

LAPORAN PENELITIAN

SKEMA PENELITIAN

KEILMUAN - KOM - Dosen Pemula

AREA PENELITIAN :
Sains dan Teknologi era industri 4.0 dan masyarakat 5.0

**AI-DRIVEN USABILITY TESTING: OPTIMIZING MOBILE APP
PERFORMANCE WITH COMBINATION USABILITY CRITERIA AND
SENTIMENT ANALYSIS USING RNN MODELS**



Oleh :

Ketua Peneliti:

Hasan Basri, S.Kom., M.Kom.
NIP. 199205172024061003

Anggota :

Wahyu Noviani Purwanti, S.Si., M.Si.
NIP. 197111291998022001

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS TERBUKA
2025

LEMBAR PENGESAHAN LAPORAN PENELITIAN
LEMBAGA PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT
UNIVERSITAS TERBUKA

1.	a.	Judul Penelitian	:	AI-Driven Usability Testing: Optimizing Mobile App Performance With Combination Usability Criteria And Sentiment Analysis Using RNN Models
	b.	Skema Penelitian	:	KOM - Dosen Pemula
	c.	Area Penelitian	:	Kompetitif Keilmuan
2.		Ketua Peneliti		
	a.	Nama Lengkap & Gelar	:	Hasan Basri, S.kom., M.Kom.
	b.	NIP/NIDN	:	199205172024061003/ 0317059202
	c.	Golongan Kepangkatan	:	III/B
	d.	Jabatan Akademik	:	Dosen Asisten Ahli
	e.	Fakultas	:	FST
	f.	Unit Kerja	:	UT Pusat
	g.	Program Studi	:	S1 Sistem Informasi
3.		Anggota		
	a.	Nama Anggota I	:	Wahyu Noviani Purwanti, S.Si., M.Si.
	b.	NIP Anggota I	:	197111291998022001/ 0029117103
4.	a.	Tahun Penelitian	:	2025
	b.	Lama Penelitian	:	12 Bulan
5.		Biaya Penelitian		
	a.	Diusulkan	:	20.965.000
	b.	Disetujui	:	16.620.000
6.		Sumber Biaya	:	Universitas Terbuka
7.		Pemanfaatan Hasil Penelitian		
	a.	Seminar	:	Nasional/Regional/Internasional***)
	b.	Jurnal	:	UT/Nasional/Internasional***)
8.		Luaran Penelitian	:	Jurnal Teknik Informatika (JUTIF)

Tangerang Selatan, 24 Oktober 2025

Mengetahui,
Dekan FST

Dr. Subekti Nurmawati, M.Si.
NIP. 196705181991032001

Ketua Peneliti

Hasan Basri, S.Kom., M.Kom.
NIP. 199205172024061003

Menyetujui,
Ketua LPPM-UT

Prof. Dewi Artati Padmo Putri, Ph.D.
NIP. 196107241987102003

Kepala Pusat Penelitian Keilmuan

Dr. Mery Noviyanti, S.Si., M.Pd.
NIP. 198111242005012003

DAFTAR ISI

LEMBAR PENGESAHAN LAPORAN PENELITIAN.....	2
DAFTAR ISI	3
DAFTAR GAMBAR.....	5
DAFTAR TABEL	1
ABSTRAK	2
BAB 1 PENDAHULUAN	3
1.1 Latar Belakang.....	3
1.2 Rumusan Masalah.....	4
1.3 Tujuan Penelitian.....	4
1.4 Batasan Penelitian.....	5
1.5 Manfaat Penelitian	5
1.6 Luaran Penelitian.....	6
1.7 State of The Art.....	6
BAB 2 TINJAUAN PUSTAKA	7
2.1 Usability Testing.....	7
2.2 Sentiment Analysis.....	7
2.3 Recurrent Neural Network (RNN)	8
2.4 Penelitian Sebelumnya.....	8
2.5 Gap Analysis.....	9
BAB 3 METODE PENELITIAN.....	10
3.1 Tahapan Penelitian	10
3.2 Pengumpulan Data.....	10
3.2.1 Sumber Data.....	10
3.2.2 Teknik Pengumpulan	10
3.2.3 Sampel Data.....	11
3.3 Preprocessing Data	11
3.4 Pengembangan Model.....	11
3.4.1 Algoritma yang Digunakan.....	11
3.4.2 Pelatihan Model	11
3.4.3 Evaluasi Model.....	11
3.4.4 Akurasi.....	12
3.4.5 F1-Score	12
3.4.6 Hamming Loss.....	12
3.5 Roadmap Penelitian.....	12
3.6 Jadwal Penelitian	14

BAB IV HASIL DAN PEMBAHASAN	15
4.1 Analisis Distribusi Kelas.....	15
4.2 Data Preprocessing	16
4.2.1 Lowercasing.....	16
4.2.2 Penghapusan Tanda Baca	16
4.2.3 Tokenisasi	17
4.2.4 Analisis Panjang Ulasan.....	17
4.2.5 Padding	17
4.2.6 Encoding Label	18
4.3 Pembagian Dataset.....	18
4.4 Pembangunan Arsitektur RNN	19
4.4.1 Embedding Layer (128 dimensi)	19
4.4.2 SpatialDropout1D (rate = 0.2)	19
4.4.3 LSTM Layer (100 unit)	19
4.4.4 Dense Layer dengan Softmax (5 kelas).....	19
4.4.5 Pengaturan Pelatihan Model	20
4.5 Pelatihan Model.....	20
4.5.1 Hasil Pelatihan	20
4.6 Evaluasi Model.....	22
BAB V KESIMPULAN	28
DAFTAR PUSTAKA.....	29
Lampiran	30
1. FGD	30

DAFTAR GAMBAR

Gambar 1 Design Penelitian	10
Gambar 2 Roadmap Penelitian	12
Gambar 3 Model Training	20
Gambar 4 Model Accuracy.....	21
Gambar 5 Model Loss	22
Gambar 6 Confusion Matrix.....	25

DAFTAR TABEL

Tabel 1 Jadwal kegiatan	14
Tabel 2 Distribusi data.....	15
Tabel 3 Classification Report	23

ABSTRAK

Analisis ulasan pengguna pada platform digital seperti Google Play Store merupakan sumber informasi penting untuk mengevaluasi kualitas aplikasi, khususnya dari sisi usability dan sentimen. Penelitian ini mengembangkan model klasifikasi berbasis Recurrent Neural Network (RNN) untuk mengotomatisasi proses pemetaan ulasan pengguna ke dalam kategori usability-sentiment. Dataset awal terdiri dari 10.000 ulasan dengan delapan label, namun analisis distribusi menunjukkan ketidakseimbangan ekstrem pada tiga kelas dengan jumlah kurang dari 70 sampel. Untuk mencegah bias model dan ketidakstabilan decision boundary, ketiga kelas tersebut dieliminasi sehingga menyisakan 9.849 ulasan pada lima kelas utama. Tahap preprocessing mencakup normalisasi teks, penghapusan tanda baca, tokenisasi, dan padding hingga panjang 17 token berdasarkan distribusi panjang teks. Label diencode menggunakan teknik label encoding dan one-hot encoding. Dataset kemudian dibagi secara stratified menjadi train (80%), validation (10%), dan test (10%) untuk menjaga konsistensi proporsi label. Model RNN dibangun menggunakan arsitektur embedding 128 dimensi, SpatialDropout1D, dan LSTM dengan 100 unit. Model dilatih selama 10 epoch dan mencapai akurasi validasi 92.59%. Pada pengujian menggunakan test set, model memperoleh akurasi keseluruhan sebesar 91.68%. Performa sangat baik ditunjukkan pada kelas mayoritas dengan F1-score di atas 0.92, sedangkan kelas minoritas seperti User Satisfaction - Negative dan Learnability - Positive menunjukkan penurunan kinerja akibat keterbatasan jumlah sampel dan tingginya ambiguitas semantik. Analisis confusion matrix mengungkap bahwa banyak ulasan pendek bersifat ambigu dan memuat lebih dari satu sinyal sentimen atau aspek usability, misalnya ulasan seperti “lumayan, tapi sering macet” yang secara bersamaan merefleksikan apresiasi ringan dan keluhan teknis. Secara keseluruhan, penelitian ini menunjukkan bahwa RNN mampu memberikan performa kuat pada kelas mayoritas, namun masih menghadapi tantangan dalam menangani kelas minoritas dan interpretasi ulasan pendek yang ambigu. Temuan ini membuka peluang perbaikan melalui penyeimbangan data, augmentasi teks, serta penggunaan arsitektur berbasis attention atau transformer seperti BERT untuk meningkatkan kemampuan pemahaman konteks pada penelitian selanjutnya.

