Analisis *Text Clustering* Kebijakan Pembukaan Daerah Wisata pada Masa Pandemi Berbasis Densitas Spasial (DBSCAN)

Rahmah Wulandari[1], Wiyli Yustanti, S.Si., M.Kom.[2]

[1,2]*Program Studi Sistem Informasi, Fakultas Teknik, Universitas Negeri Surabaya*
[1]rahmah.17051214058@mhs.unesa.ac.id

[2]wiyliyustanti@unesa.ac.id

***Abstrak*— Clustering merupakan metode pengelompokkan data ke dalam suatu kelompok atau klaster menggunakan parameter tertentu sehingga objek dalam suatu klaster memiliki tingkat kemiripan yang sama. Pada penelitian ini dilakukan analisis text clustering terhadap komentar video youtube yang membahas tentang kebijakan pembukaan daerah wisata pada masa pandemi menggunakan algoritma DBSCAN serta membandingkannya dengan algoritma K-Means. Hasil dari penelitian ini diperoleh Silhouette Score sebesar 0.732 untuk algoritma DBSCAN dan sebesar 0.637 untuk algoritma K-Means. Hasil analisis dan identifikasi topik klaster DBSCAN menunjukkan bahwa klaster yang terbentuk menggunakan DBSCAN lebih baik daripada K-Means. Hal tersebut dapat terlihat dari kata-kata yang paling sering muncul pada tiap klaster. Klaster pertama dan ketiga yang terbentuk menggunakan K-Means masih terdapat kata-kata yang sama muncul, yakni kata "moga" dan "masuk". Topik tiap klaster menggunakan DBSCAN juga lebih mudah disimpulkan daripada K-Means karena tiap klasternya terkategori dengan baik berdasarkan jenis topik dalam komentarnya. Topik pada klaster pertama menggunakan DBSCAN yaitu mengenai kebijakan karantina di Indonesia pada pembukaan negara bagi wisatawan mancanegara, sedangkan klaster kedua mengenai harapan agar cepat bangkit dan ucapan syukur atas kebijakan yang diterapkan. Selain itu DBSCAN juga menghasilkan noise sebanyak 9 noise. Maka disimpulkan bahwa penggunaan algoritma DBSCAN lebih baik daripada algoritma K-Means untuk mengelompokkan data teks berupa komentar.**

***Kata Kunci*—Text Custering, DBSCAN, K-Means**

I. PENDAHULUAN

Pariwisata secara umum diartikan sebagai fenomena sosial, budaya, dan ekonomi yang melibatkan perpindahan orang ke tempat-tempat di luar lingkungan biasa mereka untuk liburan, bisnis atau tujuan lain berturut-tururt selama tidak lebih dari satu tahun.

Pandemi COVID-19 membuat dunia menghadapi kondisi daruat dalam hal kesehatan, sosial, dan ekonomi global yang belum pernah terjadi sebelumnya. Pada 31 maret 2020, setelah WHO menetapkan COVID-19 sebagai pandemi, pemerintah Indonesia membuat kebijakan pembatasan sosial berskala besar (PSBB). Kemudian seiring berjalannya waktu, kebijakan tersebut berganti nama dan format beberapa kali hingga menjadi Pemberlakukan Pembatasan Kegiatan Masyarakat (PPKM) 4 level [1].

Pariwisata merupakan salah satu sektor yang paling terkena dampak dari adanya pemberlakuan berbagai pembatasan perjalanan oleh negara-negara yang berusaha menahan penyebaran virus. Menurut Organisasi Pariwisata Dunia (*United Nations World Tourism Ogranization*/UNWTO), pada kuartal pertama tahun 2021 kedatangan wisatawan internasional menurun 83% dibandingan periode yang sama tahun lalu [2]. Begitu pula dengan Badan Pusat Statistik mencatat kunjungan wisatawan mancanegara (wisman) ke Indonesia mengalami penurunan. Pada bulan Januari hingga Agustus 2021, tercatat total kunjungan wisman mencapai 1,06 juta. Angka tersebut menurun 69,17% jika dibandingan dengan total kunjungan wisman pada periode yang sama tahun 2020 sebanyak 3,44 juta kunjungan [3].

Pemerintah Indonesia selain dituntut untuk sigap dalam mengantisipasi adanya penyebaran dan lonjakan kasus COVID-19, juga dituntut dalam menjaga kestabilan ekonomi. Menurunnya kunjungan wisatawan mancanegara membuat perekonomian dalam sektor pariwisata melemah. Dampak dari hal itu sangat terasa oleh pelaku usaha di bidang pariwisata. Mengikuti perkembangan penyebaran kasus COVID-19 yang mulai terjadi penurunan, pemerintah mulai berencana untuk membuka daerah wisata. Pada 14 Oktober 2021, Bali resmi dibuka untuk wisatawan mancanegara (wisman). Bali merupakan daerah yang perekonomiannya sangat bergantung pada sektor pariwisata. Tingkat vaksinasi telah mencapai 99% utuk dosis pertama dan hampir 90% untuk dosis kedua serta mayoritas usaha pariwisata yang telah memperoleh sertifikasi CHSE (*Cleanliness*, *Health*, *Safety*, dan *Environmental Sustainbility*) menjadi alasan atas siapnya Bali dibuka bagi wisman. Pintu masuk Bali dibuka lagi bagi wisman dengan persyaratan dan peraturan yang sudah ditetapkan oleh pemerintah [4].

Kebijakan dan aturan yang diterapkan pemerintah biasanya dipublikasikan di berbagai media, baik media tulis hingga program acara berita. Dalam *platform* Youtube, video dalam kategori Berita diunggah oleh kanal media televisi maupun portal media *online* untuk menjangkau masyarakat yang menyukai program acara berita. Komentar pengguna dalam video dapat dianalisis untuk melihat respon individu terhadap

kebijakan yang diberitakan serta dijadikan pertimbangan bagi pembuat kebijakan. Komentar tersebut berbentuk teks, sehingga perlu dilakukan analisis *text mining*.

