SISTEM OTOMATIS KLASIFIKASI BUKTI PEMBAYARAN MENGGUNAKAN OCR DAN EMBEDDING BERT DENGAN PENDEKATAN MULTI-MODEL PEMBELAJARAN MESIN

Mitchella Sinta Larasati[1], Suryasatriya Trihandaru[2], Hanna Arini Parhusip[3]
[1,2,3]Universitas Kristen Satya Wacana
632023006@student.uksw.edu, suryasatriya@uksw.edu,
hanna.parhusip@uksw.edu

ABSTRACT

The verification process of payment receipts in school environments is still predominantly conducted manually, leading to inefficiency and a high potential for human error. This study proposes an automated system for classifying the validity of digital payment receipts by combining Optical Character Recognition (OCR), BERT (Bidirectional Encoder Representations from Transformers) embeddings, and multi-model machine learning approaches. The system integrates EasyOCR for text extraction from payment receipts, BERT for generating semantic text representations, and four classification algorithms: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Naive Bayes (NB), and Logistic Regression (LR). The dataset consists of 185 payment receipt samples, comprising 149 valid and 36 invalid instances, collected via Google Forms and stored in a SQLite database. Experimental results demonstrate that the Multi-Layer Perceptron (MLP) model achieves the highest accuracy of 97% with a test size of 0.2, followed by Logistic Regression with an accuracy of 96.2%, while Naive Bayes exhibits the lowest performance with an accuracy of 85.7%. The proposed system is successfully implemented in a Streamlit-based application, enabling real-time verification of payment receipts with an average processing time of 1.16 seconds per sample.

Keywords : *Optical Character Recognition (OCR), BERT Embedding, Machine Learning, Document Classification, Payment Verification, Multi-Model Classification.*

ABSTRAK

Proses verifikasi bukti pembayaran di lingkungan sekolah masih dilakukan secara manual, menyebabkan inefisiensi dan potensi kesalahan manusia. Penelitian ini mengembangkan sistem otomatis untuk mengklasifikasikan validitas bukti pembayaran digital menggunakan kombinasi Optical Character Recognition (OCR), embedding BERT (Bidirectional Encoder Representations from Transformers), dan model pembelajaran mesin multi-model. Sistem mengintegrasikan EasyOCR untuk ekstraksi teks, BERT untuk representasi semantik, dan empat algoritma klasifikasi: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Naive Bayes (NB), dan Logistic Regression (LR). Dataset berisi 185 sampel bukti pembayaran (149 valid, 36 tidak valid) yang dikumpulkan melalui Google Forms dan tersimpan di

SQLite. Hasil pengujian menunjukkan bahwa Multi-Layer Perceptron (MLP) mencapai akurasi tertinggi 97% pada test size 0.2, diikuti Logistic Regression dengan 96.2%, sementara Naive Bayes menunjukkan performa terendah dengan 85.7%. Sistem berhasil diimplementasikan dalam aplikasi Streamlit yang memungkinkan verifikasi real-time bukti pembayaran dengan waktu proses rata-rata 1.16 detik per sampel.
Kata kunci : Optical Character Recognition (OCR), BERT Embedding, Machine Learning, Klasifikasi Dokumen, Verifikasi Pembayaran, Multi-Model Classification

A. Pendahuluan

Perkembangan pesat transformasi digital dalam sistem keuangan telah mendorong masyarakat dari berbagai sektor untuk secara masif mengadopsi metode pembayaran elektronik, termasuk di lingkungan pendidikan seperti sekolah dan universitas. Kemudahan dan efisiensi yang ditawarkan oleh sistem pembayaran digital menjadikannya pilihan yang menarik bagi institusi dalam mengelola transaksi, salah satunya adalah pemesanan layanan lainnya seperti katering sekolah.

