# *CLUSTERING* PADA DATA SENTIMEN BPJS KESEHATAN MENGGUNAKAN ALGORITMA *AGGLOMERATIVE HIERARCHICAL CLUSTERING AVERAGE LINKAGE*

# *CLUSTERING ON SENTIMENT DATA OF BPJS KESEHATAN USING AGGLOMERATIVE HIERARCHICAL CLUSTERING AVERAGE LINKAGE ALGORITHM*

**Tinton Aji Sadewo[1], Purba Daru Kusuma[2], Casi Setianingsih[3]**

[1,2,3] Universitas Telkom, Bandung
**[1]tintonajisadewo@student.telkomuniversity.ac.id, [2]purbodaru@telkomuniversity.ac.id, [3]setiacasie@telkomuniversity.co.id**

**Abstrak**

Pada era globalisasi ini menjadikan media sosial khusunya Twitter sebagai sarana komunikasi. Masyarakat dengan mudah mendapatkan informasi dengan mudah di sosial media. Tidak hanya mudah mendapatkan informasi, masyarakat Indonesia khusunya juga dapat memberikan komentar atau opini, bertukar informasi, mengunggah foto serta video yang tersedia di Twitter. Pengguna Twitter lebih banyak menyampaikan opini di kolom komentar Twitter.

Dalam penelitian tugas akhir ini akan dilakukan *clustering* data dari opini atau komentar pengguna Twitter. Komentar atau opini yang disampaikan oleh pengguna BPJS Kesehatan di Twitter sudah sangat banyak, mulai dari komentar negatif, positif dan netral. Menjadikan Twitter wadah penampung kritik dan saran terkait dengan layanan dan program yang diberikan BPJS Kesehatan. Munculnya data yang banyak maka pengguna BPJS yang memiliki akun *Twitter* kesulitan untuk melihat kualitas layanan atau program yang diberikan oleh BPJS Kesehatan.

Untuk mempermudah pengguna melihat kualitas layanan atau program yang diberikan BPJS Kesehatan, maka pada penelitian ini dibuat sistem menggunakan metode *Agglomerative Hierarchical Average Linkage* untuk *clustering* dari data komentar pada akun resmi Twitter BPJS Kesehatan. Data dikelompokkan berdasarkan opini positif, negatif, dan netral. Data yang digunakan sudah di validasi oleh Balai Bahasa Jawa Barat yaitu berjumlah 2118 data yang dikelompokan menurut cluster dan di tampilkan di website yang di rancang pada penilitian Tugas Akhir ini. Dari hasil penelitian pada tugas akhir ini dalam clustering pada pengguna Twitter mendapat hasil sillhoutte coefficient data positif sebesar 0.9912, data negatif sebesar 0.9953, dan data netral 0.9923.

**Kata Kunci:** *Clustering, Twitter,* BPJS Kesehatan, *Agglomerative Hierarchical Average Linkage.*

**Abstract**

In this era of globalization, twitter is especially a means of communication. People easily get information on social media. Not only easy to get information, Indonesian people especially can also give comments or opinions, exchange information, upload photos and videos available on Twitter. Twitter users have more opinions in Twitter's comments field.

In the study this final task will be done *clustering* data from the opinions or comments of Twitter users. Comments or opinions submitted by BPJS Kesehatan users on Twitter have been very much, ranging from negative, positive and neutral comments. Make Twitter a container for criticism and suggestions related to services and programs provided by BPJS Kesehatan. The emergence of a lot of data, bpjs users who have Twitter accounts have difficulty to see the quality of services or programs provided by BPJS Health.

To make it easier for users to see the quality of services or programs provided by BPJS Kesehatan, the research was created using *the Agglomerative Hierarchical Average Linkage* method for *clustering* from the comment data on bpjs kesehatan's official Twitter account. The data is grouped by positive, negative, and neutral opinions. The data used has been validated by the West Java Language Hall which amounts to 2118 data grouped by cluster and displayed on the website designed in this Final Task research. From the results of the study on this final task in clustering on Twitter users got a positive data sillhoutte coefficient result of 0.9912, negative data of 0.9953, and neutral data 0.9923.

