Terakreditasi Peringkat 4
Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi, No. 225/E/KPT/2022
Masa berlaku dari Volume 8 Nomor 1 Tahun 2022 sampai dengan Volume 12 Nomor 2 Tahun 2026

Published online at: **https://ejournal.raharja.ac.id/**

JOURNAL CERITA:
Creative Education of Research in Information Technology and Artificial informatics

Vol. 11 No. 2 (2025) 241-248 e-ISSN: 2655 - 2574

Analisis Sentimen Komentar Instagram Terkait Persepsi Hidup Sehat Menggunakan Algoritma BERT-LSTM

Audiva Tartila Daning Putri[1], Yisti Vita Via*[2], Muhammad Muharrom Al Haromainy[3]
[1,2,3]Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia
Email: tartilaputri631@gmail.com[1]; **yistivia.if@upnjatim.ac.id*[2]**;
muhammad.muharrom.if@upnjatim.ac.id[3]

Putri, A.T.D., Via, Y.V., & Al Haromainy, M. M. (2025). Analisis Sentimen Komentar Instagram Terkait Persepsi Hidup Sehat Menggunakan Algoritma BERT-LSTM. *Journal Cerita: Creative Education of Research in Information Technology and Artificial Informatics*, 11(2), 241-248

DOI: https://doi.org/10.33050/cerita.v11i2.3526

ABSTRAK

Persepsi masyarakat terhadap gaya hidup sehat sangat penting bagi kesejahteraan fisik dan mental, terutama dengan meningkatnya kasus penyakit tidak menular di Indonesia. Instagram merupakan sarana yang efektif untuk menyebarkan konten edukasi, sehingga dapat dilakukan analisis sentimen terhadap persepsi gaya hidup sehat dengan memanfaatkan data komentar pada unggahan akun @ayosehat.kemkes melalui proses pengumpulan dan prapemrosesan data, pelabelan, serta pembagian data menjadi set pelatihan, validasi, dan pengujian. Dengan penggabungan algoritma BERT-LSTM dalam mengklasifikasikan sentimen menjadi positif, negatif, dan netral, hasil pengujian menunjukkan model terbaik mencapai akurasi 89,20%, dengan presisi 89,49%, recall 89,20%, dan F1-score 88,74% dengan rasio pembagian dataset 80:10:10.

Kata kunci: Hidup Sehat, Instagram, Analisis Sentimen, BERT-LSTM

ABSTRACT

The perception of a healthy lifestyle is crucial for physical and mental well-being, especially with the rising cases of non-communicable diseases in Indonesia. Instagram is an effective platform for education, enabling sentiment analysis of the perception of a healthy lifestyle using comment data from posts on the @ayosehat.kemkes account. This involves data collection and preprocessing, labeling, and splitting the data into training, validation, and testing sets. By combining the BERT-LSTM algorithm to classify sentiment into positive, negative, and neutral, the testing results showed that the best model achieved an accuracy of 89.20%, with a precision of 89.49%, recall of 89.20%, and an F1-score of 88.74%, with an 80:10:10 dataset split ratio.

