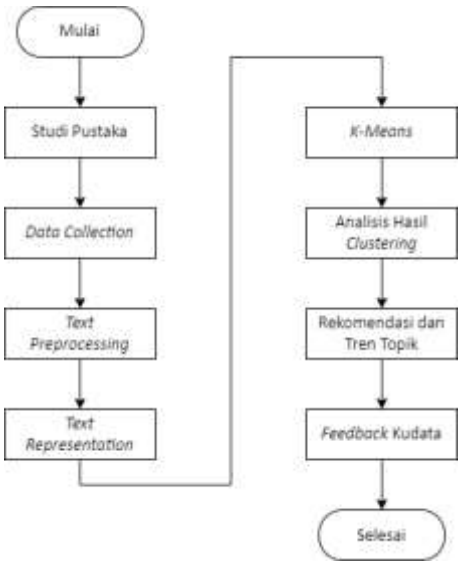
Menurut (Harahap, 2019) *K-Means* merupakan algoritme *non-hierarchical clustering* dimana setiap objek yang termasuk dalam kelompok memiliki kesamaan dan berkorelasi satu sama lain. Data yang dikelompokkan memiliki kesamaan yang lebih besar dengan tingkat perbedaan yang lebih besar juga dengan kelompok lain. Secara prinsip, *clustering* merupakan metode untuk mengelompokkan sekelompok objek menurut atribut atau karakteristik yang serupa dengan data lainnya. *Clustering* adalah salah satu metode data mining dimana algoritme ini bekerja secara tidak terawasi *(unsupervised)*, yang berarti metode ini tidak lagi memerlukan pelatihan atau panduan, bahkan *input*. Dalam *data mining*, terdapat dua jenis metode pengelompokan yang digunakan untuk mengelompokkan data, yaitu *non-hierarchical clustering* dan *hierarchical clustering*.

2.7 Evaluasi Model *Clustering*

Evaluasi hasil dari *clustering* digunakan untuk mengetahui performa atau seberapa baik suatu data yang telah *clustering*. Metrik yang digunakan untuk mengukur tersebut, seperti *homogeneity*, yang mengukur seberapa dekat sebuah algoritme dapat mengelompokan data berdasar *data point* pada kelas yang sama, kemudian *completeness*, yang mengukur tentang *clustering* dapat dikatakan sempurna ketika suatu *data point* terkumpul dalam kelas yang sama, selain itu dapat menggunakan *v-measure* yang merupakan metrik gabungan dari *homogeneity* dan *completeness* (Solikin, Kusrini, & Wibowo, 2021).

3. METODOLOGI

Model penelitian ini merupakan adopsi dari teori mengenai penerapan *text clustering* dengan metode *K-means* berdasarkan penelitian terdahulu yang dilakukkan oleh (Rosell, 2009) yang dimulai dengan mengumpulkan informasi dan mengeksplorasi karakteristik data, atau biasa disebut dengan *text mining*. Berikut merupakan alur dari penelitian yang dapat dilihat pada Gambar 1.



Gambar 1. Diagram Alur Penelitian

Metode yang digunakan dalam pengumpulan data adalah studi pustaka, kemudian observasi terhadap tabel yang memuat informasi yang diperlukan dan *scraping* atau proses pengumpulan data yang ada dalam sebuah web secara spesifik dengan menggunakan *library BeautifulSoup* dari *Python*. Pada penelitian ini dilakukan skenario pengujian dengan membagi *dataset* utama menjadi 3 rentang waktu (setiap 6 bulan) dan membagi kembali *dataset* tersebut menjadi 3 instrumen penting *survey* (deskripsi, pertanyaan dan kombinasi keduanya). Proses normalisasi data teks dilakukan melalui preprocessing yang terdiri dari *case folding, tokenizing, filtering,* dan *stemming*. Data teks hasil *preprocessing* kemudian dilakukan transformasi data ke dalam format numerik dengan melakukan pembobotan istilah atau *term weighting* yang terdiri dari TF-IDF *(Term Frequency-Inverse Document*

