Journal of Informatics and Computer Engineering Research (JICER) ISSN: 3089-6320
Vol. 2, No. 2, Desember 2025, hlm. 01-05

ANALISIS PENCARIAN TEKS EMOSI DALAM TWEET BERBAHASA INDONESIA MENGGUNAKAN TF-IDF DAN COSINE SIMILARITY

Adi Saputra[1], M Lazuardi Ferdilian[2], Hafiz Irsyad[3], Abdul Rahman[4]
[1] Prodi Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang, Palembang, Indonesia

Info Artikel

Riwayat artikel:
Received, (15 November 2025)
Revised, (1 Desember 2025)
Accepted, (2 Desember 2025)

Kata kunci:
Analisis emosi;
Twitter;
TF-IDF;
Cosine Similarity;
Bahasa Indonesia

ABSTRACT

Social media platforms such as *Twitter* present a wide range of emotional expressions from users in short and informal texts, which pose significant challenges for automated analysis. This study develops a search system for Indonesian-language tweets that identifies user emotions based on *semantic similarity* to a given text query. The method employs *Term Frequency-Inverse Document Frequency (TF-IDF)* for feature weighting and *Cosine Similarity* to measure textual similarity. *Preprocessing* stages including *normalization*, *tokenization*, *stopword removal*, and *stemming* are applied to enhance text representation accuracy. The system is tested using emotion-based queries and returns relevant tweets with high *semantic match scores*. Experimental results show that 50% of the top retrieved tweets match the expected emotional context. This approach proves effective in detecting emotions in short texts and offers potential for further development in *sentiment-driven opinion analysis* and *emotion-aware recommendation systems*.

ABSTRAK

Media sosial seperti *Twitter* menyajikan berbagai ekspresi emosional masyarakat dalam bentuk teks pendek dan informal, yang menantang untuk dianalisis secara otomatis. Penelitian ini mengembangkan sistem pencarian *tweet* berbahasa Indonesia yang mampu mengidentifikasi emosi pengguna berdasarkan kemiripan makna dengan masukan teks (*query*). Metode yang digunakan meliputi pembobotan kata menggunakan *Term Frequency-Inverse Document Frequency (TF-IDF)* dan pengukuran kemiripan menggunakan *Cosine Similarity*. Tahapan *pra-pemrosesan* seperti *normalisasi*, *tokenisasi*, penghapusan kata tidak penting (*stopword removal*), dan *stemming* dilakukan untuk meningkatkan akurasi representasi teks. Sistem diuji menggunakan *query* berbasis emosi dan menghasilkan *tweet* relevan dengan tingkat kecocokan *semantik* tinggi. Hasil pengujian menunjukkan bahwa 50% *tweet* teratas yang ditampilkan sesuai dengan konteks emosi yang diharapkan. Pendekatan ini efektif dalam mendeteksi emosi dalam teks pendek dan dapat dikembangkan lebih lanjut untuk aplikasi analisis opini dan sistem rekomendasi berbasis emosi.

Penulis Korespondensi:
Adi Saputra
Program Studi Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas MultiData Palembang, Jl. Rajawali No.14, Palembang 30113, Indonesia
Email: adisaputra_2226250036@mhs.mdp.ac.id

1. PENDAHULUAN

Media sosial telah menjadi sarana utama bagi masyarakat modern dalam menyampaikan opini, pandangan, serta emosi terhadap berbagai isu secara real-time [1]. Dari berbagai platform yang tersedia, Twitter menjadi salah satu yang paling menonjol karena formatnya yang ringkas namun padat informasi. Hal ini menjadikannya sebagai sumber data yang sangat potensial dalam analisis sentimen dan emosi [2].

Analisis emosi bertujuan untuk mengenali jenis emosi yang terkandung dalam teks, seperti marah, senang, sedih, atau takut, berdasarkan ekspresi yang dituliskan pengguna. Namun, karakteristik teks di Twitter yang cenderung pendek, informal, serta sering kali memuat singkatan, emotikon, dan simbol nonbaku, menimbulkan tantangan dalam mengekstraksi informasi emosional secara akurat [3]. Oleh karena itu, dibutuhkan pendekatan yang mampu menangkap makna semantik dari teks pendek dan tidak terstruktur tersebut secara efektif.

