Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081

Prediksi Insomnia Berdasarkan Aktivitas Pengguna Twitter Menggunakan Natural Language Processing dan Machine Learning

Trisna[1,*], Asti Herliana[2]

[1] Fakultas Teknik Informasi, Teknik Informatika, Universitas Adhirajasa Reswara Sanjaya, Bandung, Indonesia
[2] Fakultas Teknik Informasi, Sistem Informasi, Universitas Adhirajasa Reswara Sanjaya, Bandung, Indonesia
Email: [1,*]hadiyantitrisna@gmail.com, [2]asti@ars.ac.id
Email Penulis Korespondensi: hadiyantitrisna@gmail.com
Submitted: 23/07/2025; Accepted: 31/08/2025; Published: 31/08/2025

Abstrak–Insomnia adalah gangguan tidur yang banyak dialami masyarakat dan berdampak besar pada kesehatan fisik dan mental serta produktivitas. Namun, deteksi dini insomnia masih menjadi tantangan karena gejalanya sulit teridentifikasi secara langsung. Penelitian ini memanfaatkan data historis sebanyak 13.950 tweet dari 4.286 akun Twitter (1 Januari–30 April 2025) untuk memprediksi potensi insomnia menggunakan metode Natural Language Processing (NLP) dan machine learning. Label insomnia ditentukan melalui pendekatan *keyword-based* yang diverifikasi pakar, kemudian melalui tahapan preprocessing, analisis temporal, dan analisis sentimen. Dua model klasifikasi digunakan, yaitu Support Vector Machine (SVM) yang unggul dalam memisahkan kelas pada data berdimensi tinggi, dan Long Short-Term Memory (LSTM) yang unggul dalam menangkap pola berurutan serta konteks temporal. Hasil awal menunjukkan SVM memiliki akurasi 89% dan unggul pada kelas *non-insomnia* (precision 0.80, recall 0.97) namun kurang optimal pada *insomnia* (precision 0.92, recall 0.82), sedangkan LSTM memiliki akurasi 90% dan lebih baik pada *insomnia* (precision 0.98, recall 0.86) namun sedikit menurun pada *non-insomnia* (precision 0.81, recall 0.96). Oleh karena masing-masing model memiliki kekuatan berbeda, keduanya digabungkan dengan metode average probabilistic ensemble yang menghasilkan akurasi 92% dengan peningkatan seimbang di kedua kelas (non-insomnia: precision 0.82, recall 0.99; insomnia: precision 1.00, recall 0.88), sehingga lebih andal dibandingkan model tunggal dalam mendeteksi potensi insomnia.

Kata Kunci: Ensemble Learning; Insomnia; LSTM; SVM; Twitter

Abstract–Insomnia is a sleep disorder that is widely experienced by the public and has a significant impact on physical and mental health, as well as productivity. However, early detection of insomnia remains a challenge because its symptoms are difficult to identify directly. This study uses historical data of 13,950 tweets from 4,286 Twitter accounts (January 1–April 30, 2025) to predict potential insomnia using Natural Language Processing (NLP) and machine learning methods. Insomnia labels are determined through an expert-verified keyword-based approach, followed by preprocessing, temporal analysis, and sentiment analysis. Two classification models are used: Support Vector Machine (SVM), which excels at separating classes in high-dimensional data, and Long Short-Term Memory (LSTM), which excels at capturing sequential patterns and temporal context. Preliminary results showed that SVM had 89% accuracy and was superior in the non-insomnia class (precision 0.80, recall 0.97) but suboptimal in insomnia (precision 0.92, recall 0.82), while LSTM had 90% accuracy and was better in insomnia (precision 0.98, recall 0.86) but slightly inferior in non-insomnia (precision 0.81, recall 0.96). Since each model had different strengths, they were combined with a probabilistic ensemble averaging method which resulted in 92% accuracy with balanced improvements in both classes (non-insomnia: precision 0.82, recall 0.99; insomnia: precision 1.00, recall 0.88), making it more reliable than a single model in detecting potential insomnia.

Keywords: Ensemble Learning; Insomnia; LSTM; SVM; Twitter

1. PENDAHULUAN

Insomnia merupakan salah satu gangguan tidur yang kian meningkat di era digital, terutama disebabkan oleh tingginya penggunaan media sosial pada malam hari. Perubahan pola hidup, paparan cahaya dari layar gawai, serta keterlibatan emosional dalam aktivitas daring menjadi faktor yang turut memperburuk kualitas tidur masyarakat [1]. Di kawasan Asia Tenggara, prevalensi insomnia dilaporkan mencapai 67% pada kelompok remaja, sementara di Indonesia, lebih dari 28 juta penduduk diperkirakan mengalami gejala serupa [2]. Berdasarkan temuan data tersebut gangguan ini tidak hanya berdampak pada penurunan produktivitas, tetapi juga meningkatkan risiko gangguan kesehatan mental serta beban ekonomi secara luas [3], [4].

Twitter, sebagai salah satu platform media sosial berbasis teks, menawarkan sumber data yang menjanjikan untuk menelusuri ekspresi psikologis pengguna secara *real-time* [5]. Berbeda dari platform lain, Twitter memungkinkan pengguna mengekspresikan perasaan, aktivitas, dan pengalaman pribadi mereka dalam bentuk teks singkat yang terbuka untuk publik. Fitur seperti *hashtag*, *mention*, serta tingginya aktivitas unggahan pada malam hingga dini hari menjadikan Twitter sebagai representasi digital yang potensial dalam mendeteksi gangguan tidur [6].

Cuitan yang diunggah pada waktu tersebut sering mencerminkan kondisi emosional dan pola aktivitas yang berkaitan dengan insomnia. Untuk menganalisis hal ini, media sosial seperti Twitter dapat dimanfaatkan melalui analisis teks cuitan yang mencakup perasaan, aktivitas, dan waktu unggahan [7]. Namun, karena data yang dihasilkan tidak terstruktur, diperlukan teknologi seperti *Natural Language Processing* (NLP) untuk

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081

mengolahnya secara efektif [8]. Salah satu metode NLP yang relevan adalah analisis sentimen, yang digunakan untuk mengidentifikasi emosi seperti kecemasan atau kegelisahan yang sering dikaitkan dengan insomnia [9].

Sejumlah studi terdahulu mengenai analisis sentimen telah mengeksplorasi hubungan antara penggunaan media sosial dan gangguan tidur, khususnya insomnia**.** Pirdehghan et al. (2021) meneliti hubungan antara durasi penggunaan media sosial dan kualitas tidur pada remaja. Menggunakan pendekatan kuantitatif melalui kuesioner**,** mereka menemukan korelasi negatif yang signifikan, yaitu semakin lama seseorang mengakses media sosial, semakin buruk kualitas tidurnya. Hasil ini mengindikasikan adanya potensi risiko psikologis yang perlu diidentifikasi lebih lanjut melalui indikator perilaku digital.