Keywords: *Healthy Lifestyle, Instagram, Sentiment Analysis, BERT-LSTM*

Received: xx-xx-20xx | Accepted: xx-xx-20xx | Published Online: xx-xx-20xx

I. PENDAHULUAN

Menerapkan pola hidup sehat merupakan aspek fundamental untuk mencapai kesehatan optimal. Kondisi tubuh yang sehat mendukung aktivitas sehari-hari secara lancar dan produktif (Tanır, 2019). Tetapi kenyataannya, Indonesia masih menghadapi tantangan dalam penerapan gaya hidup sehat, terutama terkait Penyakit Tidak Menular (PTM). Berdasarkan data Riskesdas Kemenkes, prevalensi Diabetes Mellitus meningkat dari 6,9% pada 2013 menjadi 11,7% pada 2023. Pola makan yang tidak seimbang, gaya hidup sedentari, konsumsi makanan siap saji, kurangnya aktivitas fisik, stres, dan kurang istirahat menjadi faktor pemicu utama berbagai penyakit seperti hipertensi, diabetes, obesitas, kanker, penyakit jantung, dan hiperkolesterol (Pembengo, 2020). Survei Katadata *Insight Center* pada 2023 menunjukkan bahwa 50,5% responden memiliki ketercukupan tidur, namun 48% dari mereka memilih kopi saat bergadang (Katadata, 2023). Survei Jajak Pendapat pada 2023 juga menemukan bahwa hanya 16% dari 600 responden minum air yang cukup, sementara 67% memilih minuman manis 1-2 kali sehari. Selain itu, 60% responden memilih nasi sebagai makanan pokok dan 16% makan makanan yang digoreng setiap hari (Jakpat, 2023).

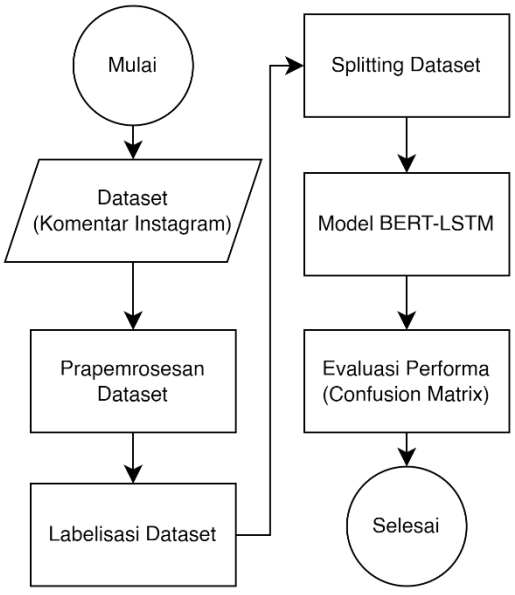
Program Gerakan Masyarakat Hidup Sehat (GERMAS) telah digagas pemerintah melalui Instruksi Presiden Nomor 1 Tahun 2017 berfokus pada edukasi kesehatan, dengan Instagram sebagai platform utama penyebaran informasi karena efektif dalam menarik perhatian pengguna (We are Social, 2023). Kampanye GERMAS oleh Kementerian Kesehatan Republik Indonesia (Kemenkes RI) mencakup berbagai inisiatif seperti edukasi melakukan aktivitas fisik, budaya konsumsi buah dan sayur, tidak merokok, tidak mengkonsumsi alkohol, melakukan pemeriksaan berkala, menjaga kebersihan lingkungan, dan menggunakan jamban (Agustina et al., 2024). Maka dari itu, penelitian ini mengusulkan analisis sentimen terhadap persepsi hidup sehat masyarakat Indonesia berdasarkan komentar di Instagram menggunakan algoritma gabungan BERT dan LSTM untuk mengevaluasi sentimen terhadap edukasi yang disampaikan oleh Kemenkes RI.

Beberapa penelitian sebelumnya menunjukkan bahwa algoritma BERT dan LSTM memberikan hasil yang sangat baik dalam klasifikasi teks. (Rai et al., 2022) menunjukkan bahwa kombinasi ini meningkatkan akurasi klasifikasi berita palsu sebesar 2,50%, dengan akurasi maksimum 88.75% untuk dataset PolitiFact. Selain itu, (Pandey & Singh, 2023) menemukan bahwa BERT-LSTM mengungguli algoritma lain dengan peningkatan skor F1 hingga 6% pada dataset campuran bahasa Inggris dan Hindi dari Twitter. Penemuan-penemuan ini menguatkan potensi kombinasi BERT dan LSTM dalam meningkatkan performa analisis teks.

