Perbandingan Kinerja TF-IDF dan Word Embeddings dalam Klasifikasi Sentimen Ulasan Aplikasi Dana: Studi Kasus Menggunakan DNN, LSTM, dan BiLSTM

Moch Rifky Aulia Adikusumah[1], M Galang Pangestu NH[2], Gevira Zahra Shofa[3]
[1,2,3]Department of Informatics, Faculty of Science and Technology, UIN Sunan Gunung Djati Bandung

Article Info

Article history:

Received -
Revised -
Accepted -

Keywords:

Sentiment Analysis

Deep Learning

TF-IDF

Mobile Application Reviews

ABSTRACT

Pertumbuhan pesat aplikasi mobile, khususnya aplikasi keuangan seperti Dana, telah menghasilkan banyak ulasan pengguna yang menjadi masukan berharga untuk meningkatkan kualitas layanan. Namun, menganalisis sentimen pada ulasan berbahasa Indonesia menghadapi tantangan unik akibat penggunaan bahasa non-formal, slang, dan campur kode. Penelitian ini mengatasi tantangan tersebut dengan membandingkan kinerja dua teknik ekstraksi fitur—Term Frequency-Inverse Document Frequency (TF-IDF) dan Word Embeddings—yang dikombinasikan dengan tiga arsitektur pembelajaran mendalam: Deep Neural Network (DNN), Long Short-Term Memory (LSTM), dan Bidirectional LSTM (BiLSTM) untuk klasifikasi sentimen ulasan aplikasi Dana. Dataset yang digunakan terdiri dari 15.000 ulasan berbahasa Indonesia yang diambil dari Google Play Store, melalui pra-pemrosesan sistematis meliputi pembersihan teks, case folding, normalisasi, tokenisasi, penghapusan stopword, dan stemming. Pelabelan sentimen dilakukan dengan pendekatan berbasis leksikon, diikuti dengan penyeimbangan data melalui oversampling. Tiga skenario dievaluasi: TF-IDF dengan DNN, Word Embeddings dengan BiLSTM, dan TF-IDF dengan BiLSTM. Hasilnya menunjukkan bahwa model TF-IDF + BiLSTM mencapai akurasi tertinggi sebesar 97,08% pada data uji. Namun, analisis kualitatif mengungkapkan bahwa TF-IDF + DNN memberikan klasifikasi yang lebih seimbang dan kontekstual, menjadikannya lebih andal untuk aplikasi dunia nyata meskipun memiliki akurasi sedikit lebih rendah, yaitu 93,26%.

Corresponding Author:

Gevira Zahra Shofa
Jurusan Teknik Informatika, UIN Sunan Gunung Djati Bandung
Email: 1227050050@student.uinsgd.ac.id

1. PENDAHULUAN

Perkembangan teknologi mobile telah mengubah cara konsumen berinteraksi dengan layanan digital. Aplikasi mobile, khususnya aplikasi finansial seperti Dana, telah menjadi bagian integral dari kehidupan sehari-hari masyarakat Indonesia. Sebagai salah satu aplikasi dompet digital terbesar di Indonesia, Dana melayani jutaan pengguna dengan berbagai fitur pembayaran, transfer, dan layanan keuangan lainnya. Ulasan pengguna pada platform aplikasi mobile menjadi sumber informasi yang sangat berharga untuk memahami kepuasan pengguna, mengidentifikasi masalah, dan meningkatkan kualitas layanan.

Analisis sentimen merupakan teknik dalam Natural Language Processing (NLP) yang bertujuan untuk mengklasifikasikan opini atau emosi yang terkandung dalam teks [1]. Dalam konteks ulasan aplikasi mobile, analisis sentimen dapat membantu pengembang aplikasi untuk memahami persepsi pengguna terhadap produk mereka secara otomatis dan dalam skala besar. Hal ini menjadi sangat penting mengingat volume ulasan yang sangat besar dan terus bertambah setiap hari, sehingga analisis manual menjadi tidak praktis dan memakan waktu.