Sakib et al. (2021) mengembangkan model prediksi insomnia berbasis psikolinguistik menggunakan data cuitan Twitter**.** Penelitian ini memanfaatkan fitur bahasa seperti penggunaan kata emosional dan waktu unggahan, serta menerapkan machine learning untuk klasifikasi. Hasilnya, model mereka mencapai akurasi 78,8%**,** namun pendekatan temporal dalam analisis data belum sepenuhnya dioptimalkan.

Adiwibawa et al. (2023) meneliti hubungan antara intensitas penggunaan media sosial dan tingkat insomnia pada mahasiswa Indonesia. Dengan metode survei dan analisis regresi, penelitian ini menunjukkan adanya hubungan signifikan antara frekuensi penggunaan media sosial pada malam hari dan peningkatan gejala insomnia. Meski demikian, studi ini terbatas pada pendekatan kuantitatif tanpa eksplorasi langsung data media sosial.

Mengingat pentingnya konteks waktu dalam analisis gangguan tidur, diperlukan algoritma yang mampu menangkap pola temporal dalam data sekuensial [10]. *Long Short-Term Memory* (LSTM) merupakan salah satu arsitektur *deep learning* yang mampu mempertahankan konteks waktu dan mengenali pola jangka panjang [11]. Selain itu, Gleeson et al. (2020) membuktikan bahwa Long Short-Term Memory (LSTM) sebuah arsitektur *deep learning* yang mampu mempertahankan konteks waktu dan mengenali pola jangka panjang—lebih akurat dibanding model konvensional dalam mendeteksi gangguan tidur dari data perangkat *wearable*. Hal ini menunjukkan relevansi LSTM untuk menganalisis data Twitter yang bersifat kronologis.

Untuk meningkatkan performa klasifikasi gejala insomnia, algoritma tambahan seperti *Support Vector Machine* (SVM) dapat digunakan. SVM dikenal efektif dalam mengolah data berlabel, terutama pada dataset berukuran kecil hingga menengah dengan fitur kompleks [12]. Studi Rani et al. (2022) menunjukkan bahwa Support Vector Machine (SVM) efektif untuk klasifikasi insomnia akut dan kronis dengan akurasi 81% dari data *actigraphy*. SVM unggul dalam mengolah data berlabel pada dataset berukuran kecil hingga menengah dengan fitur kompleks, sehingga berpotensi digunakan dalam deteksi insomnia berbasis fitur linguistik dan temporal dari media sosial.

Studi-studi sebelumnya telah memberikan kontribusi penting dalam memahami hubungan antara media sosial dan gangguan tidur. Namun, sebagian besar penelitian tersebut masih memiliki keterbatasan dalam menggabungkan konten linguistik, konteks temporal, dan algoritma prediktif secara terpadu. Misalnya, model prediksi insomnia berbasis psikolinguistik oleh Sakib et al. (2021) hanya mencapai akurasi 78,8%**,** sementara pendekatan berbasis SVM oleh Rani et al. (2022) untuk deteksi insomnia dari data actigraphy menghasilkan akurasi 81%. Hingga saat ini, belum ada penelitian berbasis data Twitter yang secara simultan mengintegrasikan analisis teks, waktu unggahan, dan kombinasi model pembelajaran mesin untuk mencapai akurasi di atas 90%. Penelitian ini bertujuan untuk mengisi celah tersebut dengan mengembangkan model prediksi insomnia berbasis NLP**,** analisis temporal**,** dan ensemble learning dari SVM dan LSTM**,** yang diharapkan mampu melampaui capaian akurasi sebelumnya dan memberikan model prediksi yang lebih andal untuk deteksi dini insomnia.

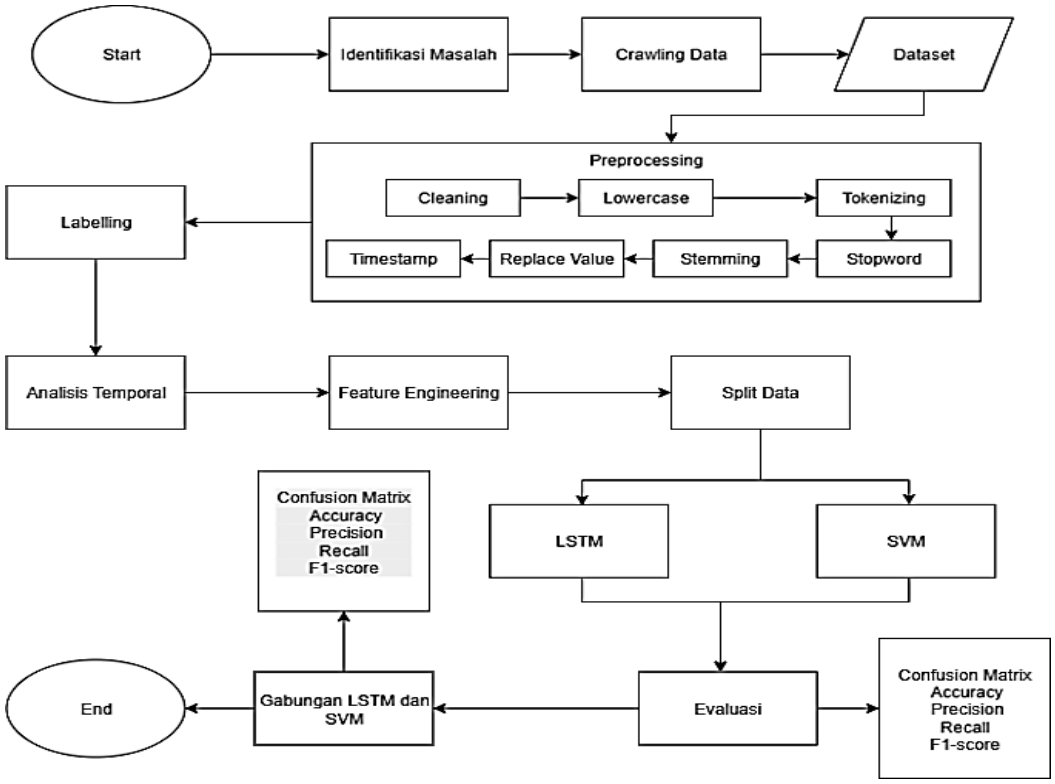
2. METODOLOGI PENELITIAN

2.1 Alur Penelitian

Penelitian ini bertujuan memprediksi potensi insomnia berdasarkan aktivitas pengguna di media sosial Twitter melalui analisis data teks yang diperoleh dari tweet. Metode yang digunakan meliputi pengumpulan data, *preprocessing*, pelabelan, ekstraksi fitur, pemodelan menggunakan *machine learning*, dan evaluasi model.

