TESIS

EVALUASI BERT DENGAN INVERSE-FREQUENCY WEIGHTED FOCAL LOSS UNTUK KLASIFIKASI KEBUTUHANPERANGKAT LUNAK PADA AGILE USER STORY



Oleh:
Anisatul Mukaromah
24206051016

PROGRAM STUDI INFORMATIKA
PROGRAM MAGISTER FAKULTAS SAINS DAN TEKNOLOGI
UIN SUNAN KALIJAGA

YOGYAKARTA
2026

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

HALAMAN PENGESAHAN

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

PENGESAHAN TUGAS AKHIR

Nomor : B-2412/Un.02/DST/PP.00.9/10/2026

Tugas Akhir dengan judul : EVALUASI BERT DENGAN INVERSE-FREQUENCY WEIGHTED FOCAL LOSS UNTUK KLASIFIKASI KEBUTUHANPERANGKAT LUNAK PADA AGILE USER STORY

yang dipersiapkan dan disusun oleh:

Nama : ANISATUL MUKAROMAH, S.Kom
Nomor Induk Mahasiswa : 24206051016
Telah diujikan pada : Jumat, 28 Agustus 2026
Nilai ujian Tugas Akhir : A/B

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR

Ketua Sidang

Dr. Ir. Bambang Sugiantoro, M.T., IPU., ASEAN Eng.
SIGNED

Valid ID: 6abfa05adaddf

Penguji I

Dr. Agung Fatwanto, S.Si., M.Kom.
SIGNED

Valid ID: 6abf6ddae6d27

Penguji II

Dr. Ir. Sumarsono, S.T., M.Kom.
SIGNED

Valid ID: 6abcb01c42e7

Yogyakarta, 28 Agustus 2026
UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 6ac32061c8c24

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

SURAT PERNYATAAN KEASLIAN

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama	: Anisatul Mukaromah
NIM	: 24206051016
Jenjang	: Magister
Program Studi	: Informatika

menyatakan bahwa naskah tesis ini secara keseluruhan adalah hasil penelitian/karya saya sendiri, kecuali pada bagian-bagian yang dirujuk sumbernya.

Yogyakarta, 26 Agustus 2026
Saya yang menyatakan,



Anisatul Mukaromah
NIM: 24206051016

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

NOTA DINAS PEMBIMBING

NOTA DINAS PEMBIMBING

Kepada Yth.,
Dekan Fakultas Sains dan Teknologi
UIN Sunan Kalijaga
Yogyakarta

Assalamu'alaikum wr. wb.

Disampaikan dengan hormat, setelah melakukan bimbingan, arahan dan koreksi terhadap penulisan tesis yang berjudul:

EVALUASI BERT DENGAN INVERSE-FREQUENCY WEIGHTED FOCAL LOSS UNTUK KLASIFIKASI KEBUTUHANPERANGKAT LUNAK PADA AGILE USER STORY

Yang ditulis oleh:

Nama	: Anisatul Mukaromah
NIM	: 24206051016
Jenjang	: Magister
Program Studi	: Informatika

Saya berpendapat bahwa tesis tersebut sudah bisa diajukan kepada Program Studi Magister Informatika UIN Sunan Kalijaga Yogyakarta untuk diujikan dalam rangka memperoleh gelar Magister Informatika.

Wassalamu'alaikum wr. wb.

Yogyakarta, 26 Agustus 2026
Pembimbing,

Dr. Ir. Bambang Sugiantoro, M.T., IPU., ASEAN Eng.
NIP. 19751024 200912 1 002

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

ABSTRAK

Klasifikasi kebutuhan perangkat lunak pada *agile user story* menghadapi distribusi kelas tidak seimbang karena jumlah *functional requirements* (FR) lebih besar daripada setiap kategori *non-functional requirements* (NFR). Penelitian ini mengevaluasi BERT-base-uncased dengan Inverse-Frequency Weighted Focal Loss untuk klasifikasi satu kelas FR dan tujuh kategori NFR. Dataset terdiri atas 10.006 *user story* berbahasa Inggris yang dibagi secara berstrata menjadi 8.004 data *training*, 1.001 data *validation*, dan 1.001 data *test* menggunakan seed 42. Bobot kelas dihitung dari frekuensi pada *training set* dan dinormalisasi sehingga jumlah bobot delapan kelas sama dengan delapan. Focal Loss menggunakan gamma 2. Hyperparameter optimization berbasis *grid search* menguji learning rate 2e-5, 3e-5, dan 5e-5 serta batch size 16 dan 32. Konfigurasi dan checkpoint dipilih berdasarkan Macro F1 pada *validation set*, kemudian konfigurasi terpilih dilatih ulang dari BERT dasar dan dievaluasi pada *test set* independen. Learning rate 2e-5 dan batch size 32 memperoleh validation Macro F1 tertinggi sebesar 0,8913. Evaluasi *test* menghasilkan Accuracy 0,9351, Weighted F1 0,9358, Macro F1 0,9027, dan Matthews Correlation Coefficient 0,8875. F1 per kelas berada pada rentang 0,8323–0,9639; FR memperoleh F1 tertinggi dan Usability terendah. Dari 65 kesalahan, 32 data FR diprediksi sebagai NFR, 14 data NFR diprediksi sebagai FR, dan 19 data NFR diprediksi sebagai kategori NFR lain. Kesalahan dua arah terbesar terjadi pada pasangan FR–Usability dan FR–Security. Hasil menunjukkan performa empiris pipeline pada dataset dan konfigurasi yang digunakan.