Kata Kunci : *Usability Testing, Sentiment Analysis, Multi-Class Classification, Recurrent Neural Network, LSTM, Mobile Application Evaluation*

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Dalam era digital yang semakin berkembang, aplikasi mobile telah menjadi bagian tak terpisahkan dari kehidupan sehari-hari. Hingga kuartal kedua tahun 2024, terdapat lebih dari 2,26 juta aplikasi yang tersedia di Google Play Store, mencakup berbagai kategori seperti game, edukasi, bisnis, gaya hidup, dan alat (Clement, 2024). Tingginya tingkat persaingan ini mendorong pengembang untuk menghadirkan aplikasi yang tidak hanya inovatif, tetapi juga memiliki tingkat usability yang tinggi. Hal ini menjadi kunci dalam memberikan pengalaman pengguna yang optimal dan meningkatkan daya saing aplikasi di pasar yang sangat kompetitif.

Usability menjadi faktor kunci yang memengaruhi keberhasilan aplikasi mobile. Aplikasi dengan usability yang baik memungkinkan pengguna untuk menyelesaikan tugas secara efisien, meminimalkan kesalahan, dan memberikan tingkat kepuasan yang tinggi (Weichbroth, 2020) (Capilla et al., 2020). Namun, metode konvensional seperti usability testing langsung, yang melibatkan pengguna dalam skenario terkontrol, seringkali terkendala oleh waktu, biaya, dan sumber daya yang besar (Ahmad et al., 2023).

Pada tahun pertama, penelitian menggabungkan Usability Testing dan Sentiment Analysis menggunakan Support Vector Machine (SVM) untuk mengevaluasi aplikasi mobile berdasarkan ulasan pengguna. Metode ini terbukti efisien dalam mengidentifikasi kekuatan dan kelemahan aplikasi, meskipun terdapat keterbatasan, seperti kinerja SVM pada data kompleks dan analisis kriteria usability implisit, seperti sindiran.

Untuk menjawab tantangan tersebut, penelitian lanjutan ini bertujuan mengintegrasikan pendekatan berbasis Artificial Intelligence (AI), khususnya Recurrent Neural Networks (RNN), dalam analisis usability. RNN dikenal karena kemampuannya dalam menangani data urutan, seperti teks ulasan pengguna, sehingga diharapkan dapat memberikan hasil analisis yang lebih akurat dibandingkan metode sebelumnya (Hassan & Shoaib, 2020) (Arras et al., 2019). Pendekatan ini akan mengembangkan model yang mampu menganalisis ulasan

pengguna secara lebih mendalam, termasuk menangkap pola sentimen yang kompleks, serta memberikan wawasan usability yang lebih relevan. Dengan memanfaatkan ulasan pengguna secara otomatis dan berbasis data nyata, penelitian ini diharapkan mampu memberikan solusi yang hemat biaya dan sumber daya, sekaligus mendukung pengembang dalam menciptakan aplikasi yang lebih user-centric.

Penelitian ini tidak hanya memberikan kontribusi dalam bidang usability testing, tetapi juga memperluas penerapan teknologi AI dalam pengembangan aplikasi mobile. Pendekatan yang diusulkan berpotensi menjadi standar baru dalam evaluasi perangkat lunak, yang tidak hanya dapat diterapkan pada aplikasi mobile, tetapi juga pada website dan jenis aplikasi lainnya. Lebih jauh lagi, jika penelitian ini dikembangkan, pendekatan tersebut dapat mendukung evaluasi kepuasan mahasiswa Universitas Terbuka dalam pembelajaran e-learning, sehingga memberikan wawasan yang lebih mendalam bagi institusi dalam meningkatkan kualitas layanan pendidikan daring.

1.2 Rumusan Masalah

Berdasarkan hasil penelitian sebelumnya yang berhasil menunjukkan potensi kombinasi usability testing dengan sentiment analysis dalam menganalisis ulasan pengguna, beberapa masalah berikut muncul untuk diteliti lebih lanjut:

- a. Sejauh mana model Recurrent Neural Network (RNN) dapat meningkatkan keandalan klasifikasi dari kombinasi kriteria usability dan sentimen ulasan pengguna dibandingkan metode yang digunakan sebelumnya?
- b. Apa saja tantangan teknis yang dihadapi dalam penerapan RNN untuk mengklasifikasikan ulasan pengguna, terutama dalam menangani bahasa informal, ambiguitas, dan bias data?
- c. Bagaimana kombinasi usability testing berbasis sentiment analysis dan RNN dapat memberikan wawasan yang lebih mendalam mengenai aspek usability aplikasi mobile?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut:

- a. Menilai sejauh mana algoritma RNN dapat meningkatkan keandalan dan akurasi klasifikasi dari kombinasi usability testing criteria dan sentimen dibandingkan dengan

metode sebelumnya.

- b. Mengidentifikasi dan mengatasi tantangan teknis dalam penerapan RNN, khususnya dalam menangani ulasan pengguna yang menggunakan bahasa informal, mengandung ambiguitas, atau bias.
- c. Menghasilkan pendekatan berbasis AI yang lebih adaptif dan mendalam untuk memberikan wawasan yang lebih kaya terhadap aspek usability aplikasi mobile.

1.4 Batasan Penelitian

Batasan dalam penelitian ini adalah sebagai berikut:

- a. Penelitian ini dibatasi pada dua aplikasi mobile yang dipilih dari kategori yang sama atau sejenis di Google Play Store. Pemilihan aplikasi dilakukan berdasarkan jumlah ulasan yang signifikan untuk memastikan data yang representatif.
- b. Ulasan pengguna yang dianalisis dibatasi pada ulasan yang diposting pada bulan tertentu. Hal ini bertujuan untuk menjaga konsistensi data dan meminimalkan pengaruh temporal.
- c. Penelitian ini mengintegrasikan kombinasi dari usability testing dan sentiment analysis dengan algoritma klasifikasi Recurrent Neural Network (RNN). Teknik sentiment analysis lainnya, seperti lexicon-based atau machine learning non-RNN, tidak akan menjadi fokus utama.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

- a. Penelitian ini memberikan kontribusi dalam pengembangan metode usability testing berbasis sentiment analysis yang dioptimalkan dengan Recurrent Neural Network (RNN) untuk meningkatkan keakuratan dan efisiensi pengujian usability pada aplikasi mobile
- b. Penelitian ini menyediakan pendekatan berbasis data yang efisien bagi pengembang untuk memahami pengalaman pengguna, mengidentifikasi kekuatan dan kelemahan aplikasi, serta membuat keputusan yang lebih tepat dalam perbaikan aplikasi.
- c. Dengan memanfaatkan teknologi AI, penelitian ini membantu meningkatkan kualitas aplikasi mobile yang lebih responsif terhadap kebutuhan pengguna, mengurangi biaya dan waktu pengujian konvensional, serta memberikan panduan bagi penelitian lanjutan.

1.6 Luaran Penelitian

Dalam penelitian ini penulis ingin memiliki target luaran sebagai berikut:

- a. Terbit dalam Jurnal nasional terakreditasi

1.7 State of The Art

Penelitian ini menawarkan inovasi Metodologi Usability Testing. Usability testing konvensional, seperti kuesioner, heuristic evaluation, atau expert reviews, sering kali membutuhkan banyak waktu, biaya, dan sumber daya (Ahmad et al., 2023). Penelitian ini memanfaatkan ulasan pengguna dari Google Play Store sebagai sumber data, memungkinkan evaluasi usability yang lebih efisien, berbasis data nyata, dan skalabel.

Sentiment analysis telah terbukti memberikan wawasan tentang opini pengguna secara otomatis dari teks ulasan (Chahar & Dubey, 2024). Penelitian sebelumnya menunjukkan bahwa kombinasi usability criteria dan sentiment analysis dapat menjadi metode alternatif yang menjanjikan untuk usability testing, namun masih menghadapi kendala dalam keakuratan klasifikasi, terutama pada ulasan dengan bahasa informal, ambiguitas, atau bias.

RNN, algoritma deep learning untuk data sekuensial, memungkinkan pengenalan pola temporal dalam ulasan pengguna. Penerapannya pada sentiment analysis diharapkan meningkatkan akurasi dibandingkan pendekatan tradisional seperti SVM atau Naïve Bayes. Kombinasi usability criteria dan sentiment analysis dengan RNN menciptakan evaluasi yang lebih adaptif dan representatif. Pendekatan ini menawarkan metode usability testing berbasis AI yang holistik, efisien, dan relevan untuk industri mobile, memberikan kontribusi signifikan dalam pengembangan aplikasi yang user-centric dan berkualitas.