Namun, analisis sentimen pada teks bahasa Indonesia menghadapi tantangan khusus yang berbeda dari bahasa Inggris. Tantangan ini meliputi variasi bahasa non-formal, penggunaan slang words dan bahasa gaul, struktur kalimat yang kompleks, serta keterbatasan sumber daya linguistik seperti leksikon dan corpus yang berkualitas [2]. Selain itu, dalam konteks ulasan aplikasi mobile, pengguna sering menggunakan bahasa campuran (code-switching) antara bahasa Indonesia dan bahasa daerah, singkatan, dan emoticon yang memperumit proses analisis.

Penelitian terdahulu dalam bidang analisis sentimen telah mengeksplorasi berbagai pendekatan, mulai dari metode berbasis leksikon hingga teknik machine learning tradisional seperti Support Vector Machine (SVM) dan Naive Bayes [3]. Metode berbasis leksikon mengandalkan kamus kata-kata yang telah diberi label sentimen, namun pendekatan ini memiliki keterbatasan dalam menangani konteks dan makna implisit. Sementara itu, metode machine learning tradisional bergantung pada feature engineering manual yang membutuhkan domain expertise yang mendalam.

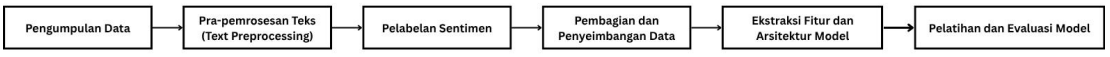
Dengan kemajuan dalam deep learning, arsitektur neural network seperti Deep Neural Network (DNN), Long Short-Term Memory (LSTM), dan Bidirectional LSTM (BiLSTM) telah menunjukkan performa yang superior dalam tugas klasifikasi teks [4]. DNN mampu mempelajari representasi fitur yang kompleks secara otomatis, sedangkan LSTM dan BiLSTM memiliki kemampuan untuk memahami dependensi jangka panjang dalam sekuens teks. Penelitian sebelumnya menunjukkan bahwa kombinasi teknik feature extraction yang tepat dengan arsitektur deep learning dapat meningkatkan akurasi klasifikasi sentimen secara signifikan [5].

Lebih lanjut, pemilihan metode feature extraction juga menjadi faktor krusial dalam keberhasilan model analisis sentimen. Term Frequency-Inverse Document Frequency (TF-IDF) merupakan metode tradisional yang telah terbukti efektif dalam representasi teks, sementara Word Embeddings menawarkan representasi semantik yang lebih kaya dan dapat menangkap hubungan makna antar kata [6]. Kombinasi antara metode feature extraction yang tepat dengan arsitektur deep learning yang sesuai dapat menghasilkan model yang optimal untuk domain aplikasi tertentu.

Tujuan dari penelitian ini adalah membandingkan kinerja tiga arsitektur neural network yang berbeda untuk klasifikasi sentimen ulasan aplikasi Dana: TF-IDF dengan DNN, Word Embeddings dengan LSTM, dan TF-IDF dengan BiLSTM. Penelitian ini diharapkan dapat memberikan wawasan tentang teknik feature extraction dan arsitektur model yang paling efektif untuk analisis sentimen dalam bahasa Indonesia, khususnya pada domain ulasan aplikasi mobile finansial. Kontribusi utama penelitian ini adalah evaluasi komprehensif yang tidak hanya mempertimbangkan metrik kuantitatif seperti akurasi, tetapi juga analisis kualitatif untuk memahami perilaku model dalam kondisi nyata.

2. METODE

Metodologi penelitian ini disusun secara sistematis yang mencakup beberapa tahapan utama, mulai dari pengumpulan data, pra-pemrosesan teks, pelabelan sentimen, hingga pemodelan dan evaluasi. Alur penelitian diilustrasikan pada Gambar 1.



Gambar 1. *Alur Diagram Metode*

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah ulasan (review) pengguna aplikasi Dana di Google Play Store. Proses pengumpulan data dilakukan dengan teknik *scraping* menggunakan library google-play-scraper pada Python. Sebanyak 15.000 ulasan yang paling relevan (*most relevant*) dalam bahasa Indonesia (lang='id') dikumpulkan untuk membentuk dataset awal.