Namun, di balik kemudahan tersebut terdapat tantangan signifikan, khususnya terkait dengan proses verifikasi bukti pembayaran. Bukti pembayaran yang dikirim oleh pengguna biasanya dalam bentuk gambar atau tangkapan layar hasil transfer dari aplikasi perbankan. Proses verifikasi ini, terutama dalam sistem layanan daring seperti pemesanan katering sekolah, umumnya masih dilakukan secara manual oleh pihak pengelola. Mereka harus memeriksa satu per satu bukti transfer yang diunggah melalui Google Form lalu mencocokkannya dengan data pemesanan yang tersedia. Hal ini tidak hanya memakan waktu, tetapi juga meningkatkan risiko kesalahan manusia, terutama ketika volume transaksi meningkat secara signifikan (Irianti et al., 2025).

Melihat permasalahan tersebut, dibutuhkan solusi yang dapat mengotomatisasi proses verifikasi bukti pembayaran digital secara cerdas, fleksibel, dan akurat. Salah satu pendekatan yang menjanjikan adalah dengan memanfaatkan teknologi Optical Character Recognition (OCR) untuk mengekstraksi teks dari gambar, sehingga informasi penting dari bukti pembayaran dapat diproses dengan cepat (Simanjorang, 2022). Namun,

sering kali teks hasil ekstraksi ini tidak memiliki struktur yang jelas dan mengandung noise. Kombinasi angka, symbol dan teks dalam dokumen transaksi semakin mempersulit proses klarfikasi dan identifikasi informasi penting. Keterbatasan ini menghambat pengembangan sistem verifikasi otomatis yang akuran dann efisien.

Sehingga, dibutuhkan sebuah pendekatan cerdas yang tidak hanya mampu mengekstrak teks, tetapi juga memahami konteks sematik atau makna dari teks yang terekstrak tersebut, sekaligus menangani noise dan keragaman format bukti pembayaran. Dalam penelitian ini dikaji potensi penggunaan embedding teks berbasis BERT (Bidirectional Encoder Representations from Transformers) yang dikembangkan oleh AI dari Google (Aljabar, 2024). BERT memungkinkan representasi semantic dari dokumen tidak terstruktur, memahami makna sebuah kata dengan mempertimbangkan konteks di sekelilingnya—baik yang muncul sebelum maupun sesudah kata tersebut dalam kalimat (Hanum et al., 2024). Setelah representasi vektor dari teks diperoleh melalui embedding BERT, proses klasifikasi dilakukan menggunakan algoritma machine learning seperti Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Naive Bayes, dan Logistic Regression. Tujuannya adalah untuk mengidentifikasi apakah suatu bukti pembayaran merupakan bukti valid (berisi informasi transfer yang sah dan lengkap) atau tidak valid (berisi teks yang tidak terkait atau tidak sesuai).

Tidak berhenti sampai di situ, sistem juga dituntut untuk dapat memverifikasi kebenaran nominal transfer, yaitu apakah jumlah uang yang tertera pada bukti transfer sesuai dengan data pesanan pengguna. Untuk mendukung hal ini, digunakan pendekatan Regular Expression (Regex) yang berfungsi mengekstraksi informasi numerik seperti nominal rupiah dari teks hasil OCR. Nilai nominal tersebut kemudian dibandingkan dengan nilai referensi yang tersimpan dalam sistem, seperti database SQLite atau spreadsheet pesanan, untuk memastikan kesesuaian antara pembayaran dan tagihan.

Beberapa penelitian terkini menunjukkan bahwa kombinasi teknologi OCR, NLP, dan machine learning mampu mengotomatisasi