Text mining adalah disiplin ilmu dalam bidang *data mining* yang mempelajari tentang pengolahan data teks yang dilakukan secara otomatis dengan tujuan untuk menggali informasi yang baru dari kumpulan data dalam jumlah besar. *Text mining* memberikan solusi pada permaslahan seperti pengelompokan dan analisa *unstructured text* dalam jumlah besar. Akan tetapi, memproses komentar untuk mengekstrak informasi yang bermakna sangat menantang, setidaknya karena terdapat dua alasan: (i) penggunaan kata dan atau ejaan yang tidak baku, dan (ii) masalah code-switching [5]. Penyatuan satuan-satuan linguistik seperti frase, kata, dan morfem suatu Bahasa ke dalam penggunaan Bahasa lain yang berdeda dikenal sebagai *code-switching* [6].

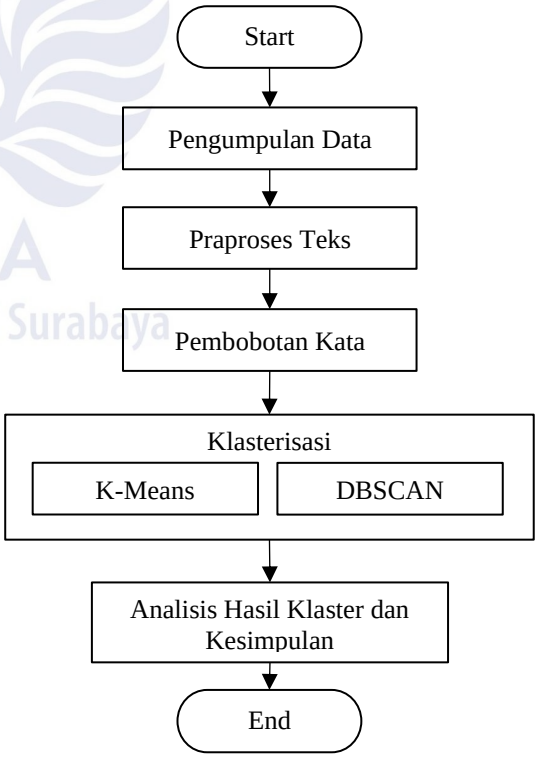
Fenomena *code-switching* menyebabkan kemungkinan terjadinya masalah karena variasi tata bahasa dan ejaan. Negara Indonesia dinobatkan sebagai negara trilingual teratas di dunia serta peringkat dua sebagai negara bilingual [7]. Bahasa Indonesia adalah satu-satunya bahasa resmi, bahasa inggris juga digunakan dalam pendidikan formal dan bisnis. Selain itu, masih terdapat kurang lebih 700 bahasa daerah dan bahasa jawa merupakan bahasa daerah yang paling banyak digunakan. Dalam thesis Christina [8], ditemukan bahwa penguna sosial media di Indonesia, menggunakan beragam bahasa, khususnya bahasa gaul, Bahasa Indonesia, dan Bahasa Inggris formal dan informal untuk mengekspresikan perasaan solidaritas dengan komunitas mereka. Keberagaman suku, bahasa daerah, dan budaya mendukung fenomena code-switching terjadi saat berkomentar terhadap suatu *issue* di media sosial, dalam konteks *platform* youtube.

Salah satu teknik yang umum digunakan dalam penelitian *text mining* adalah *clustering*. *Clustering* adalah teknik yang digunakan untuk mengelompokkan data ke dalam suatu klaster menggunakan parameter tertentu sehingga objek dalam satu klaster memiliki tingkat kemiripan yang sama[9].Salah satu metode *clustering* yang ada adalah DBSCAN (*Density Based Spatial Clustering Application with Noise*). Algoritma DBSCAN merupakan algoritma non-parametik dalam pembelajaran tak terarah, sehingga tidak memerlukan asumsi untuk beroperasi. Secara konseptual, DBSCAN dapat membentuk klaster yang longgar, acak, dan lebih mudah membentuk cluster jika terdapat *noise* atau *outlier* pada klaster tersebut[10][11][12].

Penelitian sebelumnya tentang *Clustering* opini masyarakat tentang bencana alam di Indonesia menggunakan algoritma DBSCAN dan K-Medoids menemukan bahwa algoritma DBSCAN menghasilkan *cluster* yang mampu menangani *noise*/*outlier*, hasil *cluster* lebih akurat, dan baik untuk data dengan jumlah besar jika dibandingkan dengan K-Medoids [9]. Peneltian lain yang menggunakan DBSCAN untuk *clustering* dilakukan pada konten buatan pengguna Weibo membuktikan bahwa metode ini efektif digunakan. Pusat *clustering* pada penelitian ini mewakili minat pengguna hingga batas tertentu dengan menyempurnakan teks tema iklan yang akan direkomendasikan. Selain itu juga didapatkan kesamaan antara *clustering center* dengan tema yang direkomendasikan, dan ategori iklan dengan kemiripan tinggi yang direkomendasikan kepada pengguna [13]. Penelitian lainnya yang bertujuan membentuk segmen-segmen pengguna *e-money* menggunakan algoritma DBSCAN membentuk 2 segmen dan menghasilkan nilai *Silhouette Coefficient* sebesar 0,26 dan 17 *noise*. [10]. Penelitian sejenis tentang *text clustering* menggunakan metode DBSCAN dan membandingkannya dengan metode K-Means pada data *tweets* membuktikan keunggulan DBSCAN dalam mengelompokkan data *tweets* yang dialamatkan kepada beberapa layanan ekspedisi karena menghasilkan nilai *Silhouette Coefficient* yang lebih tinggi [14].

Pada penelitian ini, algoritma DBSCAN akan digunakan untuk analisis *text clustering* mengenai kebijakan pembukaan daerah wisata pada masa pandemi. Selain melakukan analisis text clustering, penelitian ini juga akan membandingkan hasil *clustering* algoritma DBSCAN dengan algoritma K-Means. Tujuannya yakni agar dapat memberikan informasi algoritma mana yang memiliki performa lebih baik dalam mengklasterisasi data teks. DBSCAN memiliki keunggulan yaitu jauh lebih efisien dalam mencari klaster yang bentuknya berubah-ubah serta dapat menemukan cluster yang bentuknya tak tentu. Sedangkan K-Means memiliki keunggulan dalam mengelompokkan data dalam jumlah besar serta cukup sederhana dan cepat dalam proses klasterisasi. Hasil analisis *text clustering* dari penelitian ini diharapan dapat menjadi bahan evaluasi bagi pembuat kebijakan.

