

**KLASTERISASI *TWEET* DENGAN ALGORITMA K-MEANS
UNTUK KASUS PENYEBARAN PENYAKIT MENULAR AKUT
(STUDI KASUS INFEKSI SALURAN PERNAPASAN AKUT,
KEBAKARAN HUTAN DAN POLUSI)**

SKRIPSI



MOHAMMAD ICHWAN SATRYO

PROGRAM STUDI S1 SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS AIRLANGGA

SURABAYA

2022

**KLASTERISASI *TWEET* DENGAN ALGORITMA K-MEANS
UNTUK KASUS PENYEBARAN PENYAKIT MENULAR AKUT
(STUDI KASUS INFEKSI SALURAN PERNAPASAN AKUT,
KEBAKARAN HUTAN DAN POLUSI)**

SKRIPSI



MOHAMMAD ICHWAN SATRYO

NIM. 081511633046

PROGRAM STUDI S1 SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS AIRLANGGA

SURABAYA

2022

**KLASTERISASI *TWEET* DENGAN ALGORITMA K-MEANS
UNTUK KASUS PENYEBARAN PENYAKIT MENULAR AKUT
(STUDI KASUS INFEKSI SALURAN PERNAPASAN AKUT,
KEBAKARAN HUTAN DAN POLUSI)**

SKRIPSI

Sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Komputer Bidang
Sistem Informasi pada Fakultas Sains dan Teknologi

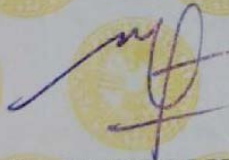
Universitas Airlangga

Mohammad Ichwan Satryo

NIM. 081511633046

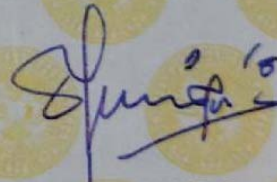
Disetujui oleh:

Pembimbing I,



Taufik ST, M.Kom
NIP. 197101042008121001

Pembimbing II,



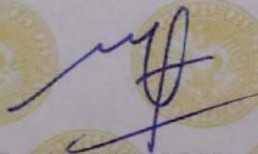
Purbandini, S.Si., M.Kom.
NIP. 197107122008122001

LEMBAR PENGESAHAN NASKAH SKRIPSI

Judul : Klasterisasi Dengan Tweet Dengan Algoritma K-Means
Untuk Kasus Penyebaran Penyakit Menular Akut (Studi
Kasus Infeksi Saluran Pernapasan Akut, Kebakaran
Hutan, dan Polusi)
Penyusun : Mohammad Ichwan Satryo
NIM : 081511633046
Pembimbing I : Taufik, S.T., M.Kom.
Pembimbing II : Purbandini, S.Si., M.Kom.
Tanggal Seminar : 28 Desember 2022

Disetujui oleh:

Pembimbing I,



Taufik, S.T., M.Kom.
NIP. 197101042008121001

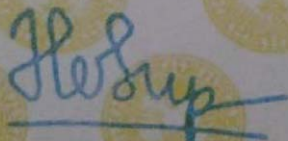
Pembimbing II,



Purbandini, S.Si., M.Kom.
NIP. 197107122008122001

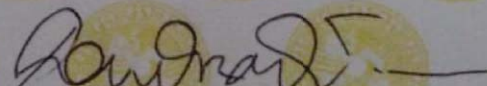
Mengetahui,

Ketua Departemen Matematika
Fakultas Sains dan Teknologi
Universitas Airlangga



Dr. Herry Suprajitno, S.Si., M.Si.
NIP. 196804041994031020

Koordinator Program Studi
SI Sistem Informasi
Fakultas Sains dan Teknologi
Universitas Airlangga



Dr. Rimuljo Hendradi, S.Si., M.Si.
NIP. 197102111997021001

SURAT PERNYATAAN TENTANG ORISINALITAS

Yang bertanda tangan dibawah ini, saya:

Nama : Mohammad Ichwan Satryo

NIM : 081511633046

Program Studi : S1-Sistem Informasi

Fakultas : Sains dan Teknologi

Jenjang : Sarjana (S1)

Menyatakan bahwa saya tidak melakukan kegiatan plagiat dalam penulisan skripsi saya yang berjudul:

“KLASTERISASI *TWEET* DENGAN ALGORITMA K-MEANS UNTUK KASUS PENYEBARAN PENYAKIT MENULAR AKUT (STUDI KASUS INFEKSI SALURAN PERNAPASAN AKUT, KEBAKARAN HUTAN DAN POLUSI)”

Apabila suatu saat nanti terbukti melakukan tindakan plagiat, maka saya akan menerima sanksi yang telah ditetapkan.

Demikian surat pernyataan ini saya buat dengan sebenar-benarnya.

Surabaya, Januari 2023



Mohammad Ichwan Satryo
NIM. 081511633046

PEDOMAN PENGGUNAAN SKRIPSI

Skripsi ini tidak dipublikasikan, namun tersedia di perpustakaan dalam lingkungan Universitas Airlangga, diperkenankan untuk dipakai sebagai referensi kepustakaan, tetapi pengutipan harus seizin penulis dan harus menyebutkan sumbernya sesuai kebiasaan ilmiah.

Dokumen skripsi ini merupakan hak milik Universitas Airlangga.

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah SWT karena berkat rahmat dan karunia-Nya penulis dapat menyelesaikan penyusunan skripsi yang berjudul judul “Klasterisasi Tweet dengan metode algoritma K-Means untuk Kasus Penyebaran Penyakit Menular Akut (Studi Kasus Infeksi Saluran pernapasan akut, Kebakaran Hutan dan Polusi).

Penulis tak lupa mengucapkan terima kasih kepada Taufik, S.T., M.Kom., selaku dosen pembimbing I dan Purbandini, S.Si., M.Kom. selaku dosen pembimbing II, yang dengan sabar dan ikhlas membimbing dan memberikan ilmunya kepada penulis sampai terselesaikannya skripsi ini. Penulis juga berterima kasih kepada orang tua, keluarga, orang terdekat, teman-teman S1 Sistem Informasi 2015, dan pihak-pihak yang telah memberikan semangat dan dukungannya kepada penulis.

Penulis menyadari bahwa dalam penyusunan skripsi ini banyak terdapat kekurangan, untuk itu penulis senantiasa bersedia dan terbuka dalam menerima saran serta kritik yang bersifat membangun demi kesempurnaan skripsi ini. Semoga skripsi ini dapat memberikan manfaat dan wawasan yang berguna, Amin.

Surabaya, Desember 2022

Mohammad Ichwan Satryo

UCAPAN TERIMAKASIH

Segala puji syukur kehadiran Tuhan Yang Maha Esa atas curahan anugerah dan karunia-Nya sehingga skripsi dengan judul “Klasterisasi Tweet Dengan Algoritma K-Means untuk Kasus Penyebaran Penyakit Menular Akut (Studi Kasus Infeksi Saluran Pernapasan Akut, Kebakaran Hutan, dan Polusi)” ini dapat diselesaikan.

Dalam pelaksanaan dan penyusunan skripsi ini, penulis banyak menemui kendala, namun dengan adanya bantuan dari berbagai pihak, akhirnya penelitian ini dapat terselesaikan. Oleh karena itu, penulis tidak lupa mengucapkan terima kasih kepada:

1. Allah SWT yang senantiasa memberikan segala berkah, rahmat, dan karunia-Nya dan Rasulullah Muhammad SAW yang telah menjadi suri tauladan sehingga penulis dapat terus menjalani roda kehidupan dengan sebaik-baiknya.
2. Hariyanto dan Devi, selaku orang tua penulis yang telah memberikan dukungan penuh dalam bentuk doa, kasih sayang, dan materi sekaligus menjadi semangat dan motivasi bagi penulis dalam menyelesaikan studi dan skripsi dengan baik.
3. Ira Puspitasari, S.T., M.T., Ph.D. selaku dosen wali yang dengan sabar membantu penulis dalam menyelesaikan studi S1 Sistem Informasi Universitas Airlangga.
4. Taufik, S.T., M.Kom., selaku dosen pembimbing I yang dengan sabar memahami, membimbing, membantu, dan memberikan ilmu selama penulisan hingga skripsi ini terselesaikan.
5. Purbandini, S.Si., M.Kom., selaku dosen pembimbing II yang dengan sabar memahami, membimbing, membantu, dan memberikan ilmu selama penulisan hingga skripsi ini terselesaikan.
6. Ahmad Alif Robit Al Hazmi yang telah membantu dengan berbagi ilmu pengetahuan selama penyusunan skripsi.

7. Oviary Satryo, yang telah memberi bantuan secara moral dan semangat selama penyusunan skripsi ini.
8. Muhammad Adam Bahtiar, Mohammad Agung Nugroho, Abdul Aziz Al Khakim, Jundan Barofil dan seluruh teman-teman S1 Sistem Informasi Universitas Airlangga angkatan 2015 yang telah mendukung satu sama lain dan membantu dengan berbagi ilmu pengetahuan selama penulisan skripsi.
9. Teman-teman serta berbagai pihak yang telah memberikan bantuan dan informasi selama proses penulisan skripsi.

Mohammad Ichwan Satryo. 2022. **Klasterisasi *Tweet* dengan Algoritma K-Means untuk Kasus Penyebaran Penyakit Menular akut (Studi Kasus Infeksi Saluran Pernapasan Akut, kebakaran hutan, dan polusi)**. Skripsi ini di bawah bimbingan Taufik, S.T., M.Kom dan Purbandini, S.Si., M.Kom. Program Studi S1 Sistem Informasi. Fakultas Sains dan Teknologi, Universitas Airlangga.

ABSTRAK

Berkembangnya media sosial mendorong munculnya situs *microblogging* seperti Twitter yang dapat berperan sebagai sistem sensor dengan menjadikan penggunaannya sebagai sensor dan *tweet* yang memiliki informasi tentang suatu lokasi. Volume dan kecepatan *tweet* pada saat suatu kejadian berlangsung sangat tinggi sehingga masyarakat yang terdampak dan petugas profesional mengalami kesulitan dalam pemrosesan informasi. Pada penelitian ini dilakukan klasterisasi *tweet* dengan algoritma K-Means untuk kasus penyebaran penyakit menular akut (studi kasus ISPA, kebakaran hutan, dan polusi). Penelitian ini menggunakan data teks dari media sosial Twitter dengan kata kunci tentang penyebaran penyakit menular akut (studi kasus ISPA, kebakaran hutan dan polusi). Sebelum proses klasterisasi, dilakukan tahap pertama adalah praproses seperti *case folding*, penghapusan simbol khusus *tweet*, penghapusan angka dan tanda baca, penghapusan *white space* dan *single character*, *tokenizing*, penghapusan *stopword*, normalisasi, dan *stemming* sehingga mengubah data *tweet* menjadi *Term Document Matix* (TDM). Tahap klasterisasi dengan K-Means pada TDM memperoleh klaster terbaiknya pada klaster dengan jumlah *term* sebesar 20 dan nilai *k* sebesar 8 sehingga menghasilkan nilai *Sum Squared Error* sebesar 10.912,70 pada ISPA. Pada kebakaran hutan dengan nilai *k* sebesar 6 menghasilkan *Sum Square Error* (SSE) sebesar 56.717,76 dan polusi dengan nilai *k* sebesar 8 menghasilkan *Sum Square Error* (SSE) sebesar 75.899,30. Pola *tweet* dari hasil klasterisasi terkait penyebaran penyakit menular akut (studi kasus ISPA, kebakaran hutan, dan polusi) terbanyak merupakan bulan Juli, Agustus, dan September. Sedangkan untuk jumlah *tweet* terkait ISPA, kebakaran hutan, dan polusi terendah pada bulan Januari. Dibandingkan dengan data dari Kemenkes 2020 terbanyak pada bulan Juli, Agustus, dan September.

