KOMPARASI ALGORITMA *K-MEANS* DAN *AGGLOMERATIVE HIERARCHICAL* UNTUK KLASTERISASI KALIMAT DALAM NASKAH PIDATO

SKRIPSI

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:
DEAR RESISTA
NIM D1041201062

FAKULTAS TEKNIK
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Dear Resista

NIM : D1041201062

menyatakan bahwa dalam skripsi yang berjudul "Komparasi Algoritma *K-Means* dan *Agglomerative Hierarchical* untuk Klasterisasi Kalimat dalam Naskah Pidato" tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi manapun. Sepanjang pengetahuan Saya, tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Demikian pernyataan ini dibuat dengan sebenar-benarnya. Saya sanggup menerima konsekuensi akademis dan hukum di kemudian hari apabila pernyataan yang dibuat ini tidak benar.

Pontianak, 24 Juli 2025



Dear Resista
NIM D1041201062

KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS TANJUNGPURA
FAKULTAS TEKNIK
Jalan. Prof. Dr. H. Hadari Nawawi Pontianak 78124
Telepon (0561) 740186, WA: +6282152280907
Email : ft@untan.ac.id Website: http://teknik.untan.ac.id

HALAMAN PENGESAHAN

KOMPARASI ALGORITMA *K-MEANS* DAN *AGGLOMERATIVE HIERARCHICAL* UNTUK KLASTERISASI KALIMAT DALAM NASKAH PIDATO

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:

Dear Resista
NIM D1041201062

Telah dipertahankan di depan Penguji Skripsi pada tanggal 24 Juli 2025
dan diterima sebagai salah satu persyaratan untuk memperoleh gelar sarjana.

Susunan Penguji Skripsi:

Ketua,

Prof. Dr. Herry Sujaini, S.T., M.T.
NIP 196806291997021001

Penguji Utama,

Dr. Arif Bijaksana Putra Negara, S.T., M.T.
NIP 197208081998021002

Sekretaris,

Niken Candraningrum, S.T., M.Cs.
NIP 199309272022032010

Penguji Pendamping,

Helen Sastypratiwi, S.T., M.Eng.
NIP 198601172012122004

Pontianak, 24 Juli 2025
Dekan,

Dr. Ing. Ir. Slamet Widodo, M.T., IPM
NIP 196712231992031002

Dipindai dengan CamScanner

HALAMAN PERSEMBAHAN

Dengan segala rasa syukur dan hormat, karya ini dipersembahkan untuk mereka yang selalu menjadi sumber kekuatan dan inspirasi di setiap langkah. Kepada diri sendiri, terima kasih telah berani memulai dan menjalani setiap proses dengan penuh keteguhan. Tidak mudah untuk sampai di titik ini, namun setiap tantangan berhasil dilewati dengan doa, usaha, dan semangat.

Kepada kedua orang tua tercinta, bapak dan mamak terima kasih atas doa-doa yang tiada henti, kasih sayang yang tulus, dan dukungan yang tanpa batas. Skripsi ini adalah bentuk kecil dari rasa hormat untuk segala pengorbanan yang telah diberikan. Semoga hasil ini menjadi awal dari masa depan yang lebih baik.

Terima kasih juga kepada keluarga di rumah, yang selalu menjadi tempat pulang terbaik dan sumber penyemangat terbesar. Kalian adalah rumah, dengan cinta, dukungan, dan kehangatan yang selalu menguatkan di setiap langkah.

Kepada Bapak Haried Novriando, S.T., M.Eng., terima kasih atas segala arahan, ilmu, dan perhatian selama masa perkuliahan. Bimbingan dari bapak telah memberikan pengaruh besar dalam perjalanan akademik saya.

Rasa hormat dan terima kasih sebesar-besarnya saya sampaikan kepada Bapak Prof. Dr. Herry Sujaini, S.T., M.T., dan Ibu Niken Candraningrum, S.T., M.Cs., sebagai pembimbing tugas akhir. Terima kasih atas ilmu, nasihat, dan bimbingan selama proses penyelesaian tugas akhir ini. Mohon maaf atas kekhilafan saya selama masa bimbingan dan semoga apa yang telah diberikan menjadi bekal yang berharga untuk masa depan.