II. METODE PENELITIAN



Gbr 1. Kerangka Penelitian

Pada penelitian ini, metode *text clustering* dengan algoritma DBSCAN akan dibandingkan dengan algoritma K-Means. Tahapan yang dilakukan dalam penelitian ini digambarkan seperti pada Gbr 1.

A. Pengumpulan Data

Dalam pengumpulan data, diambil data komentar pengguna youtube pada video berita yang membahas mengenai kebijakan pembukaan daerah wisata di masa pandemi. Kebijakan yang dimaksud lebih spesifik yakni pembukaan daerah wisata untuk wisatawan mancanegara (wisman) di daerah Bali. Data komentar yang diambil merupakan komentar yang diunggah pengguna dalam rentang waktu bulan September hingga November 2021.

B. Praproses Teks

Praproses teks merupakan tahapan penting dalam *text mining*. Tahap Praproses mengolah data mentah menjadi data bersih sehingga dapat mempermudah proses *clustering*. Tahapan praproses teks terdiri dari.

1. *Case Folding*, pada tahapan ini komentar akan diubah menjadi teks dengan huruf kecil (non kapital) dan tanda baca akan dihilangkan.
2. *Cleaning*, pada tahap ini komentar akan dibersihan dari hal-hal yang tidak diperlukan seperti karakter angka, tanda baca, HTML, alamat URL, emotikon, dan spasi berlebih.
3. *Slangword Replacement*, pada tahap ini kata-kata slang dalam komentar akan diubah menjadi kata baku berdasarkan kamus *slangword* yang telah dikumpulkan.
4. *Tokenizing*, pada tahap ini, komentar akan dipisahkan dan dipecahkan menjadi kata per kata.
5. *Stopword Removing*, pada tahap ini kata dalam komentar akan dihilangkan jika terdapat kata dalam daftar kamus *Stopswords*. Daftar kamus *stopwords* didapatkan dari *library* Sastrawi.
6. *Stemming*, pada tahap ini kata-kata dalam komentar akan dibuah bentuknya menjadi kata dasar dengan menghilangkan kata imbuhan. *Stemming* menggunakan *library* Sastrawi.

C. Pembobotan Kata

Setelah tahapan praposes teks yang menghasilkan kumpulan term atau kata, tahapan selanjutkan akan dilakukan pembobotan kata yang nantinya tiap kata akan diberikan bobot atau nilai. Bobot atau nilai tersebut akan mengindikasikan pentingnya sebuah kata terhadap komentar. Tujuannya yaitu untuk mengetahui kemiripan dan ketersediaan suatu kata dalam komentar. Semakin banyak kata tersebut muncul maka semakin tinggi bobot atau nilai kata tersebut. Pada proses pembobotan kata, metode yang digunakan adalah metode TF-IDF.

Term Frequency-Inverse Document Frequency (TF-IDF) yaitu sebuah metode algoritma yang berguna untuk menghitung bobot atau nilai setiap kata yang umum digunakan. TF-IDF mengevaluasi seberapa penting suatu kata dalam dokumen. Hal tersebut bergantung pada berapa kali kata itu muncul dalam dokumen [9]. Persamaan yang membentuk TF-IDF dapat dilihat pada persamaan 1 dan 2 di bawah ini.

$$W_{i,j} = TF_{i,j} \times IDF_j \qquad (1)$$

$$IDF_j = log\left(\frac{N}{DF_j}\right) \qquad (2)$$

Keterangan:
$W_{i,j}$ = bobot dari kata ke j pada komentar ke i
DF_j = banyaknya komentar yang mengandung kata j
$TF_{i,j}$ = jumlah kemunculan kata ke j pada komentar ke i
IDF_j = inverse document frequency pada kata ke j
N = jumlah keseluruhan komentar

D. Klasterisasi

Setelah melewati tahap pembobotan kata, hasil proses perhitungan dari TF-IDF akan dibentuk menjadi suatu vektor. Setelah itu dilanjutkan dengan tahap klasterisasi. Penelitian ini menggunakan algoritma K-Means dan DBSCAN untuk klasterisasi data.

Penggunaan algoritma K-Means cukup sensitif untuk inisialisasi cluster centroid karena dilakukan secara random. Algoritma K-Means menggunakan mean sebagai pusat *cluster*. Berikut adalah langkah-langkah dari algoritma K-Means.

a. Inisialisasi nilai k secara *random* untuk *centroid*, nilai k ditetapkan berdasarkan hasil penghitungan s*ilhouette method*. Nilai silhouette yang tertinggi diambil sebagai nilai k.
b. Setiap data dibagi menjadi k cluster dan pusat cluster diperoleh dengan menggunakan *Euclidean Distance* seperti pada persamaan 3.

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \qquad (3)$$

Keterangan:
d_{ij} = jarak antar objek *i* dengan *j*
x_{ik} = nilai objek *i* pada variabel ke-*k*
x_{jk} = nilai objek *j* pada variabel ke-*k*
n = banyaknya variabel yang diamati

c. Setiap pusat *cluster* dihitung ulang berdasarkan nilai rata-rata pada cluster yang diperoleh.
d. Ulangi langkah dua dan tiga jika ada perubahan pada grup *cluster*. Proses akan berhenti jika tidak ada perubahan pada *cluster*.

DBSCAN merupakan metode clustering yang membangun cluster berdasarkan kepadatan, *cluster* yang tidak termasuk objek dianggap *noise*. Praktik DBSCAN membutuhkan waktu yang sangat lama karena penggunaan metode ini dilakukan dengan mencari epsilon dan minimum poin secara acak untuk mendapatkan klaster tertentu. Langkah-langkah untuk melengkapi algoritma DBSCAN adalah sebagai berikut.

a. Inisialisasi parameter Min Pts dan dan parameter Eps.
b. Tentukan titik awal atau p secara acak.

c. Hitung Eps atau semua jarak titik yang kepadatan atau densitasnya dicapai terhadap p menggunakan rumus *Euclidian Distance* seperti dalam persamaan 3.
d. Terbentuk sebuah *cluster* ketika titik yang memenuhi Eps lebih dari MinPts dan titik p sebagai *core point*.
e. Ulangi langkah 3 dan 4 sampai terproses pada semua titik. Jika p merupakan titik *border* dan tidak ada yang kepadatan atau densitasnya dapat dicapai p, maka proses dilanjutkan ke titik yang lain.