proses ekstraksi dan klasifikasi informasi dari dokumen keuangan seperti bukti pembayaran, invoice, atau laporan transaksi. Salah satu pendekatan yang relevan adalah CUTIE (Convolutional Universal Text Information Extractor), yang mampu memahami dokumen berbasis layout seperti kuitansi dengan menggabungkan informasi spasial dan semantik menggunakan CNN, tanpa memerlukan pretraining tambahan (Zhao et al., n.d.). Sementara itu, penerapan BERT dalam industri perbankan telah dibuktikan dalam studi "Automated Document Classification using BERT in Banking Industry", yang menunjukkan efektivitas klasifikasi otomatis terhadap dokumen keuangan untuk mengurangi beban kerja manual dan kesalahan klasifikasi (Gopalakrishnan, 2021). Model KPI-BERT juga memperluas pemanfaatan transformer untuk ekstraksi entitas dan relasi dalam laporan keuangan, memperlihatkan keunggulan BERT dalam memahami konteks finansial yang kompleks (Hillebrand et al., 2022). Di sisi lain, kajian sistematis oleh Journal of Big Data menegaskan pentingnya pendekatan hybrid berbasis AI (OCR, NLP, dan deep learning) dalam menangani dokumen tak terstruktur seperti formulir atau bukti pembayaran digital (Mahadevkar et al., 2025). Keunggulan BERT yang telah diadaptasi ke domain keuangan, seperti FinBERT, turut mendukung akurasi pemrosesan informasi dari teks keuangan (8. FINBERT).

Dengan menggabungkan OCR, embedding BERT, teknik Regex, dan klasifikasi machine learning, diharapkan sistem dapat memproses bukti pembayaran secara otomatis dari gambar mentah hingga menghasilkan keputusan yang akurat mengenai validitas dan kecocokan nominal transfer. Pendekatan ini tidak hanya meningkatkan efisiensi dan kecepatan proses verifikasi, tetapi juga mengurangi ketergantungan pada pemeriksaan manual yang rentan terhadap kesalahan.

B. Metode Penelitian

Penelitian ini fokus pada pengembangan sistem otomatis untuk mengklasifikasikan validitas bukti pembayaran digital (misalnya hasil tangkapan layar atau foto transfer bank). Sistem ini dirancang untuk mengidentifikasi apakah bukti pembayaran yang diunggah oleh pengguna tergolong valid atau tidak

valid berdasarkan informasi yang terekstraksi dari dokumen tidak terstruktur. Desain penelitian terdiri dari beberapa tahap utama, yaitu :

a. Pengumpulan Data, mencakup pengumpulan bukti pembayaran valid dan tidak valid yang dikumpulkan melalui Google Forms.
b. Ekstraksi Teks dengan OCR (Optical Character Recognition) untuk mengubah citra menjadi teks.
c. Pembersihan dan Pemrosesan Teks menggunakan ekspresi reguler (RegEx) untuk mengambil elemen penting seperti nama bank, tanggal, jumlah transfer, dan nomor rekening.
d. Pembuatan Embedding Menggunakan Model BERT, untuk mengonversi teks hasil OCR menjadi representasi vektor numerik.
e. Klasifikasi Menggunakan Multi-Model Machine Learning (SVM, MLP, Naive Bayes, Logistic Regression).
f. Evaluasi Model menggunakan metrik akurasi, presisi, recall, dan F1-score untuk menentukan model terbaik.

Data dan Sumber Data

Data penelitian terdiri atas bukti pembayaran valid, yaitu pembayaran yang telah diverifikasi pihak sekolah sebagai sah dan sesuai tanggal pemesanan, serta data tidak valid berupa hasil OCR dari gambar yang tidak sesuai, seperti nota pembayaran. Seluruh data disimpan dalam basis data SQLite (mqtt_data.db), dengan setiap entri berisi teks hasil ekstraksi OCR dan label klasifikasi (valid atau invalid).

Arsitektur Sistem

Arsitektur sistem menggambarkan alur terintegrasi mulai dari unggahan bukti pembayaran hingga hasil klasifikasi yang ditampilkan pada aplikasi berbasis Streamlit. Tahapan sistem meliputi input data, ekstraksi teks, pemrosesan dan penyimpanan sementara, pembentukan embedding, klasifikasi multi-model, serta visualisasi hasil.

Input Data

Pengguna mengunggah bukti pembayaran melalui Google Form yang terintegrasi dengan Google Drive untuk penyimpanan gambar dan Google Sheet untuk pencatatan metadata. Integrasi ini memanfaatkan

ekosistem Google Workspace yang telah digunakan oleh sekolah.