2.2 Pra-pemrosesan Teks (Text Preprocessing)

Untuk menyiapkan data teks mentah sebelum dianalisis oleh model, dilakukan serangkaian langkah pra-pemrosesan secara sistematis agar data menjadi bersih, seragam, dan representatif terhadap konteks sentimen. Adapun tahapan-tahapan tersebut dijelaskan sebagai berikut:

a. Pembersihan data awal

Langkah pertama adalah pembersihan data awal, yaitu menghapus kolom-kolom yang tidak relevan untuk analisis sentimen, seperti *reviewId*, *userName*, *userImage*, dan *replyContent*. Penghapusan ini bertujuan untuk menyederhanakan dataset dan fokus pada informasi yang berkaitan langsung dengan isi ulasan.

b. Text cleaning

Selanjutnya dilakukan proses text cleaning, yang mencakup penghapusan elemen-elemen yang dianggap sebagai noise dalam teks. Ini meliputi penghapusan emotikon, *mentions* (@username), *hashtags* (#topic), karakter non-alfanumerik, angka, dan tanda baca.

c. Case folding

Proses case folding kemudian dilakukan dengan mengubah semua huruf dalam teks ulasan menjadi huruf kecil (*lowercase*) untuk menyeragamkan penulisan kata dan memudahkan proses analisis selanjutnya.

d. Normalisasi kata

Berikutnya adalah normalisasi kata, yaitu mengubah kata-kata tidak baku atau *slang* ke dalam bentuk baku menggunakan kamus *slang* yang telah didefinisikan secara manual. Misalnya, kata "bgt" diubah menjadi "banget", dan "gk" menjadi "tidak".

e. Tokenisasi

Berikutnya adalah normalisasi kata, yaitu mengubah kata-kata tidak baku atau *slang* ke dalam bentuk baku menggunakan kamus *slang* yang telah didefinisikan secara manual. Misalnya, kata "bgt" diubah menjadi "banget", dan "gk" menjadi "tidak".

f. Filtering

Langkah berikutnya adalah filtering atau penghapusan *stopword*, yaitu kata-kata umum yang tidak memiliki bobot sentimen. Proses ini menggunakan tiga sumber daftar *stopword*: (1) daftar *stopword* bahasa Indonesia dari library Sastrawi, (2) daftar *stopword* bahasa Inggris dari library NLTK, dan (3) daftar *stopword* kustom berisi kata-kata domain-spesifik seperti "dana", "aplikasi", "transaksi", "akun", dan kata umum lainnya yang dianggap tidak relevan dalam penentuan sentimen.

g. Stemming

Terakhir, dilakukan proses stemming, yaitu mengubah setiap kata ke bentuk dasarnya (*root word*) menggunakan library Sastrawi. Stemming membantu mengurangi variasi kata dan menyederhanakan korpus untuk meningkatkan konsistensi dalam analisis.

Tahapan pra-pemrosesan teks ini dirancang untuk memastikan bahwa data masukan memiliki kualitas yang optimal sebelum digunakan dalam tahap pemodelan. Dengan data yang bersih dan terstruktur, performa model dalam melakukan analisis sentimen diharapkan menjadi lebih akurat dan dapat diandalkan.

2.3 Pelabelan Sentimen

Karena dataset awal tidak dilengkapi dengan label sentimen, proses pelabelan dilakukan secara otomatis menggunakan pendekatan berbasis leksikon. Metode ini memanfaatkan kamus kata-kata positif dan negatif untuk menghitung skor sentimen dari setiap ulasan.

Setiap kata dalam teks ulasan yang telah melalui tahap pra-pemrosesan akan dicocokkan dengan daftar kata dalam leksikon positif dan negatif. Untuk setiap kecocokan dengan kata positif, skor sentimen akan bertambah, sedangkan kecocokan dengan kata negatif akan mengurangi skor tersebut. Skor akhir dari setiap ulasan diperoleh dengan menjumlahkan seluruh kontribusi kata-kata tersebut.