E. Analisis Hasil Klaster dan Kesimpulan

Pada tahap terakhir, hasil klasterisasi akan dianlisis dan diidentifikasi topiknya. Sehingga akan terbentuk sebuah kesimpulan atas hasil analisis *text clustering* berdasarkan apa yang dituju dalam penelitian ini.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Penelitian ini mengambil dataset komentar video youtube dengan judul atau topik tentang kebijakan pembukaan daerah wisata di masa pandemi. Dataset yang digunakan diambil melalui Youtube API dengan proses *crawling* menggunakan *google sheet*. Sebanyak 1414 data komentar diperoleh dari unggahan pengguna sejak bulan September 2021 hingga November 2021. Contoh data tersebut dapat dilihat pada tabel I dibawah ini.

TABEL I
COTOH DATA MENTAH

Name	Comment	Time	Likes	Reply Count
Kang Gorengan	Ppkm level 3 😂	2021-11-20T06:31:20Z	0	0
San Nada	Katanya mau tahun baru PPKm level 3 saja sudah di berlakukan lagi	2021-11-20T06:00:33Z	0	0
pucet gaming	Semoga hindu dibali makin banyak penganutnya ,bali tetap hindu ....,dan makin berkembang daerah nya ..... Karena hindu yang tersisa tinggal ada dibali..... Salam saudara sebangsa....	2021-11-20T05:20:36Z	0	0
Hobi Browsing	Bali harus terus menjaga tradisi dan budaya tanpa perlu dirubah rubah... karna tradisi dan budaya, dunia selalu mengenalnya	2021-11-19T08:48:37Z	1	0
filmrebahin	Kok orang china duluan lebih tau ya?....	2021-11-18T21:53:35Z	0	0
Sutris Setiawan	Pak lutut lagi	2021-11-18T07:23:47Z	0	0
Oktio Dwi Laksana Putra	Bali emang nggak ada duanya, salam semeton dari lombok	2021-11-18T06:51:52Z	0	0
vanz gaming	Sy sebagai warga Bali asli SDH merasa lumpuh setengah badan KRNA covid ini, aplgi d Bali dh 100% vaksin 2x,,, buat ap vaksin 2x KLO msih takut2an,, kami tidak menginginkan bantuan beras yg gak seberapa yg kami inginkan pariwisata buka dan kami bisa kembali bekerja🙏	2021-11-18T03:51:53Z	0	0
Ruby 루비	Cuma media indonesia yang ngehebohin media china karena media China heboh	2021-11-18T01:01:58Z	0	0
…	…	…	…	…
Dika?	Min pin dong	2021-10-12T04:00:33Z	0	0

B. Praproses Teks

Sebelum diolah, komentar-komentar tersebut melewati tahap praposes terlebih dahulu, tahap pertama yaitu *cleaning*. Pada tahap *cleaning* dibersihkan dari emotikon, angka, tanda baca, spasis berlebih, serta alamat URL. Tahap *Case Foling* sudah termasuk ke dalam tahap *Cleaning*, yang mana semua huruf dalam komentar akan diubah menjadi huruf kecil. Contoh hasil tahap ini dapat dilihat pada tabel II.

TABEL II
CONTOH HASIL TAHAP CLEANING & TOKENIZING

Komentar Mentah	Hasil *Cleaning* & *Case Folding*
Harus dengan syarat yang ketat bagi warga asing masuk Indonesia agar tdk menjadi boomerang,semoga perekonomian bidang pariwisata kembali normal.	harus dengan syarat yang ketat bagi warga asing masuk indonesia agar tdk menjadi bomerang semoga perekonomian bidang pariwisata kembali normal
Selagi ada karantina sepertinya mw dibuka pariswisata tuk wisatawan asing agak berat untuk dtg... Karena biaya hotel karantina mahal.. Tp ini langkahnya bagus untuk	selagi ada karantina sepertinya mw dibuka pariswisata tuk wisatawan asing agak berat untuk dtg karena biaya hotel karantina mahal tp ini langkahnya bagus untuk

Komentar Mentah	**Hasil *Cleaning* & *Case Folding***
mendongkrak ekonomi dibali smg smwnya aman terkendalii..	mendongkrak ekonomi dibali smg smwnya aman terkendali

Komentar mentah masih terdapat kata *slang* atau ejaan kata yang tidak baku, berikut juga akronim singkatan sebuah kata. Oleh karena itu, *slangword removal* sangat diperlukan dalam praproses teks. Kumpulan *slangword* atau bahasa gaul dalam Bahasa Indonesia semakin beragam seiring berjalannya waktu. Selain itu, kamus *Slangword* pada penelitian ini juga berisi kumpulan kata berbahasa inggris dan bahasa daerah yang umum digunakan dalam kondisi *code-switching*. Berikut adalah cuplikan kamus yang telah dikumpulkan.

TABEL III
CUPLIKAN KAMUS SLANGWORD

Slangword	**Formal**
mw	mau
dtg	dating
smg	semoga
tuk	untuk
tp	tapi
kopid	covid
tourist	turis
mehong	mahal
good	bagus
banget	sekali
holiday	liburan
ora	tidak
mbujuk	berbohong
…	…

Penerapan tahap *slangword replacement* membuat makna komentar lebih dipahami serta mempermudah tahapan selanjutnya saat praproses teks. Hasil penerapan *slangword replacement* dapat dilihat pada tabel IV.

TABEL IV
CONTOH HASIL TAHAP SLANGWORD REPLACEMENT

Hasil *Cleaning*	**Hasil *Slangword Replacement***
harus dengan syarat yang ketat bagi warga asing masuk indonesia agar tdk menjadi bomerang semoga perekonomian bidang pariwisata kembali normal	harus dengan syarat yang ketat bagi warga asing masuk indonesia agar tidak menjadi bomerang semoga perekonomian bidang pariwisata kembali normal
selagi ada karantina sepertinya mw dibuka pariswisata tuk wisatawan asing agak berat untuk dtg karena biaya hotel karantina mahal tp ini langkahnya bagus untuk mendongkrak ekonomi dibali smg smwnya aman terkendali	selagi ada karantina sepertinya mau dibuka pariswisata untuk wisatawan asing agak berat untuk datang karena biaya hotel karantina mahal tapi ini langkahnya bagus untuk mendongkrak ekonomi di bali semoga semuanya aman terkendali

Tahap selanjutnya yakni *tokenizing*, yang mana sebelum kata dalam kamus *stopwords* dihilangkan, komentar akan dipisah atau dipecah menjadi kata perkata seperti pada tabel V di bawah ini.

