Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)
Vol. 2, No. 1 Februari 2023
e-ISSN: 2963-7813; p-ISSN: 2963-8178, Hal 73-89

Cluster Text Random Opinion Tweet In Yogyakarta Using Automatic Clustering

Rabiatul Adawiyah
Politeknik Pratama
Email: wilyanyatul@gmail.com

Abstract

Tweet Besides making computations difficult, the data obtained is also inefficient and complicated to interpret. Therefore, it is necessary to explore how to overcome these problems. This study proposes an approach to find the global optimum and make automatic grouping by analyzing moving averages, namely K-Means Automatic Clustering. So the purpose of this study was to explore and evaluate high-dimensional data from a collection of tweets, namely random opinion text tweets in Yogyakarta. The K-means Automatic Clustering algorithm is used for clusters based on the data attributes that have been obtained. Pre-processing experiments were carried out among others. Cleansing, Case folding, Tokenizing, Filtering, Stemming. Then look for the variance cluster to find the global optimum as an ideal cluster by identifying the moving variance by placing λ as the threshold (Global Optimum). So that the ideal cluster value is 0.332975. That is, the closer the cluster value obtained to number 1, the more the cluster search finds the optimum point. This research can be utilized in exploring and evaluating high-dimensional data, so that it becomes a consideration in providing approximate patterns from unstructured data sets with Visualization.

Keywords: *Cluster, Tweet, K-Means, Automatic Clustering, Data, Opinion*

Abstract

Twitter adalah platform digital yang memfasilitasi penggunanya untuk saling berkomunikasi berupa tulisan, foto dan video satu sama lain tanpa dibatasi ruang dan waktu, Hal itulah yang menjadikan twitter tempat penampungan terbesar data text yang masih belum diketahui polanya atau tidak terstruktur dengan variabel berdimensi tinggi. Disamping mengakibatkan komputasi sulit dilakukan, data yang didapatkan juga tidak efisien dan rumit untuk diinterpretasikan. Oleh karena itu, diperlukan eksplorasi bagaimana mengatasi permasalahan tersebut. Penelitian ini, mengusulkan pendekatan untuk menemukan global optimum dan membuat pengelompokan otomatis dengan menganalisis moving average yaitu dengan K-Means Automatic Clustering. Sehingga tujuan penelitian ini adalah mengeksplorasi dan mengevaluasi data berdimensi tinggi dari kumpulan tweet, yaitu text tweet random opinion di yogyakarta. Algoritma K-means Automatic Clustering digunakan untuk cluster berdasarkan atribut data yang telah didapatkan. Percobaan Pre-processing dilakukan diantaranya. Cleansing, Case folding, Tokenizing, Filtering, Stemming. Selanjutnya mencari variance cluster untuk menemukan global optimum sebagai klaster ideal dengan mengidentifikasi varian bergerak dengan menempatkan λ sebagai ambang batas (Global Optimum). Sehingga

Received Desember 27, 2022; Revised Januari 22, 2023; Februari 06, 2023
* Rabiatul Adawiyah, wilyanyatul@gmail.com

diperoleh nilai cluster ideal yaitu 0.332975. Artinya, semakin dekat nilai cluster yang diperoleh pada angka 1, maka menunjukkan pencarian cluster menemukan titik optimum. Penelitian ini dapat dimanfaatkan dalam melakukan eksplorasi dan mengevaluasi data yang berdimensi tinggi, sehingga menjadi pertimbangan dalam memberikan perkiraan pola dari kumpulan data yang tidak terstruktur dengan Visualisasi.

Kata kunci : Cluster, Tweet, K-Means, Automatic Clustering, Data, Opinion

INTRODUCTION

Media sosial atau sering juga disebut sebagai sosial media adalah platform digital yang memfasilitasi penggunanya untuk saling berkomunikasi atau membagikan konten berupa tulisan, foto dan video, dengan menyediakan fasilitas bersosialisasi satu sama lain yang dilakukan secara daring, sehingga memungkinkan manusia untuk saling berinteraksi tanpa dibatasi ruang dan waktu. Contoh media sosial yang popular digunakan dalam beberapa dekade salah satunya adalah Twitter [1]. Twitter menjadi salah satu tempat orang untuk meluapkan seluruh pikiran dan perasaannya melalui tulisan bahkan sampai dijadikan pengganti buku diary [2]. Hal itulah yang menjadikan twitter tempat penampungan terbesar data text yang masih belum diketahui polanya atau tidak terstruktur.

Text mining merupakan area penelitian yang termasuk dalam lingkup text analytics, dimana fokus utamanya adalah menemukan pengetahuan baru dan berguna dari sumber data tekstual. Implementasi text mining menggunakan text from berita online dalam mengetahui pola dan memprediksi [3]. Text mining sendiri dapat didefinisikan sebagai proses semi otomatis menggunakan komputer yang digunakan untuk mengekstraksi pola dari kumpulan data tidak terstruktur yang sangat besar. Sebuah data random opini tweet khususnya di daerah Yogyakarta dapat dimanfaatkan untuk diketahui pola datanya dengan metode Text clustering yang merupakan bagian dari text mining, algoritma ini menerapkan Unsupervised Learning untuk mengelompokkan data tekstual (tweet) ke dalam cluster yang memiliki karakteristik yang sama [4]. Hasil pengelompokan tweet diharapkan dapat mengetahui pola informasi dari data yang ada di tweet khusunya di Yogyakarta, Indonesia.