Berdasarkan skor yang diperoleh, label sentimen kemudian ditentukan menggunakan aturan berikut: ulasan diberi label **positif** jika skor total lebih besar dari nol, **negatif** jika skor total kurang dari nol, dan **netral** jika skor total sama dengan nol. Dengan pendekatan ini, setiap ulasan dapat diklasifikasikan secara sederhana namun cukup efektif ke dalam tiga kategori sentimen utama tanpa memerlukan anotasi manual.

2.4. Pembagian dan Penyeimbangan Data

Dataset yang telah dilabeli kemudian dibagi menjadi data latih (80%) dan data uji (20%). Berdasarkan analisis distribusi kelas, ditemukan adanya ketidakseimbangan jumlah data antar kelas sentimen. Untuk mengatasi hal ini, teknik oversampling diterapkan pada data latih. Kelas minoritas (negatif dan netral) digandakan secara acak hingga jumlah sampelnya setara dengan kelas mayoritas (positif). Langkah ini bertujuan agar model tidak bias terhadap kelas mayoritas selama proses pelatihan.

2.5 Ekstraksi Fitur dan Arsitektur Model

Penelitian ini dirancang untuk membandingkan performa tiga skenario berbeda yang menggabungkan metode ekstraksi fitur dan arsitektur model deep learning dalam klasifikasi sentimen ulasan teks. Ketiga skenario ini dirancang untuk mengevaluasi sejauh mana kombinasi teknik representasi teks dan struktur jaringan memengaruhi akurasi model.

a. Skenario 1: TF-IDF + Deep Neural Network (DNN)

Menggabungkan metode ekstraksi fitur **TF-IDF** dengan arsitektur **Deep Neural Network (DNN)**. Dalam skenario ini, ulasan teks diubah menjadi representasi numerik menggunakan *TfidfVectorizer*, dengan membatasi jumlah fitur pada 10.000 kata dengan frekuensi kemunculan tertinggi. Vektor hasil ekstraksi kemudian menjadi input bagi model DNN sekuensial yang terdiri dari tiga lapisan tersembunyi dengan jumlah unit sebesar 2048, 1024, dan 512, masing-masing menggunakan fungsi aktivasi ReLU. Untuk mengurangi overfitting, setiap lapisan diikuti oleh lapisan Dropout dengan rasio 0.7. Lapisan output menggunakan fungsi aktivasi Softmax untuk mengklasifikasikan ulasan ke dalam tiga kelas sentimen: positif, negatif, dan netral.

b. Skenario 2: Word Embeddings + BiLSTM

Menggunakan pendekatan berbasis **word embeddings** yang dikombinasikan dengan arsitektur **Bidirectional LSTM (BiLSTM)**. Dalam tahap ekstraksi fitur, teks diubah menjadi sekuens integer menggunakan *Tokenizer* dengan batas 20.000 kata teratas, kemudian di-*padding* menjadi panjang sekuens tetap sebanyak 200 token. Setiap token kemudian dipetakan ke dalam ruang vektor padat berdimensi 256 melalui lapisan *Embedding*. Arsitektur model terdiri dari dua lapisan BiLSTM bertingkat, masing-masing dengan 256 dan 128 unit,

yang memungkinkan model menangkap informasi konteks dari dua arah (maju dan mundur). Untuk regularisasi, digunakan lapisan Dropout dengan rasio 0.5 setelah masing-masing lapisan BiLSTM, dan diakhiri dengan lapisan output Softmax.

c. Skenario 3: TF-IDF + BiLSTM

Menguji kombinasi **TF-IDF** dengan arsitektur **BiLSTM**, yang secara arsitektural identik dengan Skenario 2 namun dengan perbedaan pada metode ekstraksi fitur. Di sini, representasi TF-IDF yang sama seperti pada Skenario 1 digunakan kembali, namun sebelum dimasukkan ke dalam model, vektor 2D tersebut diubah ke format 3D agar kompatibel sebagai input untuk lapisan BiLSTM. Tujuan dari skenario ini adalah mengevaluasi efektivitas BiLSTM saat digunakan dengan representasi fitur tradisional seperti TF-IDF, dibandingkan dengan representasi berbasis embedding.

Dengan membandingkan ketiga skenario ini, penelitian bertujuan mengidentifikasi kombinasi optimal antara teknik representasi teks dan arsitektur jaringan untuk klasifikasi sentimen yang lebih akurat dan efisien.

2.6 Pelatihan dan Evaluasi Model

Setiap model dilatih menggunakan data latih yang telah diseimbangkan. Proses pelatihan menggunakan optimizer 'adam' dan fungsi loss 'sparse_categorical_crossentropy'. Mekanisme EarlyStopping diterapkan untuk memonitor akurasi validasi (val_accuracy) dan menghentikan pelatihan jika tidak ada peningkatan setelah 10 epoch untuk mencegah *overfitting*. Kinerja model dievaluasi pada data uji menggunakan metrik akurasi, laporan klasifikasi (presisi, recall, F1-score), dan confusion matrix.

3. HASIL PENELITIAN

Bagian ini menyajikan hasil kuantitatif dari ketiga skenario pemodelan yang telah diuji, diikuti dengan analisis kualitatif berdasarkan hasil inferensi pada data baru.

3.1 Hasil Kinerja Kuantitatif

Evaluasi dilakukan pada data uji yang belum pernah dilihat oleh model sebelumnya. Rangkuman akurasi dari ketiga model disajikan pada Tabel 1.

Tabel 1. *Perbandingan Akurasi Pengujian Model*

Skenario	Ekstraksi Fitur	Arsitektur Model	Akurasi (Data Uji)
1	TF-IDF	DNN	93.26%
2	Word Embeddings	BiLSTM	85.81%
3	TF-IDF	BiLSTM	97.08%

Dari Tabel 1, model **TF-IDF + BiLSTM** menunjukkan akurasi tertinggi secara keseluruhan (97.08%), diikuti oleh model **TF-IDF + DNN** (93.26%). Model **Word Embeddings + LSTM** menghasilkan akurasi terendah (85.81%).

3.2 Analisis Kualitatif dan Inferensi

Untuk memahami perilaku praktis setiap model, pengujian dilakukan pada beberapa kalimat ulasan baru yang belum pernah ada di dataset. Hasil prediksi disajikan pada Tabel 2.

Tabel 5. *Hasil Prediksi pada Teks Uji Baru*

Teks Uji	TF-IDF + DNN	Word Embeddings + LSTM	TF-IDF + BiLSTM
sangat kecewa sekali di hp saya lag	negatif	positif	positif

aplikasi bagus semoga bisa tangung jawab gak cuma pamer menumenu menjanjikan	positif	positif	positif
akun dana saya tidak bisa dibuka tidak tau kenapa ya?	netral	positif	positif
aplikasi sampah tidak berguna sama sekali sering bermasalah dan gangguan	negatif	positif	positif
keren aplikasinya sangat membantu	positif	netral	positif
saya harap kedepannya ada fitur pinjaman online di dana	netral	positif	positif

Berikut adalah analisis kualitatif dari Tabel 2: Hasil Prediksi pada Teks Uji Baru berdasarkan ketiga model yang dibandingkan:

a. TF-IDF + DNN

Model TF-IDF + DNN menunjukkan performa paling seimbang dengan hasil klasifikasi yang konsisten terhadap konteks kalimat. Model ini berhasil mengenali ulasan negatif, netral, dan positif secara proporsional, menunjukkan kemampuannya dalam menangkap nuansa emosional dalam teks pendek.

b. Word Embeddings + LSTM

Model Word Embeddings + LSTM cenderung bias terhadap sentimen positif. Lima dari enam teks uji diklasifikasikan sebagai positif, termasuk kalimat yang seharusnya bernada negatif atau netral. Hal ini mengindikasikan kurangnya sensitivitas model terhadap ekspresi negatif dalam konteks kalimat.

c. TF-IDF + BiLSTM

Model TF-IDF + BiLSTM menunjukkan bias yang lebih ekstrem dengan mengklasifikasikan seluruh ulasan sebagai positif, tanpa membedakan sentimen yang sebenarnya. Meskipun metrik akurasinya tinggi, model ini gagal melakukan generalisasi pada data uji yang lebih realistis dan bervariasi.