Tidak lupa, terima kasih untuk teman-teman Informatika 20 dan teman saya Jilan, atas dukungan, canda tawa, dan kebersamaan selama perjalanan ini.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT atas limpahan rahmat dan hidayah-Nya. Sholawat serta salam senantiasa tercurah kepada Nabi Muhammad SAW, sehingga penulis dapat menyelesaikan skripsi dengan judul "Komparasi Algoritma *K-Means* dan *Agglomerative Hierarchical* untuk Klasterisasi Kalimat dalam Naskah Pidato". Skripsi ini disusun untuk memenuhi salah satu syarat memperoleh gelar Sarjana (S1) pada Fakultas Teknik, Program Studi Informatika, Universitas Tanjungpura.

Penulis menyadari bahwa terselesaikannya skripsi ini tidak terlepas dari bantuan, dukungan, dan doa dari banyak pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih yang tulus kepada semua yang telah berperan serta dalam memberikan bantuan baik moral maupun materi. Secara khusus, penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada Bapak Prof. Dr. Herry Sujaini, S.T., M.T. selaku dosen Pembimbing I dan Ibu Niken Candraningrum, S.T., M.Cs. selaku dosen Pembimbing II, serta kepada Bapak Dr. Arif Bijaksana Putra Negara, S.T., M.T. selaku dosen Penguji I dan Ibu Helen Sastypratiwi, S.T., M.Eng. selaku dosen Penguji II.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna, untuk itu kritik dan saran yang konstruktif sangat di harapkan demi perbaikan kedepannya. Semoga skripsi ini dapat memberikan manfaat serta menjadi referensi yang berguna dalam pengembangan ilmu, khususnya dalam bidang klasterisasi gaya bahasa di naskah pidato.

Pontianak, Juli 2025
Penulis,

Dear Resista

ABSTRAK

Penyampaian pesan dalam pidato bergantung pada susunan kalimat yang efektif dan bervariasi. Namun, keberagaman struktur kalimat dalam naskah pidato membuat pengelompokan secara manual menjadi kurang efisien. Dengan menggunakan teknik pemrosesan bahasa alami, proses ini dapat dilakukan secara sistematis melalui pendekatan klasterisasi. Penelitian ini bertujuan untuk membandingkan performa algoritma *K-Means* dan *Agglomerative Hierarchical* dalam klasterisasi kalimat pada naskah pidato. Data dikumpulkan dari situs Korpus Nusantara dan Jago Berpidato. Tahapan *pre-processing* mencakup *data cleaning, Case Folding, Stopword Removal*, tokenisasi, ekstraksi fitur menggunakan TF-IDF dengan parameter N-Gram, serta reduksi dimensi dengan PCA. Evaluasi dilakukan menggunakan metrik *Davies-Bouldin Index* (DBI) dan *Silhouette Score*. Hasil pengujian menunjukkan bahwa algoritma *Agglomerative Hierarchical* dengan metode *average linkage* menghasilkan klasterisasi yang lebih optimal dengan nilai DBI sebesar 0.1419 dan *Silhouette Score* sebesar 0.7742. Sebagai perbandingan, algoritma *K-Means* menghasilkan DBI sebesar 0.8551 dan *Silhouette Score* sebesar 0.5557. Selain itu, penggunaan parameter n-gram juga berpengaruh terhadap kualitas klasterisasi. Representasi trigram memberikan hasil terbaik khususnya pada *Agglomerative Hierarchical* dibandingkan dengan unigram (DBI/*Silhouette Score:* 0.6934/0.3002) dan bigram (DBI/*Silhouette Score:* 0.2192/0.6625). Pada penelitian ini, kombinasi *Agglomerative Hierarchical* dengan representasi fitur trigram terbukti paling efektif dalam mengelompokkan kalimat berdasarkan kemiripan gaya bahasa dalam naskah pidato.

