

**SISTEM IMPLEMENTASI ANALISIS KLASTERISASI DAN
SENTIMEN DATA UNTUK MENGETAHUI TOPIK YANG
PALING BANYAK DIBAHAS MENGGUNAKAN METODE K-
MEANS (STUDI KASUS DATA TWITTER DAN YOUTUBE
TENTANG PPKM)**

**Tesis
untuk memenuhi sebagian persyaratan
mencapai derajat Sarjana S-2 Program Studi
Magister Sistem Informasi**



**Dikky Nayosa
30000320410010**

SEKOLAH PASCASARJANA

**SEKOLAH PASCASARJANA
UNIVERSITAS DIPONEGORO
SEMARANG
2023**

HALAMAN PERSETUJUAN

Tesis dengan judul :

SISTEM IMPLEMENTASI ANALISIS KLASTERISASI DAN SENTIMEN DATA UNTUK MENGETAHUI TOPIK YANG PALING BANYAK DIBAHAS MENGGUNAKAN METODE K-MEANS (STUDI KASUS DATA TWITTER DAN YOUTUBE TENTANG PPKM)

Oleh:
Dikky Nayosa
30000320410010

Telah dilakukan pembimbingan tesis dan dinyatakan layak untuk mengikuti ujian tesis pada Program Studi Magister Sistem Informasi Sekolah Pascasarjana Universitas Diponegoro.

Semarang,
Menyetujui,

Pembimbing I

Pembimbing II

Dr. Rahmat Gernowo, MSI
NIP.196511231994031003

Dr.Ir. R. Rizal Isnanto, S.T., M.M.,
M.T., IPU, ASEAN Eng.
NIP.197007272000121001

**PERNYATAAN PERSETUJUAN
PUBLIKASI TESIS UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademik Universitas Diponegoro, saya yang bertanda tangan di bawah ini :

Nama : Dikky Nayosa
NIM : 30000320410010
Program Studi : Magister Sistem Informasi
Program : Sekolah Pascasarjana
Jenis Karya : Tesis

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Diponegoro Hak Bebas Royalti Noneksklusif atas karya ilmiah saya yang berjudul :

**SISTEM IMPLEMENTASI ANALISIS KLASTERISASI DAN
SENTIMEN DATA UNTUK MENGETAHUI TOPIK YANG
PALING BANYAK DIBAHAS MENGGUNAKAN METODE K-
MEANS (STUDI KASUS DATA TWITTER DAN YOUTUBE
TENTANG PPKM)**

beserta perangkat yang ada. Dengan Hak bebas Royalti Noneksklusif ini Program Studi Magister Sistem Informasi Sekolah Pascasarjana Universitas Diponegoro berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*) merawat, dan mempublikasikan tesis saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik hak cipta.

Dibuat di : Semarang

Pada tanggal : 7 Juni 2023

Yang menyatakan

SEKOLAH PASCASARJANA

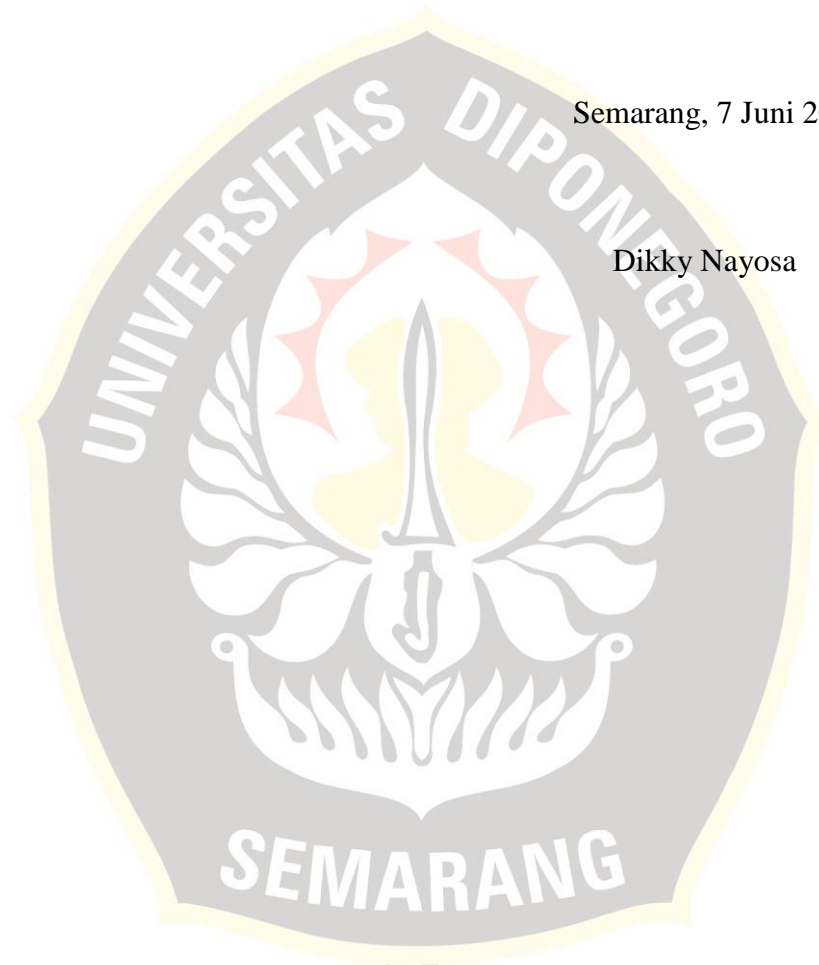
Dikky Nayosa
NIM 30000320410010

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam tesis ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar akademik di suatu perguruan tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Semarang, 7 Juni 2023

Dikky Nayosa



SEKOLAH PASCASARJANA

**SISTEM IMPLEMENTASI ANALISIS KLASTERISASI DAN SENTIMEN
DATA UNTUK MENGETAHUI TOPIK YANG PALING BANYAK
DIBAHAS MENGGUNAKAN METODE K-MEANS (STUDI KASUS
DATA TWITTER DAN YOUTUBE TENTANG PPKM)**

ABSTRAK

Pada tahun 2021, pemerintah menerapkan kebijakan Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) yang menarik perhatian publik saat ini. Pemberlakuan PPKM ini mendapat tanggapan yang beragam dari masyarakat. Twitter dan Youtube adalah platform media sosial yang digunakan oleh pengguna untuk menyampaikan pendapat secara langsung. Konten yang terdapat di dalam Twitter dan Youtube juga sangat beragam. Dalam konteks ini, dibutuhkan suatu metode pendeteksian topik secara otomatis, seperti metode *Mini Batch K-means Clustering*, yang dapat mempermudah pengguna dalam mengakses informasi. Penelitian ini menggunakan pendekatan *Mini Batch*, di mana hanya sekelompok kecil data yang digunakan dalam proses pengelompokan (*clustering*). Hasil penelitian menunjukkan bahwa untuk data tweet dengan kata kunci PPKM, diperoleh 12 kelompok cluster berdasarkan pengujian menggunakan *Sum of Squared Error* (SSE). Setelah proses pengelompokan data, hasil clustering akan divisualisasikan menggunakan *Word Cloud*, dan sistem akan menampilkan persentase kata-kata yang sesuai dengan *Word Cloud* tersebut.

Kata kunci : Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM), clustering, K-Means, Mini Batch K-Means Clustering

SEMARANG
SEKOLAH PASCASARJANA

IMPLEMENTATION SYSTEM FOR CLUSTER ANALYSIS AND DATA SENTIMENT USING THE K-MEANS METHOD TO DETERMINE THE MOST DISCUSSED TOPICS

ABSTRACT

The Community Activities Restrictions Enforcement (CARE) government rule, that is currently a matter of public concern, will be enforced in 2021. Community Activities Restrictions Enforcement (CARE) highlights the community's pros and cons. Twitter and YouTube are social media that facilitate direct user expression. The material offered on social media platforms Twitter and YouTube is likewise quite diversified. Therefore, an automatic approach for topic detection is required, such as Mini Batch K-means Clustering, which facilitates user access to information. This study employs the Mini Batch method, which utilizes just a limited set of data for the clustering procedure. Based on testing with the Sum of Squared Error, this study's clustering results for tweet data including the phrase Community Activities Restrictions Enforcement (CARE) produced 12 cluster groups. The clustering results will be represented using Word Cloud, and the system will display the percentage of words based on Word Cloud.

Keywords : Community Activities Restrictions Enforcement (CARE), clustering, K-Means, Mini Batch K-Means Clustering



SEKOLAH PASCASARJANA

DAFTAR ISI

	Halaman
Halaman Judul	i
Halaman Pengesahan	ii
Halaman Pernyataan Persetujuan Publikasi	iii
Halaman Pernyataan	iv
Daftar Isi	v
Daftar Gambar	vii
Daftar Tabel	viii
Abstrak	ix
Abstract	x
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Tujuan Penelitian	2
1.3 Manfaat Penelitian	3
BAB II TINJAUAN PUSTAKA DAN DASAR TEORI	4
2.1 Tinjauan Pustaka	4
2.2 Dasar Teori	10
2.1.1 Data Mining	10
2.1.2 Text Mining	11
2.1.3 <i>Term Frequency – Inverse Dokumen Frequency (TF-IDF)</i>	12
2.1.4 <i>Sum of Squared Errors (SSE)</i>	14
2.1.5 Klasterisasi K-means	14
2.1.6 <i>Mini Batch K-Means</i>	16
2.1.7 <i>Word Cloud</i>	18
2.1.8 Pemberlakuan Pembatasan Kegiatan Masyarakat	18
2.1.9 Kerangka Kerja Streamlit	19
BAB III METODE PENELITIAN	21
3.1 Bahan dan Alat Penelitian	21
3.1.1 Bahan penelitian	21
3.1.2 Alat penelitian	22
3.2 Prosedur Penelitian	22
3.3 Kerangka Sistem Informasi	25
BAB IV HASIL PENELITIAN DAN PEMBAHASAN	28
4.1 Hasil dan Tampilan Sistem	28
4.2 Pengumpulan Data	32
4.3 Prapengolahan Data	33
4.3.1 <i>Case Folding</i>	33
4.3.2 <i>Filtering</i>	34

4.3.3 <i>Removing Stopword</i>	34
4.3.4 <i>Stemming</i>	35
4.3.5 <i>Tokenizing</i>	36
4.4 Implementasi Model	37
4.5 <i>Word Cloud</i>	44
4.6 Pengujian dengan dataset lain	46
4.6.1 Pengujian dataset gempa cianjur	46
4.6.2 Pengujian dataset nikahan anak presiden	50

BAB V KESIMPULAN DAN SARAN	53
5.1 Kesimpulan	53
5.2 Saran	53

DAFTAR PUSTAKA	55
----------------------	----



SEKOLAH PASCASARJANA

DAFTAR GAMBAR

	Halaman
Gambar 2.2 Akar Ilmu Data Mining	11
Gambar 2.2 Ilustrasi Algoritma K-Means (Yudi Agusta, 2007)	16
Gambar 3.1 Tahapan prapengolahan data.....	23
Gambar 3.2 Kerangka sistem informasi.....	25
Gambar 4.1 Tampilan halaman depan sistem	28
Gambar 4.2 Tampilan halaman utama sistem.....	28
Gambar 4.3 Tampilan jendela direktori komputer.....	29
Gambar 4.4 Tampilan data berhasil di <i>upload</i>	29
Gambar 4.5 Tampilan hasil klasterisasi	30
Gambar 4.6 Tampilan <i>word cloud</i> hasil klasterisasi	30
Gambar 4.7 Tampilan persentase kata hasil klasterisasi.....	31
Gambar 4.8 Grafik nilai SSE data komentar youtube.....	41
Gambar 4.9 Grafik nilai SSE data twitter yang mengandung kata PPKM ...	42
Gambar 4.10 <i>Word cloud</i> topik data komentar youtube	44
Gambar 4.11 <i>Word cloud</i> dari data twitter.....	45
Gambar 4.12 Dataset gempa cianjur berhasil diunggah	47
Gambar 4.13 Hasil Klasterisasi dataset gempa cianjur.....	47
Gambar 4.14 <i>Word cloud</i> dataset gempa cianjur	48
Gambar 4.15 Persetase Kata dataset gempa cianjur	48
Gambar 4.16 Dataset nikahan anak presiden berhasil diunggah	49
Gambar 4.17 Hasil klasterisasi dataset nikahan anak presiden.....	49
Gambar 4.18 <i>Word cloud</i> dataset nikahan anak Presiden.....	50
Gambar 4.19 Persetase kata dataset Nikahan anak Presiden	50

SEMARANG

SEKOLAH PASCASARJANA

DAFTAR TABEL

	Halaman
Tabel 2.1 Hasil penelitian terdahulu	4
Tabel 4.1 Hasil pencarian atribut data youtube	32
Tabel 4.2 Hasil pencarian atribut data twitter	32
Tabel 4.3 Struktur data.....	32
Tabel 4.4 Hasil tahap teks komentar ke <i>case folding</i>	33
Tabel 4.5 Hasil tahap <i>case folding</i> ke <i>filtering</i>	34
Tabel 4.6 Hasil tahap <i>filtering</i> ke <i>removing stopword</i>	34
Tabel 4.7 Hasil tahap <i>removing stopword</i> ke <i>stemming</i>	35
Tabel 4.8 Hasil tahap <i>tokenizing</i>	35
Tabel 4.9 Hasil tahap prapengolahan data	36
Tabel 4.10 Pustaka (<i>library</i>) yang digunakan.....	37
Tabel 4.11 Hasil klaster data komentar youtube.....	42
Tabel 4.12 Hasil klaster data cuitan twitter yang mengandung kata PPKM..	43
Tabel 4.13 Hasil pencarian atribut pengujian menggunakan dataset lain.....	46
Tabel 4.14 Struktur data dataset lain.....	46



SEKOLAH PASCASARJANA

DAFTAR ARTI LAMBANG DAN SINGKATAN

DAFTAR ARTI LAMBANG

Lambang	Arti Lambang
tf	Suatu kata atau <i>term</i> terdapat dalam dokumen atau tidak
Ln	Fungsi <i>inverse</i>
n	Jumlah dokumen
W	Bobot
K	Jumlah klaster
X_i	Data ke i
X_n	Data ke n
C_k	Centroid klaster
μ_k	Titik awal klaster
ϵ	Anggota dari

DAFTAR SINGKATAN

Singkatan	Kepanjangan Singkatan
TF	<i>Term</i> frekuensi
IDF	<i>Inverse</i> dokumen frekuensi
SSE	<i>Sum of Squared Errors</i>
PSBB	Pembatasan Sosial Berskala Besar
PPKM	Pemberlakuan Pembatasan Kegiatan Masyarakat
CSV	<i>Comma Separated Values</i>

URL	<i>Uniform Resource Locators</i>
-----	----------------------------------



SEKOLAH PASCASARJANA



SEKOLAH PASCASARJANA