BAB II

TINJAUAN PUSTAKA

2.1 Kajian Pustaka

2.1.1 Analisis Sentimen

Analisis sentimen merupakan suatu poses analisis teks yang digunakan untuk mendapatkan sentimen atau opini dari internet dan platform media sosial. Analisis sentimen bertujuan untuk mendapatkan opini dari pengguna platform tersebut. Analisis sentimen akan membangun sistem untuk mengenali dan mengekstraksi opini ke dalam bentuk teks. Dengan bantuan analisis sentimen, informasi yang didapatkan belum terstruktur sehingga dapat diubah menjadi data yang lebih terstruktur (LP2M, 2022).

Analisis sentimen disebut juga sebagai *opinion mining.* Entitas dari analisis sentimen dapat berupa prosuk, jasa, individu, masalah, peristiwa atau topik tertentu. Analisis sentimen hamper digunakan dan diperlukan untuk menganalisis suatu produk, jasa, industry dan topik lainnya (Asshiddiqi & Lhaksmana, 2020).

2.1.2 *Data Mining*

Data mining merupakan proses dalam menemukan pola yang menarik dan pengetahuan dari sejumlah besar data. Sumber data dapat berasal dari database, data warehouse, web, tempat penyimpanan informasi lain atau data yang dialirkan ke dalam system. Data mining juga dikenal sebagai *Knowledge Discovery in Databases (KDD),* selain itu data mining juga dikenal sebagai langkah penting dalam proses penemuan pengetahuan (Han, J., Pei, J. & Tong, H., 2022).

Kegiatan data mining merupakan proses iteratif yang melibatkan mekanisme umpan balik sehinga dapat dilakukan revisi berkelanjutan. Berdasarkan tujuan analisis utamanya, data mining dapat dibedakan menjadi 2 yaitu (Vercellis, 2011):

1. Interpretasi

   Interpretasi mengidentifikasi pola regular di dalam data dan mengekspresikan pola menggunakan aturan dan kriteria yang mudah dipahami. Aturan yang didapatkan harus orisinal agar dapat meningkatkan level pengetahuan dan pemahaman terhadap sistem.

2. Prediksi

   Prediksi mengantisipasi nilai sebuah variabel acak di masa depan atau mengestimasi kemungkinan yang akan terjadi di masa depan.

2.1.3 *Clustering*

Clustering merupakan proses mengorganisasikan objek-objek ke dalam kelompok-kelompok *(cluster)* yang anggota kelompoknya memiliki kemiripan di beberapa karakteristiknya. *Clustering* bertujuan untuk mengelompokan beberapa data atau objek ke dalam klaster sehingga klaster tersebut akan berisi data yang mirip (Vercellis, 2011).

Cluster yang baik memiliki 2 kriteria yaitu (Subakti, dkk, 2020):

1. Homogenitas Internal (*within cluster*), yaitu kesamaan antar anggota di dalam satu klaster atau kelompok.
2. Heterogenitas External (*between cluster*), yaitu perbedaan antar klaster atau kelompok yang satu dengan yang lainnya.

Menurut Jiawei Han, Micheline Kamber dan Jian Pei (2011) metode clustering terbagi menjadi 4, yaitu:

Tabel 2. 1 Metode Clustering

Metode	Karakteristik
Hierarchical Methods	• Mengorganisasikan objek data ke dalam suatu kelompok yang berbentuk hierarki atau "tree". • Agglomerative, Divisive
Partitioning Methods	• Mengorganisasikan objek-objek yang mirip dari suatu himpunan ke dalam beberapa kelompok atau klaster • Menentukan jumlah cluster • K-Means, K-Medoids
Density-based Methods	• Mengelompokkan objek data berdasarkan area kepadatannya • DBSCAN, DENCLUE
Grid-based methods	• Mengkuantisasi/pemetaan ruang objek data menjadi sejumlah sel yang membentuk struktur kisi/ grid • STING, CLIQUE

2.1.4 K-Means

K-means clustering merupakan metode klaster non hirarki untuk mempartisi objek ke dalam satu atau lebih klaster atau kelompok berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang mirip dikelompokan dalam satu klaster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam klaster yang berbeda. Tahapan algoritma K-Means (Vercellis, 2011).:

1. Tentukan nilai k (jumlah klaster) dan centroid dari masing-masing klaster secara acak.

2. Hitung jarak tiap data terhadap centroid menggunakan rumus *Ecludian Distance*

$$dist(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \text{ .....(2.1)}$$

3. Bentuk cluster dengan memanfaat hasil *ecludian distance* terdekat.
4. Hitung nilai ratio sebagai bahan perbandingan untuk stop iterasi dengan rumus:

$$Ratio = bcv/wcv \text{ …….(2.2)}$$

- Nilai bcv *(between cluster variation)* merupakan jarak antar centroid terpilih dengan rumus *ecludian distance.*
- Nilai wcv *(within cluster variation)* merupakan jumlah kuadra jarak terdekat setiap data.

$$wcv = \sqrt{\sum_{i=1}^{j}(\text{jarak terdekat setiap data})^2} \text{ …….(2.3)}$$

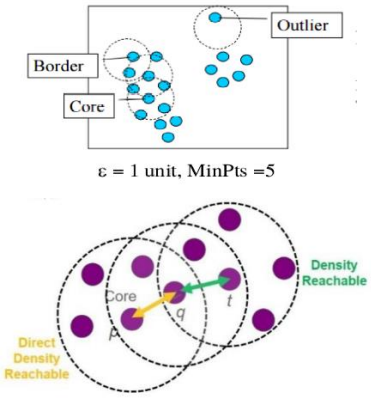
5. Cari nilai centroid baru dengan memanfaatkan rata-rata nilai dari setiap anggota klaster.
6. Ulangi langkah 1-5 sampai anggota cluster tidak ada yang berpindah cluster atau nilai ratio baru ≤ nilai ratio lama.

2.1.5 DBSCAN

Density-Based Clustering of Application with Noise atau DBSCAN merupakan algoritma yang menumbuhkan area-area kepadatan yang cukup tinggi ke dalam cluster dan menemukan cluster-cluster dalam bentuk yang sembarang dalam suatu database spatial yang memuat noise (Id, 2017).

DBSCAN menggunakan 2 parameter input sebelum melakukan clustering yaitu epsilon (eps) dan minimum points (minPts). Epsilon adalah jarak maksimal

antara dua data dalam satu cluster dan minimum points adalah banyaknya data minimal dalam jarak epsilon agar terbentuk suatu cluster (David, 2020).



Gambar 2. 1 Gambaran DBSCAN
(sumber: https://algotech.netlify.app/blog/dbscan-clustering/)

Berikut adalah langkah-langkah dari DBSCAN:

1. Tentukan nilai minPts dan epsilon (eps) yang akan digunakan.
2. Pilih data awal “p” secara acak.
3. Hitung jarak antara data “p” terhadap semua data menggunakan Euclidian distance.
4. Ambil semua amatan yang *density-reachable* dengan amatan “p”.
5. Jika amatan yang memenuhi nilai epsilon lebih dari jumlah minimal amatan dalam suatu kelompok maka amatan “p” dikategorikan sebagai *core points* dan kelompok terbentuk.
6. Jika amatan “p” adalah *border points* dan tidak ada amatan *density-reachable* dengan amatan “p”, maka lanjutkan pada amatan lainnya.
7. Ulangi langkah 3 sampai 6 semua amatan diproses.

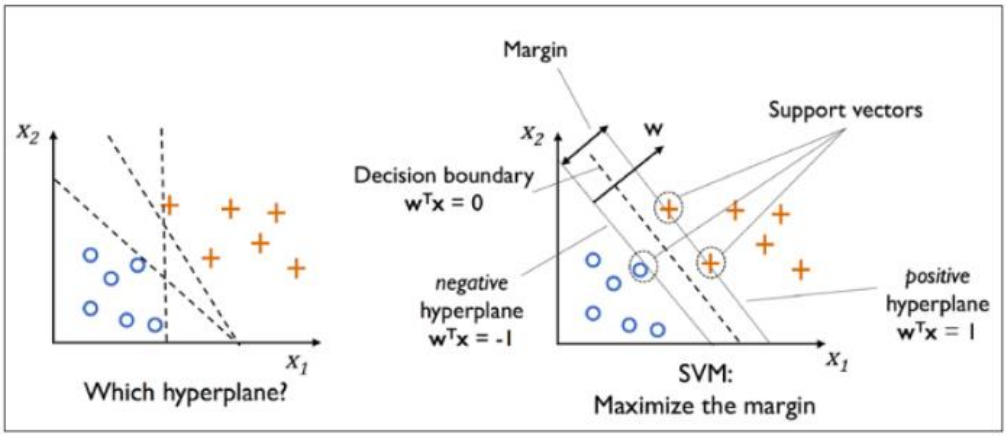
2.1.6 Klasifikasi

Klasifikasi merupakan suatu proses yang dilakukan untuk mengkategorikan data yang ada sesuai dengan atribut atau kelasnya. Metode klasifikasi akan