Namun, Sering terjadi masalah pada proses Text Clustering dimana banyaknya data tekstual yang tersedia biasanya berukuran sangat besar (big data) dan memiliki variabel berdimensi tinggi yang mengakibatkan komputasi sulit dilakukan sehingga hasilnya tidak efisien dan rumit untuk diinterpretasikan. Oleh karena itu, diperlukan eksplorasi bagaimana mengatasi permasalahan tersebut dari sudut pandang statistik [5][6].

Metode yang umum digunakan dan memiliki performa yang baik di area Text Clustering untuk mengelompokkan tweet adalah K-Means. Dimana tweet diubah menjadi bentuk numerik atau model ruang vector yang nantinya akan membentuk document-term-matrix (DTM) yang memuat bobot setiap kata pada setiap tweet. DTM tersebut menjadi masukan bagi algoritma K-Means yang akan mengelompokkan teks dengan menghitung jarak dari vektor tweet ke centroid (titik tengah) seluruh cluster dan menempatkan vektor tweet tersebut ke dalam cluster terdekat. Beberapa penelitian yang berfokus pada analisis data Twitter telah melakukan Text Clustering menggunakan algoritma K-Means untuk mendapatkan cluster atau grup yang terdapat dalam kumpulan tweet pada berbagai topik [4] [7].

Namun, Tugas menemukan cluster yang baik adalah masalah yang sangat kritis dalam clustering. biasanya peneliti mencobanya dengan jumlah cluster yang berbeda. Tapi hal ini membuat sangat sulit, terutama jika kasus pengelompokan tidak mudah diamati seperti data yang saat ini digunakan yaitu random opini. Algoritma genetika diusulkan untuk mencari cluster yang optimal [8]. Sehingga, tetap mengharuskan peneliti untuk memberikan jumlah cluster secara apriori. Algoritma Automatic Clustering akan secara otomatis mencari jumlah klaster yang tepat dan mengklasifikasikan objek ke dalam klaster ini secara bersamaan. Dalam penelitian ini, mengusulkan pendekatan yang lebih baik untuk menemukan global optimum dan membuat pengelompokan otomatis dengan menganalisis moving average yaitu dengan K-Means Automatic Clustering [9] Berdasarkan penjelasan sebelumnya, tujuan dari penelitian ini adalah untuk mengeksplorasi dan mengevaluasi masalah data berdimensi tinggi dari kumpulan tweet, yaitu text tweet random opinion di yogyakarta dengan K-Means Automatic clustering.

RELATED WORK

Barakbah_IES_2004 [9], Penelitian ini menjelaskan gambaran bagaimana menemukan global optimum pada cluster, yaitu dengan menganalisis moving varian cluster untuk setiap tahap konstruksi cluster, kemudian mengamati pola untuk menemukan global optimum serta menghindari terjadinya optimum lokal. Pada penelitian ini memperkenalkan dua batasan, yaitu valley-tracing dan hill-climbing, untuk menemukan global optimum. Selain itu penelitian ini juga menganalisis kemungkinan untuk membuat automatic clustering. Hasil percobaan menunjukkan pendekatan yang efektif.

M. F. Tyas, A. Kurnia, and A. M. Soleh [4], Makalah ini bertujuan untuk menguji apakah metode sampling untuk mengelompokkan tweet dapat menghasilkan pengelompokan yang efisien dengan menggunakan seluruh dataset. Setelah pra-pemrosesan, terdapat lima ukuran sampel dipilih dari 28300 tweet yaitu 250, 500, 2500, 10000 dan 20000 untuk melakukan pengelompokan K-Means. Hasil penelitian ini menunjukkan bahwa dari 10 iterasi, tiga topik klaster utama muncul 90%-100% dengan ukuran sampel 2500, 10000 dan 20000. Sedangkan ukuran sampel 250 dan 500 cenderung menghasilkan 20%-60% kemunculan dari tiga topik klaster utama. Ini berarti bahwa sekitar 8% hingga 35% dari tweet yang digunakan dapat menghasilkan cluster yang representatif dan perhitungan yang efisien yang empat kali lebih cepat daripada menggunakan seluruh kumpulan data.

Alfian M, Ridho Barakbah A, Winarno [10], Penelitian ini mengusulkan metode Evolving Clustering time series mengadaptasi pengetahuan model yang ada di lingkungan nyata yang terus berkembang tanpa mengelompokkan ulang data. Penelitian ini juga mengusulkan ekstraksi fitur dengan fitur stemming berbasis ruang vektor untuk meningkatkan stemming bahasa Indonesia. Penerapan sistem terdiri dari tujuh tahap, (1) Akuisisi Data, (2) Pipeline Data, (3) Ekstraksi Fitur Kata Kunci, (4) Agregasi Data, (5) Predefined Cluster menggunakan algoritma Automatic Clustering, (6) Evolving Clustering , dan (7) Hasil Clustering Berita. Hasil eksperimen menunjukkan bahwa Automatic Clustering menghasilkan 388 cluster sebagai predefined cluster dari 3.000 berita. Salah satunya adalah cluster yang tidak diketahui. Pengembangan pengelompokan berlangsung selama dua hari untuk mengelompokkan berita melalui streaming, menghasilkan total 611 kelompok. Pengelompokan yang berkembang berjalan dengan baik, baik memperbarui model maupun

menambah model. Sehingga hasil dari penelitian ini yaitu Performansi algoritma Evolving Clustering cukup baik, dibuktikan dengan nilai akurasi cluster sebesar 88%.

