JUMANJI | Vol. 4 | No. 2 | Halaman 60 - 68
ISSN (p): 2598-8050 Februari 2021
ISSN (e): 2598-8069

Klasifikasi Fasilitas Umum di Jawa Tengah pada Twitter dengan Algoritma Agglomerative Hierarchical Clustering dan Naive Bayes Classifier

NISA'UL HAFIDHOH[1], ALDILAN NOORMAN SISSANDHY[2]

[1,2] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Email: nisa@dsn.dinus.ac.id

ABSTRAK

Provinsi Jawa Tengah memiliki wilayah yang luas dengan penduduk yang tersebar di berbagai kabupaten dan kota. Pemerintah pun senantiasa meningkatkan pembangunan daerah di berbagai wilayah tersebut. Beberapa keluhan serta masukan terkait fasilitas umum banyak disampaikan masyarakat untuk pemerintah Jawa Tengah, salah satunya melalui media sosial Twitter. Pada penelitian ini informasi terkait fasilitas umum yang ada pada sosial media Twitter diklasifikasikan ke dalam lima kelas, yaitu infrastruktur, layanan, transportasi, fasilitas dan lainnya. Klasifikasi dilakukan melalui dua tahap yaitu tahap clustering dengan algoritma Agglomerative Hierarchical Clustering dan tahap klasifikasi dengan algoritma Naïve Bayes Classification. Terapat 450 data yang diklasifikasi dengan akurasi global sebesar 78%.

***Kata kunci**: Twitter, Fasilitas umum, Jawa Tengah, Klasifikasi, Agglomerative Hierarchical Clustering , Naïve Bayes Classifier.*

ABSTRACT

Central Java has a large area with a population spread across various districts and cities. The government also continues to increase regional development in these areas. Many complaints and suggestions related to public facilities were conveyed by the public to the Central Java government, one of which was through the social media Twitter. In this study, information related to public facilities on Twitter social media is classified into five classes, namely infrastructure, services, transportation, facilities and others. Classification is carried out in two stages, namely the clustering stage with the Agglomerative Hierarchical Clustering algorithm and the classification stage with the Naïve Bayes Classification algorithm. There are 450 data classified with a global accuracy of 78%.

***Keywords**: Twitter, Public facility, Central java, Classification, Agglomerative Hierarchical Clustering, Naïve Bayes Classifier.*

1. PENDAHULUAN

Provinsi Jawa Tengah memiliki luas wilayah 32.800 km^2 atau sekitar 25,04% dari pulau Jawa. Provinsi Jawa Tengah memiliki 29 kabupaten serta 6 kota dengan kepadatan penduduk pada tahun 2019 mencapai 987,26 jiwa/km^2 yang tersebar di berbagai kabupaten serta kota (BPS, 2020). Dengan jumlah tersebut, pengaduan dari masalah yang dihadapi penduduk Jawa Tengah terhadap pembangunan yang dilakukan pemerintah Jawa Tengah cukup beragam. Salah satu hal yang sering diadukan masyarakat adalah keberadaan dan kondisi fasilitas umum yang banyak digunakan masyarakat. Fasilitas umum tersebut dapat berupa infrastruktur, transportasi umum, layanan umum dari pemerintah maupun fasilitas lainnya. Hal tersebut juga dapat menjadi tolak ukur pembangunan daerah dari suatu pemerintah.

Pemerintah Jawa Tengah telah melakukan upaya dengan menyediakan website layanan dimana masyarakat dapat menyampaikan keluhan atas kinerja pembangunan yang ada di sekitarnya. Akan tetapi masyarakat lebih suka mengutarakan pendapatnya melalui media sosial karena penggunaan yang mudah. Seiring banyaknya pengguna media sosial di Indonesia, menarik pemerintah untuk memanfaatkan media sosial sebagai media komunikasi dengan masyarakat (Carley, Malik, Kowalchuk, Pfeffer, & Landwehr, 2015). Instansi pemerintah serta para *public figure* pun juga banyak yang menggunakan media sosial karena lebih dekat dengan masyarakat yang banyak menjadi penggna media sosial seperti Facebook, Twitter, Instagram dan lainnya. Pemerintah Jawa Tengah serta pemangku jabatan pun memiliki beberapa akun resmi pada berbagai media sosial yang dikelola oleh pemerintah Jawa tengah atau individu, salah satu media sosial yang digunakan adalah Twitter.

