Jurnal Rekayasa Sistem Informasi dan Teknologi
Volume 3, No 3 – Februari 2026
e-ISSN : 3025-888X

PENERAPAN *MULTILINGUAL BERT* UNTUK KLASIFIKASI BAHASA INDONESIA DAN ATAU BAHASA MALAYSIA PADA TEKS PENDEK MEDIA SOSIAL

Moch. Chaidar Chanif[1*], Imam Much Ibnu Subroto [2]
[1,2]Universitas Islam Sultan Agung Semarang
chaidarchanif@std.unissula.ac.id[1]
imam@unissula.ac.id[2]

Received: 12-12-2025	Revised: 14-01-2026	Approved: 31-01-2026

ABSTRAK

Bahasa Indonesia dan Bahasa Malaysia memiliki akar linguistik yang sama sehingga sering menunjukkan kemiripan kosakata dan struktur kalimat, namun juga menyimpan perbedaan makna yang dapat menimbulkan ambiguitas, khususnya pada teks pendek di media sosial. Kondisi ini menjadi tantangan dalam pengklasifikasian bahasa secara otomatis. Penelitian ini bertujuan mengimplementasikan model Multilingual Bidirectional Encoder Representations from Transformers (mBERT) untuk membedakan Bahasa Indonesia dan Bahasa Malaysia pada teks pendek dari platform Twitter. Data dikumpulkan melalui web scraping dengan panjang teks 1–20 kata, menghasilkan total 56.701 data, dengan distribusi Bahasa Indonesia (48,47%), Bahasa Malaysia (14,31%), dan campuran keduanya (37,22%). Proses penelitian mencakup preprocessing (pembersihan teks, case folding, normalisasi, tokenisasi), pembagian data latih dan uji (80:20), fine-tuning mBERT, serta evaluasi menggunakan akurasi, precision, recall, dan F1-score. Hasil pengujian menunjukkan bahwa model mBERT mencapai akurasi 95,8%, precision 97,9%, recall 95,1%, dan F1-score 96,5%, dengan performa stabil pada kedua kelas bahasa. Kesimpulan penelitian ini adalah mBERT efektif dan andal dalam mengklasifikasikan bahasa yang memiliki kemiripan tinggi pada teks pendek media sosial, sehingga berpotensi diterapkan pada pengolahan bahasa alami untuk bahasa-bahasa serumpun lainnya.

***Kata Kunci**: Klasifikasi Bahasa; Bahasa Indonesi; Bahasa Malaysia; mBERT, Teks Pendek, Media Sosial, NLP.*

PENDAHULUAN

Bahasa Indonesia dan Bahasa Malaysia merupakan bahasa serumpun yang memiliki kesamaan leksikal, tata bahasa, dan struktur kalimat, namun sejumlah kosakata menunjukkan perbedaan makna yang signifikan dan berpotensi menimbulkan ambiguitas dalam komunikasi lintas negara. Studi linguistik kontrastif menemukan bahwa meskipun kedua bahasa berasal dari akar yang sama, variasi bentuk dan makna kosakata tetap ada dan membutuhkan pemahaman konteks yang lebih dalam (Husyandi, 2025). Perbedaan semantik tersebut seringkali menimbulkan salah tafsir, terutama dalam konteks komunikasi yang minim penjelasan.

Fenomena tersebut semakin menonjol di media sosial, khususnya Twitter, yang menjadi ruang interaksi publik bagi masyarakat Indonesia dan Malaysia. Karakteristik Twitter yang didominasi oleh teks pendek, tidak baku, dan penuh singkatan menyebabkan konteks percakapan menjadi terbatas, sehingga meningkatkan potensi kesalahpahaman antar pengguna dari kedua negara. Hal ini diperparah oleh tingginya jumlah pengguna Twitter di Indonesia dan Malaysia. Pada tahun 2024, Indonesia tercatat memiliki sekitar 27,1 juta pengguna Twitter, sementara Malaysia memiliki sekitar 5,71 juta pengguna atau sekitar 17% dari total penduduknya (Review, 2025). Tingginya intensitas interaksi ini memperbesar risiko konflik digital akibat perbedaan pemaknaan bahasa.

Permasalahan ini menjadi semakin kompleks karena Bahasa Indonesia dan Bahasa Malaysia masih tergolong sebagai *low-resource languages* dalam bidang *Natural Language Processing* (NLP). Model NLP umum sering mengalami kesulitan dalam membedakan kedua bahasa tersebut, khususnya pada teks informal media sosial yang sarat dengan variasi bahasa dan konteks terbatas (Maxwell-Smith dkk., 2021) Kondisi ini menunjukkan perlunya pengembangan model klasifikasi bahasa yang lebih spesifik dan kontekstual.

