e- ISSN: 2828–299X

# Perbandingan Metode K-means and *Agglomerative Nesting* untuk *Clustering* Data *Digital Marketing* di Twitter

**Nandya Arifa Wulandari[1], Hasih Pratiwi[2], Sri Sulistijowati Handayani[3]**
[1,2,3]Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret
E-mail: *[1]nandya.01@student.uns.ac.id, [2]hasihpratiwi@staff.uns.ac.id, [3]rr_ssh@staff.uns.ac.id

***Abstrak*** *– Banyak masyarakat yang meraup keuntungan besar dari penjualan online dan digital marketing atau promosi di berbagai platform media sosial. Twitter merupakan salah satu media sosial paling berpengaruh di dunia, banyak brand yang memanfaatkan platform ini untuk melakukan digital marketing. Clustering dapat diimplementasikan di berbagai aspek yang berhubungan dengan pengelompokan data. Metode yang digunakan pada penelitian ini adalah metode K-means dan Agglomerative Nesting (AGNES) clustering (single linkage, average linkage, complete linkage, dan ward linkage). Data yang digunakan merupakan data API Twitter dengan keyword "Rp 0" yang merupakan trending pada tanggal 10, bulan 10 tahun 2022 Hasil penelitian menunjukan bahwa metode terbaik untuk melakukan clustering pada data digital marketing Twitter adalah metode AGNES dengan complete linkage dengan K=2 dan Silhouette_Score 0.754428965. Jumlah data cluster C0 sebanyak 423 data tweet dan cluster C1 sebanyak 726 data tweet.*

***Kata Kunci*** *— Clustering, K-means, Agglomerative Nesting, Digital Marketing*

## 1. PENDAHULUAN

Sekarang konsep *data mining* banyak digunakan di berbagai aspek. *Data mining* adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola hubungan dalam set data berukuran besar. Dalam data mining salah satu metode yang banyak digunakan adalah metode *clustering*. Definisi *clustering* menurut konsep data mining yaitu pengelompokan berbagai data atau objek ke dalam *cluster* (kelompok) sehingga setiap *cluster* berisi data yang mirip dan berbeda dari objek di *cluster* lainnya yang dapat dipraktikkan.[1].

Dalam *clustering* terdapat beberapa metode salah satu metode *clustering* yang sering digunakan yaitu metode K-means, metode K-means adalah salah satu metode *clustering* berbasis jarak yang membagi data dalam beberapa *cluster*. Metode K-means hanya bekerja pada atribut numerik[2]. Metode K-means bertujuan untuk mengelompokkan data ke dalam *cluster*. Data yang memiliki kemiripan karakteristik dimasukan dalam satu *cluster*, namun jika berbeda dikelompokan ke dalam *cluster* yang lain. K-means termasuk metode *data mining* yang melakukan pemodelan tanpa pengawasan (*supervised*)[3].

*Agglomerative Nesting* (AGNES) salah satu metode *hierarchical clustering*. Metode *Agglomerative Nesting* (AGNES) adalah metode yang memroses pengelompokan dimulai dari setiap objek sebagai sebuah *cluster* kemudian secara rekursif menggabungkan *cluster* terdekat ke dalam *cluster* yang lebih besar. *Cluster* tunggal menjadi akar hierarki[4]. Apabila terdapat jumlah data sebanyak n, dan k dianggap sebagai jumlah cluster, sehingga besarnya n sama dengan k (n = k). Selanjutnya, penelitian ini menggunakan Euclidean Distance Space untuk menghitung jarak antar cluster berdasarkan jarak rata-rata antar objek. Berdasarkan dari hasil perhitungan tersebut pilih jarak yang paling minimal kemudian gabungkan, maka besarnya n adalah n-1 (n= n - 1). Jarak cluster di-update ketika 2 (dua) cluster digabungkan.Dalam *Agglomerative Nesting* (AGNES) terdapat metode seperti *single linkage, ward linkage, average linkage,* dan *complete linkage.*

Pada penelitian sebelumnya telah dilakukan uji analisis tingkat ketepatan *digital marketing* untuk media sosial Facebook menggunakan metode K-Means di Negara Thailand. Variabel yang digunakan dalam penelitian ini adalah jumlah *like, comment, dan share* [5]. Selain Facebook juga ada sosial media lainnya seperti Instagram, Twitter, Tiktok, dan lain-lain. Menurut data, Twitter merupakan salah satu media sosial yang paling memengaruhi masyarakat di dunia. Hingga tahun 2021 terdapat 15,7 juta pengguna Twitter di Indonesia, angka ini menduduki peringkat keenam di dunia sebagai negara dengan user Twitter terbanyak[6]. Pada penelitian kali ini penulis menggunakan data Twitter untuk melakukan analisis *clustering digital marketing* Twitter di Indonesia menggukanan metode K-means dan *Agglomerative Nesting* (AGNES)