ORIGINALITY

Pada bagian ini, originality dari pada penelitian sebelumnya yaitu Data yang digunakan adalah data random opinion twitter di yogyakarta, Indonesia. Data memiliki 5 atribut atau fitur yang diantaranya mencangkup twitter_id, name, created_at, followers_count dan text. Data ini adalah hasil dari crawling data twitter menggunakan API (Application Programming Interface) di Google Colaboratory dengan dibatasi pada data tweet yogyakarta. Periode pengambilan data dimulai dari tanggal 13 Januari 2023 dengan kata kunci Yogyakarta. Dalam penelitian ini mengusulkan pendekatan untuk menemukan global optimum dan membuat pengelompokan otomatis dengan menganalisis moving average yaitu dengan K-Means Automatic Clustering. Sehingga tujuan dari penelitian ini adalah untuk mengeksplorasi dan mengevaluasi masalah data berdimensi tinggi dari kumpulan tweet, yaitu text tweet random opinion di yogyakarta dengan Automatic clustering. Secara rinci dapat disajikan pada Gambar 1.

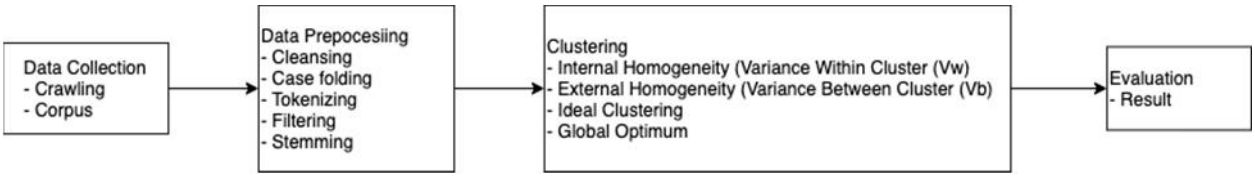
	twitter_id	name	created_at	followers_count	text
0	1613750000000000000	pukiskejumanies	01/13/23	159	b"'@TXTjajan wvg lullaby ver yeonjun 230k dp 1...
1	1613750000000000000	divin_com	01/13/23	6	b"'@Outstandjing As a Jogjanese, sepakat overr...
2	1613750000000000000	christian_swjy	01/13/23	110	b"'@ProfWeinstein Adhitia Sofyan - Sesuatu di ...
3	1613750000000000000	Nanaaa16293860	01/13/23	523	b"'Rdy BO jogja cash in room\xf0\x9f\x92\xa6 #...
4	1613750000000000000	Tiens_Jogja	01/13/23	283	b"'Jual Herbal Patah Tulang di Jogja, WA: 0896...
...	...	...	...	...	...
478	1613760000000000000	sejutarasasa	01/13/23	203	b"'@schfess jelas bali lah. sbg warga jogja, w...
479	1613760000000000000	ailulice	01/13/23	232	b'\'RT @fizylya: Si paling "dalem dek" pulkam ...
480	1613760000000000000	sejutarasasa	01/13/23	203	b"'ng gumuk pasir muk gosong km nder'"
481	1613760000000000000	venus_znx	01/13/23	3075	b"'RT @Viooktaviaa8: Hari ini aku masih availj...
482	1613760000000000000	rachajulian_	01/13/23	962	b"'RT @txtdarinonktpYK: Saya mengusulkan ini u...

483 rows × 5 columns

Gambar 1. Data Preparation

SYSTEM DESIGN

Sehingga solusi yang kami sajikan pada penelitian Cluster Text Random Opinion Tweet In Yogyakarta Using Automatic Clustering terdiri dari beberapa tahapan yaitu Data Collection, Data Preprocessing, Featrure Extraxtion, Clustering dan Evaluasi. Secara rinci dapat disajikan pada Gambar 2.



Gambar 2. Tahapan Cluster Text Random Opinion Tweet In Yogyakarta Using Automatic Clustering

Tahapan-tahapan ini diperlukan untuk mengestrak pola sehingga menghasilkan kualitas data yang lebih baik dan bagus dari kumpulan data yang tidak terstruktur, sehingga menghasilkan informasi penting yang terdapat pada data. Tahapan secara detail akan dijelaskan pada point selanjutnya.

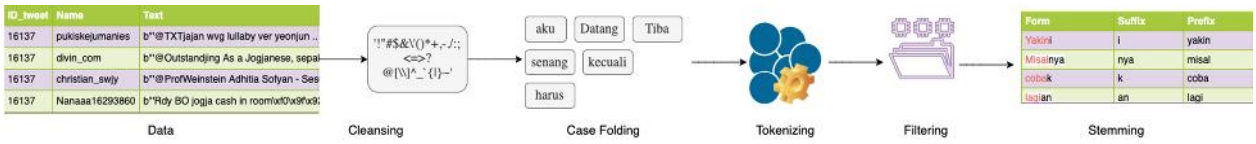
1.1 Data Collection

Dari gambaran besar desain sistem, proses dari penelitian ini dibagi dalam empat bagian inti yang antara lain, Tahapan Data crawling atau Web crawling adalah proses dalam mencari dan memindai data yang berada di halaman website. Di mana data dapat berupa text, artikel, gambar, video, ataupun dokumen. Cara kerja web crawler, proses akan dilakukan berdasarkan daftar link halaman yang sudah ada dalam suatu program dan sudah dipindai sebelumnya dari sitemap website dengan mempertimbangkan beberapa hal mengenai link yang akan di crawing. Salah satu pertimbangannya ialah seberapa penting dan relevannya sebuah halaman website serta versi terbaru. Pada penelitian ini data yang diperoleh dari proses crawling akan disimpan dalam bentuk file CSV (Comma Separated Values), tahapan selanjutnyam yaitu Data akan dilakukan Pre-Processing.

Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)
Vol. 2, No. 1 Februari 2023
e-ISSN: 2963-7813; p-ISSN: 2963-8178, Hal 58-72

1.2 Data Pre-Processing

Pre-processing adalah suatu manipulasi atau proses menyeleksi data agar lebih terstruktur dengan melalui serangkaian tahapan sebelum dilakukan input kedalam model dengan tujuan agar kompatibel dengan library yang digunakan. Ada beberapa tahapan yang dilalui diantaranya.



Gambar 2. Tahapan Data Pre-Processing

a. Cleansing, yaitu tahapan untuk melakukan pembersihan pada keseluruhan data yang terdapat pemberian tanda atau tanda baca seperti berikut ini '!"#$&\'()*+,-./:;<=>?@[\\]^_`{|}~' Sehingga menjadi lebih fokus pada isian data.
b. Case folding, yaitu tahapan untuk mengubah semua huruf menjadi huruf kecil (lowercase), contoh DaTA SCIence menjadi data science. Sementara itu, tulisan lain yang seperti tanda baca dan spasi dianggap delimenter. Delimenter ini bisa juga dihapus atau diabaikan. Dan pada penelitian ini untuk delimenter dihapus berdasarkan tahapan pertama yaitu cleansing.
c. Tokenizing, yaitu tahapan analisis dengan cara memecah kalimat-kalimat menjadi kata atau bisa disebut token. Dengan tokenizing, dapat membedakan mana antara pemisah kata atau bukan.
d. Filtering, yaitu tahapan untuk mengambil kata yang penting dari hasil tokenizing. Biasannya kata umum yang muncul dan tidak memiliki makna disebut stopword. Seperti, dan, yang, serta, setelah dan lainnya. Sehingga proses filtering pada stopword dapat mengurasi ukuran index dan waktu pemrosesan. Selain juga dapat mengurangi level noise.
e. Stemming, yaitu tahapan untuk memperkecil jumlah index yang berbeda dari suatu data sehingga sebuah kata yang memiliki suffix maupun prefix akan kembali ke bentuk dasarnya.

1.3 Clustering

Clustering merupakan sebuah teknik untuk membagi populasi atau titk-titik data sebuah nomer dari grup data seperti titik-titik data dalam grup yang sama lebih mirip dengan titik-titik data lain dalam grup dan tidak sama pada titik-titik data di grup-grup lain. Sedangkan cluster yang baik adalah ketika anggota cluster memiliki tingkat kesamaan yang tinggi satu sama lain (homogenitas internal) dan tidak seperti anggota cluster lain (homogenitas eksternal). Cluster ideal memiliki Vw minimum untuk menyatakan homogenitas internal dan Vb maksimum untuk menyatakan homogenitas eksternal. Tetapi, Menemukan klaster ideal sangat sulit karena kita tidak dapat menerapkan min(V) secara langsung untuk menemukan global optimum sebagai klaster ideal, untuk menemukan optimal global sebagai klaster ideal dengan mengidentifikasi varian bergerak dengan menempatkan λ sebagai ambang batas (Global Optimum).

a. Internal homogeneity (Variance within cluster (Vw))
b. External homogeneity (Variance betweet cluster Vb))
c. Ideal Cluster
d. Global Optimum

1.4 Evaluation

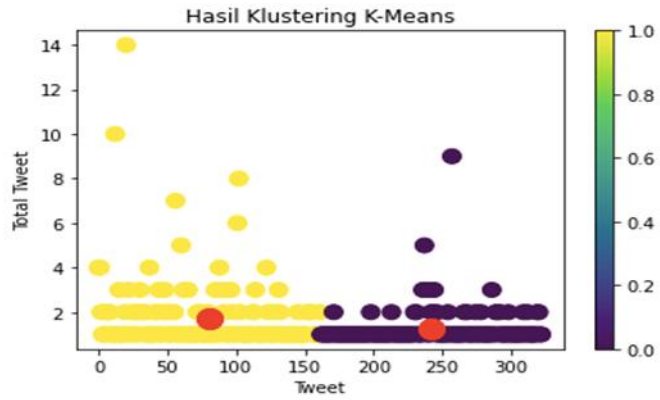
Pengujian dilakukan pada platform Jupyter Notebook. Bahasa pemrograman yang digunakan adalah bahasa pemrograman Python dengan Library yang digunakan merupakan library dasar seperti pandas, numpy dan scikit learn. Pandas merupakan library python untuk mempermudah pengolahan data dalam bentuk baris dan kolom seperti format data csv, sql maupun no-sql. Numpy merupakan library python untuk membantu pengolahan matriks dan operasi matematisnya. Library scikit-learn atau sklearn adalah modul untuk bahasa pemograman python yang dibangun diatas NumPy, SciPy, dan matplotlib, fungsinya dapat membantu melakukan processing data ataupun melakukan training data untuk kebutuhan machine learning seperti pemodelan data. Library Keras merupakan library jaringan syaraf tiruan tingkat tinggi yang ditulis dengan bahasa python dan mampu berjalan di atas TensorFlow, atau Theano. Library ini menyediakan fitur yang digunakan dengan fokus mempermudah pengembangan lebih dalam tentang Deep Learning.