Twitter telah berkembang sebagai sumber akan beragam informasi, pengguna aktif harian Twitter di Indonesia pada awal tahun 2020 mencapai 166 juta pengguna, meningkat 24% dari tahun 2019 (Jati, 2020). Para pengguna twitter dapat menuliskan dan mengakses opini secara *real-time* tentang berbagai macam topik dan informasi, membahas isu-isu yang ada, hingga memberikan reaksi tentang suatu informasi yang mereka dapatkan sehari-hari (Bharti, Pradhan, Babu, & Jena, 2016). Hal ini dapat menjadi bahan evaluasi serta masukan bagi pemerintah terkait opini dari masyarakat terhadap pemerintah. Pemerintah dapat mengetahui pendapat, saran serta keluhan masyarakat terkait informasi penyelenggaraan kegiatan, fasilitas umum, kondisi masyarakat serta informasi lainnya untuk membantu menyelenggarakan pembangunan daerah yang lebih baik. Dari banyaknya *tweet* yang ada, dapat dilihat berbagai macam opini atau tanggapan pengguna twitter akan suatu informasi, sehingga dapat dikelompokkan sesuai kategori opini pengguna twitter. Pemerintah pun memerlukan cara untuk memilah informasi yang ada agar dapat dimanfaatkan dengan tepat.

Pengelempokkan data berdasarkan ciri-ciri objek dapat dilakukan dengan metode klasifikasi (Wibawa, Purnama, Akbar, & Dwiyanto, 2018). Klasifikasi dapat dilakukan dengan cara manual maupun bantuan teknologi. Klasifikasi berbagai *tweet* tentang suatu topik yang sesuai adalah hal yang sulit jika dilakukan secara manual. Untuk melakukan klasifikasi dengan bantuan teknologi dapat menggunakan beberapa algoritma yang sudah ada, seperti *K-means, Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), Maximum Entropy (ME)* untuk *Supervised learning* (Wibawa, Purnama, Akbar, & Dwiyanto, 2018). Sedangkan untuk *Unsupervised learning* dapat menggunakan *clustering* atau klasifikasi untuk data yang belum ditentukan atau berlabel (Unnisa, Ameen, & Raziuddin, 2016). Salah satu metode yang dapat digunakan dalam proses pengelompokan opini atau *clustering* data tersetruktur maupun tidak tersetruktur adalah *Agglomerative hierarchical clustering (AHC)*. AHC dapat mengurangi biaya komputasi dengan membangun hirarki dari *centroid* dibanding data mentah sehingga cocok untuk *clustering* data dunia nyata yang belum berlabel (Bouguettaya, Yu, Liu, Zhou, & Song,

2015). Oleh karena itu, dalam penelitian ini dipilih AHC untuk melakukan *clustering* data set yang random seperti pada Twitter.

Setelah *tweet* dikelompokkan berdasar hasil *clustering*, selanjutnya perlu dilakukan proses klasifikasi agar hasil pengelompokkan tersebut dapat sesuai dengan kategorinya. Pada penelitian ini akan digunakan metode *Naïve Bayes Classifier* (NBC) untuk proses klasifikasi. Metode ini dipilih karena memiliki waktu pemrosesan yang singkat serta metode ini cocok untuk melakukan klasifikasi data teks yang berasal dari sosial media Twitter juga sosial media yang lain (Ibrahim & Yusoff, 2015). Dari hasil klasifikasi tersebut dapat dilihat kinerja dan keluhan terhadap fasilitas umum di Jawa Tengah yang dapat digunakan untuk membantu mengevaluasi kinerja pemerintah Jawa tengah dan mengetahui keluhan warga Jawa tengah yang ada pada sosial media twitter.