Kata kunci: BERT; Focal Loss; inverse-frequency weighting; klasifikasi kebutuhan perangkat lunak; non-functional requirements.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

MOTTO

اَعْلَمُوْٓا اَنَّمَا الْحَيٰوةُ الدُّنْيَا لَعِبٌ وَّلَهْوٌ وَّزِيْنَةٌ وَّتَفَاخُرٌۢ بَيْنَكُمْ وَتَكَاثُرٌ فِى
الْاَمْوَالِ وَالْاَوْلَادِ ۗ كَمَثَلِ غَيْثٍ اَعْجَبَ الْكُفَّارَ نَبَاتُهٗ ثُمَّ يَهِيْجُ فَتَرٰىهُ
مُصْفَرًّا ثُمَّ يَكُوْنُ حُطَامًا ۗ وَفِى الْاٰخِرَةِ عَذَابٌ شَدِيْدٌ ۙ وَّمَغْفِرَةٌ مِّنَ اللّٰهِ
وَرِضْوَانٌ ۗ وَمَا الْحَيٰوةُ الدُّنْيَآ اِلَّا مَتَاعُ الْغُرُوْرِ ﴿٢٠﴾

"Ketahuilah, sesungguhnya kehidupan dunia itu hanyalah permainan dan senda gurauan, perhiasan dan saling berbangga di antara kamu serta berlomba dalam kekayaan dan anak keturunan, seperti hujan yang tanam-tanamannya mengagumkan para petani; kemudian (tanaman) itu menjadi kering, dan kamu lihat warnanya kuning kemudian menjadi hancur. Dan di akhirat (nanti) ada azab yang keras dan ampunan dari Allah serta keridaan-Nya. Dan kehidupan dunia tidak lain hanyalah kesenangan yang palsu."

Al-Hadid (57:20)

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur, tesis ini penulis persembahkan kepada:

Allah SWT, atas segala rahmat, pertolongan, kemudahan, dan kekuatan yang diberikan dalam setiap proses yang penulis lalui hingga dapat menyelesaikan pendidikan dan tesis ini.

Diri sendiri, yang telah memilih untuk terus berusaha, bertahan, dan menyelesaikan setiap proses yang dijalani. Terima kasih karena tidak menyerah ketika menghadapi berbagai kesulitan dan tetap melangkah hingga sampai pada titik ini.

Kedua orang tua dan keluarga, yang senantiasa memberikan doa, kasih sayang, dukungan, serta kepercayaan kepada penulis. Terima kasih atas segala pengorbanan dan dukungan yang selalu menyertai setiap langkah penulis dalam menempuh pendidikan.

Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Kalijaga Yogyakarta, sebagai tempat penulis menimba ilmu, memperoleh pengalaman, serta mengembangkan pengetahuan selama menempuh pendidikan magister.

Semoga karya sederhana ini dapat menjadi bagian dari ilmu yang bermanfaat serta menjadi awal untuk terus belajar, berkembang, dan memberikan manfaat bagi orang lain.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

KATA PENGANTAR

Assalamu 'alaikum Warahmatullahi Wabarakatuh

Alhamdulillahi Rabbil ‘Alamin. Segala puji dan syukur penulis panjatkan ke hadirat Allah SWT atas limpahan rahmat, karunia, dan pertolongan-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul “Evaluasi BERT dengan Inverse-Frequency Weighted Focal Loss untuk Klasifikasi Kebutuhan Perangkat Lunak pada Agile User Story”. Shalawat dan salam semoga senantiasa tercurahkan kepada Nabi Muhammad SAW, beserta keluarga, para sahabat, dan seluruh umatnya hingga akhir zaman. Semoga kita semua termasuk dalam golongan umat yang memperoleh syafaat beliau di hari akhir.

Tesis ini disusun sebagai salah satu persyaratan untuk memperoleh gelar Magister pada Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Kalijaga Yogyakarta. Di samping sebagai bagian dari proses akademik, penelitian ini diharapkan dapat menambah pengetahuan dan wawasan serta menjadi salah satu referensi yang bermanfaat bagi para pembaca.

Proses penyusunan hingga penyelesaian tesis ini tidak terlepas dari bantuan, bimbingan, dukungan, dan doa dari berbagai pihak. Oleh karena itu, dengan rasa hormat dan penuh rasa terima kasih, penulis menyampaikan penghargaan kepada:

1. Allah SWT atas segala rahmat, pertolongan, dan kemudahan yang senantiasa diberikan kepada penulis selama menjalani pendidikan hingga dapat menyelesaikan tesis ini.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

2. Kedua orang tua dan seluruh keluarga yang senantiasa memberikan doa, dukungan, motivasi, serta bantuan, baik secara moral maupun material, selama penulis menempuh pendidikan.
3. Prof. Noorhaidi Hasan, S.Ag., M.A., M.Phil., Ph.D., selaku Rektor Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
4. Prof. Dr. Dra. Hj. Khurul Wardati, M.Si., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
5. Dr. Ir. Sumarsono, S.T., M.Kom., selaku Ketua Program Studi Magister Informatika, serta Ir. Muhammad Didik Rohmad Wahyudi, S.T., M.T., selaku Sekretaris Program Studi Magister Informatika Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
6. Dr. Ir. Bambang Sugiantoro, M.T., IPU., ASEAN Eng., selaku dosen pembimbing yang telah berkenan meluangkan waktu untuk memberikan bimbingan, arahan, masukan, dan koreksi selama proses penelitian hingga penyusunan tesis ini.
7. Seluruh dosen dan staf Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Kalijaga Yogyakarta, yang telah memberikan ilmu, bantuan, dan pengalaman selama penulis menjalani pendidikan.
8. Teman-teman seperjuangan Program Studi Magister Informatika yang telah menjadi bagian dari perjalanan akademik serta senantiasa memberikan dukungan selama masa perkuliahan hingga proses penyusunan tesis.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

9. Seluruh pihak yang turut memberikan bantuan dan dukungan selama penulis menempuh pendidikan magister hingga menyelesaikan tesis ini, yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari bahwa tesis ini masih memiliki keterbatasan dan jauh dari kesempurnaan. Oleh karena itu, penulis dengan terbuka menerima kritik dan saran yang membangun sebagai bahan evaluasi dan perbaikan pada masa mendatang. Penulis berharap penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan serta dapat menjadi bahan rujukan bagi penelitian selanjutnya yang berkaitan dengan bidang yang dibahas dalam tesis ini.