Adapun hasil dari percobaan yang telah dilakukan menggunakan alat uji Python dengan Automatic cluster ditunjukkan pada Point 5, Gambar 3 dan seterusnya.

EXPERIEMENT AND ANALYSIS

Penelitian ini menggunakan algoritma K-means Automatic Cluster untuk menemukan anggota cluster yang memiliki tingkat kesamaan yang tinggi satu sama lain (homogenitas internal). Tahapan yang pertama dilakukan yaitu membaca data, kemudian menganalisa data dengan melihat data yang terdapat kolom kosong dan perlu dibersihkan, selanjutnya step to clean the data yaitu case folding, Tokenizing, Filtering dan Stemming, tahapan ini sangat diperlukan untuk menghasilkan data yang siap pada proses cluster. Langkah selanjutnya melakukan grouby data unique dan mengubah dalam bentuk array 2D. Selanjutnya untuk menemukan Cluster ideal, maka pertama mencari V terlebih dahulu, kemudian Vw minimum untuk menyatakan homogenitas internal dan Vb maksimum untuk menyatakan homogenitas eksternal. Namun, untuk Menemukan klaster ideal sangat sulit karena kita tidak dapat menerapkan min(V) secara langsung. Sehingga untuk menemukan optimal global sebagai klaster ideal yaitu mengidentifikasi varian bergerak dengan menempatkan λ sebagai ambang batas (Global Optimum). Tahapan terakhir yaitu Visualisasi berdasarkan hasil cluster terbaik.

Adapun rekapitulasi hasil pengujian menemukan pencarian Varince cluster dengan K-means Automatic Clustering ditunjukkan pada Gambar 3.

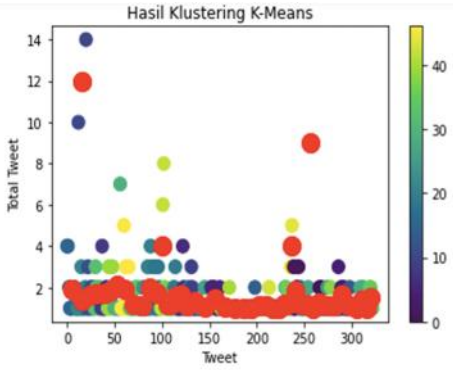



```
[541.0910441958775, 549.0333984403043]
Vw : 545.0745925863844
Vb : 545.0498500497974
V : 1.000045394997512
```

Gambar 3. Mencari Variance Cluster dengan 2 Centroid

Gambar 3, menunjukkan diagram hasil cluster K-means dari proses pencarian variance cluster dengan 2 centroid, yaitu centroid pertama ada pada jarak posisi 541.0910441958775 dan centroid kedua ada di posisi 549.0333984403043. sehingga diperoleh nilai Vw 545.0745925863844, Vb 545.0498500497974 dan V yaitu 1.000045394997512. Gambar kuning dan ungu menunjukkan sebaran data dimana sumbu Y adalah Tweet atau text tweet, artinya data tweet yang dilakukan oleh akun atau name yang sama sebanyak 14x dan sumbu X adalah sebaran data followers_account dari akun name. Proses ini akan berlanjut sampai akhirnya nilai Centroid menjadi 0 dan memenuhi sebaran data dengan nilai Vw, Vb dan V adalah 0. Sehingga dari proses ini akan ditemukan cluster optimum berada pada centroid ke berapa. Selanjutnya hasil pencarian variance cluster dengan centroid pertengahan ditunjukan pada gambar 4 dan centroid yang memenuhi sebaran data pada gambar 5.

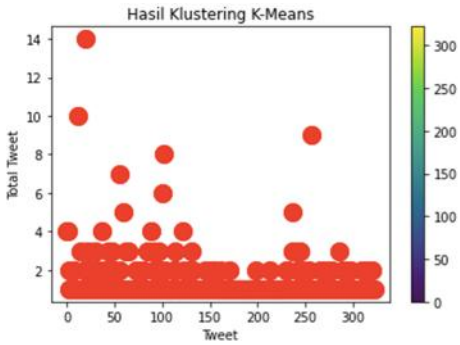
Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)
Vol. 2, No. 1 Februari 2023
e-ISSN: 2963-7813; p-ISSN: 2963-8178, Hal 58-72




```
[0.9022837979917011, 0.9265586534810941, 1.2380952380952381, 1.3333333333333333, 1.4409048917822658, 2.529255185302
041, 0.7916657254884908, 0, 1.0175362458312085, 1.9125469510066093, 0.0, 1.3306989120662942, 1.2380952380952381, 1.
2372654380025327, 1.9254587978639015, 2.6392326060784472, 1.096973741020287, 0.5742328856959301, 1.1098423247619926
, 0.40110309028469177, 0.6521674010996967, 1.4906765204924288, 0.6496640686841063, 1.792179414608689, 2.22222222222
22223, 1.2372654380025327, 2.3951131766905704, 0.7896400225647541, 0.5077507731413826, 2.8063487146752983, 0.700000
0000000001, 1.4285714285714284, 1.1394827389844508, 0.6496640686841063, 1.8423952355083786, 0.2546440075000701, 0.6
496640686841063, 0.6496640686841063, 1.4632034743640738, 0.6165686219530543, 2.793742653802631, 1.2380952380952381,
0.8309661425814783, 1.2092550755585658, 0.0, 0.7000000000000001, 1.24996095988393]
Vw : 1.4310083636254847
Vb : 54.17298022759312
V : 0.026415537000428813
```