Pada Gambar 1 menunjukkan alur kerja penelitian berbasis analisis data teks menggunakan metode Long Short-Term Memory (LSTM) dan Support Vector Machine (SVM). Proses dimulai dari tahap identifikasi masalah, kemudian dilanjutkan dengan crawling data untuk mengumpulkan dataset. Setelah data terkumpul, tahap preprocessing dilakukan untuk membersihkan dan menyiapkan data melalui beberapa langkah seperti *cleaning*, *lowercase*, *tokenizing*, *stopword removal*, *stemming*, *replace value*, dan *timestamp* agar data siap digunakan. Selanjutnya dilakukan labelling untuk memberi tanda pada data sesuai kategori yang diinginkan, diikuti oleh analisis temporal guna melihat pola data berdasarkan waktu. Tahap berikutnya adalah feature engineering, yang berfungsi mengekstraksi fitur penting dari data sebelum dilakukan split data untuk membagi dataset menjadi data latih dan data uji. Data kemudian diproses menggunakan dua model yaitu LSTM dan SVM. Hasil dari kedua model ini dievaluasi menggunakan Confusion Matrix dengan metrik evaluasi *accuracy*, *precision*, *recall*, dan *F1-score*. Selanjutnya, hasil dari LSTM dan SVM digabungkan untuk memperoleh performa model terbaik. Proses ini diakhiri dengan analisis keseluruhan yang menunjukkan kinerja model gabungan dalam mengklasifikasikan data teks secara akurat dan efisien.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081



Gambar 1. Alur Penelitian

2.2 Pengumpulan Data

Data dikumpulkan menggunakan metode *crawling* dari platform Twitter melalui Twitter API dengan bantuan pustaka Python seperti Tweepy dan Twint. Periode pengambilan data berlangsung dari 1 Januari hingga 30 April 2025 [13]. Data yang dikumpulkan adalah tweet yang mengandung kata kunci terkait insomnia, seperti "insomnia" dan "susah tidur". Data disimpan dalam format CSV berisi teks tweet dan metadata terkait waktu unggah [14].

2.3 Preprocessing Data

Data yang diperoleh diproses untuk meningkatkan kualitas melalui beberapa tahapan *preprocessing* teks. Proses *cleaning* dilakukan dengan menghapus elemen-elemen yang tidak relevan seperti URL, mention, hashtag, angka, dan karakter khusus yang tidak diperlukan [15]. Teks dinormalisasi menggunakan *lowercase* dengan mengubah seluruh huruf menjadi format kecil agar konsisten [16]. Proses *tokenization* memecah kalimat menjadi token kata menggunakan fungsi *word_tokenize* dari pustaka NLTK [14], [17]. Kata-kata umum yang kurang bermakna dihapus melalui *stopword removal* berdasarkan daftar *stopword* bahasa Indonesia dan Inggris dari NLTK [14]. Kata-kata tersisa direduksi ke bentuk dasar dengan teknik *stemming* menggunakan pustaka Sastrawi [14]. Nilai string kosong dalam data diganti dengan nilai terdekat di atasnya untuk mencegah gangguan pada analisis. Format waktu unggahan tweet dikonversi ke dalam *datetime* menggunakan pustaka pandas guna memastikan informasi temporal dapat dianalisis secara sistematis dan akurat.

2.4 Pelabelan

Tweet diberi label berdasarkan dua kriteria utama. Pertama, berdasarkan keberadaan kata kunci "insomnia" dalam isi tweet; jika tweet mengandung kata kunci tersebut, maka diberi label "Ya", sedangkan jika tidak mengandung, diberi label "Tidak". Kedua, berdasarkan waktu unggah tweet, yang dikategorikan ke dalam lima rentang waktu, yaitu Pagi, Siang, Sore, Malam, dan Larut Malam (antara pukul 22:00 hingga 05:00). Tweet yang diunggah pada rentang waktu Larut Malam diberi label "Ya" untuk menandai potensi adanya indikasi insomnia, sementara tweet pada rentang waktu lainnya diberi label "Tidak".

2.5 Analisi Temporal

Analisis temporal dilakukan untuk mengidentifikasi pola aktivitas pengguna berdasarkan jam dan hari unggahan. Aktivitas di malam hari (22:00–04:00) dianalisis lebih lanjut sebagai periode kritis potensial insomnia. Visualisasi heatmap dan pengelompokan berdasarkan waktu serta uji *chi-square* dilakukan untuk melihat hubungan signifikan antara waktu unggah dan potensi insomnia.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081

2.6 Ekstrasi Fitur

Ekstraksi fitur teks menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) dengan memilih 3.000 kata dengan bobot tertinggi [18]. Fitur temporal seperti jam, hari, dan bulan dari waktu unggah juga dimasukkan. Untuk model LSTM, teks ditokenisasi menggunakan Tokenizer Keras dan diproses menjadi urutan angka yang dipadatkan (padding) agar panjang data seragam.

2.7 Pembagian Data

Dataset yang terdiri dari 13.950 sampel dibagi menjadi data pelatihan sebesar 80% (11.160 sampel) dan data pengujian sebesar 20% (2.790 sampel) menggunakan fungsi *train_test_split* dari pustaka *scikit-learn* dengan stratifikasi label. Stratifikasi dilakukan untuk memastikan distribusi kelas tetap seimbang pada kedua set data, sehingga representatif dalam proses pelatihan maupun pengujian [19], [20].

2.8 Model

Dalam penelitian ini digunakan dua model *machine learning* untuk klasifikasi tweet insomnia. Model pertama adalah *Support Vector Machine* (SVM) dengan *kernel linear* [21] yang memanfaatkan fitur gabungan berupa TF-IDF dan fitur temporal [18]. Model kedua menggunakan *Long Short-Term Memory* (LSTM) yang mampu menangkap konteks urutan kata dalam tweet [22]. Model LSTM ini terdiri dari *embedding layer* berdimensi 100 [23], diikuti oleh LSTM layer dengan 128 unit dan dropout sebesar 0,2, serta *fully connected layer* yang menggunakan aktivasi sigmoid [24]. Model LSTM dikompilasi dengan fungsi *loss binary crossentropy* dan *optimizer Adam* dengan learning rate 0,001. Untuk menghindari *overfitting*, digunakan teknik *early stopping* selama pelatihan yang berlangsung maksimal 10 epoch dengan batch size 64 [25].