Frequency) dan dilakukan reduksi dimensi menggunakan PCA *(Principal Component Analysis)* untuk mendapatkan *feature* yang maksimal. Proses transformasi tersebut direpresentasikan sebagai masukan data untuk algoritme *K-means*. Langkah selanjutnya adalah mengelompokan dengan menggunakan metode *K-means clustering* dan untuk penentuan kecenderungan cluster dilakukan sebagai upaya dalam memperoleh jumlah *cluster* yang optimal dengan menggunakan metrik *Silhouette Score*. Hasil akhir dari penelitian adalah analisis deskriptif setiap *cluster*, pembobotan instrumen *survey*, serta rekomendasi dan tren topik *survey* untuk menjawab kendala Kudata dalam melakukan kategori *survey*, membuat template *survey*, dan mengetahui topik *survey* yang banyak digunakan oleh pengguna, serta tren topik *survey* dari waktu ke waktu.

4. HASIL DAN PEMBAHASAN

Proses pengumpulan data dengan melakukan observasi *internal database* Kudata dengan tabel *surveys*. Tabel ini mencakup sejumlah variabel yang penting dan relevan dengan topik penelitian yang sedang dijalani. Dalam proses pengumpulan data, peneliti berhasil mengakses dan menggunakan sebanyak 1913 baris data yang mewakili berbagai aspek dalam penelitian, salah satunya kolom untuk URL *Google Forms* dan *timestamp survey*. Sebelum dilakukan scraping, dilakukan pembagian menjadi 3 *dataset* berdasarkan interval waktu 6 bulan untuk kebutuhan analisis data mengenai tren atau pola hasil *clustering* yang terbentuk, terdiri dari *dataset* 1 (09/11/2021 sampai 30/04/2022), *dataset* 2 (01/05/2022 sampai 31/10/2022) dan *dataset* 3 (01/11/2022 sampai 04/04/2023). Berikut merupakan tabel 1 untuk pembagian *dataset*.

Tabel 1. *View Database* Kudata

Dataset 1	*Dataset* 2	*Dataset* 3
https://docs.google.com/forms/d/e/1FAIpQLSeKVq...	https://docs.google.com/forms/d/e/1FAIpQLSfDax...	https://docs.google.com/forms/d/e/1FAIpQLScIgg...
https://docs.google.com/forms/d/e/1FAIpQLSeBVK...	https://docs.google.com/forms/d/e/1FAIpQLSfXXC...	https://docs.google.com/forms/d/e/1FAIpQLSdDrl...
...	...	...

Kemudian menggunakan *scraping* dengan *library BeautifulSoup* dari *Python* dan didapatkan masing-masing data instrumen *survey*, meliputi deskripsi, pertanyaan dan kombinasi keduanya, pada setiap *dataset*, seperti pada tabel 2.

Tabel 2. Hasil *Scraping*

Dataset 1	*Dataset* 2	*Dataset* 3
var FB_PUBLIC_LOAD_DATA_ = [null,["Assala muala...	var FB_PUBLIC_LOAD_DATA_ = [null,["Perkena lkan...	var FB_PUBLIC_LOAD_DATA_ = [null,["Selamat pag...
var FB_PUBLIC_LOAD_DATA_ = [null,["Assala muala...	var FB_PUBLIC_LOAD_DATA_ = [null,["Assala muala...	var FB_PUBLIC_LOAD_DATA_ = [null,["Dengan horm...
...	...	...

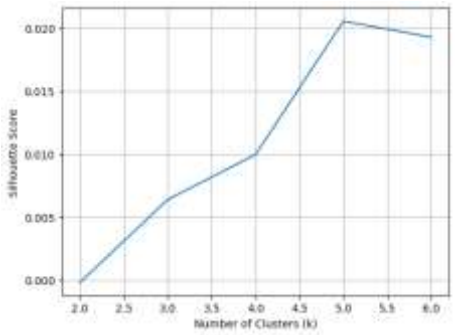
Untuk mendapatkan *clustering* secara optimal dengan menggunakan algoritme K-means, dilakukan beberapa metode, diantaranya *text preprocessing*, seperti *case folding, tokenizing, filtering dan stemming*. Selain itu, juga menggunakan *text representation* dengan TF-IDF *(Term Frequency-Inverse Document Frequency)*, yang direpresentasikan dalam format COO *(Coordinate List)*, kemudian melakukan reduksi dimensi data menggunakan PCA *(Principal Component Analysis)* dengan mempertahankan 95% komponen terhadap keragaman data asli. Komponen hasil reduksi tersebut bertujuan untuk mengatasi masalah *curse of dimensionality* yang dapat mempengaruhi kinerja algoritme dan menganalisis data yang lebih kompleks dengan efisien yang dapat dilihat pada tabel 3.