Oleh karena itu, dilakukannya penelitian ini bertujuan untuk mengimplementasikan algoritma gabungan BERT dan LSTM dalam melakukan analisis sentimen terhadap komentar masyarakat Indonesia mengenai hidup sehat yang diambil dari unggahan konten edukasi di Instagram dan mengevaluasi performa model melalui matrik akurasi, presisi, *recall*, dan F1-*score* dari model gabungan BERT dan LSTM dalam klasifikasi sentimen terkait persepsi hidup sehat di Indonesia.

II. METODE PENELITIAN

Penelitian ini terdapat beberapa tahapan yang harus dilakukan agar dapat terselesaikan dengan baik. Alur dari metode penelitian ini dipaparkan pada Gambar 1.



Gambar 1. Alur Metode Penelitian

A. Pengumpulan Data

Pada penelitian ini, dataset berupa teks komentar Instagram didapatkan dari bantuan *web scraper* bernama Apify dengan menggunakan aktor *Instagram Comments Scraper*. Data yang diambil adalah komentar pada unggahan dari 6 Juni 2017 hingga 8 September 2024 pada akun @ayosehat.kemkes.

Tabel 1. Perbandingan Hasil Splitting Data

	Data Latih		Data Validasi		Data Uji	
Rasio 80:10:10	Total	2589	Total	324	Total	324
	Negatif	305	Negatif	38	Negatif	38
	Positif	986	Positif	124	Positif	123
	Netral	1298	Netral	162	Netral	163
Rasio 70:15:15	Total	2265	Total	486	Total	486
	Negatif	266	Negatif	58	Negatif	57
	Positif	863	Positif	185	Positif	185
	Netral	1136	Netral	243	Netral	244
Rasio 60:20:20	Total	954	Total	647	Total	648
	Negatif	228	Negatif	77	Negatif	76
	Positif	740	Positif	246	Positif	247
	Netral	974	Netral	324	Netral	325

B. Prapemrosesan Dataset

Prapemrosesan dataset merupakan tahap untuk mengurangi gangguan pada data yang akan diproses ke dalam model. Berikut adalah langkah-langkah yang diambil untuk melakukan prapemrosesan data:

1) *Translating* atau penerjemahan dilakukan dengan menyelaraskan bahasa yang ada dalam teks ke bahasa Inggris.
2) *Text Cleaning* untuk menghilangkan username(@), URL, simbol, dan angka yang tidak diperlukan.
3) *Remove Special Characters* untuk menghapus atau menghilangkan karakter-karakter khusus.
4) *Case Folding* untuk mengubah semua karakter dalam teks menjadi bentuk huruf *lowercase* yang konsisten.
5) *Stopwords Removal* untuk menghapus kata ganti, kata depan, kata penghubung, kata penunjuk, kata bantu, dan kata keterangan.

C. Pemilahan Dataset

Pemilahan dataset adalah tahap untuk memastikan bahwa dataset bersih, hanya mengandung data yang relevan dengan persepsi gaya hidup sehat, dan bebas dari entri yang tidak berguna atau duplikat yang dapat mempengaruhi hasil analisis atau model yang akan dibangun. Dalam tahap ini, didapatkan sampel dataset sebanyak 3.237 teks komentar yang telah bersih dan relevan untuk penelitian.

D. Labelisasi Dataset

Sampel dataset yang telah melalui prapemrosesan data akan diproses lebih lanjut untuk melakukan labelisasi menggunakan TextBlob yang berfokus pada analisis sentimen dokumen teks dalam bahasa Inggris. Pada penelitian ini, label yang digunakan untuk klasifikasi sentimen adalah negatif, netral, dan positif, TextBlob membantu untuk mengukur subjektivitas dan polaritas. Subjektivitas berkisar dari 0 hingga 1, di mana 0 menunjukkan objektivitas (fakta) dan 1 menunjukkan subjektivitas (opini). Polaritas mengukur sentimen dengan nilai < 0 (negatif), > 0 (positif), dan 0 menunjukkan sentimen netral. Hasil dari labelisasi ini akan dijadikan acuan dalam melatih model BERT-LSTM melakukan klasifikasi analisis sentimen.