Kata kunci: Klasterisasi, N-Gram, *K-Means*, *Agglomerative Hierarchical, Davies-Bouldin Index, Silhouette Score*

ABSTRACT

The effectiveness of speech communication depends on the use of varied and well-structured sentences. However, the diversity of sentence structures in speech transcripts makes manual grouping inefficient. By applying natural language processing techniques, this process can be conducted systematically using a clustering approach. This study aims to compare the performance of the K-Means and Agglomerative Hierarchical algorithms in clustering sentences from speech transcripts. The data were collected from the Korpus Nusantara and Jago Berpidato websites. The pre-processing stages included data cleaning, Case Folding, Stopword Removal, tokenization, feature extraction using TF-IDF with N-Gram parameters, and dimensionality reduction using PCA. The evaluation was carried out using the Davies-Bouldin Index (DBI) and Silhouette Score metrics. The test results show that the Agglomerative Hierarchical algorithm with the average linkage method produces more optimal clustering, with a DBI value of 0.1419 and a Silhouette Score of 0.7742. In comparison, the K-Means algorithm resulted in a DBI of 0.8551 and a Silhouette Score of 0.5557. In addition, the use of n-gram parameters also affects the quality of clustering. The trigram representation yielded the best results, particularly for Agglomerative Hierarchical, compared to unigram (DBI/Silhouette Score: 0.6934/0.3002) and bigram (DBI/Silhouette Score: 0.2192/0.6625). In this study, the combination of Agglomerative Hierarchical and trigram feature representation proved to be the most effective in clustering sentences based on similarities in language style within speech transcripts.

Keywords: Clustering, N-Gram, K-Means, Agglomerative Hierarchical, Davies-Bouldin Index (DBI), Silhouette Score

DAFTAR ISI

Halaman Pernyataan ii
Halaman Pengesahan iii
Halaman Persembahan iv
Kata Pengantar v
Abstrak vi
Abstract vii
Daftar Isi viii
Daftar Tabel x
Daftar Gambar xi
Kode Program xii

Bab I Pendahuluan 1

1.1 Latar Belakang 1
1.2 Perumusan Masalah 3
1.3 Tujuan Penelitian 3
1.4 Pembatasan Masalah 3
1.5 Sistematika Penulisan 3

Bab II Tinjauan Pustaka 5

2.1 Penelitian Terkait 5
2.2 Data Mining 6
2.3 Klasterisasi 7
2.4 Ekstraksi Fitur 8
2.4.1 TF-IDF (Term Frequency-Inverse Document Frequency) 9
2.4.2 Penggunaan N-Gram dalam TF-IDF 10
2.5 PCA (Principal Component Analysis) 10
2.6 Algoritma *K-Means* 11
2.7 Algoritma Agglomerative Hierarchical 12
2.8 Metrik Evaluasi 14
2.8.1 Davies-Bouldin Index (DBI) 14
2.8.2 Silhouette Score 15

Bab III Metodologi Penelitian 16

3.1 Diagram Alir Penelitian 16
3.1.1 Pengumpulan Data 18
3.1.2 Pre-Processing Data 18
3.1.2.1 *Data Cleaning* 18
3.1.2.2 *Case Folding* 18
3.1.2.3 *Stopword Removal* 19
3.1.2.4 *Tokenizing* 19
3.1.3 Ekstraksi Fitur (TF-IDF) 20

3.1.4 Reduksi Dimensi (PCA) 28
3.1.5 Implementasi Algoritma 33
3.1.5.1 Algoritma *K-Means* 33
3.1.5.2 Algoritma *Agglomerative Hierarchical* 36
3.1.6 Evaluasi Hasil Klasterisasi 39
3.1.7 Analisis Hasil dan Perbandingan Algoritma 40