Selain faktor linguistik, perkembangan interaksi digital lintas negara di Asia Tenggara turut mendorong meningkatnya kompleksitas komunikasi berbasis teks. Penelitian oleh (Ma dkk., 2021)menegaskan bahwa bahasa-bahasa di Asia Tenggara, termasuk Bahasa Indonesia dan Bahasa Malaysia, masih menghadapi keterbatasan sumber daya NLP, terutama pada data informal media sosial. Kondisi ini menyebabkan model NLP pralatih sering kali kurang optimal dalam memahami variasi bahasa, ejaan tidak baku, dan konteks percakapan yang terbatas, sehingga berdampak langsung pada akurasi tugas klasifikasi bahasa.

Di sisi lain, karakteristik teks media sosial yang bersifat dinamis dan kontekstual menuntut model yang tidak hanya mampu mengenali kata secara leksikal, tetapi juga memahami hubungan makna antar kata dalam konteks kalimat. Studi oleh (Ruder dkk., 2019) menekankan bahwa pendekatan berbasis *pre-trained language models* menjadi fondasi utama dalam pengembangan sistem NLP *modern*, khususnya untuk bahasa dengan keterbatasan data. Model berbasis *transformer* dinilai lebih adaptif dalam menangani teks pendek dan bervariasi, sehingga relevan untuk digunakan dalam klasifikasi bahasa pada platform seperti Twitter.

Penelitian terdahulu menunjukkan bahwa pendekatan berbasis *transformer* lebih unggul dibandingkan metode tradisional dalam klasifikasi teks pendek. Model *Bidirectional Encoder Representations from Transformers* (BERT) mampu menangkap konteks semantik secara lebih mendalam meskipun jumlah kata terbatas (Singh dkk., 2021). Salah satu variannya, *Multilingual* BERT (mBERT), dirancang untuk mendukung lebih dari 100 bahasa, termasuk Bahasa Indonesia dan Bahasa Malaysia. Studi sebelumnya membuktikan bahwa mBERT efektif dalam mendeteksi bahasa Melayu pada media sosial, baik dalam skenario *zero-shot* maupun pada data dengan campuran bahasa (Guo dkk., 2024). Selain itu, penerapan model multibahasa juga terbukti mampu meningkatkan performa klasifikasi teks pendek secara signifikan dibandingkan metode tradisional (Putra & Purwarianti, 2020).

Dalam beberapa tahun terakhir, perhatian penelitian NLP juga semakin tertuju pada klasifikasi bahasa yang memiliki kedekatan linguistik tinggi. Studi oleh (Zhao dkk., 2021) menegaskan bahwa bahasa serumpun cenderung menghasilkan ambiguitas yang lebih besar dalam pemodelan bahasa otomatis dibandingkan bahasa yang tidak berkerabat. Hal ini disebabkan oleh kesamaan kosakata dasar dan pola sintaksis yang sulit dibedakan tanpa pemahaman konteks yang kuat.

Seiring berkembangnya teknologi *deep learning,* model *transformer* multibahasa mulai banyak digunakan untuk mengatasi keterbatasan tersebut. Penelitian oleh (Lu & Li, 2020) menunjukkan bahwa model multibahasa seperti mBERT memiliki kemampuan transfer lintas bahasa yang efektif, terutama pada bahasa dengan sumber daya terbatas. Namun demikian, performa model sangat bergantung pada kesesuaian domain data, sehingga adaptasi terhadap karakteristik media sosial menjadi faktor penting.

Pengembangan lanjutan dari mBERT, seperti XLM-RoBERTa, juga menunjukkan

peningkatan performa yang signifikan dalam berbagai tugas klasifikasi multibahasa. Penelitian oleh (Conneau dkk., 2019) membuktikan bahwa model ini mampu menangani bahasa-bahasa dengan kemiripan linguistik tinggi secara lebih stabil dibandingkan model multibahasa generasi sebelumnya. Temuan ini mengindikasikan bahwa pendekatan transformer multibahasa merupakan solusi yang menjanjikan untuk klasifikasi Bahasa Indonesia dan Bahasa Malaysia.

Lebih lanjut, penelitian terkini menyoroti pentingnya penyesuaian model terhadap domain media sosial. (Nguyen dkk., 2020) menunjukkan bahwa fine-tuning model BERT pada data Twitter secara signifikan meningkatkan akurasi klasifikasi dibandingkan penggunaan model pralatih tanpa adaptasi domain. Hal ini relevan mengingat karakteristik Twitter yang sangat berbeda dari teks formal.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengembangkan sistem klasifikasi otomatis yang mampu membedakan Bahasa Indonesia dan Bahasa Malaysia pada teks pendek Twitter. Diharapkan penelitian ini dapat memberikan kontribusi dalam pengembangan NLP untuk bahasa-bahasa serumpun serta membantu mengurangi potensi kesalahpahaman dalam komunikasi digital.