TABEL V
CONTOH HASIL TOKENIZING

Hasil *Slangword Replacement*	**Hasil *Tokenizing***
harus dengan syarat yang ketat bagi warga asing masuk indonesia agar tidak menjadi bomerang semoga perekonomian bidang pariwisata kembali normal	['harus', 'dengan', 'syarat', 'yang', 'ketat', 'bagi', 'warga', 'asing', 'masuk', 'indonesia', 'agar', 'tidak', 'menjadi', 'bomerang', 'semoga', 'perekonomian', 'bidang', 'pariwisata', 'kembali', 'normal']
selagi ada karantina sepertinya mau dibuka pariswisata untuk wisatawan asing agak berat untuk datang karena biaya hotel karantina mahal tapi ini langkahnya bagus untuk mendongkrak ekonomi di bali semoga semuanya aman terkendali	['selagi', 'ada', 'karantina', 'sepertinya', 'mau', 'dibuka', 'pariswisata', 'untuk', 'wisatawan', 'asing', 'agak', 'berat', 'untuk', 'datang', 'karena', 'biaya', 'hotel', 'karantina', 'mahal', 'tapi', 'ini', 'langkahnya', 'bagus', 'untuk', 'mendongkrak', 'ekonomi', 'di', 'bali', 'semoga', 'smwnya', 'aman', 'terkendali']

Tahap *stopwords removal* pada penelitian ini menggunakan *library* yang sudah tersedia di bahasa *python*, yakni Sastrawi. Selain menggunakan *library*, penelitian ini juga menggunakan kamus *stopwords* yang telah dikumpulkan. Kamus *stopwords* berisi kata-kata yang tidak diperlukan atau tidak memiliki arti khusus seperti kata konjugasi. Cuplikan kamus *stopwords* dapat dilihat pada tabel VI di bawah ini.

TABEL VI
CUPLIKAN KAMUS STOPWORDS

Stopwords		
Ada	Tapi	di
Yang	Untuk	ke
dengan	dan	agak
dari	dong	doang
kalau	karena	pun

Hasil tokenasi yang sudah didapatkan akan diproses dalam tahap *stopwords removal*. Komentar yang telah bersih dari stopwords akan menjadi lebih ringkas tetapi makna atau inti komentar tetap tidak hilang. Tabel VII merupakan conntoh hasil penerapan proses *stopwords removal*.

TABEL VII
CONTOH HASIL STOPWORDS REMOVAL

Hasil *Tokenizing*	**Hasil *Stopwords Removal***
['harus', 'dengan', 'syarat', 'yang', 'ketat', 'bagi', 'warga', 'asing', 'masuk', 'indonesia', 'agar', 'tidak', 'menjadi', 'bomerang', 'semoga', 'perekonomian', 'bidang', 'pariwisata', 'kembali', 'normal']	syarat ketat warga asing masuk indonesia bomerang semoga perekonomian bidang pariwisata normal

Hasil *Tokenizing*	Hasil *Stopwords Removal*
['selagi', 'ada', 'karantina', 'sepertinya', 'mau', 'di', 'buka', 'pariswisata', 'untuk', 'wisatawan', 'asing', 'agak', 'berat', 'untuk', 'datang', 'karena', 'biaya', 'hotel', 'karantina', 'mahal', 'tapi', 'ini', 'langkahnya', 'bagus', 'untuk', 'mendongkrak', 'ekonomi', 'di', 'bali', 'semoga', 'semuanya', 'aman', 'terkendali']	selagi karantina buka pariswisata wisatawan asing berat biaya hotel karantina mahal langkahnya bagus mendongkrak ekonomi bali semoga semuanya aman terkendali

Setelah tahap *stopwrods removal*, tahap selanjutnya yaitu *stemming*. Dengan menggunakan *library* Sastrawi yang sudah tersedia pada Python, hasil *stemming* dapat dilihat pada Tabel VIII. Didapatkan total 1392 data bersih dari tahap praproses teks yang siap untuk diolah setelah melakukan seleksi pada kolom yang kosong.

TABEL VIII
CONTOH HASIL TAHAP STEMMING

Hasil *Stopwords Removal*	Hasil *Stemming*
syarat ketat warga asing masuk indonesia bomerang semoga perekonomian bidang pariwisata normal	syarat ketat warga asing masuk indonesia bomerang moga ekonomi bidang pariwisata normal
selagi karantina buka pariswisata wisatawan asing berat biaya hotel karantina mahal langkahnya bagus mendongkrak ekonomi bali semoga semuanya aman terkendali	selagi karantina buka pariswisata wisatawan asing berat biaya hotel karantina mahal langkah bagus dongkrak ekonomi bal moga semua aman kendali

C. Pembobotan Kata

Data bersih hasil dari tahap praproses teks akan dihitung bobot tiap kata dengan menggunakan metode TF-IDF. Diperoleh sebanyak 2892 *features*/kata dari hasil perhitungan tersebut. Kemudian setiap kata akan dihitung frekuensinya dengan bobot terbesar di setiap komentar sehingga diperoleh matriks TF-IDF seperti pada Tabel IX. Setelah matriks TF-IDF diperoleh, kemudian akan mulai diklasterisasi.

TABEL IX
MATRIKS TF-IDF

	aamiin	**…**	**baik**	**…**	**prokes**	**…**
Komentar ke- 1	0	…	…	…	0	…
Komentar ke- …	…	…	…	…	…	…
Komentar ke- 16	0	…	0	…	0.29	…
Komentar ke- …	…	…	…	…	…	…
Komentar ke- 34	0.38	…	0.46	…	0	…
Komentar ke- …	…	…	…	…	…	…
Komentar ke- 1392	0	…	0	…	0	…

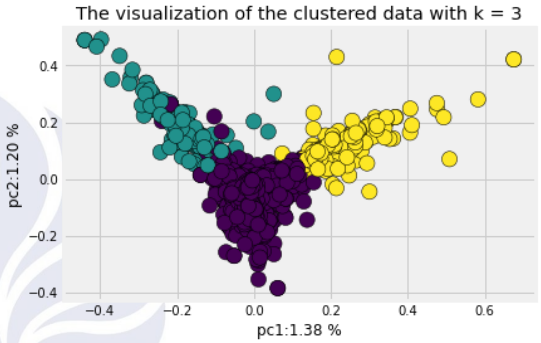
D. Klasterisasi

Klasterisasi yang pertama menggunakan algoritma K-Means. Namun sebelum menerapkan dengan algoritma K-Means, perlu dicari nilai k yang paling optimal menggunakan *Silhouette Method*. Berdasarkan tabel X, hasil perhitungan *silhouette score* menunjukkan bahwa k dengan nilai 3 merupakan yang paling optimal karena menghasilkan *silhouette score* yang paling tinggi di antara nilai k lainnya.