Bab IV Hasil dan Analisis 41

4.1 Hasil Pengumpulan Data 41
4.2 Hasil Pre-Processing Data 42
4.3 Hasil Ekstraksi Fitur (TF-IDF) 47
4.4 Hasil Reduksi Dimensi (PCA) 49
4.5 Hasil Implementasi Algoritma 50
4.5.1 K-Means 51
4.5.2 Agglomerative Hierarchical 54
4.6 Evaluasi Hasil Klasterisasi 59
4.7 Analisis Hasil dan Perbandingan Algoritma 61
4.7.1 Analisis Hasil 61
4.7.2 Perbandingan Algoritma 71

Bab V Kesimpulan dan Saran 75

5.1 Kesimpulan 75
5.2 Saran 75

Daftar Pustaka 77

DAFTAR TABEL

Tabel 3. 1 Data Cleaning 18
Tabel 3. 2 Case Folding 19
Tabel 3. 3 Stopword Removal 19
Tabel 3. 4 Tokenizing 19
Tabel 3. 5 Term Frequency (TF) Unigram 21
Tabel 3. 6 Inverse Document Frequency (IDF) 21
Tabel 3. 7 TF-IDF (Unigram) 22
Tabel 3. 8 Term Frequency (TF) Bigram 23
Tabel 3. 9 Inverse Document Frequency (IDF) Bigram 24
Tabel 3. 10 TF-IDF (Bigram) 25
Tabel 3. 11 Term Frequency (TF) Trigram 25
Tabel 3. 12 Inverse Document Frequency (IDF) Trigram 26
Tabel 3. 13 TF-IDF (Trigram) 27
Tabel 3. 14 Standarisasi Data 29
Tabel 3. 15 Perhitungan Matriks Kovarians 30
Tabel 3. 16 Fitur Kalimat 34
Tabel 3. 17 Perhitungan Centroid 35
Tabel 3. 18 Iterasi Ulang (iterasi kedua) 36
Tabel 3. 19 Fitur Kalimat 38
Tabel 3. 20 Perhitungan Jarak 38
Tabel 3. 21 Matriks jarak awal 38
Tabel 3. 22 Jarak antar klaster baru 39
Tabel 4. 1 Data Kalimat 42
Tabel 4. 2 Sebelum Pre-Processing Data 43
Tabel 4. 3 Hasil Data Cleaning 44
Tabel 4. 4 Hasil Case Folding 44
Tabel 4. 5 Hasil Stopword Removal 45
Tabel 4. 6 Hasil Tokenizing 46
Tabel 4. 7 Representasi TF-IDF 47
Tabel 4. 8 Kata dengan Skor TF-IDF tertinggi 48
Tabel 4. 9 Bigram dengan Skor TF-IDF tertinggi 48
Tabel 4. 10 Trigram dengan Skor TF-IDF tertinggi 49
Tabel 4. 11 Data setelah PCA 50
Tabel 4. 12 Cophenetic Correlation Coefficient 55
Tabel 4. 13 Evaluasi Klasterisasi Berdasarkan Variasi N-Gram 60
Tabel 4. 14 Distribusi Klaster 66
Tabel 4. 15 Perbandingan Algoritma 72

DAFTAR GAMBAR

Gambar 3. 1 Diagram alir penelitian 17
Gambar 3. 2 Flowchart algoritma k-means 33
Gambar 3. 3 Flowchart algoritma agglomerative hierarchical 37
Gambar 4. 1 Korpus Nusantara (kortara) 41
Gambar 4. 2 Jago berpidato 41
Gambar 4. 3 Dbi unigram (k-means) 52
Gambar 4. 4 Dbi bigram (k-means) 53
Gambar 4. 5 Dbi trigram (k-means) 53
Gambar 4. 6 Dbi unigram (agglomerative hierarchical) 57
Gambar 4. 7 Dbi bigram (agglomerative hierarchical) 58
Gambar 4. 8 Dbi trigram (agglomerative hierarchical) 58
Gambar 4. 9 Scatter plot unigram 61
Gambar 4. 10 Scatter plot bigram 62
Gambar 4. 11 Scatter plot trigram 63
Gambar 4. 12 Dendogram unigram 64
Gambar 4. 13 Dendogram bigram 64
Gambar 4. 14 Dendogram Trigram 65
Gambar 4. 15 Word cloud klaster 1 67
Gambar 4. 16 Bar chart variasi n-gram 71