Gambar 4. Mencari Variance Cluster dengan 47 Centroid

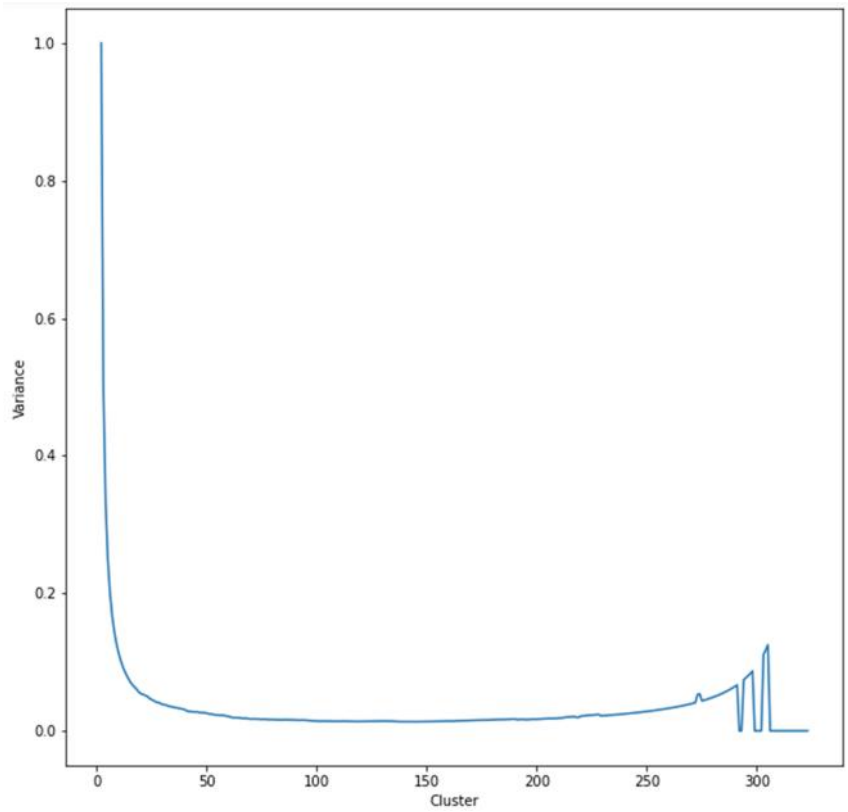
Gambar 4, menunjukkan diagram hasil cluster K-means dari proses pencarian variance cluster dengan 47 centroid, sehingga diperoleh nilai Vw 1.4310083636254847, Vb 54.1729022759312 dan V yaitu 0.026415537000428813. Proses ini adalah pencarian variance cluster yang berada pada centroid pertengahan untuk menemukan cluster optimum. Selanjutnya hasil cenroid yang memenuhi sebaran data ditunjukan pada gambar 5.




```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Vw : 0
Vb : 0
V : nan
```

Gambar 5. Mencari Variance Cluster dengan 322 Centroid

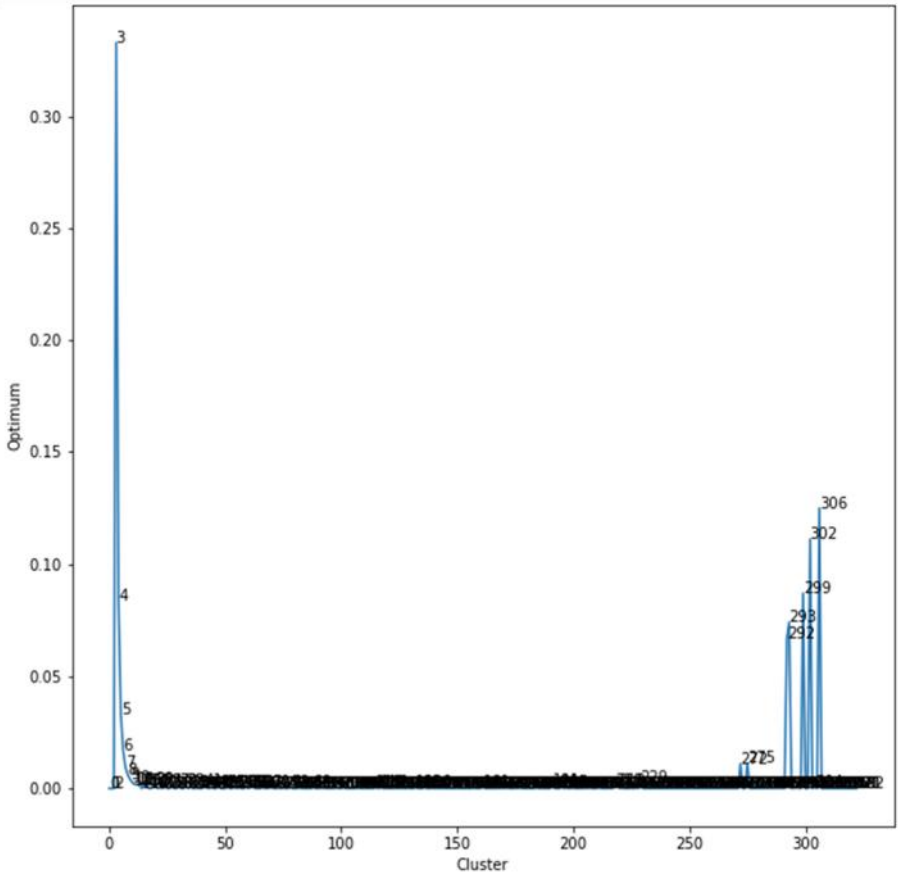
Gambar 5, menunjukkan diagram hasil cluster K-means dari proses pencarian variance cluster dengan 322 centroid, hasil ini adalah akhir dari proses pencarian variance cluster dengan nilai Centroid yang diperoleh 0 dan memenuhi sebaran data dengan nilai Vw, Vb dan V adalah 0. Sehingga dari proses pencarian variance cluster yang dimulai dengan 2 centroid sampai memenuhi sebaran data yang ada akan ditemukan gambaran variance cluster yang ditunjukan pada gambar 6.