**Keywords:** *Clustering, Twitter,* BPJS Kesehatan, *Agglomerative Hierarchical Average Linkage.*

## 1. PENDAHULUAN

Masyarakat Indonesia mengenal Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan sebagai Lembaga yang baru muncul di publik, namun lembaga ini sudah dibentuk pada tahun 1968 dengan sebutan yang berbeda dengan sekarang. Banyak masyarakat Indonesia yang mengenal BPJS ini dengan sebutan Askes.[1] Namun pada tahun 2011 membuat UU Nomor 24 Tahun 2011 tentang Badan Penyelenggara Jaminan Sosial (BPJS) maka setelah terbitnya Undang-Undang tersebut PT. Askes (Persero) berganti nama menjadi BPJS Kesehatan. Dengan adanya Program-program dari pemerintah tentang Jaminan Kesehatan Nasional-Kartu Indonesia Sehat (JKN-KIS) yang dilaksanakan oleh BPJS Kesehatan, pemerintah memiliki tujuan menjamin kesehatan yang komprehensif, adil dan merata kepada seluruh rakyat Indonesia.[2]

Seiring berkembangnya media sosial, BPJS Kesehatan memberikan informasi melalui akun resmi di Twitter. Pengguna lebih mudah memberikan berbagai opini baik pujian maupun keluhan di media

sosial Twitter BPJS Kesehatan. Opini dari pengguna harusnya menjadi salah satu prioritas utama bagi BPJS Kesehatan sebagai tolak ukur untuk meningkatkan kualitas layanan atau program BPJS Kesehatan. Banyaknya opini yang disampaikan pengguna layanan BPJS Kesehatan menyulitkan pengguna baru untuk melihat taraf kualitas layanan BPJS Kesehatan berdasarkan kelompok sentimennya. Akibat opini dari pengguna BPJS Kesehatan yang tidak di prioritaskan maka akan berdampak pada penurunan pengguna.

Maka untuk mempermudah dalam pengelompokan opini yang ada di BPJS kesehatan, penulis akan membuat sebuah sistem berbasis web menggunakan *clustering* dengan metode *Agglomerative Hierarchical Clustering Average Linkage*. Metode *Average Linkage* memiliki kinerja yang stabil di bandingkan metode *Complete Linkage* dan *Single Lingkage*,[3] pada penelitian ini dilakukan clustering data yang ada pada sentiment pengguna BPJS Kesehatan yang memiliki akun twitter menggunakan *Agglomerative Hierarchical Clustering Average Linkage* sebagai cara untuk menyelesaikan permasalahan dalam penelitian clustering data sentiment BPJS Kesehatan di Twitter.

## 2. TINJAUAN PUSTAKA

### 2.1 Media Sosial

Media sosial adalah sekumpulan individu, organisasi, entitas lain yang terhubung melalui internet.[4] Internet semakin kesini semakin berkembang membuat media sosial dapat menghubungkan pengguna ke berbagai daerah maupun negara yang ada di dunia maya. Contoh dari aktivitas di dunia maya merupakan *chating*, kirim pesan elektronik, komentar di twitter itulah beberapa momen yang ada di media sosial.

### 2.2 Text Mining

*Text mining* ialah suatu istilah untuk menambang data dalam bentuk *text* yang dijalankan secara otomatis oleh perangkat komputer untuk mencari suatu informasi yang berkualitas dari kumpulan teks tersusun dalam sebuah dokumen atau data. [5] Pada *text mining* data yang diolah adalah data tekstual untuk diproses dan dianalisis menjadi suatu informasi dengan metode *classification*, *clustering*, dan *information retrieval*.[6] Tahapan utama dalam melakukan *text mining* ini dengan menemukan sekumpulan kata yang mewakili dari isi dokumen maka selanjutnya dilakukan analisis

keterhubungan antar data-data atau dokumen menggunakan metode statistika semacam klasifikasi, analisis kelompok, dan asosiasi.