KODE PROGRAM

Kode Program 4. 1 Read Data..... 42
Kode Program 4. 2 Data Cleaning 43
Kode Program 4. 3 Case Folding 44
Kode Program 4. 4 Stopword Removal 45
Kode Program 4. 5 Tokenizing 46
Kode Program 4. 6 TF-IDF 47
Kode Program 4. 7 Reduksi Dimensi (PCA) 49
Kode Program 4. 8 Implementasi Algoritma K-Means 51
Kode Program 4. 9 Cophenetic Correlation Coefficient (CCC) 54
Kode Program 4. 10 Implementasi Agglomerative Hierarchical..... 55

BAB I
PENDAHULUAN

1.1 Latar Belakang

Di era digital saat ini, perkembangan teknologi informasi telah membawa perubahan signifikan dalam berbagai bidang, termasuk dalam pengolahan data tekstual. Pendekatan komputasi seperti *Natural Language Processing* (NLP) memungkinkan mesin untuk memproses dan memahami teks secara efisien. Salah satu bentuk teks yang memiliki nilai penting dalam komunikasi formal adalah naskah pidato. Pidato tidak hanya menyampaikan informasi, tetapi juga membangun hubungan emosional dan mempengaruhi pendengar. Dalam naskah pidato, kalimat disusun dengan tujuan tertentu, bisa bersifat persuasif, deskriptif, naratif, atau imperatif yang menjadikannya kompleks untuk dianalisis secara manual. Proses analisis kalimat dalam pidato secara manual membutuhkan waktu dan keahlian linguistik, sehingga diperlukan pendekatan otomatis yang dapat mengelompokkan kalimat berdasarkan kemiripan.

Salah satu pendekatan yang dapat digunakan untuk menangani kompleksitas tersebut adalah klasterisasi, yaitu teknik dalam *data mining* yang mengelompokkan data tanpa memerlukan label awal. Dalam konteks NLP, klasterisasi memungkinkan pengelompokan kalimat berdasarkan kemiripan struktur atau gaya bahasa secara otomatis. Kalimat sebagai unit linguistik terkecil menjadi fokus utama dalam pengolahan ini. Dengan menerapkan klasterisasi terhadap kalimat-kalimat dalam pidato, proses analisis dapat dilakukan lebih sistematis dan efisien, serta membuka peluang untuk memahami pola komunikasi formal secara lebih dalam melalui pendekatan berbasis data.

Penerapan klasterisasi telah digunakan dalam berbagai bidang untuk mengelompokkan data berdasarkan kemiripan karakteristik. Azzahra & Wijayanto, (2022) membandingkan algoritma *K-Means* dan *Agglomerative Hierarchical* dalam mengelompokkan provinsi berdasarkan cakupan pelayanan kesehatan maternal, dan menemukan bahwa *Agglomerative Hierarchical* dengan *average linkage* memberikan hasil evaluasi yang lebih baik. Sementara itu, Rozaq, (2022) menunjukkan bahwa *K-Means* unggul dari sisi efisiensi waktu saat digunakan untuk mengelompokkan data kebutuhan rumah. Kedua penelitian tersebut

menunjukkan bahwa efektivitas algoritma klasterisasi bergantung pada karakteristik data yang dianalisis. Oleh karena itu, penting untuk mengevaluasi performa masing-masing algoritma dalam konteks data tekstual seperti naskah pidato, yang memiliki struktur dan kompleksitas berbeda dibandingkan data numerik.

