

STORYTELLING HASIL KLASTERISASI BI K-MEANS PADA DATA UMKM MENGGUNAKAN TRANSFORMERS

Tubagus Mohammad Akhriza*¹, Tubagus Arief Armanda², Tubagus M. Adrie Admira² dan Latifah²

⁽¹⁾STMIK Pradnya Paramita (STIMATA)

Jl. Laksda Adi Sucipto No.249a, Pandanwangi, Malang, Jawa Timur 65126

⁽²⁾STMIK Jakarta STI&K

Jl. BRI No. 17, Radio Dalam, Kebayoran Baru, Jakarta Selatan 12140

akhriza@stimata.ac.id, tb_armanda@yahoo.com, adrie.admira@jak-stik.ac.id, latifah@jak-stik.ac.id

*Corresponding Author

ABSTRAK

Klasterisasi terhadap data hasil pendampingan unit usaha mikro, kecil dan menengah (UMKM) diselenggarakan dengan tujuan untuk mengetahui pola pengusaha mikro dalam branding produk mereka di platform digital. Artikel ini menjelaskan penggunaan metode klasterisasi Bisecting K-means (Bi K-means) dan membandingkan hasil klasterisasinya dengan hasil dari metode K-means tradisional. Metode Elbow terhadap nilai Silhouette dan Within cluster sum-of-square, serta visualisasi Silhouette klaster-klaster diterapkan untuk mendapatkan k klaster yang optimal untuk mengklaster UMKM. Selanjutnya, untuk meningkatkan interpretasi hasil klasterisasi, sebuah pendekatan transformers berbasis GPT-3.5 digunakan untuk membangkitkan narasi storytelling dari tiap klaster yang terbentuk, selain penerangan klaster berdasarkan strategi branding yang dijalankan oleh UMKM.

Kata Kunci: UMKM, Klasterisasi, K-means, Bisecting K-means, Large language model, GPT

PENDAHULUAN

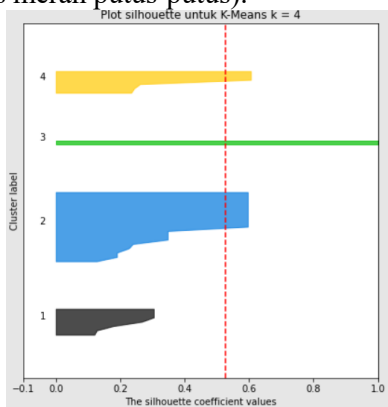
Usaha mikro kecil & menengah (UMKM) menjadi salah satu harapan masyarakat Indonesia untuk bangkit pasca dilanda pandemi Covid-19 [1], [2]. Pembatasan interaksi sosial secara fisik dan luring berskala besar di Indonesia mendorong pengusaha mikro belajar memanfaatkan teknologi informasi untuk melakukan promosi usaha mereka, yaitu salah satunya adalah menggunakan platform digital untuk melakukan branding dan pemasaran digital terhadap produk barang dan/atau jasa yang dihasilkan [3], [4]. Pemerintah juga mendorong pengusaha UMKM yang terdampak pandemi untuk menggunakan platform digital sebagai media branding produk, dan menargetkan sebanyak 30 juta UMKM naik (onboard) ke platform digital di akhir 2024 [5], [6]. Platform digital dimaksud termasuk media sosial seperti Facebook (FB), Instagram (IG), Tiktok (TT), WhatsApp dan lokapasar elektronik (e-marketplace (EM)) seperti Shopee, Tokopedia, Go-food, Grab-food dan lain-lain [7], [8].

Sejalan dengan gerakan pemerintah ini, institusi non-pemerintah di wilayah

Malang, Jawa Timur juga menyelenggarakan kegiatan pendampingan dan pelatihan branding usaha bagi UMKM, antara lain Mercy Corps Indonesia, dan Ikatan Ahli Informatika Indonesia (IAII) DPW Jawa Timur bersama beberapa perguruan tinggi di dalamnya yaitu Universitas Merdeka Pasuruan, STIKI Malang, Politeknik NSC Surabaya dan STMIK PPKIA Pradnya Paramita (STIMATA) Malang [8]. Program pendampingan dimaksud diberi nama CAMELIA, dan diikuti oleh 43 pengusaha perempuan se-Malang raya. dengan materi yang diberikan adalah logo (LG), kemasan (KM) dan ads copywriting (AC) sebagai komponen merek yang digunakan dalam mem-branding produk.

Dalam rangka memahami tingkat keberhasilan kegiatan pendampingan dimaksud, studi berbasis eksploitasi data (data mining) sudah dilakukan di [8], yaitu dengan menerapkan teknik klasterisasi K-means terhadap data UMKM mengenai aktivitas branding mereka di FB, IG, TT dan EM, serta penggunaan unsur branding LG, KM dan AC di produk yang dihasilkan. Jumlah k klaster yang optimal, yaitu $k = 4$,

sudah diperoleh dengan menerapkan grafik *Elbow* terhadap nilai *Silhouette* dan *Within cluster sum-of-squared error* (WCSS). Namun demikian plot visualisasi nilai *Silhouette* menunjukkan masih ada kluster dengan nilai *Silhouette* di bawah rerata *Silhouette* seluruh kluster, seperti dalam Gambar 1. Kluster dengan label 1 memiliki nilai *Silhouette* jauh di bawah nilai rerata (garis merah putus-putus).