### 2.3 Preprocessing

*Preprocessing* adalah langkah awal mengolah data yang diperoleh agar mudah diproses oleh sistem dengan mudah. *Preprocessing* memiliki beberapa proses dengan tujuan untuk mengubah data menjadi sesuai dan mudah untuk diproses dan juga untuk membuang sebagian data yang tidak di perlukan sistem. Langkah *preprocessing* ini sangat penting saat dilakukan analisis mengenai teks yang ada pada media sosial khusus nya twitter, dikarenakan media sosial Sebagian besar berisi teks atau kata-kata atau kalimat yang tidak memiliki ejaan yang benar atau tidak formal yang memiliki noise besar. [7]


Gambar 1. Preprocessing

Langkah-langkah *preprocessing* secara signifikan mempengaruhi kinerja dari pembelajaran mesin.[8] Menghapus noise atau menghilangkan gangguan pada data adalah salah satu proses harus dilakukan agar mendapatkan hasil yang diharapkan secara maksimal.

Pada penelitian tugas akhir ini data sudah dilakukan preprocessing antara lain *case folding*, *tokenizing*, *filtering* dan akan mendapatkan hasil *clustering* yang maksimal pada penelitian ini maka dilakukan tahap stopword yang memiliki tujuan menghilangkan kata-kata yang sering sering muncul seperti “yang”, “dan”, “atau” dsb. Kata-kata itu perlu di hilangkan karena tidak memiliki makna dan tidak akan berpengaruh pada proses *clustering*.[9] Proses berikutnya adalah *stemming* yang memiliki tujuan menghilangkan kata imbuhan pada awalan maupun akhiran untuk mendapatkan kata dasar dan meningkatkan kinerja *clustering*.

### 2.4 Clustering

*Clustering* artinya proses pengelompokan sehingga seluruh data pada partisi memiliki persamaan sesuai matriks masing-masing.[10] Tujuan *clustering* adalah untuk mengenali dengan jelas perbedaan dalam suatu data set dan memberikan nama kelompok pada setiap penelitian. Mengelompokkan data *clustering* dilakukan tanpa memandang kelas tertentu, *clustering* digunakan untuk memeberikan label di kelas data yang belum dimengerti. Oleh karena itu, *clustering* terdekat diklasifikasikan sebagai metode *unsupervised*

*learning*.[11] *Clustering* juga dapat diterapkan pada data *mining*, secara khusus *clustering* dirancang untuk menemukan kelompok dalam sauatu dokumen. [12]

*Clustering* memiliki prinsip antara lain memaksimalkan kesamaan data antar *cluster* dan meminimalkan kesamaan data antar *cluster*. *Clustering* dapat dikelompokan dalam bidang dua dimensi dimana data set dikelombokkan menjadi beberapa *cluster* yang memiliki pusat *cluster* bertanda positif (+).[13]

### 2.5 Pembobotan TF-IDF

Pembobotan TF-IDF adalah pembobotan kata menggunakan frekuensi kata yang muncul dalam suatu dokumen yang dibentuk dalam suatu vector yaitu vector space model.[14] Supaya mendapatkan bobot dari nilai $tf(wtf)$ dapat dihitung dengan persamaan rumus:

$$wtf = 1 + \log(tf) \quad (1)$$

tf: Frekuensi kemunculan term dalam dokumen.

Untuk menghitung nilai Idf, dapat menggunakan rumus berikut:

$$idf = \log\left(\frac{N}{df}\right) \quad (2)$$

N: Jumlah total dokumen yang digunakan.

df: Jumlah dokumen di mana term yang dipilih muncul.

Untuk menghitung nilai bobot TfIdf dapat digunakan rumus sebagai berikut:

$$TF - IDF = Wtf \times idf \quad (3)$$

### 2.6 Euclidean Distance

*Euclidean distance* merupakan jarak antara dua titik pada garis lurus. Metode perhitungan ini menggunakan rumus *Pythagoras*, yaitu perhitungan jarak antar data yang sering digunakan dalam *machine learning*. [15] Rumus dari *Euclidiean Distance* sebagai berikut:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}\left(x_{ik} - x_{jk}\right)^2} \quad (4)$$

Dimana:

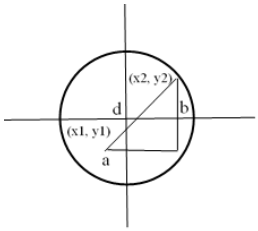
dij= perhitungan jarak untuk kemiripan

n= jumlah vector

xik= vector citra masukan

xjk= vector citra pembanding

Dari persamaan 4, bentuk dari perhitungan jarak *Euclidean Distance* adalah lingkaran yang ditunjukkan pada Gambar 2.



Gambar 2. Euclidean Distance

Keterangan:

a= $x_2 - x_1$

b= $y_2 - y_1$

rumus Pythagoras

1. $a^2 + b^2 = d^2$ (5)
2. $d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$(6)

### 2.7 Aglomerative Hierarchical Clustering

*Agglomerative Hierarchical Clustering* (AHC) merupakan prosedur analisis clustering, yang mengadopsi metode bottom-up, sehingga struktur hirarki dimulai dari setiap node atau inti menjadi satu *cluster*, dan kemudian menggabungkan node atau inti menjadi satu. sebagai *cluster* yang lebih besar sampai hanya terdapat satu cluster yang tersisa. [16]Untuk menyatukan node dalam satu *cluster, Agglomerative Hierarchical Clustering* menghitung persamaan dengan ukuran jarak antar simpul menggunakan tiga jenis perhitungan jarak yang popular, antara *lain single-linkage*(jarak dekat), *complete-linkage*(jarak jauh), *average-linkage*(jarak rata-rata).[16] Pada penelitian ini penulis menggunakan metode *Average Linkage*.

#### 2.7.1 Average Linkage

Pada perhitungan dengan metode *average linkagae*, jarak antara dua *cluster* dianggap sebagai jarak rata-rata antara data untuk semua anggota satu *cluster* dan data untuk semua anggota *cluster* lain.

Rumus jarak *Average Linkage*:

$$d_{(IJ)K} = \frac{\sum_a \sum_b d_{ab}}{N_{IJ} N_K} \qquad (2.7)$$

$d_{ab}$ = jarak antara objek i pada cluster (IJ) dan objek b pada cluster K

$N_{IJ}$ = jumlah data pada cluster (IJ)

$N_K$= jumlah elemen pada cluster (IJ) dan pada K.

### 2.8 Silhouette Coefficient

*Silhouette Coefficient* adalah metode evaluasi digunakan untuk melihat kekuatan dan kualitas *cluster* serta kualitas suatu objek dalam sebuah *cluster*.[17] Cara menghitung *silhouette coefficient* :

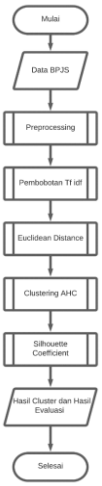
$$SC = \frac{b_i - a_i}{max(a_i b_i)} \qquad (8)$$

## 3. METODELOGI PENELITIAN

### 3.1 Data Penelitian

Data yang digunakan studi kasus penelitian ini adalah data sentimen dari BPJS Kesehatan yang ada di media sosial Twitter dari akun yang tidak *privat* menggunakan kata kunci "BPJS Kesehatan" pada tanggal 29 Januari 2020 sampai 7 Juni 2020 untuk data uji. Data masukan yang digunakan berjumlah 2116 tweet, dengan 716 tweet memiliki label positif, 702 tweet memiliki label negatif, dan 705 tweet memiliki label netral.

### 3.2 Diagram alir




Gambar 3. Diagram Alir

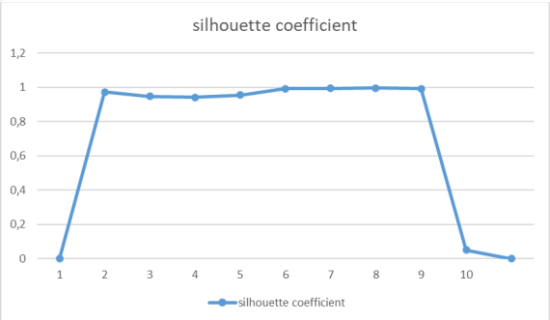
Tahapan dari sistem umum terdiri dari:

1. Sistem menerima masukan dataset berupa file .csv yang memiliki isi komentar dari akun Twitter BPJS.
2. Kemudian sistem melakukan proses berikutnya yaitu *preprosesing* yang terdiri dari *stop removal, stemming, TF-IDF*. Salah satu proses *preprocessing* yaitu TF-IDF dilakukan agar bisa mendapatkan nilai dan mengetahui seberapa besar bobot suatu kata. Hasil *summarize* (peringkasan) didapatkan dengan cara melakukan pemilihan atau sorting dari TF-IDF, dan hasil sorting diurutkan dari yang besar untuk mendapatkan hasil ringkasan yang sesuai dengan presentase yang peneliti inginkan.
3. Proses selanjutnya adalah proses penghitungan jarak antar dokumen menggunakan metode *Euclidean Distance*.
4. Kemudian dilakukan proses *clustering* menggunakan metode *average linkage* yaitu mencari jarak rata-rata dari antar dokumen, apabila memiliki jarak rata-rata yang sama maka akan digabungkan menjadi satu *cluster* yang sama.
5. Proses terakhir, yaitu dilakukan tahap evaluasi menggunakan *silhouette coefficient* agar mengetahui hasil dari cluster itu bagus atau kurang optimal.

## 4. HASIL DAN PEMBAHASAN

### 4.1 Hasil Pengujian Average Linkage Data Sentiment Negatif

Pada pengujian menggunakan parameter jarak *average linkage* dan data negatif untuk mendapatkan 10 cluster dan mendapatkan titik potong dimulai dari 683 sampai 673.

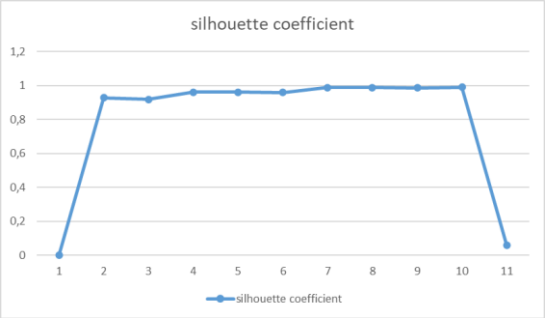



Gambar 4. Hasil Pengujian Data Negatif

Hasil dari pengujian *clustering* data sentiment BPJS Kesehatan menggunakan *average linkage* data negatif dapat dilihat pada gambar 4 hasil yang optimal terdapat pada jumlah *cluster* 8 dan titik potong 675 yang memiliki rata-rata dari *silhouette coefficient* keseluruhan datanya adalah 0.9953. Sedangkan untuk hasil paling rendah terdapat pada jumlah *cluster* 1 dan memiliki titik potong 683 yang memiliki rata-rata *silhouette coefficient* keseluruhan data adalah 0. Dapat disimpulkan bahwa pada jumlah *cluster* 8 memiliki hasil paling optimal dalam pengelompokan data sentiment BPJS Kesehatan data negatif karena mendapatkan hasil *silhouette coefficient* paling optimal. Mengenai jumlah cluster, hasil 1 belum optimal karena memiliki *silhouette coefficient* yang paling rendah.

### 4.2 Hasil Pengujian Average Linkage Data Sentiment Positif

Pada pengujian menggunakan parameter jarak *average linkage* data negatif untuk mendapatkan 11 *cluster* dan mendapatkan titik potong dimulai dari 685 sampai 674.



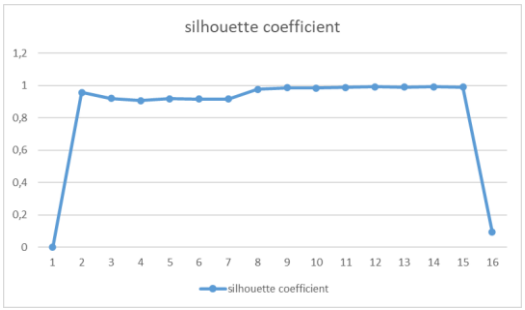

Gambar 5. Hasil Pengujian Data Positif

Hasil dari pengujian *clustering* data sentiment BPJS Kesehatan menggunakan *average linkage* data positif dapat dilihat pada gambar 5 hasil yang optimal terdapat pada jumlah *cluster* 10 dan titik potong 675 yang memiliki rata-rata dari *silhouette coefficient* keseluruhan datanya adalah 0.9912. Sedangkan untuk hasil paling rendah terdapat pada jumlah *cluster* 1 dan memiliki titik potong 683 yang memiliki rata-rata *silhouette coefficient* keseluruhan data adalah 0. Dapat disimpulkan bahwa pada jumlah *cluster* 10 memiliki hasil paling optimal dalam pengelompokan data sentiment BPJS Kesehatan data positif karena mendapatkan hasil *silhouette coefficient* paling optimal. Mengenai jumlah cluster, hasil 1 belum optimal karena memiliki *silhouette coefficient* yang paling rendah.