Algoritma yang sering digunakan dalam klasterisasi adalah *K-Means* dan *Agglomerative Hierarchical*. *K-Means* merupakan algoritma klasterisasi berbasis jarak yang membagi data ke dalam sejumlah klaster dan dalam prakteknya, algoritma ini umum digunakan karena relatif cepat dan mudah beradaptasi (Yahya, 2022). *Agglomerative Hierarchical* adalah metode pengelompokkan data yang menggabungkan dua buah klaster yang mempunyai kemiripan. Pada metode ini jumlah klaster tidak perlu ditentukan secara spesifik (Muttaqin, 2023). Algoritma *K-Means* dan *Agglomerative Hierarchical* memiliki konsep yang berbeda. *K-Means* merupakan metode non-hirarki yang bekerja dengan mengelompokkan data ke dalam sejumlah klaster yang telah ditentukan sebelumnya berdasarkan jarak *centroid*. Sementara itu *Agglomerative Hierarchical* bekerja secara hierarkis dengan menggabungkan klaster-klaster kecil menjadi klaster yang lebih besar, struktur hierarki dapat memberikan pemahaman terhadap hubungan antar gaya bahasa.

Penelitian ini bertujuan untuk membandingkan performa algoritma *K-Means* dan *Agglomerative Hierarchical* dalam mengklasterisasi kalimat dalam naskah pidato. Selain membandingkan algoritma klasterisasi, penelitian ini juga melakukan eksperimen terhadap parameter n-gram yang digunakan dalam proses ekstraksi fitur, untuk melihat pengaruhnya terhadap hasil klasterisasi kalimat dalam naskah pidato. Eksperimen ini dilakukan sebagai bagian dari upaya untuk memperoleh hasil klasterisasi yang optimal. Hasil dari penelitian ini diharapkan dapat memberikan gambaran mengenai perbandingan performa kedua algoritma dalam klasterisasi kalimat pidato, sehingga diketahui pendekatan mana yang lebih efektif berdasarkan kualitas hasil klaster yang terbentuk. Selain itu, penelitian juga memberikan gambaran mengenai tipe-tipe kalimat yang muncul dalam naskah pidato, sehingga dapat dimanfaatkan dalam analisis lebih lanjut di bidang linguistik maupun pengembangan teknologi bahasa berbasis teks.

1.2 Perumusan Masalah

Berdasarkan penjabaran latar belakang penelitian, maka membawa peneliti pada pertanyaan atau rumusan masalah berikut:

1. Bagaimana perbandingan performa algoritma *K-Means* dan *Agglomerative Hierarchical* dalam klasterisasi kalimat pada naskah pidato berdasarkan metrik evaluasi?
2. Bagaimana pengaruh penggunaan parameter n-gram terhadap hasil klasterisasi?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk membandingkan performa algoritma *K-Means* dan *Agglomerative Hierarchical* dalam mengelompokkan kalimat dalam naskah pidato menggunakan metrik evaluasi. Penelitian juga bertujuan untuk menganalisis pengaruh penggunaan parameter n-gram terhadap hasil klasterisasi.

1.4 Pembatasan Masalah

Agar penulisan Proposal Skripsi ini lebih terarah dan tidak menyimpang dari permasalahan, maka penulis membuat batasan masalah sebagai berikut:

1. Penelitian akan menggunakan data Naskah Pidato dari situs web Korpus Nusantara dan Jago Berpidato
2. Representasi teks yang dilakukan menggunakan TF-IDF dengan parameter n-gram yaitu unigram, bigram, dan trigram.
3. Metrik evaluasi hasil klasterisasi dibatasi pada metrik *Davies-Bouldin Index* dan *Silhouette Score*, untuk mengukur kualitas pembagian klaster dan divisualisasikan menggunakan *scatter plot* dan dendogram.
4. Analisis hasil lebih lanjut dilakukan pada kombinasi algoritma dan representasi fitur dengan performa terbaik berdasarkan metrik evaluasi *Davies-Bouldin Index* dan *Silhouette Score.*

1.5 Sistematika Penulisan

Sistematika penulisan dalam laporan penelitian ini disusun dalam lima bab yang terdiri sebagai berikut:

1. BAB I PENDAHULUAN

Bab ini berisikan mengenai latar belakang, perumusan masalah, tujuan penelitian, pembatasan masalah, dan sistematika penulisan penelitian ini.