OCR dan Ekstraksi Teks

Ekstraksi teks dari gambar dilakukan menggunakan EasyOCR untuk mengenali karakter pada bukti pembayaran digital. Teks hasil OCR kemudian dibersihkan menggunakan Regular Expression (RegEx) untuk mengekstraksi informasi penting seperti nama bank, nominal transfer, nomor rekening, dan tanggal transaksi.

Pemrosesan Teks dan NLP

Teks hasil OCR diproses menggunakan pendekatan Natural Language Processing (NLP), meliputi tokenisasi, normalisasi, penghapusan kata tidak relevan, serta Named Entity Recognition (NER). Proses ini menghasilkan teks bersih yang siap digunakan pada tahap pembentukan embedding.

Penyimpanan Sementara

Hasil OCR, teks yang telah diproses, dan hasil klasifikasi disimpan dalam database SQLite sebagai penyimpanan sementara. Selain itu, data juga dikelola dalam Pandas DataFrame untuk mendukung pemrosesan komputasi yang efisien.

Embedding dan Klasifikasi

Teks bersih direpresentasikan dalam bentuk vektor menggunakan embedding BERT untuk menangkap konteks semantik. Vektor ini kemudian diklasifikasikan menggunakan beberapa model pembelajaran mesin, yaitu SVM, MLP, Logistic Regression, dan Naive Bayes. Model dengan performa terbaik disimpan untuk digunakan pada tahap implementasi.

Visualisasi Hasil

Hasil klasifikasi ditampilkan melalui aplikasi Streamlit yang menunjukkan status validitas bukti pembayaran beserta skor probabilitas. Data hasil klasifikasi disimpan kembali ke database untuk keperluan audit dan pelacakan.

Evaluasi Model

Evaluasi dilakukan dengan variasi proporsi data latih dan uji (0.2–0.8). Metrik yang digunakan meliputi akurasi, presisi, recall, dan F1-score untuk menilai performa model secara menyeluruh, terutama pada kondisi data tidak seimbang. Model dengan nilai evaluasi terbaik dipilih sebagai model akhir sistem.