menggunakan data latih untuk menghitung cara terbaik untuk memetakan data input ke label tertentu dan menggunakan data uji yang menguji performa dari model yang terbentuk. Metode klasifikasi memiliki banyak algoritma diantaranya Decision Tree, Support Vector Machine, KNN, Naïve Baiyes, dan sebagainya (Wahyuni, dkk, 2020).

2.1.7 *Support Vector Machine*

Support Vector Machine (SVM) termasuk algoritma supervised learning yang digunakan untuk klasifikasi. SVM menerima masukkan berupa data yang menghasilkan keluaran berupa garis yang akan memisahkan kelas-kelas dari data tersebut. Tujuan dari SVM adalah untuk menemukan hyperplane dalam ruang dimensi-N yang secara jelas membagi titik-titik data (Wahyuni, dkk, 2020).



Gambar 2. 2 *Hyperplane* pada SVM
(Sumber: https://medium.com/@samsudiney/penjelasan-sederhana-tentang-apa-itu-svm-149fec72bd02)

Tujuan SVM adalah menemukan plane yang memiliki margin maksimum, jarak maksimum antara titik data kedua kelas. Hyperplanes adalah batas dari *decision* yang membantu mengklasifikasikan titik data. Titik data yang berada di kedua sisi hyperplane dapat dikaitkan dengan kelas yang berbeda (Wahyuni, dkk, 2020).

Berikut adalah langkah-langkah dari SVM (Plaosan, 2021):

1. Data dinotasikan dengan $\overrightarrow{X_l} \in R^d$, label dinotasikan dengan $y_i \in \{-1, +1\}$ untuk *i* = 1,2,..,*l*, dimana *l* merupakan banyak data. Diasumsikan kedua class -1 dan +1 yag terpisah oleh hyperplane berdimensi d yang didefinisikan sebagai berikut:

$$\vec{w}\vec{x} + b = 0 \text{ …….(2.4)}$$

Pattern $\overrightarrow{x_i}$ yang merupakan sampel negatif (-1) dapat dirumuskan dengan persamaan:

$$\vec{w}\vec{x} + b \leq -1 \text{ …….(2.5)}$$

Pattern $\overrightarrow{x_i}$ yang merupakan sampel positif (+1) dapat dirumuskan dengan persamaan:

$$\vec{w}\vec{x} + b \geq +1 \text{ …….(2.6)}$$

2. Margin terbesar dapat dihitung dengan memaksimalkan nilai jarak antara hyperplane dan titik terdekatnya. Dapat dirumuskan dengan *Quadratic Programing* (QP) *problem*, dengan persamaan sebagai berikut:

$$\min_{\vec{w}} \tau(w) = \frac{1}{2}||\vec{w}||^2 \text{ …….(2.7)}$$

$$y_i(\overrightarrow{x_i}.\vec{w} + b) - 1 \geq 0, \forall_i \text{ …….(2.8)}$$

3. Untuk memecahkan masalah QP menggunakan teknik komputasi *Lagrange Multiplier*:

$$L(\vec{w}, b, \alpha) = \frac{1}{2}|\vec{w}|^2 - \sum_{i=l}^{l} \alpha_i(y_i((\overrightarrow{x_i}.\vec{w} + b) - 1) \ (i = 1,2,..,l) \text{ …….(2.9)}$$

α_i merupakan *Lagrange Multipliers* yang bernilai nol atau positif ($\alpha_i \geq 0$). Nilai optimal dapat dihitung dengan meminimalkan *L* terhadap $\vec{w}$ dan b, dan memaksimalkan *L* terhadap α_i. Dengan memperhatikan sifat bahwa titik

optimal gradient L = 0, dapat dimodifikasi sebagai maksimalisasi yang hanya mengandungn α_i saja, dengan persamaan:

Maximize:

$$\sum_{i=l}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=l}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i \vec{x}_j \text{ .....(2.10)}$$

Subject to:

$$\alpha_i \geq 0(i = 1,2, \dots, l) \ \sum_{i=l}^{l} \alpha_i \, y_i = 0 \text{ .....(2.11)}$$

4. Dari hasil perhitungan didaptkan α_i yang bernilai positif. Data yang berkorelasi dengan $\alpha_i$yang positif disebut *support vector*.