Tabel 3. Hasil PCA pada Setiap Skenario Pengujian

Skenario	Deskr-ipsi	Pertany-aan	Kombi-nasi	Total Kom-ponen
1	212	206	209	300
Kombi-nasi 1 dan 2	722	678	693	1351
Kom-binasi 1, 2 dan 3	980	905	924	1911

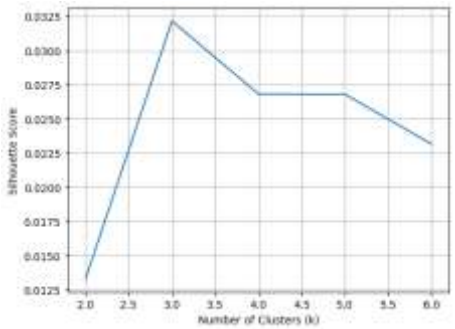
Penentuan nilai optimal k menggunakan *Silhouette Score* dengan penentuan *range* dari k adalah 2 hingga 7 *cluster*, hal tersebut digunakan karena pertimbangan kompleksitas dan memerlukan sumber daya yang lebih besar jika

menggunakan *range* yang lebih tinggi. Hasil *clustering* menggunakan *K-means* sebagai berikut, untuk *dataset* 1, pada instrumen deskripsi, jumlah *cluster* yang optimal adalah 5 dan *Silhouette Score* tertinggi adalah sekitar 0,020 yang dapat diamati pada Gambar 2.



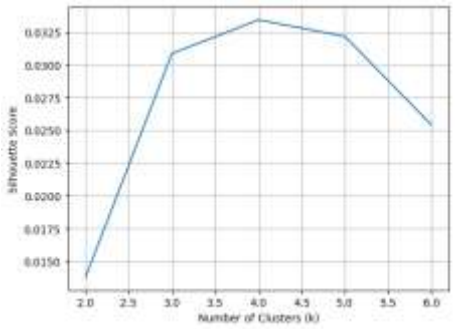
Gambar 2. Hasil *Silhouette Score* untuk Setiap Nilai dari *k* pada Instrumen Deskripsi *Dataset* 1

Instrumen pertanyaan, jumlah *cluster* yang optimal adalah 3 dan *Silhouette Score* tertinggi adalah sekitar 0,032 yang dapat diamati pada Gambar 3.



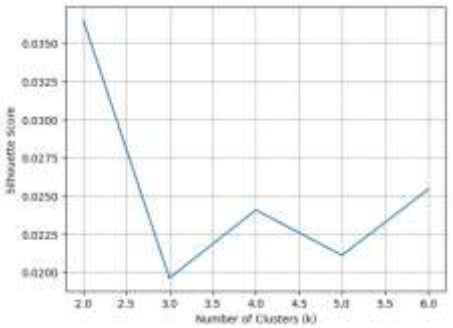
Gambar 3. Hasil Silhouette *Score* untuk Setiap Nilai dari *k* pada Instrumen Pertanyaan *Dataset* 1

Untuk instrumen kombinasi (deskripsi dan pertanyaan), jumlah *cluster* yang optimal adalah 4 dan *Silhouette Score* tertinggi adalah sekitar 0,033 yang dapat diamati pada Gambar 4.



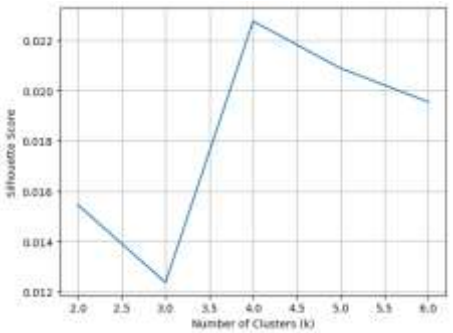
Gambar 4. Hasil Silhouette *Score* untuk Setiap Nilai dari *k* pada Instrumen Kombinasi *Dataset* 1

Kemudian untuk kombinasi *dataset* 1 dan 2, pada instrumen deskripsi, jumlah *cluster* yang optimal adalah 2 dan *Silhouette Score* tertinggi adalah sekitar 0,019 yang dapat diamati pada Gambar 5.



