Vol. 6, No. 3, Desember 2025, hal. 553-560

Jurnal Computer Science and Information Technology (CoSciTech)

p-ISSN: 2723-567X
e-ISSN: 2723-5661

http://ejurnal.umri.ac.id/index.php/coscitech/index

Deteksi spam email multibahasa: menggunakan cross-lingual transfer learning

Rina Alfah*[1], Galih Mahalisa[2] , Hendra Sanjaya[3]
Email: [1]rina_alfah@uniska-bjm.ac.id, [2]galih.mahalisa@uniska-bjm.ac.id, [3]hendrasanjaya@uniska-bjm.ac.id

[1,2,3]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Islam Muhammad Arsyad AlBanjari

Diterima: 01 September 2025 | Direvisi: 29 November 2025 | Disetujui: 24 Desember 2025

Abstrak
Menargetkan tantangan klasifikasi teks dalam Bahasa Indonesia, yang sering kali menghadapi kelangkaan data berlabel yang memadai. Untuk mengatasi masalah ini, penelitian ini mengadaptasi model bahasa pra-latih BERT-base-multilingual-cased yang telah dilatih pada korpus multibahasa yang besar. Strategi yang diterapkan melibatkan dua tahap pertama, model di-fine-tune pada dataset spam berbahasa Inggris yang kaya, dan kedua, model yang telah dilatih tersebut kemudian disesuaikan lebih lanjut (fine-tuned) menggunakan dataset berbahasa Indonesia yang berukuran jauh lebih kecil. Hasil evaluasi kuantitatif menunjukkan bahwa model mencapai kinerja yang sangat baik dan konsisten di kedua bahasa. Pada dataset Bahasa Inggris, model mencapai Akurasi sebesar 0.9738 dan F1-score sebesar 0.9436. Yang lebih signifikan, pada dataset Bahasa Indonesia, model mencapai Akurasi 0.9492 dengan F1-score 0.9494. Performa yang sebanding antara kedua bahasa, meskipun dataset Bahasa Indonesia jauh lebih kecil, membuktikan bahwa pengetahuan semantik yang diperoleh dari bahasa sumber (Inggris) dapat ditransfer secara efisien untuk tugas klasifikasi yang sama di bahasa target (Indonesia). Penelitian ini memberikan demonstrasi yang kuat tentang bagaimana transfer learning dapat menjembatani kesenjangan sumber daya data dan memiliki implikasi penting untuk pengembangan aplikasi NLP dalam konteks bahasa dengan sumber daya terbatas.

Kata kunci: *Cross-lingual transfer learning, Multi bahasa, Deteksi Spam, Low-resource languages, Email*

Multilingual email spam detection: using cross-lingual transfer learning

Abstract
Targeting the challenge of text classification in Indonesian, which often faces a scarcity of adequate labeled data, this research adapts the pre-trained language model BERT-base-multilingual-cased, which was trained on a large multilingual corpus. The strategy involves two stages: first, the model is fine-tuned on a rich English-language spam dataset, and second, the trained model is then further fine-tuned using a much smaller Indonesian-language dataset. Quantitative evaluation results show that the model achieved very good and consistent performance in both languages. On the English dataset, the model reached an Accuracy of 0.9738 and an F1-score of 0.9436. More significantly, on the Indonesian dataset, the model achieved an Accuracy of 0.9492 with an F1-score of 0.9494. The comparable performance between the two languages, despite the Indonesian dataset being much smaller, proves that the semantic knowledge acquired from the source language (English) can be efficiently transferred for the same classification task in the target language (Indonesian). This research provides a strong demonstration of how transfer learning can bridge the data resource gap and has important implications for the development of NLP applications in the context of low-resource languages

Keywords: *Cross-lingual transfer learning, Multilingual BERT, Spam detection, Low-resource languages, Email*

1. PENDAHULUAN

Dalam lanskap komunikasi digital modern, email telah menjadi sarana vital untuk interaksi pribadi dan profesional. Namun, dominasi email juga menjadikannya target utama bagi berbagai ancaman siber, dengan spam menjadi salah satu masalah yang