Akhir kata, penulis menyampaikan terima kasih kepada seluruh pihak yang telah turut menjadi bagian dalam proses penyelesaian tesis ini. Semoga setiap kebaikan, bantuan, dan dukungan yang telah diberikan memperoleh balasan yang lebih baik dari Allah SWT.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

DAFTAR ISI

HALAMAN SAMPUL ..... **i**
HALAMAN PENGESAHAN ..... **ii**
PERNYATAAN KEASLIAN ..... **iii**
NOTA DINAS PEMBIMBING ..... **iv**
ABSTRAK ..... **v**
MOTTO ..... **vi**
HALAMAN PERSEMBAHAN ..... **vii**
KATA PENGANTAR ..... **viii**
DAFTAR ISI ..... **xi**
DAFTAR TABEL ..... **xiv**
DAFTAR GAMBAR ..... **xv**
DAFTAR SINGKATAN ..... **xvi**
BAB I PENDAHULUAN **1**
A. Latar Belakang 1
B. Rumusan Masalah 7
C. Batasan Masalah 7
D. Tujuan Penelitian 8
E. Manfaat Penelitian 9
BAB II KAJIAN PUSTAKA DAN LANDASAN TEORI **10**
A. Kajian Pustaka 10
B. Landasan Teori 18
1. Functional dan Non-Functional Requirements 18
2. Agile User Story 19
3. Transformer dan BERT 20
4. Class Imbalance 22
5. Inverse-Frequency Class Weighting 22
6. Focal Loss 24

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

7. Hyperparameter Optimization dan Fine-Tuning 25
8. Evaluasi Klasifikasi Multikelas 26
BAB III METODE PENELITIAN 29
A. Data Penelitian 29
1. Sumber dan Karakteristik Data 29
2. Distribusi dan Pemetaan Label 31
3. Pembagian Data 32
B. Alat dan Lingkungan 34
1. Perangkat dan Lingkungan 34
2. Model dan Bahan Penelitian 35
C. Tahapan Penelitian 36
1. Alur Penelitian 36
2. Inverse-Frequency Weighted Focal Loss 37
3. Hyperparameter Optimization 39
4. Konfigurasi Training dan Pemilihan Checkpoint 40
5. Pengendalian Reproduksibilitas 43
6. Prosedur Evaluasi 44
7. Pengendalian Validitas Prosedur 45
BAB IV HASIL DAN PEMBAHASAN 47
A. Hasil Hyperparameter Optimization 47
B. Hasil Retraining Konfigurasi Terpilih 50
C. Hasil Evaluasi pada Test Set 52
D. Evaluasi Per Kelas 54
E. Analisis Confusion Matrix 56
F. Pembahasan Berdasarkan Fokus Penelitian 58
1. Performa Enam Konfigurasi pada Validation Set 58
2. Performa Konfigurasi Terpilih pada Test Set 59
3. Pola Kesalahan Mayoritas dan Minoritas 59
4. Implikasi terhadap Penggunaan Pipeline 60

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

G. Keterbatasan Hasil 61
BAB V PENUTUP 63
A. Kesimpulan 63
B. Saran 65
DAFTAR PUSTAKA 67
LAMPIRAN 71
CV/CP 72

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

DAFTAR TABEL

Tabel 2. 1	Penelitian terdahulu yang relevan	16
Tabel 3. 1	Definisi dan pemetaan label penelitian	31
Tabel 3. 2	Distribusi kelas pada keseluruhan dataset	32
Tabel 3. 3	Distribusi kelas pada setiap subset	33
Tabel 3. 4	Lingkungan perangkat keras dan perangkat lunak	34
Tabel 3. 5	Bobot inverse-frequency berdasarkan training set	38
Tabel 3. 6	Ruang hyperparameter optimization	40
Tabel 3. 7	Konfigurasi training model	41
Tabel 4. 1	Hasil enam konfigurasi pada validation set	48
Tabel 4. 2	Rerata deskriptif hasil HPO	49
Tabel 4. 3	Perkembangan metrik validation selama retraining	51
Tabel 4. 4	Metrik agregat pada test set	53
Tabel 4. 5	Classification report pada test set	54

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

DAFTAR GAMBAR

Gambar 3. 1	Alur pelaksanaan penelitian	37
Gambar 4. 1	Hubungan batch size, learning rate, dan Macro F1	49
Gambar 4. 2	Confusion matrix model pada test set	56

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

DAFTAR SINGKATAN

ANN	:	*Artificial Neural Network*
BERT	:	*Bidirectional Encoder Representations from Transformers*
CM	:	*Compatibility*
CNN	:	*Convolutional Neural Network*
CPU	:	*Central Processing Unit*
CSV	:	*Comma-Separated Values*
CUDA	:	*Compute Unified Device Architecture*
F1	:	*F1 Score*
FP16	:	*16-bit Floating-Point*
FR	:	*Functional Requirements*
GPU	:	*Graphics Processing Unit*
HPO	:	*Hyperparameter Optimization*
MCC	:	*Matthews Correlation Coefficient*
MN	:	*Maintainability*
NFR	:	*Non-Functional Requirements*
NLP	:	*Natural Language Processing*
PE	:	*Performance*
PO	:	*Portability*
RL	:	*Reliability*
SE	:	*Security*
SVM	:	*Support Vector Machine*
TF-IDF	:	*Term Frequency–Inverse Document Frequency*
US	:	*Usability*
W&B	:	*Weights & Biases*