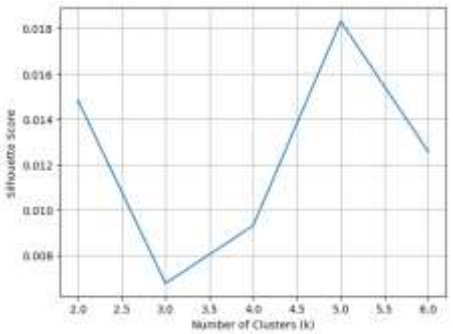
Gambar 5. Hasil *Silhouette Score* untuk Setiap Nilai dari *k* pada Instrumen Deskripsi *Dataset* 1 dan 2

Instrumen pertanyaan, jumlah *cluster* yang optimal adalah 4 dan *Silhouette Score* tertinggi adalah sekitar 0,022 yang dapat diamati pada Gambar 6.



Gambar 6. Hasil *Silhouette Score* untuk Setiap Nilai dari *k* pada Instrumen Pertanyaan *Dataset* 1 dan 2

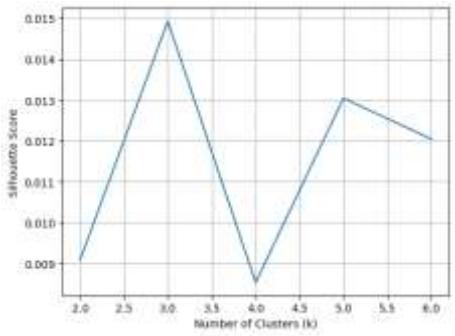
Untuk instrumen kombinasi, jumlah *cluster* yang optimal adalah 5 dan *Silhouette Score* tertinggi adalah sekitar 0,018 yang dapat diamati pada Gambar 7 berikut.



Gambar 7. Hasil *Silhouette Score* untuk Setiap Nilai dari *k* pada Instrumen Kombinasi *Dataset* 1, 2 dan 3

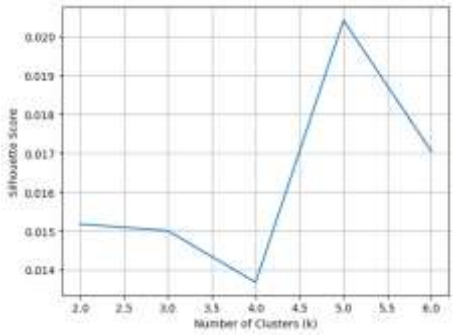
Kemudian untuk kombinasi *dataset* 1, 2 dan 3. Pada instrumen deskripsi, jumlah *cluster* yang optimal adalah 3 dan *Silhouette Score* tertinggi

adalah sekitar 0,014 yang dapat diamati pada Gambar 8.



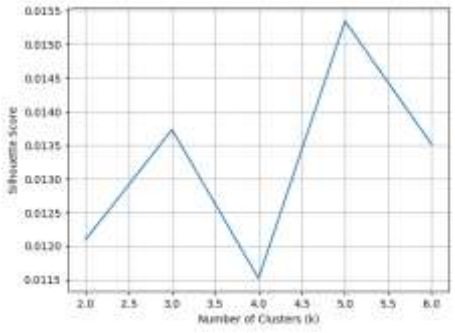
Gambar 8. Hasil *Silhouette Score* untuk Setiap Nilai dari *k* pada Instrumen Deskripsi *Dataset* 1, 2 dan 3

Instrumen pertanyaan, jumlah *cluster* yang optimal adalah 5 dan *Silhouette Score* tertinggi adalah sekitar 0,020 yang dapat diamati pada Gambar 9.



Gambar 9 Hasil *Silhouette Score* untuk Setiap Nilai dari *k* pada Instrumen Pertanyaan *Dataset* 1, 2 dan 3

Untuk instrumen kombinasi, jumlah *cluster* yang optimal adalah 5 dan *Silhouette Score* tertinggi adalah sekitar 0,015 yang dapat diamati pada Gambar 10.