2.9 Evaluasi

Evaluasi model menggunakan *confusion matrix*, accuracy, *precision*, *recall*, dan *F1-score*. *Confusion matrix* menjelaskan jumlah *True Positive*, *True Negative*, *False Positive*, dan *False Negative* [13], [14]. Model LSTM juga dievaluasi menggunakan loss dan accuracy pada data testing.

2.10 Model Gabungan

Prediksi probabilitas dari SVM dan LSTM digabungkan menggunakan rata-rata probabilitas untuk menentukan label akhir dengan threshold 0.5. Model *ensemble* dievaluasi menggunakan metrik yang sama untuk meningkatkan akurasi dan robustness prediksi insomnia.

3. HASIL DAN PEMBAHASAN

3.1 Scrapping Data

Pengumpulan data dilakukan melalui platform Google Colab dengan memanfaatkan teknik scraping dan bahasa pemrograman Python. Dalam proses ini, digunakan pustaka *Twitter API* untuk mengekstrak tweet dari pengguna yang berkaitan dengan topik insomnia. Data dikumpulkan selama periode empat bulan, mulai dari 1 Januari hingga 30 April 2025, dengan total sebanyak 13.950 baris data berhasil diperoleh. Seperti yang ditampilkan pada Tabel 1, data mentah tersebut kemudian diseleksi dengan hanya mempertahankan atribut-atribut yang relevan dengan tujuan penelitian, yaitu isi tweet (*full_text*), waktu unggahan (*created_at*), serta ID dan nama pengguna (*user_id_str* dan *username*). Dataset yang telah diseleksi ini kemudian disimpan dengan nama dataset_insomnia dalam format file *.csv*.

Tabel 1. Sampel dataset Insomnia

No	conversation_id_str	created_at	full_text	User_id_str	username
1	1899971867433450000	Wed Mar 12 23:52:47 +0000 2025	Jeleknya aku adalah ketika stress datang insomnia kambuh	1174255927836450000	TasyaKamillahh
2	1899965465197530000	Wed Mar 12 23:27:21 +0000 2025	yg insomnia mana suaranyaaaaas hadirrrr	1656376221553560000	mbaktata_
…	….	…	...	…	…
13950	1900118496056640000	Thu Mar 13 09:35:26 +0000 2025	@tattyhassan insomnia teruk badan lesu sampai gigil. berbuka je muntah hehe	119647297	k1m1eee

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081

3.2 Preprocessing

Tahap *preprocessing* dilakukan untuk membersihkan dan mempersiapkan data hasil *crawling* dari Twitter agar siap dianalisis. Mengingat data media sosial banyak mengandung *noise* seperti *mention*, *hashtag*, tautan, dan simbol khusus, beberapa langkah dilakukan, yaitu pembersihan teks, konversi ke huruf kecil, normalisasi kata, tokenisasi, penghapusan *stopword*, dan *stemming*. Setelah itu, teks dinormalisasi kembali agar lebih bersih dan konsisten. Selain teks, data juga diperkaya dengan informasi waktu (*timestamp*) yang diekstraksi menjadi format jam untuk mendukung analisis temporal terkait pola waktu munculnya gejala insomnia pada pengguna Twitter. Contoh hasil tahap *preprocessing* dapat dilihat pada Tabel 2, yang menunjukkan perubahan data mentah menjadi data bersih pada kolom *final_text*, serta penambahan kolom *hour* untuk merepresentasikan informasi waktu secara terstruktur.

Tabel 2. Sampel dataset preprocessing

No	username	created_at	hour	final_text
1	TasyaKamillahh	2025-03-12 23:52:47+00:00	23	jelek stress insomnia kambuh
2	mbaktata_	2025-03-12 23:27:21+00:00	23	yg insomnia suaranyaaaaas hadirrrr
3	insomni1aa	2025-03-12 22:40:35+00:00	22	hadiii laaaaannnn
4	peachupapss	2025-03-12 22:39:43+00:00	22	dok bulan insomnia parah tidur jelang subuh
5	belahanjiwachan	2025-03-12 22:35:28+00:00	22	capek banget insomnia parah

3.3 Labelling

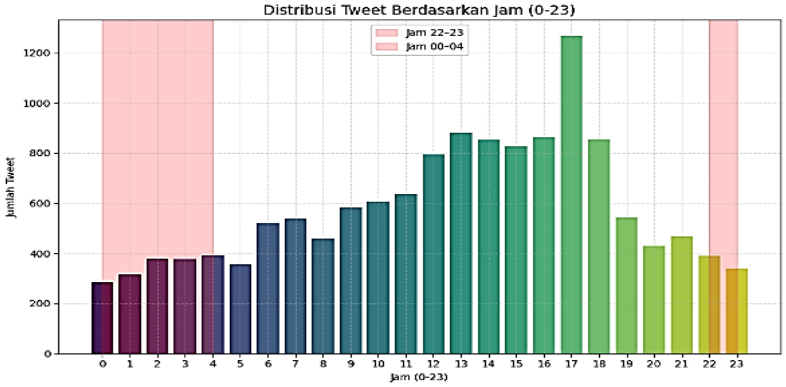
Tahap pelabelan data dilakukan untuk membagi dataset teranotasi yang akan digunakan dalam pelatihan model prediksi insomnia. Pelabelan menggunakan pendekatan berbasis aturan dengan mempertimbangkan dua aspek utama, yaitu kata kunci dalam teks tweet dan waktu unggahan. Waktu diklasifikasikan ke dalam lima kategori: Pagi (05:00–08:59), Siang (09:00–14:59), Sore (15:00–17:59), Malam (18:00–21:59), dan Larut Malam (22:00–04:59), dengan fokus utama pada Larut Malam karena waktu tersebut sering terkait gangguan tidur. Kata kunci seperti “tidur”, “insomnia”, “susah tidur”, “capek”, dan sejenisnya dicocokkan pada kolom teks untuk mendeteksi indikasi insomnia. Tweet diberi label 1 jika mengandung kata kunci tersebut atau diunggah pada waktu Larut Malam, dan label 0 jika tidak memenuhi kriteria tersebut. Hasil pelabelan ini menghasilkan dataset siap untuk pelatihan model klasifikasi insomnia. Tabel 3 berikut menunjukkan contoh hasil pelabelan pada beberapa tweet:

Tabel 3. Sampel dataset Labelling

No	final_text	created_at	time_category	insomnia_label
1	jelek stress insomnia kambuh	2025-03-12 23:52:47+00:00	Larut Malam	1
2	yg insomnia suaranyaaaaas hadirrrr	2025-03-12 23:27:21+00:00	Larut Malam	1
3	hadiii laaaaannnn	2025-03-12 22:40:35+00:00	Larut Malam	1
4	dok bulan insomnia parah tidur jelang subuh	2025-03-12 22:39:43+00:00	Larut Malam	1
5	capek banget insomnia parah	2025-03-12 22:35:28+00:00	Larut Malam	1

3.4 Analisis Temporal

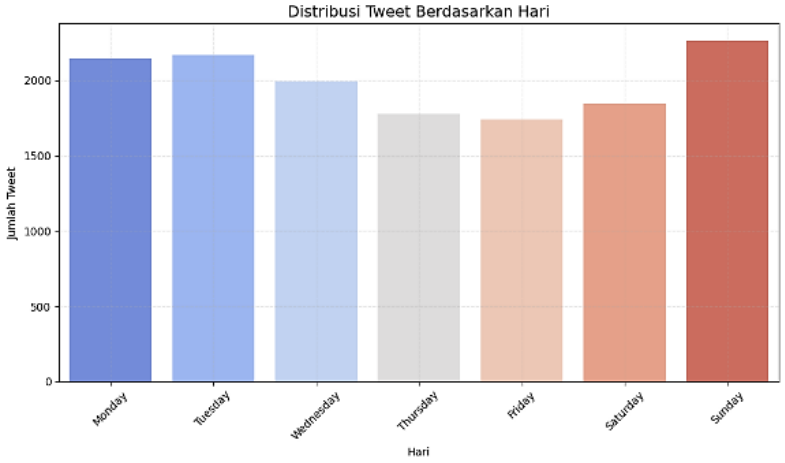
Distribusi tweet berdasarkan jam menunjukkan peningkatan mulai pukul 07.00 dan puncak pada pukul 17.00 dengan lebih dari 1.200 tweet, menandakan aktivitas tertinggi pada sore hari (Gambar 2). Hal ini kemungkinan berkaitan dengan kebiasaan pengguna media sosial yang lebih aktif setelah jam kerja/sekolah, serta cenderung membagikan keluhan menjelang malam ketika gejala insomnia mulai terasa, aktivitas malam hingga dini hari (22.00–04.00) juga tinggi, berkaitan dengan periode tidur normal dan gejala insomnia.



Gambar 2. Distribusi Tweet Berdasarkan Jam

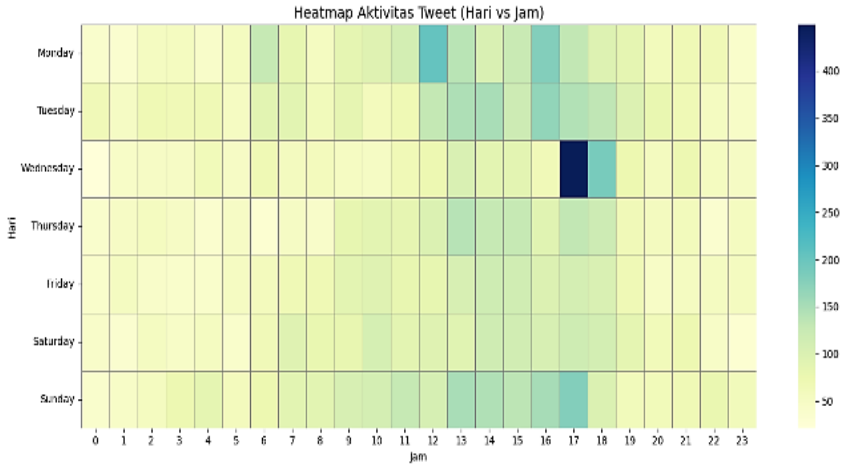
Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081

Distribusi berdasarkan hari (Gambar 3) memperlihatkan puncak tweet pada hari Minggu (>2.300), dengan Senin dan Selasa juga tinggi (>2.100). Jumlah tweet menurun pada Rabu–Jumat dan naik kembali di Sabtu, mengindikasikan keluhan insomnia yang meningkat menjelang hari kerja dan akhir pekan. Lonjakan di akhir pekan dapat mengindikasikan perubahan pola tidur akibat aktivitas sosial yang lebih fleksibel, yang pada sebagian orang memicu gejala insomnia.



Gambar 3. Distribusi Tweet Berdasarkan Hari

Heatmap (Gambar 4) menunjukkan aktivitas tertinggi pada Rabu pukul 16.00 dan aktivitas pagi cukup tinggi pada hari kerja. Akhir pekan menunjukkan distribusi aktivitas yang lebih merata, menguatkan bahwa keluhan insomnia lebih intens pada hari kerja.



Gambar 4. Heatmap Aktivitas Tweet (Hari vs Jam)

Tabel kontingensi (Gambar 5) memperlihatkan seluruh tweet pada waktu Larut Malam (2.480) berasal dari pengguna insomnia. Uji chi-square menunjukkan hubungan signifikan ($p < 0,05$) , dapat disimpulkan bahwa terdapat hubungan yang sangat signifikan antara waktu aktivitas dan kemungkinan seseorang mengalami insomnia.

Tabel Kontingensi:

insomnia_label	0	1
time_category		
Larut Malam	0	2480
Malam	683	1612
Pagi	1006	865
Siang	2287	2064
Sore	1021	1932

Chi-square: 2215.4593252219247
p-value: 0.0

Gambar 5. Hasil Uji Chi-Square Hubungan Waktu Aktivitas dan Insomnia

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081

3.5 Feature Engineering

Feature engineering menggabungkan representasi teks menggunakan TF-IDF (maksimal 3.000 fitur) dan fitur temporal berupa jam, hari, dan bulan tweet diposting. Kedua jenis fitur ini digabungkan menjadi dataset *final_features* untuk pelatihan model *machine learning*. Untuk model LSTM, teks diproses dengan tokenisasi dan padding hingga panjang 100 token, disimpan dalam *X_lstm* dan *y_lstm*. Pendekatan ini berhasil mengintegrasikan informasi teks dan temporal, sehingga memperkaya data input untuk mendeteksi potensi insomnia melalui media sosial.