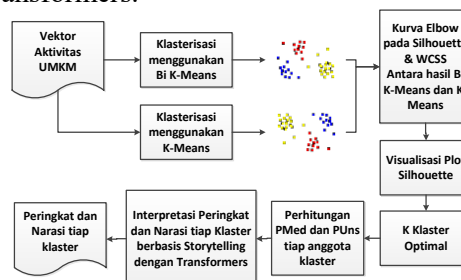
Gambar 1. Plot *Silhouette* K-means dengan $k=4$ [8]

Situasi ini terjadi karena *K-means* harus mencari terlebih dulu k buah *centroid* $c_{i=1..k}$ dari titik-titik data yang ada. Tiap *centroid* c_i diasumsikan mewakili sebuah kluster K_i dan mencari titik-titik data lain untuk dijadikan anggota klasternya jika titik-titik data dimaksud memiliki jarak terdekat dengan *centroid* kluster c_i . Pada praktiknya, akurasi titik-titik data di dalam dan antar kluster hasil dari *K-means* sangat bergantung pada k buah c_i pertama yang dipilih. Masalah lainnya adalah proses interpretasi hasil klusterisasi yang masih dilakukan secara manual, yaitu mempelajari karakteristik tiap obyek di dalam sebuah kluster sebelum menyimpulkan label yang paling sesuai untuk sebuah kluster.

Artikel ini mengusulkan pendekatan *Bisecting K-means* (disingkat *Bi K-means*) untuk memperbaiki hasil klusterisasi *K-means* sehingga nilai *Silhouette* tiap kluster tidak terlalu berada di bawah rerata *Silhouette*; artinya, kualitas tiap klasternya tinggi. Selain itu, artikel ini mengusulkan proses interpretasi tiap kluster dengan pendekatan *Generative Pre-trained Transformers* (GPT) milik OpenAI.

METODE PENELITIAN

Penelitian dilakukan dengan tujuan untuk menentukan k kluster optimal untuk mengelompokkan UMKM berdasarkan vektor aktivitas UMKM, dan kemudian menginterpretasi hasil klusterisasi dengan pendekatan *Transformers*. Alur penelitian diberikan dalam Gambar 2. Setelah klusterisasi dengan *Bi K-Means* dan *K-Means* dilakukan, hasil klusterisasi dianalisis dengan metode *Elbow* terhadap nilai *Silhouette* dan WCSS. Visualisasi plot *Silhouette* dibuat untuk membantu meningkatkan hasil analisis. Setelah diperoleh jumlah k kluster optimal, kluster-kluster diperingkat dan diinterpretasi terhadap persentase penggunaan media sosial (*PMed*) dan persentase penggunaan unsur (*PUns*) oleh UMKM dalam *branding* produk mereka. Peringkat dan narasi storytelling yang menginterpretasi tiap kluster diperoleh dengan pendekatan *transformers*.



Gambar 2. Alur penelitian

Klusterisasi menggunakan *Bi K-Means*

Bi K-means tidak menginisiasi pemilihan jumlah k kluster melalui pemilihan k buah *centroid* sebagaimana yang dilakukan *K-means*, namun secara dinamis mengelola jumlah kluster dengan membagi dua kluster-kluster dengan varian yang tinggi secara iteratif. Varian dari kluster dapat diukur dengan WCSS atau *Silhouette*. Kluster dengan varian yang tinggi berisi titik-titik data yang kemiripan satu dengan lainnya rendah, sehingga dibelah dua oleh *Bi K-means* [9], [10]. Mengacu pada referensi ini, *pseudo-code* klusterisasi *Bi K-means* dijelaskan sebagai berikut:

1. Diberikan dataset D berisi n titik data d
2. # k adalah jumlah kluster akhir yang diinginkan
Masukkan k

3. #dictionary untuk klaster: KLAS dengan kunci = centroid c , dan value = anggota c yaitu titik-titik data d
 $KLAS = \text{dict}[c, [d]]$
4. #inisiasi memilih dua centroid dari semua obyek d di D
 $C = [c1, c2]$ # $c1, c2$ anggota D
5. #Lakukan K-means dengan $k = 2$
 Untuk setiap obyek d di D :
 Untuk setiap centroid c di C :
 hitung jarak $Dist(d, c)$
 Jika $Dist(d, c1)$ terkecil:
 Tambahkan d ke $KLAS[c1]$
 Selain itu:
 Tambahkan d ke $KLAS[c2]$
6. $K = \text{worst}(c1, c2)$ #kunci KLAS dengan nilai $WCSS$ tertinggi atau *Silhouette* terkecil
7. #Bentuk D baru, yaitu semua anggota d di $KLAS[K]$ dengan varian tinggi
 $D = [d \text{ for } d \text{ in } \text{Klaster}[K]]$
8. #hapus entri K dari dict KLAS
 $KLAS.remove[K]$
9. Jika $KLAS.size < k$:
 kirim D baru untuk diproses kembali dari ke langkah 4
10. #Jika $KLAS.size = k$
 Return KLAS
 Selesai

Di dalam prosedur nomer 5, fungsi $Dis(d, c)$ mewakili suatu fungsi penghitung jarak antara titik data d dengan c yang umumnya berbentuk rumus jarak Euclidean, Manhattan, atau rumus lainnya.

Di dalam prosedur nomer 6, disebutkan bahwa untuk menentukan klaster KLAS terburuk karena masih memiliki varian yang tinggi antara $KLAS[c1]$ dan $KLAS[c2]$, ditulis $\text{worst}(c1, c2)$, dapat digunakan metrik $WCSS$ dengan nilai tertinggi (yaitu mendekati 1) atau *Silhouette* dengan nilai terendah (yaitu mendekati -1). Perbedaan $WCSS$ dan *Silhouette* untuk kriteria pembagian klaster adalah bahwa makin besar $WCSS$ (mendekati 1), maka kualitas klaster makin rendah karena *sum-of-squared error* di dalam klaster tersebut besar; sebaliknya, makin besar *Silhouette* (mendekati 1) maka kualitas klaster makin tinggi karena jarak antara klaster makin jauh namun jarak titik dalam satu klaster adalah makin dekat[9], [10].