Kata kunci: Klasterisasi, *Tweet*, *Data mining*, TDM, ISPA.

Mohammad Ichwan Satryo. 2022. **Clustering and Tweet using K-Means Algorithm for Dissemination Issues of Infectious Disease with Direct Acute (A Case Study of ARI, forest fire, and pollution)**. This thesis supervised by Taufik, S.T., M.Kom. and Purbandini, S.Si., M.Kom. Bachelor of Computer Science (Information System). Faculty of Science and Technology, Universitas Airlangga.

ABSTRACT

The proliferation of social media fosters the microblogging sites such as Twitter as a sensor system with their users acting as sensors and their tweet conveying information with a geographic location. Volume and velocity of tweets posted during events today tend to be extremely high, making it hard for event-affected communities and professional responders to process the information in a timely manner. This study aims to cluster tweet data using the K-Means algorithm for dissemination issues of infectious disease with direct contact (a case study of ARI, forest fire and pollution). This study uses text data from Twitter with keywords related to dissemination issues of infectious disease with direct acute (a case study of ARI, forest fire, and pollution). Before the clustering process, preprocessing steps will be performed such as case folding, deletion of special tweet symbols, deletion of numbers and punctuation, deletion white space and single character, *tokenizing*, stopwords, normalization, and stemming so that transform tweet data into Term Document Matrix (TDM). Clustering step with K-Means algorithm on TDM uses 20 as the best number of term and 8 as the best k value so that Sum Squared Error (SSE) value is 10.912,70 at ARI, at forest fire 6 as the best k value so that Sum Squared Error (SSE) value is 56.717,76 and pollution 8 as the best k value so that Sum Squared Error (SSE) value is 75.899,30. The pattern of clustering results related to the spread of acute infectious acute (a case study of ARI, forest fire, and pollution) the highest were in July, August and September. While the number of tweets related to ARI, forest fires, and pollution was the lowest in January. Compared to data from the Ministry of Health in 2020, most were in July, August and September.

Keywords: *Clustering, Tweet, Data mining, TDM, ARI*

DAFTAR ISI

	Halaman
LEMBAR JUDUL	i
LEMBAR PENGESAHAN NASKAH SKRIPSI ... Error! Bookmark not defined.	
SURAT PERNATAAN TENTANG ORSINALITAS	iv
PEDOMAN PENGGUNAAN SKRIPSI	v
KATA PENGANTAR	vi
UCAPAN TERIMA KASIH	vii
ABSTRAK	ix
ABSTRACT	x
DAFTAR ISI	xii
DAFTAR GAMBAR	xiii
DAFTAR TABEL	xvii
BAB I PENDAHULUAN	1
1.1. Latar Belakang.....	1
1.2 Rumusan Masalah.....	4
1.3 Tujuan Penelitian	4
1.4 Manfaat Penelitian	5
1.5 Batasan Masalah	5
BAB II TINJAUAN PUSTAKA	6
2.1 <i>Web Scraping</i>	6
2.2 <i>Data Mining</i>	6

2.3	Praproses Data	7
2.4	Pembuatan Term Document Metrix.....	13
2.5	Klasterisasi.....	13
2.6	Pemilihan Data Optimal.....	17
2.7	Penelitian Sebelumnya.....	188
BAB III METODE PENELITIAN		211
3.1.	Waktu Penelitian.....	211
3.2.	Objek Penelitian.....	211
3.3.	Tahapan Penelitian.....	222
BAB IV PEMBAHASAN DAN HASIL		311
4.1.	<i>Scraping Twitter</i>	31
4.2	Praproses Data	40
4.3	Klasterisasi dengan K-Means	477
4.4	Evaluasi Hasil Analisis	59
BAB V KESIMPULAN DAN SARAN		66
5.1	Kesimpulan.....	66
5.2	Saran.....	66
DAFTAR PUSTAKA		68
LAMPIRAN.....		71
Lampiran 1:Daftar Kata Stopword.....		71
Lampiran2:Daftar Kata Normalisasi.....		73