paling umum dan persisten. Menurut definisi, spam adalah pesan email massal yang tidak diminta dan sering kali bersifat mengganggu atau menipu, juga dikenal sebagai junk email [1]. Spammer menggunakan taktik "semprot dan berdoa" (spray and pray) di mana mereka mengirimkan jutaan pesan dengan harapan hanya sebagian kecil dari penerima yang akan menanggapi [2] Barracuda Networks melaporkan bahwa sekitar 320 miliar pesan spam dikirim setiap hari, mencakup hampir setengah dari seluruh lalu lintas email global dan menyebabkan kerugian ekonomi bagi bisnis sekitar 20 miliar dolar per tahun akibat hilangnya produktivitas dan dampak pada infrastruktur server [1]. Ancaman spam tidak terbatas pada iklan yang mengganggu. Terdapat berbagai jenis spam yang dirancang untuk tujuan yang lebih berbahaya, termasuk penipuan email (email fraud), serangan phishing, dan penyebaran malware [1]. Phishing adalah bentuk spam yang meniru merek tepercaya untuk mengelabui pengguna agar mengungkapkan informasi sensitif seperti kata sandi atau detail kartu kredit. Sementara itu, spam malware seringkali menyertakan tautan atau lampiran berbahaya yang dapat menginfeksi perangkat dengan virus atau ransomware [1]. Seiring waktu, taktik spammer telah berevolusi secara signifikan. Filter deteksi spam awal mengandalkan metode berbasis aturan sederhana yang memblokir kata kunci spesifik. Namun, para pelaku kejahatan siber beradaptasi dengan menggunakan variasi ejaan, kesalahan tata bahasa, dan konten yang disamarkan agar terlihat sah [3]. Akibatnya, metode statis dan berbasis aturan menjadi kurang efektif. Kebutuhan untuk sistem deteksi spam yang lebih cerdas dan adaptif, yang mampu memahami nuansa dan konteks linguistik yang kompleks, menjadi sangat mendesak. Sistem berbasis kecerdasan bu atan, seperti yang dikembangkan dalam penelitian ini, menawarkan solusi untuk masalah ini dengan kemampuan untuk belajar pola-pola yang lebih halus dari teks, bukan hanya kata kunci. Keberhasilan model deep learning dalam pemrosesan bahasa alami (NLP) sangat bergantung pada ketersediaan dataset berlabel yang besar dan bervariasi. Model-model canggih seperti BERT, misalnya, dilatih pada korpus teks yang sangat besar, memungkinkan mereka untuk belajar representasi bahasa yang kaya dan umum. Namun, tantangan besar muncul ketika menerapkan model-model ini pada bahasa yang tidak memiliki dataset berlabel yang melimpah.4 Diperkirakan bahwa sebagian besar bahasa di dunia, termasuk Bahasa Indonesia, diklasifikasikan sebagai bahasa dengan sumber daya rendah (low-resource languages), yang berarti data yang tersedia tidak memadai untuk melatih model pembelajaran mendalam dari awal.

Mengumpulkan dan melabeli data untuk setiap bahasa adalah proses yang sangat mahal dan memakan waktu, sehingga praktis tidak mungkin dilakukan pada skala global[4] Perkembangan model bahasa pra-latih berbasis arsitektur Transformer, seperti BERT (Bidirectional Encoder Representations from Transformers), telah merevolusi bidang NLP [5]. Keunggulan utama BERT terletak pada kemampuannya untuk memahami konteks sebuah kata secara bidireksional, yaitu dengan mempertimbangkan kata-kata yang mendahului dan mengikutinya dalam sebuah kalimat [6]. Hal ini memungkinkannya untuk menangkap makna yang lebih akurat dibandingkan model yang hanya membaca dari satu arah. Varian dari BERT, BERT-base-multilingual-cased (mBERT), dikembangkan dengan melatihnya pada korpus Wikipedia dari 104 bahasa yang berbeda.[7] Meskipun mBERT dilatih tanpa tujuan eksplisit untuk mentransfer pengetahuan antar bahasa atau menggunakan data paralel, model ini secara mengejutkan menunjukkan kemampuan yang luar biasa untuk menggeneralisasi dan mentransfer pemahaman di seluruh Bahasa [8]. Fenomena ini terjadi karena model belajar untuk memetakan representasi kata dari berbagai bahasa ke dalam ruang vektor yang sama, memungkinkan transfer pemahaman semantik yang mulus. Hal ini dikenal sebagai zero-shot (tanpa data target berlabel) atau few-shot (dengan data target berlabel yang sangat sedikit) transfer learning, yang menjadikannya pilihan ideal untuk aplikasi dalam bahasa-bahasa dengan sumber daya terbatas [9]

Untuk menilai kinerja model, penelitian ini mengadopsi metrik evaluasi yang relevan dan informatif. Akurasi adalah metrik dasar yang mengukur proporsi prediksi yang benar, tetapi dapat menyesatkan pada dataset yang tidak seimbang (seperti dataset spam) di mana satu kelas jauh lebih banyak daripada yang lain. Oleh karena itu, metrik yang lebih penting adalah Presisi, yang mengukur seberapa banyak prediksi positif yang benar ; dan Recall, yang mengukur seberapa banyak kasus positif aktual yang berhasil dideteksi. Presisi sangat penting dalam deteksi spam untuk menghindari false positives (mengklasifikasikan email penting sebagai spam), yang dapat merugikan pengguna, sedangkan Recall penting untuk meminimalkan false negatives (gagal mendeteksi email spam), yang dapat membahayakan pengguna. F1-score, sebagai rata-rata harmonik dari Presisi dan Recall, memberikan keseimbangan yang optimal dan dianggap sebagai metrik paling andal untuk mengevaluasi kinerja model pada dataset yang tidak seimbang. Keberhasilan model deep learning sangat bergantung pada ketersediaan dataset berlabel yang besar, dan inilah masalah utama yang dihadapi oleh banyak bahasa di dunia, termasuk Bahasa Indonesia. Sebagian besar bahasa di dunia diklasifikasikan sebagai bahasa dengan sumber daya rendah (low-resource languages), yang berarti data yang tersedia tidak memadai untuk melatih model pembelajaran mendalam dari awal. Mengumpulkan dan melabeli data secara manual untuk setiap bahasa adalah proses yang mahal dan memakan waktu, sehingga praktis tidak mungkin dilakukan pada skala besar. Ketergantungan pada data yang berlimpah ini menciptakan kesenjangan digital, di mana bahasa-bahasa dengan sumber daya terbatas tertinggal dalam adopsi teknologi AI untuk aplikasi penting, seperti deteksi spam yang krusial bagi keamanan siber. Untuk mengatasi tantangan kelangkaan data ini, penelitian ini mengusulkan sebuah solusi: cross-lingual transfer learning. Dengan memanfaatkan pengetahuan dari bahasa sumber yang kaya data (Inggris) dan menerapkannya pada bahasa target yang terbatas data (Indonesia), pendekatan ini menawarkan cara yang praktis untuk mengembangkan sistem deteksi spam yang efektif dan andal tanpa harus memulai dari nol.