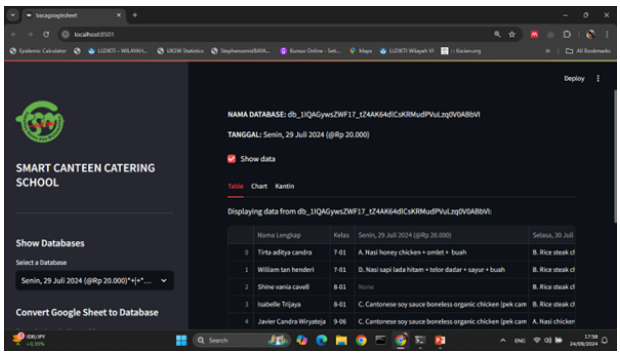
C. Hasil Penelitian dan Pembahasan

Implementasi Sistem

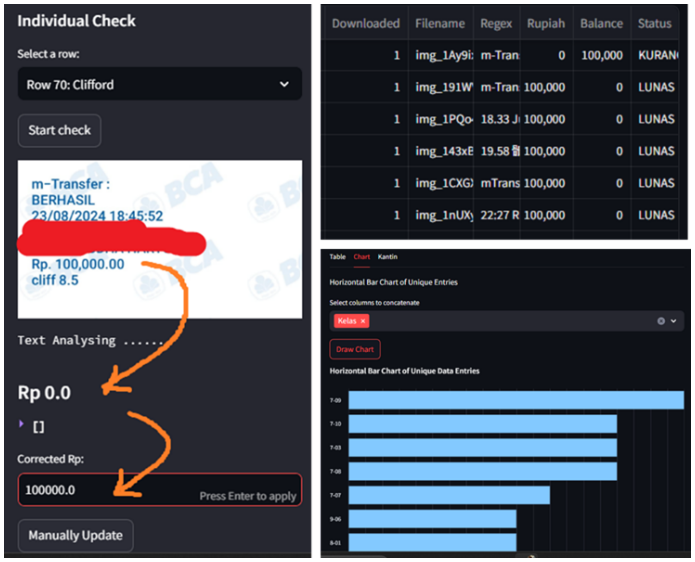
Implementasi sistem dilakukan dengan membangun aplikasi berbasis Streamlit yang terhubung dengan basis data SQLite (mqtt_data.db). Sistem ini bertujuan untuk mengklasifikasikan validitas bukti pembayaran digital secara otomatis melalui tahapan: pengumpulan data, ekstraksi teks menggunakan OCR, pemrosesan linguistik (NLP), pembentukan embedding menggunakan BERT, dan klasifikasi dengan beberapa model pembelajaran mesin.

Sistem dibangun menggunakan bahasa pemrograman Python dengan pustaka utama seperti EasyOCR, spaCy, Transformers (Hugging Face), dan scikit-learn. Arsitektur sistem memungkinkan alur kerja otomatis mulai dari proses unggah data hingga hasil klasifikasi yang ditampilkan secara real-time di antarmuka Streamlit.

Gambar Tampilan antarmuka aplikasi Streamlit sistem klasifikasi bukti pembayaran.



Gambar Contoh hasil ekstraksi teks dan klasifikasi validitas



	Rupiah	Balance	Status	Validate
) R	60000	0	LUNAS	valid
) R	100000	0	LUNAS	valid
NC	60000	0	LUNAS	valid
'ud	100000	0	LUNAS	valid
09:	100000	0	LUNAS	valid

Hasil Ekstraksi dan Pemrosesan Teks

Tahap ini menyajikan hasil ekstraksi teks dari gambar bukti pembayaran menggunakan EasyOCR, yang mampu membaca

berbagai format bukti pembayaran, seperti tangkapan layar aplikasi perbankan dan foto struk ATM. Teks hasil OCR dibersihkan menggunakan Regular Expression (RegEx) untuk mengekstraksi pola penting, seperti nominal transfer, tanggal transaksi, dan nama bank. Selanjutnya, pemrosesan Natural Language Processing (NLP) dilakukan melalui tokenisasi, normalisasi, serta Named Entity Recognition (NER) untuk menstandarkan teks dan mengidentifikasi entitas penting. Hasil pemrosesan disimpan dalam tabel preprocessed_data pada database SQLite dan dimuat ke dalam Pandas Data Frame untuk analisis lebih lanjut.
Pembentukan Embedding dan Klasifikasi

Pembentukan embedding dilakukan menggunakan model BERT (Bidirectional Encoder Representations from Transformers) untuk menghasilkan representasi vektor yang menangkap makna semantik teks secara kontekstual. Vektor embedding ini digunakan sebagai input bagi empat model pembelajaran mesin, yaitu Support Vector Machine (SVM), Neural Network (NN), Naive Bayes (NB), dan Logistic Regression (LR).

SVM dipilih karena kemampuannya menangani data berdimensi tinggi dan menentukan hyperplane optimal dengan margin maksimum, sehingga efektif untuk klasifikasi dokumen (Ovirianti et al., 2022). Neural Network digunakan karena kemampuannya mempelajari pola non-linear kompleks melalui proses backpropagation dan arsitektur yang fleksibel, yang terbukti meningkatkan akurasi klasifikasi (Skalka et al., 2025). Naive Bayes digunakan sebagai model probabilistik yang efisien dan umum dijadikan baseline dalam klasifikasi teks (Sari et al., 2023). Logistic Regression dipilih karena sifatnya yang sederhana, cepat, dan interpretatif, serta efektif dalam mengolah fitur embedding dengan penerapan regularisasi L2 untuk menghindari overfitting (Briggs et al., 2012).

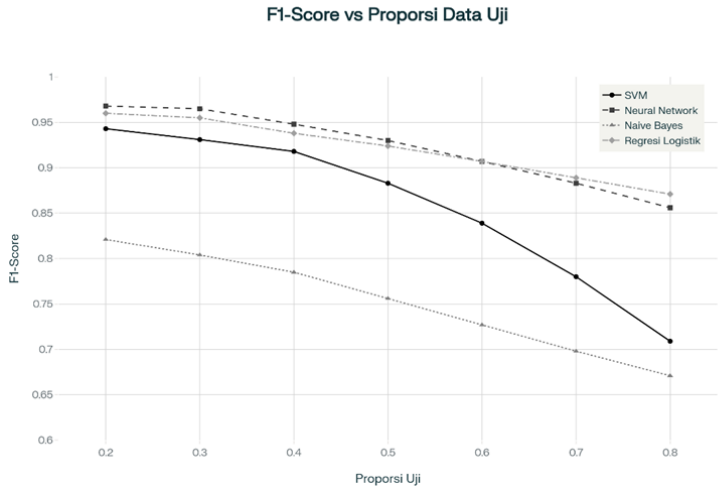
Masing-masing model diuji menggunakan beberapa variasi rasio data latih dan uji (0.2, 0.3, 0.5, 0.8, 0.9) untuk menemukan konfigurasi terbaik.

Tabel Hasil pengujian performa model dengan variasi proporsi data uji

Gambar Tabel Akurasi Antar odel klKasifikasi.

Test Size	SVM (Acc)	SVM (F1)	NN (Acc)	NN (F1)	NB (Acc)	NB (F1)	LR (Acc)	LR (F1)
0.2	0.946	0.943	0.970	0.968	0.857	0.821	0.962	0.960
0.3	0.934	0.931	0.968	0.965	0.841	0.804	0.957	0.955
0.4	0.922	0.918	0.952	0.948	0.823	0.785	0.941	0.938
0.5	0.889	0.883	0.935	0.930	0.798	0.756	0.928	0.924
0.6	0.847	0.839	0.914	0.907	0.776	0.727	0.912	0.907
0.7	0.791	0.780	0.892	0.883	0.752	0.698	0.895	0.889
0.8	0.723	0.709	0.867	0.856	0.729	0.671	0.878	0.8

Gambar Grafik F1-Score Terhadap Variasi Proporsi Data Uji



Dari hasil di atas, terlihat bahwa NN memberikan performa terbaik dengan akurasi tertinggi mencapai 97%

Analisis Hasil

Berdasarkan hasil pengujian :

Model dengan kinerja terbaik secara konsisten adalah Neural Network dan Logistic Regression, yang sama-sama menunjukkan akurasi tinggi hingga 97% pada test size 0.2 dan 0.3.

a. Neural Network mampu menyesuaikan bobot antar lapisan untuk menangkap hubungan non-linear dari vektor embedding BERT.
b. Logistic Regression tetap unggul karena data hasil embedding BERT relatif sudah terdistribusi baik, sehingga pemisahan linear masih efektif.

Naive Bayes menunjukkan performa terendah di hampir semua variasi data uji, terutama karena asumsi independensi antar fitur tidak cocok untuk data berbasis konteks semantik seperti hasil embedding BERT.

SVM mengalami penurunan tajam pada test size 0.5 ke atas, bahkan pada test size 0.8 dan 0.9 model gagal memprediksi kelas minoritas (“invalid”). Hal ini menunjukkan bahwa margin optimal SVM sulit dicapai ketika data latih menjadi terlalu sedikit atau tidak seimbang antar kelas.

Secara umum, semakin besar proporsi data uji, performa seluruh model cenderung menurun, karena jumlah data latih berkurang dan generalisasi model menjadi kurang stabil.

F1-Score menunjukkan pola yang mirip dengan akurasi, di mana Neural Network tetap paling stabil di semua proporsi data, membuktikan

kemampuannya dalam menjaga keseimbangan antara presisi dan recall.