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

BAB I
PENDAHULUAN

A. Latar Belakang

Kebutuhan perangkat lunak menjadi dasar dalam menentukan fungsi, batasan, dan kualitas sistem yang akan dikembangkan. Kebutuhan fungsional atau *functional requirements* (FR) menjelaskan fungsi serta layanan yang harus disediakan, sedangkan *non-functional requirements* (NFR) berkaitan dengan karakteristik kualitas dan batasan yang menyertai pelaksanaan fungsi tersebut. NFR seperti *Security*, *Performance*, *Reliability*, dan *Usability* dapat memengaruhi keputusan arsitektur serta penerimaan pengguna. Meskipun penting, batas antara FR dan NFR tidak selalu dinyatakan secara tegas dan konsep NFR sendiri memiliki persoalan definisi serta klasifikasi (Glinz, 2007).

Dalam pengembangan agile, kebutuhan sering didokumentasikan sebagai *user story*. Bentuk ini membantu menyampaikan kebutuhan secara ringkas dari sudut pandang pengguna, tetapi atribut kualitas dapat dinyatakan secara eksplisit maupun tersirat. (Gilson, Galster and Georis, 2019) menunjukkan bahwa informasi atribut kualitas dalam *user story* tidak selalu dapat dikenali melalui kata kunci. Kajian (Raharjana, Siahaan and Fatichah, 2021) serta (Amna and Poels, 2022) juga memperlihatkan berkembangnya penggunaan natural language processing pada *user story*, disertai tantangan berupa keterbatasan dataset, variasi data, dan proses evaluasi.

Klasifikasi manual terhadap kumpulan kebutuhan yang besar memerlukan waktu dan konsistensi penelaahan. Oleh sebab itu, berbagai

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

penelitian mengembangkan klasifikasi otomatis FR dan NFR menggunakan machine learning. Pendekatan awal banyak menggunakan fitur leksikal, sintaktis, Bag of Words, atau TF-IDF bersama algoritma klasik (Kurtanović and Maalej, 2017; Canedo and Mendes, 2020). Perkembangan deep learning dan pretrained language model memungkinkan teks direpresentasikan secara kontekstual dengan ketergantungan lebih kecil pada rekayasa fitur manual.

Klasifikasi otomatis pada penelitian ini diposisikan sebagai alat bantu analisis, bukan pengganti penilaian manusia. Prediksi dapat membantu menyaring dan mengelompokkan backlog sebelum pemeriksaan analis. Namun, kebutuhan kualitas yang salah diprediksi sebagai FR dapat luput dari perhatian, sedangkan FR yang salah diprediksi sebagai NFR dapat menambah beban pemeriksaan. Konsekuensi tersebut membuat analisis per kelas dan arah kesalahan sama pentingnya dengan skor agregat.

Bidirectional Encoder Representations from Transformers (BERT) merupakan pretrained language model berbasis Transformer encoder yang membentuk representasi bidirectional dari konteks token (Devlin *et al.*, 2019). Kemampuan fine-tuning memungkinkan BERT disesuaikan untuk klasifikasi teks dengan menambahkan classification head. Dalam konteks kebutuhan perangkat lunak, model seperti BERT relevan karena suatu atribut kualitas dapat bergantung pada hubungan antarkata dan keseluruhan konteks kalimat.

(Alhaizaey and Al-Mashari, 2025b) mengevaluasi BERT, DistilBERT, RoBERTa, dan XLNet untuk klasifikasi FR serta tujuh kategori NFR pada agile user story. Penelitian tersebut menggunakan dataset 10.006 user story dengan delapan kelas, yaitu FR, *Compatibility*,

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

Maintainability, *Performance*, *Portability*, *Reliability*, *Security*, dan *Usability*. Artikel pendamping menjelaskan bahwa dataset disusun dari beberapa sumber akademik dan repositori serta melalui proses pelabelan dan tinjauan ahli (Alhaizaey and Al-Mashari, 2025a).

Penelitian acuan menunjukkan kelayakan pretrained language model untuk tugas tersebut, tetapi pertanyaan lanjutan dapat dipersempit pada perilaku satu pipeline dan protokol evaluasinya. Perbandingan banyak arsitektur menjawab pilihan model, sedangkan evaluasi variasi fine-tuning pada satu model menjawab konfigurasi mana yang dipilih dalam ruang tertentu dan bagaimana kesalahannya tersebar. Penelitian ini mengambil arah kedua agar proses pemilihan dan hasil per kelas dapat dibahas lebih rinci.

Dataset tersebut memiliki distribusi yang tidak seimbang. FR mencakup 6.457 data atau 64,53%, sedangkan 3.549 data lain tersebar pada tujuh kelas NFR. Kelas NFR terkecil, *Maintainability*, hanya memiliki 379 data, sedangkan *Usability* memiliki 780 data. Ketidakseimbangan dapat membuat kelas dengan jumlah lebih besar lebih sering berkontribusi terhadap pembaruan parameter model. Namun, kesulitan klasifikasi tidak hanya ditentukan oleh frekuensi karena tumpang tindih karakteristik, variasi teks, dan kualitas label juga berpengaruh (He and Garcia, 2009; Krawczyk, 2016).