Analisis dari Tabel 2 menunjukkan adanya trade-off antara akurasi kuantitatif dan ketepatan kualitatif. Meskipun model TF-IDF + BiLSTM secara metrik mungkin unggul, performa kualitatifnya sangat bias dan tidak mencerminkan klasifikasi yang adil. Sebaliknya, model TF-IDF + DNN, meskipun lebih sederhana, memberikan hasil yang lebih stabil, kontekstual, dan seimbang, serta lebih layak diterapkan pada data dunia nyata.

4. CONCLUSION

Penelitian ini membandingkan tiga skenario kombinasi ekstraksi fitur dan arsitektur deep learning dalam tugas klasifikasi sentimen teks ulasan aplikasi, yaitu: TF-IDF + DNN, Word Embeddings + LSTM, dan TF-IDF + BiLSTM. Evaluasi dilakukan melalui dua pendekatan, yaitu analisis kuantitatif menggunakan metrik akurasi dan analisis kualitatif berdasarkan inferensi pada data uji baru.

Berdasarkan hasil evaluasi kuantitatif, model TF-IDF + BiLSTM mencatat akurasi tertinggi sebesar 97.08%, diikuti oleh TF-IDF + DNN dengan akurasi 93.26%, dan Word Embeddings + LSTM dengan akurasi 85.81%. Namun, hasil analisis kualitatif menunjukkan bahwa model TF-IDF + BiLSTM dan Word Embeddings + LSTM cenderung memiliki bias terhadap kelas positif, sehingga kurang andal dalam mengenali ulasan negatif dan netral. Sebaliknya, model TF-IDF + DNN mampu mengklasifikasikan sentimen dengan lebih proporsional dan kontekstual, menunjukkan keseimbangan yang baik dalam menangkap nuansa emosional dari teks pendek.

Dengan demikian, meskipun TF-IDF + BiLSTM unggul dalam metrik kuantitatif, model TF-IDF + DNN terbukti lebih stabil dan layak diterapkan dalam konteks nyata, terutama ketika model perlu menangani data ulasan yang beragam dan tidak terstruktur. Hasil ini juga menyoroti pentingnya mempertimbangkan aspek kualitatif dalam mengevaluasi model pembelajaran mesin, bukan hanya bergantung pada metrik akurasi semata.

DAFTAR PUSTAKA

[1] Winarni, L., Amalia, S., Lestari, D.P., et al. (2023). Sentiment analysis of Indonesian datasets based on a hybrid

deep-learningstrategy. *Journal of Big Data*, 10, 82. https://doi.org/10.1186/s40537-023-00782-9

[2] Hadi, H.U., Sari, I.P., & Wahyuni, E.S. (2022). Emotion classification of Indonesian Tweets using Bidirectional LSTM. *Neural Computing and Applications*, 34, 16499-16517. https://doi.org/10.1007/s00521-022-08186-1

[3] Bera, S., Shrivastava, V.K. (2023). Sentiment analysis from textual data using multiple channels deep learning models. *Journal of Electrical Systems and Information Technology*, 10, 36. https://doi.org/10.1186/s43067-023-00125-x

[4] Khattak, F.K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2023). Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review*, 56(11), 13358-13390. https://doi.org/10.1007/s10462-023-10651-9

[5] Elbagir, S., & Yang, J. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3), 483. https://doi.org/10.3390/electronics9030483

[6] Xiaoyan, Z., Qing, Y., Ao, L., & Xing, W. (2022). GloVe‐CNN‐BiLSTM Model for Sentiment Analysis on Text Reviews. *Journal of Sensors*, 2022, 7212366. https://doi.org/10.1155/2022/7212366