Gambar 10 Hasil *Silhouette Score* untuk Setiap Nilai dari *k* pada Instrumen Kombinasi *Dataset* 1, 2 dan 3

Hasil analisis yang komprehensif mengenai pengaruh skenario pengujian yang diberikan terhadap hasil *clustering* melalui analisis deskriptif setiap *cluster* untuk pembentukan topik, hasilnya sebagai berikut, untuk *dataset* 1 terbentuk 36 topik, kombinasi *dataset* 1 dan 2 terbentuk 35 topik, serta kombinasi *dataset* 1, 2 dan 3 terbentuk 43 topik.

Pembobotan instrumen *survey* juga dilakukan dengan pendekatan frekuensi kemunculan topik. Nilai tersebut kemudian digunakan untuk menentukan bobot *cluster* dengan mengakumulasikannya. Semakin tinggi nilai tersebut, semakin signifikan pengaruhnya. Dari hasil penelitian disimpulkan bahwa instrumen kombinasi memiliki pengaruh yang signifikan terhadap pembentukan topik, karena memiliki bobot tertinggi di setiap skenario pengujian. Untuk rekomendasi dan tren topik survey dapat disimpulkan terdapat 16 topik yang di rekomendasikan. Hasil dari rekomendasi tersebut dapat dilihat pada tabel.

Table 4. Hasil Rekomendasi Topik

Rekomendasi Topik
perilaku konsumen (9) , reputasi merek (6) , kepuasan mahasiswa (5), interaksi sosial (5), pemilihan produk (5), pengalaman konsumen (5), persepsi kredibilitas (5) , pengetahuan produk (4), tingkat kepuasan (4), komunitas (3), konten informatif (3), networking (3), olahraga (2), penggunaan teknologi (2), kesehatan mental (2), dan kualitas produk (2)

Selain itu juga terdapat juga topik lainnya yang teridentifikasi dan dapat menjadi pertimbangan karena menciptakan tren baru di setiap skenario pengujian, namun topik-topik tersebut memiliki pengaruh yang kurang signifikan.

Table 5. Hasil Rekomendasi Topik Unik

Rekomendasi Topik Unik
efektivitas administrasi (*dataset* 1), kualitas kedai kopi (kombinasi *dataset* 1 dan 2), perkembangan UMKM (kombinasi *dataset* 1, 2 dan 3), dan minat dalam pariwisata (kombinasi *dataset* 1, 2 dan 3)

5. KESIMPULAN DAN SARAN

Dari hasil penelitian dan analisis yang telah dilakukan, dapat disimpulkan bahwa proses pengumpulan data dilakukan melalui observasi pada database internal Kudata dengan tabel *surveys*. Data yang diperoleh terdiri dari 1913 baris yang mencakup berbagai aspek penelitian, termasuk kolom URL *Google Forms* dan *timestamp survey*. Sebelum dilakukan scraping, data URL *Google Forms* dibagi menjadi 3 dataset berdasarkan interval waktu 6 bulan. Metode *scraping* menggunakan *library BeautifulSoup* dari *Python* digunakan untuk

mendapatkan data instrumen *survey*, seperti deskripsi, pertanyaan, dan kombinasi keduanya, pada setiap *dataset*. Untuk mendapatkan *clustering* yang optimal, metode yang dilakukan meliputi *text preprocessing*, seperti *case folding, tokenizing, filtering, dan stemming*, serta menggunakan *text representation* dengan TF-IDF dan reduksi dimensi data menggunakan PCA. Penentuan nilai optimal *k* menggunakan *Silhouette Score* dengan *range k* dari 2 hingga 7. Hasil *clustering* menunjukkan jumlah *cluster* optimal dan *Silhouette Score* tertinggi untuk setiap instrumen dan kombinasi *dataset*. Hasil analisis menyimpulkan bahwa instrumen kombinasi memiliki pengaruh yang signifikan terhadap pembentukan topik. Selain itu, terdapat 16 topik yang direkomendasikan, termasuk perilaku konsumen, tingkat kepuasan, reputasi merek, olahraga, dan kesehatan mental. Namun, metode *clustering* yang digunakan masih perlu ditingkatkan untuk memisahkan dan mengelompokkan data dengan lebih efektif, seperti mempertimbangkan metode *clustering* lainnya atau melakukan penyesuaian parameter dan fitur yang digunakan dalam analisis.