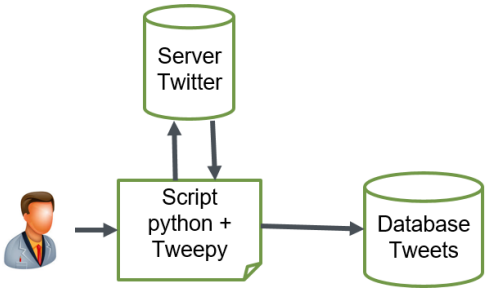
2. METODE

2.1. Metode Pengambilan Data

Pada penelitian ini data yang digunakan terdiri dari tiga jenis data yaitu :

a. Data *Tweet*
   Data *tweet* didapatkan dari proses *crawling* menggunakan bahasa pemrograman python dengan bantuan library Tweepy untuk mengakses Twitter API. Total data tweet yang digunakan sebesar 450 data. Data diambil dari akun resmi gubernur Jawa Tengah @ganjarpranowo dengan kata kunci seperti infrastruktur, fasilitas umum, pelayanan, dll.
b. Data *Stopword*
   Data *stopword* yang digunakan berjumlah 753 kata data dari tweet-tweet yang telah digunakan dalam penelitian sebelumnya.
c. Data Kata Dasar
   Data kata dasar diperoleh melalui kamus bahasa Indonesia online. Dengan data yang berjumlah 30.342 kata dasar.

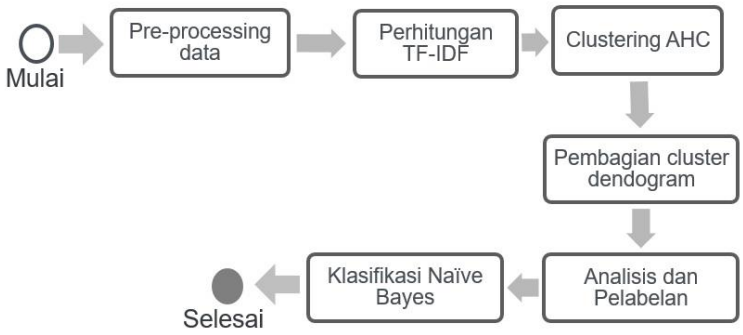
Pengumpulan data *tweet* menggunakan proses *crawling* untuk mengambil data dari media sosial Twitter seperti pada Gambar 1. Pengambilan data menggunakan *script* program dengan bahasa pemrograman Python dan tambahan *library* Tweepy untuk menghubungkan ke Twitter API. Pengambilan data dari server Twitter sesuai dengan hashtag @ganjarpranowo dengan kata kunci seperti infrastruktur, fasilitas umum, pelayanan dan lainnya. Data mentah yang berupa kumpulan *tweets* dalam bentuk .csv yang selanjutnya disimpan dalam Database Tweets.



Gambar 1 Proses Crawling Data

2.2. Metode Klasifikasi Opini

Pada penelitian ini untuk metode klasifikasi opini melalui proses *clustering* menggunakan algoritma *Agglomerative Hierarchical Clustering* (AHC) dan proses klasifikasi menggunakan algoritma *Naïve Bayes Classification* (NBC) seperti pada Gambar 2.



Gambar 2 Metode Klasifikasi Opini

Penjelasan metode klasifikasi opini pada Gambar 2 yang digunakan dalam penelitian ini adalah sebagai berikut :

a. Tahap awal dimulai dengan *pre-processing* data *tweets* yang didapat Dari hasil *crawling* data. Dalam *text pre-processing* terdiri dari tahapan *tokenisasi, stopword removal,* dan *stemming.*
b. Hasil dari proses *pre-processing* kemudian dilakukan perhitungan proses pembobotan dengan menghitung *tf-idf* dari hasil proses *stemming*.
c. Setelah mendapatkan hasil proses pembobotan setiap kata dengan *tf-idf* dilakukan perhitungan algoritma AHC metode *average linkage* dengan *Eucledian distance*. Proses ini dilakukan berulang-ulang hingga membentuk sebuah dendogram. Dendogram tersebut tersusun dari berbagai *cluster* yang dimulai dari *cluster-cluster* yang memiliki poin individu level yang paling bawah. Selanjutnya pengulangan akan menghasilkan sebuah penggabungan antara *cluster* satu dengan *cluster* lain yang memiliki tingkat atau sifat kesamaan yang paling tinggi.
d. Setelah mendapat hasil perhitungan algoritma AHC yang sudah berbentuk *dendogram* dari beberapa *cluster,* dilakukan pelabelan sesuai dengan isi *tweets,* yang dibagi menjadi 5 kelas sesuai dengan topik pembahasan.
e. Setelah mendapatkan data training yang sudah memiliki label atau kelas maka akan dilakukan proses klasifikasi dengan algoritma *Naive Bayes Classifier* dengan masukan data uji yang label atau kelas dari data tersebut dihiraukan.