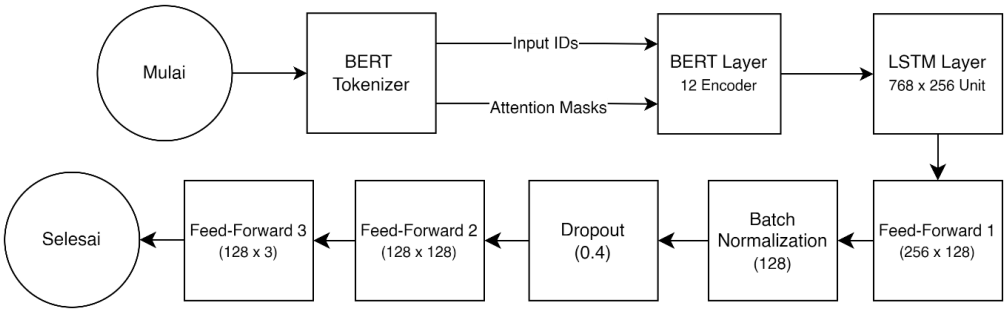
E. Spliting Dataset

Dalam tahapan *Splitting* Dataset, pembagiannya didasari oleh metode holdout yang membagi dataset sampel menjadi tiga bagian, yaitu data pelatihan (*training data*), data validasi (*validation data*), dan data pengujian (*test data*)

dengan dilakukan pada 3 rasio perbandingan. Dengan total dataset yang didapatkan dari tahapan-tahapan sebelumnya, penyebaran data untuk masing-masing bagian dapat dipaparkan seperti Tabel 1.

F. Arsitektur Model

Setelah seluruh dataset berhasil melalui prapemrosesan, tahapan selanjutnya adalah membangun model BERT-LSTM. Pada tahap ini, model akan dikembangkan dan disusun untuk dapat mengolah data dengan efektif. Proses ini melibatkan pemilihan arsitektur model, konfigurasi lapisan-lapisan yang dibutuhkan, dan penyesuaian parameter. Arsitektur dari model BERT-LSTM dapat dilihat pada Gambar 2.



Gambar 2. Arsitektur Model BERT-LSTM

Sesuai pada alur Gambar 2, pertama-tama dataset akan diproses melalui tahap BERT Tokenizer untuk mendapatkan input IDs dan attention masks. Setelah itu, proses embedding dilakukan oleh BERT yang terdiri dari *token embedding, segment embedding, dan position embedding*. Ketiga *embedding* ini digabungkan untuk menangkap informasi kontekstual serta posisi relatif setiap token dalam urutan input. Setelah embedding, input diproses melalui 12 *encoder layers* karena penelitian ini menggunakan arsitektur BERT Base.

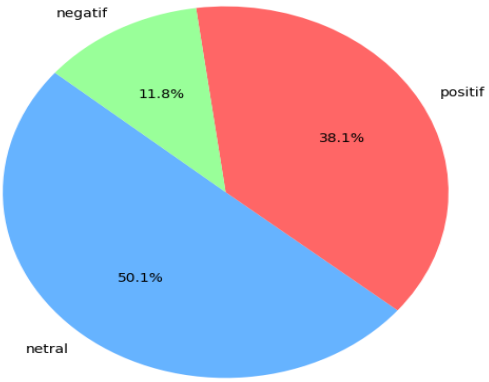
Hasil ekstraksi fitur dari BERT kemudian dikirimkan ke lapisan LSTM. Dalam penelitian ini, digunakan dua lapisan LSTM dengan ukuran 768 unit pada lapisan pertama dan 256 unit pada lapisan kedua. Lapisan pertama bertujuan untuk menangkap fitur dasar dari data sekuensial, sedangkan lapisan kedua membentuk fitur yang lebih abstrak berdasarkan hasil dari lapisan pertama.