6. DAFTAR PUSTAKA

Averina, A., Hadi, H., & Siswantoro, J. (2022). Analisis Sentimen Multi-Kelas Untuk Film Berbasis Teks Ulasan Menggunakan Model Regresi Logistik. TEKNIKA, 123-128.

Berry, M. W., & Kogan, J. (2010). Text Mining: Applications and Theory. United Kingdom: Wiley.

Chen, Z. L. (2022). Research and Application of Clustering Algorithm for Text Big Data. Computational Intelligence and Neuroscience, 1-8.

Handoyo, R., Rumami M, R., & Nasution, S. M. (2014). PERBANDINGAN METODE CLUSTERING MENGGUNAKAN METODE SINGLE LINKAGE DAN K - MEANS PADA PENGELOMPOKAN DOKUMEN. JSM STMIK Mikroskil, 73-82.

Harahap, B. (2019). Penerapan Algoritma K-Means Untuk Menentukan Bahan Bangunan Laris (Studi Kasus Pada UD. Toko Bangunan YD Indarung). Regional Development Industry & Health Science, Technology and Art of Life, 394-403.

Indraloka, D. S., & Santosa, B. (2017). Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia. JURNAL SAINS DAN SENI ITS, 51-56.

Josi, A., Abdillah, L. A., & Suryayusra. (2022). PENERAPAN TEKNIK WEB SCRAPING PADA MESIN PENCARI ARTIKEL ILMIAH. UNIVERSITAS BINA DARMA, 159-164.

Jumeilah, F. S. (2017). Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian. JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi), 19-25.

Kodinariya, T. M., & Makwana, D. R. (2013). Review on determining number of cluster in K-Means clustering. International Journal of Advance Research in Computer Science and Management Studies, 90-95.

Maryanto, B. (2017). BIG DATA DAN PEMANFAATANNYA DALAM BERBAGAI SEKTOR. Media Informatika, 14-19.

Maulida. (2020). TEKNIK PENGUMPULAN DATA DALAM METODOLOGI PENELITIAN. Jurnal Ilmiah Islam dan Sosial, 32-39.

Munawar, & Silitonga, Y. R. (2019). SISTEM PENDETEKSI BERITA HOAX DI MEDIA SOSIAL DENGAN TEKNIK DATA MINING SCIKIT LEARN. Jurnal Ilmu Komputer, 173-179.

Perkovic, L. (2012). Introduction to Computing Using Python: an Application Development Focus. Hoboken: J. Wiley & Sons.

Putri, E. K., & Setiadi, T. (2014). PENERAPAN TEXT MINING PADA SISTEM KLASIFIKASI EMAIL SPAM MENGGUNAKAN NAIVE BAYES. Jurnal Sarjana Teknik Informatika, 73-83.

Rosell, M. (2009). Text Clustering Exploration. Stockholm: Universitetsservice US AB.

Saitta, S., Raphael, B., & Smith, I. (2008). A Comprehensive Validity Index for Clustering. Intelligent Data Analysis, Vol. 12, No. 6, 529-548.

Sammut, C., & Webb, G. I. (2011). Encyclopedia of Machine Learning. Boston: Springer.

Setiawan, D. F., Tristiyanto, & Hijriani, A. (2020). APLIKASI WEB SCRAPING DESKRIPSI PRODUK. Jurnal TEKNOINFO, 41-47.

Solikin, A. F., Kusrini, & Wibowo, F. W. (2021). Evaluasi Cluster Data Interkomparasi Anak Timbangan Dengan Algoritma Self Organizing Maps. SISFOTENIKA, 208-219.

Turland, M. (2010). php|architect's Guide to Web Scraping. Toronto: Marco Tabini & Associates, Inc.

Tutupary, S. E., & Aldianto, L. (2014). THE BENEFITS OF MANAGEMENT INFORMATION SYSTEM ON THE EFFECTIVENESS AND EFFICIENCY OF THE ONLINE BUSINESS. JOURNAL OF BUSINESS AND MANAGEMENT, 835-849.

Witten , I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Massachusetts: Morgan Kaufmann.

Yudiarta, N. G., Sudarma, M., & Ariastina, W. G. (2018). Penerapan Metode Clustering Text Mining Untuk Pengelompokan Berita Pada Unstructured Textual Data. Majalah Ilmiah Teknologi Elektro, 339-344.