DAFTAR GAMBAR

Nomor	Judul Gambar	Halaman
Gambar 2.1	Contoh <i>Tokenizing</i>	8
Gambar 2.2	Knee Point Pada Jumlah Klaster Optimal.....	17
Gambar 3.1	Tahapan Penelitian.....	21
Gambar 3.2	Praproses Data	23
Gambar 4.1	Proses import Library.....	31
Gambar 4.2	Proses link ke Twitter dan memasukan search term.....	32
Gambar 4.3	Halaman Search pada Twitter.....	33
Gambar 4.4	Proses scraping Tweet yang ditampilkan pada halaman chrome.....	33
Gambar 4.5	Script Autoscroll.....	34
Gambar 4.6	Code untuk menyimpan hasil scrape ke csv.....	34
Gambar 4.7	Raw Data hasil Scrapinig.....	35
Gambar 4.8	Proses Penggabungan hasil Scraping.....	36
Gambar 4.9	Regular Expresion yang digunakan untuk kombinasi file.....	36
Gambar 4.10	Proses penggabungan dan penyimpanan file Scraping.....	36
Gambar 4.11	Tweet yang berisikan Handle, Username dan Timestamp.....	37
Gambar 4.12	Proses Penhapusan username dan handle.....	37
Gambar 4.13	Output proses penghapusan Username dan Handle.....	37
Gambar 4.14	Proses Penghapusan Mention.....	38
Gambar 4.15	Hasil Output dari proses Penghapusan Mention.....	38
Gambar 4.16	Proses penghapusan tweet duplikat.....	38
Gambar 4.17	Proses dan Output <i>Case Folding</i>	39
Gambar 4.18	Proses Penghapusan simbol khusus Tweet.....	40
Gambar 4.19	Output Proses pembersihan Khusus di Twitter.....	40
Gambar 4.20	Proses dan Output Penghapusan Angka dan tanda baca.....	40
Gambar 4.21	Proses dan Output Penghapusan White Space dan single Character.....	41

Gambar 4.22 Proses dan Output dari Tokenisasi.....	42
Gambar 4.23 Contoh Kata Stopword.....	42
Gambar 4.24 Proses dan Output dari Stopword Removal.....	43
Gambar 4.25 Contoh List kata yang dinormalisasi.....	43
Gambar 4.26 Proses dan Output normalisasi.....	44
Gambar 4.27 Proses Stemming.....	44
Gambar 4.28 Output proses Stemming.....	45
Gambar 4.29 Contoh pembuatan <i>Term document matrix</i> ISPA.....	47
Gambar 4.30 Grafik pengaruh jumlah dimensi matriks terhadap SSE ISPA.....	48
Gambar 4.31 Grafik pengaruh jumlah dimensi matriks terhadap Koefisien Silhouette ISPA.....	49
Gambar 4.32 Grafik pengaruh jumlah dimensi matriks terhadap SSE ISPA.....	49
Gambar 4.33 Grafik pengaruh jumlah dimensi matriks terhadap Koefisien Silhouette ISPA.....	49
Gambar 4.34 Grafik metode elbow untuk ISPA.....	50
Gambar 4.35 Grafik analisis silhouette untuk ISPA.....	51
Gambar 4.36 Contoh data hasil klasterisasi ISPA dengan nilai k Sebesar 8.....	51
Gambar 4.37 Contoh pembuatan <i>Term document matrix</i> Kebakaran hutan.....	51
Gambar 4.38 Grafik pengaruh jumlah dimensi matriks terhadap SSE Kebakaran Hutan.....	52
Gambar 4.39 Grafik pengaruh jumlah dimensi matriks terhadap Koefisien Silhouette Kebakaran Hutan	53
Gambar 4.40 Grafik pengaruh jumlah dimensi matriks terhadap SSE Kebakaran Hutan.....	53
Gambar 4.41 Grafik pengaruh jumlah dimensi matriks terhadap Koefisien Silhouette Kebakaran Hutan.....	53
Gambar 4.42 Grafik metode elbow untuk Kebakaran Hutan.....	54
Gambar 4.43 Grafik analisis silhouette untuk Kebakaran Hutan.....	55