Gambar 6. Visualialisasi Variance Cluster

Gambar 6, menunjukkan hasil visualisasi dari proses yang dilalui sebelumnya yang ditunjukan pada gambar 3, 4 dan 5. Berdasarkan hasil dari grafik ini akan terlihat jelas arah untuk menemukan cluster optimum. Dimana sumbu X adalah banyaknya centroid dari proses pencarian variance cluster yang dimulai dengan 2 centroid sampai centroid memenuhi sebaran data yang ada sebanyak 322 centroid dan sumbu Y adalah nilai variance cluster, grafik ini menunjukkan semakin dekat nilai variance yang diperoleh pada angka 1, maka menunjukkan pencarian variance cluster menemukan titik cluster optimum yang ditunjukan pada gambar 7 dan 8.

	Cluster	Optimum
0	2	0.000000
1	3	0.332975
2	4	0.083620
3	5	0.033310
4	6	0.016808
...	...	...
316	318	0.000000
317	319	0.000000
318	320	0.000000
319	321	0.000000
320	322	0.000000

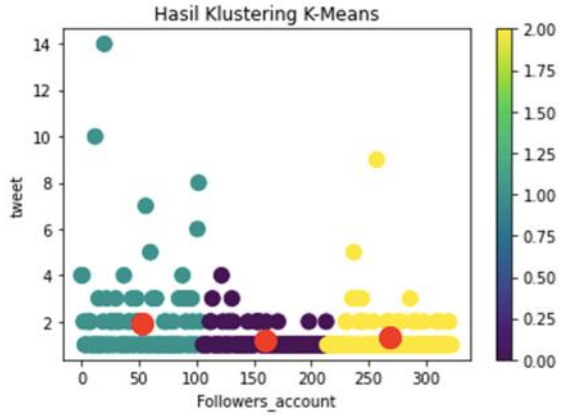


Gambar 7. Optimum Cluster dan **Gambar 8.** Visualisasi Optimum Cluster

Gambar 7, menunjukan hasil nilai cluster dan nilai optimum, hasil ini diperoleh berdasarkan proses perhitungan dari hasil nilai pencarian variance cluster yang ditunjukkan gambar 3, 4 dan 5 serta gambar 6 adalah visualisasi pencarian variance cluster. Sedangkan untuk gambar 8 adalah visualisasi dari gambar 7, terlihat bahwa cluster optimum yaitu 3 dengan nilai optimum 0.332975. Artinya, semakin dekat nilai cluster yang diperoleh pada angka 1, maka menunjukkan pencarian cluster menemukan titik optimum. Sehingga berdasarkan hasil dari pencarian variance cluster dan cluster optimum yang telah didapatkan maka selanjutnya akan dilakukan proses clustering dengan cluster terbaik yang dihasilkan. Proses dapat ditunjukan pada gambar 9.

Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)
Vol. 2, No. 1 Februari 2023
e-ISSN: 2963-7813; p-ISSN: 2963-8178, Hal 58-72

```
kmeans = KMeans(n_clusters = cluster_terbaik, random_state=123, init='random')
kmeans.fit(df_val)
kmeans.cluster_centers_
df['kluster'] = kmeans.labels_
data_ = df.values
output = plt.scatter(df_val[:,0], df_val[:,1], s = 100, c = df.kluster, marker = "o", alpha = 1)
centers = kmeans.cluster_centers_
plt.scatter(centers[:,0], centers[:,1], c='red', s=200, alpha=1 , marker="o")
plt.title("Hasil Klustering K-Means")
plt.xlabel("Followers_account")
plt.ylabel("tweet")
plt.colorbar(output)
plt.show()
```



Gambar 9. Visualisai Cluster Terbaik

Gambar 9, menunjukkan proses untuk melakukan cluster menggunakan K-means automatic clustering dengan jumlah cluster yaitu cluster terbaik berdasarkan hasil dari pencarian variance cluster dan optimum cluster, sehingga terlihat gambar visualisasi sebaran data, dimana untuk sumbu Y adalah Tweet atau text tweet, artinya data tweet yang dilakukan oleh akun atau name yang sama sebanyak 14x dan sumbu X adalah sebaran data followers_account dari akun name. Berdasarkan hasil tersebut, sebaran data ada 3 cluster yaitu, data warna tosca dengan 1 centroid yang ditandai warna merah, data ungu dengan 1 centroid yang ditandai warna merah dan data kuning dengan 1 centroid yang ditandai warna merah. Untuk Ukuran angka dan warna di samping grafik menunjukkan data, artinya warna ungu menunjukkan data berada pada nilai antara 0.50 sampai 0.00, warna tosca menunjukkan data berada pada rentang nilai 1.25 sampai 0.50 dan warna kuning berada pada rentang nilai 2.00 dampai 1.75. Terlihat bahwa hasil menunjukkan sebaran data yang ada memiliki tingkat

kesamaan atau menemukan anggota cluster yang memiliki tingkat kesamaan yang tinggi satu sama lain (homogenitas internal).