Pada penelitian (Ansari dkk., 2020), sebelumnya telah dilakukan untuk mengatasi tantangan dalam klasifikasi bahasa yang saling berkerabat. Salah satu penelitian menggunakan pendekatan pembelajaran mesin berbasis karakter n-gram untuk membedakan antara Bahasa Portugis dan Spanyol yang memiliki kemiripan leksikal tinggi. Meskipun pendekatan ini efektif dalam beberapa kasus, hasil penelitian menunjukkan bahwa metode berbasis fitur sederhana memiliki keterbatasan dalam menangkap konteks semantik yang lebih kompleks.

Penelitian (Jauhiainen dkk., 2019), melakukan kajian komprehensif terhadap berbagai pendekatan deteksi bahasa, dari metode statistik hingga *deep learning*, dan menekankan bahwa dalam konteks bahasa yang memiliki kemiripan morfologis dan sintaktis, pemodelan berbasis konteks sangat penting. Dalam kasus Bahasa Indonesia dan Malaysia, pendekatan ini memungkinkan sistem untuk mengenali nuansa makna melalui hubungan antar kata dalam satu kalimat.

Salah satu pendekatan yang populer dalam tugas-tugas klasifikasi bahasa adalah model berbasis transformer seperti BERT karena kemampuannya dalam memahami konteks secara mendalam. Penelitian menunjukkan bahwa *Multilingual BERT* (mBERT) mampu melakukan *zero-shot cross-lingual transfer*, yaitu kemampuan untuk mentransfer pengetahuan antar bahasa tanpa perlu pelatihan ulang, dengan hasil yang kompetitif dalam klasifikasi multibahasa(Pires dkk., 2019). Sementara itu, Pengembangan lebih lanjut dilakukan melalui model XLM-RoBERTa (XLM-R), yang menunjukkan peningkatan akurasi dalam klasifikasi bahasa, terutama pada dataset dengan kemiripan linguistik yang tinggi. Model ini memperkuat pendekatan berbasis *transformer* sebagai solusi yang andal dalam klasifikasi lintas bahasa (Wu & Dredze, 2020).

Dalam penelitian (Singh dkk., 2021), mengeksplorasi kemampuan mBERT dalam mengklasifikasikan bahasa pada data sosial media dan menemukan bahwa akurasi model dapat ditingkatkan dengan teknik *fine-tuning* pada data domain-spesifik. Temuan ini memperkuat gagasan bahwa adaptasi model ke konteks lokal, seperti gaya bahasa di media sosial Indonesia dan Malaysia, merupakan faktor penting dalam keberhasilan klasifikasi.

Penelitian mengenai pemrosesan teks media sosial menunjukkan bahwa

keberadaan teks *code-mixed* dan bahasa informal menjadi tantangan utama dalam tugas klasifikasi bahasa dan pemahaman teks. (Hidayatullah dkk., 2025)mengembangkan model *pre-trained transformer* yang secara khusus dirancang untuk menangani teks campuran Bahasa Indonesia, Jawa, dan Inggris. Penelitian ini menunjukkan bahwa model yang dilatih dengan data *code-mixed* mampu menghasilkan representasi kontekstual yang lebih baik dibandingkan model multibahasa umum. Hasil evaluasi mereka memperlihatkan peningkatan performa yang signifikan pada berbagai tugas klasifikasi teks, sehingga menegaskan pentingnya penyesuaian model terhadap karakteristik linguistik lokal.

Pendekatan lain dalam menangani teks *code-mixed* dikemukakan oleh (Takawane dkk., 2023) melalui penerapan *language augmentation* pada model BERT untuk klasifikasi teks. Penelitian ini memanfaatkan informasi bahasa pada tingkat kata untuk memperkaya data pelatihan, sehingga model dapat lebih adaptif terhadap variasi bahasa yang muncul pada media sosial. Hasil penelitian menunjukkan bahwa strategi augmentasi bahasa mampu meningkatkan akurasi klasifikasi secara konsisten dibandingkan pendekatan BERT standar tanpa augmentasi, khususnya pada teks pendek yang bersifat informal.

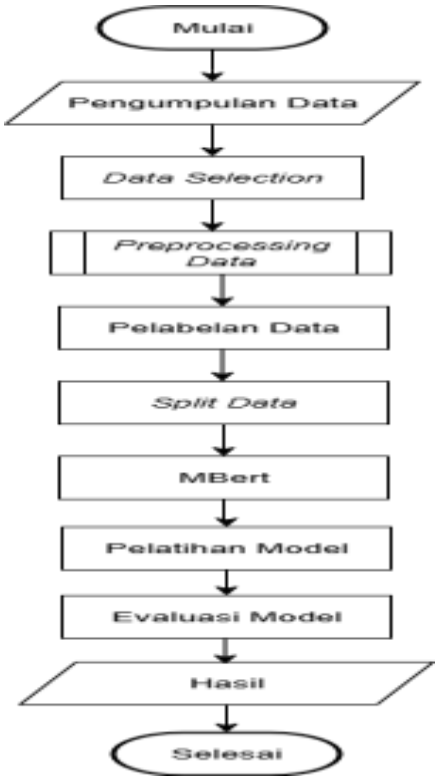
Fokus pada identifikasi bahasa di tingkat token dibahas dalam penelitian yang dipublikasikan pada (Hidayatullah dkk., 2024), yang mengkaji *word-level language identification* pada teks campuran Bahasa Indonesia, Jawa, dan Inggris. Penelitian ini membandingkan pendekatan berbasis aturan dan statistik dengan model berbasis *transformer*, dan menemukan bahwa model *transformer* lebih unggul dalam menangkap konteks linguistik lokal pada teks pendek. Temuan ini menegaskan bahwa pendekatan kontekstual sangat diperlukan untuk membedakan bahasa yang memiliki kedekatan leksikal dan sering muncul secara bercampur dalam satu kalimat.

