

NLP Assignment

Brian Alexander - 2702282351

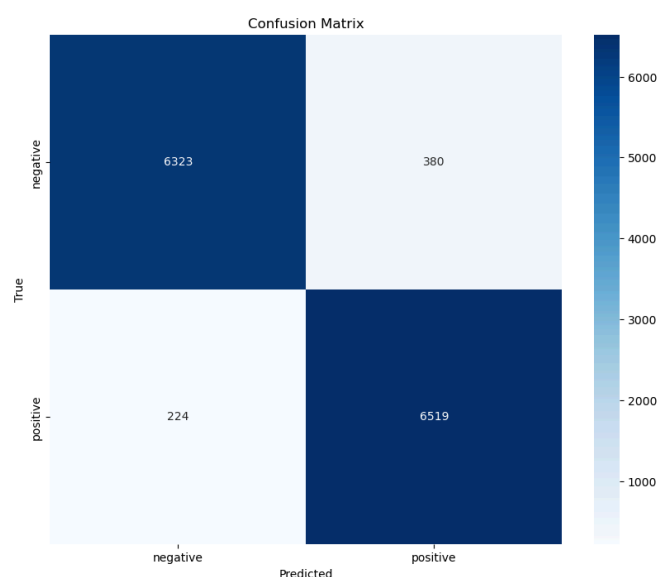
Kinsley Reynard Tanjung - 2702349432

Jason Christian Budhihartono - 2702326593

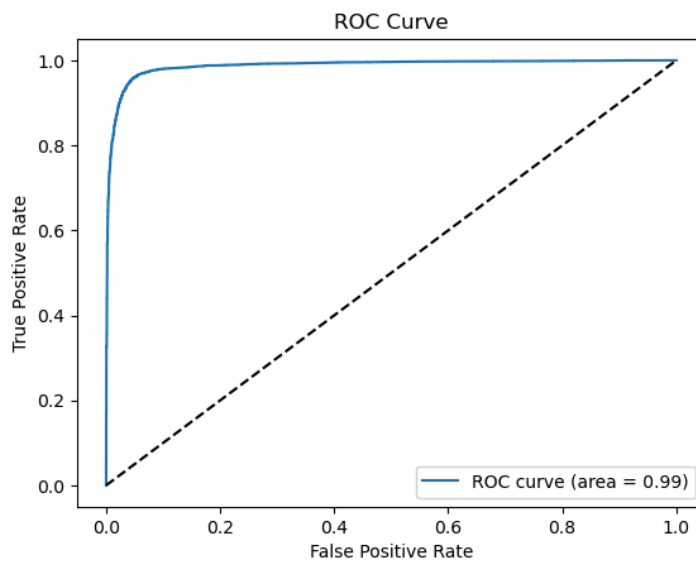
1. Model yang kami pilih adalah model Indobert, salah satu turunan dari arsitektur BERT (Bidirectional Encoder Representations from Transformer) yang dikhususkan untuk Bahasa Indonesia. Kenapa kami memilih model ini? Karena BERT sendiri sampai sekarang masih menjadi salah satu model *state-of-the-art* atau model yang populer dalam dunia NLP (*Natural Language Processing*). Apa lagi dengan adanya terobosan Indobert yaitu turunan dari BERT yang dibuatkan untuk teks berbahasa Indonesia sehingga ini menjadi salah satu poin menarik bagi kami untuk mencoba arsitektur model indobert ini untuk mengetahui apakah model ini setara dengan model BERT pada teks berbahasa Indonesia.
2. Untuk Indobert sendiri, sebenarnya cara kerjanya sama persis seperti bert sehingga kami akan menjelaskan bagaimana bert sendiri dalam mengenali dan memahami sebuah teks yang diberikan bukan hanya itu, bahkan bert bisa tahu arti dari sebuah kata dalam hal konteks dimana bert bisa membedakan arti dari sebuah kata meski katanya sama persis namun dalam konteks yang berbeda. Bert mengenalkan algoritma bernama bidirectional, dimana model ini tidak hanya membaca sebuah kalimat dari kiri ke kanan seperti pada umumnya, namun model ini membaca sebuah kalimat dari kiri ke kanan dan sebaliknya yang tentunya akan meningkatkan kemampuan model untuk memahami konteks pada kalimat tersebut. Kemudian Bert dibuat dengan arsitektur transformer sehingga ada mekanisme attention dimana sebuah kata akan dibandingkan atau dihubungkan dengan kata lainnya untuk memahami konteks pada sebuah kalimat. Berbeda dengan transformer, bert hanya mengandalkan encoder karena tujuan dari bert sendiri hanya fokus memahami arti pada sebuah kalimat.