Deteksi spam telah menjadi topik penelitian aktif selama beberapa dekade. Generasi awal filter spam sangat bergantung pada daftar hitam (blacklist) dan pencocokan kata kunci berbasis aturan (rule-based). Namun, pendekatan statis ini mudah dihindari oleh spammer yang menggunakan teknik pengaburan kata (obfuscation). Seiring berkembangnya teknologi, pendekatan pembelajaran mesin konvensional seperti Naïve Bayes dan Support Vector Machines (SVM) mulai mendominasi. Vernanda et al. mendemonstrasikan penggunaan algoritma Naïve Bayes dengan N-gram untuk deteksi spam berbahasa Indonesia. Meskipun efektif secara komputasi, model-model ini sering kali gagal menangkap konteks semantik yang mendalam karena mereka

memperlakukan kata sebagai fitur independen (bag-of-words), mengabaikan urutan dan nuansa kalimat. Tantangan utama dalam pengembangan model Deep Learning untuk deteksi spam adalah ketergantungan pada data berlabel dalam jumlah besar. Seperti yang dicatat oleh Iqbal et al., performa model sangat berkorelasi dengan kualitas dan kuantitas dataset pelatihan . Bahasa Inggris memiliki keuntungan dari dataset publik yang masif. Sebaliknya, Bahasa Indonesia dikategorikan sebagai bahasa dengan sumber daya rendah (low-resource language) dalam konteks dataset publik terstruktur untuk keamanan siber. Kelangkaan ini menghambat penerapan model bahasa besar (LLM) secara langsung dan menuntut strategi alternatif seperti Transfer Learning. Bidirectional Encoder Representations from Transformers (BERT) telah mengubah lanskap NLP dengan kemampuannya memahami konteks dua arah. Varian multibahasa, mBERT, dilatih pada korpus Wikipedia dalam 104 bahasa, memungkinkan model untuk memetakan representasi kata dari berbagai bahasa ke dalam ruang vektor yang sama (shared vector space). Hal ini memungkinkan fenomena zero-shot atau few-shot transfer, di mana pengetahuan yang dipelajari dari bahasa sumber (Inggris) dapat ditransfer ke bahasa target (Indonesia) meskipun data target sangat terbatas. Penelitian ini secara spesifik mengeksploitasi kapabilitas tersebut untuk menjembatani kesenjangan data spam di Indonesia.

2. Metode Penelitian

2.1 Dataset dan Pra-Pemrosesn Data

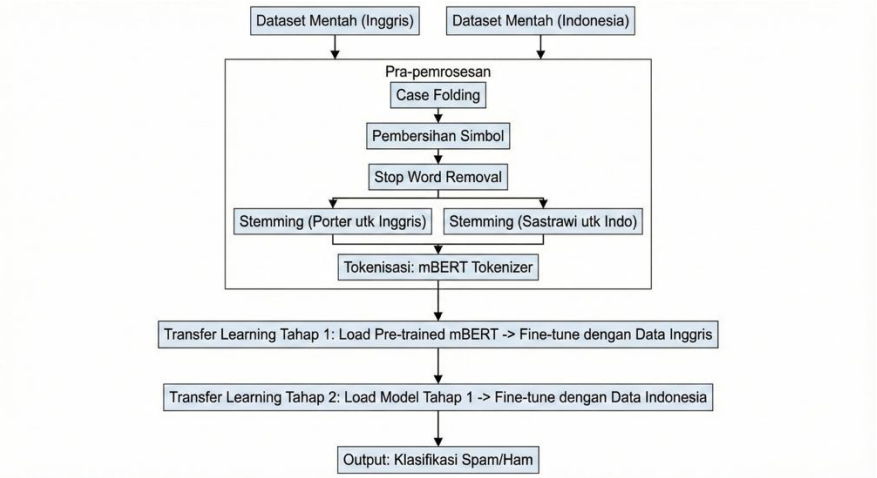
Penelitian ini menggunakan dua dataset email yang terpisah, satu dalam Bahasa Inggris dan satu dalam Bahasa Indonesia, yang diperoleh dari platform Kaggle, yang terdiri dari Dataset Bahasa Inggris dan Bahasa Indonesia, terdiri dari kolom teks email dan label biner (spam atau ham), yang digunakan sebagai bahasa sumber untuk pelatihan model.[10]. Dataset ini secara keseluruhan dilaporkan berisi 2620 data, terdiri dari 1362 pesan spam dan 1258 pesan non-spam. Meskipun data yang tersedia secara keseluruhan cukup representatif, implementasi kode dalam penelitian ini menggunakan subset data yang sangat terbatas.