Penanganan class imbalance dapat dilakukan pada tingkat data atau algoritma. Pendekatan tingkat data mengubah distribusi menggunakan oversampling atau undersampling, sedangkan pendekatan tingkat algoritma mengubah kontribusi contoh pada fungsi objektif (Johnson and Khoshgoftaar, 2019). Penelitian ini menggunakan pendekatan tingkat algoritma berupa Inverse-Frequency Weighted

Focal Loss. Bobot inverse-frequency memberi kontribusi relatif lebih besar kepada kelas dengan frekuensi lebih kecil. Focal Loss menurunkan kontribusi contoh yang mudah diprediksi melalui faktor yang bergantung pada probabilitas kelas aktual (Lin *et al.*, 2017). Penggabungan keduanya membuat loss dipengaruhi oleh frekuensi kelas dan tingkat kesulitan contoh.

Pendekatan tersebut dipilih karena kedua mekanisme bekerja pada sisi yang berbeda. Inverse-frequency weighting ditentukan sebelum training berdasarkan distribusi label, sedangkan focal factor berubah untuk setiap contoh mengikuti keluaran model. Sebuah contoh dari kelas berfrekuensi kecil memperoleh class weight lebih besar, tetapi kontribusi akhirnya tetap dipengaruhi oleh tingkat keyakinan prediksi. Sebaliknya, contoh dari kelas mayoritas tidak selalu diabaikan karena contoh yang sulit masih dapat menghasilkan focal factor relatif besar. Hubungan ini menjadikan pipeline layak dievaluasi pada data multikelas yang tidak seimbang, tetapi tidak secara otomatis menjamin peningkatan performa.

Performa fine-tuning BERT juga dipengaruhi oleh hyperparameter. Learning rate menentukan besar pembaruan parameter, sedangkan batch size menentukan jumlah contoh dalam satu langkah training. (Devlin *et al.*, 2019) menggunakan beberapa learning rate dan batch size dalam prosedur fine-tuning BERT. Penelitian ini mengadaptasi ruang tersebut secara terarah dengan menguji learning rate 2e-5, 3e-5, dan 5e-5 serta batch size 16 dan 32. Enam kombinasi dievaluasi pada validation set dan konfigurasi dipilih berdasarkan Macro F1.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

Pemilihan ruang yang terbatas merupakan kompromi antara kebutuhan eksplorasi dan biaya komputasi. Penelitian tidak berusaha menemukan semua kemungkinan konfigurasi BERT. Sebaliknya, pencarian diarahkan pada nilai yang memiliki dasar dari prosedur fine-tuning model aslinya dan dapat dijalankan secara konsisten pada sumber daya yang tersedia. Dengan membatasi pencarian pada dua hyperparameter, hubungan antara konfigurasi dan hasil lebih mudah ditelusuri, sedangkan parameter lain dipertahankan sama.

Macro F1 digunakan karena menghitung rata-rata F1 dengan kontribusi yang sama dari setiap kelas. Pada data tidak seimbang, Accuracy dan Weighted F1 dapat lebih banyak dipengaruhi kelas dengan support besar. Penggunaan Macro F1 tidak menggantikan metrik lain, sehingga penelitian juga melaporkan Accuracy, Weighted F1, Matthews Correlation Coefficient (MCC), Precision, Recall, F1 setiap kelas, serta *confusion matrix* (Sokolova and Lapalme, 2009; Jurman, Riccadonna and Furlanello, 2012).

Pemisahan validation dan test juga menjadi bagian penting dari rancangan. Validation set digunakan ketika model masih berada dalam proses pemilihan, sedangkan test set baru digunakan setelah konfigurasi dan checkpoint ditetapkan. Pemisahan fungsi tersebut memberikan gambaran performa akhir pada data yang tidak ikut menentukan pilihan model. Selain nilai agregat, evaluasi per kelas dan arah kesalahan diperlukan karena dua model dengan Accuracy yang serupa dapat menghasilkan pola kesalahan yang berbeda pada kelas NFR.

Dataset penelitian dipilih dengan pertimbangan kesinambungan metodologis. Objek, skema label, dan distribusi yang sama dengan penelitian (Alhaizaey and Al-Mashari, 2025b) memungkinkan pipeline

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

BERT yang ditetapkan dievaluasi tanpa mencampurkan perubahan tersebut dengan pergantian domain atau taksonomi. Dataset juga menyediakan jumlah data yang lebih besar daripada sejumlah penelitian user story terdahulu dan secara langsung mendukung tugas klasifikasi satu kelas FR serta tujuh kategori NFR.

Penelitian ini karena itu membatasi klaim pada dataset, pembagian data, model, dan konfigurasi yang digunakan. Fokusnya bukan membuktikan bahwa setiap komponen pipeline lebih unggul daripada alternatif lain. Penelitian mengevaluasi enam konfigurasi learning rate dan batch size pada validation set, menilai konfigurasi terpilih pada test set independen, dan menganalisis pola kesalahan antara FR dan tujuh kelas NFR. Dengan fokus tersebut, kontribusi penelitian terletak pada proses yang jelas, pemisahan fungsi validation dan test, pelaporan per kelas, serta pembahasan hasil yang dapat diaudit.

Kebutuhan akan evaluasi semacam ini muncul karena satu skor agregat belum cukup untuk menjelaskan perilaku model pada delapan kelas. Pipeline dapat menghasilkan nilai Accuracy tinggi karena sebagian besar data berasal dari FR, sementara beberapa kategori NFR tetap sulit dikenali. Sebaliknya, peningkatan perhatian terhadap kelas berfrekuensi kecil dapat disertai kesalahan FR menjadi NFR. Analisis confusion matrix diperlukan untuk mengetahui arah pertukaran tersebut dan melengkapi interpretasi Macro F1.