Dalam konteks analisis sentimen yang berfokus pada klasifikasi teks dengan tiga label keluaran (positif, negatif, dan netral), arsitektur model menggunakan *Dense* layer pada lapisan output yang memiliki 3 neuron. Seperti yang terlihat pada Gambar 2, alur lapisan klasifikasi ini terdiri dari Dense layer pertama dengan 128 neuron dan fungsi aktivasi ReLU, dilanjutkan dengan Batch Normalization untuk menormalkan output. Selanjutnya, ada lapisan Dropout dengan tingkat *dropout* 0.4 untuk mengurangi *overfitting*, dan *dense layer* kedua dengan 128 neuron serta fungsi aktivasi ReLU. Lapisan terakhir dari lapisan model ini adalah *output layer* dengan fungsi aktivasi Softmax, yang memiliki neuron sesuai jumlah label (3 neuron untuk klasifikasi positif, negatif, dan netral). Kombinasi dari lapisan-lapisan ini memungkinkan model untuk mempelajari representasi secara kompleks.

III. HASIL DAN PEMBAHASAN

A. Dataset

Dataset teks komentar yang berhasil dilakukan prapemrosesan data, pemilahan data, dan labelisasi data dalam penelitian ini dipaparkan pada Gambar 3.



Gambar 3. Dataset

Berdasarkan diagram yang ditampilkan pada Gambar 3, dari 3.237 data teks menghasilkan sentimen positif sebanyak 38,1% atau 1.233 data, sentimen netral sebanyak 50,1% atau 1.623 data, dan sentimen negatif sebanyak 11,8% atau 381 data. Berdasarkan nilai polaritas yang ditemukan dengan TextBlob menunjukkan bahwa sentimen netral tidak memiliki emosi dan memiliki nilai subjektivitas sebesar 0.

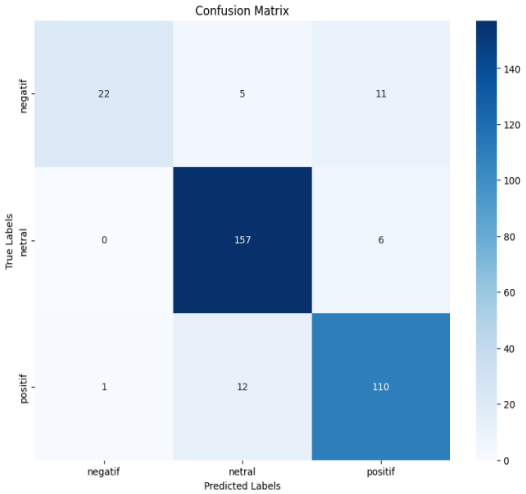
B. Perbandingan Confusion Matrix

Dalam memastikan kinerja sistem kombinasi model klasifikasi teks dengan BERT-LSTM, peneliti menilai efek dari berbagai pembagian rasio pada pembagian dataset menjadi data latih, data validasi, dan data uji pada akurasi model BERT dalam memprediksi klasifikasi teks. Proses pelatihan dan pengujian model dilakukan dengan menggunakan parameter-parameter seperti yang dipaparkan pada Tabel 2.

Tabel 2. Parameter BERT-LSTM

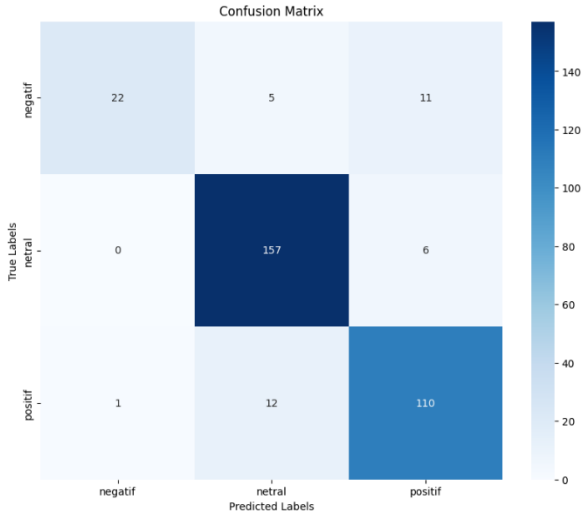
Parameter	Value
Learning Rate	1e-4
Batch size	32
Epoch	10
Optimizer	Adam
Dropout Rate	0.4