## 2. METODE PENELITIAN

Pada bagian metodologi penelitian ini diuraikan langkah-langkah sebagai berikut:

2.1 Pengambilan Data

e- ISSN: 2828–299X

Pengambilan data penelitian ini dilakukan pada Oktober tanggal 10 tahun 2022. Data yang digunakan pada penelitian ini adalah data API Twitter dengan *keyword* "Rp 0" yang diambil menggunakan bantuan *web* https://netlytic.org/. *Keyword* yang digunakan "Rp 0" karena merupakan trending Twitter pada tanggal 10 bulan 10 yang mana dimanfaatkan para penjual untuk mempromosikan produknya dan dimanfaatkan pembeli untuk menemukan harga terbaik. *Dataset* berisi 1835 data yang dengan menggunakan kolom *favorite_count*, kolom *retweet_count*, dan kolom *user_followers*_count.

Tabel 1. Variabel

| Atribut | Keterangan |
|---|---|
| favorite_count | Jumlah suka dalam satu tweet |
| retweet_count | Jumlah retweet dalam satu tweet |
| user_followers_count | Jumlah pengikut dari pengguna yang memposting tweet |

2.2 Tahap Pengolahan Data

Pada tahap ini dilakukan persiapan data. Sebelum masuk ke tahap *clustering* data perlu melalui tahap *preprocessing*. Tahap *preprocessing* berfungsi mengubah data yang tidak terstruktur menjadi terstruktur[6]. Proses *preprocessing* pada analisis ini adalah pertama dilakukan data *cleaning* dengan menghilangkan kolom atau variabel yang tidak digunakan, selanjutnya melihat tipe data masing masing variabel, mengubah tipe data variabel "id" dari *int* ke *object* agar tidak ikut terproses dalam *clustering*, melakukan pengecekan data *null* atau data yang tidak memiliki nilai, dan memastikan distribusi data, deskripsi data dan *outlier*, menghilangkan *outlier* pada masing-masing variabel. *Outlier* (data pencilan) adalah contoh data yang memiliki karakteristik menonjol secara signifikan dari data pengamatan lain dan berupa nilai ekstrim baik untuk satu variabel atau sekelompok variabel[8].

2.3 Tahap *Clustering*

Setelah data siap selanjutnya adalah tahap *clustering*. Pada tahap ini data yang sudah siap dimodelkan menggunakan metode K-Means dan *Agglomerative Nesting* (AGNES)

a. K-means

Langkah-langkah melakukan *clustering* K-means dengan metode K-Means yaitu memilih jumlah *cluster* k, menginisialisasi k pusat *cluster* ini bisa dilakukan dengan berbagai cara. Namun yang paling sering dilakukan adalah dengan cara random. Pusat-pusat *cluster* diberi nilai awal dengan angka-angka random, selanjutnya mengalokasikan semua data/objek ke *cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke *cluster* tertentu ditentukan dari jarak antara data dengan pusat *cluster*. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster*. Jarak terdekat antara satu data dengan satu *cluster* tertentu untuk menentukan suatu data masuk dalam *cluster* mana. Untuk menghitung jarak semua data ke setiap titik pusat *cluster* dapat menggunakan teori jarak *Euclidean* yang dirumuskan pada persamaan 1[1].

$$D_{(i,j)} = \sqrt{(x_{1i} - \mu_{1j})^2 + (x_{2i} - \mu_{2j})^2 + \cdots + (x_{ki} - \mu_{kj})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

dengan:

$D_{(i,j)}$ = Jarak data ke i ke pusat *cluster j*,

$X_{ki}$ = Data ke i pada atribut data ke *k*,

$X_{kj}$ = Titik pusat ke j pada atribut ke *k*.

Menghitung kembali pusat *cluster* dengan keanggotaan *cluster* yang sekarang. Pusat *cluster* adalah rata-rata dari semua data/objek dalam *cluster* tertentu. Jika dikehendaki bisa juga menggunakan median dari *cluster* tersebut. Jadi rata-rata (mean) bukan satu-satunya ukuran yang bisa dipakai. Langkah terakhir menugaskan lagi setiap objek memakai pusat *cluster* yang baru. Jika pusat *cluster* tidak berubah lagi maka proses *clustering* selesai, jika pusat *cluster* berubah maka kembali ke langkah ke 3 (mengalokasikan semua data) sampai pusat *cluster* tidak berubah.

b. *Agglomerative Nesting* (AGNES).