2.1.8 *Confusion Matrix*

Confusion Matrix merupakan pengukuran performa klasifikasi dimana keluaran dapat berupa dua kelas atau lebih. *Confusion Matrix* memiliki 4 kombinasi berbeda dari nilai prediksi dan nilai actual pada suatu tabel. *Confusion Matrix* dapat menghitung nilai *accuracy*, *precision*, *recall* dan *F-1 score* (Anggreany, 2020). [16]

1. Accuracy merupakan gambaran seberapa akurat model dalam mengklasifikasi dengan benar. Persamaan akurasi:

$$akurasi = \frac{prediksi\ benar}{prediksi\ benar + prediksi\ salah} \text{ .....(2.12)}$$

2. Precision merupakan gambaran akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Persamaan *precision*:

$$precision = \frac{jumlah\ prediksi\ benar\ untuk\ kelas\ positif}{total\ prediksi\ kelas\ positif} \text{ .....(2.13)}$$

3. Recall atau sensitivity merupakan gambaran keberhasilan suatu model dalam menemukan kembali sebuah informasi. Persamaan *recall*:

$$recall = \frac{jumlah\ prediksi\ benar\ untuk\ kelas\ positif}{total\ data\ kelas\ positif} \text{ .....(2.14)}$$

4. F-1 Score merupakan perbandingan rata-rata precision dan recall yang dibobotkan. Accuracy digunakan sebagai acuan performansi algoritma jika dataset kita memiliki jumlah data False Negatif dan False Positif yang sangat mendekati (symmetric). Namun jika jumlahnya tidak mendekati, maka sebaiknya kita menggunakan F1 Score sebagai acuan. Persamaan *F-1 score*:

$$F1\ score = 2 \times \frac{recall \times precision}{recall + precision} \text{ …….(2.15)}$$

2.2 Penelitian Terdahulu

Penelitian terdahulu ini menjadi salah satu acuan penulis dalam melakukan penelitian sehingga penulis dapat memperkaya teori yang digunakan dalam mengkaji penelitian yang dilakukan. Dari penelitian terdahulu, penulis menemukan judul yang sama seperti judul penelitian penulis. Namun penulis mengangkat beberapa penelitian sebagai referensi untuk memperkaya bahan kajian pada penelitian penulis. Berikut ulasan dari penelitian terdahulu:

Tabel 2. 2 Penelitian Terdahulu

No	Judul, Penulis	Tujuan	Hasil
1	Implementasi Algoritma K-Means Clustering Pada Analisis Sentimen Keluhan Pengguna Indosat Try Iryanto Saputra, Rini Arianty (2019)	Pada penelitian ini, akan dilakukan analisis sentimen terhadap konsumen pengguna provider Indosat, menggunakan data tweet sejumlah 300 data acak yang di kumpulkan dari bulan desember 2018 hingga bulan april 2019. Data yang dianalisis adalah	Penelitian ini berhasil menampilkan kelompok dari anggota masing-masing cluster yang berbentuk wordcloud ke dalam 3 buah wordcloud berbeda, pada wordcloud cluster 0 anggotanya berbicara tentang jaringan Indosat yang parah, pada wordcloud cluster 1 anggotanya berbicara tentang