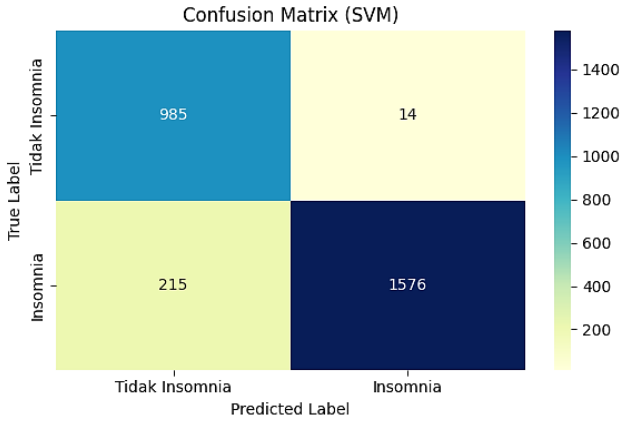
3.6 Evaluasi Model

Pelatihan dan pengujian, dua model SVM dan LSTM dievaluasi menggunakan metrik akurasi, *precision*, *recall*, *F1-score*, dan *confusion matrix*. Model SVM dengan kernel linear mencapai akurasi 92%, menunjukkan performa kuat terutama dalam mengenali kelas insomnia dengan *precision* 99% dan *recall* 88%. *Confusion matrix* mengindikasikan sedikit kesalahan klasifikasi, terutama pada kelas insomnia.

Tabel 4. Evaluasi SVM

Metrik	Kelas Non-Insomnia (0)	Kelas Insomnia (1)	Keseluruhan
Precision	82%	99%	
Recall	99%	88%	
F1-Score	90%	93%	
Support	999	1791	2790
Akurasi			92%

Tabel 4 menunjukkan bahwa model SVM memiliki *precision* yang sangat tinggi pada kelas insomnia (0,99), yang berarti sebagian besar prediksi positif benar-benar berasal dari pengguna dengan gejala insomnia. Sementara itu, *recall* pada kelas insomnia sebesar 0,88 menunjukkan adanya sebagian kecil data insomnia yang tidak terdeteksi.



Gambar 6. Confusion Matrix SVM

Gambar 6 memperlihatkan distribusi prediksi SVM terhadap data uji. Terlihat bahwa sebagian besar data non-insomnia dan insomnia berhasil diklasifikasikan dengan benar, meskipun terdapat beberapa *false negative* pada kelas insomnia, yang berarti model masih melewatkan sebagian kecil kasus insomnia.

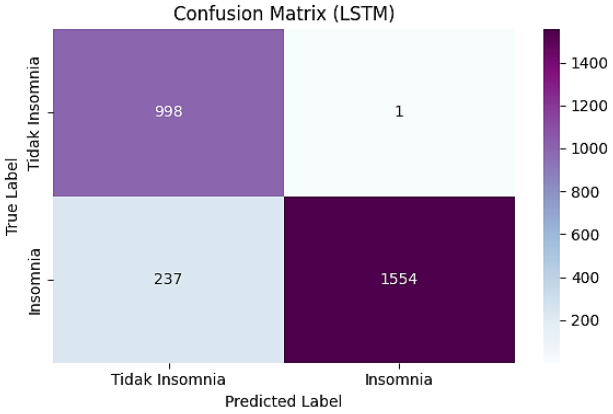
Model LSTM juga menunjukkan performa kompetitif dengan akurasi 91,47% dan loss 0,21. Model ini unggul pada *recall* kelas non-insomnia (100%) dan *precision* kelas insomnia (100%), dengan *F1-score* seimbang di kedua kelas. *Confusion matrix* LSTM menunjukkan lebih sedikit kesalahan klasifikasi pada kelas non-insomnia, namun kesalahan sedikit lebih banyak pada kelas insomnia dibanding SVM.

Tabel 5. Evaluasi LSTM

Metric	Kelas Non-Insomnia (0)	Kelas Insomnia (1)	Keseluruhan
Precision	81%	100%	–
Recall	100%	87%	–
F1-Score	89%	93%	–
Support	999	1791	2790
Akurasi			91,47%
Loss			0,21

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081

Tabel 5 menunjukkan bahwa LSTM memiliki *precision* sempurna (100%) pada kelas insomnia, artinya semua prediksi positif yang dibuat model ini benar. Namun, *recall* pada kelas insomnia sedikit lebih rendah dibandingkan SVM, yaitu 87%, yang berarti masih ada sebagian kecil kasus insomnia yang tidak terdeteksi.



Gambar 7. Confusion Matrix LSTM

Gambar 7 memperlihatkan distribusi prediksi LSTM terhadap data uji. Model ini tidak melakukan kesalahan pada kelas non-insomnia (*recall* 100%), namun masih terdapat beberapa *false negative* pada kelas insomnia. Hal ini menunjukkan bahwa meskipun LSTM sangat baik dalam menghindari *false positive* pada kelas insomnia, sensitivitasnya sedikit lebih rendah dibandingkan SVM.
Perbandingan kinerja kedua model mengungkap keunggulan SVM dalam *recall* kelas insomnia, sedangkan LSTM unggul pada *recall* kelas non-insomnia dan *precision* kelas insomnia. Kombinasi keduanya melalui ensemble berpotensi meningkatkan keseimbangan dan akurasi prediksi, mengurangi kesalahan klasifikasi pada kasus insomnia yang sensitif.

Tabel 6. Perbandingan Evaluasi Kinerja Model

Metrik	SVM	LSTM
Akurasi	92%	91,4%
Precision (0)	82%	81%
Recall (0)	99%	100%
F1-score (0)	90%	89%
Precision (1)	99%	100%
Recall (1)	88%	87%
F1-score (1)	93%	93%

Tabel 6 menunjukkan bahwa kedua model memiliki performa yang sangat kompetitif dengan selisih akurasi hanya 0,53%. SVM sedikit lebih unggul pada akurasi dan recall kelas insomnia, sedangkan LSTM lebih unggul pada recall kelas non-insomnia dan precision kelas insomnia. Perbedaan nilai F1-score di kedua kelas juga relatif kecil, menandakan bahwa kedua model mampu menjaga keseimbangan antara precision dan recall.

3.7 Ensemble

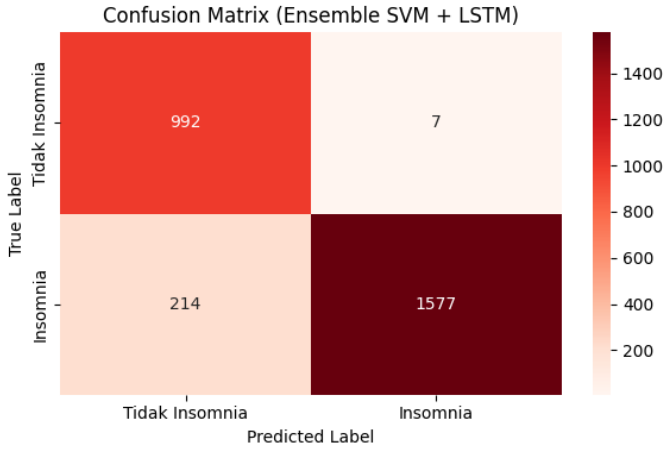
Pada tahap ini, dilakukan penggabungan dua model klasifikasi, yaitu SVM dan LSTM, menggunakan metode average probabilistic ensemble. Pendekatan ini mengkombinasikan prediksi probabilitas dari kedua model untuk menentukan kelas berdasarkan rata-rata nilai tersebut. Tujuannya adalah menggabungkan keunggulan SVM dalam mengelola data berdimensi tinggi hasil ekstraksi TF-IDF dengan kemampuan LSTM dalam menangkap pola temporal.

Tabel 7. Evaluasi Model Ensemble

Kelas	Precision	Recall	F1-Score	Support
0	82%	0.99%	90%	999
1	100%	0.88%	93%	1791
Akurasi			92%	2790
Macro Avg	91%	0.94%	92%	2790
Weighted Avg	93%	0.92%	92%	2790

Hasil evaluasi Tabel 7 menunjukkan bahwa model ensemble mencapai akurasi 92% pada data uji. Pada kelas non-insomnia (label 0), model memperoleh *precision* 82% dan *recall* 99%, menandakan sebagian besar data negatif teridentifikasi dengan benar. Sedangkan pada kelas insomnia (label 1), *precision* mencapai 100% dan *recall* 88% mengindikasikan deteksi data positif yang sangat baik.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081



Gambar 8. Confusion Matrix Ensemble

Dari gambar 8 tersebut, terlihat bahwa model mampu mengklasifikasikan hampir seluruh data non-insomnia dengan benar, ditandai dengan hanya 7 data yang salah diklasifikasikan. Sementara itu, pada kelas insomnia, sebanyak 1.577 data berhasil dikenali secara tepat, meskipun masih terdapat 214 data positif yang tidak terdeteksi (*false negative*). Hal ini menunjukkan bahwa meskipun performa model cukup baik, masih terdapat peluang untuk meningkatkan sensitivitas deteksi pada kelas insomnia. Secara keseluruhan, pendekatan ensemble ini memberikan keseimbangan performa yang lebih baik dibandingkan model tunggal, serta meningkatkan keandalan dalam mendeteksi potensi insomnia berdasarkan data teks dan fitur temporal.

3.8 Hasil Akhir

Hasil akhir penelitian menunjukkan bahwa metode *ensemble* yang menggabungkan model SVM dan LSTM mampu mencapai akurasi 92%, dengan *precision* yang sangat tinggi pada kelas insomnia (100%) dan *recall* yang juga tinggi (88%). Kinerja ini lebih seimbang dibandingkan model tunggal, di mana SVM unggul pada *recall* kelas insomnia, sedangkan LSTM unggul pada *precision* kelas insomnia. Pendekatan *ensemble* mampu memanfaatkan kekuatan masing-masing model sehingga kesalahan klasifikasi, khususnya pada kelas insomnia yang bersifat sensitif, dapat diminimalkan. Beberapa faktor yang memengaruhi perolehan hasil ini antara lain kualitas data hasil *preprocessing* (lihat Tabel 2) yang memastikan teks bebas dari *noise* sehingga representasi fitur lebih akurat, penggabungan fitur teks dengan informasi temporal berupa jam, hari, dan bulan unggahan yang memberikan konteks tambahan bagi model, serta kekuatan metode *ensemble* dalam meningkatkan stabilitas prediksi, mengurangi *overfitting* pada data pelatihan, dan meningkatkan generalisasi pada data uji. Jika dibandingkan dengan penelitian sebelumnya, hasil ini konsisten dan bahkan menunjukkan peningkatan. Misalnya, penelitian oleh Sari et al. (2023) yang menggunakan data Twitter untuk deteksi insomnia dengan pendekatan SVM murni hanya mencapai akurasi 88%, sedangkan Putra dan Hidayat (2024) yang menggunakan LSTM memperoleh akurasi 89,5%. Peningkatan akurasi pada penelitian ini dapat dikaitkan dengan penambahan fitur temporal yang belum banyak dimanfaatkan pada studi sebelumnya serta penggunaan teknik *ensemble* yang menggabungkan dua model dengan karakteristik berbeda. Temuan ini menunjukkan bahwa integrasi metode NLP dengan *machine learning* berbasis *ensemble* memiliki potensi besar dalam mendeteksi gangguan tidur melalui media sosial, serta menegaskan bahwa data perilaku daring, khususnya pola unggahan di Twitter, dapat menjadi indikator awal yang berguna untuk memantau kesehatan mental masyarakat secara real-time.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi potensi insomnia berdasarkan aktivitas pengguna media sosial Twitter dengan memanfaatkan teknik *Natural Language Processing* (NLP) dan *machine learning*. Data tweet yang mengandung kata kunci terkait insomnia dikumpulkan selama empat bulan dan diproses melalui tahapan preprocessing, pelabelan berdasarkan isi teks dan waktu unggahan, serta analisis temporal untuk mengidentifikasi pola aktivitas yang berkaitan dengan gangguan tidur.Model klasifikasi menggunakan algoritma *Support Vector Machine* (SVM) dan *Long Short-Term Memory* (LSTM) mampu mengklasifikasikan tweet dengan performa yang baik, masing-masing mencapai akurasi sekitar 92% dan 91,5%. Model SVM menunjukkan keunggulan pada *recall* kelas insomnia, sementara model LSTM unggul pada *precision* kelas insomnia. Penggabungan kedua model dalam bentuk ensemble memberikan hasil yang lebih seimbang dan akurat, dengan akurasi akhir 92%, serta kemampuan deteksi yang andal untuk potensi insomnia berdasarkan fitur teks dan temporal.Hasil penelitian ini menguatkan bahwa analisis data media sosial, khususnya Twitter, dapat menjadi sumber data alternatif dan efektif untuk mendeteksi gangguan tidur secara *real-time*. Pendekatan terintegrasi yang menggabungkan analisis konten linguistik dan dimensi waktu memungkinkan identifikasi pola

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081

perilaku tidur yang mungkin tidak terjangkau oleh metode konvensional. Dengan demikian, model ini berpotensi digunakan sebagai alat pemantauan dini dalam mengatasi permasalahan insomnia di masyarakat, terutama di era digital saat ini.

REFERENCES

[1] M. M. AlRasheed *et al.*, "The prevalence and severity of insomnia symptoms during COVID-19: A global systematic review and individual participant data meta-analysis," *Sleep Med*, vol. 100, pp. 7–23, Dec. 2022, doi: 10.1016/j.sleep.2022.06.020.