Berdasarkan hal ini cluster menggunakan K-means Automatic Cluster adalah tahapan terbaik untuk data yang memiliki ukuran sangat besar (big data) dan variabel berdimensi tinggi serta pola dari kumpulan data tidak terstruktur mengakibatkan komputasi sulit dilakukan. Sehingga penelitian ini mengusulkan pendekatan untuk menemukan global optimum dan membuat pengelompokan otomatis dengan menganalisis moving average yaitu dengan K-Means Automatic Clustering [9]. Berdasarkan hasil yang telah dilakukan dari penelitian ini yaitu data telah dilakukan eksplorasi dan evaluasi data yang berdimensi tinggi dari kumpulan tweet, yaitu text tweet random opinion di yogyakarta dengan Automatic clustering, terlihat bahwa sebaran data yang dihasilkan adalah sangat baik. Berdasarkan proses yaitu mencari variance cluster, kemudian optimum cluster sehingga dihasilkan cluster terbaik. Terakhir melakukan cluster dengan K-means Automatic Clustering.

CONCLUSION

Penelitian ini mengeksplorasi dan mengevaluasi data yang berdimensi tinggi dari kumpulan tweet, yaitu text tweet random opinion di yogyakarta dengan Automatic clustering. Algoritma K-means Automatic Clustering digunakan untuk cluster berdasarkan atribut data yang telah didapatkan. Pengaturan atribut data dibatasi pada data yang memiliki variabel berdimensi tinggi serta pola dari kumpulan data tidak terstruktur. Berdasarkan sebaran data yang telah digambarkan sebanyak 2 atribut terpilih yaitu atribut tweet dan followers_count sebagai acuan untuk cluster. Percobaan Pre-processing dilakukan dalam manipulasi atau proses menyeleksi data agar lebih terstruktur dengan melalui serangkaian tahapan sebelum dilakukan input kedalam model dengan tujuan agar kompatibel dengan library yang digunakan. Ada beberapa tahapan yang dilalui diantaranya. Cleansing, Case folding, Tokenizing, Filtering, Stemming. Kemudian selanjutnya tahapan mencari variance cluster untuk menemukan global optimum sebagai klaster ideal dengan mengidentifikasi varian bergerak dengan menempatkan λ sebagai ambang batas (Global Optimum). Sehingga diperoleh nilai cluster ideal yaitu 0.332975. Artinya, semakin dekat nilai cluster yang diperoleh pada angka 1, maka menunjukkan pencarian cluster menemukan titik optimum.

Sistem yang telah dibangun dapat dimanfaatkan untuk melakukan eksplorasi dan mengevaluasi data data yang berdimensi tinggi, sehingga menjadi pertimbangan dalam memberikan perkiraan pola dari kumpulan data tidak terstruktur dengan Visualisasi.

FUTURE WORK

Penelitian ini menggunakan data yang tidak lumayan banyak dalam proses cluster. sehingga hasil cluster optimum berada dibawah 0.5. berdasarkan hal itu proses yang dihasilkan atau data yang dilakukan proses cluster masih kurang optimal, yang dapat dilakukan pada penelitian ini selanjutnya yaitu global optimum dan cluster ideal yang diperoleh dapat ditingkatkan dengan penambahan jumlah atribut dan jumlah data yang digunakan serta bisa dibandingkan dengan algoritma cluster yang lain.

REFERENCES

"Media sosial - Wikipedia bahasa Indonesia, ensiklopedia bebas".

"Mengapa banyak orang yang nyaman curhat di twitter? - Sosial : Diskusi Komunikasi - Dictio Community".

F. Khairani, A. Kurnia, M. N. Aidi, and S. Pramana, "Predictions of Indonesia Economic Phenomena Based on Online News Using Random Forest," SinkrOn, vol. 7, no. 2, pp. 532–540, Apr. 2022, doi: 10.33395/sinkron.v7i2.11401.

M. F. Tyas, A. Kurnia, and A. M. Soleh, "TEXT CLUSTERING ONLINE LEARNING OPINION DURING COVID-19 PANDEMIC IN INDONESIA USING TWEETS," BAREKENG: Jurnal Ilmu Matematika dan Terapan, vol. 16, no. 3, pp. 939–948, Sep. 2022, doi: 10.30598/barekengvol16iss3pp939-948.

K. N. Aini, H. Murfi, K. Nur'aini, I. Najahaty, L. Hidayati, and S. Nurrohmah, "Combination of Singular Value Decomposition and K-means Clustering Methods for Topic Detection on Twitter", doi: 10.13140/RG.2.1.4081.2886.

"Elementary Survey Sampling, 7th ed.".

K. E. Setiawan, A. Kurniawan, A. Chowanda, and D. Suhartono, "Clustering models for hospitals in Jakarta using fuzzy c-means and k-means," Procedia Comput Sci, vol. 216, pp. 356–363, 2023, doi: 10.1016/j.procs.2022.12.146.

C. A. Murthy and N. Chowdhury, "In search of optimal clusters using genetic algorithms," 1996.

"ridho.lecturer.pens.ac.id: papers: Barakbah_IES_2004".

M. Alfian, A. Ridho Barakbah, and I. Winarno, "INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Indonesian Online News Extraction and Clustering Using Evolving Clustering." [Online]. Available: www.joiv.org/index.php/joiv