Setelah didapatkan nilai k yang paling optimal berdasarkan *silhouette method*, maka parameter atau nilai k tersebut diterapkann pada proses klasterisasi K-Means. Gbr 2 adalah hasil plot klasterisasi menggunakan algoritma K-means dengan nilai k = 3.

TABEL X
HASIL SILHOUETTE METHOD

Nuber of Cluster (k)	***Silhouette Score***
2	0.617
3	**0.637**
4	0.354
5	0.222
6	0.260
7	0.288
8	0.156
9	0.097



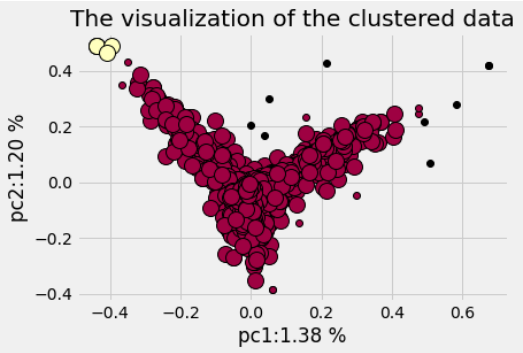
Gbr 2. Hasil Plot Cluster K-Means

Proses klasteriasi yang kedua yakni menggunakan algoritma DBSCAN. Mengunakan *Silhouette Method*, akan ditemukan nilai parameter yang optimal. Uji coba dengan range min Pts dari 2-20 dan nilai epsilon dari 0,01-0,1 diperoleh hasil yang dapat dilihat pada Tabel XII.

TABEL XII
HASIL NILAI PARAMETER DBSCAN YANG OPTIMAL

Epsilon	0.08
Min Pts	5
Number of Cluster	2
Estimate Number of Noise	9
Silhouette Score	0.732

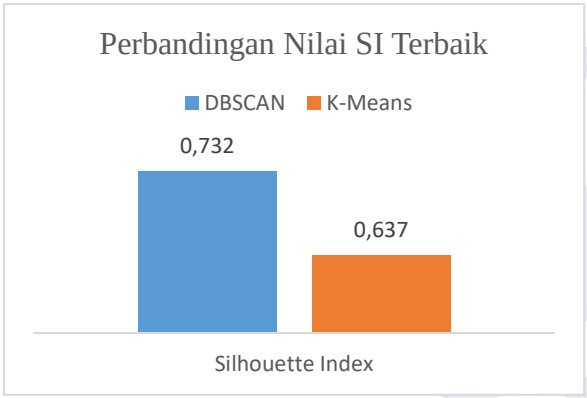
Setelah didapatkan parameter yang paling optimal berdasarkan *silhouette method*, maka parameter tersebut diterapkan pada proses klasterisasi menggunakan algoritma DBSCAN. Keunggulan algoritma DBSCAN yaitu dapat mengidenitifikasi noise pada data. Dapat dilihat pada Gbr 3, titik yang berwarna hitam merupakan *noise* yang ditemukan.



Gbr 3. Hasil Plot Cluster DBSCAN

E. Validitas Klaster

Validitas klister bertujuan untuk mendapatkan nilai klister terbaik dari beberapa percobaan yang telah dilakukan dengan menggunakan *Silhouette Index* (SI). Nilai *Silhouette Index* terbaik adalah yang paling besar atau paling mendekati 1. Perbandingan nilai SI terbaik dari uji coba data menggunakan DSBCAN dan K-Means dapat dilihat pada gambar 4.

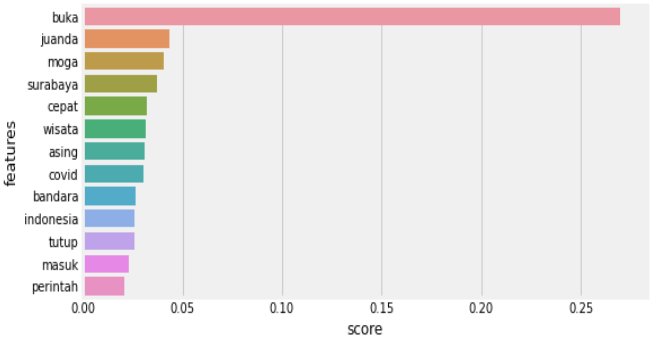


Gbr 4. Perbandingan Nilai SI Terbaik

F. Analisis dan Identifikasi Topik Klaster

Untuk menganalisis konten hasil tiap *cluster*, maka dibuatlah *word cloud* untuk tiap-tiap *cluster* yang telah dihasilkan. Pola sebaran kata pada *word cloud* dapat diamati, sehingga bisa diidentifikasi isi kontennya. Tiap klaster yang dibentuk oleh algoritma K-Means dapat dilihat pada Gbr 5, 7, dan 9.

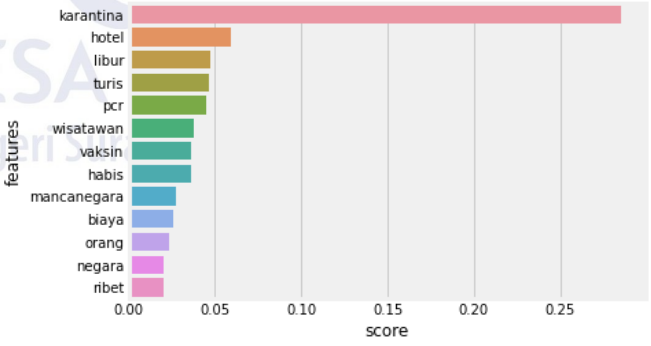
Gbr 5. Word Cloud Klaster 1 K-Means



Gbr 6. Frekuensi Kemunculan Kata pada Klaster 1 K-Means

Anggota klaster pertama berdasarkan hasil plot *word cloud* Gbr 5 diidentifikasi berisi komentar mengenai harapan dibukanya bandara juanda Surabaya, selain itu juga berisi informasi tentang wisata asing yang masuk ke Indonesia. Terlihat dari kata yang paling sering muncul dalam klaster 1 ini yakni kata “buka”, “juanda”, “moga”, “Surabaya”, “cepat”, “wisata”, “asing”.