Dari penerapan parameter-parameter tersebut pada model, akan dilakukan 3 skenario pengujian yang akan menguji tentang bagaimana pengaruh pembagian rasio *splitting* dataset terhadap akurasi model. Pengujian ini diharapkan dapat memberikan pemahaman tentang pemahaman tentang bagaimana variasi dalam pembagian rasio dataset antara data latih, validasi, dan uji mempengaruhi akurasi model.



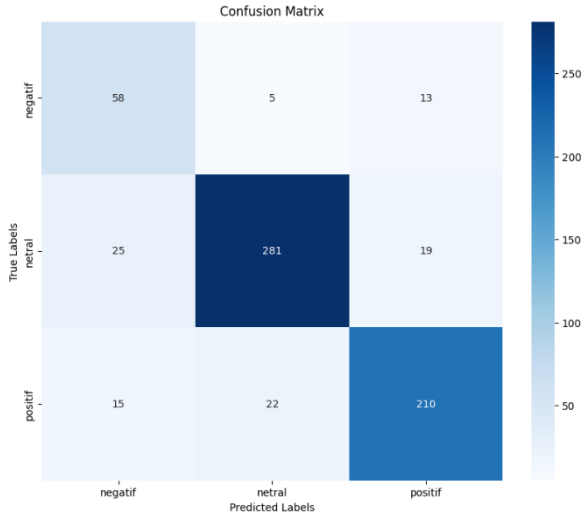
Gambar 4. Confusion Matrix Rasio 80:10:10

Confusion matrix yang ditampilkan pada Gambar 4 menunjukkan performa model dalam mengklasifikasikan data menjadi tiga kelas, yaitu negatif, netral, dan positif. Model berhasil memprediksi 22 dari 38 data negatif dengan benar, sementara 5 data negatif salah diklasifikasikan sebagai netral, dan 11 lainnya salah diklasifikasikan sebagai positif. Untuk kelas netral, performa model sangat baik, dengan 157 prediksi benar dari total 163 data netral, tetapi 6 data salah diklasifikasikan sebagai positif. Pada kelas positif, model memprediksi 110 dari 123 data positif dengan benar, sementara 12 data salah diklasifikasikan sebagai netral dan 1 data salah diklasifikasikan sebagai negatif. Secara keseluruhan, matrix ini menunjukkan performa yang sangat baik pada kelas netral dan positif, tetapi terdapat kesalahan yang lebih signifikan pada prediksi untuk kelas negatif.