BAB 2 TINJAUAN PUSTAKA

2.1 Usability Testing

Usability testing merupakan proses evaluasi untuk menentukan sejauh mana suatu sistem, produk, atau layanan dapat digunakan oleh pengguna tertentu untuk mencapai tujuan dengan efektivitas, efisiensi, dan kepuasan (Hertzum, 2020). Jakob Nielsen (2012) mengidentifikasi lima aspek utama usability testing, yaitu:

- a. Learnability: Kemudahan pengguna untuk memahami sistem.
- b. Efficiency: Kecepatan pengguna dalam menyelesaikan tugas.
- c. Memorability: Kemudahan pengguna mengingat cara menggunakan sistem.
- d. Error Tolerance: Toleransi sistem terhadap kesalahan pengguna.
- e. Satisfaction: Tingkat kepuasan pengguna.

Metode konvensional seperti heuristic evaluation dan expert reviews sering digunakan, tetapi memerlukan waktu, biaya, dan sumber daya yang besar (Privitera, 2019) (Fernández & Macías, 2021). Oleh karena itu, pendekatan berbasis data seperti sentiment analysis menjadi alternatif yang lebih efisien untuk mengevaluasi usability aplikasi mobile.

2.2 Sentiment Analysis

Sentiment analysis adalah teknik untuk mengekstraksi opini, emosi, atau sentimen dari teks, yang sering digunakan dalam analisis ulasan produk atau media sosial (Nandwani & Verma, 2021). Tiga pendekatan utama dalam sentiment analysis (Mohamad Sham & Mohamed, 2022), adalah sebagai berikut:

- a. Lexicon-based, menggunakan kamus kata dengan nilai sentimen tertentu.
- b. Machine Learning-based, menggunakan algoritma seperti SVM, Naïve Bayes, dan Random Forest.
- c. Hybrid, menggabungkan metode leksikon dan machine learning.

Dalam penelitian sebelumnya dalam memanfaatkan ulasan pengguna untuk mengevaluasi pengalaman pengguna terhadap aplikasi menghadapi tantangan, terutama dalam menangani:

- a. Bahasa informal atau slang dalam ulasan.

- b. Ambiguitas, seperti penggunaan sarkasme.
- c. Singkatan dalam penulisan seperti untuk menulis kata “yang” pengguna hanya menuliskan “yng” atau “yg”

Penelitian sebelumnya menggunakan SVM menunjukkan hasil yang menjanjikan, tetapi masih memiliki keterbatasan dalam akurasi pada data yang kompleks dan heterogen.

2.3 Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) adalah algoritma deep learning yang dirancang untuk data sekuensial seperti teks. RNN memiliki kemampuan untuk menangkap pola temporal dan hubungan antar kata dalam ulasan pengguna. Variasi RNN, seperti Long Short-Term Memory (LSTM) dan Gated Recurrent Unit (GRU), menawarkan solusi untuk masalah vanishing gradient pada data sekuensial yang panjang (Noh, 2021). Keunggulan RNN meliputi (Zhang et al., 2020) :

- a. Kemampuan untuk menganalisis konteks dari urutan kata dalam teks.
- b. Efektivitas dalam menangani data yang memiliki ketergantungan temporal.

RNN bekerja dengan menggunakan koneksi berulang (recurrent) dalam jaringan neuron. Setiap langkah waktu (time step) memproses input saat ini dan menyimpan informasi dari langkah sebelumnya dalam memori tersembunyi (hidden state). Hidden state ini memungkinkan RNN untuk mempertimbangkan konteks dari data sebelumnya ketika memproses data saat ini (Hartvigsen et al., 2020).

Dalam konteks Sentiment Analysis, RNN diharapkan mampu mengatasi tantangan Sentiment Analysis pada ulasan pengguna, terutama yang mengandung pola bahasa informal atau ambigu.

2.4 Penelitian Sebelumnya

Penelitian sebelumnya menggunakan pendekatan kombinasi usability testing dan sentiment analysis untuk mengevaluasi aplikasi mobile berdasarkan ulasan pengguna di Google Play Store. Dalam penelitian tersebut, Support Vector Machine (SVM) digunakan untuk mengklasifikasikan sentimen positif dan negatif serta menghubungkannya dengan kriteria usability seperti Learnability, Efficiency, Error Tolerance, dan Satisfaction. Hasil penelitian menunjukkan akurasi SVM sebesar 75,20% setelah tuning parameter, namun masih terdapat keterbatasan dalam menangani ulasan dengan bahasa informal dan ambiguitas.

2.5 Gap Analysis

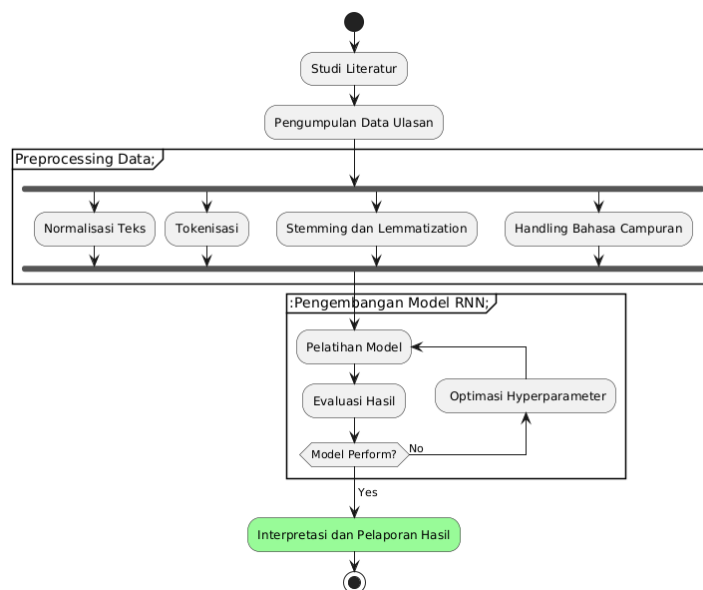
Celah penelitian yang diidentifikasi meliputi:

- a. Keterbatasan metode konvensional seperti SVM dalam menangani data ulasan yang kompleks.
- b. Minimnya penelitian yang mengintegrasikan sentiment analysis dengan usability criteria secara komprehensif.
- c. Tantangan dalam menangani komentar pengguna yang sering menggunakan singkatan, bahasa campuran (Indonesia dengan bahasa daerah atau bahasa Inggris), yang dapat menyulitkan analisis otomatis
- d. Penggunaan gaya bahasa ironi seperti sarkasme oleh pengguna, yang dapat mempersulit algoritma dalam menginterpretasikan sentimen sebenarnya dari komentar tersebut.

BAB 3 METODE PENELITIAN

3.1 Tahapan Penelitian

Penelitian ini menggunakan desain eksperimen kuantitatif yang bertujuan untuk mengevaluasi efektivitas algoritma RNN dalam mengklasifikasikan kombinasi Usability Testing Criteria dan Sentiment Analysis pada aplikasi mobile. Tahapan penelitian meliputi pengumpulan data ulasan pengguna, preprocessing data, pengembangan model, dan evaluasi hasil. Adapun tahapannya dapat dilihat pada gambar sebagai berikut:



Gambar 1 Design Penelitian

3.2 Pengumpulan Data

3.2.1 Sumber Data

Data yang digunakan dalam penelitian ini diperoleh dari ulasan pengguna aplikasi mobile di Google Play Store. Ulasan yang diambil mencakup komentar yang ditulis dalam bahasa Indonesia, baik yang menggunakan bahasa formal maupun informal.

3.2.2 Teknik Pengumpulan

Data dikumpulkan dengan teknik crawling menggunakan library Python seperti BeautifulSoup, Selenium atau google-play-scraper . Data yang diambil meliputi:

- Komentar pengguna

- b. Rating bintang
- c. Waktu komentar

Komentar yang tidak relevan, seperti spam atau ulasan tanpa teks, akan dihapus selama preprocessing.

3.2.3 Sampel Data

Data ulasan diambil dari aplikasi dalam kategori tertentu (misalnya, perbankan atau pendidikan) dengan jumlah minimal 10.000 ulasan. Sampel dipilih untuk memastikan representasi yang memadai dan keakuratan hasil analisis.