### 4.3 Hasil Pengujian Average Linkage Data Sentiment Netral

Pada pengujian menggunakan parameter jarak *average linkage* data negatif untuk mendapatkan 16 *cluster* dan

mendapatkan titik potong dimulai dari 672 sampai 656.




Gambar 6. Hasil Pengujian Data Netral

Hasil dari pengujian *clustering* data sentiment BPJS Kesehatan menggunakan *average linkage* data netral dapat dilihat pada gambar 6 hasil yang optimal terdapat pada jumlah *cluster* 14 dan titik potong 658 yang memiliki rata-rata dari *silhouette coefficient* keseluruhan datanya adalah 0.9923. Sedangkan untuk hasil paling rendah terdapat pada jumlah *cluster* 1 dan memiliki titik potong 672 yang memiliki rata-rata *silhouette coefficient* keseluruhan data adalah 0. Dapat disimpulkan bahwa pada jumlah *cluster* 14 memiliki hasil paling optimal dalam pengelompokan data sentiment BPJS Kesehatan data netral karena mendapatkan hasil *silhouette coefficient* paling optimal. Mengenai jumlah cluster, hasil 1 belum optimal karena memiliki *silhouette coefficient* yang paling rendah.

### 4.3 Hasil Analisis

Pada analisis pengujian average linkage dilakukan dengan membandingkan ketiga data berdasarkan nilai rata-rata hasil pengujian silhouette coefficient paling tinggi. Untuk standar nilai dari silhouette coefficient berada pada -1 sampai 1. Saat nilai hasil silhouette coefficient mendekati 1 maka data yang di proses pada cluster yang optimal. Sedangkan pada nilai hasil silhouette coefficient mendekati -1 atau o maka data yang di proses pada cluster kurang optimal.

Tabel 1 Analisis Hasil

| Data | Silhouette Coefficient |
|---|---|
| Negatif | 0.9953 |
| Positif | 0.9912 |
| Netral | 0.9923 |

Pada table 1 bahwa untuk nilai rata-rata silhouette coefficient paling tinggi terdapat pada data negatif dengan memiliki nilai 0.9953. untuk nilai silhouette coefficient data negatif sangat mendektai nilai 1 sehingga disimpulkan bahwa cluster yang terbentuk dengan optimal. Untuk nilai silhouette coefficient dipengaruhi dari jarak kedekatan antar data penelitian ini dihitung dengan Euclidean Distance. Maka semakin dekat jarak antar data maka akan mendapatkan

hasil silhouette coefficient yang optimal. Pada penelitian ini yang dapat mempengarui tingkat ketepatan clustering pada data yaitu jumlah data yang di uji dan keberagaman data. Data dengan nilai silhouette coefficient paling optimal yaitu pada data negatif yang memiliki jumlah cluster 8 dan titik potong 683.

## 5. KESIMPULAN DAN SARAN

### 5.1 Kesimpulan

Berdasarkan pengujian dan analisis dari *clustering* pada data sentimen bpjs Kesehatan menggunakan algoritma *agglomerative hierarchical clustering average linkage* maka kesimpulan dari penelitian ini sebagai berikut ini:

1. Pada penerapan metode agglomerative hierarchical clustering average linkage mendapatkan hasil pengujian pada penelitian ini dengan data negatif yang memiliki nilai jarak terbaik dan paling optimal. Pada data negatif memberikan hasil yang optimal pada jumlah cluster 8 serta titik potong 683.
2. Tingkat akurasi pada metode agglomerative hierarchical clustering average linkage yang di uji dengan silhouette coefficient pada data negatif adalah nilai rata-rata paling tinggi dengan nilai 0.9953.