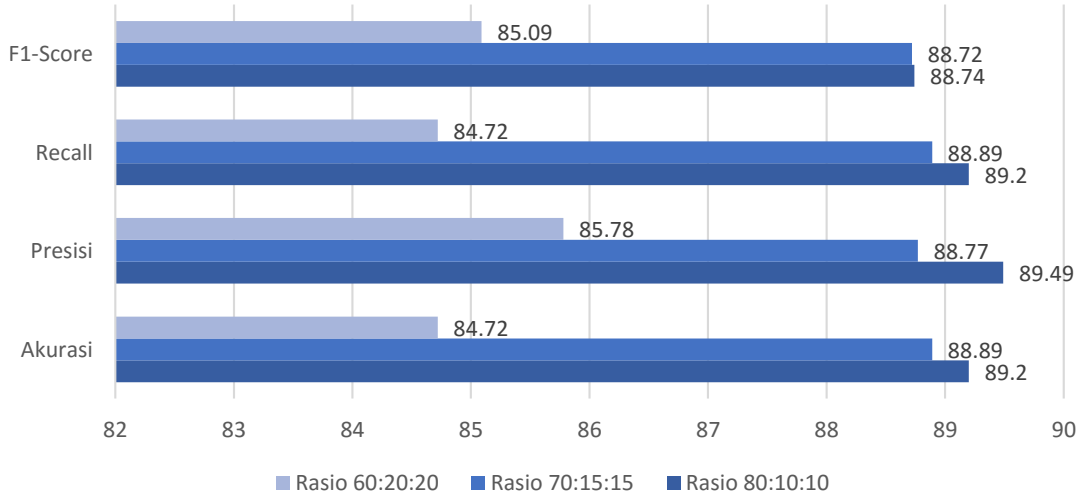


Gambar 5. Confusion Matrix Rasio 70:15:15

Confusion matrix yang ditampilkan pada Gambar 5 menunjukkan bahwa model berhasil mengklasifikasikan 38 sampel negatif dengan benar, namun salah mengklasifikasikan 9 sampel sebagai netral dan 10 sampel sebagai positif. Untuk kelas netral, model menunjukkan performa yang sangat baik dengan 236 sampel diklasifikasikan dengan benar, sementara hanya 4 sampel salah diklasifikasikan sebagai negatif dan 4 sampel lainnya sebagai positif. Pada kelas positif, 158 sampel berhasil diklasifikasikan dengan benar, tetapi terdapat 11 sampel yang salah diklasifikasikan sebagai negatif dan 16 sampel sebagai netral.



Gambar 6. Confusion Matrix Rasio 60:20:20



Gambar 7. Grafik Perbandingan

Confusion matrix yang ditampilkan pada Gambar 6 menunjukkan performa model dalam mengklasifikasikan data menjadi tiga kelas, yaitu negatif, netral, dan positif. Model berhasil mengklasifikasikan 58 sampel negatif dengan benar, namun salah mengklasifikasikan 5 sampel sebagai netral dan 13 sampel sebagai positif. Untuk kelas netral, model menunjukkan performa yang sangat baik dengan 281 sampel diklasifikasikan dengan benar, sementara 25 sampel salah diklasifikasikan sebagai negatif dan 19 sampel lainnya sebagai positif. Pada kelas positif, 210 sampel berhasil diklasifikasikan dengan

benar, tetapi terdapat 15 sampel yang salah diklasifikasikan sebagai negatif dan 22 sampel sebagai netral. Hasil ini menunjukkan bahwa model memiliki akurasi tinggi dalam mengklasifikasikan kelas netral dan positif, tetapi menghadapi lebih banyak kesulitan dalam membedakan kelas negatif dari yang lainnya.

C. Perbandingan Performa

Berdasarkan ketiga skenario tersebut, dapat ditemukan matrik akurasi, presisi, *recall*, dan F1-*score*. Dari ketiga skenario pembagian data yang ditampilkan pada Gambar 7, perbedaan performa model terlihat jelas pada berbagai metrik evaluasi, termasuk Akurasi, Presisi, *Recall*, dan F1-*Score*. Pada skenario 80:10:10, model mencapai akurasi tertinggi yaitu 89.20%, diikuti oleh skenario 70:15:15 dengan 88.89%, dan skenario 60:20:20 dengan 84.72%. Ini menunjukkan bahwa semakin banyak data yang digunakan untuk pelatihan, semakin baik kemampuan model dalam melakukan prediksi yang benar secara keseluruhan.

Metrik presisi yang mengukur seberapa baik model meminimalkan *false positives*, juga paling tinggi pada skenario 80:10:10 dengan 89.49%. Model dalam skenario ini lebih mampu memprediksi kelas positif secara akurat dibandingkan dengan dua skenario lainnya. Presisi pada skenario 70:15:15 sedikit lebih rendah di angka 88.77%, sementara skenario 60:20:20 memiliki penurunan yang lebih nyata pada 85.78%.