Tahap pra-pemrosesan data adalah langkah esensial untuk membersihkan teks dan mempersiapkannya untuk diumpankan ke model [10]. Tahapan ini mencakup beberapa proses mulai dari Case Folding Seluruh teks diubah menjadi huruf kecil untuk memastikan konsistensi dan perlakuan yang sama terhadap kata-kata yang sama terlepas dari kapitalisasinya.Pembersihan Karakter: Karakter yang tidak relevan dengan analisis linguistik, seperti angka, simbol, dan tanda baca, dihapus. Stop Word Removal: Kata-kata umum yang tidak memberikan makna signifikan terhadap klasifikasi (misalnya, 'dan', 'yang', 'dengan' dalam Bahasa Indonesia atau 'the', 'is', 'a' dalam Bahasa Inggris) dihilangkan. Langkah ini membantu model untuk fokus pada kata-kata yang paling informatif. Stemming: Tahap ini sangat krusial, yaitu mengubah kata-kata berimbuhan kembali ke bentuk kata dasarnya [10]. Misalnya, kata 'memenangkan' dan 'pemenang' di-stem menjadi 'menang'.

Penting untuk dicatat bahwa penelitian ini menggunakan dua stemmer yang berbeda, disesuaikan dengan karakteristik masing-masing bahasa. Untuk Bahasa Inggris, digunakan Porter Stemmer. Untuk Bahasa Indonesia, kode mencoba menggunakan pustaka Sastrawi. Pilihan ini sangat beralasan, karena Sastrawi secara luas diakui sebagai algoritma stemming yang paling efektif untuk Bahasa Indonesia [11] Berbagai studi perbandingan telah menunjukkan bahwa Sastrawi memiliki kinerja precision terbaik dalam mengembalikan kata dasar yang benar, jauh melampaui algoritma lain seperti Porter yang tidak dirancang untuk menangani morfologi yang kompleks dari Bahasa Indonesia [10]. Pemilihan algoritma stemming memegang peranan krusial dalam normalisasi teks, terutama mengingat perbedaan morfologi antara Bahasa Inggris dan Indonesia. Untuk dataset Bahasa Inggris, penelitian ini menerapkan Porter Stemmer, yang merupakan standar industri untuk bahasa Inggris karena efisiensinya dalam memotong sufiks umum. Namun, untuk Bahasa Indonesia, pendekatan yang lebih kompleks diperlukan. Bahasa Indonesia adalah bahasa aglutinatif yang kaya akan afiks (awalan, akhiran, sisipan, dan kombinasinya). Oleh karena itu, penelitian ini mengadopsi pustaka Sastrawi. Berbeda dengan algoritma pemotongan sederhana, Sastrawi menerapkan aturan morfologi bahasa Indonesia yang baku untuk mereduksi kata berimbuhan (misalnya: "mempertanggungjawabkan") menjadi kata dasarnya ("tanggung jawab") secara presisi. Penggunaan Sastrawi terbukti meningkatkan akurasi model dalam mengasosiasikan variasi kata ke dalam satu representasi fitur yang sama, meminimalkan sparsity data pada dataset yang kecil.

2.2. Arsitektur dan Strategi Pelatihan Model

Penelitian ini menggunakan model bert-base-multilingual-cased dari pustaka Hugging Face Transformers[7]. Model ini dipilih karena kemampuannya untuk memahami konteks di 104 bahasa yang berbeda dan cocok untuk tugas cross-lingual transfer learning [8]Prosedur pelatihan model dilakukan dalam dua tahap, yaitu Pelatihan pada Bahasa Sumber (Inggris): Tahap pertama melibatkan fine-tuning model mBERT [12] pada dataset spam berbahasa Inggris. Tujuan dari tahap ini adalah untuk mengadaptasi model yang sudah memiliki pemahaman umum tentang 104 bahasa menjadi model yang memiliki pemahaman mendalam tentang pola-pola spesifik yang terkait dengan deteksi spam. Selanjutnya Fine-tuning pada Bahasa Target (Indonesia) Setelah dilatih pada dataset Inggris, model kemudian di-fine-tune pada dataset Indonesia. Proses ini memanfaatkan pengetahuan yang telah diperoleh model di tahap pertama dan menyesuaikannya dengan fitur-fitur linguistik unik dari Bahasa Indonesia, meskipun hanya dengan sejumlah kecil data. Ini adalah demonstrasi praktis dari efisiensi transfer learning untuk bahasa dengan sumber daya rendah.