D. Kesimpulan

Berdasarkan hasil penelitian dan implementasi sistem verifikasi bukti pembayaran otomatis berbasis Optical Character Recognition (OCR), Regular Expression (Regex), BERT Embedding, dan beberapa model Machine Learning (SVM, NN, Naive Bayes, dan Logistic Regression), penelitian ini telah menghasilkan beberapa temuan penting. Sistem berhasil mengotomatisasi proses verifikasi bukti pembayaran yang sebelumnya dilakukan secara manual oleh pengelola kantin sekolah. Proses otomatis ini memungkinkan pembacaan teks dari gambar bukti transfer melalui OCR, pembersihan dan ekstraksi pola penting menggunakan Regex, serta representasi semantik teks dengan BERT untuk mendukung klasifikasi berbasis machine learning. Pendekatan kombinatif OCR–Regex–BERT–ML terbukti efektif dalam mendeteksi validitas bukti pembayaran, dengan model pembelajaran mesin yang mampu membedakan antara bukti pembayaran valid dan tidak valid berdasarkan fitur teks hasil ekstraksi dengan tingkat akurasi tinggi pada data uji.

Dari hasil pengujian berbagai model, algoritma Neural Network (NN) menunjukkan performa terbaik ditinjau dari metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Hal ini menunjukkan bahwa model NN paling optimal dalam menangani data teks hasil embedding dari BERT dan mampu mengekstrak fitur-fitur kompleks dari representasi semantik yang dihasilkan. Penggunaan embedding BERT memberikan peningkatan signifikan terhadap performa model dibandingkan representasi teks konvensional seperti TF-IDF. Peningkatan ini disebabkan oleh kemampuan BERT dalam memahami konteks semantik secara dua arah (bidirectional), sehingga informasi pada bukti pembayaran dapat diinterpretasikan secara lebih akurat dan komprehensif. Integrasi sistem ke dalam platform Streamlit memungkinkan proses klasifikasi dilakukan secara interaktif dan real-time oleh pengguna, mendukung implementasi sistem verifikasi otomatis yang praktis dan user-friendly di lingkungan sekolah.

Secara keseluruhan, penelitian ini membuktikan bahwa penerapan teknologi NLP modern seperti BERT, dikombinasikan dengan pendekatan OCR dan Regex, dapat menjadi solusi efektif dalam mengotomatiskan proses administratif seperti verifikasi bukti pembayaran di sekolah. Teknologi ini mengurangi beban kerja manual, meningkatkan efisiensi operasional, dan meminimalkan kesalahan dalam proses verifikasi. Sebagai tindak lanjut, pengembangan sistem dapat diarahkan pada peningkatan dataset dengan menambahkan lebih banyak sampel bukti pembayaran dari berbagai bank dan metode transfer. Penambahan model deep learning berbasis transformer lainnya seperti RoBERTa atau DistilBERT juga dapat dipertimbangkan untuk membandingkan performa dan menemukan model yang lebih optimal. Integrasi dengan sistem keuangan sekolah secara langsung akan menghasilkan solusi yang lebih komprehensif dan efisien, memungkinkan sinkronisasi otomatis data pembayaran dan laporan keuangan real-time. Selain itu, implementasi mekanisme feedback dan continuous learning dapat membantu model beradaptasi dengan pola pembayaran baru dan meningkatkan akurasi prediksi seiring waktu.

DAFTAR PUSTAKA

Aljabar, A. (2024). Mengungkap Opini Publik : Pendekatan BERT-based- caused untuk Analisis Sentimen pada Komentar Film. *Journal of System and Computer Engineering (JSCE)*, *5*(1), 36–43.

Baek, J., et al. (2019). What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. International Conference on Computer Vision (ICCV), 4716-4726

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media

Briggs, A., Morrison, M., & Coleman, M. (2012). Research methods in educational leadership and management.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. Diakses dari

https://arxiv.org/pdf/1810.04805.pdf

Friedl, J. E. (2006). Mastering Regular Expressions (3rd ed.). O'Reilly Media

Gopalakrishnan, K. (2021). *Available online www.jsaer.com Automated Document Classification using BERT in Banking Industry*. *8*(12), 224–226.