Salah satu pendekatan representasi teks yang paling umum digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF), yaitu metode yang memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya dalam suatu dokumen dibandingkan dengan seluruh korpus [4]. TF-IDF mengubah teks menjadi vektor numerik, yang kemudian dapat dianalisis menggunakan metode Cosine Similarity. Cosine Similarity mengukur kesamaan antar dua teks berdasarkan sudut di antara dua vektor dalam ruang multidimensi [5].*TF-IDF* dan *Cosine Similarity* dalam menganalisis konten media sosial. Wahid dan Azhari [6] menggunakannya dalam peringkasan sentimen secara ekstraktif terhadap *tweet*, menunjukkan bahwa pendekatan ini mampu mempertahankan inti emosi dari teks asli. Nugraha dan Sebastian [7] menerapkan metode serupa untuk mengidentifikasi tren akun media sosial berdasarkan kesamaan konten yang dipublikasikan. Tiwari et al. [8] mengimplementasikan pendekatan *TF-IDF* secara skala besar untuk analisis sentimen terhadap data Twitter dalam bahasa Inggris, dan menunjukkan kinerja yang cukup stabil dalam berbagai topik.

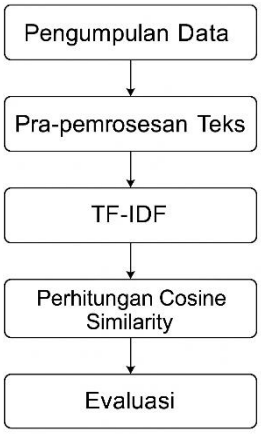
Berbagai penelitian terdahulu menunjukkan bahwa kombinasi metode TF-IDF dan Cosine Similarity efektif dalam menganalisis teks pendek di media sosial. Wahid dan Azhari [6], misalnya, menerapkan pendekatan ini bersama SentiStrength untuk melakukan peringkasan sentimen pada tweet berbahasa Indonesia dan mencapai relevansi emosi sekitar 60%. Nugraha dan Sebastian [7] juga menggunakan metode serupa untuk mengidentifikasi tren konten pada akun Twitter melalui kemiripan antar tweet. Sedangkan Tiwari et al. [8] mengimplementasikan analisis sentimen berbasis TF-IDF dalam skala besar dan melaporkan kinerja yang stabil di berbagai topik.

Sementara itu, pendekatan berbasis pembelajaran mendalam menunjukkan akurasi yang lebih tinggi. Hilmiaji et al. [9] menggunakan CNN untuk klasifikasi emosi dan mencapai akurasi hingga 82%, sedangkan Alfarizi et al. [10] memanfaatkan LSTM dengan dukungan TF-IDF dan menghasilkan akurasi sebesar 85%. Meski demikian, metode klasik seperti TF-IDF dan Cosine Similarity tetap relevan karena kesederhanaannya dan kemampuannya dalam menangani teks pendek secara efisien.

Jika dibandingkan dengan studi-studi sebelumnya, sistem yang dikembangkan dalam penelitian ini mampu mencapai tingkat relevansi sebesar 50% dalam menampilkan tweet yang sesuai dengan emosi yang diminta pengguna. Hasil ini sedikit lebih rendah dibandingkan penelitian oleh Wahid dan Azhari [6] yang menggunakan pendekatan hybrid TF-IDF, Cosine Similarity, dan SentiStrength, dengan relevansi sekitar 60% terhadap konteks emosional. Sementara itu, pendekatan berbasis pembelajaran mendalam seperti CNN dan LSTM menunjukkan performa lebih tinggi, masing-masing mencapai akurasi hingga 82% dan 85% [9][10]. Meskipun demikian, metode yang digunakan dalam penelitian ini tetap memiliki keunggulan dari sisi kesederhanaan implementasi, efisiensi komputasi, serta fleksibilitas terhadap berbagai jenis input tanpa ketergantungan pada struktur model yang kompleks atau sumber daya eksternal seperti leksikon emosi. Dengan demikian, meskipun tingkat akurasi belum menyamai metode deep learning, pendekatan ini dapat menjadi alternatif ringan dan praktis dalam pencarian informasi bermuatan emosi pada teks pendek seperti tweet.