$$WCSS(K_i) = \sum_{j=1}^{|K_i|} Dis(d_j, c_i) \quad (1)$$

$WCSS$ dihitung dengan rumus (1), dimana c_i adalah centroid klaster K_i dan d_j adalah semua anggota klaster K_i (selain c_i). Koefisien *Silhouette* sil sebuah titik $d_i \in K_i$ dihitung dengan (2):

$$sil(d_i) = \frac{b(d_i) - a(d_i)}{\max\{a(d_i), b(d_i)\}} \quad (2)$$

dengan $a(d_i)$ adalah nilai tengah jarak antara suatu titik $d_i \in K_i$ dengan setiap titik lainnya di klaster yang sama, yaitu $d_j \in K_i$, sebagaimana ditulis dalam (3); sedangkan $b(d_i)$ adalah nilai tengah jarak terkecil antara d_i dengan semua titik di semua klaster yang tidak mengandung d_i , disimbolkan dengan K_j , sebagaimana ditulis dalam (4)

$$a(d_i) = \frac{1}{|K_i| - 1} \sum_{i \neq j, i=1}^{|K_i|} Dis(d_i, d_j) \quad (3)$$

$$b(d_i) = \min_{j \neq i} \frac{1}{|K_j|} \sum_{j=1}^{|K_j|} Dis(d_j, d_i) \quad (4)$$

Menurut (2), agar $sil(d_i)$ mendekati 1, dibutuhkan nilai $a(d_i) \ll b(d_i)$ sehingga $b(d_i) - a(d_i)$ di (2) mendekati 0. Nilai $a(d_i)$ menunjukkan disimilaritas atau ketidakmiripan d_i di klasternya sendiri, jadi makin mendekati nol maka makin sesuai d_i berada di klasternya sendiri; sebaliknya, $b(d_i)$ yang besar berarti d_i makin tidak sesuai berada di klaster lainnya.

Bi K-means (dan K-Means) melakukan klasterisasi terhadap data aktivitas UMKM yang telah ditransformasi menjadi bentuk vektor aktivitas UMKM sebagai berikut:

$$\begin{aligned} & (\text{wilayah}) \\ & = [IG, FB, TT, EM, LG, KM, AC] \end{aligned}$$

dengan wilayah terdiri atas kecamatan Pakis, Tumpang, Jabung, Poncokusumo di Malang Timur, dimana tiap kecamatan terdiri atas beberapa grup pengusaha yaitu Pakis-1 s.d. Pakis-16, Tumpang-1 s.d.

Tumpang-7, Jabung-1 s.d. Jabung-6, dan Ponokusumo-1 s.d. 2, ditambah sebuah unit di kecamatan Blimbing [8]. Untuk pengkodean pada fitur-fitur media sosial IG, FB, TT, dan EM diberi angka 1 jika sudah digunakan untuk melakukan *branding* dan 0 jika belum; sedangkan LG, KM, dan AC diberi angka 1 jika komponen merek tersebut sudah digunakan dalam produk, dan 0 jika belum.

Hanya 32 dari 43 pengusaha yang memberikan informasi mengenai media sosial dan unsur merek yang mereka gunakan dalam mem-*branding* produknya, dan keseluruhan data primer ini diberikan dalam Tabel 1. Metode klasterisasi Bi K-means yang digunakan adalah *BisectingKMeans* dari putaka pemrograman *scikit-learn* yang diimplementasikan dalam bahasa Python 3.11. Metriks *Silhouette* dan WCSS juga tersedia di pustaka dimaksud. Teknik pemilihan *centroid* inisial yang diterapkan adalah *init* = 'k-means++' yaitu tidak memilih secara acak, namun secara statistik memilih titik-titik data yang tidak mirip, sedangkan parameter *n_init* = 50 yaitu jumlah iterasi untuk memilih titik bakal *centroid* yang berbeda.

Tabel 1. Data Primer Medsos dan Unsur Merek yang Digunakan untuk Branding [8]

wilayah	IG	FB	TT	MP	LG	KM	AC
blimbing	1	1	1	1	0	0	0
jabung-1	1	1	1	1	1	1	1
jabung-2	1	1	1	1	1	1	1
jabung-3	1	1	0	1	0	0	0
jabung-4	1	1	1	1	0	0	0
jabung-5	1	1	1	1	0	0	0
jabung-6	1	1	1	1	0	0	1
pakis-1	1	0	0	1	0	0	0
pakis-2	0	0	1	1	0	0	1
pakis-3	0	0	1	1	0	0	0
pakis-4	1	1	0	1	0	0	1
pakis-5	1	0	0	0	0	0	0
pakis-6	0	0	0	0	0	0	0
pakis-7	1	0	0	0	0	0	0
pakis-8	1	0	0	0	0	0	0
pakis-9	1	1	0	0	0	0	0
pakis-10	1	0	0	0	0	0	1
pakis-11	0	0	0	0	0	0	1
pakis-12	1	0	0	0	0	0	1
pakis-13	1	0	0	0	0	0	0
pakis-14	1	0	0	0	0	0	0
pakis-15	0	0	0	1	0	0	0
pakis-16	0	0	0	0	0	0	0
ponco-1	1	0	0	0	0	0	0
ponco-2	1	0	0	0	0	0	0

tump-1	0	0	0	0	0	0	0
tump-2	1	0	0	0	1	1	0
tump-3	1	0	0	0	1	0	1
tump-4	1	0	0	0	0	0	0
tump-5	1	1	0	0	0	0	0
tump-6	1	0	0	0	0	0	0
tump-7	1	0	0	0	0	0	1

Program Bi K-Means diulangi untuk nilai $k = 2$ sampai dengan 10, kemudian nilai WCSS dan *Silhouette* di tiap k dicetak ke dalam diagram garis *Elbow*. Nilai k yang optimal pada penggunaan nilai WCSS adalah pada saat grafik *Elbow* mulai turun secara melandai di sebuah nilai k dan dari semua k , terpilih k yang paling kecil; sedangkan nilai k optimal pada penggunaan *Silhouette* adalah di nilai k yang mulai naik secara melandai, dan dari semua k , terpilih k yang paling kecil [9], [10].