5.

A. Indobert



Classification Report:				
	precision	recall	f1-score	support
negative	0.97	0.94	0.95	6703
positive	0.94	0.97	0.96	6743
accuracy			0.96	13446
macro avg	0.96	0.96	0.96	13446
weighted avg	0.96	0.96	0.96	13446

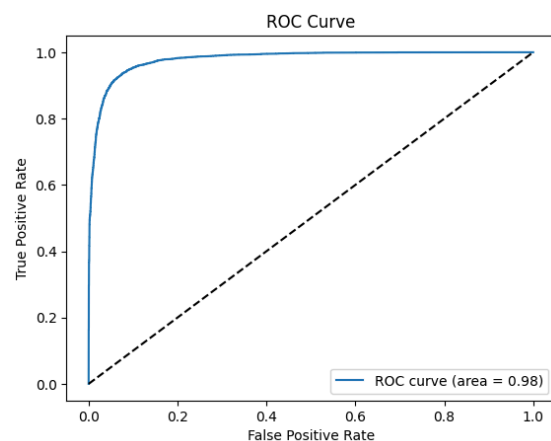
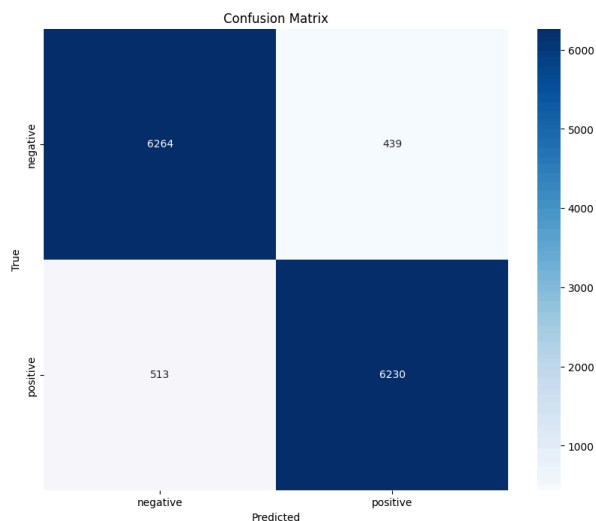


Gambar-gambar di atas merupakan evaluasi dari model Indobert dari confusion matrix, classification report, dan roc curve. Jadi, performa dari model indobert terhadap dataset kami cukup baik dan menandakan model yang kami gunakan bisa belajar dengan baik dan memahami konteks dari kalimat yang diberikan. Ini bisa terjadi karena Indobert merupakan model transformer yang canggih dan dapat memahami konteks dari kalimat berkat algoritmanya, kemudian kami menggunakan model pre-trained, yang sebetulnya sudah ditrain dengan sangat baik sebelumnya jadi hanya perlu dilakukan finetuning dan penyesuaian input data sehingga hasilnya bisa sebaik ini. Peningkatan yang bisa dilakukan pada model ini ialah meningkatkan variasi dataset, karena dataset yang kami gunakan sebetulnya masih dalam Bahasa Inggris dan kami lakukan translate sehingga bahasa yang digunakan cenderung formal, dengan meningkatkan variasi bahasa yang lebih terkini dan gaul.

B. Indobert-lite

Indobert-lite merupakan versi ringan dari model indobert dengan mengurangi jumlah parameter train sehingga mengurangi beban komputasi dari model ini tanpa menghilangkan performa dari model indobert.

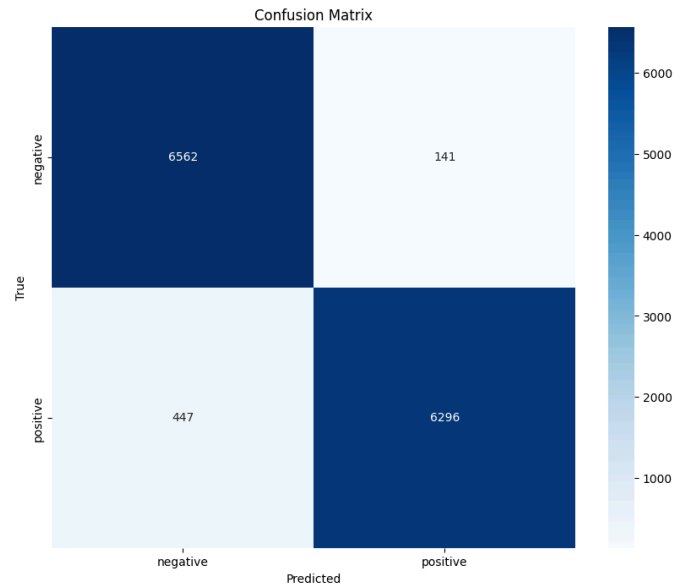
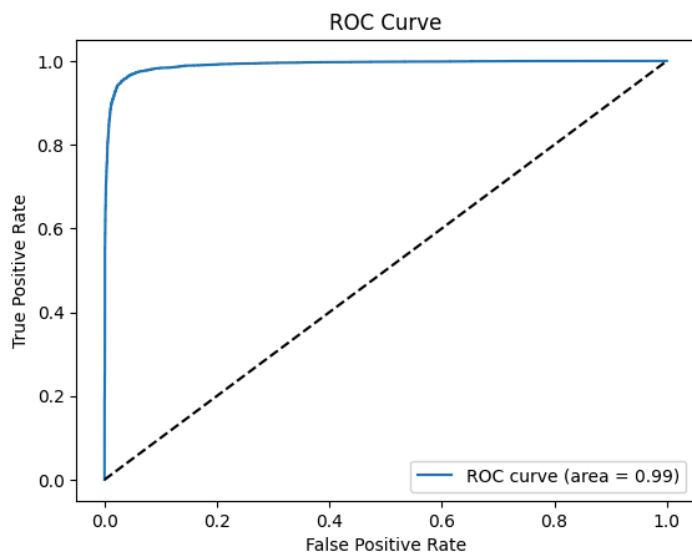
Classification Report:				
	precision	recall	f1-score	support
negative	0.92	0.93	0.93	6703
positive	0.93	0.92	0.93	6743
accuracy			0.93	13446
macro avg	0.93	0.93	0.93	13446
weighted avg	0.93	0.93	0.93	13446



Gambar-gambar di atas merupakan evaluasi dari model Indobert-lite dari confusion matrix, classification report, dan roc curve. Jadi, performa dari model indobert terhadap dataset kami cukup baik dan menandakan model yang kami gunakan bisa belajar dengan baik dan memahami konteks dari kalimat yang diberikan. Ini bisa terjadi karena walaupun ini merupakan versi ringan dari Indobert tetapi indobert sendiri cukup powerful sehingga meskipun dikurangi parameternya, model ini masih bisa sedikit mempertahankan performanya. Namun ketika diujikan pada data real, model ini gagal memprediksi class positive yang menurut kami terjadi karena kualitas dataset yang belum maksimal sehingga peningkatan yang bisa dilakukan pada model ini ialah meningkatkan variasi dataset, karena dataset yang kami gunakan sebetulnya masih dalam Bahasa Inggris dan kami lakukan translate sehingga bahasa yang digunakan cenderung formal, kurangnya variasi kata yang menandakan class positif juga bisa menyebabkan model ini sulit untuk memahami kata yang belum pernah dilihat.

C. Distilbert (Indo version)

Jadi distilbert merupakan salah satu turunan bert juga dimana ia melakukan distillation atau meniru performa dari bert namun tetap bisa menghemat beba komputasi. Cara kerjanya seperti murid yang belajar dari gurunya.