3.3 Preprocessing Data

Preprocessing data dilakukan untuk membersihkan dan mempersiapkan data ulasan pengguna agar siap dianalisis. Adapun tahapan preprocessing data meliputi:

- a. Normalisasi Teks, menghapus tanda baca, spasi berlebih, dan karakter non-ASCII.
- b. Tokenisasi, memecah teks ulasan menjadi kata-kata individu.
- c. Stemming dan Lemmatization, mengubah kata-kata menjadi bentuk dasar untuk konsistensi analisis.
- d. Handling Bahasa Campuran, mengidentifikasi bahasa campuran dan konversi singkatan jika memungkinkan.

3.4 Pengembangan Model

3.4.1 Algoritma yang Digunakan

Penelitian ini menggunakan Recurrent Neural Network (RNN), khususnya arsitektur Long Short-Term Memory (LSTM) untuk menangani ketergantungan temporal dalam data ulasan pengguna. LSTM dipilih karena kemampuannya untuk menangani data sekuensial dengan panjang urutan yang bervariasi.

3.4.2 Pelatihan Model

Pada tahap ini, data dibagi menjadi data pelatihan (80%) dan data pengujian (20%). Representasi teks dilakukan menggunakan word embeddings seperti Word2Vec atau GloVe. Parameter model dioptimalkan menggunakan algoritma Adam Optimizer dengan fungsi loss categorical cross-entropy.

3.4.3 Evaluasi Model

Kinerja model dalam penelitian ini dievaluasi menggunakan tiga metrik utama yang memberikan gambaran komprehensif tentang performanya:

3.4.4 Akurasi

Mengukur persentase prediksi yang benar, memberikan indikasi keseluruhan keandalan model.

3.4.5 F1-Score

Menyediakan keseimbangan yang optimal antara precision dan recall, terutama penting dalam dataset yang tidak seimbang.

3.4.6 Hamming Loss

Mengidentifikasi jumlah kesalahan klasifikasi pada tugas multi-label, membantu menganalisis seberapa sering prediksi tidak sesuai dengan label yang sebenarnya.

3.5 Roadmap Penelitian

Penelitian ini dirancang untuk dilaksanakan selama tiga tahun, dengan setiap tahun berfokus pada tahapan yang berbeda sesuai dengan roadmap penelitian yang telah dibuat. Saat ini, penelitian berada pada tahun kedua yang berfokus pada pengembangan dan pelatihan model RNN berbasis data ulasan pengguna. Berikut adalah roadmap penelitian secara keseluruhan:



Gambar 2 Roadmap Penelitian

Usability testing menggunakan sentiment analysis dan SVM classification. Tahun pertama, penelitian difokuskan pada eksplorasi awal teknologi Sentiment

Analysis yang dikombinasikan dengan usability testing criteria dengan menggunakan Support Vector Machine (SVM). Pada tahap ini, dasar-dasar sentiment analysis dibangun untuk memahami pola sentimen pengguna dari ulasan aplikasi. Hasil penelitian pada tahun pertama menunjukkan kinerja yang baik, di mana pendekatan SVM berhasil mengklasifikasikan sentimen dengan akurasi yang memadai. Temuan ini membuktikan potensi besar dari penggunaan kombinasi Usability testing criteria dan Sentiment Analysis menggunakan SVM dalam evaluasi usability aplikasi mobile, sekaligus membuka peluang untuk pengembangan lebih lanjut pada tahap berikutnya.

Optimizing mobile app usability testing through sentiment analysis and RNN classification of playstore review. Pada tahun kedua, penelitian berfokus pada optimalisasi pengujian usability dengan menggunakan Usability Testing Criteria yang dikombinasikan dengan Sentiment Analysis menggunakan Recurrent Neural Network (RNN). Pendekatan ini bertujuan untuk meningkatkan akurasi dalam menganalisis sentimen dan memahami pola bahasa kompleks seperti informalitas, campuran bahasa, serta ironi. Tahun ini merupakan inti dari penelitian, di mana pengembangan model RNN diharapkan memberikan kontribusi signifikan terhadap hasil yang lebih presisi.

Integrating multimodal data for Enhanced Usability testing: Combining Sentiment analysis, RNN, and User Interaction Data. Penelitian akan memperluas cakupan dengan mengintegrasikan data multimodal, seperti ulasan teks dan data interaksi pengguna. Tujuannya adalah menciptakan model usability testing yang lebih komprehensif, efektif, dan mampu memberikan wawasan mendalam tentang kebutuhan pengguna.

3.6 Jadwal Penelitian

Penelitian ini akan dilaksanakan di Program Studi Sistem Informasi Fakultas Sains dan Teknologi Universitas Terbuka dengan sampel ulasan pengguna aplikasi mobile pada google playstore.

Tabel 1 Jadwal kegiatan

Kegiatan	Sub-Kegiatan	Bulan							
		3	4	5	6	7	8	9	10
Studi Literatur	Studi Literatur								
Pengumpulan Data	Pengumpulan Data								
Preprocessing Data	Normalisasi teks								
	Tokenisasi								
	Lemmatization								
	Penanganan bahasa campuran								
Pengembangan Model RNN	Mendesain arsitektur model								
	Implementasi model								
Pelatihan dan Optimasi Model	Pelatihan awal model								
	Optimasi hypermeter								
Evaluasi dan Validasi	Evaluasi model								
	Validasi model dengan data baru								
Interpretasi dan Penyusunan Laporan	Interpretasi hasil								
	Penulisan Laporan Penelitian								
	Penulisan Jurnal								

BAB IV HASIL DAN PEMBAHASAN

4.1 Analisis Distribusi Kelas

Tahap awal penelitian dimulai dengan meninjau distribusi label pada dataset berisi 10.000 ulasan Google Play Store. Pemeriksaan melalui fungsi `value_counts()` menunjukkan bahwa dataset terdiri atas delapan kelas dengan sebaran yang sangat tidak merata. Tabel 2 distribusi awal label menunjukkan pola sebagai berikut:

Tabel 2 Distribusi data

Label	Jumlah
User Satisfaction - Positive	4585
Effectiveness - Positive	2950
Error Rate - Negative	1933
User Satisfaction - Negative	196
Learnability - Positive	185
Effectiveness - Negative	68
Others - Others	53
Learnability - Negative	30

Dalam pembelajaran mesin, terutama pada kasus multi-class text classification, distribusi kelas yang tidak seimbang menjadi persoalan serius. Kelas dengan jumlah sampel yang sangat sedikit cenderung menimbulkan beberapa masalah metodologis, antara lain:

- Model sulit mempelajari pola bahasa, contoh yang sangat sedikit membuat model tidak memiliki variasi yang cukup untuk mengenali karakteristik teks dari kelas minor.
- Risiko overfitting meningkat, model lebih mudah “hafal” kelas mayoritas dan mengabaikan kelas kecil, sehingga performa pada data baru menurun.
- F1-score makro menjadi buruk, kelas dengan dukungan sangat kecil akan selalu menghasilkan F1-score rendah, dan ini menurunkan performa keseluruhan model.
- Decision boundary tidak stabil, pada algoritma berbasis neural networks, kelas minor dapat menyebabkan batas keputusan menjadi keruh dan sulit dipelajari secara konsisten.

Berdasarkan pertimbangan tersebut, tiga kelas dengan jumlah data paling kecil diputuskan untuk dieliminasi, yaitu:

- a. Learnability - Negative (30 data)
- b. Others - Others (53 data)
- c. Effectiveness - Negative (68 data)

Total data yang dihapus: 151 ulasan

Jumlah data setelah eliminasi: 9.849 ulasan

Jumlah kelas yang digunakan: 5 kelas utama

Menghapus kelas minor bukan sekadar tindakan pembersihan data, tetapi merupakan langkah metodologis untuk memastikan model dapat dilatih secara stabil dan menghasilkan performa klasifikasi yang lebih akurat. Dengan mempertahankan lima kelas utama yang memiliki representasi data kuat, penelitian ini memperoleh fondasi yang lebih solid untuk membangun model berbasis RNN yang reliabel dan generalizable.

4.2 Data Preprocessing

Setelah proses penyaringan kelas, tahap berikutnya adalah menyiapkan data teks agar dapat diproses oleh model RNN. Preprocessing dilakukan secara sistematis untuk memastikan kualitas input dan mengurangi noise dalam ulasan pengguna. Tahapan lengkapnya dijelaskan sebagai berikut.