2. METODE

Penelitian ini menggunakan pendekatan eksperimental untuk mengembangkan sistem pencarian dan analisis emosi pada *tweet* berbahasa Indonesia. Metodologi yang diterapkan terdiri dari beberapa tahapan, yaitu: pengumpulan data, *pra-pemrosesan* teks, representasi data dengan *TF-IDF*, pengukuran kemiripan menggunakan *Cosine Similarity*, dan evaluasi hasil dapat dilihat pada Gambar 1.



Gambar 1. Flowchart Metode

2.1 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini merupakan bagian dari benchmark IndoNLU, yaitu EmoT (Emotion Twitter), yang pertama kali dikembangkan oleh Saputri et al. [11] dan kemudian diintegrasikan dalam benchmark IndoNLU oleh Wilie et al. [12]. Dataset ini terdiri dari sekitar 4.000 tweet berbahasa Indonesia yang telah diklasifikasikan ke dalam lima kategori emosi utama, yaitu marah, takut, senang, cinta, dan sedih. Dataset ini tersedia secara terbuka dengan lisensi MIT, dan telah banyak digunakan dalam penelitian pengolahan bahasa alami (NLP) berbahasa Indonesia, khususnya dalam tugas klasifikasi emosi dan analisis sentimen.

2.2 Pra-pemrosesan Teks

Sebelum dianalisis lebih lanjut, *tweet* yang terkumpul diproses melalui beberapa tahap pembersihan dan *normalisasi*. Tahapan tersebut meliput:

1. Lowercase: Mengubah seluruh teks menjadi huruf kecil.
2. Tokenisasi: Memecah kalimat menjadi unit kata.
3. Penghapusan tanda baca dan simbol: Menghapus karakter non-alfabet.
4. Stopword Removal: Menghapus kata-kata umum yang tidak berkontribusi terhadap makna (contoh: "dan", "yang", "di").
5. Stemming: Mengembalikan kata ke bentuk dasar (contoh: "menangis" → "tangis").

Langkah ini bertujuan mengurangi noise pada data dan meningkatkan kualitas hasil ekstraksi fitur.

2.3 Representasi Menggunakan TF-IDF

Setiap tweet diubah menjadi vektor fitur menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Rumus TF-IDF dapat dilihat pada persamaan 1:

$$TF - IDF(t,d) = tf(t,d) \times \log\left(\frac{N}{df(t)}\right) \quad (1)$$

Dengan:

1. tf(t,d): frekuensi kata ttt dalam dokumen d
2. df(t)df(t)df(t): jumlah dokumen yang mengandung kata ttt
3. N: total jumlah dokumen dalam korpus

TF-IDF memberi bobot tinggi untuk kata yang sering muncul di suatu dokumen namun jarang ditemukan di dokumen lain.

2.4 Penghitungan Kemiripan dengan Cosine Similarity

Pengguna dapat memasukkan kalimat pertanyaan (*query*) yang juga melalui proses pembersihan dan *stemming*, lalu diubah menjadi *vektor* menggunakan *TF-IDF*. Kemudian dilakukan penghitungan *Cosine Similarity* antara vektor *query* dan seluruh *tweet* dalam *dataset* menggunakan persamaan 2.

$$simirality(A,B) = \frac{A.B}{||A||||B||} \quad (2)$$

Nilai kemiripan ini digunakan untuk mengurutkan *tweet* dari yang paling relevan hingga yang paling tidak relevan terhadap *query*. Sistem menampilkan 10 *tweet* dengan nilai kemiripan tertinggi sebagai hasil pencarian.

3. HASIL DAN PEMBAHASAN

Penelitian ini menguji efektivitas sistem pencarian *tweet* berdasarkan kemiripan *semantik* antara *query* dan *tweet* dalam *korpus*. *Query* uji yang digunakan adalah "saya merasa takut akan penyakit"**,** yang mewakili emosi *fear* (takut). Sistem menghasilkan daftar *tweet* yang paling mirip berdasarkan nilai *Cosine Similarity* antara *vektor TF-IDF* dari *query* dan setiap *tweet* yang telah melalui proses *pra-pemrosesan*.