Pemanfaatan model *transformer* multibahasa pada data media sosial juga diteliti dalam konteks analisis sentimen oleh penelitian di (Hashmi dkk., 2024). Studi ini menunjukkan bahwa model multibahasa seperti mBERT dan mBART mampu menangani variasi bahasa dan struktur kalimat tidak baku pada teks Twitter dengan lebih baik dibandingkan metode tradisional. Penelitian tersebut menekankan bahwa kemampuan *contextual embedding* pada *transformer* berperan penting dalam memahami makna teks yang ambigu akibat campuran bahasa dan gaya bahasa informal.

Selain itu, penelitian oleh (Patankar & Phadke, 2025) mengusulkan kerangka kerja hibrida CNN–*Transformer* untuk pengenalan emosi pada teks *code-mixed*. Meskipun studi ini berfokus pada bahasa Inggris–Hindi, pendekatan yang digunakan relevan untuk konteks bahasa serumpun. Hasil penelitian menunjukkan bahwa penggabungan fitur lokal dari CNN dan representasi kontekstual dari *transformer* mampu meningkatkan performa klasifikasi emosi secara signifikan. Temuan ini memperkuat argumen bahwa arsitektur *transformer,* baik secara mandiri maupun hibrida, efektif dalam menangani kompleksitas linguistik pada teks media sosial.

METODE PENELITIAN

Adapun metode penelitian yang digunakan oleh penelitian tersebut, yaitu sebagai berikut:



Gambar 1 *Flowchart* Perancangan Sistem

1. Pengumpulan Data

Pada tahap ini, data yang digunakan dalam penelitian ini diperoleh melalui proses pengumpulan dari media sosial Twitter. Pengumpulan menggunakan pendekatan *web scraping* untuk mendapatkan data berupa teks yang relevan dengan penelitian, dan data dibatasi dalam rentang waktu dari 1 November 2024 sampai 30 Mei 2025.

2. *Data Selection*

Setelah data berhasil dikumpulkan, dilakukan proses seleksi data untuk menyaring cuitan yang relevan, serta menghapus data yang bersifat spam, duplikat, atau tidak mengandung informasi bahasa yang bisa diidentifikasi.

3. Preprocessing Data

Data teks yang sudah diseleksi selanjutnya diproses agar siap untuk dianalisis oleh model. Tahapan ini meliputi pembersihan teks (menghapus tanda baca, simbol, URL), konversi ke huruf kecil (*case folding*), dan normalisasi teks agar konsisten. Tujuan dari preprocessing adalah meningkatkan kualitas input yang akan diterima oleh model.

4. Pelabelan Data

Setelah teks dibersihkan, dilakukan pelabelan semi otomatis terhadap masing-masing data. Label yang digunakan adalah ID untuk Bahasa Indonesia dan MY untuk Bahasa Malaysia. Proses ini penting untuk memastikan bahwa model dapat belajar dari data yang sudah diklasifikasi dengan benar.

5. Split Data

Setelah pelabelan data, dataset dibagi menjadi dua bagian dengan proporsi 80% untuk data *training*, dan 20% untuk *testing.* Data *training* digunakan untuk melatih model, sedangkan data *testing* berfungsi mengukur performa model pada data baru yang belum pernah dilihat sebelumnya.

6. MBert

Pada tahap ini digunakan arsitektur *Multilingual* BERT (mBERT) sebagai model dasar karena kemampuannya dalam merepresentasikan berbagai bahasa, termasuk Bahasa Indonesia dan Bahasa Malaysia. Model mBERT memanfaatkan mekanisme *transformer* untuk menghasilkan representasi kontekstual dari teks pendek media sosial yang telah melalui proses *preprocessing*, sehingga mampu menangkap perbedaan linguistik antarbahasa.

7. Pelatihan Model

Setelah tahap mBERT, dilakukan pelatihan model dengan membagi data menjadi 80% data latih dan 20% data uji, Model mBERT digunakan sebagai model dasar dengan penambahan lapisan klasifikasi untuk membedakan Bahasa Indonesia dan Bahasa Malaysia. Pelatihan dilakukan dengan pembekuan bobot awal dan dilanjutkan *fine-tuning,* menggunakan optimizer Adam dan fungsi kerugian *categorical cross-entropy*, serta evaluasi kinerja berdasarkan metrik akurasi.