3. HASIL DAN PEMBAHASAN

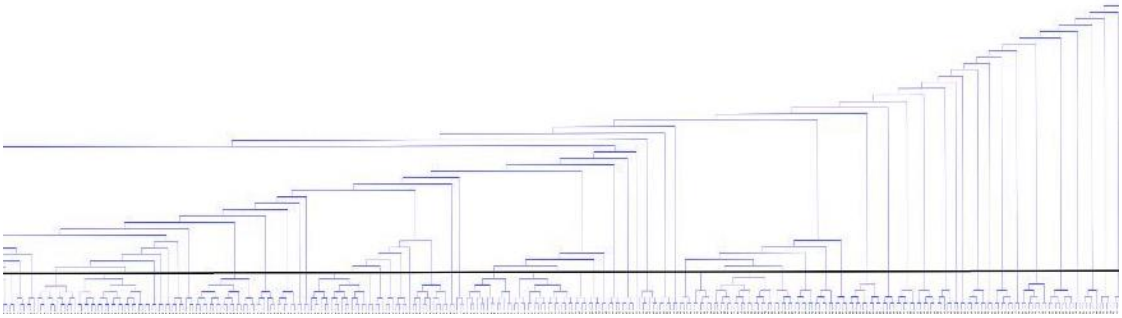
3.1 Analisis Data

Pada penelitian ini digunakan data sebanyak 450 *tweets* yang didapat melalui *crawling* dari akun sosial media Twitter resmi milik gubernur Jawa Tengah @ganjarpranowo. Data mentah tersebut masih acak dan belum sesuai dengan kategori, oleh karena itu diproses pada tahap awal yaitu *text pre-processing.* Contoh hasil *tweet* yang telah dilakukan *pre-processing* dapat dilihat pada Tabel 1.

Tabel 1. Contoh hasil *pre-processing tweets*

Tweet	Hasil pre-processing
@ganjarpranowo jam kerja jg bos, pegawe kelurahan kok jam 9 kdg ijek sepi	ganjarpranowo jam kerja bos pegawe kelurahan kok jam kdg ijek sepi
@ganjarpranowo Di Jepara sekarang banyak pabrik asing tapi jalan belum di perlebaran jadi macetnya kaya Jakarta pak ganjar Pranowo,,	ganjarpranowo jepara sekarang banyak pabrik asing jalan perlebaran jadi macetnya kaya pak ganjar pranowo
@ganjarpranowo Layanan transportasi antara kota dlm provinsi pak. Saya akui jalannya sdh bagus, tp kualitas angkutan umumnya jelek sekali	ganjarpranowo layanan kota provinsi pak akui jalannya bagus kualitas angkutan umumnya jelek sekali
@ganjarpranowo Jalannya sragen memprihatinkan nggih bopo...?? Tolong dibikin seperti jalan jalan didaerah purwokert… https://t.co/eAwaWwzwL0	ganjarpranowo jalannya sragen memprihatinkan nggih bopo dibikin jalan jalan didaerah
@ganjarpranowo Pak Ganjar Praniwo, sepertinya jalan utama Demak perlu dpt perhatian lebih, nacet krn perbaikan tentunya bisa di atur	ganjarpranowo pak ganjar praniwo jalan utama demak perlu perhatian lebih nacet perbaikan tentunya atur

Selanjutnya dilakukan proses perhitungan tf-idf terhadap term-term dari hasil *pre-processing* tersebut yang bertujuan untuk menentukan bobot setiap dokumen tweets. Setelah melalui perhitungan dan mendapatkan nilai tf-idf proses selanjutnya adalah *clustering* setiap dokumen-dokumen yang ada sehingga dapat dibagi ke beberapa kelas dengan algoritma AHC menggunakan rumus jarak *Eucledian* dengan *average linkage*. Hasil dari perhitngan *AHC* akan membentuk sebuah dendogram seperti pada Gambar 3.