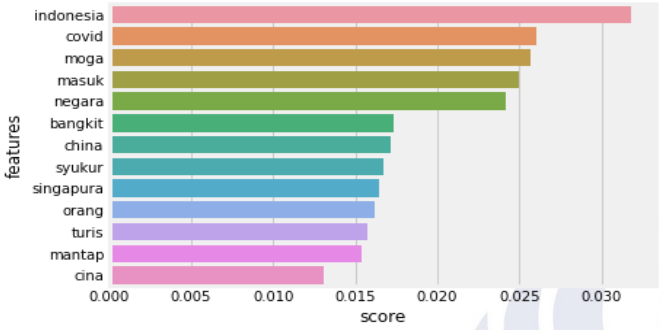
Gbr 7. Word Cloud Klaster 2 K-Means



Gbr 8. Frekuensi Kemunculan Kata pada Klaster 2 K-Means

Gbr 7 menunjukkan visualisasi anggota klaster kedua. Anggota klaster kedua diidentifikasi berisi tentang persyaratan bagi wisatawan mancanegara yang masuk ke Indonesia. Hal tersebut terwakili oleh kata “karantina”, “turis” “wisatawan”, “hotel”, “pcr”, dan “vaksin” sebagai kata yang paling sering muncul dalam anggota klaster kedua seperti dalam Gbr 8.

Gbr 9. Word Cloud Klaster 3 K-Means

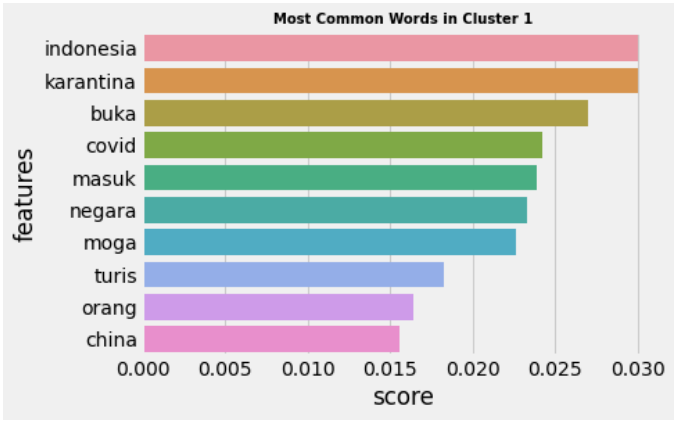


Gbr 10. Frekuensi Kemunculan Kata pada Klaster 3 K-Means

Dapat dilihat pada Gbr 9, klaster ketiga memiliki anggota yang berisi komentar tentang harapan untuk Indonesia agar bangkit saat covid masuk. Kata "Indonesia", "moga", "covid, "masuk", dan "bangkit" merupakan kata yang paling sering muncul seperti dalam Gbr 10.

Selanjutnya adalah hasil analisis dan identifikasi topik klaster yang dihasilkan dari proses klasterisasi menggunakan algoritma DBSCAN. Word Cloud tiap klaster dapat dilihat pada Gbr 11 dan Gbr 13.

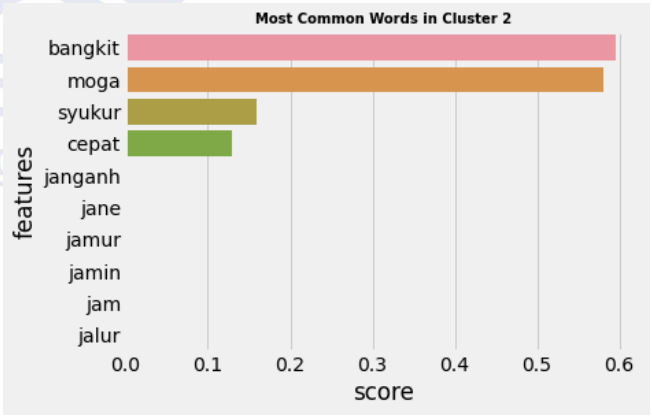
Gbr 11. Word Cloud Klaster 1 DBSCAN



Gbr 12. Frekuensi Kemunculan Kata pada Klaster 1 DBSCAN

Gbr 11 menunjukkan visualisasi anggota klaster pertama DBSCAN. Anggota klaster pertama diidentifikasi berisi tentang kebijakan karantina di Indonesia pada pembukaan negara bagi wisatawan mancanegara. Hal tersebut terwakili oleh kata "indonesia", "karantina" "buka", "covid", "negara", dan "turis" sebagai kata yang paling sering muncul dalam anggota klaster pertama seperti dalam Gbr 12.

Gbr 13. Word Cloud Klaster 2 DBSCAN



Gbr 14. Frekuensi Kemunculan Kata pada Klaster 2 DBSCAN

Anggota klaster kedua berdasarkan hasil plot word cloud dalam Gbr 12 diidentifikasi berisi komentar mengenai harapan agar cepat bangkit dan ucapan syukur atas kebijakan yang diterapkan. Terlihat dari kata yang paling sering muncul dalam klaster kedua ini yakni kata "bangkit", "moga", "cepat", dan "syukur".

Kata "moga" sangat mempresentasikan sebuah harapan seperti yang tertera pada Kamus Besar Bahasa Indonesia (KBBI), yang mana "Moga" merupakan kata dasar dari kata "semoga" yang berarti mudah-mudahan; hendaknya.

Berdasarkan analisis dan identifikasi topik hasil klaster yang sudah dijelaskan di atas, maka akan dibuat tabel perbandingan untuk mengetahui metode atau algoritma mana yang lebih baik. Berikut adalah tabel perbandingan antar metode atau algoritma.

TABEL XIII
PERBANDINGAN TOPIK ANTAR METODE

Nomor Klaster	DBSCAN	K-Means
1	Kebijakan karantina di Indonesia pada pembukaan negara bagi wisatawan mancanegara.	Harapan dibukanya bandara juanda Surabaya dan informasi tentang wisata asing yang masuk ke Indonesia.
2	Harapan agar cepat bangkit dan ucapan syukur atas kebijakan yang diterapkan.	Persyaratan bagi wisatawan mancanegara yang masuk ke Indonesia.
3	-	Harapan untuk Indonesia agar bangkit saat masa pandemi.