Gambar 4.44 Contoh data hasil klasterisasi Kebakaran hutan dengan nilai k Sebesar 6.....	55
Gambar 4.45 Contoh pembuatan <i>Term document metrix</i> Polusi.....	55
Gambar 4.46 Grafik pengaruh jumlah dimensi matriks terhadap SSE Polusi.....	56
Gambar 4.47 Grafik pengaruh jumlah dimensi matriks terhadap Koefisien Silhouette Polusi.....	57
Gambar 4.48 Grafik pengaruh jumlah dimensi matriks terhadap SSE Polusi.....	57
Gambar 4.49 Grafik pengaruh jumlah dimensi matriks terhadap Koefisien Silhouette Polusi.....	57
Gambar 4.50 Grafik metode elbow untuk Polusi.....	58
Gambar 4.51 Grafik analisis silhouette untuk Polusi.....	59
Gambar 4.52 Contoh data hasil klasterisasi Polusi dengan nilai k Sebesar 8.....	59
Gambar 4.53 Jumlah <i>Tweet</i> ISPA per bulan 2019.....	63
Gambar 4.54 Jumlah <i>Tweet</i> Kebakaran Hutan per bulan 2019.....	64
Gambar 4.55 Jumlah kasus Kebakaran hutan per bulan 2019.....	65
Gambar 4.56 Jumlah <i>Tweet</i> Polusi per bulan 2019.....	65

DAFTAR TABEL

Nomor	Judul Tabel	Halaman
Tabel 2.1	Kombinasi Awalan yang Tidak Diizinkan	12
Tabel 2.2	Aturan Pemenggalan Awalan yang Kompleks.....	12
Tabel 2.3	Metode Klasterisasi (Han et. al., 2012)	14
Tabel 2.4	Range Nilai dan Interpretasi Silhouette Klaster.....	18
Tabel 3.1	Kata Kunci ISPA yang Digunakan untuk Pencarian Tweet.....	21
Tabel 3.2	Daftar Kata Kunci Kebakaran Hutan yang Digunakan untuk Pencarian Tweet.....	21
Tabel 3.3	Daftar Kata Kunci polusi yang Digunakan untuk Pencarian Tweet	22
Tabel 3.4	Daftar Atribut Data Tweet yang Diperoleh dari Hasil Akuisisi Data	22
Tabel 3.5	Contoh Data Tweet yang Diperoleh dari Hasil Akuisisi Data	23
Tabel 3.6	Contoh Data Tweet Sebelum dan Sesudah <i>Case Folding</i>	24
Tabel 3.7	Contoh Data Tweet Sebelum dan Sesudah Penghapusan simbol khusus padat Tweet.....	25
Tabel 3.8	Contoh Data Tweet Sebelum dan Sesudah Penghapusan Angka dan Tanda baca	26
Tabel 3.9	Contoh Data Tweet Sebelum dan Sesudah White Space dan Single Character	26
Tabel 3.10	Contoh Data Tweet Sebelum dan Sesudah Penghapusan <i>Tokenizing</i>	27
Tabel 3.11	Contoh Data Tweet Sebelum dan Sesudah Stopword.....	28
Tabel 3.12	Contoh Data Tweet Sebelum dan Sesudah Normalisasi.....	29
Tabel 3.13	Contoh Data Tweet Sebelum dan Sesudah Stemming.....	29
Tabel 4.1	Daftar Parameter <i>Class</i> KMeans yang digunakan.....	47
Tabel 4.2	Hasil Klasterisasi ISPA.....	60
Tabel 4.3	Sebaran Frekuensi Kata Kunci ISPA pada Klaster.....	60