Gambar 3 Potongan hasil dendogram

Dari perhitungan tersebut mendapatkan hasil 72 *cluster* dengan batas 5 kategori yaitu infrastruktur, pelayan, transportasi, fasilitas, dan topik lainnya. Setiap *cluster* memiliki anggota dengan kemiripan bobot tweet yang terdekat, sehingga didapat hasil data dan label tiap *cluster* seperti pada Tabel 2. 450 data yang sudah memiliki label dibagi menjadi 400 data *training* dan 50 data *testting* untuk melakukan klasifikasi dengan algoritma *Naive Bayes Classifier* (NBC). Untuk setiap *term* dari data testing dilakukan perhitungan probabilitas dengan menggunakan algoritma NBC terhadap setiap kategori kelas yang ada.

Tabel 2. Jumlah Data yang Masuk *Cluster*

Label	Jumlah data
C1 (infrastruktur)	124
C2 (Pelayanan)	52
C3 (Transportasi)	69
C4 (Fasilitas)	13
C5 (Lainnya)	192
Total	**450**

Berikutnya dengan mencari nilai probabilitas terbesar di antara semua hasil yang telah dihitung, maka sudah dapat dilihat data *testing* tersebut masuk ke dalam kelas yang mana. Contoh klasifikasi data dapat dilihat pada Tabel 3.

Tabel 3. Contoh Hasil Klasifikasi

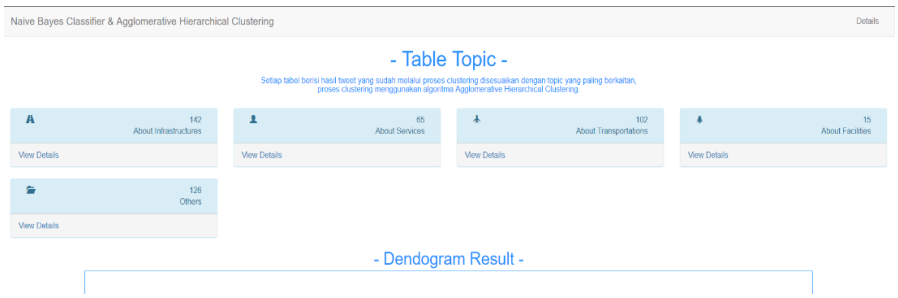
Kategori	Probabilitas tiap Kategori
Infrastruktur	9.84074547912623
Pelayanan	3.112222773250325
Transportasi	8.468124925284326
Fasilitas	7.658335823640726
Lainnya	7.395849657518828
Classifying =>@ganjarpranowo Jalannya sragen memprihatinkan nggih bopo...?? Tolong dibikin seperti jalan jalan didaerah purwokerto https://t.co/eAwaWwzwL0'	
HASIL = Infrastruktur	

3.2 Implementasi Program

Dalam penelitian ini program yang dihasilkan adalah dalam bentuk website yang digunakan sebagai media untuk melakukan proses perhitungan dan pengolahan data, mulai dari proses *preprocessing*, *TF-IDF, clustering Agglomerative Hierarchical clustering* hingga klasifikasi menggunakan algoritma *Naïve Bayes Classifier*. Semua proses tersebut dilakukan secara otomatis dengan tampilan yang *user friendly*. Data yang sudah masuk *cluster* sesuai topik digunakan untuk memudahkan dalam mendaptkan infomasi keluhan tentang fasilitas umum yang ada di Jawa Tengah.

