JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 198-204

Clustering Tweet Publik Menggunakan Metode TF-IDF dan K-Means dengan Modifikasi Preprocessing

Rama Achmad Fadillah[1], Muhammad Sabiilul Hikam Azzuhrie[2], Rayhan Numa Shaquille Then[3], Rafif Octaviano Hibabullah[4], Perani Rosyani[5]

[12345]Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1]adenfadill28@gmail.com, [2]billhikam2003@gmail.com, [3]rafifoctaviano.h@gmail.com, [5]dosen00837@unpam.co.id

Abstrak-Pertumbuhan platform media sosial, khususnya Twitter, telah menghasilkan data teks tidak terstruktur dalam jumlah besar yang berisi berbagai opini dan topik publik. Analisis manual terhadap data tersebut tidak efisien dan tidak praktis. Penelitian ini mengusulkan pendekatan clustering teks otomatis untuk mengelompokkan tweet ke dalam topik yang bermakna menggunakan TF-IDF (Term Frequency-Inverse Document Frequency) untuk representasi fitur dan algoritma K-Means untuk clustering tidak terawasi. Penelitian menggunakan dataset TweetTopic yang berisi tweet berbahasa Inggris berlabel berbagai topik. Teknik preprocessing yang ditingkatkan termasuk pembersihan data, case folding, tokenisasi, penghapusan stopword, dan stemming diterapkan untuk meningkatkan kualitas teks. Jumlah cluster optimal (K=3) ditentukan menggunakan Metode Elbow dan analisis Silhouette Score, sesuai dengan tiga topik utama: politik, olahraga, dan hiburan. Hasil eksperimen menunjukkan Silhouette Score sebesar 0,64, mengindikasikan pemisahan dan kualitas cluster yang baik. Dibandingkan dengan penelitian baseline menggunakan CountVectorizer dengan preprocessing minimal, pendekatan TF-IDF dengan preprocessing komprehensif menunjukkan peningkatan signifikan dalam koherensi cluster dan interpretabilitas topik. Temuan penelitian memberikan wawasan praktis untuk pemantauan media sosial, analisis isu publik, dan aplikasi text mining.

Kata kunci: Text Mining, Clustering, K-Means, TF-IDF, Natural Language Processing, Analisis Twitter, Pemodelan Topik

Abstract-*The rapid growth of social media platforms, particularly Twitter, has generated vast amounts of unstructured text data containing diverse public opinions and topics. Manual analysis of such data is inefficient and impractical. This study proposes an automated text clustering approach to group tweets into meaningful topics using TF-IDF (Term Frequency-Inverse Document Frequency) for feature representation and the K-Means algorithm for unsupervised clustering. The research utilizes the TweetTopic dataset containing English tweets labeled with various topics. Enhanced preprocessing techniques including data cleaning, case folding, tokenization, stopword removal, and stemming were applied to improve text quality. The optimal number of clusters (K=3) was determined using the Elbow Method and Silhouette Score analysis, corresponding to three main topics: politics, sports, and entertainment. Experimental results show a Silhouette Score of 0.64, indicating good cluster separation and quality. Compared to baseline research using CountVectorizer with minimal preprocessing, the TF-IDF approach with comprehensive preprocessing demonstrated significant improvements in cluster coherence and topic interpretability. The findings provide practical insights for social media monitoring, public issue analysis, and text mining applications.*

Keywords: *Text Mining, Clustering, K-Means, TF-IDF, Natural Language Processing, Twitter Analysis, Topic Modeling*

1. PENDAHULUAN

Pertumbuhan eksponensial platform media sosial telah mengubah cara informasi dibuat, dibagikan, dan dikonsumsi secara global. Twitter, sebagai salah satu platform mikroblog utama, menghasilkan jutaan tweet setiap hari yang mencakup berbagai topik, seperti politik, olahraga, hiburan, dan peristiwa terkini (Wang et al., 2020). Volume data teks yang besar dan tidak terstruktur ini menghadirkan tantangan sekaligus peluang dalam analisis otomatis serta ekstraksi wawasan berbasis data.

Pendekatan manual tradisional dalam analisis tweet cenderung memakan waktu, bersifat subjektif, dan tidak praktis jika diterapkan pada data berskala besar. Oleh karena itu, clustering teks

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 198-204

otomatis menjadi solusi yang relevan karena mampu mengelompokkan dokumen berdasarkan kemiripan tanpa memerlukan data berlabel, sehingga mendukung penemuan topik dan analisis tren secara efisien (Chen et al., 2023). Namun demikian, karakteristik teks tweet yang bersifat informal—ditandai dengan penggunaan singkatan, emoji, hashtag, serta bahasa tidak baku—menimbulkan tantangan tersendiri dalam tahap preprocessing dan representasi fitur.

Penelitian sebelumnya terkait clustering tweet umumnya masih mengandalkan metode representasi teks dasar, seperti CountVectorizer dengan preprocessing yang terbatas (Surianto, 2023; Fani et al., 2021). Meskipun pendekatan tersebut mampu menghasilkan cluster awal, metode ini belum sepenuhnya menangkap hubungan semantik antar istilah dan kurang efektif dalam menangani noise yang lazim terdapat pada teks media sosial. Sebagai alternatif, metode Term Frequency–Inverse Document Frequency (TF-IDF) menawarkan keunggulan dengan memberikan bobot pada istilah berdasarkan tingkat kepentingannya dalam suatu dokumen relatif terhadap keseluruhan korpus, sehingga berpotensi meningkatkan kualitas hasil clustering.

Berdasarkan kondisi tersebut, penelitian ini bertujuan mengembangkan sistem clustering tweet berbasis TF-IDF dan algoritma K-Means untuk membentuk cluster topik yang lebih bermakna. Penelitian ini juga membandingkan pendekatan baseline dengan pendekatan yang dimodifikasi melalui penerapan pipeline preprocessing yang lebih komprehensif, penggunaan representasi fitur TF-IDF, serta evaluasi cluster yang lebih sistematis. Fokus analisis diarahkan pada tiga topik publik utama, yaitu politik, olahraga, dan hiburan.

Secara metodologis, penelitian ini mencakup beberapa tahapan utama, yaitu preprocessing teks tweet, ekstraksi fitur menggunakan TF-IDF, penentuan jumlah cluster optimal, proses clustering menggunakan algoritma K-Means, serta evaluasi kualitas cluster menggunakan metrik evaluasi internal. Selain itu, hasil clustering divisualisasikan menggunakan teknik Principal Component Analysis (PCA) dan wordcloud untuk membantu interpretasi topik. Dataset tweet berlabel digunakan sebagai pembanding (ground truth) guna menilai kinerja pendekatan yang diusulkan.

Dengan mengintegrasikan aspek preprocessing yang lebih optimal, representasi fitur yang informatif, serta evaluasi yang komprehensif, penelitian ini diharapkan dapat memberikan kontribusi baik secara teoretis maupun praktis dalam bidang analisis media sosial, khususnya dalam pengelompokan topik tweet secara otomatis.

2. METODE PENELITIAN

2.1 Dataset

Dataset yang digunakan adalah TweetTopic Dataset yang berisi tweet berbahasa Inggris dengan berbagai label topik seperti politics, sports, dan entertainment. Label tidak digunakan dalam proses clustering, tetapi dimanfaatkan untuk evaluasi kesesuaian hasil clustering dengan topik sebenarnya (ground truth).

2.2 Preprocessing Teks

Tahap preprocessing bertujuan mengurangi noise pada tweet (URL, mention, hashtag, emoji, tanda baca, angka, dan karakter non-alfabet), menyeragamkan huruf (case folding), memecah kalimat menjadi token (tokenization), menghapus kata umum (stopword removal), serta mengubah kata ke bentuk dasar (stemming/lemmatization).

Tabel 1. Contoh Hasil Preprocessing

Sebelum	Sesudah (clean)
Great match last night!! 🔥 ⚽ https://t.co/xyz	great match last night

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 198-204

Sebelum	Sesudah (clean)
@user New movie trailer looks amazing!!! 😍	new movie trailer looks amazing
The president announced a new policy today #government	president announced new policy today
This new song is sooo good!!! 🎵💕	new song good

2.3 Ekstraksi Fitur TF-IDF

Teks yang telah dipreproses diubah menjadi vektor numerik menggunakan TF-IDF dengan pengaturan n-gram untuk menangkap frasa dua kata serta pembatasan jumlah fitur agar komputasi efisien.

Tabel 2. Parameter TF-IDF

Parameter	Nilai
max_features	5000
ngram_range	(1, 2)
use_idf	True
smooth_idf	True
sublinear_tf	True

2.4 Penentuan Jumlah Cluster

Jumlah cluster (K) ditentukan menggunakan Elbow Method (berdasarkan WCSS) dan Silhouette Score. Secara umum, titik elbow dan skor silhouette terbaik mengindikasikan K=3, sesuai fokus topik politik, olahraga, dan hiburan.

Tabel 3. WCSS (Elbow Method) untuk Beberapa Nilai K

K	WCSS
2	4781
3	3104
4	2950
5	2894
6	2866

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 198-204

2.5 Clustering dengan K-Means

Clustering dilakukan menggunakan K-Means dengan K=3. Algoritma menginisialisasi centroid, menghitung jarak tiap data ke centroid, mengelompokkan data ke centroid terdekat, lalu memperbarui centroid hingga konvergen.

Tabel 4. Parameter K-Means

Parameter	Nilai
n_clusters	3
init	k-means++
max_iter	300
n_init	10
random_state	42

2.6 Evaluasi

Evaluasi internal menggunakan Silhouette Score untuk menilai seberapa baik pemisahan cluster. Selain itu, dilakukan perbandingan dengan penelitian lama (baseline) yang menggunakan CountVectorizer dan preprocessing minimal, guna melihat dampak modifikasi metode.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Clustering (K = 3)

Model K-Means menghasilkan tiga cluster yang dapat diinterpretasikan sebagai politik, olahraga, dan hiburan berdasarkan kata dominan. Distribusi jumlah tweet per cluster ditunjukkan pada Tabel 5.

Tabel 5. Distribusi Tweet per Cluster dan Kata Dominan

Cluster	Jumlah Tweet	Kata Dominan (contoh)	Dugaan Topik
0	4.532	government, president, vote	Politik
1	3.912	match, team, win, score	Olahraga
2	4.001	movie, trailer, music, actor	Hiburan

3.2 Analisis TF-IDF

TF-IDF memberikan bobot lebih tinggi pada kata yang penting pada suatu topik dan menurunkan pengaruh kata umum yang tidak informatif. Contoh bobot kata dominan pada korpus ditunjukkan pada Tabel 6.

Tabel 6. Contoh Bobot TF-IDF Kata Dominan

Kata	Bobot TF-IDF
government	0.812

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 198-204

Kata	Bobot TF-IDF
match	0.745
movie	0.701
team	0.664
policy	0.603
trailer	0.577
win	0.554
music	0.531
protest	0.498
score	0.466

3.3 Evaluasi Silhouette Score

Nilai Silhouette Score penelitian baru sebesar 0,64 menunjukkan kualitas cluster yang baik dan pemisahan yang cukup jelas. Nilai ini meningkat dibanding penelitian lama (baseline) yang memperoleh 0,39, mengindikasikan bahwa kombinasi TF-IDF dan preprocessing yang lebih lengkap berdampak positif terhadap kualitas cluster.

Tabel 7. Perbandingan Penelitian Lama vs Penelitian Baru

Aspek	Penelitian Lama (Baseline)	Penelitian Baru (Modifikasi)
Representasi fitur	CountVectorizer	TF-IDF
Preprocessing	Minimal	Lengkap (cleaning + stopword + stemming)
Evaluasi	Terbatas	Elbow + Silhouette
Visualisasi	Tidak ada	PCA, Wordcloud
Nilai Silhouette	0,39	0,64
Interpretasi topik	Kurang jelas	Jelas dan terpisah

3.4 Visualisasi dan Interpretasi

Visualisasi PCA 2D digunakan untuk melihat pemisahan cluster secara geometris, sedangkan wordcloud menampilkan kata dominan tiap cluster. Secara umum, ketiga cluster menunjukkan pemisahan yang cukup jelas meskipun terdapat beberapa titik yang berdekatan karena kemiripan konteks antar-topik.

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

a. Kombinasi preprocessing komprehensif dan representasi TF-IDF secara signifikan meningkatkan kualitas clustering tweet dibandingkan pendekatan baseline.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 198-204

b. Nilai optimal K=3 sesuai dengan tiga topik utama (politik, olahraga, hiburan) dengan Silhouette Score 0,64 mengindikasikan kualitas cluster yang baik.

c. TF-IDF memberikan representasi fitur yang lebih efektif daripada CountVectorizer untuk data teks media sosial.

d. Visualisasi PCA dan WordCloud membantu interpretasi hasil clustering secara lebih intuitif.

e. Pendekatan yang diusulkan berpotensi untuk aplikasi praktis dalam pemantauan media sosial dan analisis isu publik.Saran pengembangan meliputi penggunaan dataset yang lebih besar dan multibahasa, penerapan metode clustering lain (mis. DBSCAN atau HDBSCAN), serta pemanfaatan embedding modern (Word2Vec/GloVe/BERT) untuk representasi yang lebih kontekstual.

4.2 Saran

a. Eksperimen dengan dataset yang lebih besar dan multibahasa untuk meningkatkan generalisasi model.

b. Penggunaan metode clustering alternatif seperti DBSCAN atau Spectral Clustering untuk perbandingan performa.

c. Implementasi teknik embedding modern seperti Word2Vec atau BERT untuk representasi teks yang lebih kontekstual.

d. Pengembangan dashboard interaktif untuk visualisasi dan analisis real-time.

e. Integrasi dengan analisis sentimen untuk pemahaman yang lebih komprehensif terhadap opini publik.

REFERENCES

Andi STMIK TIME, Juliandy, C., & STMIK TIME, D. (2022). Clustering Analysis of Tweets About COVID-19 Using the K-Means Algorithm. Sinkron: Jurnal dan Penelitian Teknik Informatika, 8(1). https://doi.org/10.33395/sinkron.v8i1.12145

Astutik, D. K., Indrasetianingsih, A., & Fitriani, F. (2022). Penerapan Text Mining pada Analisis Sentimen Pengguna Twitter Layanan Transportasi Online Menggunakan Metode DBSCAN dan K-Means. J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika, 15(1). https://doi.org/10.36456/jstat.vol15.no1.a5983

Br Sembiring, T. A., & Hasibuan, M. S. (2023). Text Clustering in Karo Language Using TF-IDF Weighting and K-Means Clustering. Jurnal Teknik Informatika (JUTIF), 4(5), 1257-1265. https://doi.org/10.52436/1.jutif.2023.4.5.1462

Chen, Z., Mi, C., Duo, S., He, J., & Zhou, Y. (2023). ClusTop: An unsupervised and integrated text clustering and topic extraction framework. arXiv preprint.

Darwis, M., Pranoto, G. T., Wicaksana, Y. E., & Yaddarabullah. (2022). Implementation of TF-IDF Algorithm and K-Means Clustering Method to Predict Words or Topics on Twitter. Jurnal Informatika dan Sains (JISA).

Fani, S. M., Imro'ah, N., & Aprizkiyandari, S. (2021). Penerapan Text Mining Data Tweet Tokopedia Menggunakan K-Means Clustering. BIMASTER: Buletin Ilmiah Matematika, Statistika dan Terapannya. https://doi.org/10.26418/bbimst.v13i3.77706

Kusumaningtyas, K., Habibi, M., Dwijayanti, I., & Sumiyarini, R. (2023). Tweet Analysis of Mental Illness Using K-Means Clustering and Support Vector Machine. Telematika: Jurnal Telematika dan Teknologi Informasi, 20(3), 295-308. https://doi.org/10.31315/telematika.v20i3.9820

Lengari, C. G., & Puspitasari, I. (2024). Identifying Twitter Topics Using K-Means Clustering and Association Rule Mining for Improved Insights. Indonesian Journal of Artificial Intelligence and Data Mining, 8(1), 67-75.

Remawati, D., Wijayanto, H., Utami, Y. R. W., & Raharja, B. D. (2023). Pengelompokkan Film Trending di YouTube Menggunakan TF-IDF dan K-Means Clustering. Jurnal Sistem Informasi Triguna Dharma (JURSI TGD). https://doi.org/10.53513/jursi.v4i1.10614

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 198-204

Santi, & Februariyanti, H. (2023). Implementation of Clustering on Tweet Uploading Side Effects of COVID-19 Post Vaccination Using K-Means Algorithm. Jurnal Teknik Informatika (JUTIF). https://doi.org/10.52436/1.jutif.2023.4.4.704

Surianto, D. F. (2023). Clustering Tweets Data on Twitter Social Media using K-Means Method. Journal of Security, Computer, Information, Embedded, Network, and Intelligence System, 1(2), 44-51. https://doi.org/10.61220/scientist.v1i2.20232

Wang, L., Gao, C., Wei, J., Ma, W., Liu, R., & Vosoughi, S. (2020). An Empirical Survey of Unsupervised Text Representation Methods on Twitter Data. arXiv preprint.

Wiguna, R. A. R., & Rifai, A. I. (2021). Analisis Text Clustering Masyarakat di Twitter Mengenai Omnibus Law Menggunakan Orange Data Mining. Journal of Information Systems and Informatics, 3(1), 1-12. https://doi.org/10.33557/journalisi.v3i1.78