4.2.1 Lowercasing

Seluruh teks ulasan dikonversi ke huruf kecil. Langkah ini dilakukan untuk menyatukan representasi kata, sehingga model tidak membedakan kata yang sama seperti “Bagus” dan “bagus”. Normalisasi sederhana ini membantu mengurangi ukuran kosakata tanpa kehilangan makna.

4.2.2 Penghapusan Tanda Baca

Tanda baca dan karakter non-alfabet dihilangkan menggunakan ekspresi reguler `re.sub(r'[\W_]+', '', x)`. Tujuannya adalah mengurangi gangguan (noise) dari karakter yang tidak memberikan kontribusi signifikan terhadap pemaknaan sentimen atau kriteria usability. Hasilnya berupa teks yang lebih bersih dan seragam.

4.2.3 Tokenisasi

Teks yang telah dibersihkan kemudian dipetakan menjadi urutan angka menggunakan Tokenizer dari Keras. Tahap ini membentuk representasi numerik dari setiap kata berdasarkan frekuensi kemunculannya. Hasil tokenisasi menunjukkan bahwa:

- a. ukuran kosakata mencapai 5873 kata unik
- b. setiap ulasan berubah menjadi daftar integer, misalnya:
- c. “sangat membantu” menjadi [2, 8]

Representasi ini penting karena model RNN hanya dapat bekerja dengan input numerik, bukan teks mentah.

4.2.4 Analisis Panjang Ulasan

Sebelum menentukan panjang input untuk model, dilakukan analisis terhadap distribusi panjang teks. Hasil analisis menunjukkan bahwa ulasan pengguna cenderung sangat pendek:

- a. median panjang ulasan = 3 kata
- b. rata-rata = 6 kata
- c. persentil ke-90 = 17 kata

Temuan ini menggambarkan bahwa mayoritas ulasan berupa frasa singkat seperti “bagus”, “tidak bisa login”, atau “susah dibuka”. Oleh karena itu, panjang maksimum urutan (`max_len`) dipilih sebesar 17, agar cakupan 90% ulasan tetap terjaga tanpa padding berlebihan.

4.2.5 Padding

Untuk memastikan seluruh input memiliki dimensi yang sama, setiap urutan ditambah (padding) di bagian akhir:

```
pad_sequences(..., maxlen=17, padding='post')
```

Hasil akhir menghasilkan matriks input berukuran:

- 9849 baris (jumlah ulasan setelah filtering)
- 17 token per baris

Struktur ini sesuai dengan kebutuhan arsitektur RNN yang memerlukan input berukuran tetap.

4.2.6 Encoding Label

Tahap terakhir preprocessing adalah mengonversi label kategori ke format numerik yang dapat diproses model. Proses dilakukan dalam dua langkah:

a. LabelEncoder

Mengubah label teks menjadi angka, misalnya:

“Effectiveness - Positive” menjadi 0,

“Error Rate - Negative” menjadi 1, dan seterusnya.

b. One-hot encoding

Angka label kemudian diubah menjadi vektor biner untuk digunakan pada lapisan softmax. Contoh:

“User Satisfaction - Positive” menjadi [0 0 0 0 1]

Jumlah kelas akhir yang digunakan adalah 5 kelas utama, sesuai dengan hasil penyaringan sebelumnya.

4.3 Pembagian Dataset

Setelah proses preprocessing selesai, dataset dibagi menjadi tiga subset utama yang digunakan untuk pelatihan, validasi, dan pengujian model. Proses ini dilakukan dengan pendekatan stratified split untuk memastikan bahwa proporsi setiap label tetap konsisten pada seluruh subset, sehingga model tidak bias terhadap kelas tertentu. Pembagian data dilakukan dengan skema sebagai berikut:

- a. Data latih (Training Set): 7.879 sampel (sekitar 80%). Dataset ini digunakan sebagai bahan utama untuk mempelajari pola dan representasi teks.
- b. Data validasi (Validation Set): 985 sampel (sekitar 10%). Subset ini digunakan untuk memantau performa model pada data yang tidak terlihat selama pelatihan, sekaligus mencegah overfitting.
- c. Data uji (Test Set): 985 sampel (sekitar 10%). Digunakan sebagai evaluasi akhir untuk mengukur kemampuan generalisasi model terhadap data baru.

Penggunaan stratified split sangat penting dalam konteks klasifikasi multi-kelas, karena memastikan bahwa distribusi label pada dataset tetap proporsional dan representatif. Dengan demikian, setiap kelas memperoleh peluang yang sama untuk

dilatih dan dievaluasi, terutama kelas minor yang jumlahnya relatif sedikit.

4.4 Pembangunan Arsitektur RNN

Model klasifikasi yang dikembangkan pada penelitian ini menggunakan pendekatan Recurrent Neural Network (RNN) dengan varian Long Short-Term Memory (LSTM). Arsitektur ini dipilih karena memiliki kemampuan unggul dalam menangkap pola urutan kata pada teks ulasan pengguna, termasuk konteks yang muncul secara bertahap. Secara keseluruhan, arsitektur model terdiri dari beberapa lapisan berikut.

4.4.1 Embedding Layer (128 dimensi)

Lapisan ini bertugas mengubah setiap kata menjadi vektor representasi berdimensi 128. Embedding bekerja seperti memetakan kata ke dalam ruang makna, di mana kata yang mirip secara semantik akan berada pada koordinat yang berdekatan. Tahap ini menjadi fondasi penting agar model dapat memahami hubungan antar-kata secara lebih halus.

4.4.2 SpatialDropout1D (rate = 0.2)

Lapisan ini diterapkan untuk mengurangi risiko overfitting. Dengan mematikan sebagian fitur embedding secara acak, model belajar menjadi lebih robust dan tidak terlalu bergantung pada pola tertentu yang mungkin hanya muncul pada data latih.

4.4.3 LSTM Layer (100 unit)

Lapisan inti dari arsitektur ini adalah LSTM dengan 100 unit. LSTM membantu model “mengingat” informasi penting dari urutan kata sebelumnya sambil memproses kata berikutnya. Secara intuitif, LSTM bekerja seperti equalizer yang mengatur intensitas sinyal makna pada setiap langkah, mereduksi noise, dan mempertahankan informasi yang relevan sepanjang kalimat.

Lapisan ini sangat cocok untuk teks pendek seperti ulasan Google Play Store, karena mampu menangkap konteks yang sering muncul secara singkat namun padat makna.

4.4.4 Dense Layer dengan Softmax (5 kelas)

Lapisan terakhir berupa Dense Layer dengan aktivasi softmax, yang menghasilkan probabilitas untuk lima kelas kombinasi usability dan sentimen. Setiap output mewakili peluang ulasan masuk ke dalam salah satu kategori seperti Effectiveness, Error Rate, User Satisfaction, dan seterusnya.

4.4.5 Pengaturan Pelatihan Model

Model dikompilasi menggunakan:

- Loss function: `categorical_crossentropy`, sesuai untuk multi-class classification
- Optimizer: Adam, karena adaptif dan bekerja baik pada data teks
- Metrics utama: Accuracy, sebagai metrik umum untuk menilai performa model pada seluruh kelas

Arsitektur ini dirancang untuk memberikan keseimbangan antara kompleksitas model dan kapasitas generalisasinya, sehingga mampu mempelajari pola teks pendek khas ulasan pengguna secara efektif.

4.5 Pelatihan Model

Model RNN berbasis LSTM dilatih selama 10 epoch dengan ukuran batch 64. Selama proses pelatihan, performa model diamati melalui perubahan nilai accuracy dan loss pada data latih serta validasi. Hasil lengkap dapat dilihat pada log pelatihan.