a. Halaman menu

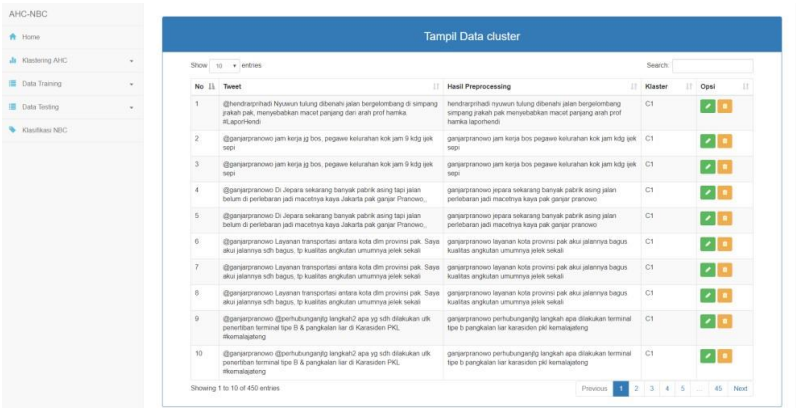
Pada halaman menu terdapat menu-menu topik yang sudah terkategori sesuai dengan masing-masing topik pembahasan .



Gambar 4 Halaman menu utama

b. Halaman details

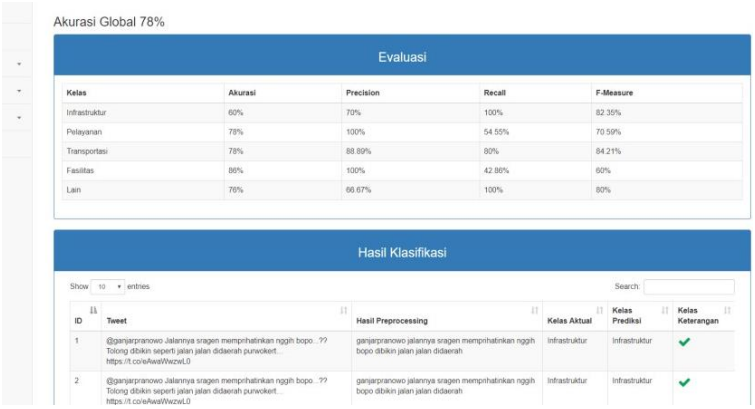
Halaman ini memiliki menu *side-bar* yang berisi semua sistem yang ada pada website dari menginputkan data *tweet*, melakuakn perhtungan *AHC*, menambahkan data *training*, data *testing*, dan melakukan uji akurasi.



Gambar 5 Halaman details

c. Halaman akurasi

Pada menu halaman klasifikasi NBC berisi hasil perhitungan metode *Naïve Bayes Classifier* dengan perhitungan akurasi menggunakan *conffusion matrix.*



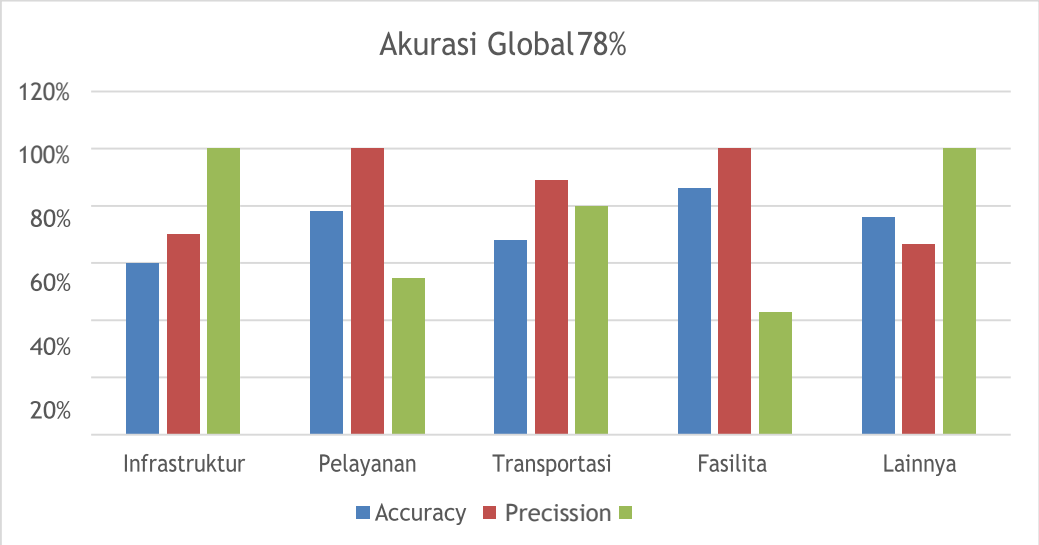
Gambar 6 Halaman akurasi

3.3 Pengujian Hasil

Pengujian hasil klasifikasi menggunakan *confusion matrix* untuk mencari nilai akurasi, *precission*, dan *recall*. Dari penelitian yang telah dilakukan, dihasilkan nilai akurasi global sebesar 78% dengan rincian sesuai Tabel 4 dan Gambar 7.