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa dataset berupa kumpulan komentar dari video youtube yang membahas tentang kebijakan dibukanya daerah wisata di masa pandemi, sebanyak 1414 komentar yang diambil dalam jangka waktu bulan September 2021 hingga November 2021 diperoleh 2 klaster, 9 noise dan Silhouette Score sebesar 0.732 untuk algoritma DBSCAN, sedangkan untuk algoritma K-Means diperoleh 3 klaster dan nilai Silhouette Score sebesar 0.637.

Karakteristik topik dalam klaster pertama DBSCAN yang terbentuk setelah diidentifikasi berisi tentang kebijakan karantina di Indonesia pada pembukaan negara bagi wisatawan mancanegara, sedangkan klaster kedua mengenai harapan agar cepat bangkit dan ucapan syukur atas kebijakan yang diterapkan. Karakteristik konten klaster pertama yang diperoleh menggunakan K-Means berisi informasi mengenai harapan dibukanya bandara juanda Surabaya, klaster kedua tentang persyaratan bagi wisatawan yang masuk ke Indonesia, dan klaster ketiga tentang harapan untuk Indonesia agar bangkit saat covid masuk.

Hasil analisis dan identifikasi topik klaster DBSCAN menunjukkan bahwa klaster yang terbentuk menggunakan DBSCAN lebih baik daripada K-Means. Hal terebut dapat dilihat dari kata-kata yang paling sering muncul pada tiap klaster, Klaster pertama dan ketiga yang terbentuk menggunakan K-Means masih terdapat kata-kata yang sama muncul, yakni kata "moga" dan "masuk". Topik tiap klaster menggunakan DBSCAN juga lebih mudah disimpulkan daripada K-Means karena tiap klasternya terkategori dengan baik berdasarkan jenis topik dalam komentarnya. Selain itu juga DBSCAN menghasilkan noise sebanyak 9 noise. Maka dapat disimpulkan bahwa penggunaan algoritma DBSCAN lebih baik daripada algoritma K-Means untuk mengelompokkan data teks berupa komentar.

B. Saran

Saran yang dapat diberikan berdasarkan penelitian ini ialah:

1. Perlu dilakukan penelitian dengan menggunakan metode atau algoritma *clustering* yang lain sebagai perbandingan metode yang menghasilkan klaster lebih baik lagi.
2. Perlu dilakukan penelitian dengan metode pengukuran jarak yang lain sebagai perbandingan beberapa metode pengukuran jarak.
3. Saat tahap praproses teks diharap lebih teliti sehingga data yang dihasilkan jauh lebih bersih lagi.

REFERENSI

[1] Gonta-Ganti Istilah Pembatasan Kegiatan Masyarakat [online]. Viewed Nov. 16, 2021. Available : https://www.cnnindonesia.com/nasional/20210722070140-20-670613/gonta-ganti-istilah-pembatasan-kegiatan-masyarakat

[2] UNWTO . International Tourism and covid-19 . [online]. Viewed ed Nov. 15, 2021. Available : https://www.unwto.org/international-tourism-and-covid-19

[3] Badan Pusat Statistik, "Perkembangan Pariwisata dan Transportasi Nasional Agustus 2021," *Ber. Resmi Stat.*, vol. 11, no. 73, pp. 1–16, 2021.

[4] Bali Siap Menyambut Wisatawan Mancanegara [Online]. Viewed accessed Nov. 16, 2021. Available : https://pedulicovid19.kemenparekraf.go.id/bali-siap-menyambut-wisatawan-mancanegara/

[5] A. M. Barik, R. Mahendra, and M. Adriani, "Normalization of indonesian-english code-mixed twitter data, " *W-NUT@EMNLP 2019 - 5th Work. Noisy User-Generated Text, Proc.*, no. November, pp. 417–424, 2019, doi: 10.18653/v1/d19-5554.

[6] C. Myers-Scotton, "Common and Uncommon Ground: Social and Structural Factors in Codeswitching," *Lang. Soc.*, vol. 22, no. 4, pp. 475–503, 1993, doi: 10.1017/S0047404500017449.

[7] 2015. M.-S. Team, SwiftKey. SwiftKey Emoji Report,". 2015. Available : https://www.scribd.com/doc/262594751/SwiftKey-Emoji-Report

[8] C. Skujins, "*Indonesian / English Code-Switching On Social Media*," thesis, Flinders University, 2017.

[9] Mustakim, M. Zakiy Fauzi2, Mustafa, A. Abdullah, and Rohayati, Clustering of Public Opinion on Natural Disastersin Indonesia Using DBSCAN and K-MedoidsAlgorithms," *J. Phys. Conf. Ser.*, vol. 1783, no. No. 1, 2021.

[10] W. Rohalidyawati and M. Rita Rahmawati, "Segmentasi Pelanggan E-Money dengan Menggunakan Algoritma DBSCAN (Density Based Spatial Clustering Applications With Noise) di Provinsi DKI Jakarta," *J. Gaussian*, vol. 9, pp. 162–169, 2020.

[11] R. Adha, N. Nurhaliza, U. Soleha, P. Studi, S. Informasi, and F. Sains, "Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia," vol. 18, no. 2, pp. 206–211, 2021.

[12] F. Setiawan, C. Setianingsih, U. Telkom, and S. Coefficient, "Clustering pada Data Sentimen Transportasi Online Menggunakan Algoritma DBSCAN Clustering On Sentiment Data Online," in *e-Proceeding Eng.*, 2021, vol. 8, no. 5, pp. 6089–6096.

[13] Y. Huang, W. J. Huang, X. L. Xiang, and J. J. Yan, "An empirical study of personalized advertising recommendation based on DBSCAN clustering of sina weibo user-generated content," *Procedia Comput. Sci.*, vol. 183, pp. 303–310, 2021, doi: 10.1016/j.procs.2021.02.063.

[14] D. P. Isnarwaty and Irhamah, "Text clustering pada akun twitter layanan ekspedisi JNE , J&T, dan Pos Indonesia menggunakan metode Density-Based Spatial Clustering of Applications with Noise (DBSCAN)," *J. Sains dan Seni*, vol. 8, no. 2, pp. 2–9, 2019.