Metode *Elbow* kadang tidak dapat menemukan nilai k optimal karena bentuk kurva *Elbow* yang melandai. Untuk mengatasi situasi ini digunakan *plot* visual *Silhouette* dengan cara memvisualisasikan komposisi keanggotaan tiap klaster dan membandingkannya dengan nilai rata-rata *Silhouette*, seperti diilustrasikan dalam Gambar 1. Hasil analisis menggunakan *Elbow* dan *plot* visual antar hasil klasterisasi Bi K-Means dan K-means, digunakan untuk menentukan k klaster optimal. Setelah k klaster optimal diperoleh, dilakukan pemingkatan klaster dan interpretasi karakter tiap klaster berdasarkan PMed dan PUns dengan pembahasan sebagai berikut.

Persentase Penggunaan Media dan Unsur Merek dalam Branding

Program *Python* untuk klasterisasi telah disusun untuk selain menghasilkan klaster dan anggota-anggota tiap klaster, juga merangkum penggunaan media sosial dan unsur merek oleh pengusaha dalam kegiatan *branding* di tiap klaster. studi ini mengusulkan rumus perhitungan penggunaan media dan unsur merek sebagaimana dalam rumus (5) dan (6).

$$PMed = \frac{nMed}{nMedia \times nK} \quad (5)$$

$$PUns = \frac{nUns}{nUnsur \times nK} \quad (6)$$

$PMed$ adalah persentase media yang digunakan untuk *branding*, yang dihitung dari jumlah media yang digunakan oleh semua anggota klaster ($nMed$), dibagi jumlah semua media yang ada ($nMedia$) dikali jumlah anggota dalam klaster yang dievaluasi (nK). $PUns$ adalah persentase unsur merek yang digunakan dalam *branding* yang dihitung dari jumlah unsur yang digunakan oleh semua anggota klaster ($nUns$) dibagi jumlah semua unsur yang ada ($nUnsur$) dikali nK .

Sebagai contoh, misalnya terbentuk klaster dengan informasi sebagaimana dalam Tabel 2. Maka $nK = 3$ yaitu jumlah anggota dalam klaster, $nMedia = 4$ yaitu jumlah media yang ada (FB, IG, TT, EM), sedangkan $nP = 4$ yaitu banyaknya media digunakan oleh semua anggota klaster (FB = 1, IG = 2, dan EM = 1); sehingga persentase media digunakan oleh klaster ini adalah $PMed = 4 / (4 * 3) = 33.33\%$. Dengan cara yang sama diperoleh $PUns = 3 / (3 * 3) = 33.33\%$.

Tabel 2. Klaster Percontohan

o	B	G	T	M	G	M	C

Storytelling berbasis *Transformers*

Setelah klasterisasi selesai dan terbentuk k klaster, maka tahapan berikutnya adalah melakukan interpretasi tiap klaster, dan untuk melakukan tugas tersebut, artikel ini mengusulkan penggunaan pendekatan *Transformer* yaitu dengan membuat *storytelling* terhadap karakter UMKM di tiap klaster. Keutamaan dari pembuatan *storytelling* dari hasil klasterisasi adalah interpretasi terhadap klaster bukan sekedar berupa label klaster yang bersifat diskrit misalnya unit usaha yang baik, cukup atau kurang dalam kegiatan *branding* di platform digital, namun berupa penjelasan naratif dan lebih mudah diinterpretasikan.

Transformers dimaksud adalah sebuah arsitektur model kecerdasan artifisial yang didesain untuk mempelajari dan mengingat deretan atau urutan data yang

panjang sebagai sebuah konteks informasi. *Transformers* diusulkan untuk menggantikan mekanisme perulangan dalam *recurrent neural network* atau jaringan saraf berulang sepenuhnya dengan mekanisme atensi sehingga menciptakan keterkaitan yang erat antara data masuk dan informasi yang keluar [11]. *Transformers* menjadi dasar pengembangan model bahasa alamiah skala besar (*large language model*) yang pralatih seperti *Generative Pre-trained Transformers* (GPT) dan *Bidirectional Encoder Representations from Transformers* (BERT). Perbedaan antara keduanya pada arsitektur dan pendekatan dijelaskan sebagai berikut. GPT adalah model berbasis *Transformer* yang menggunakan arsitektur dekoder. GPT hanya memerhatikan konteks sebelumnya dalam pemrosesan teks dan menghasilkan teks yang dihasilkan secara generatif, sedangkan BERT adalah model berbasis *Transformer* yang menggunakan arsitektur encoder. BERT menggabungkan konteks sebelum dan sesudah token target untuk pemrosesan teks dan menggunakan pendekatan "*masked language modeling*" untuk pemodelan bahasa [11]–[13].