Tabel 4. Hasil akurasi

Jumlah Kesamaan Tiap kelas	
Kelas	Nilai
Infrastruktur	60%
Pelayanan	78%
Transportasi	68%
Fasilitas	86%
Lainnya	76%
Nilai Precission Tiap Kelas	
Kelas	Nilai
Infrastruktur	70%
Pelayan	100%
Transportasi	88.89%
Fasilitas	100%
Lainnya	66.67%
Nilai Recall Tiap Kelas	
Kelas	Nilai
Infrastruktur	100%
Pelayan	54.55%
Transportasi	80%
Fasilitas	42.86%
Lainnya	100%



Gambar 7 Diagram akurasi

4. KESIMPULAN

Berdasarkan hasil penelitian tentang implementasi algoritma *Agglomerative Hierarchical Clustering* dan *Naive Bayes Classifier* untuk klasifikasi *tweet* tentang fasilitas umum di Jawa Tengah dapat ditarik sebuah kesimpulan :

1. Hasil *crawling* data dari Twitter akun resmi gubernur Jawa Tengah @ganjarpranowo dengan kata kunci seperti infrastruktur, fasilitas umum, pelayanan, dll telah berhasil dilakukan *clustering* menggunakan *Agglomerative Hierarchical Clustering* dikelompokkan data dengan kelas infrastruktur, pelayanan, transportasi, fasilitas, dan topik lainnya.
2. Dari jumlah 450 *tweet* yang sudah dibagi menjadi 400 data training dengan dan 50 data tersting, selanjutnya berhasil dilakukan proses klasifikasi dengan *Naive Bayes Classifier.*
3. Telah dilakukan perhitungan akurasi dengan *conffusion matrix* terhadap metode klasifikasi menggunakan algoritma *Agglomerative hierarchical clustering* dan *Naive bayes classifier* dan mendapatkan hasil akurasi global sebesar 78%. Hal ini menunjukkan bahwa implementasi algoritma Agglomerative Hierarchical Clustering dan Naïve Bayes Classifier cukup akurat unutk melakukan klasifikasi pada tweets seputar fasilitas umum di Jawa Tengah.

DAFTAR RUJUKAN

Bharti, S. K., Pradhan, R., Babu, K. S., & Jena, S. K. (2016). Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks Volume 2, Issue 3*, 108-121.

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient Agglomerative Hierarchical Clustering. *Expert Systems with Applications*, 2785-2797.

BPS. (2020). *Provinsi Jawa Tengah Dalam Angka 2020.* Badan Pusat Statistik Provinsi Jawa Tengah.

Carley, K. M., Malik, M., Kowalchuk, M., Pfeffer, J., & Landwehr, P. (2015). *Twitter Usage in Indonesia.* Pittsburgh: Center for the Computational Analysis of Social and Organizational Systems.

Ibrahim, M. N., & Yusoff, M. Z. (2015). Twitter Sentiment Classification Using Naive Bayes Based on Trainer Perception. *IEEE Conference on e-Learning, e-Management and e-Services (IC3e)* (hal. 187-189). Melaka: IEEE.

Jati, A. S. (2020, Mei 04). *Jumlah Pengguna Twitter Meningkat, Tapi...* Retrieved from detikInet: https://inet.detik.com/cyberlife/d-5001786/jumlah-pengguna-twitter-meningkat-tapi

Unnisa, M., Ameen, A., & Raziuddin, S. (2016). Opinion Mining on Twitter Data using Unsupervised Learning Technique. *International Journal of Computer Applications*, 12-19.

Wibawa, A. P., Purnama, M. G., Akbar, M. F., & Dwiyanto, F. A. (2018). Metode-metode Klasifikasi. *Seminar Ilmu Komputer dan Teknologi Informasi* (pp. 134-138).