3.1 Hasil Pencarian

Tabel 1. Hasil pencarian 10 tweet berdasarkan kemiripan tertinggi terhadap query "saya merasa takut akan penyakit"

No	*Tweet* Asli	Label Emosi	*Similarity*
1	jadi takut buat nikah...buat punya anak...takut kalo saya nikah...	fear	0.3065
2	dek, saya cuma takut, perasaan saya akan melebihi batas...	fear	0.3051
3	seluruh hidup saya...saya dedikasikan dan saya peruntukkan...	love	0.2884
4	yakinlah ada sesuatu yang menantimu selepas banyak kesabaran...	happy	0.2355
5	saya dok, enggan menikah karena lebih mengutamakan karir...	fear	0.1872
6	saya ajukan PayLater, namun ditolak dengan alasan...	anger	0.1831

7	bangun bangun ngeliat post an jay kek gitu...	sadness	0.1775
8	Jika ingat diriMu, hati ini plong gitu ku akan...	love	0.1641
9	jujur iya aku takut, soalnya aku takut keciduk...	fear	0.1640
10	minum obat kamu jangan sampai sakit ye...	fear	0.1572

3.2 Interpretasi Hasil

Dari 10 *tweet* dengan skor kemiripan tertinggi, sebanyak 5 *tweet* (50%) terklasifikasi sebagai *fear*, sesuai dengan emosi yang diwakili oleh *query*. Temuan ini menunjukkan bahwa sistem mampu mengenali kesamaan *semantik* yang relevan untuk mendeteksi emosi serupa. Beberapa *tweet* yang tidak berlabel *fear* (misalnya *love* atau *happy*) juga menunjukkan konteks emosional yang secara tidak langsung berkaitan, yang mengindikasikan bahwa *Cosine Similarity* cukup sensitif terhadap makna kontekstual dalam teks pendek.

3.3 Analisis Hasil

Kehadiran *tweet* dengan label *love* dan *happy* dalam hasil pencarian mengindikasikan bahwa model terkadang menangkap ekspresi yang bersifat konotatif atau ambigu secara emosional. Meskipun demikian, nilai *similarity* tertinggi tetap didominasi oleh *tweet* dengan emosi *fear*, yang menunjukkan keberhasilan sistem dalam memetakan konteks emosi utama dari *query* ke dalam ruang *vektor*. Hal ini menegaskan bahwa pendekatan berbasis *TF-IDF* dan *Cosine Similarity* cukup efektif dalam mengidentifikasi kemiripan makna emosional dalam teks pendek.

3.4 Kaitan dengan Penelitian Sebelumnya

Hasil ini mendukung temuan Wahid dan Azhari [6], serta Nugraha dan Sebastian [7], yang menyatakan bahwa kombinasi *TF-IDF* dan *Cosine Similarity* efektif dalam menilai kemiripan makna pada teks pendek seperti *tweet*. Penelitian ini memperluas konteks tersebut dengan menerapkannya pada identifikasi emosi, bukan sekadar pencocokan topik, sehingga menunjukkan potensi metode ini dalam analisis makna yang lebih mendalam dan kontekstual.

4. EVALUASI HASIL

Kombinasi TF-IDF dan Cosine Similarity telah berhasil digunakan sebagai dasar sistem pencarian berbasis makna dan emosi. TF-IDF mampu memberi bobot penting pada kata-kata yang khas terhadap suatu emosi, Ssedangkan Cosine Similarity menentukan kedekatan antara teks berdasarkan kemiripan distribusi kata.
Namun demikian, terdapat beberapa batasan:

1. **Ketidaksensitifan semantik:** TF-IDF tidak mempertimbangkan sinonim atau hubungan antar kata. Misalnya, "cemas" dan "takut" tidak dianggap serupa jika tidak terdapat dalam dokumen yang sama.
2. **Keterbatasan konteks:** Cosine Similarity hanya melihat pola numerik (vektor) tanpa memahami konteks makna sebenarnya.
3. **Akurasi terbatas:** Dari 10 tweet teratas, hanya 50% yang secara eksplisit sesuai emosi, menunjukkan bahwa sistem masih dapat ditingkatkan.

Evaluasi kuantitatif lebih lanjut menggunakan metrik Information Retrieval seperti Precision@10, Recall@10, dan F1-Score disarankan untuk penelitian selanjutnya. Selain itu, sistem ini dapat dibandingkan dengan baseline berbasis machine learning untuk mengukur peningkatan performa.

5. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem pencarian tweet berbahasa Indonesia yang mampu mengidentifikasi kemiripan semantik antara input query pengguna dengan tweet dalam korpus menggunakan pendekatan TF-IDF dan Cosine Similarity. Proses yang dilakukan meliputi pengumpulan data tweet berbahasa Indonesia yang mengandung ekspresi emosi, pra-pemrosesan teks seperti normalisasi, tokenisasi, penghapusan stopword, dan stemming, serta representasi teks ke dalam vektor numerik menggunakan TF-IDF. Selanjutnya, pengukuran kemiripan antara query dan tweet dilakukan dengan Cosine Similarity untuk menghasilkan daftar tweet yang paling relevan secara makna dan muatan emosional.

Hasil pengujian menunjukkan bahwa dari 10 tweet dengan nilai kemiripan tertinggi, sebanyak 5 tweet atau 50% terklasifikasi sesuai dengan emosi yang diwakili oleh query, yakni emosi takut (fear). Temuan ini mengindikasikan bahwa sistem mampu mengenali kesamaan makna yang terkait dengan konteks emosional secara cukup akurat, meskipun beberapa tweet yang tidak memiliki label emosi yang sama tetap memiliki relevansi kontekstual. Dengan demikian, pendekatan yang digunakan terbukti efektif dalam mendeteksi dan mencocokkan muatan emosi dalam teks pendek di media sosial, dan dapat dikembangkan lebih lanjut untuk aplikasi analisis opini dan sistem rekomendasi berbasis emosi.

6. REFERENSI

[1] Addiga, A., & Bagui, S. (2022). Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency. *Journal of Computer and Communications*, 10, 117-128.

[2] A. Mee, E. Homapour, F. Chiclana, and O. Engel, "Sentiment Analysis Using TF-IDF Weighting of UK MPs' Tweets on Brexit," *Knowledge-Based Systems*, vol. 228, p. 107238, 2021.

[3] S. Rahmawati and M. Habibi, "Public Sentiments Analysis about Indonesian Social Insurance Administration Organization on Twitter," *International Journal on Informatics for Development*, vol. 9, no. 2, pp. 87–93, 2020.

[4] M. T. Haque Khan and M. T. Islam, "A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data," *arXiv preprint*, arXiv:2110.00859, 2021. [Online]. Available: https://arxiv.org/abs/2110.00859

[5] GeeksforGeeks, "Cosine Similarity," 2025. [Online]. Available:https://www.geeksforgeeks.org/cosine-similarity/

[6] D. H. Wahid and S. N. Azhari, "Peringkasan Sentimen Ekstraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *Indonesian Journal of Computing and Cybernetics Systems*, vol. 10, no. 2, pp. 207–218, 2016.

[7] K. A. Nugraha and D. Sebastian, "Analisis Trend Akun Media Sosial Twitter Menggunakan TF-IDF dan Cosine Similarity," in *Seminar Nasional ReTII*, pp. 103–110, 2018.

[8] S. Tiwari, A. Verma, P. Garg, and D. Bansal, "Social Media Sentiment Analysis on Twitter Datasets," in *Proc. 6th Int. Conf. on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2020, pp. 355–359.

[9] N. Hilmiaji, K. M. Lhaksmana, dan M. D. Purbolaksono, "Identifying Emotion on Indonesian Tweets using Convolutional Neural Networks," *2021 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, pp. 81–86, 2021. doi: 10.1109/ICoDSA53046.2021.9568793.

[10] M. I. Alfarizi, L. Syafa'ah, dan M. Lestandy, "Emotional Text Classification Using TF-IDF and LSTM," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, pp. 353–359, 2021. doi: 10.29207/resti.v5i2.3107.

[11] G. D. N. Saputri, S. R. S. Wibowo, and M. Adriani, "Emotion Classification on Indonesian Twitter Dataset," in *2018 Int. Conf. on Asian Language Processing (IALP)*, Yogyakarta, Indonesia, 2018, pp. 42–45. doi: 10.1109/IALP.2018.8629155.

[12] B. Wilie, K. A. Vincentio, G. I. Winata, S. Cahyawijaya, Z. Lim, S. Soleman, R. Mahendra, P. Fung, and A. Purwarianti, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7626–7643.