8. Evaluasi Model

Evaluasi model dilakukan menggunakan data uji untuk mengukur kinerja dan kemampuan generalisasi dalam mengklasifikasikan teks pendek berbahasa Indonesia dan Bahasa Malaysia. Kinerja model dievaluasi menggunakan metrik akurasi, *precision, recall*, dan *F1-score* yang diperoleh dari *confusion matrix* dan *classification report.* Hasil evaluasi ini menunjukkan performa model *Multilingual* BERT dalam membedakan kedua bahasa serta menjadi dasar pengembangan model selanjutnya.

9. Hasil

Pada tahap Hasil, ditampilkan keluaran akhir berupa hasil evaluasi model, yang mencakup nilai akurasi, *precision, recall*, dan *F1-score*, serta *confusion matrix*, untuk menunjukkan efektivitas *Multilingual BERT* dalam mengklasifikasikan teks pendek ke dalam Bahasa Indonesia dan Bahasa Malaysia. Hasil ini menjadi dasar dalam penarikan kesimpulan penelitian.

10. *Deployment Model*

Model *Multilingual* BERT yang telah dilatih dan dievaluasi diimplementasikan ke dalam aplikasi web berbasis Streamlit untuk melakukan klasifikasi teks pendek berbahasa Indonesia dan Bahasa Malaysia secara langsung. Aplikasi ini memungkinkan pengguna memasukkan teks dan memperoleh hasil klasifikasi secara otomatis, sehingga memudahkan pemanfaatan dan pengujian model.

HASIL DAN PEMBAHASAN

1. Pengumpulan Data

no	full_text	id	my
1	ohhh taeyong pernah ddk malaysia ke no wonder he said back in malaysia aaaa comel	0	1
2	i keep thinking about how smart you are siss pose comel tapi still jaga batas	0	1
3	comel je	0	1
4	topic suho eat nasi lemak he pronounce so cuteeee sayang can u make it exolbunnyzen	0	1
5	wait for meee i got little legs macam kucen munchkin aaaa comel	0	1
6	comel je kan baca jongdae hes so cute	0	1
7	alicia comel gila	0	1
8	haha comel	0	1
10	comel giler pls jgn comel2 satgi kita minat awak	0	1
11	kucing comel awak tak	0	1
13	yihhh comel	0	1
14	argh comel mak nak kahwin hahah	0	1
15	comel gila hello choi seungcheol	0	1

Gambar 2 Dataset yang telah dikumpulkan

Pada Gambar di atas ditunjukkan bahwa dataset pada penelitian ini diperoleh melalui proses *web scraping* platform Twitter menggunakan pustaka *snscrape*. Data yang dikumpulkan berupa teks pendek berbahasa Indonesia dan/atau Bahasa Malaysia dengan panjang 1–20 kata sesuai batasan penelitian. Sebelum pelabelan, dilakukan validasi awal menggunakan model *fastText* pralatih *lid.176.bin* untuk memastikan kesesuaian bahasa. Proses pengumpulan data menggunakan kata kunci umum yang sering digunakan pada kedua bahasa guna menangkap variasi bahasa formal dan nonformal. Dataset disimpan dalam format CSV/XLSX dengan kolom *full_text* sebagai isi teks, serta label *id* dan *my* yang menunjukkan keberadaan Bahasa Indonesia dan Bahasa Malaysia dalam teks.

2. Peprocessing Data

Tabel 1 Contoh Hasil *Preprocessing*

No	Clean text
1	ohhh taeyong pernah duduk malaysia ke no wonder he said back in malaysia comel
2	i keep thinking about how smart you are siss pose comel tapi still jaga batas
3	comel je
4	wait for me i got little legs macam kucing munchkin comel
5	argh comel mak nak kahwin

Tabel tersebut menampilkan hasil *preprocessing* data teks yang telah melalui tahap pembersihan dan normalisasi sebelum digunakan dalam proses analisis. *Preprocessing* meliputi penghapusan tanda baca, simbol, serta karakter berulang yang tidak memiliki makna semantik, diikuti dengan konversi seluruh teks ke huruf kecil (*case folding*) untuk menjaga konsistensi data. Selain itu, dilakukan normalisasi ringan pada kata tidak baku tanpa menghilangkan unsur bahasa campuran yang masih relevan. Hasil akhir berupa *clean text* yang lebih terstruktur dan siap digunakan sebagai input pada model *Multilingual* BERT.