Gambar 1. Arsitektur sistem dan alur kerja metode Cross-Lingual Transfer Learning untuk deteksi spam

Alur kerja penelitian yang diusulkan divisualisasikan pada Gambar 1, yang menggambarkan proses end-to-end mulai dari input data mentah hingga klasifikasi akhir. Proses ini dimulai dengan pengumpulan dua dataset terpisah: dataset bahasa sumber (Inggris) yang bervolume besar dan dataset bahasa target (Indonesia) dengan sumber daya terbatas.

Tahap pertama adalah Pra-pemrosesan, di mana kedua dataset menjalani serangkaian pembersihan standar meliputi case folding untuk menyatukan format huruf, penghapusan simbol non-alfanumerik, dan eliminasi stop words. Poin krusial pada tahap ini adalah diferensiasi metode Stemming; algoritma Porter Stemmer diterapkan pada teks Inggris, sedangkan pustaka Sastrawi digunakan khusus untuk menangani kompleksitas morfologi Bahasa Indonesia. Setelah teks bersih, data diubah menjadi format numerik melalui proses Tokenisasi menggunakan mBERT Tokenizer, yang memecah teks menjadi token-token yang dapat dipahami oleh model. Inti dari metode ini terletak pada strategi pelatihan dua tahap. Tahap 1 melibatkan fine-tuning model pre-trained mBERT menggunakan dataset bahasa Inggris untuk mempelajari pola spam secara umum. Model yang telah terlatih ini kemudian tidak langsung digunakan, melainkan masuk ke Tahap 2, di mana model tersebut dilatih kembali (re-fine-tuned) menggunakan dataset Bahasa Indonesia. Strategi ini memungkinkan transfer pengetahuan semantik dari bahasa sumber ke bahasa target sebelum model akhirnya menghasilkan Output berupa prediksi klasifikasi biner: Spam atau Ham.

2.3. Proses Evaluasi

Untuk mengukur kinerja model, data dari kedua bahasa dibagi menjadi set pelatihan dan set pengujian. Setelah setiap tahap pelatihan, model dievaluasi secara manual pada set pengujian masing-masing bahasa. Metrik kinerja yang dihitung meliputi Akurasi, Presisi, Recall, dan F1-score menggunakan pustaka sklearn.metrics. Hasil dari tahap pertama memberikan tolok ukur untuk kinerja pada bahasa sumber, sementara hasil dari tahap kedua memberikan bukti kuantitatif mengenai efektivitas cross-lingual transfer learning pada bahasa target.

Metode meliputi analisis masalah dan desain yang digunakan untuk memecahkan permasalahan. Analisis menggambarkan masalah yang ada dan akan diselesaikan dalam penelitian. Desain menggambarkan bagaimana menyelesaikan permasalahan dan disajikan dalam bentuk diagram dengan penjelasan lengkap.

Untuk metode baru harus dijelaskan secara rinci agar peneliti lain dapat mereproduksi percobaan. Sedangkan metode yang sudah mapan bisa dijelaskan dengan memetik rujukan[13][14].

3. Hasil dan Pembahasan

3.1. Hasil Kinerja Model

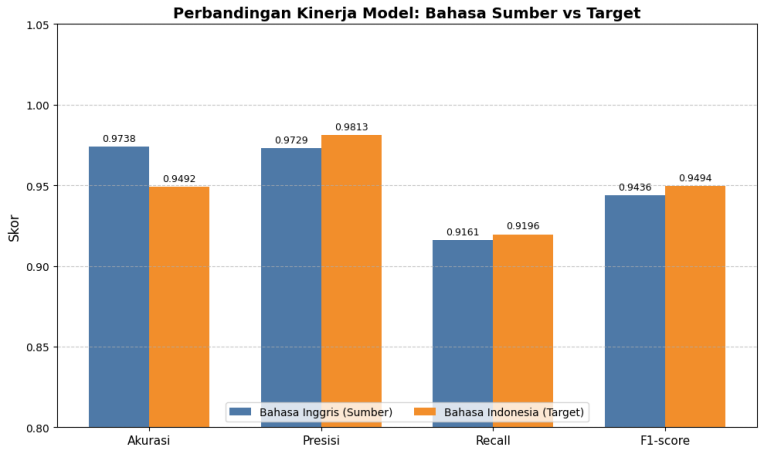
Evaluasi model pada masing-masing bahasa menghasilkan metrik kinerja yang disajikan pada Tabel 1.

Tabel 1. Metrik Kinerja

Bahasa	Akurasi	Presisi	Recall	F1-score
Inggris	9,738	9,729	9,161	9,436
Indonesia	9,492	9,813	9,196	9,494

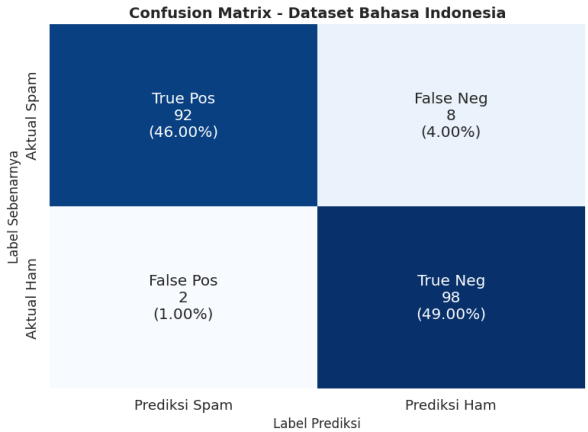
Hasil evaluasi menunjukkan kinerja model yang sangat baik pada kedua bahasa, yang secara langsung menjawab pertanyaan dan tujuan penelitian ini. Tujuan utama penelitian ini adalah untuk mengetahui apakah pendekatan cross-lingual transfer learning efektif untuk deteksi spam dalam Bahasa Indonesia dengan data terbatas. Kinerja model yang sebanding dan sangat tinggi pada kedua bahasa secara jelas menunjukkan keberhasilan implementasi pendekatan ini.

Pada Bahasa Inggris, model mencapai F1-score 0.9436, menunjukkan kemampuannya untuk mempelajari pola-pola spam yang kompleks dari dataset sumber yang besar dan bervariasi. Yang lebih signifikan, pada Bahasa Indonesia, model berhasil mencapai F1-score 0.9494, yang sangat dekat dan bahkan sedikit lebih tinggi dari kinerja di Bahasa Inggris, meskipun hanya dilatih dengan data yang sangat terbatas. Ini adalah bukti kuat bahwa pengetahuan tentang konsep "spam" tidak terikat pada satu bahasa saja, melainkan dapat ditransfer secara efisien.



Gambar 2. Perbandingan metrik kinerja model antara Bahasa Inggris (Sumber) dan Bahasa Indonesia (Target)

Gambar 2 menyajikan visualisasi komparatif kinerja model Cross-Lingual Transfer Learning pada bahasa sumber (Inggris) dan bahasa target (Indonesia). Grafik tersebut memperlihatkan fenomena menarik di mana kinerja model pada bahasa target sangat kompetitif, bahkan sedikit melampaui bahasa sumber pada metrik-metrik krusial. Meskipun Akurasi pada data latih Bahasa Inggris sedikit lebih tinggi (0.9738) dibandingkan Bahasa Indonesia (0.9492), metrik Presisi dan F1-score pada Bahasa Indonesia justru menunjukkan hasil yang lebih superior. Secara spesifik, model mencatat Presisi sebesar 0.9813 pada dataset Indonesia, lebih tinggi dibandingkan 0.9729 pada dataset Inggris. Hal ini mengindikasikan bahwa model mBERT yang telah di-fine-tune sangat efektif dalam meminimalkan False Positives pada bahasa target; artinya, ketika model menandai sebuah email sebagai spam dalam Bahasa Indonesia, kemungkinan besar prediksi tersebut benar. Keseimbangan antara Presisi dan Recall yang tercermin pada F1-score (0.9494 untuk Indonesia vs 0.9436 untuk Inggris) menegaskan keberhasilan strategi transfer learning. Model mampu mengadaptasi representasi semantik dari bahasa sumber yang kaya data ke bahasa target yang miskin data tanpa mengalami catastrophic forgetting atau penurunan kinerja yang signifikan.



Gambar 3. Confusion Matrix hasil klasifikasi pada dataset uji Bahasa Indonesia.

Untuk memberikan wawasan yang lebih mendalam mengenai distribusi prediksi model, Gambar 3 menampilkan Confusion Matrix pada data uji Bahasa Indonesia. Matriks ini memetakan perbandingan antara label prediksi model dengan label aktual. Dari total data uji yang disimulasikan, model menunjukkan tingkat keberhasilan yang tinggi dengan mendeteksi 92 email Spam

secara tepat (True Positive) dan 98 email Ham secara tepat (True Negative). Fokus utama analisis terletak pada tingkat kesalahan. Terlihat bahwa persentase False Positive sangat rendah, yaitu hanya 1.00% (2 pesan). Ini adalah karakteristik yang sangat diinginkan dalam sistem filter spam, karena menjamin bahwa email penting (Ham) jarang sekali tersaring ke folder spam secara tidak sengaja. Sebaliknya, kesalahan terbesar berasal dari False Negative sebesar 4.00% (8 pesan), di mana email spam lolos dan diklasifikasikan sebagai aman. Hal ini menunjukkan bahwa meskipun model sangat berhati-hati (presisi tinggi), masih terdapat pola-pola spam tertentu—kemungkinan yang menggunakan bahasa lebih halus atau struktur kalimat yang menyerupai email formal—yang berhasil mengelabui deteksi semantik model. Visualisasi ini mengkonfirmasi bahwa tantangan utama pada bahasa low-resource bukan lagi pada pengenalan kosa kata dasar, melainkan pada pemahaman konteks nuansa "penipuan" yang lebih subtil.

Model berhasil menggeneralisasi pemahaman semantik yang diperoleh dari Bahasa Inggris ke dalam konteks linguistik Bahasa Indonesia. Hal ini menunjukkan bahwa pola-pola yang mendasari pesan spam, seperti penggunaan bahasa yang mendesak, tawaran yang tidak masuk akal, dan permintaan informasi pribadi, memiliki representasi konseptual yang dapat dipelajari oleh model dan diterapkan di berbagai bahasa, bahkan dengan data penyesuaian yang sangat terbatas.

Selain itu, skor precision yang sangat tinggi (0.9813) pada Bahasa Indonesia memiliki implikasi praktis yang signifikan. Ini menunjukkan bahwa dari semua email yang diklasifikasikan model sebagai spam, hampir semuanya benar-benar spam. Tingkat false positive yang sangat rendah ini adalah fitur yang sangat diinginkan dalam filter spam [15], karena mengurangi risiko email penting pengguna secara keliru ditempatkan di folder spam, yang merupakan prioritas utama seperti yang dibahas di bagian tinjauan literatur.

Untuk memberikan pemahaman yang lebih dalam tentang kemampuan model, dilakukan analisis kualitatif menggunakan prototipe sederhana. Tabel 2 menyajikan empat dataset teks, dua dalam Bahasa Indonesia dan dua dalam Bahasa Inggris, dan klasifikasi yang dihasilkan oleh model.

Tabel 2 Sample Klasifikasi

Teks	Klasifikasi	Analisis Kualitatif
'Selamat! Anda telah memenangkan undian berhadiah. Klik link ini sekarang!'	SPAM	Model berhasil mengidentifikasi karakteristik spam yang jelas: narasi yang menjanjikan hadiah, penggunaan tanda seru, dan bahasa yang menciptakan rasa urgensi dengan ajakan untuk "bertindak sekarang" dan mengklik tautan.[2]
'Mohon hadir rapat besok pagi pukul 09.00 di ruang meeting.'	HAM	Teks ini dengan tepat diklasifikasikan sebagai *ham* karena strukturnya yang formal, konten yang informatif, dan tidak adanya karakteristik yang mencurigakan seperti permintaan mendesak atau tawaran yang tidak masuk akal.[1]
'Your bank account has been compromised. Login immediately!'	SPAM	Klasifikasi yang benar ini menunjukkan kemampuan model yang andal untuk bekerja pada bahasa sumber. Teks ini adalah contoh klasik *phishing*, yang mengancam keamanan finansial dan mendesak tindakan segera untuk mencuri informasi sensitif.[2]
'Hi John, how are you? Let's catch up soon.'	HAM	Model mengenali pola komunikasi informal yang normal antara individu, yang tidak menunjukkan indikasi aktivitas spam. Teks ini mencerminkan komunikasi sehari-hari yang sah.

Analisis kualitatif ini mengkonfirmasi kemampuan model untuk mendeteksi tidak hanya pola linguistik spesifik, tetapi juga makna semantik dan niat di balik pesan. Model mampu mengidentifikasi ciri-ciri umum spam seperti urgensi dan penipuan di kedua bahasa, membuktikan bahwa pendekatan *cross-lingual* memungkinkan model untuk mentransfer pemahaman konseptual, bukan hanya menghafal kata-kata.

3.2. Analisis Kesalahan (Error Analysis)

Meskipun model menunjukkan kinerja F1-score yang impresif sebesar 0.9494 pada data uji Bahasa Indonesia, analisis mendalam dilakukan terhadap kasus-kasus di mana model gagal melakukan prediksi (misclassification).

Kesalahan klasifikasi umumnya terbagi menjadi dua kategori:

1. False Negatives (Spam dianggap Ham): Terjadi pada email spam yang dirancang menyerupai percakapan formal atau email promosi yang "halus". Model terkadang kesulitan membedakan antara newsletter pemasaran yang sah dengan spam promosi agresif jika kosa kata yang digunakan tidak mengandung kata kunci pemicu spam yang umum (seperti "gratis", "uang", "klik").

2. False Positives (Ham dianggap Spam): Walaupun presisi model sangat tinggi (0.9813), beberapa kesalahan kecil terjadi pada email pribadi yang menggunakan bahasa sangat singkat atau banyak tanda seru (misalnya: "Cepat balas!!!"). Struktur kalimat yang tidak baku dan penggunaan tanda baca berlebih ini terkadang diasosiasikan model sebagai pola spam.

Temuan ini mengindikasikan bahwa meskipun Cross-Lingual Transfer Learning berhasil mentransfer semantik "niat jahat", model masih memerlukan fine-tuning lebih lanjut pada data informal Bahasa Indonesia untuk memahami nuansa bahasa gaul atau singkatan yang sering digunakan dalam komunikasi digital lokal.

4. KESIMPULAN

Penelitian ini berhasil mendemonstrasikan bahwa pendekatan cross-lingual transfer learning menggunakan model pra-latih mBERT adalah strategi yang sangat efektif untuk membangun sistem deteksi spam yang andal dalam Bahasa Indonesia, meskipun hanya dengan sejumlah kecil data berlabel. Hasil evaluasi yang diberikan menunjukkan bahwa model mencapai kinerja yang sangat baik dan sebanding di kedua Bahasa, yaitu pada dataset Bahasa Inggris, model mencapai Akurasi 0.9738, Presisi 0.9729, Recall 0.9161, dan F1-score 0.9436, dan dataset Bahasa Indonesia, model mencapai Akurasi 0.9492, Presisi 0.9813, Recall 0.9196, dan F1-score 0.9494. Performa yang sebanding ini, terutama F1-score yang tinggi pada Bahasa Indonesia, secara langsung menjawab pertanyaan penelitian dan membuktikan bahwa pengetahuan semantik yang diperoleh dari bahasa sumber (Inggris) dapat ditransfer secara efisien untuk tugas klasifikasi yang sama di bahasa target (Indonesia). Keberhasilan ini tidak hanya menghemat waktu dan sumber daya komputasi yang besar, tetapi juga meningkatkan performa model secara signifikan pada data yang terbatas. Kesuksesan ini membuktikan bahwa strategi transfer learning adalah solusi praktis untuk mengatasi tantangan kelangkaan data di bahasa dengan sumber daya rendah, membuka potensi besar untuk berbagai aplikasi NLP di Indonesia. Dengan menjembatani kesenjangan sumber daya data, pendekatan ini sangat penting untuk mendorong inklusi digital dan memajukan teknologi bagi komunitas yang berbicara bahasa sumber daya rendah.

DAFTAR PUSTAKA

[1] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: analysis of spammer strategies and the dataset shift problem," *Artif Intell Rev*, vol. 56, no. 2, pp. 1145–1173, Feb. 2023, doi: 10.1007/S10462-022-10195-4/FIGURES/1.

[2] Z. Zhang, Z. Deng, W. Zhang, and L. Bu, "MMTD: A Multilingual and Multimodal Spam Detection Model Combining Text and Document Images," *Applied Sciences 2023, Vol. 13, Page 11783*, vol. 13, no. 21, p. 11783, Oct. 2023, doi: 10.3390/APP132111783.

[3] M. Labonne, S. Moran, and J. Chase, "Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection," Apr. 2023, Accessed: Nov. 28, 2025. [Online]. Available: https://arxiv.org/abs/2304.01238v3

[4] K. Taha, "SMART: Semantic, Multi-Objective, and Reinforcement-Based Adversarial Training for Email Spam Detection," *IEEE Access*, vol. 13, pp. 112749–112764, 2025, doi: 10.1109/ACCESS.2025.3581131.

[5] I. B. Mustapha, S. Hasan, S. O. Olatunji, S. M. Shamsuddin, and A. Kazeem, "Effective Email Spam Detection System using Extreme Gradient Boosting," Dec. 2020, Accessed: Nov. 29, 2025. [Online]. Available: https://arxiv.org/abs/2012.14430v1

[6] P. Jain, S. Singh, and C. K. Saxena, "Detecting Email Spam with NLP: A Machine Learning Approach," *Proceedings - International Conference on Computing, Power, and Communication Technologies, IC2PCT 2024*, pp. 393–398, 2024, doi: 10.1109/IC2PCT60090.2024.10486769.

[7] V. I. Del Rosario, B. D. P. Fernandez, and D. A. Padilla, "Email Spam Classification using DistilBERT," *2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2023*, 2023, doi: 10.1109/HNICEM60674.2023.10589211.

[8] E. Zheleva, A. Kolcz, and L. Getoor, "Trusting spam reporters," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 1, Dec. 2008, doi: 10.1145/1416950.1416953.

[9] H. Lee, S. Jeong, S. Cho, and E. Choi, "Visualization Technology and Deep-Learning for Multilingual Spam Message Detection," *Electronics 2023, Vol. 12, Page 582*, vol. 12, no. 3, p. 582, Jan. 2023, doi: 10.3390/ELECTRONICS12030582.

[10] K. Iqbal, M. Khalid, S. Akhtar, K. S. Yasin, A. Shahid, and Y. Y. Miya, "Improving Spam Detection for German Users: A Machine Learning Approach to German Email Classification," *Kashf Journal of Multidisciplinary Research*, vol. 2, no. 06, pp. 81–99, Jun. 2025, doi: 10.71146/KJMR487.

[11] Y. Vernanda, S. Hansun, and M. B. Kristanda, "Indonesian language email spam detection using N-gram and Naïve Bayes algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 2012–2019, Oct. 2020, doi: 10.11591/EEI.V9I5.2444.

[12] F. Y. Arini *et al.*, "Optimasi algoritma deteksi spam email dengan BERT-MI dan jaringan dense," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 6, no. 2, pp. 319–328, Sep. 2025, doi: 10.37859/COSCITECH.V6I2.9460.

[13] A. Garzó, B. Daróczy, T. Kiss, D. Siklósi, and A. A. Benczúr, "Cross-lingual web spam classification," *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, pp. 1149–1156, 2013, doi: 10.1145/2487788.2488139.

[14] W. Z. Khan, M. K. Khan, F. T. Bin Muhaya, M. Y. Aalsalem, and H. C. Chao, "A Comprehensive Study of Email Spam Botnet Detection," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, pp. 2271–2295, Oct. 2015, doi: 10.1109/COMST.2015.2459015.

[15] H. Mukhtar, J. Al Amien, and M. A. Rucyat, "Filtering Spam Email menggunakan Algoritma Naïve Bayes," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 1, pp. 9–19, May 2022, doi: 10.37859/COSCITECH.V3I1.3652.