Recall mengikuti pola yang sama dengan akurasi, di mana skenario 80:10:10 memiliki nilai tertinggi (89.20%) dan skenario 60:20:20 berada di posisi terendah dengan 84.72%. Recall mengukur kemampuan model dalam mendeteksi semua kasus positif, dan data pelatihan yang lebih sedikit tampaknya menyebabkan model kehilangan lebih banyak kasus positif.

F1-*Score*, yang merupakan rata-rata harmonis dari presisi dan *recall*, juga menunjukkan hasil terbaik pada skenario 80:10:10 dengan 88.74%, diikuti oleh 70:15:15 dengan 88.72%. Skenario 60:20:20 menunjukkan penurunan yang lebih besar pada 85.09%. Ini menunjukkan bahwa skenario 80:10:10 memberikan keseimbangan yang lebih baik antara presisi dan recall, membuatnya menjadi skenario dengan performa terbaik di semua aspek.

IV. KESIMPULAN

1. Analisis persepsi hidup sehat masyarakat Indonesia melalui komentar Instagram menggunakan algoritma BERT-LSTM menunjukkan bahwa platform media sosial dapat menjadi sumber data yang efektif untuk memahami pandangan publik tentang gaya hidup sehat.
2. Kombinasi algoritma BERT-LSTM terbukti memiliki kinerja yang tinggi dalam klasifikasi sentimen positif, netral, dan negatif, dengan akurasi terbaik mencapai 89,20%. Ini mengindikasikan bahwa pendekatan ini mampu menangkap nuansa sentimen masyarakat terkait informasi kesehatan.
3. Pemilihan rasio data 80:10:10 untuk pelatihan, validasi, dan pengujian dataset memberikan hasil yang paling optimal dalam memprediksi sentimen dibandingkan rasio lainnya, membuktikan bahwa jumlah data pelatihan yang lebih besar berkontribusi pada peningkatan performa model.

DAFTAR PUSTAKA

[1]. Agustina, D., Rahayu, S., Jasmine, S. T., Sabila, W., & Ramadani, S. (2024). Tinjauan Aplikasi Perencanaan dan Evaluasi Program Germas: Meningkatkan Kesehatan Masyarakat. *Jurnal Pendidikan Tambusai*, *8*(2).

[2]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

[3]. Jakpat. (2023). *Indonesia Consumer Financial Habit 2023 on Gen Z, Millennials and Gen X*. blog.jakpat.net. https://blog.jakpat.net/indonesia-

consumer-financial-habit-2023-on-gen-z-millennials-and-gen-x/

[4]. Katadata. (2023). *Survei Pola Gaya Hidup Sehat*. databoks.katadata.co.id. https://databoks.katadata.co.id/publikasi/2023/08/15/survei-pola-gaya-hidup-sehat

[5]. Pandey, R., & Singh, J. P. (2023). BERT-LSTM model for sarcasm detection in code-mixed social media post. *Journal of Intelligent Information Systems*. https://doi.org/10.1007/s10844-022-00755-z

[6]. Pembengo, N. (2020). *Deteksi Dini Faktor Risiko PTM untuk Mencegah Penyakit Tidak Menular*. 13 Januari. https://dinkes.gorontaloprov.go.id/deteksi-dini-faktor-risiko-ptm-untuk-mencegah-penyakit-tidak-menular/

[7]. Rai, N., Kumar, D., Kaushik, N., Raj, C., & Ali, A. (2022). Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering*. https://doi.org/10.1016/j.ijcce.2022.03.003

[8]. Tanır, H. (2019). Determination of Healthy Life Style Behaviours of the Students in Middle-Adolescence. *World Journal of Education*. https://doi.org/10.5430/wje.v9n1p70

[9]. We are Social. (2023). *Special Report Digital 2023*. wearesocial.com. https://wearesocial.com/id/blog/2023/01/digital-2023/