4.5.1 Hasil Pelatihan

Berdasarkan Gambar 3 output pelatihan, alur pembelajaran model menunjukkan pola yang stabil dan konsisten.

```

... Starting model training for 10 epochs with a batch size of 64...
Epoch 1/10
124/124 ----- 10s 59ms/step - accuracy: 0.5925 - loss: 1.0291 - val_accuracy: 0.8284 - val_loss: 0.4758
Epoch 2/10
124/124 ----- 5s 37ms/step - accuracy: 0.9056 - loss: 0.3173 - val_accuracy: 0.9127 - val_loss: 0.2953
Epoch 3/10
124/124 ----- 6s 45ms/step - accuracy: 0.9330 - loss: 0.2217 - val_accuracy: 0.9076 - val_loss: 0.2819
Epoch 4/10
124/124 ----- 6s 44ms/step - accuracy: 0.9482 - loss: 0.1694 - val_accuracy: 0.9107 - val_loss: 0.2607
Epoch 5/10
124/124 ----- 5s 38ms/step - accuracy: 0.9533 - loss: 0.1502 - val_accuracy: 0.9137 - val_loss: 0.2853
Epoch 6/10
124/124 ----- 7s 54ms/step - accuracy: 0.9611 - loss: 0.1196 - val_accuracy: 0.9117 - val_loss: 0.2638
Epoch 7/10
124/124 ----- 5s 37ms/step - accuracy: 0.9710 - loss: 0.1052 - val_accuracy: 0.9198 - val_loss: 0.2769
Epoch 8/10
124/124 ----- 5s 37ms/step - accuracy: 0.9766 - loss: 0.0907 - val_accuracy: 0.9168 - val_loss: 0.3078
Epoch 9/10
124/124 ----- 7s 55ms/step - accuracy: 0.9818 - loss: 0.0682 - val_accuracy: 0.9198 - val_loss: 0.2989
Epoch 10/10
124/124 ----- 5s 37ms/step - accuracy: 0.9836 - loss: 0.0659 - val_accuracy: 0.9259 - val_loss: 0.2877
Model training complete.

```

Gambar 3 Model Training

- Epoch 1 (Pada tahap ini model baru mulai mengenali pola dasar teks.)

Train accuracy: 59.25%

Validation accuracy: 82.84%

Validation loss masih tinggi (0.4758)

- b. Epoch 2 - 4 (Model mulai belajar representasi kata dengan baik dan sudah mampu menggeneralisasi pola umum pada data validasi.)

Train accuracy meningkat drastis dari 90.56% menjadi 94.82%

Validation accuracy naik di kisaran 90.7% - 91.2%

Validation loss turun signifikan hingga 0.2607

- c. Epoch 5 - 6 (Ini merupakan titik di mana performa validasi mencapai nilai terbaiknya.)

Train accuracy: 95.33% sampai 96.11%

Validation accuracy: 91.17% sampai 91.37%

Validation loss mulai fluktuatif (0.2853 sampai 0.2638)

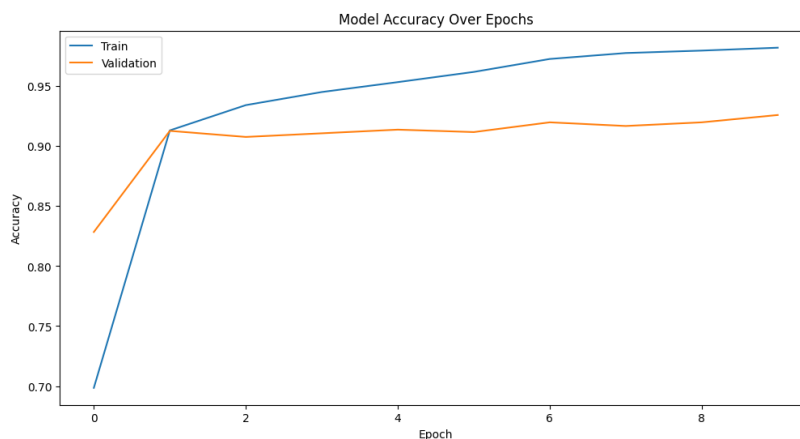
- d. Epoch 7 - 10

Train accuracy sangat tinggi (97.10% sampai 98.36%)

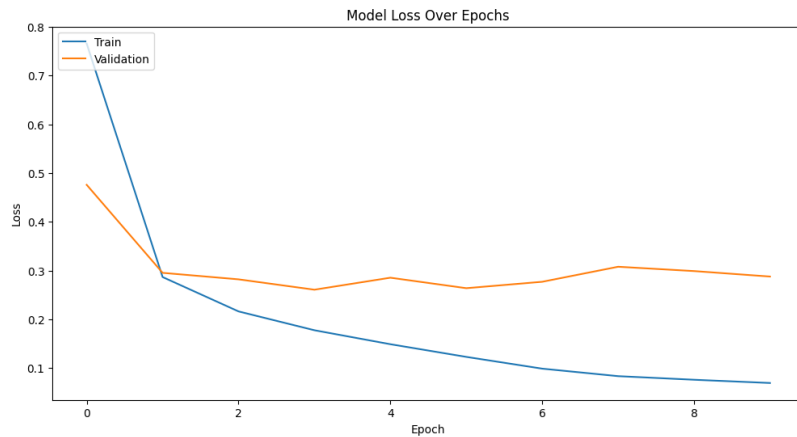
Validation accuracy berada di 91.68% sampai 92.59%

Validation loss sedikit meningkat (0.2769 sampai 0.3078)

Mulai epoch ke-6 ke atas, terlihat bahwa training accuracy terus naik mendekati 98%, tetapi validation accuracy hanya naik sedikit dan validation loss mulai fluktuatif. Kondisi ini menandakan bahwa model mulai menghafal pola pada kelas mayoritas. Overfitting yang terjadi tidak ekstrem, tetapi terlihat dari kesenjangan yang semakin besar antara performa pada data latih dan data validasi. Ini sangat wajar terjadi pada dataset dengan panjang teks pendek dan distribusi kelas tidak seimbang.



Gambar 4 Model Accuracy



Gambar 5 Model Loss

Dari Gambar 4 dan 5, pola kurva Validation accuracy terbaik: 92.59% (epoch ke-10), Validation loss terbaik: 0.2607 (epoch ke-4). Jika menggunakan early stopping, titik optimal pelatihan kemungkinan berada pada epoch 4 - 6, yaitu saat model mencapai keseimbangan terbaik antara akurasi dan generalisasi.

Dari pelatihan tersebut model memahami pola bahasa ulasan dengan cepat (terlihat dari lonjakan akurasi pada epoch awal). Setelah epoch ke-5, model mengalami penurunan kemampuan generalisasi. Namun performa validasi tetap tinggi di atas 91%, menunjukkan model cukup stabil dan efektif. Overfitting yang terjadi masih dalam batas aman dan tidak berdampak drastis pada hasil evaluasi.

4.6 Evaluasi Model

Setelah proses pelatihan selesai, model diuji menggunakan 985 data uji yang mewakili 10% dari keseluruhan dataset. Data uji ini sepenuhnya tidak terlihat selama pelatihan sehingga memberikan gambaran paling objektif tentang kemampuan generalisasi model.

Akurasi keseluruhan model mencapai akurasi sebesar 91.68% pada data uji. Nilai ini tergolong sangat baik, terutama mengingat karakteristik ulasan Google Play Store yang penuh dengan variasi bahasa, termasuk:

- banyaknya singkatan seperti “gk”, “bgt”, atau “ga bisa login”
- kalimat yang sangat pendek dan tidak baku
- penggunaan campuran bahasa Indonesia dan bahasa daerah
- emosi yang diekspresikan secara spontan dan tidak terstruktur

Dengan segala bentuk “noise” tersebut, keberhasilan model mempertahankan akurasi di atas 90% menunjukkan bahwa arsitektur RNN-LSTM yang digunakan mampu mengekstraksi pola makna secara efektif, bahkan dari teks yang singkat dan

tidak formal. Akurasi ini menjadi indikasi awal bahwa pendekatan berbasis RNN memiliki potensi kuat untuk diterapkan dalam evaluasi usability aplikasi mobile berbasis ulasan pengguna secara otomatis.

Evaluasi performa model berdasarkan classification report pada Tabel 3 memberikan gambaran lebih rinci mengenai kemampuan model dalam mengenali masing-masing kelas. Tabel 3 merangkum nilai precision, recall, F1-score, serta jumlah sampel (support) untuk setiap kategori.

Tabel 3 Classification Report

Label	Precision	Recall	F1-score	Suport
Effectiveness - Positive	0.92	0.93	0.92	295
Error Rate - Negative	0.96	0.92	0.94	194
Learnability - Positive	0.52	0.78	0.62	18
User Satisfaction - Negative	0.36	0.40	0.38	20
User Satisfaction - Positive	0.95	0.94	0.94	458

Model menunjukkan performa yang sangat baik pada kelas-kelas dengan jumlah data cukup besar. Tiga kelas berikut memiliki nilai F1-score di atas 0.92:

- a. User Satisfaction - Positive
- b. Effectiveness - Positive
- c. Error Rate - Negative

Kinerja tinggi pada kategori tersebut menunjukkan bahwa model mampu mengenali pola bahasa yang jelas, baik dalam bentuk pujian maupun keluhan eksplisit. Komentar seperti “bagus sekali”, “fiturnya membantu”, atau “sering error saat login” memiliki struktur yang mudah dipelajari sehingga menghasilkan prediksi yang akurat.