No	Judul, Penulis	Tujuan	Hasil
		kalimat berbahasa Indonesia.	permintaan perbaikan jaringan sinyal Indosat, dan pada wordcloud cluster 2 anggotanya berbicara tentang jaringan sinyal parah Indosat pada daerah Bogor.
2	Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019 Imam Kurniawan, Ajib Susanto (2019)	Tujuan dari penelitian ini memperoleh analisis dokumen text untuk mendapatkan sentimen positif atau negatif. Metode yang digunakan K-Means untuk melakukan klastering pada data latih dan Naive Bayes classifier untuk mengklasifikasi pada data testing.	Hasil dari pembobotan ini berupa sentimen positif dan negatif. Data diambil dari Twitter mengenai pemilu presiden 2019 sebanyak 500 data tweet. Dari hasil pengujian 100 dan 150 data uji diperoleh akurasi rata-rata 93.35% dan error rate sebesar 6.66%.
3	Kombinasi *K-Means* Dan *Support Vector Machine* (SVM) Untuk Memprediksi Unsur SARA Pada *Tweet* Wiga Maulana Baihaqi, Muliasari Pinilih, Miftakhul Rohmah (2018)	Penelitian ini bertujuan untuk membuat *corpus* kalimat yang mengandung unsur SARA yang didapatkan dari twitter, kemudian melabeli kalimat dengan label mengandung unsur SARA dan tidak, serta melakukan *sentiment* klasifikasi. Algoritme yang digunakan untuk proses pelabelan adalah k-*means*, sedangkan *Support Vector Machine* (SVM) digunakan untuk proses klasifikasi.	Hasil akurasi yang diperoleh untuk meningkatkan hasil akurasi, data hasil proses k-*means* diolah kembali dengan validasi pakar bahasa, hasil yang diperoleh menjadi 139 *tweet* positif SARA dan 62 *tweet* negatif SARA, hasil akurasi meningkat menjadi 70,15% dan 71,14%. Dari hasil yang didapatkan, twitter dapat dijadikan sumber untuk membuat *corpus* mengenai kalimat SARA, dan metode yang diusulkan berhasil untuk

No	Judul, Penulis	Tujuan	Hasil
			proses pelabelan dan sentimen klasifikasi, akan tetapi masih perlu peningkatan hasil akurasi.
4	Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI Muhammad Fadli Asshiddiqi, Kemas Muslim Lhaksmana (2020)	Dari banyaknya opini tersebut akan menjadi sebuah data yang bisa menentukan kepuasan masyarakat terhadap kinerja PSSI apakah hasilnya cenderung postif atau negatif dengan cara melakukan sentimen analisis. Untuk meunujang sentimen analisis diperlukan Algoritma klasifikasi, algoritma pada penelitian ini menggunakan Decision Tree dan Support Vector Machine.	Hasil komposisi data terbaik untuk melakukan pengujian adalah 80%:20% dengan mendapatkan hasil nilai akurasi 87.45%, precison 87.72%, recall 91.74% dan F1-Score 89.69% pada Decision Tree dengan TF-IDF sedangkan untuk Support Vector Machine dengan TF-IDF komposisi data terbaik adalah 80%:20% mendapatkan hasil nilai akurasi akurasi 94.36%, precison 96.78%, recall 94.30% dan F1-Score 95.53%. Maka pada kasus ini sentimen analisis pada komentar instagram akan lebih baik jika menggunakan Support Vector Machine (SVM) dengan TF-IDF.
5	Sentimen Analisis pada Data Tweet Pengguna Twitter Terhadap Produk Penjualan Toko Online Menggunakan Metode K-Means	Bertujuan untuk memberikan suatu keputusan dari opini konsumen terhadap produk penjualan sehingga menjadi peluang bagi produsen dalam mempromosikan dan memasarkan produknya kepada	Dengan pengujian berdasarkan kata tweet diperoleh tingkat akurasi sebesar 92.80 % sedangkan pengujian berdasarkan tweet harian diperoleh tingkat akurasi sebesar 89.80 %.

No	Judul, Penulis	Tujuan	Hasil
	Andris Faesal, Aziz Muslim, Aditya Hastami Ruger, Kusrini (2020)	konsumen. Pra-prosesing untuk mencari kata-kata yang sering muncul didalam tweet. Langkah terakhir dengan menggunakan metode K-Means sebagai proses pengelompokan sebayak 3 cluster yaitu kata dengan kemunculan sering, sedang dan jarang digunakan didalam tweet.	
6	Modfikasi DBSCAN (*Density-Based Spatial Clustering With Noise*) pada Objek 3 Dimensi Ibnu Daqiqil Id dan Evfi Mahdiyah (2017)	Algoritma DBSCAN merupakan algoritma pengelompokan data spasial berdasarkan kerapatan objek 2 dimensi. Pada paper ini akan dibahas bagaimana melakukan clustering pada objek 3 dimensi menggunakan algoritma DBSCAN. Modifikasi yang dilakukan adalah dengan merubah mekanisme penghitungan jarak kerapatan antara objek.	Berdasarkan hasil pengujian didapatkan rata-data *Silhouette Coefficient* adalah 0.550 dengan *eps*=0.2 dan *minpts*= 3. Dari data tersebut, cluster yang dihasilkan memiliki stuktur yang baik dan tidak sensitive terhadap noise.
7	Penerapan SVM Dan Information Gain Pada Analisis Sentimen Pelaksanaan Pilkada Saat Pandemi Aliffia Kulsumarwati, Intan Purnamasari,	Untuk itu maka dilakukan penerapan *data mining* dengan algoritma *Support Vector Machine* dan seleksi fitur *information gain* untuk menganalisis berbagai tanggapan masyarakat mengenai pelaksanaan pilkada 2020. Data yang digunakan merupakan	Hasil klasifikasi data *tweet* dengan *Support Vector Machine* menggunakan kernel linear menghasilkan nilai akurasi yang besar yaitu 92%, *precision* 90%, dan *recall* 92%.

No	Judul, Penulis	Tujuan	Hasil
	Budi Arif Dermawan (2021))	*tweet* dari aplikasi Twitter sebanyak 496 data. Sebelum tahap *data mining*, dilakukan pembagian data menjadi 80% *data traning* dan 20% *data testing*.	
8	Implementasi Analisis Clustering Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode K-Means Yan Watequlis Syaifudin, Rizki Andi Irawan (2018)	Pada penelitian ini melakukan analisis sentimen terhadap 10 pantai yang ada di Indonesia. Dataset yang digunakan dari twitter dengan 500 tweet. Melakukan klasifikasi terhadap opini dengan menggunakan SVM.	Hasil akurasi dari algoritma SVM sebesar 74,39%. Selanjutnya data opini dari kuesioner ditambahkan untuk mengelompokkan pantai berdasarkan ketersediaan sumber daya, fasilitas, akses, kesiapan masyarakat, potensi pasar dan posisi pariwisata. Dalam proses pengelompokan data ini digunakan metode K-Means.
9	Implementasi Algoritma *K – Means* Untuk *Clustering* Sentimen Pada Opini Kualitas Pelayanan Jasa Penerbangan Syarifah Iin Safitri, Cucu Suhery, Syamsul Bahri (2021)	Penelitian ini bertujuan untuk membangun sebuah sistem yang digunakan untuk mengelompokkan opini positif dan opini negatif yang terdapat pada *website* Skytrax menggunakan algoritma *K-Means*. Data yang digunakan merupakan data opini maskapai penerbangan Garuda Indonesia, Air Asia dan Lion Air yang terdapat pada *website* Skytrax dengan jumlah	Keluaran yang dihasilkan berupa data opini yang telah dikelompok menjadi kelompok negatif dan kelompok positif. Perhitungan akurasi Dengan membandingkan nilai rating opini pada *website* Skytrax dengan hasil *clustering* sentimen pada algoritma *K-Means*. Persentase keberhasilan sistem *clustering* sentimen maskapai Lion Air 60,5%, Air Asia 52,8%, dan Garuda Indonesia 71,8%.

No	Judul, Penulis	Tujuan	Hasil
		keseluruhan 1060 data opini dari tahun 2015 sampai dengan 2019.	
10	Staged Text Clustering Algorithm Based On K-Means And Hierarchical Agglomeration Clustering Youjin Rong, Yi'an Liu (2020)	SC-KH *(Staged text clustering algorithm based on K-Means and HAC)* diusulkan untuk mengatasi kekurangan akurasi rendah algoritma K-Means dan kompleksitas waktu yang tinggi dari *Hierarchical Agglomeration Clustering*	Hasil menunjukkan bahwa kinerja SC-KH lebih baik disbanding K-Means dan *Hierarchical Agglomeration Clustering,* dan kualitas klastering meningkat.

Perbedaan penelitian yang dilakukan adalah melakukan perbandingan kombinasi algoritma K-Means dengan SVM dan DBSCAN dengan SVM, untuk melihat sentimen pengguna aplikasi PeduliLindungi berdasarkan komentar Google PlayStore dan App Store.