3. Evaluasi Model

a. Evaluasi Akurasi dan Loss

Tabel 2 Evaluasi Akurasi Dan Loss

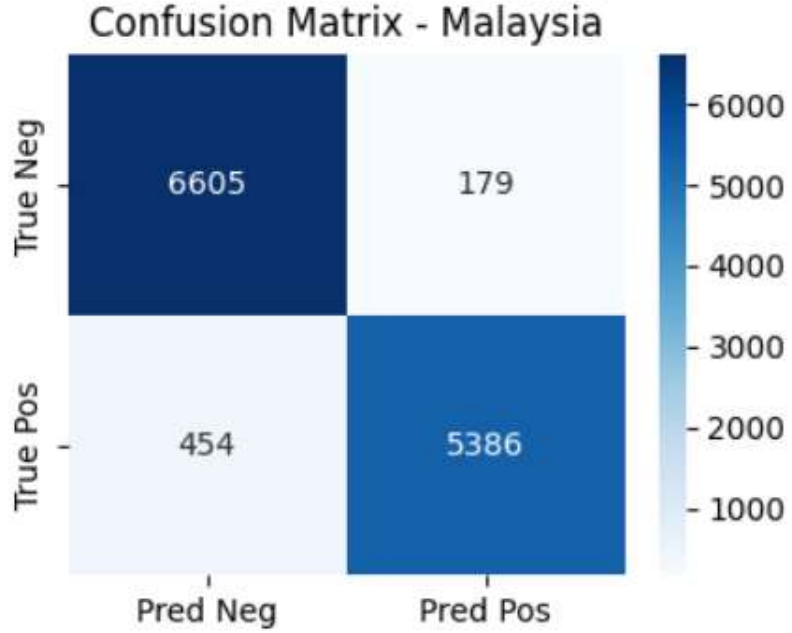
Dataset	Akurasi%	Loss	Precission	Recall	F1-Score
Training Set	97,25	0,0622	-	-	-
Test Set	95,81	0,1329	0,9798	0,9515	0,9654

Tabel ini menampilkan Hasil evaluasi kinerja model *Multilingual* BERT pada data latih dan data uji. Tabel ini menunjukkan nilai akurasi dan *loss* pada data latih, serta metrik evaluasi berupa akurasi, *precision, recall*, dan *F1-score* pada data uji. Metrik evaluasi lengkap hanya ditampilkan pada data uji karena lebih merepresentasikan kemampuan generalisasi model dalam mengklasifikasikan teks pendek berbahasa Indonesia dan Bahasa Malaysia.

b. Confusion Matrix

Gambar 3 *Confusion Matrix* Label Bahasa Indonesia

Gambar diatas menampilkan *confusion matrix* untuk label Bahasa Indonesia (id). Dari total 12.624 data uji, terdapat 9.688 data yang seharusnya berlabel Bahasa Indonesia. Model berhasil mengklasifikasikan 9.389 data sebagai *true positive* (prediksi benar), sedangkan 299 data salah diprediksi sebagai *false negative* (terlewat sebagai Bahasa Indonesia). Pada sisi lain, terdapat 2.936 data yang memang tidak berlabel Bahasa Indonesia, dengan 2.810 data berhasil dikenali sebagai *true negative*, dan 126 data yang salah diklasifikasikan sebagai *false positive*. Nilai *recall* yang tinggi (sekitar 0,97) mengindikasikan bahwa model jarang melewatkan teks yang seharusnya masuk kategori Bahasa Indonesia.



Gambar 4 *Confusion Matrix* Label Bahasa Malaysia

Gambar diatas menunjukkan *confusion matrix* untuk label Bahasa Malaysia (my). Dari total 12.624 data uji, terdapat 5.840 data yang seharusnya berlabel Bahasa Malaysia. Model berhasil mengklasifikasikan 5.386 data sebagai *true positive*, sedangkan 454 data salah diprediksi sebagai *false negative*. Sementara itu, dari 6.784 data yang tidak berlabel Bahasa Malaysia, sebanyak 6.605 data berhasil dikenali sebagai *true negative*, dan 179 data salah diklasifikasikan sebagai *false positive*. Nilai *recall* yang relatif tinggi (sekitar 0,92) menunjukkan bahwa model cukup baik dalam mengenali teks Bahasa Malaysia, meskipun masih ada sebagian kecil teks yang terlewat, terutama pada kalimat yang sangat mirip dengan Bahasa Indonesia.

c. Classification Report

Tabel 3 *Classification Report*

Kelas	Deskripsi	Precision	Recall	F1-Score
ID	Bahasa Indonesia	0.99	0.97	0.98
MY	Bahasa Malaysia	0.97	0.92	0.94

Tabel diatas Hasil evaluasi kinerja model *Multilingual* BERT pada masing-masing kelas bahasa. Pada kelas Bahasa Indonesia, model mencapai nilai *precision* sebesar 0,99, *recall* sebesar 0,97, dan *F1-score* sebesar 0,98 yang menunjukkan kemampuan klasifikasi yang sangat baik. Sementara itu, pada kelas Bahasa Malaysia diperoleh nilai *precision* sebesar 0,97, *recall* sebesar 0,92, dan *F1-score* sebesar 0,94, yang mengindikasikan masih terdapat sebagian kecil data yang belum terdeteksi secara optimal. Secara keseluruhan, nilai *macro average F1-score* sebesar 0,96 menunjukkan bahwa model memiliki performa yang baik dan relatif seimbang dalam mengklasifikasikan kedua bahasa.