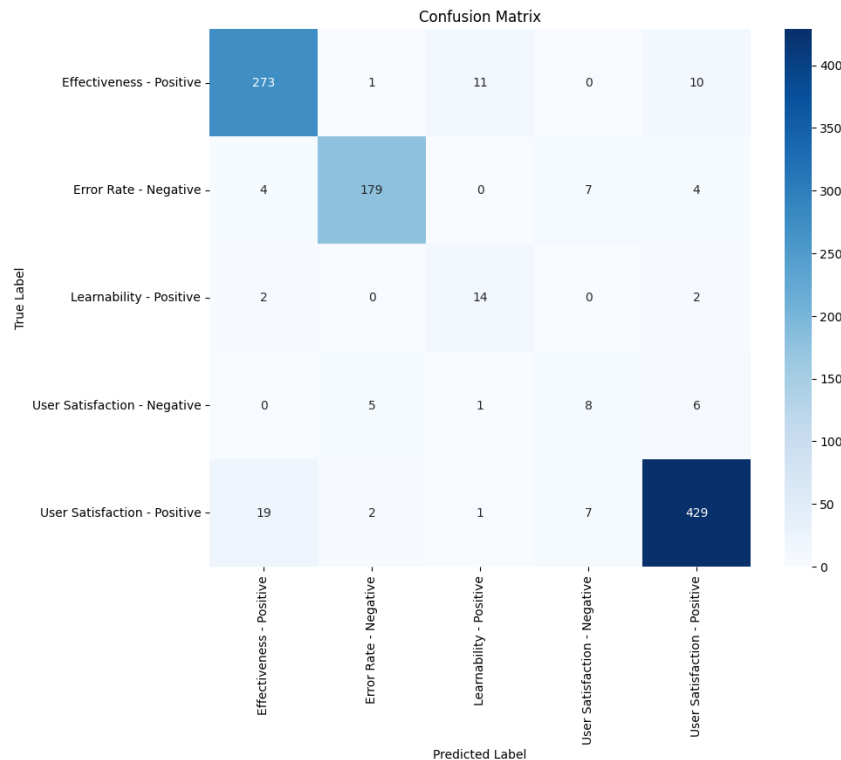
Sebaliknya, model cenderung kesulitan dalam mengklasifikasikan dua kelas minoritas:

- a. Learnability - Positive
- b. User Satisfaction - Negative

Kedua kategori ini memiliki nilai F1-score yang lebih rendah, yaitu 0.62 dan 0.38. Terdapat beberapa faktor penyebab utama:

- a. Jumlah sampel yang sangat sedikit, dengan hanya 18 dan 20 data pada masing-masing kelas, model tidak memiliki cukup variasi contoh untuk membangun representasi pola bahasa yang kuat.
- b. Tumpang tindih makna dengan kelas lain, ulasan untuk aspek Learnability atau User Satisfaction sering kali ditulis dengan gaya bahasa umum, tanpa kata kunci yang khas. Contoh “aplikasinya cukup oke” dapat mencerminkan satisfaction, learnability, atau bahkan sekadar penilaian umum.
- c. Komentar negatif yang bersifat halus atau implisit, kelas “User Satisfaction - Negative” sangat dipengaruhi oleh ekspresi ketidakpuasan yang tidak selalu ditulis secara eksplisit. Misalnya “lumayan... tapi sering macet”. Kalimat seperti ini dapat masuk ke Error Rate - Negative atau User Satisfaction - Negative, tergantung konteks yang tersirat.

Kesulitan membedakan komentar negatif yang halus inilah yang menyebabkan model lebih sering salah mengklasifikasikan kelas minoritas, seperti yang terlihat pada Gambar 6 confusion matrix.



Gambar 6 Confusion Matrix

Confusion matrix memberikan gambaran paling jelas mengenai pola kesalahan yang terjadi selama proses prediksi. Dari analisis matriks tersebut, terlihat bahwa sebagian besar kekeliruan model berasal dari dua kelas dengan jumlah data paling sedikit, yaitu User Satisfaction - Negative dan Learnability - Positive.

Kategori User Satisfaction - Negative menunjukkan performa terlemah di antara seluruh kelas. Hanya 8 dari 20 data yang berhasil diklasifikasikan dengan benar, sisanya salah dipetakan sebagai User Satisfaction - Positive sebanyak 6 kali. Error Rate - Negative sebanyak 5 kali, fenomena ini mencerminkan dua tantangan utama mengenai

- a. Ambiguitas Bahasa pada Komentar Negatif Halus
- b. Banyak ulasan negatif tidak dituliskan secara langsung, misalnya:

“lumayan, tapi kadang suka macet”

Kalimat seperti ini dapat dibaca sebagai keluhan teknis (error) atau ekspresi ketidakpuasan secara umum (user satisfaction), sehingga model kesulitan menempatkan konteks secara tepat.

Bias terhadap Kelas Mayoritas, karena data “User Satisfaction - Positive” jumlahnya sangat besar, model cenderung “menarik” prediksi ke arah kelas tersebut ketika konteksnya tidak jelas.

Kategori lainnya Learnability - Positive, model masih dapat memberikan hasil yang cukup baik 14 dari 18 ulasan terdeteksi dengan benar, namun 4 data lainnya salah dipetakan ke kelas mayoritas, terutama kategori positif yang lebih umum. Dua hal ini menjadi penyebab utamanya:

- a. Minimnya Indikator Bahasa yang Khas untuk Learnability
- b. Ulasan mengenai kemudahan belajar aplikasi sering ditulis secara umum, seperti:

“cukup mudah digunakan”

Ungkapan seperti ini mudah tertukar dengan kategori “User Satisfaction - Positive”.

- c. Keterbatasan Data Latih, dengan hanya 18 contoh, variasi pola bahasa yang dapat dipelajari model sangat terbatas, sehingga model rentan menggeneralisasi secara berlebihan.

Dari hasil analisis classification report dan confusion matrix, dapat disimpulkan bahwa terdapat dua penyebab utama yang mendorong munculnya kesalahan prediksi pada model:

- a. Ambiguitas makna dalam ulasan yang pendek dan tidak eksplisit. Ulasan pengguna di Google Play Store sering kali dituliskan secara singkat, emosional, dan tanpa konteks yang jelas. Banyak komentar negatif yang disampaikan secara halus atau tidak langsung, sehingga model kesulitan membedakan apakah pengguna sedang mengeluhkan aspek usability, performa teknis, atau sekadar mengekspresikan kekecewaan umum.
- b. Ketidakseimbangan jumlah data antar kelas. Kelas mayoritas mendominasi proses pembelajaran, membuat model memiliki bias alami untuk memprediksi kategori-kategori yang lebih sering muncul. Sementara itu, kelas minoritas seperti Learnability - Positive dan User Satisfaction - Negative tidak memiliki cukup variasi contoh untuk membentuk representasi yang kuat di dalam ruang fitur.

Dua temuan ini menjadi fondasi penting untuk pengembangan model pada penelitian berikutnya. Perbaikan dapat diarahkan melalui:

- a. penambahan atau augmentasi data pada kelas minoritas,
- b. penerapan strategi pembobotan kelas selama pelatihan,
- c. serta eksplorasi arsitektur yang lebih canggih, seperti model dengan

mekanisme perhatian (attention mechanism) atau pendekatan berbasis transformer yang mampu menangkap konteks kalimat secara lebih mendalam.

BAB V KESIMPULAN

Hasil evaluasi menunjukkan beberapa poin penting mengenai kinerja model RNN yang digunakan. Model RNN menunjukkan performa yang sangat baik pada kelas-kelas mayoritas. Kategori seperti User Satisfaction - Positive, Effectiveness - Positive, dan Error Rate - Negative memiliki jumlah data besar sehingga pola bahasanya dapat dipelajari dengan stabil. Hal ini tercermin dari nilai F1-score yang konsisten di atas 0.92. Model masih mengalami kesulitan pada kelas minoritas dengan jumlah data kurang dari 30 contoh. Kategori seperti User Satisfaction - Negative dan Learnability - Positive tidak memiliki cukup variasi bahasa, sehingga representasi yang terbentuk di dalam ruang embedding cenderung lemah dan mudah tertukar dengan kelas lain. Indikasi overfitting ringan terlihat setelah epoch ke-6. Akurasi pelatihan terus meningkat, sedangkan akurasi validasi cenderung stagnan. Ini menunjukkan bahwa model mulai ‘menghafal’ pola kelas mayoritas alih-alih memahami generalisasi data secara menyeluruh. Karakter ulasan yang sangat pendek membuat konteks sulit ditangkap oleh RNN. Banyak ulasan hanya terdiri dari satu hingga tiga kata. Pola ini tidak memberikan cukup sinyal semantik bagi model untuk membedakan sentimen atau aspek usability yang dimaksudkan pengguna.