### 5.2 Saran

Berdasarkan kesimpulan, peneliti memiliki saran untuk pembaca antara lain:

1. Pada penelitian selanjutnya dapat menggunakan cara pengelompokan Hierarchical yang lain untuk mengelompokan data sentiment. Dengan menambahkan variasi fitur dan jumlah data yang lebih banyak sehingga mendapatkan hasil yang meningkatkan kompleksitas data yang akan di gunakan untuk clustering data sentiment.
2. Dapat menggunakan metode pengukuran dan parameter jarak yang lain untuk meningkatkan akurasi yang lebih optimal dari hasil penelitian ini.

## REFERENSI


[1] T. Yuniarto, "Badan Penyelenggara Jaminan Sosial Kesehatan," 2020. https://kompaspedia.kompas.id/baca/profil/lembaga/badan-penyelenggara-jaminan-sosial-kesehatan.

[2] BPJS, "Sejarah Perjalanan Jaminan Sosial di Indonesia," *2018*, 2020. https://bpjs-kesehatan.go.id/bpjs/index.php/pages/detail/2013/4.

[3] R. P. Justitia, N. Hidayat, and E. Santoso, "Implementasi Metode Agglomerative Hierarchical Clustering Pada Segmentasi Pelanggan Barbershop ( Studi Kasus : RichDjoe Barbershop Malang )," vol. 5, no. 3, pp. 1048–1054, 2021.

[4] M. A. Laagu and A. Setyo Arifin, "Analysis the Issue of Increasing National Health Insurance (BPJS Kesehatan) Rates through Community Perspectives on Social Media: A Case Study of Drone Emprit," *Proceeding - ICoSTA 2020 2020 Int. Conf. Smart Technol. Appl. Empower. Ind. IoT by Implement. Green Technol. Sustain. Dev.*, 2020, doi: 10.1109/ICoSTA48221.2020.1570615599.

[5] R. Feldman and J. Sanger, *The Text Mining Handbook : Advanced Approaches to Analyzing Unstructured Data*. Cambridge: Cambridge University Press, 2007.

[6] W.Berry and J. Kogan, "Text Mining: Applications and Theory," 2010.

[7] S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[8] S. B. Kotsiantis and D. Kanellopoulos, "Data preprocessing for supervised leaning," *Int. J. …*, vol. 1, no. 2, pp. 1–7, 2006, doi: 10.1080/02331931003692557.

[9] A. I. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification," *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 6, pp. 22–32, 2018, [Online]. Available: https://sites.google.com/site/ijcsis/.

[10] A. Nugraha, M. Arista Harum Perdana, H. Agus Santoso, J. Zeniarja, A. Luthfiarta, and A. Pertiwi, "Determining the Senior High School Major Using Agglomerative Hierarchial Clustering Algorithm," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 225–228, 2018, doi: 10.1109/ISEMANTIC.2018.8549834.

[11] I. Pramudiono, *Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data*. 2007.

[12] D. Sailaja, M. Kishore, B. Jyothi, and N. Prasad, "An Overview of Pre-Processing Text Clustering Methods.," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 3, pp. 3119–3124, 2015.

[13] N. Pitaloka, "Pengelompokan Data Menggunakan Hierarchical Clustering (AHC)," 2009.

[14] H. Zayuka, S. M. Nasution, and Y. Purwanto, "Perancangan Dan Analisis Clustering Data Menggunakan Metode K-Medoids Untuk Berita Berbahasa Inggris Design and Analysis of Data Clustering Using K-Medoids Method For English News," *e-Proceeding Eng.* , vol. 4, no. 2, pp. 2182–2190, 2017.

[15] F. H. Saad, O. I. E. Mohamed, and R. E. Al-qutaish, "C Omparison of H Ierarchical a Gglomerative a Lgorithms F or C Lustering M Edical," vol. 3, no. 3, pp. 1–15, 2012.

[16] S. Viriyavisuthisakul, P. Sanguansat, P. Charnkeitkong, and C. Haruechaiyasak, "A comparison of similarity measures for online social media Thai text classification," *ECTI-CON 2015 - 2015 12th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol.*, pp. 0–5, 2015, doi: 10.1109/ECTICon.2015.7207106.

[17] A. Jaiswal and N. Janwe, "Hierarchical Document Clustering: A Review," *Int. J. Comput. Appl.*, pp. 37–41, 2011.