4. Hasil

Tabel 4 Contoh Hasil

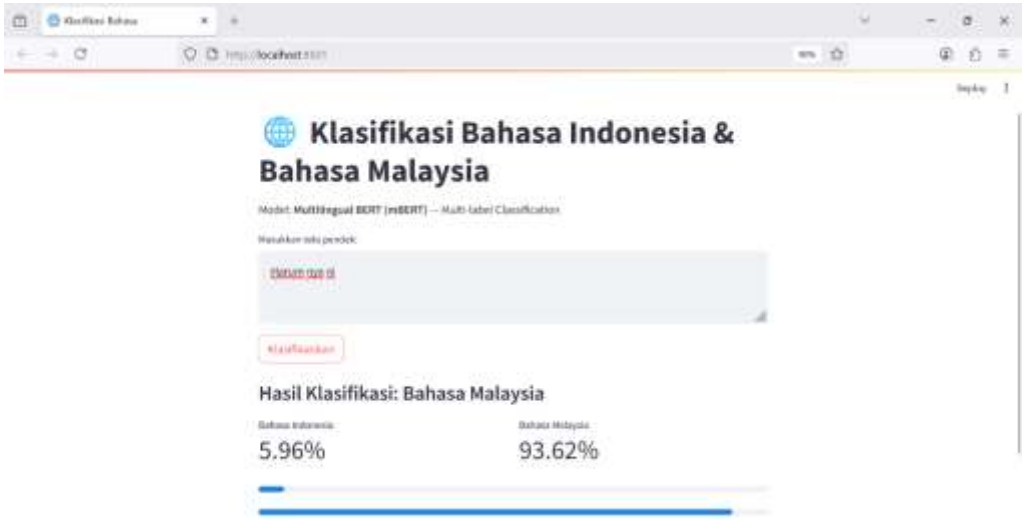
No	Input Teks	Prediksi Bahasa	Probabilitas (%)
1	“Harum nye ni”	Bahasa Malaysia	93.62
2	“bauun nye”	Bahasa Malaysia	98.43
3	“saya nak makan”	Campuran	ID: 63.86 / MY: 74.40
4	“saya lagi makan”	Campuran	ID: 88.41 / MY: 68.59
5	“jom la makan di rumah gue”	Campuran	ID: 87.51 / MY: 51.54
6	“saya makan bakso di dekat warung rumah”	Bahasa Indonesia	99.59

Dari hasil tabel di atas terlihat bahwa sistem mampu mengenali bahasa dengan cukup baik, bahkan ketika teks mengandung singkatan atau gaya bahasa informal. Misalnya, pada teks *“bauun nye”*, sistem secara tepat mengklasifikasikannya sebagai Bahasa Malaysia dengan tingkat keyakinan 98,43%. Sementara pada teks *“saya lagi makan”* yang berpotensi ambigu karena kosakata umum digunakan di kedua bahasa, sistem memberikan prediksi campuran dengan probabilitas cukup tinggi pada keduanya.

5. Deployment Model

Gambar 5 Halaman Awal

Desain halaman awal ini dibuat bersih, terpusat, dan responsif sehingga dapat diakses baik dari perangkat desktop maupun *mobile*. Tidak ada elemen yang mengganggu fokus, sehingga perhatian pengguna tertuju langsung pada proses memasukkan teks dan melakukan prediksi. Pendekatan minimalis ini bertujuan untuk mempermudah interaksi pengguna dengan sistem, sekaligus mempercepat proses penggunaan tanpa memerlukan banyak navigasi.



Gambar 6 Halaman Hasil Klasifikasi

Gambar diatas Tampilan antarmuka aplikasi web klasifikasi bahasa berbasis *Multilingual* BERT (mBERT). Aplikasi ini memungkinkan pengguna memasukkan teks pendek untuk diklasifikasikan ke dalam Bahasa Indonesia atau Bahasa Malaysia. Pada contoh pengujian, sistem menampilkan hasil prediksi berupa probabilitas masing-masing kelas, yaitu 5,96% untuk Bahasa Indonesia dan 93,62% untuk Bahasa Malaysia, sehingga teks diklasifikasikan sebagai Bahasa Malaysia. Tampilan ini menunjukkan implementasi model yang bersifat interaktif dan mudah digunakan untuk pengujian secara langsung.