Selain itu, evaluasi juga menemukan bahwa satu kalimat ulasan sering kali bersifat ambigu dan berpotensi masuk ke dua label berbeda secara bersamaan. Contoh kasus ambiguitas “Lumayan, tapi sering macet.” Kalimat ini secara bersamaan menunjukkan apresiasi ringan (positive) sekaligus keluhan teknis (error negative). Model dapat salah menafsirkan apakah ulasan ini lebih tepat dikategorikan sebagai User Satisfaction - Negative atau justru Error Rate - Negative. “Bagus sih, cuma kadang lemot.” Bagian pertama adalah positive sentiment, sedangkan bagian kedua merupakan performance error. Ini menyulitkan RNN yang membaca kalimat secara berurutan tanpa mekanisme fokus konteks.

Ambiguitas seperti ini memperkuat alasan mengapa model berbasis konteks mendalam seperti attention mechanism atau transformer-based models (misalnya BERT atau IndoBERT) lebih cocok untuk penelitian lanjutan. Dengan kemampuan memahami hubungan antarfrasa, model tersebut dapat menangkap nuansa positif dan negatif dalam satu kalimat secara lebih akurat.

DAFTAR PUSTAKA

- Ahmad, A., Bashir, M. S., Raza, M. S., & Babool, A. (2023). A Dataset for Analyzing Crowdsourced Feedback in Usability Testing. *2023 International Conference on Frontiers of Information Technology (FIT)*, 250-255. <https://doi.org/10.1109/FIT60620.2023.00053>
- Arras, L., Osman, A., Müller, K.-R., & Samek, W. (2019). Evaluating Recurrent Neural Network Explanations. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 113-126. <https://doi.org/10.18653/v1/W19-4813>
- Capilla, R., Kazman, R., Romera, C., & Carrillo, C. (2020). Usability implications in software architecture: The case study of a mobile app. *Software: Practice and Experience*, 50(12), 2145-2168. <https://doi.org/10.1002/spe.2883>
- Chahar, R., & Dubey, A. K. (2024). Machine learning-based Sentiment Analysis: A Comprehensive Review. *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, 52-57. <https://doi.org/10.1109/ICC-ROBINS60238.2024.10533989>
- Clement, J. (2024). *Number of available gaming apps in the Google Play Store from 1st quarter 2015 to 2nd quarter 2024*. Statista. <https://www.statista.com/statistics/780229/number-of-available-gaming-apps-in-the-google-play-store-quarter/>
- Fernández, J., & Macías, J. A. (2021). Heuristic-Based Usability Evaluation Support. *Proceedings of the XXI International Conference on Human Computer Interaction*, 1-9. <https://doi.org/10.1145/3471391.3471395>
- Hartvigsen, T., Sen, C., Kong, X., & Rundensteiner, E. (2020). Learning to Selectively Update State Neurons in Recurrent Networks. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 485-494. <https://doi.org/10.1145/3340531.3412018>
- Hassan, J., & Shoaib, U. (2020). Multi-class Review Rating Classification using Deep Recurrent Neural Network. *Neural Processing Letters*, 51(1), 1031-1048. <https://doi.org/10.1007/s11063-019-10125-6>
- Hertzum, M. (2020). Usability Testing: A Practitioner's Guide to Evaluating the User Experience. In *Synthesis Lectures on Human-Centered Informatics* (Vol. 1, Issue 1). <https://doi.org/10.2200/S00987ED1V01Y202001HCI045>
- Jakob Nielsen. (2012). *Usability 101: Introduction to Usability*. <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>
- Mohamad Sham, N., & Mohamed, A. (2022). Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches. *Sustainability*, 14(8), 4723. <https://doi.org/10.3390/su14084723>
- Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1), 81. <https://doi.org/10.1007/s13278-021-00776-6>
- Noh, S.-H. (2021). Analysis of Gradient Vanishing of RNNs and Performance Comparison. *Information*, 12(11), 442. <https://doi.org/10.3390/info12110442>
- Privitera, M. B. (2019). Heuristic analysis, cognitive walkthroughs & expert reviews. In *Applied Human Factors in Medical Device Design* (pp. 165-180). Elsevier. <https://doi.org/10.1016/B978-0-12-816163-0.00010-4>
- Weichbroth, P. (2020). Usability of Mobile Applications: A Systematic Literature Study. *IEEE Access*, 8, 55563-55577. <https://doi.org/10.1109/ACCESS.2020.2981892>
- Zhang, P., Xue, J., Lan, C., Zeng, W., Gao, Z., & Zheng, N. (2020). EleAtt-RNN: Adding Attentiveness to Neurons in Recurrent Neural Networks. *IEEE Transactions on Image Processing*, 29, 1061-1073. <https://doi.org/10.1109/TIP.2019.2937724>

Lampiran

1. FGD

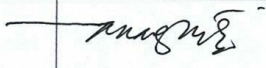
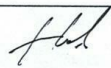
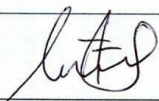
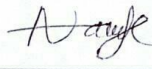
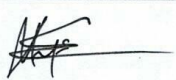


DAFTAR KEHADIRAN

Kegiatan :

Tempat :

Tanggal :

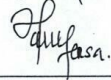
No	Nama	Keterangan	Ttd
1	Xuang Mantayo		
2	Hasan Basri		
3			
4			
5	Archangela Chiriani ug		
6	Zalfa Rizky Pinaomi p.		
7	Nasywa Mahdiyyah		
8	Fathma Aulia Puteri		
9	Nova Lina Anggraeny		
10	Navisa Lailatin Arifan		
11	Davina Fairuz		

DAFTAR KEHADIRAN

Kegiatan :

Tempat :

Tanggal :

No	Nama	Keterangan	Ttd
12	Maria Goreti Godekvacang		
13	Shafira Putri Salsabila		
14	Si Syeren Permang		
15	M. Kayyis, Abdulloh		
16	Muhamad Raafi		
17	Alfian		
18	Friansa Jagoto		
19			
20			
21			
22			



**KEMENTERIAN PENDIDIKAN TINGGI, SAINS,
DAN TEKNOLOGI**
UNIVERSITAS TERBUKA
LEMBAGA PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT
Jalan Cabe Raya, Pondok Cabe, Pamulang, Tangerang Selatan 15437
Telepon: (021) 74909411 ext. 1201, Faksimile: (021) 7434790,
E-mail: lgpm@ecampus.ut.ac.id, Laman: www.lgpm.ut.ac.id

SURAT TUGAS

NOMOR : B/1566/UN31.LPPM/PT.01.04/2025

Ketua Lembaga Penelitian dan Pengabdian Kepada Masyarakat Universitas Terbuka, dengan ini menugaskan kepada:

No	Nama	NIP	Berangkat dari
1.	Hasan Basri, S.Kom., M.Kom.	199205172024061003	Kota tangerang Selatan
2.	Ihsan Alparisi	051598332	Sukabumi
3.	Renaldi Saepul Rohmat	051462175	Sukabumi

Mulai : 27/05/2025 Sampai: 27/05/2025
Selama: : 1
Tujuan : Jakarta
Lokasi : Cyber University Jakarta

FGD dan Konsultasi terkait penelitian yang berjudul AI-Driven Usability Testing: Optimizing Mobile App Performance With Combination Usability Criteria And Sentiment Analysis Using RNN Models nomor: B/349/UN31.LPPM/PT.01.03/2025 yang di ketuai oleh Hasan Basri, S.Kom., M.Kom..

Surat Tugas ini dibuat untuk dilaksanakan dengan rasa penuh tanggung jawab.

Tangerang Selatan, 21/05/2025

Tembusan:

1. Dekan FST
2. LPPM



Dokumen telah ditandatangani secara elektronik yang diterbitkan oleh Universitas Terbuka sesuai UU ITE No. 11 tahun 2008 dan Peraturan Rektor Universitas Terbuka No. 1243 tahun 2021. Informasi dokumen dapat dilihat dengan aplikasi "Acrobat Reader" dan pembuktian keasliannya dengan aplikasi ProDSign UT.