Model-model bahasa alamiah berbasis *Transformers* yang di-generate oleh peneliti dan praktisi diunggah ke sebuah repositori yaitu Huggingface.com. Di sini dapat ditemukan turunan GPT dan BERT dalam Bahasa Indonesia, seperti IndoGPT dan IndoBERT, dan semuanya dapat digunakan secara gratis dan bebas untuk di-*fine-tune* kembali menjadi model lainnya [11]–[13]. Tiap model dibangun dengan sebuah tugas yang beragam seperti *text classification*, *text generation*, *feature extraction*, *sentence similarity* dan *summarization*. IndoGPT adalah turunan dari GPT2 dan dibangun untuk tugas *text generation*, sedangkan IndoBERT adalah turunan dari BERT dibangun untuk tugas *feature extraction*; sedangkan untuk menghasilkan narasi *storytelling* dengan konteks yang kompleks dibutuhkan kemampuan pembangkitan teks (*text generation*) yang dilatih dengan korpus berskala besar. IndoGPT belum dilengkapi dengan kemampuan ini, dan hal ini ditunjukkan melalui sebuah contoh

prompt input = “peran pemasaran digital bagi umkm adalah”; dan dijawab dengan

output = “peranan pemasaran digital bagi umkm adalah memberikan solusi bagi para pelaku usaha di bidang perdagangan dan UKM untuk memasarkan produknya melalui pemasaran online ini. Pada kesempatan ini juga diadakan program acara yang di ikuti oleh sekitar 150 UKM serta peserta yang berasal dari berbagai negara di”

Di sini kalimat pertama dari output terlihat cukup relevan dengan *prompt input* yang diberikan, namun kalimat berikutnya terasa memiliki konteks yang melebar dari *prompt input* yang diajukan. Model bahasa Indonesia lain yang tersedia di Huggingface.com, menurut observasi dan eksperimen, memiliki konteks yang kurang sesuai untuk diterapkan dalam pembangkitan *storytelling* di studi ini karena dibentuk dari sumber yang tidak diketahui atau tidak disebutkan.

Sebagai solusi, studi ini menggunakan model GPT-3.5 yang diproduksi oleh OpenAI, yaitu model yang menjadi mesin (*engine*) dari ChatGPT. Model ini bersifat multi-bahasa dan dilengkapi dengan berbagai alat analisis, baik statistika seperti regresi dan statistik deskriptif, data mining seperti K-means dan klustering hirarki, bahkan manajemen seperti analisis SWOT. Untuk *prompt input* “peran pemasaran digital bagi umkm adalah” dijawab oleh GPT-3.5 dengan output:

“Sangat penting dalam membantu UMKM memperluas jangkauan pasar, meningkatkan visibilitas dan kehadiran mereka di dunia online, serta meningkatkan penjualan dan profitabilitas bisnis mereka secara keseluruhan. Berikut adalah beberapa peran penting dari pemasaran digital bagi UMKM: 1. Meningkatkan visibilitas online: Dengan menggunakan strategi pemasaran digital seperti SEO (Search Engine Optimization) dan pemasaran konten, UMKM dapat meningkatkan visibilitas mereka di mesin pencari seperti Google, Bing, dan Yahoo;

dst.” Di sini terlihat bahwa jawaban GPT-3.5 akurat baik dari sisi tata bahasa Indonesia, maupun konteks yang sedang dibicarakan, dibandingkan dengan IndoGPT.

Narasi *storytelling* harusnya mengacu pada sebuah fakta yang dihasilkan dari data [14], [15], yaitu dalam studi ini digunakan data hasil klusterisasi yang telah dirangkum menjadi persentase penggunaan media (*PMed*) dan penggunaan unsur merek (*PUns*) dalam *branding*. GPT-3.5 digunakan untuk menyelesaikan tugas pembangkitan teks menggunakan API *chat completion*. Berdasarkan dokumentasi resmi model GPT di OpenAI.com [16], *Pseudo-code* untuk pembangkitan narasi *storytelling* adalah sebagai berikut:

1. meng-*import library* openai
2. memasukkan API-Key dari openai
3. menentukan role (system/user/asistant)
4. menentukan konten konteks untuk narasi yaitu ‘pendampingan UMKM di Indonesia’
5. menentukan temperatur
6. membaca hasil pengukuran persentase penggunaan media (*PMed*) dan unsur merek (*PUns*) tiap klaster yang dihasilkan oleh klusterisasi Bi K-Means
7. memasukkan *prompt* perintah untuk menghasilkan narasi dari hasil pengukuran *PMed* dan *PUns*
8. mengirimkan *prompt* perintah ke API *ChatCompletion* untuk mendapatkan respons dari model
9. menampilkan respons dari model yang berupa narasi *storytelling* dari hasil klusterisasi

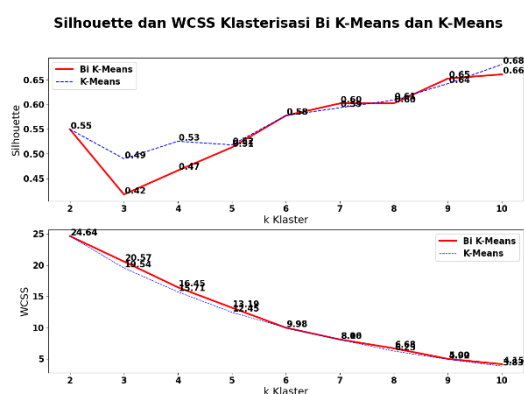
Pseudo-code ini menjelaskan penggunaan *application programming interface* (API) *chat completion* yang disediakan oleh GPT-3.5. Menurut dokumentasi GPT dan API *Chat completion* [17], temperature 0,2 membuat hasil response lebih deterministik, sedangkan 0,8 membuat lebih acak. Perintah dalam kode di atas adalah perintah untuk menghasilkan narasi yang diinginkan dari hasil *PMed* dan *PUns*. Dalam eksperimen, digunakan perintah untuk menghasilkan peringkat tiap klaster berdasarkan *PMed* dan *PUns*.

HASIL DAN PEMBAHASAN

Hasil Klasterisasi Bi K-Means

Pembahasan dimulai dari hasil klasterisasi Bi K-Means dan membandingkannya dengan hasil dari K-Means tradisional. Grafik *Elbow* dari hasil Bi K-Means dan K-Means diperlihatkan dalam Gambar 3, dimana nilai *Silhouette* dan WCSS hasil klasterisasi diberikan masing-masing di bagian atas dan bawah gambar.