Hanum, A. R., Zetha, I. A., Fajrina, J. N., Wulandari, R. A., Putri, C., Andina, S. P., Yudistira, N., & Brawijaya, U. (2024). *MENDETEKSI BERITA HOAKS PERFORMANCE ANALYSIS OF THE BERT TEXT CLASSIFICATION ALGORITHM*. *11*(3), 537–546. https://doi.org/10.25126/jtiik2024118093

Hillebrand, L., Deußer, T., Dilmaghani, T., & Kliem, B. (2022). *KPI-BERT : A Joint Named Entity Recognition and Relation Extraction Model for Financial Reports*.

Irianti, A., Halimah, Sutedi, & Agariana, M. (2025). Integration of BERT and SVM in Sentiment Analysis of Twitter/X Regarding Constitutional Court Decision No. 60/PUU-XXII/2024. *Jurnal Teknik Inforatika (JUTIF)*, *6*, 469–482. https://doi.org/10.52436/1.jutif.2025.6.2.4068

Kaesmetan, Y. R., & Kalengkongan, W. W. (2025). Klasifikasi Sentimen Publik Terkait Stunting Di Indonesia Menggunakan BERT Dan SVM Classification of Public Sentiment Related to Stunting in Indonesia Using BERT and SVM. *Jurnal of Business and Audit Information System (JBASE)*, *8*(2), 11–23. dx.doi.org/10.23965/jbase.v8i2.8960

Mahadevkar, S. V, Patil, S., Kotecha, K., Soong, L. W., & Choudhury, T. (2025). Exploring AI - driven approaches for unstructured document analysis and future horizons. *Journal of Big Data*, *2024*. https://doi.org/10.1186/s40537-024-00948-z

Morrison, J. (2012). Tutorial on logistic-regression calibration and fusion. arXiv preprint arXiv:1211.4104. Diakses dari https://arxiv.org/pdf/1211.4104.pdf

Ovirianti, M., et al. (2022). Support Vector Machine Using A Classification Algorithm. Jurnal

Polgan, 8(2), 1-10. Diakses dari https://jurnal.polgan.ac.id/index.php/sinkron/article/view/11597

Przybyła-Kasperek, M., et al. (2024). A multi-layer perceptron neural network for varied conditional classification problems. PLOS ONE, 19(12), e0316186. Diakses dari https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0316186

Sari, B., Sembiring, B., Pandia, M., Sembiring, H., & Margaretta, D. (2023). *Naïve Bayes Classifier and Decision Tree Algorithms for Classifying Payment Data*. *4*(1), 592–600. https://doi.org/10.30865/klik.v4i1.963

Sembiring, R. W., et al. (2023). Naïve Bayes Classifier and Decision Tree Algorithms for Classification Tasks. KLIK: Jurnal Teknologi Informasi, 2(3), 45-58. Diakses dari https://djournals.com/index.php/klik/article/view/963

Simanjorang. (2022). Strategi Pemulihan Umkm Pada Masa New Normal Dan Industri 4.0 Di Desa Pulau Gambar. *Jurnal Pengabdian Kepada Masyarakat Nusantara (JPkMN)*, *2*(2), 97–103.

Skalka, J., Przybyła-kasperek, M., & Dagien, V. (2025). *Cross-national survey data on student attitudes toward artificial intelligence*. *62*. https://doi.org/10.1016/j.dib.2025.112022

Smith, R. (2007). An Overview of the Tesseract OCR Engine. International Conference on Document Analysis and Recognition, 629-633

Wilie, B., et al. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. Proceedings of the 2020 Empirical Methods in Natural Language Processing, 6987-7007.

Zhao, X., Niu, E., Wu, Z., & Wang, X. (n.d.). *Information Extractor*.