KESIMPULAN

Berdasarkan hasil penelitian, model Multilingual BERT (mBERT) terbukti mampu mengklasifikasikan teks berbahasa Indonesia dan Bahasa Malaysia secara efektif pada data media sosial. Model yang diusulkan menghasilkan akurasi sebesar 95–96% dengan nilai F1-score rata-rata 0,96, yang menunjukkan performa yang stabil pada kedua kelas bahasa. Meskipun demikian, masih terdapat kesalahan klasifikasi pada teks yang memiliki kemiripan kosakata tinggi antara kedua bahasa. Secara keseluruhan, hasil ini menunjukkan bahwa mBERT memiliki potensi yang baik untuk diterapkan dalam tugas klasifikasi bahasa serumpun pada teks pendek media sosial.

DAFTAR PUSTAKA

Ansari, M. Z., Ahmad, T., & Fatima, A. (2020). *Feature Selection on Noisy Twitter Short Text Messages for Language Identification*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Unsupervised Cross-lingual Representation Learning at Scale*. http://arxiv.org/abs/1911.02116

Guo, X., Adnan, H. M., & Abidin, M. Z. Z. (2024). Detecting Offensive Language on Malay Social Media: A Zero-Shot, Cross-Language Transfer Approach Using Dual-Branch mBERT. *Applied Sciences (Switzerland)*, *14*(13). https://doi.org/10.3390/app14135777

Hashmi, E., Yayilgan, S. Y., & Shaikh, S. (2024). Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers. *Social Network Analysis and Mining*, *14*(1), 1–15. https://doi.org/10.1007/s13278-024-01245-6

Hidayatullah, A. F., Apong, R. A., Lai, D. T. C., & Qazi, A. (2024). Word Level Language Identification in Indonesian-Javanese-English Code-Mixed Text. *Procedia Computer Science*, *244*, 105–112. https://doi.org/10.1016/j.procs.2024.10.183

Hidayatullah, A. F., Apong, R. A., Lai, D. T. C., & Qazi, A. (2025). Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer. *Social Network Analysis and Mining*, *15*(1), 1–17. https://doi.org/10.1007/s13278-025-01444-9

Husyandi, M. (2025). Analisis Komparatif Kosakata Bahasa Indonesia Dan Bahasa Melayu Malaysia Dalam Episode Perdana Serial Drama "Bidaah." *Jurnal Bahasa Asing*, *18*(1), 74–84. https://doi.org/10.58220/jba.v18i1.110

Jauhiainen, T., Lindén, K., & Jauhiainen, H. (2019). Language model adaptation for language and dialect identification of text. In *Natural Language Engineering* (Vol. 25, Nomor 5, hal. 561–583). Cambridge University Press. https://doi.org/10.1017/S135132491900038X

Lu, Y. J., & Li, C. Te. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 505–514. https://doi.org/10.18653/v1/2020.acl-main.48

Ma, N., Politowicz, A., Mazumder, S., Chen, J., Liu, B., Robertson, E., & Grigsby, S. (2021). Semantic Novelty Detection in Natural Language Descriptions. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 866–882. https://doi.org/10.18653/v1/2021.emnlp-main.66

Maxwell-Smith, Z., Kohler, M., & Suominen, H. (2021). *Scoping natural language processing in Indonesian and Malay for education applications*.

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). *BERTweet : A pre-trained language model for English Tweets*. 9–14.

Patankar, S., & Phadke, M. (2025). A CNN-transformer framework for emotion recognition in code-mixed English–Hindi data. *Discover Artificial Intelligence*, *5*(1), 1–13. https://doi.org/10.1007/s44163-025-00400-y

Pires, T., Schlinger, E., & Garrette, D. (2019). *How multilingual is Multilingual BERT?* https://github.com/google-research/bert

Putra, I. F., & Purwarianti, A. (2020, September 8). Improving Indonesian Text Classification

Using Multilingual Language Model. *2020 7th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2020*. https://doi.org/10.1109/ICAICTA49861.2020.9429038

Review, W. P. (2025). *Twitter Users by Country – Indonesia (27.1 million users in 2024)*.

Ruder, S., Vulić, I., & Søgaard, A. (2019). *11640-Article (PDF)-21826-1-10-20190813*. *65*, 569–630.

Singh, G., Sharma, S., Kumar, V., Kaur, M., Baz, M., & Masud, M. (2021). Spoken Language Identification Using Deep Learning. *Computational Intelligence and Neuroscience*, *2021*. https://doi.org/10.1155/2021/5123671

Takawane, G., Phaltankar, A., Patwardhan, V., Patil, A., Joshi, R., & Takalikar, M. S. (2023). Language augmentation approach for code-mixed text classification. *Natural Language Processing Journal*, *5*(November), 100042. https://doi.org/10.1016/j.nlp.2023.100042

Wu, S., & Dredze, M. (2020). *Are All Languages Created Equal in Multilingual BERT?*

Zhao, S., Gupta, R., Song, Y., & Zhou, D. (2021). Extremely small BERT models from mixed-vocabulary training. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 2753–2759. https://doi.org/10.18653/v1/2021.eacl-main.238