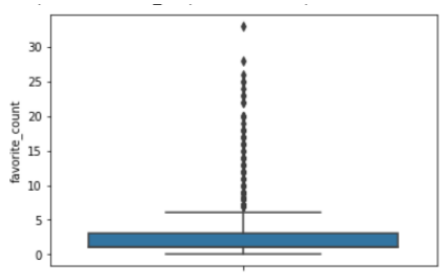
Pada tahap ini data yang sudah siap dimodelkan menggunakan metode *Agglomerative Nesting* (AGNES). Metode perhitungan dalam analisis *clustering Agglomerative Nesting* yaitu menghitung matriks jarak dengan menggunakan jarak Euclidean, menetapkan setiap objek adalah sebuah *cluster*, menggabungkan dua *cluster* berdasarkan ukuran jarak dan metode penggabungan yang digunakan, memperbarui matriks jarak dan menghitung jarak antar *cluster* baru dengan *cluster* awal, mengulangi langkah ke 3 dan 4 sampai terbentuk sejumlah *cluster* yang telah ditentukan[9].

e- ISSN: 2828–299X

# 3. HASIL DAN PEMBAHASAN

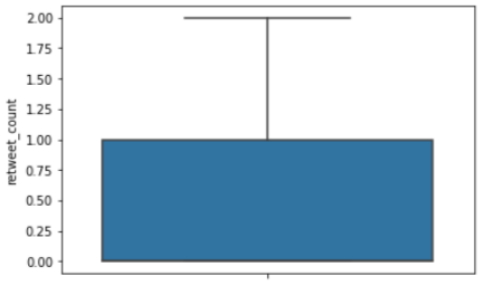
## 3.1 *Preprocessing*

Peralatan yang digunakan untuk analisis ini yaitu laptop, internet, dan software excel. Bahasa Pemrograman untuk pengolahan data ini adalah bahasa python. Kolom yang tidak diperlukan dihapus hingga tersisa kolom variabel *favorite_count*, kolom *retweet_count*, dan kolom *user_followers*_count. Kolom favorite_count mewakili jumlah likes yang didapatkan setiap tweet, kolom retweet_count mewakili jumlah retweet atau balasan setiap tweet, dan kolom *user_followers*_count mewakili jumlah pengikut dari akun atau user yang memposting tweet tweet tersebut, Terdapat beberapa *outlier* pada setiap variabel sehingga perlu dilakukan penghapusan data *outlier*. Setelah melalui tahap *Pre-Processing* data yang tersisa berjumlah 1149 data, data setiap variabelnya dapat dilihat pada *boxplot* Gambar 1, Gambar 2, dan Gambar 3



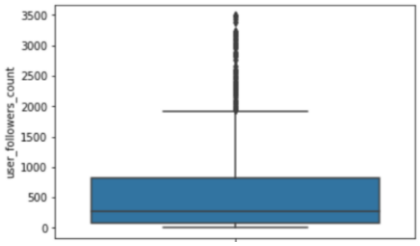

Gambar 1. *Boxplot* Favorite_count

Pada Gambar 1 menunjukan bahwa pada variabel *favorite_count* masih ada data yang melebihi batas atas namun masih ditoleransi karena tidak terlalu ekstrim.




Gambar 2. *Boxplot* Retweet_count

Pada Gambar 2 Variabel retweet_count sudah tidak terlihat *outlier* pada boxplot




Gambar 3. user_followers_count

Pada Gambar 3 menunjukkan bahwa pada variabel user_followers_count masih terdapat beberapa data yang melebihi batas atas namun masih ditoleransi karena tidak terlalu ekstrim. Untuk deskripsi data dapat dilihat pada Tabel 2.