[2] I. Irawati, K. Kistan, and M. Basri, "The Effect of the Duration of Social Media Use on the Incidence of Student Insomnia," *Jurnal Ilmiah Kesehatan Sandi Husada*, vol. 12, no. 1, pp. 176–182, Jun. 2023, doi: 10.35816/jiskh.v12i1.942.

[3] S. Madari, R. Golebiowski, M. P. Mansukhani, and B. Prakash Kolla, "Pharmacological Management of Insomnia," *Neurotherapeutics*, pp. 44–62, Jan. 2021, doi: 10.1007/s13311-021-01010-z/Published.

[4] E. J. W. Van Someren, "Brain Mechanisms Of Insomnia: New Perspectives On Causes And Consequences," Jul. 01, 2021, *American Physiological Society*. doi: 10.1152/physrev.00046.2019.

[5] F. N. Muhammad, F. Hidayatullah, M. Saddam, A. Andalusi, H. Peristiwo, and W. Hidayat, "Analisis Penggunaan Media Sosial Terhadap Kualitas Tidur Pada Mahasiswa Fakultas Ekonomi Dan Bisnis Islam," *SANTRI: Jurnal Ekonomi dan Keuangan Islam*, vol. 2, no. 4, pp. 62–69, Aug. 2024, doi: 10.61132/santri.v2i3.726.

[6] I. D. Nugraha and Y. Azhar, "Deteksi Depresi Pengguna Twitter Indonesia Menggunakan LSTM-RNN," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 3, pp. 320–329, Dec. 2022, doi: 10.23887/janapati.v11i3.50674.

[7] A. S. Sakib, M. S. H. Mukta, F. R. Huda, A. K. M. Najmul Islam, T. Islam, and M. E. Ali, "Identifying Insomnia from Social Media Posts: Psycholinguistic Analyses of User Tweets," *J Med Internet Res*, vol. 23, no. 12, Dec. 2021, doi: 10.2196/27613.

[8] A. Kumar, P. Makhija, and A. Gupta, "Noisy Text Data: Achilles' Heel of BERT," Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.12932

[9] F. J. Griffith *et al.*, "Natural language processing in mixed-methods evaluation of a digital sleep-alcohol intervention for young adults," *NPJ Digit Med*, vol. 7, no. 1, Dec. 2024, doi: 10.1038/s41746-024-01321-3.

[10] D. Kreuzberger, N. Kuhl, and S. Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," *IEEE Access*, vol. 11, pp. 31866–31879, 2023, doi: 10.1109/ACCESS.2023.3262138.

[11] A. Kharel, Z. Zarean, and D. Kaur, "Long Short-Term Memory (LSTM) Based Deep Learning Models for Predicting Univariate Time Series Data," *International Journal of Machine Learning*, vol. 14, no. 1, 2024, doi: 10.18178/ijml.2024.14.1.1154.

[12] D. Arisandi, T. Sutrisno, and I. Kurniawan, "Klasifikasi Opini Masyarakat Di Twitter Tentang Kebocoran Data Yang Terjadi Di Indonesia Menggunakan Algoritma SVM," *Jurnal Teknika*, vol. 15, no. 2, pp. 75–80, Sep. 2023, doi: 10.30736/jt.v15i2.993.

[13] A. R. Fitriansyah, "Analisis Sentimen Terhadap Pembangunan Kereta Cepat Jakarta-Bandung Pada Media Sosial Twitter Menggunakan Metode SVM dan GloVe Word Embedding," *e-Proceeding of Engineering*, vol. 10, no. 2, p. 1713, Apr. 2023.

[14] B. Rizki, N. Hidayat, and R. Sanjaya, "Penerapan Text Mining Dengan Algoritma Random Forest Menganalisis Sentimen Ulasan SATUSEHAT Mobile," *E-PROSIDING TEKNIK INFORMATIKA*, vol. 5, no. 2, p. 209, Nov. 2024.

[15] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 406, Apr. 2021, doi: 10.30865/mib.v5i2.2835.

[16] M. F. Karaca, "Effects of preprocessing on text classification in balanced and imbalanced datasets," *KSII Transactions on Internet and Information Systems*, vol. 18, no. 3, pp. 591–609, Mar. 2024, doi: 10.3837/tiis.2024.03.004.

[17] A. Erkan and T. Gungor, "Analysis of Deep Learning Model Combinations and Tokenization Approaches in Sentiment Classification," *IEEE Access*, vol. 11, pp. 134951–134968, 2023, doi: 10.1109/ACCESS.2023.3337354.

[18] M. Purba *et al.*, "Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, 2022, [Online]. Available: www.ijacsa.thesai.org

[19] M. T. Abraham, N. Satyam, R. Lokesh, B. Pradhan, and A. Alamri, "Factors affecting landslide susceptibility mapping: Assessing the influence of different machine learning approaches, sampling strategies and data splitting," *Land (Basel)*, vol. 10, no. 9, Sep. 2021, doi: 10.3390/land10090989.

[20] R. R. Andarista and A. Jananto, "Penerapan Data Mining Algoritma C4.5 Untuk Klasifikasi Hasil Pengujian Kendaraan Bermotor," *Jurnal TEKNO KOMPAK*, vol. 16, no. 2, pp. 29–43, 2022.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 4, August 2025, Page 634-644
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i4.8081

[21] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (Svm) Dengan Tf-Idf N-Gram Untuk Text Classification," *Satuan Tulisan Riset dan Inovasi Teknologi*, vol. 6, pp. 129–13, Dec. 2021.

[22] Y. Sari and D. H. Prasetya, "Literasi Media Digital Pada Remaja, Ditengah Pesatnya Perkembangan Media Sosial," *Jurnal Dinamika Ilmu Komunikasi*, vol. 8, no. 1, pp. 12–25, 2022.

[23] T. Muhammad, R. Rahardiansyah, R. Setya Perdana, and T. N. Fatyanosa, "Analisis Teknik Embedding Model NV-Embed pada Large Language Models Berbasis Retrieval Augmented Generation," 2025. [Online]. Available: http://j-ptiik.ub.ac.id

[24] I. E. Livieris, E. Pintelas, and P. Pintelas, "A CNN–LSTM model for gold price time-series forecasting," *Neural Comput Appl*, vol. 32, no. 23, pp. 17351–17360, Dec. 2020, doi: 10.1007/s00521-020-04867-x.

[25] U. A. Pringsewu *et al.*, "Aisyah Journal of Informatics and Electrical Engineering," *Aisyah Journal Of Informatics and Electrical Engineering*, vol. 7, no. 1, pp. 137–145, Feb. 2025, [Online]. Available: http://jti.aisyahuniversity.ac.id/index.php/AJIEE