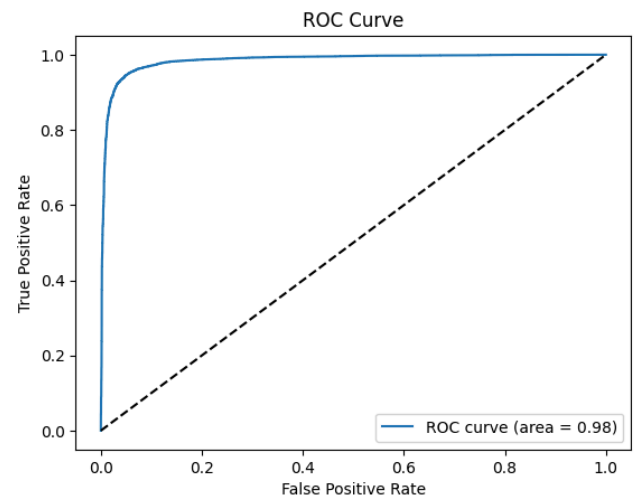
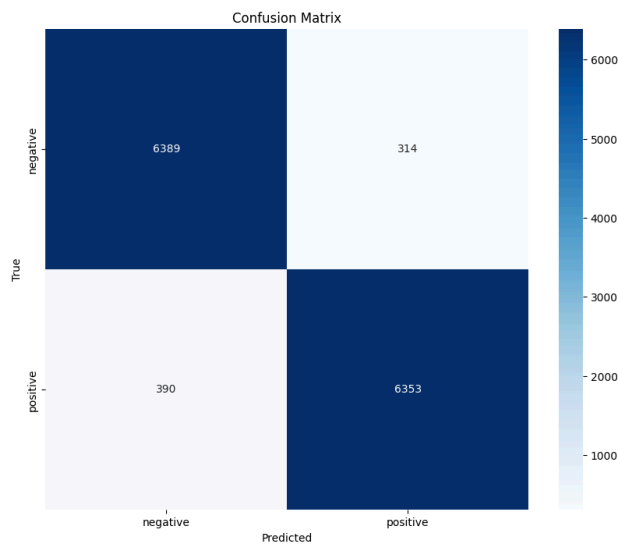


Classification Report:				
	precision	recall	f1-score	support
negative	0.94	0.98	0.96	6703
positive	0.98	0.93	0.96	6743
accuracy			0.96	13446
macro avg	0.96	0.96	0.96	13446
weighted avg	0.96	0.96	0.96	13446

Gambar di atas merupakan hasil evaluasi dari model DistilIndoBERT menggunakan confusion matrix, classification report, dan roc curve. Model distilbert ini memiliki performa yang seimbang bahkan sedikit melampaui model indobert, ini bisa terjadi karena meskipun parameter dikurangi, distilbert tetap bisa meimbangi performa dari model bert karena meniru dan mempelajari dari model bert itu sendiri. Peningkatan yang bisa dilakukan adalah meningkatkan variasi pada dataset untuk memastikan kata-kata yang unik dan baru bisa diprediksi juga dengan baik oleh model ini.

D. mBERT (Multilingual BERT)

mBERT sendiri adalah salah satu turunan dari BERT dimana model BERT dilatih pada dataset multi-language, artinya model ini memiliki kemampuan dalam mengenali dan mengklasifikasi bahasa-bahasa yang berbeda.



Classification Report:				
	precision	recall	f1-score	support
negative	0.94	0.95	0.95	6703
positive	0.95	0.94	0.95	6743
accuracy			0.95	13446
macro avg	0.95	0.95	0.95	13446
weighted avg	0.95	0.95	0.95	13446

Gambar di atas merupakan hasil evaluasi dari model mBERT menggunakan confusion matrix, classification report, dan roc curve. Model ini memiliki performa yang sedikit buruk jika dibandingkan dengan hasil dari model indobert. Hal ini dikarenakan model ini dilatih dengan tujuan bisa menggeneralisasi berbagai bahasa, sehingga performanya akan berkurang namun fleksibel untuk digunakan dengan menyesuaikannya dengan dataset. Peningkatan bisa dilakukan dengan memperbaiki dataset yang lebih baik sehingga bisa mewakili tata bahasa Indonesia.

KESIMPULAN

Menurut kami, semua model yang kami ujikan memiliki performa dan arsitektur yang cukup mirip dan seimbang, hanya saja kami lebih merasa dataset yang kami gunakan belum maksimal, karena dataset ini aslinya merupakan bahasa Indonesia yang kami terjemahkan. Sehingga tata bahasanya cukup formal dan belum bisa dikatakan mewakili tata bahasa yang baik dan benar.