Celah yang ditangani penelitian bukan ketiadaan metode klasifikasi NFR, melainkan kebutuhan akan evaluasi terarah dan dapat ditelusuri pada kombinasi BERT, inverse-frequency weighting, dan Focal Loss dengan pemisahan validation serta test. Hasilnya diharapkan memberi gambaran yang cukup rinci untuk menilai pipeline pada dataset

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

yang dipilih sekaligus menunjukkan aspek yang masih memerlukan pengujian lanjutan.

B. Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah penelitian adalah sebagai berikut.

1. Bagaimana performa enam konfigurasi learning rate dan batch size BERT-base-uncased dengan Inverse-Frequency Weighted Focal Loss pada validation set?
2. Bagaimana performa konfigurasi terpilih pada test set berdasarkan Macro F1, Accuracy, Weighted F1, MCC, serta Precision, Recall, dan F1 setiap kelas?
3. Bagaimana pola kesalahan klasifikasi antara kelas mayoritas FR dan tujuh kelas NFR pada test set?

C. Batasan Masalah

Penelitian dibatasi agar metode, hasil, dan kesimpulannya tetap sesuai dengan eksperimen yang dilakukan. Batasan penelitian adalah sebagai berikut.

1. Data penelitian dibatasi pada 10.006 agile user story berbahasa Inggris dengan satu kelas FR dan tujuh kategori NFR dari dataset (Alhaizaey and Al-Mashari, 2025a, 2025b).
2. Model dibatasi pada BERT-base-uncased dengan pipeline Inverse-Frequency Weighted Focal Loss. Hyperparameter yang dicari hanya learning rate dan batch size pada enam kombinasi yang telah ditetapkan.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

3. Evaluasi menggunakan satu pembagian berstrata 80:10:10 dan satu run untuk setiap konfigurasi dengan seed 42. Penelitian tidak mengukur signifikansi statistik atau variasi lintas seed.
4. Pemilihan konfigurasi dan checkpoint menggunakan Macro F1 pada validation set, sedangkan test set digunakan setelah proses pemilihan selesai.
5. Penelitian mengevaluasi pipeline sebagai satu kesatuan dan tidak mengisolasi dampak Focal Loss, weighting, preprocessing, maupun rasio split melalui baseline atau ablation.
6. Kesimpulan dibatasi pada dataset, model, pembagian data, dan lingkungan penelitian sehingga tidak digeneralisasi langsung ke domain atau sumber kebutuhan lain.

D. Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut.

1. Mengevaluasi performa enam konfigurasi learning rate dan batch size BERT-base-uncased dengan Inverse-Frequency Weighted Focal Loss pada validation set serta menentukan konfigurasi berdasarkan Macro F1.
2. Mengevaluasi performa konfigurasi terpilih pada test set menggunakan Macro F1, Accuracy, Weighted F1, MCC, serta Precision, Recall, dan F1 setiap kelas.
3. Menganalisis pola kesalahan antara kelas mayoritas FR dan tujuh kelas NFR berdasarkan confusion matrix test set.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

E. Manfaat Penelitian

Penelitian diharapkan menambah bukti empiris mengenai penerapan BERT dengan Inverse-Frequency Weighted Focal Loss pada klasifikasi kebutuhan perangkat lunak tidak seimbang. Hasil agregat, per kelas, dan confusion matrix dapat memperkaya pemahaman mengenai perilaku pipeline pada satu kelas FR dan tujuh kategori NFR. Dokumentasi batas interpretasi juga membantu membedakan reproduksibilitas satu run dari kestabilan temuan lintas seed.

Penelitian menyediakan prosedur yang dapat ditelusuri mulai dari pembagian data, pembentukan bobot kelas, HPO, pemilihan checkpoint, retraining, hingga evaluasi test. Prosedur dan artefak tersebut dapat menjadi rujukan bagi penelitian lanjutan yang menggunakan dataset atau tugas serupa. Hasil per kelas dan pola confusion matrix juga dapat membantu menentukan kategori yang memerlukan pemeriksaan lebih lanjut apabila pipeline dikembangkan menjadi alat bantu klasifikasi kebutuhan.

Manfaat praktis tersebut tidak berarti model siap digunakan langsung dalam seluruh organisasi. Penerapan pada lingkungan nyata tetap memerlukan validasi pada data organisasi, penyesuaian taksonomi, pemeriksaan kualitas label, serta evaluasi risiko kesalahan sesuai kebutuhan penggunaannya.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

BAB V
PENUTUP

A. Kesimpulan

Penelitian ini mengevaluasi BERT-base-uncased dengan Inverse-Frequency Weighted Focal Loss untuk klasifikasi delapan kelas kebutuhan perangkat lunak pada agile user story. Dataset berjumlah 10.006 data dibagi secara berstrata menjadi 80% training, 10% validation, dan 10% test menggunakan seed 42. Hyperparameter optimization menguji enam kombinasi learning rate dan batch size, sedangkan Macro F1 validation digunakan untuk memilih konfigurasi dan checkpoint.

Enam konfigurasi menghasilkan Macro F1 validation antara 0,882316 dan 0,891309. Learning rate 2e-5 dan batch size 32 memperoleh Macro F1 tertinggi sebesar 0,891309 sehingga dipilih untuk retraining. Konfigurasi learning rate 5e-5 dan batch size 32 memperoleh Accuracy, Weighted F1, dan MCC tertinggi, tetapi tidak dipilih karena kriteria seleksi yang ditetapkan adalah Macro F1. Selisih Macro F1 dua konfigurasi teratas hanya 0,000835 dan tidak menjadi bukti kestabilan peringkat karena setiap konfigurasi hanya dijalankan satu kali.