Dari grafik *Elbow* diperoleh beberapa catatan, pertama, nilai WCSS dari kedua metode dapat dikatakan sama, dengan bentuk siku (*Elbow*) yang melandai turun di semua titik sehingga cukup sulit untuk menentukan k kluster yang paling optimal. Idealnya, nilai k optimal terlihat pada siku dari grafik yang menonjol di sebuah k sebelum kemudian melandai di k berikutnya. Untuk memperjelas penentuan k , kurva *Elbow* nilai *Silhouette* digambarkan, dan dari sini terlihat bahwa *Silhouette* K-Means sempat terlihat meningkat di $k = 4$, namun turun di $k = 5$, dan naik tajam di $k = 6$. *Silhouette* Bi K-means di $k = 4$ tidak membentuk siku, namun terus menanjak hingga $k = 6$. Dari grafik *Elbow* ini, ditetapkan bahwa klasterisasi akan ditinjau pada $k = 4$ dan 6.

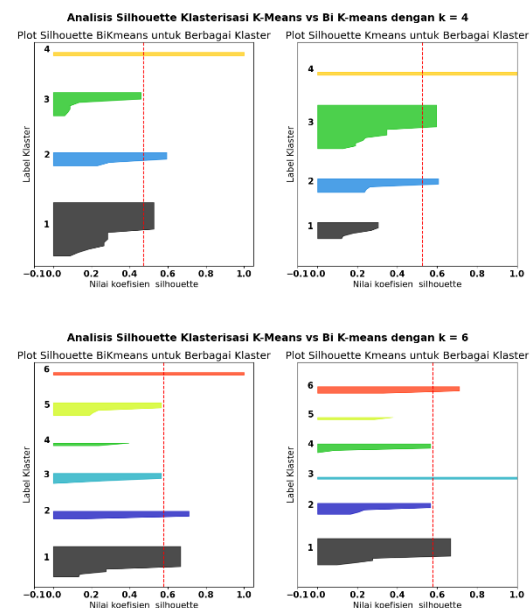


Gambar 3. Plot *Elbow* kedua metode (sumber: pengolahan sendiri)

Untuk mendapatkan wawasan (*insight*) yang lebih jelas, digunakan *plot* visual *Silhouette* dengan cara memvisualisasikan keanggotaan tiap kluster untuk $k = 4$ dan 6 dari kedua metode, dan membandingkannya dengan nilai tengah

Silhouette (garis merah putus-putus) sebagaimana ditunjukkan dalam Gambar 4.

Gambar 4 memberikan beberapa wawasan bahwa pada $k = 4$, Bi-KMeans dan K-Means memiliki jumlah keanggotaan yang tidak sama di beberapa kluster yang dihasilkan. *Silhouette* antar kluster hasil Bi K-Means hampir seimbang satu dengan lainnya dan dekat dengan nilai tengah *Silhouette*; sementara kluster hasil K-Means tidak seimbang khususnya pada kluster berlabel 1, yang memiliki nilai yang jauh di bawah nilai tengah meskipun kurva *Elbow* K-Means menunjukkan nilai *Silhouette* yang lebih tinggi dari Bi K-Means di $k = 4$. Dari sini disimpulkan untuk $k = 4$, hasil Bi K-Means lebih efektif untuk mendapatkan wawasan kluster-kluster, dibandingkan hasil K-Means. Untuk $k = 6$, meskipun penomoran label kluster tampak berbeda, namun nilai *Silhouette* hasil kedua metode sama persis dan observasi menemukan bahwa anggota-anggota kluster hasil kedua metode sama persis.



Gambar 4. Visualisasi *Silhouette* tiap Kluster kedua metode (sumber: pengolahan sendiri)

Tabel 3 dan 4 masing-masing memperlihatkan keanggotaan di dalam empat dan enam kluster hasil Bi K-Means yang berisi informasi mengenai kode kluster, jumlah anggota kluster (n), anggota-anggota kluster, dan *PMed* dan *PUns*.

Tabel 3. Keanggotaan tiap Klaster $k=4$

cluster 0, n = 5 anggota : ['pakis-6', 'pakis-11', 'pakis-15', 'pakis-16', 'tumpang-1'] penggunaan media: 5.00%, unsur: 6.67%
cluster 1, n = 17 anggota : ['pakis-1', 'pakis-5', 'pakis-7', 'pakis-8', 'pakis-9', 'pakis-10', 'pakis-12', 'pakis-13', 'pakis-14', 'poncokusumo-1', 'poncokusumo-2', 'tumpang-2', 'tumpang-3', 'tumpang-4', 'tumpang-5', 'tumpang-6', 'tumpang-7'] penggunaan media: 29.41%, unsur: 13.73%
cluster 2, n = 8 anggota : ['blimbing', 'jabung-3', 'jabung-4', 'jabung-5', 'jabung-6', 'pakis-2', 'pakis-3', 'pakis-4'] penggunaan media: 81.25%, unsur: 12.50%
cluster 3, n = 2 anggota : ['jabung-1', 'jabung-2'] penggunaan media: 100.00%, unsur: 100.00%

Tabel 4. Keanggotaan tiap Klaster $k=6$

cluster 0, n = 5 anggota : ['pakis-6', 'pakis-11', 'pakis-15', 'pakis-16', 'tumpang-1'] penggunaan media: 5.00%, unsur: 6.67%
cluster 1, n = 4 anggota : ['pakis-10', 'pakis-12', 'tumpang-3', 'tumpang-7'] penggunaan media: 25.00%, unsur: 41.67%
cluster 2, n = 13 anggota : ['pakis-1', 'pakis-5', 'pakis-7', 'pakis-8', 'pakis-9', 'pakis-13', 'pakis-14', 'poncokusumo-1', 'poncokusumo-2', 'tumpang-2', 'tumpang-4', 'tumpang-5', 'tumpang-6'] penggunaan media: 30.77%, unsur: 5.13%
cluster 3, n = 2 anggota : ['pakis-2', 'pakis-3'] penggunaan media: 50.00%, unsur: 16.67%
cluster 4, n = 6 anggota : ['blimbing', 'jabung-3', 'jabung-4', 'jabung-5', 'jabung-6', 'pakis-4'] penggunaan media: 91.67%, unsur: 11.11%
cluster 5, n = 2 anggota : ['jabung-1', 'jabung-2'] penggunaan media: 100.00%, unsur: 100.00%

Hasil *Storytelling* berbasis *Transformers*

Program pembentuk narasi *storytelling* dijalankan dengan *prompt* berupa perintah untuk melakukan pemeringkatan dan menghasilkan narasi hasil klastering, mengacu pada Tabel 3 dan 4. Hasil ini menunjukkan kemampuan GPT-3.5 dalam membantu menginterpretasi hasil klasterisasi secara naratif, sekaligus memberikan peringkat klaster sesuai perintah yang diberikan.

Berikut ini adalah narasi *storytelling* yang dihasilkan oleh GPT-3.5 untuk $k = 4$

Pemeringkatan klaster menurut strategi branding dari tinggi ke rendah adalah sebagai berikut: Klaster 3, Klaster 2, Klaster 1, Klaster 0

Klaster 3:

Klaster ini terdiri dari 2 bisnis dengan tingkat penggunaan media sosial dan unsur merek paling tinggi di antara semua klaster. Kedua bisnis dalam klaster ini menggunakan media sosial dengan persentase 100% dan unsur merek dengan persentase 100%. Hal ini menunjukkan komitmen yang sangat tinggi dalam membangun kehadiran merek mereka di media sosial dan mengintegrasikan unsur merek ke dalam kegiatan branding. Klaster ini menempati peringkat tertinggi dalam strategi branding karena penggunaan media dan unsur merek yang sangat intens.

Klaster 2:

Klaster ini terdiri dari 8 bisnis dengan persentase penggunaan media sosial sebesar 81.25% dan penggunaan unsur merek sebesar 12.50%. Bisnis dalam klaster ini menunjukkan tingkat komitmen yang cukup tinggi dalam memanfaatkan media sosial sebagai bagian dari strategi branding. Namun, penggunaan unsur merek masih perlu ditingkatkan untuk mencapai intensitas yang lebih tinggi.

Klaster 1:

Klaster ini terdiri dari 17 bisnis dengan persentase penggunaan media sosial sebesar 29.41% dan penggunaan unsur merek sebesar 13.73%. Meskipun ada upaya dalam penggunaan media sosial, persentase penggunaan masih relatif rendah dibandingkan klaster lainnya. Penggunaan unsur merek juga perlu ditingkatkan untuk mencapai intensitas yang lebih tinggi. Klaster ini menempati peringkat menengah dalam hal strategi branding karena penggunaan media sosial dan unsur merek yang cukup, tetapi masih memiliki ruang untuk peningkatan.

Klaster 0:

Klaster ini terdiri dari 5 bisnis dengan persentase penggunaan media sosial sebesar 5.00% dan penggunaan unsur merek sebesar 6.67%. Penggunaan media sosial dan unsur merek dalam klaster ini tergolong rendah dibandingkan klaster lainnya. Bisnis dalam klaster ini perlu meningkatkan intensitas penggunaan media sosial dan penggunaan unsur merek sebagai bagian dari strategi branding mereka. Klaster ini menempati peringkat terendah dalam hal strategi branding.

Berikut ini adalah narasi *storytelling* yang untuk $k = 6$:

Pemeringkatan klaster menurut strategi branding dari tinggi ke rendah adalah sebagai berikut: Klaster 5, Klaster 4, Klaster 1, Klaster 3, Klaster 2, Klaster 0

Klaster 5:

Klaster ini terdiri dari 2 bisnis yang memiliki tingkat penggunaan media sosial dan unsur merek paling tinggi di antara semua klaster. Kedua bisnis dalam klaster ini menggunakan media sosial dengan persentase 100% dan unsur merek dengan persentase 100%. Hal ini menunjukkan komitmen yang sangat tinggi dalam membangun kehadiran merek mereka di media sosial dan mengintegrasikan unsur merek ke dalam kegiatan branding. Klaster ini menempati peringkat tertinggi dalam strategi branding karena penggunaan media dan unsur merek yang sangat intens.

Klaster 4:

Klaster ini terdiri dari 6 bisnis dengan persentase penggunaan media sosial sebesar 91.67% dan penggunaan unsur merek sebesar 11.11%. Bisnis dalam klaster ini menunjukkan tingkat penggunaan media sosial yang tinggi dalam strategi branding mereka. Namun, penggunaan unsur merek masih perlu ditingkatkan untuk mencapai intensitas yang lebih tinggi.

Klaster 1:

Klaster ini terdiri dari 4 bisnis dengan persentase penggunaan media sosial sebesar 25.00% dan penggunaan unsur merek sebesar 41.67%. Meskipun penggunaan media sosial dalam klaster ini masih relatif rendah, bisnis dalam klaster ini menunjukkan komitmen yang tinggi dalam membangun dan mengintegrasikan unsur merek ke dalam strategi branding mereka.

Klaster 3:

Klaster ini terdiri dari 2 bisnis dengan persentase penggunaan media sosial sebesar 50.00% dan penggunaan unsur merek sebesar 16.67%. Meskipun penggunaan media sosial dan unsur merek dalam klaster ini cukup seimbang, masih ada ruang untuk peningkatan intensitas penggunaan keduanya.

Klaster 2:

Klaster ini terdiri dari 13 bisnis dengan persentase penggunaan media sosial sebesar 30.77% dan penggunaan unsur merek sebesar 5.13%. Penggunaan media sosial dalam klaster ini cukup konsisten, namun penggunaan unsur merek masih perlu ditingkatkan untuk mencapai intensitas yang lebih tinggi.

Klaster 0:

Klaster ini terdiri dari 5 bisnis dengan persentase penggunaan media sosial sebesar 5.00% dan penggunaan unsur merek sebesar 6.67%. Penggunaan media sosial dan unsur merek dalam klaster ini masih relatif rendah dibandingkan klaster lainnya. Bisnis dalam klaster ini perlu meningkatkan intensitas penggunaan media sosial dan penggunaan unsur

merek sebagai bagian dari strategi branding mereka.

PENUTUP

Hasil eksperimen dalam studi ini menghasilkan beberapa wawasan. Pertama, kombinasi penggunaan metode *Elbow* terhadap nilai *Silhouette* dan WCSS, serta plot visual nilai *Silhouette* dapat membantu menentukan jumlah k klaster optimal dalam penggunaan Bi K-Means dan K-Means, terutama ketika kurva *Elbow* turun atau naik melandai sehingga tidak dapat menunjukkan posisi k yang optimal. Di dalam eksperimen, klasterisasi Bi K-Means dinyatakan lebih baik karena menghasilkan klaster-klaster dengan nilai *Silhouette* yang seimbang, daripada klasterisasi K-Means. Pembangkitan narasi *storytelling* dari hasil klasterisasi menggunakan *Transformers* seperti GPT-3.5 dapat membantu meningkatkan interpretasi terhadap hasil klasterisasi selain dapat menentukan peringkat klaster sesuai persentase penggunaan media dan unsur merek dalam branding.

GPT-3.5 menetapkan tarif atas penggunaan API-nya, meskipun relatif terjangkau; namun sebagai saran, penelitian selanjutnya dapat menelaah beberapa *Transformers* untuk Bahasa Indonesia yang tersedia di Huggingface.com, atau mengembangkan model transformers seperti GPT atau BERT untuk menyelesaikan tugas yang sejenis dengan API *chat completion* milik OpenAI.

DAFTAR PUSTAKA

- [1] T. Rumayanto, S. N. Sanusi, and S. Sihombing, "Strategi Pengembangan UMKM Pasca COVID-19," *J. Educ.*, 2022, [Online]. Available: <https://jonedu.org/index.php/joe/article/view/698>
- [2] F. D. Pranatasari, "AGILE MARKETING SEBAGAI SOLUSI BUSINESS DISRUPTION MENUJU KEBANGKITAN BISNIS UMKM PASCA COVID-19," *Modus*, vol. 33, no. 2, pp. 196–211, Jul. 2021, doi: 10.24002/modus.v33i2.4662.
- [3] T. A. Auliandri, A. Purmiyati, M. Mustain, A. Setyawan, G. A. Pramesti,

- and G. A. Nastiti, "Manfaat Penggunaan Teknologi Pada UMKM dalam Penguatan Ekonomi Jawa Timur Selama Pandemi Covid-19," *INOVIS J. Inov. Bisnis dan Manaj. Indones.*, vol. 5, no. 3, 2022, doi: 10.31842/jurnalnobis.v5i3.240.
- [4] A. Aminy and K. Fithriasari, "ANALISIS DAMPAK COVID-19 BAGI UMKM DI JAWA TIMUR," *Semin. Nas. Off. Stat.*, vol. 2020, no. 1, 2021, doi: 10.34123/semnasoffstat.v2020i1.646.
- [5] W. Hanjarwadi, "Targetkan 30 Juta UMKM 'Onboarding' ke Platform Digital," 2021. <https://www.pajak.com/ekonomi/targetkan-30-juta-umkm-onboarding-ke-platform-digital/2/>
- [6] P. Natalia, "Onboarding UMKM Bank Indonesia," 2021. <https://onboardingumkm.id/>
- [7] F. Asharudin and S. P. Dewi, "Optimalisasi Media Digital sebagai Strategi Pemasaran UMKM pada Masa Pandemi Covid-19," *E-Dimas J. Pengabd. Kpd. Masy.*, vol. 12, no. 3, 2021, doi: 10.26877/e-dimas.v12i3.7142.
- [8] T. M. Akhriza, E. Sophia, and W. A. Dewa, "Evaluasi Pendampingan Branding berbasis Platform Digital bagi UMKM se-Malang Raya," in *SMATIKA JURNAL*, Malang: STIKI Malang, 2022. doi: 10.32664/smatika.v12i02.734.
- [9] N. Puspitasari, J. A. Widians, and N. B. Setiawan, "Segmentasi pelanggan menggunakan algoritme bisecting k-means berdasarkan model recency, frequency, and monetary (RFM)," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, 2020.
- [10] S. Dwididanti and D. A. Anggoro, "Analisis Perbandingan Algoritma Bisecting K-Means dan Fuzzy C-Means pada Data Pengguna Kartu Kredit," *Emit. J. Tek. Elektro*, vol. 22, no. 2, 2022, doi: 10.23917/emitor.v22i2.15677.
- [11] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
- [12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019.
- [13] L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds and Machines*, vol. 30, no. 4. 2020. doi: 10.1007/s11023-020-09548-1.
- [14] B. M. Winn, "The Design, Play, and Experience Framework," in *Handbook of Research on Effective Electronic Gaming in Education*, 2011. doi: 10.4018/978-1-59904-808-6.ch058.
- [15] C. N. Knaflic, *Storytelling with data: A data visualization guide for business professionals*, 1st ed. New Jersey: John Wiley & Sons, Inc., 2015.
- [16] Anonim, "GPT Models," *OpenAI.com*, 2023. <https://platform.openai.com/docs/guides/gpt>
- [17] Anonim, "Create chat completion," *OpenAI.com2*, 2023.