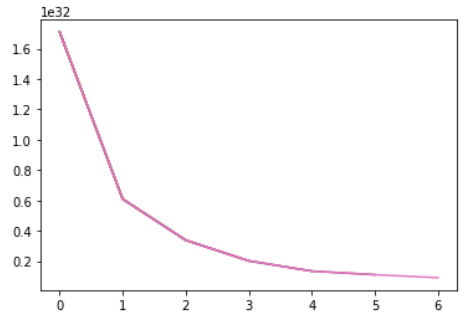
Tabel 2. Deskripsi Data

| | favorite_count | retweet_count | user_followers_count |
|---|---|---|---|
| count | 1149 | 1149 | 1149 |
| mean | 2,75 | 0,32 | 632,44 |
| std | 3,63 | 0,58 | 798,87 |
| min | 0 | 0 | 0 |
| 25% | 1 | 0 | 79 |
| 50% | 1 | 0 | 276 |
| 75% | 3 | 1 | 816 |
| max | 33 | 2 | 3492 |

e- ISSN: 2828–299X

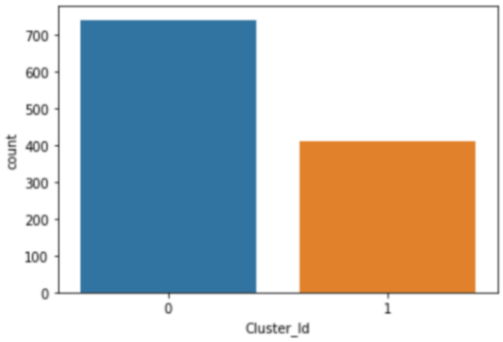
3.2 Metode K-means

Tahap selanjutnya data dilakukan *clustering* menggunakan metode K-means. Agar penetuan nilai k menjadi lebih tepat maka penentuan nilai k menggunakan metode elbow seperti yang dapat dilihat pada Gambar 5.




Gambar 5. *Elbow Method*

Dari Gambar 4 dapat dilihat bahwa garis sebelum menurun menunjukan angka 2 sehingga dapat ditentukan jumlah *k* adalah 2, kemudian dilakukan *clustering* data menggunakan metode *K-means* sehingga didapatkan jumlah masing-masing *cluster* seperti yang ditampilkan pada Gambar 6. Grafik warna biru menunjukan *cluster* 0 dan grafik warna jingga menunjukan *cluster* 1.




Gambar 6. Grafik Perbandingan Jumlah *Cluster*

Hasil *clustering* dapat dilihat pada potongan tabel data yang sudah dilakukan *clustering* K-means ditunjukan pada Tabel 3.

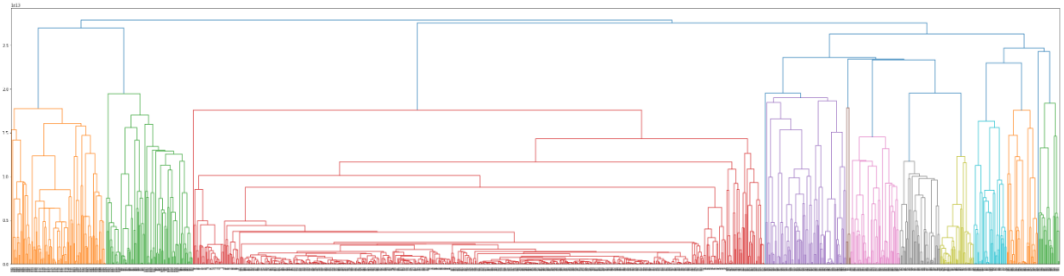
Tabel 3. Potongan data hasil *clustering* K-means

| tweetid | favorite_count | retweet_count | user_followers_count | *Cluster*_Id |
|---|---|---|---|---|
| 1578375894187400000 | 2 | 0 | 90 | 1 |
| 1578375267633850000 | 5 | 1 | 351 | 1 |
| 1578364959897300000 | 2 | 0 | 2 | 1 |
| 1578354959179190000 | 7 | 2 | 371 | 0 |
| 1578351140559000000 | 1 | 0 | 418 | 0 |
| 1578345098777020000 | 1 | 0 | 1 | 0 |
| 1578342566071710000 | 0 | 0 | 355 | 0 |

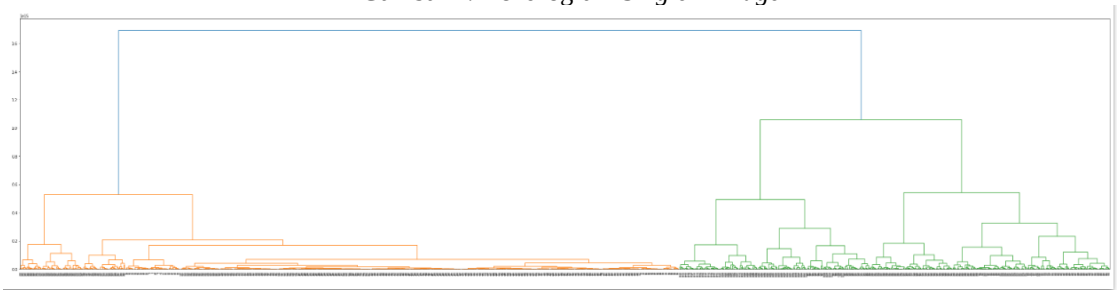
3.3 Metode *Agglomerative Nesting*

AGNES (*Agglomerative Nesting*) yang dibagi menjadi 4 metode yaitu *Single linkage*, *Average linkage*, *Complete linkage*, dan *Ward linkage*. Dendrogram hasil *clustering* masing masing metode dapat dilihat pada Gambar 7, Gambar 8, Gambar 9, dan Gambar 10.

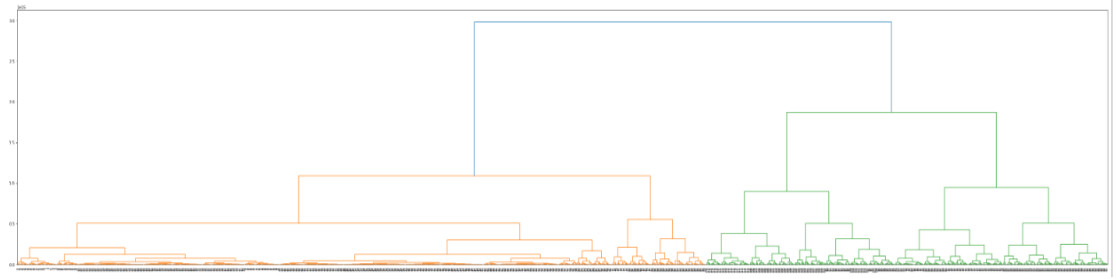
e- ISSN: 2828–299X



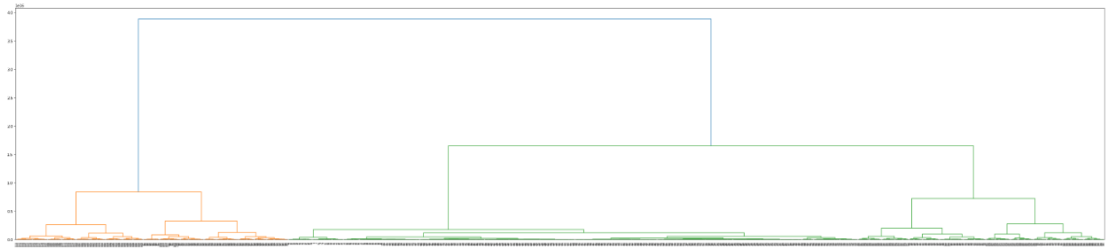
Gambar 7. Dendrogram *Single Linkage*



Gambar 8. Dendrogram A*verage Linkage*



Gambar 9. Dendrogram *Complete Linkage*



Gambar 10. Dendrogram *Ward Linkage*

Perbandingan *cluster* data dari keempat metode *Agglomerative Nesting* dapat dilihat pada Tabel 4.

Tabel 4. Hasil *Clustering* Metode *Agglomerative Nesting*

| tweetid | favorite_count | retweet_count | user_followers_count | h*cluster* | A*cluster* | C*cluster* | w*cluster* |
|---|---|---|---|---|---|---|---|
| 1578473399483240000 | 1 | 0 | 654 | 0 | 0 | 1 | 0 |
| 1578455541168320000 | 3 | 0 | 128 | 0 | 0 | 1 | 0 |
| 1578454872890170000 | 1 | 0 | 1355 | 0 | 0 | 1 | 0 |
| 1578451477512600000 | 5 | 1 | 1058 | 0 | 0 | 1 | 0 |
| 1578428025611240000 | 2 | 1 | 742 | 0 | 0 | 0 | 0 |
| 1578423223141530000 | 2 | 0 | 83 | 0 | 0 | 0 | 0 |

Pada Tabel 4 dapat dilihat potongan hasil cluster dari metode *Agglomerative Nesting*. Pada kolom hcluster merupakan *clustering Agglomerative Nesting* dengan single linkage, kolom Acluster merupakan *clustering Agglomerative Nesting* dengan average linkage, kolom Ccluster merupakan *clustering Agglomerative Nesting* dengan complete linkage, kolom wcluster merupakan *clustering Agglomerative Nesting* dengan ward linkage.

Setelah melakukan kelima metode *clustering* untuk menentukan metode yang terbaik dilakukan perbandingan nilai *silhouette_score* masing-masing metode yang dapat dilihat pada Tabel 4.

e- ISSN: 2828–299X

Tabel 5. *Silhouette_Score*

| Metode | *Silhouette_Score* |
|---|---|
| K-Means | 0.753254846 |
| *Agglomerative Nesting Single linkage* | 0.636551323 |
| *Agglomerative Nesting Average linkage* | 0.745342050 |
| *Agglomerative Nesting Complete linkage* | 0.754428965 |
| *Agglomerative Nesting Ward linkage* | 0.706613687 |

Dilihat dari Tabel 5 *silhouette_score* metode *score* K-means 0,7532, metode *Agglomerative Nesting Single linkage* 0,6365, metode *Agglomerative Nesting Average linkage 0,7453,* metode *Agglomerative Nesting Complete linkage* 0,7544, dan *Agglomerative Nesting Ward linkage* 0,7066.

## 4 SIMPULAN

Menentukan metode yang terbaik dapat dilihat memalui *silhouette_score.* Suatu data dianggap berada di *cluster* yang benar memiliki nilai silhouette coefficient lebih dari 0, sedangkan data yang dianggap tidak berada di *cluster* yang benar memiliki nilai *Silhouette coefficient* kurang dari 0. Hasil perbandingan antara metode K-means dan *Agglomerative Nesting* didapatkan bahwa *silhouette_score* tertinggi adalah dengan metode *Agglomerative Nesting complete linkage* dengan score 0.754428965 dengan jumlah *K*=2, sehingga dapat disimpulkan bahwa *cluster* terbaik untuk data *digital marketing* Twitter adalah *Agglomerative Nesting dengan complete linkage*. Pada urutan kedua metode dengan *score* tinggi terdapat metode K-means dengan *score* 0.753254846 dan nilai yang terendah yaitu *single linkage* dengan *score* 0.636551323. Sehingga kurang disarankan menggunakan single linkage untuk digunakan pada data *digital marketing*. Berdasarkan hasil cluster metode *Agglomerative Nesting complete linkage* jumlah data tweet pada Cluster 0 sebanyak 432 data dan Cluster 1 sebanyak 726 data. Cluster 0 memiliki karakteristik tweet.

## 5 SARAN

Kedepannya dapat dilakukan penelitian dengan metode *clustering* lain atau metode gabungan *clustering* seperti gabungan metode K-means dan *Agglomerative Nesting* untuk mendapatkan nilai ketepatan yang lebih tinggi. Sedangkan Penelitian lanjutan dibidang marketing dapat lebih spesifik dan berfokus mengenai karakteristik tweet yang banyak menarik perhatian dalam *trending* topik juga dapat dilakukan oleh peneliti selanjutnya.

## DAFTAR PUSTAKA

[1] B. Santosa. 2007. Data Mining. Teknik Pemanfaatan Data untuk Keperluan Bisnis, First Edition ed. Yogyakarta: Graha Ilmu.

[2] Witten, I. H., & Frank, E. 2005. Data Mining: Practical machine learning tools and techniques 2nd edition. San Francisco: Morgan Kaufmann.

[3] Agusta, Y., 2007. K-means–penerapan, permasalahan dan metode terkait. Jurnal Sistem dan informatika, 3(1), pp.47-60.

[4] Han, J, & Kamber, M., 2006 . Data Mining: Concept and Techniques 2nd Edition. San Francisco: Morgan Kaufmann Publisher.

[5] Sari, E. P., & Aras, R. A. 2021. Analisis Tingkat Ketepatan Digital Marketing Pada Facebook cOleh Online Shop di Thailand Tahun 2018. Seminar Nasional Sains dan Teknologi Informasi (SENSASI), 138 - 142.

[5] Dihni, V. A., & Bayu, D. J., 2021. Inilah 10 Negara dengan Pengguna Twitter Terbanyak, Ada Indonesia? databoks.katadata.co.id.

[6] Han, J, & Kamber, M. 2006. Data Mining: Concept and Techniques 2nd Edition. San Francisco: Morgan Kaufmann Publisher.

[7] Ghozali, I. 2011. Model persamaan struktural: Konsep dan aplikasi dengan program Amos 19.0, Semarang: Badan Penerbit Universitas Diponegoro

[8] Tan, P., Steinbach, M. & Kumar, V. 2006. Introduction to Data Mining.USA: Person Education,Inc.