Setelah konfigurasi terpilih dilatih ulang dari BERT dasar, best checkpoint dievaluasi pada 1.001 data test yang tidak digunakan dalam HPO. Model menghasilkan Accuracy 0,9351, Weighted F1 0,9358, Macro F1 0,9027, dan MCC 0,8875. Sebanyak 936 data diprediksi benar dan 65 data salah. F1 per kelas berada pada rentang 0,8323–0,9639. FR

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

memperoleh F1 tertinggi sebesar 0,9639, sedangkan US memperoleh F1 terendah sebesar 0,8323 dan SE memperoleh F1 terendah kedua sebesar 0,8611. Hasil tersebut menunjukkan bahwa performa agregat yang tinggi tidak tersebar secara seragam pada seluruh kelas.

Analisis confusion matrix menunjukkan bahwa 32 data FR diprediksi sebagai NFR, 14 data NFR diprediksi sebagai FR, dan 19 data NFR diprediksi sebagai kategori NFR lain. Pasangan kesalahan dua arah terbesar adalah FR–US sebanyak 19 dan FR–SE sebanyak 17. Arah kesalahan FR menjadi NFR lebih banyak daripada NFR menjadi FR pada test set ini. Selain itu, kelas NFR dengan support kecil tidak selalu memperoleh performa terendah; MN dan PE justru memiliki recall di atas 0,97. Dengan demikian, pola kesalahan tidak dapat dijelaskan hanya dari ukuran kelas.

Perbedaan antara metrik agregat dan per kelas menegaskan pentingnya pemilihan ukuran evaluasi. Weighted F1 yang lebih tinggi daripada Macro F1 dipengaruhi oleh besarnya support FR dan F1 FR yang tinggi. Penggunaan Macro F1 sebagai kriteria seleksi memungkinkan hasil kelas NFR tetap memengaruhi keputusan secara setara, sedangkan pelaporan Weighted F1 dan Accuracy mempertahankan gambaran performa sesuai distribusi aktual.

Hasil penelitian menjelaskan perilaku empiris pipeline pada dataset, pembagian data, konfigurasi, dan seed yang digunakan. Penelitian tidak membuktikan keunggulan Focal Loss, inverse-frequency weighting, tanpa pembersihan tambahan, atau rasio split terhadap alternatif lain karena komponen tersebut tidak diuji melalui baseline maupun ablation. Hasil juga tidak menunjukkan signifikansi statistik atau kestabilan lintas seed.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

B. Saran

Penelitian berikutnya dapat mengulangi setiap konfigurasi menggunakan beberapa seed dan melaporkan rerata serta penyebaran metrik. Pengulangan tersebut akan membantu menilai apakah peringkat konfigurasi tetap konsisten, terutama karena selisih Macro F1 dua konfigurasi teratas sangat kecil. Penentuan jumlah run perlu mempertimbangkan biaya komputasi dan tujuan inferensi yang ingin dicapai.

Apabila tujuan penelitian berikutnya adalah mengukur kontribusi setiap komponen, rancangan dapat menambahkan baseline dan ablation yang terkontrol. Focal Loss dapat dibandingkan dengan Cross-Entropy pada weighting yang sama, sedangkan inverse-frequency weighting dapat dibandingkan dengan kondisi tanpa weighting pada loss yang sama. Preprocessing dan rasio split juga dapat diuji secara terpisah. Rancangan tersebut menjawab pertanyaan kausal atau komparatif yang berbeda dari evaluasi pipeline pada penelitian ini.

Ruang HPO dapat diperluas secara bertahap dengan memasukkan weight decay, warmup ratio, jumlah epoch, dan nilai gamma. Namun, perluasan sebaiknya disertai validation protocol yang jelas dan test set yang tetap disimpan hingga seluruh keputusan model selesai. Jika beberapa pipeline dibandingkan, setiap pipeline dapat diberi ruang optimisasi yang sepadan agar perbandingannya tidak bergantung pada parameter yang hanya sesuai untuk salah satu metode.

Analisis kesalahan dapat diperdalam dengan menyimpan teks, label aktual, probabilitas prediksi, dan label prediksi untuk setiap contoh. Pemeriksaan contoh FR–US dan FR–SE dapat digunakan untuk mengidentifikasi kemiripan semantik, ambiguitas label, atau pola

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

bahasa yang menyebabkan kebingungan. Analisis semacam itu perlu membedakan interpretasi peneliti dari bukti anotasi yang tersedia.

Penelitian lanjutan juga dapat meminta penelaah domain untuk memeriksa sebagian contoh salah klasifikasi. Penelaahan tersebut dapat membantu membedakan kesalahan model dari contoh yang memang ambigu atau berpotensi memiliki lebih dari satu atribut kualitas. Jika ditemukan ambiguitas sistematis, pengembangan taksonomi multilabel dapat dipertimbangkan sebagai penelitian tersendiri.

Generalisasi dapat diuji pada sumber kebutuhan lain, bahasa lain, atau data organisasi yang memiliki taksonomi berbeda. Penggunaan ulasan aplikasi juga memungkinkan, tetapi memerlukan perumusan ulang objek, skema label, dan proses anotasi karena ulasan pengguna bukan pengganti langsung bagi dataset agile user story. Sebelum penggunaan praktis, model perlu divalidasi kembali pada data serta kebutuhan organisasi yang menjadi konteks penerapannya.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

DAFTAR PUSTAKA

Alhaizaey, A. and Al-Mashari, M. (2025a) 'An Aggregated Dataset of Agile User Stories and Use Case Taxonomy for AI-Driven Research', *International Journal of Advanced Computer Science and Applications*, 16(9), pp. 250–259.

Alhaizaey, A. and Al-Mashari, M. (2025b) 'Automated Classification and Identification of Non-Functional Requirements in Agile-Based Requirements Using Pre-Trained Language Models', *IEEE Access*, 13, pp. 87401–87417. Available at: https://doi.org/10.1109/ACCESS.2025.3570359.

Amna, A.R. and Poels, G. (2022) 'Systematic Literature Mapping of User Story Research', *IEEE Access*, 10, pp. 51723–51746. Available at: https://doi.org/10.1109/ACCESS.2022.3173745.

Aurelio, Y.S. *et al.* (2019) 'Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function', *Neural Processing Letters*, 50(2), pp. 1937–1949. Available at: https://doi.org/10.1007/s11063-018-09977-1.

Bergstra, J. and Bengio, Y. (2012) 'Random Search for Hyper-Parameter Optimization', *Journal of Machine Learning Research*, 13(10), pp. 281–305. Available at: https://jmlr.org/papers/v13/bergstra12a.html.

Binkhonain, M. and Zhao, L. (2019) 'A Review of Machine Learning Algorithms for Identification and Classification of Non-Functional Requirements', *Expert Systems with Applications: X*, 1, p. 100001. Available at: https://doi.org/10.1016/j.eswax.2019.100001.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

Bouthillier, X., Laurent, C. and Vincent, P. (2019) 'Unreproducible Research is Reproducible', in *Proceedings of the 36th International Conference on Machine Learning*. (Proceedings of Machine Learning Research), pp. 725–734.

Canedo, E.D. and Mendes, B.C. (2020) 'Software Requirements Classification Using Machine Learning Algorithms', *Entropy*, 22(9), p. 1057. Available at: https://doi.org/10.3390/e22091057.

Casillo, F., Deufemia, V. and Gravino, C. (2022) 'Detecting Privacy Requirements from User Stories with NLP Transfer Learning Models', *Information and Software Technology*, 146, p. 106853. Available at: https://doi.org/10.1016/j.infsof.2022.106853.

Devlin, J. *et al.* (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186.

Dodge, J. *et al.* (2020) 'Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping', *arXiv preprint arXiv:2002.06305* [Preprint]. Available at: https://arxiv.org/abs/2002.06305.

Gilson, F., Galster, M. and Georis, F. (2019) 'Extracting Quality Attributes from User Stories for Early Architecture Decision Making', *2019 IEEE International Conference on Software Architecture Companion (ICSA-C)*, pp. 129–136. Available at: https://doi.org/10.1109/ICSA-C.2019.00031.

Glinz, M. (2007) 'On Non-Functional Requirements', in *15th IEEE International Requirements Engineering Conference (RE 2007)*,

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

pp. 21–26. Available at: https://doi.org/10.1109/RE.2007.45.

He, H. and Garcia, E.A. (2009) 'Learning from Imbalanced Data', *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263–1284. Available at: https://doi.org/10.1109/TKDE.2008.239.

Johnson, J.M. and Khoshgoftaar, T.M. (2019) 'Survey on Deep Learning with Class Imbalance', *Journal of Big Data*, 6, p. 27. Available at: https://doi.org/10.1186/s40537-019-0192-5.

Jurman, G., Riccadonna, S. and Furlanello, C. (2012) 'A Comparison of MCC and CEN Error Measures in Multi-Class Prediction', *PLoS ONE*, 7(8), p. e41882. Available at: https://doi.org/10.1371/journal.pone.0041882.

Krawczyk, B. (2016) 'Learning from Imbalanced Data: Open Challenges and Future Directions', *Progress in Artificial Intelligence*, 5, pp. 221–232. Available at: https://doi.org/10.1007/s13748-016-0094-0.

Kurtanović, Z. and Maalej, W. (2017) 'Automatically Classifying Functional and Non-Functional Requirements Using Supervised Machine Learning', *2017 IEEE 25th International Requirements Engineering Conference (RE)*, pp. 490–495. Available at: https://doi.org/10.1109/RE.2017.82.

Lin, T.-Y. *et al.* (2017) 'Focal Loss for Dense Object Detection', *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988.

Mosbach, M., Andriushchenko, M. and Klakow, D. (2021) 'On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines', *International Conference on Learning*

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

Representations (ICLR) [Preprint]. Available at: https://arxiv.org/abs/2006.04884.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12(85), pp. 2825–2830. Available at: https://jmlr.org/papers/v12/pedregosa11a.html.

Raharjana, I.K., Siahaan, D. and Fatichah, C. (2021) 'User Stories and Natural Language Processing: A Systematic Literature Review', *IEEE Access*, 9, pp. 53811–53826. Available at: https://doi.org/10.1109/ACCESS.2021.3070606.

Rahman, K. *et al.* (2024) 'A Deep Learning Framework for Non-Functional Requirement Classification', *Scientific Reports*, 14, p. 3216. Available at: https://doi.org/10.1038/s41598-024-52802-0.

Reimers, N. and Gurevych, I. (2017) 'Reporting Score Distributions Makes a Difference: Performance Study of {LSTM}-networks for Sequence Tagging', in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348. Available at: https://doi.org/10.18653/v1/D17-1035.

Sokolova, M. and Lapalme, G. (2009) 'A Systematic Analysis of Performance Measures for Classification Tasks', *Information Processing & Management*, 45(4), pp. 427–437. Available at: https://doi.org/10.1016/j.ipm.2009.03.002.

Vaswani, A. *et al.* (2017) 'Attention Is All You Need', *Advances in Neural Information Processing Systems*. Available at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA