Indonesian Journal on Software Engineering (IJSE)
Vol. 12, No. 1, Juni 2026, hlm. 9-18 p-ISSN: 2461-0690
Akreditasi KEMENRISTEKDIKTI, No. 225/E/KPT/2022 e-ISSN: 2714-9935

INTERPRETABILITAS KLASIFIKASI TEKS KESEHATAN MENTAL BERBASIS TRANSFORMER MENGGUNAKAN LIME DAN KNOWLEDGE GRAPH

Siti Ernawati [*1], Risa Wati [2]

[1]Universitas Nusa Mandiri
Jl. Raya Jatiwaringin No.2, Kota Jakarta Timur, DKI Jakarta 13620, Indonesia
[2]Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, Kota Jakarta Pusat, DKI Jakarta 10450, Indonesia
Email: [1]siti.ste@nusamandiri.ac.id, [2]risawati.rwx@bsi.ac.id

Abstrak

Perkembangan model berbasis Transformer telah meningkatkan akurasi klasifikasi teks kesehatan mental secara signifikan, namun interpretabilitas model masih menjadi tantangan utama karena sifatnya yang black-box. Meskipun berbagai pendekatan XAI (Explainable Artificial Intelligence) telah dikembangkan, integrasi visualisasi pengetahuan konseptual untuk mendukung interpretasi hasil prediksi masih terbatas. Penelitian ini bertujuan mengembangkan kerangka post-hoc explainability dengan mengombinasikan metode LIME dan visualisasi Knowledge Graph untuk meningkatkan transparasi klasifikasi teks Kesehatan mental berbasis Transformer. Empat model Transformer yaitu BERT, RoBERTa, DistilBERT, dan XLNet, dievaluasi menggunakan metrics Accuracy, Precision, Recall, F1-Score, dan AUC. Hasil eksperimen menunjukan bahwa DistilBERT memberikan performa terbaik dengan akurasi 88.91%, precision 88.94%, recall 88.91%, F1-score 88.90%, dan AUC sebesar 92.60%. Pendekatan LIME digunakan untuk mengidentifikasi kontribusi fitur secara lokal terhadap prediksi, sementara Knowledge Graph memvisualisasikan hubungan semantik antar entitas yang berpengaruh dalam proses klasifikasi. Hasil visualisasi menunjukan bahwa fitur linguistik dengan muatan emosional tinggi memiliki peran dominan dalam penentuan kategori. Penelitian ini berkontribusi dengan menyediakan kerangka interpretabilitas berbasis visual yang mendukung transparasi model. Pendekatan ini berpotensi mendukung pengembangan sistem deteksi kesehatan mental berbasis AI yang lebih transparan, dapat dipercaya, dan berorientasi pada prinsip human-centered AI.

Kata kunci: *Explainable AI, Knowledge Graph, Kesehatan Mental, LIME, Model Transformer*

Abstract

The development of Transformer-based models has significantly improved the accuracy of classification of mental health texts, but the interpretability of the models remains a major challenge due to their black-box nature. Although various XAI (Explainable Artificial Intelligence) approaches have been developed, the integration of conceptual knowledge visualization to support the interpretation of predictive results is still limited. This study aims to develop a framework of post-hoc explainability by combining the LIME method and visualization of the Knowledge Graph to improve the transparency of the text classification of Transformer-based mental health. Four Transformer models, namely BERT, RoBERTa, DistilBERT, and XLNet, were evaluated using the metrics Accuracy, Precision, Recall, F1-Score, and AUC. The experimental results showed that DistilBERT gave the best performance with an accuracy of 88.91%, precision of 88.94%, recall of 88.91%, F1-score of 88.90%, and AUC of 92.60%. The LIME approach is used to identify the contribution of features locally to the prediction, while the Knowledge Graph visualizes semantic relationships between influential entities in the classification process. Visualization results show that linguistic features with high emotional charge have a dominant role in determining the category. This research contributes by providing a visual-based interpretability framework that supports model transparency. This approach has the potential to support the development of AI-based mental health detection systems that are more transparent, trustworthy, and oriented towards human-centered AI principles.

Keywords: *Explainable AI, Knowledge Graph, LIME, Mental Health, Transformer Model*

1. PENDAHULUAN

Perkembangan pesat *platform* komunikasi digital telah mendorong meningkatnya interaksi daring, di mana individu mengekspresikan emosi, pemikiran, serta kondisi psikologis mereka melalui media sosial. Aktifitas daring tersebut menjadikan media sosial sebagai sumber data yang potensial untuk mendeteksi indikasi gangguan kesehatan mental seseorang secara dini. Berdasarkan laporan WHO (World Health Organization) tahun 2024, memperkirakan sekitar 1 dari 8 penduduk dunia atau sekitar 970 juta orang mengalami gangguan kesehatan mental, dari jumlah tersebut, depresi dan kecemasan menjadi gangguan yang paling umum. Sebanyak 5% orang dewasa di seluruh dunia menderita depresi, dan 4% orang dewasa mengalami gangguan kecemasan (Huntington-Psychological Services, 2024). Kondisi ini menunjukan urgensi pengembangan sistem berbasis data yang mampu membantu identifikasi dini resiko gangguan mental secara otomatis dan efisien.

Dalam beberapa tahun terakhir, pendekatan berbasis NLP (*Natural Language Processing)* telah banyak digunakan untuk menganalisis pola linguistik. Model deep learning berbasis Transformer seperti BERT, DistilBERT, RoBERTa, dan XLNet telah menunjukkan kinerja yang sangat baik dalam mengidentifikasi pola linguistik yang berkaitan dengan kondisi kesehatan mental. Keunggunaln arsitektur Transformer terletak pada mekanisme *self-attention,* yang mampu menangkap dependensi jangka panjang antar kata sehingga menghasilkan representasi kontekstual yang kuat untuk memahami konteks setiap kalimat (Hedhili & Bouallagui, 2024). Namun demikian, meskipun memiliki tingkat akurasi yang tinggi, model Transformer sering dikritik karena sifatnya yang bersifat *black-box* (Saxena et al., 2024), (Hameed et al., 2025) yang membuat logika di balik prediksi model sulit untuk dijelaskan atau diinterpretasikan. Kurangnya interpretabilitas dapat menurunkan tingkat kepercayaan pengguna dan praktisi kesehatan terhadap hasil prediksi model, terutama ketika keputusan yang dihasilkan berkaitan dengan kondisi psikologis individu. Ketidakjelasan dapat menurunkan kredibilitas dan penerimaan model dalam praktik dunia nyata. Meskipun sistem pemrosesan informasi skala besar dapat mengumpulkan sejumlah besar fakta-fakta yang saling terkait, mengubah fakta-fakta kandidat tersebut menjadi pengetahuan yang bermanfaat merupakan tantangan yang sangat sulit (Pujara et al., 2013).

Berbagai pendekatan XAI *(Explainable Artificial Intellegence*) seperti LIME *(Local Interpretable Model-Agnostic Explanations)* telah dikembangkan untuk menjelaskan kata atau frasa yang mempengaruhi hasil prediksi model (Aljrees, 2024; Mustafa & Hama Saeed, 2025). Namun pendekatan tersebut umumnya bersifat *post-hoc* dan hanya menyoroti kontribusi fitur secara local terhadap satu instance prediksi, tanpa merepresentasikan keterkaitan konseptual antar fitur dalam struktur pengetahuan yang lebih luas. Sehingga, penjelasan yang dihasilkan cenderung belum mampu menggambarkan hubungan semantic antar indicator kesehatan mental secara sistematis. Di sisi lain, *Knowledge Graph* menawarkan representasi pengetahuan berbasis entitas dan relasi yang terstruktur, sehingga dapat digunakan untuk memvisualisasikan hubungan konseptual antar kata. *Knowledge Graph* digunakan sebagai pendekatan *post-hoc visualization* untuk memetakan fitur-fitur penting yang diidentifikasi oleh LIME ke dalam struktur graf relasional. Dengan pendekatan ini, hasil interpretasi lokal dari LIME diperluas ke dalam representasi visual yang menampilkan keterkaitan semantik antar indikator kesehatan mental secara lebih terstruktur.

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan untuk mengembangkan kerangka kerja *post-hoc explainability* yang menggabungkan model Transformer, dan model XAI (LIME), serta memanfaatkan *Knowledge Graph* sebagai sarana visualisasi interpretatif. Kerangka kerja yang diusulkan tidak hanya mengidentifikasi fitur linguistik yang berkontribusi terhadap hasil prediksi, tetapi juga memetakan fitur tersebut ke dalam struktur *Knowledge Graph* guna merepresentasikan hubungan semantik secara eksplisit. Pendekatan ini diharapkan dapat meningkatkan transparansi model tanpa mengubah arsitektur dasar Transformer, serta berkontribusi pada pengembangan sistem kecerdasan buatan yang transparan, dapat dipercaya, dan berorientasi pada *human-centered AI* dalam bidang analisis kesehatan mental berbasis teks.

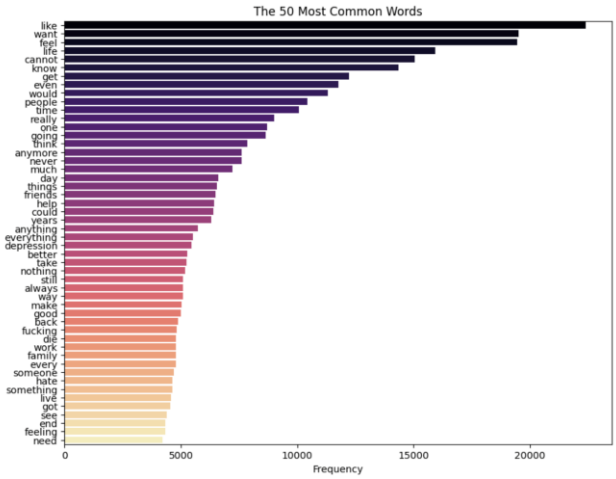
2. METODE PENELITIAN

2.1 Dataset

Penelitian ini menggunakan dataset kesehatan mental yang terdiri dari 20.339 teks berbahasa Inggris yang diperoleh dari Kaggle (Kumar, 2023). Dataset dibagi ke dalam dua kelas yaitu depression (10.351 data) dan SuicideWatch (9.988 data). Distribusi label relatif seimbang sehingga tidak diperlukan teknik resampling tambahan. Dataset dibagi dengan proporsi 70% untuk data pelatihan dan 30% untuk data pengujian guna menjaga distribusi label tetap konsisten pada kedua subset.

Gambar 1 menampilkan 50 kata dengan frekuensi kemunculan tertinggi pada dataset. Kata-kata dominan seperti like, want, feel, dan life menunjukkan fokus pembahasan pada aspek emosional dan pengalaman hidup, sedangkan kata seperti depression, help, dan die mengindikasikan adanya isu

kesehatan mental. Di sisi lain, kemunculan kata friends, family, dan better mencerminkan peran dukungan sosial dan harapan pemulihan. Hasil ini memberikan gambaran pola linguistik yang menjadi dasar dalam proses klasifikasi teks.



Gambar. 1. 50 Kata yang Sering Muncul Pada Dataset

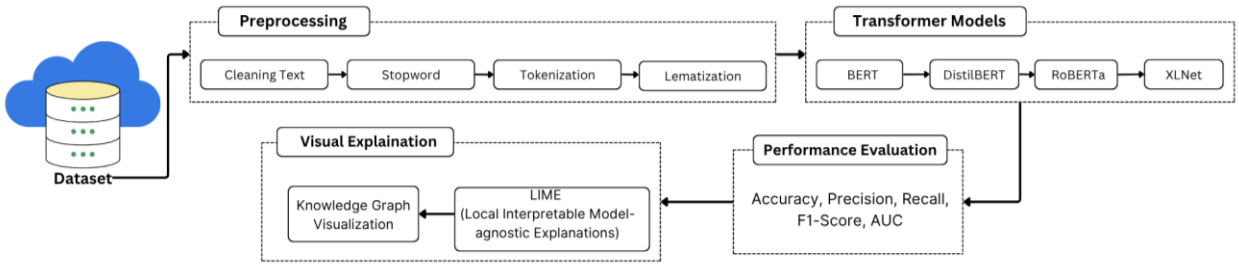
2.2 Data Preprocessing

Tahap ini merupakan proses penting dalam pengolahan data, agar kualitas data siap digunakan dalam proses pemodelan, dan analisis lebih lanjut (Hickman et al., 2022). Penelitian ini melakukan preprocessing meliputi tahap cleaning text, stopword, lemmatization, tokenization (Hassan et al., 2022). Cleaning teks dilakukan untuk menghilangkan simbol, tanda baca, karakter yang tidak relevan (Sen et al., 2020), (Alshuwaier & Alsulaiman, 2025). Stopword digunakan untuk menghapus kata-kata yang dianggap tidak memberikan makna. Lemmatization mengubah kata menjadi kata dasar, dan tokenization untuk memisahkan teks menjadi satuan kata yang terpisah. Seluruh tahapan ini dilakukan agar menghasilkan data yang bersih dan terstruktur (Mamdouh Farghaly & Abd El-Hafeez, 2023), dan sesuai untuk diproses selanjutnya. Tabel 1 menunjukkan hasil preprocessing pada salah satu sampel teks yang digunakan dalam penelitian ini.

Tabel 1. Sampel Hasil Preprocessing

Text	I want to kill myself. there is no hope at all Everyday I think
Cleaning Text	i want to kill myself there is no hope at all everyday i think
Stop Word	want kill hope everyday think
Tokenization	['want', 'kill', 'hope', 'everyday', 'think']
Lemmatization	['want', 'kill', 'hope', 'everyday', 'think']

2.3 Model Arsitektur



Gambar 2. Model Arsitektur Usulan

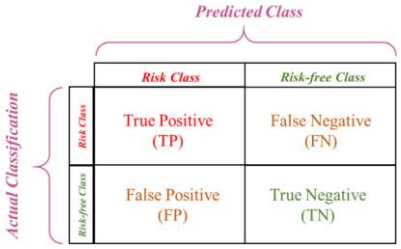
Gambar 2 menunjukan model arsitektur yang diusulkan, dimulai dari pengambilan data yang kemudian melalui tahap preprocessing untuk memastikan bahwa teks bersih, seragam, dan siap untuk digunakan dalam proses pembelajaran model. Kemudian penerapan model berbasis Transformer, yang telah terbukti efektif dalam berbagai tugas NLP (Prottasha et al., 2022), (Patwardhan et al., 2023) dan memiliki

kemampuan yang unggul dalam memahami konteks semantik antar kata dalam kalimat, serta memiliki kemampuan untuk mempelajari konteks kata secara menyeluruh melalui mekanisme self-attention (Zhao & Yu, 2021), (Rahali & Akhloufi, 2023). Model Transformer yang digunakan yaitu *BERT (Bidirectional Encoder Representations from Transformers),* merupakan model yang dikenal mampu memahami konteks dua arah (kiri dan kanan) dari teks (Hedhili & Bouallagui, 2024). DistilBERT yang merupakan versi ringan dari BERT yang lebih efisien secara komputasi, dan merupakan hasil distilasi dari BERT yang memiliki ukuran lebih kecil dan kecepatan pemrosesan lebih tinggi, namun tetap mempertahankan sebagian besar kemampuan representatif model aslinya (Özkurt, 2024). RoBERTa *(Robustly Optimized BERT Pretraining Approach)* merupakan varian BERT yang dioptimalkan dengan jumlah data pelatihan dan langkah fine-tuning yang lebih besar dan dioptimalkan (Liu et al., 2019). XLNet yang menggunakan mekanisme *permutation-based language modelling*, yang memungkinkan pembelajaran konteks dua arah tanpa menggunakan token khusus *mask*, sehingga memperbaiki masalah pre-training atau fine-tuning discrepancy dan meningkatkan kemampuan dalam menangani dependensi luas (Yang et al., 2020). Model-model ini dilatih untuk melakukan klasifikasi teks, misalnya membedakan antara kelas *Depression* dan *SuicideWatch*. Dengan keberagaman model-model tersebut, penelitian ini memiliki landasan kuat untuk mengevaluasi dan membandingkan performa masing-masing pendekatan dalam menghasilkan representasi semantik yang kaya dan kontekstual (Ma et al., 2019; Turton et al., 2020).

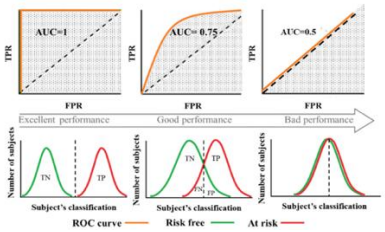
Penelitian ini juga mengevaluasi bagaimana masing-masing model bekerja untuk membuat prediksi yang akurat. Dalam pendekatan menggunakan model transformer seperti BERT, DistilBERT, RoBERTa, dan XLNet akan diuji untuk menemukan model dengan tingkat akurasi tertinggi. Hasil perbandingan ini diharapkan dapat memberikan gambaran yang menyeluruh tentang seberapa baik masing-masing teknik dalam menangani klasifikasi teks kesehatan mental. Sehingga dapat menemukan teknik yang paling cocok untuk meningkatkan akurasi model.

2.4 Evaluasi Model

Evaluasi model dilakukan menggunakan *confusion matrix,* dengan melihat performa klasifikasi berdasarkan empat komponen yaitu *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN).* Setiap baris pada matriks mewakili jumlah prediksi untuk suatu kelas, sedangkan setiap kolom mewakili jumlah data aktual dalam kelas (Sathyanarayanan, 2024). Gambar 3 menunjukkan semua kemungkinan hasil klasifikasi tersebut. Tata letak *confusion matrix* sangat berguna untuk memvisualisasikan kinerja algoritma klasifikasi. Maka akan terlihat seberapa baik model dalam mengenali kelas yang benar dan sejauh mana model melakukan kesalahan dalam prediksi. Kemudian dihitung pula metrik performa yaitu akurasi, presisi, recall, dan F1-Score (Yacouby Amazon Alexa & Axman Amazon Alexa, 2020). Oleh karena itu, *matrix confusion* menjadi alat penting untuk mengevaluasi efektivitas dan keandalan model klasifikasi yang dikembangkan dalam penelitian ini.



Gambar 3. Layout Confusion Matriks



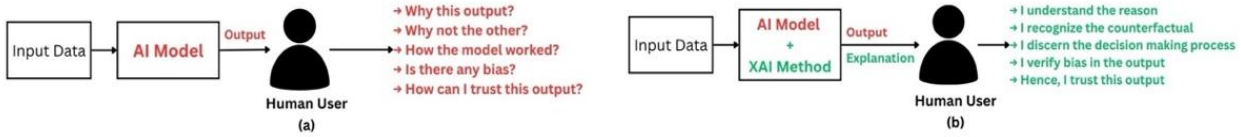
Gambar 4. Ilustrasi Metrik AUC

Digunakan juga metrik AUC *(Area Under the Curve)* untuk mengevaluasi kemampuan model dalam membedakan antara kelas secara menyeluruh. Nilai AUC yang mendekati angka satu menandakan bahwa model memiliki performa yang sangat baik dalam melakukan klasifikasi dengan tingkat kesalahan yang rendah (Google Developers, 2025). Evaluasi keseluruhan terhadap kualitas model dapat diperoleh dengan melihat contoh grafik ROC yang ditunjukkan pada Gambar 4. Nilai AUC berkisar antara 0,5 dan 1, yang berarti dengan nilai 0,5, classifier menunjukkan kinerja yang buruk, dan 1 menunjukkan classifier yang sangat baik. Karena nilai AUC yang sangat tinggi menunjukkan keandalan dan akurasi model yang tinggi, para ahli biasanya mengharapkan nilai AUC yang sangat tinggi (Aldraimli et al., 2020).

2.5 Explainable Artificial Intellegence (XAI)

XAI merupakan serangkaian teknik yang menghasilkan model yang lebih mudah dijelaskan (explainable models), dan tetap mempertahankan tingkat kinerja pembelajaran yang tinggi (akurasi prediksi), sehingga pengguna memahami, mempercayai secara tepat (Salih et al., 2025), (Mabokela et al.,

2024), (Joshi & Jain, 2025). Salah satu pihak yang mempopulerkan istilah XAI adalah Defense Advanced Research Projects Agency (DARPA) Amerika Serikat (DARPA, 2025) Definisi ini menekankan tujuan XAI dari dua sisi teknis dan berpusat pada manusia, yaitu untuk meningkatkan kejelasan model sekaligus mendukung kepercayaan pengguna dan interaksi yang efektif (Kabir et al., 2025), seperti yang ditampilkan pada gambar 5 di bawah ini.

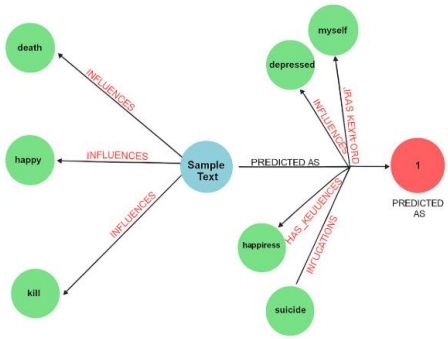


Gambar 5. Perbandingan Model AI untuk pengguna: (a) Tanpa penjelasan, sehingga kurangnya kepercayaan pengguna, (b) Dengan penjelasan metode XAI, sehingga menghasilkan keluaran yang dapat dipercaya pengguna (Kabir et al., 2025).

Dalam NLP, XAI diklasifikasikan berdasarkan cakupan penjelasan (global dan lokal) serta metode penjelasan *(post-hoc* dan *self-explaining)* (Danilevsky et al., 2020). Pendekatan global menjelaskan mekanisme model secara keseluruhan, sedangkan pendekatan lokal menjelaskan prediksi pada input tertentu. Model *self-explaining* menghasilkan penjelasan selama proses prediksi, sementara model *post-hoc* memberikan penjelasan setelah prediksi dilakukan (Jang et al., 2023). Penelitian ini menerapkan LIME untuk meningkatkan interpretabilitas model *black-box* (Abdelwahab et al., 2022), (Li et al., 2024). LIME menjelaskan prediksi dengan membentuk sampel data di sekitar instance yang dianalisis, kemudian melatih model yang dapat diinterpretasikan menggunakan bobot berdasarkan kedekatan sampel terhadap instance tersebut (Kabir et al., 2025), (Longo et al., 2023). Pendekatan ini memungkinkan identifikasi fitur yang paling berpengaruh terhadap hasil prediksi (Salih et al., 2025), (Norval & Wang, 2025; Vimbi et al., 2024).

2.6 Knowledge Graph

Penelitian ini menggunakan NetworkX pada Python untuk memvisualisasikan hubungan antar entitas dalam bentuk Knowledge Graph, yang terdiri dari node (entitas) dan edge (hubungan antar entitas). Knowledge Graph banyak diterapkan dalam kecerdasan buatan untuk mendukung proses penalaran dan pemahaman informasi (Wang et al., 2020). Dalam NLP, representasi ini menghubungkan pengetahuan terstruktur dengan informasi dalam teks sehingga membantu model memahami konteks secara lebih baik dan meningkatkan performa pada tugas seperti klasifikasi teks, analisis sentimen, question answering, dan verifikasi fakta (Kau et al., 2024).



Gambar 6. Keterkaitan Kata-Kata dalam Proses Klasifikasi

Gambar 6 menunjukkan keterkaitan antar kata dalam proses klasifikasi NLP. Node pusat berupa Sample Text (teks sampel) yang dipengaruhi oleh beberapa kata seperti death, happy, dan kill melalui relasi INFLUENCES. Berdasarkan fitur yang teridentifikasi, teks diprediksi sebagai kelas 1 (depression) melalui relasi PREDICTED AS. Kata-kata kunci seperti depressed, myself, happiness, dan suicide terhubung ke hasil prediksi melalui relasi HAS_KEYWORD dan INFLUENCES, yang menunjukkan kontribusinya dalam penentuan label klasifikasi.

3. HASIL DAN PEMBAHASAN

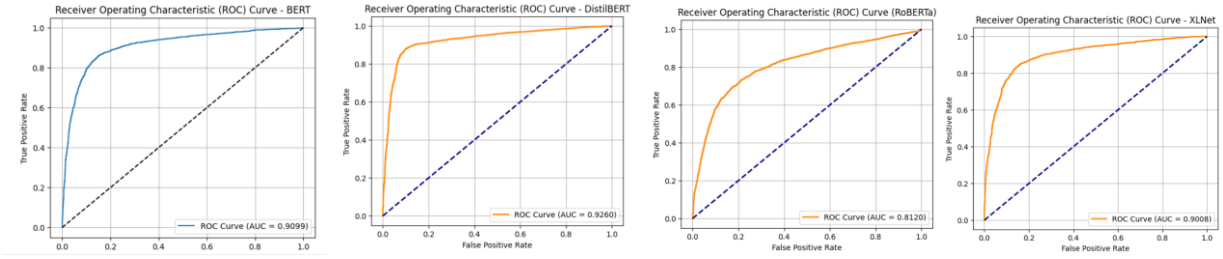
3.1 Hasil Evaluasi Model Transformer

Tabel 2 menampilkan hasil perbandingan performa empat model Transformer, yaitu BERT, DistilBERT, RoBERTa, dan XLNet dalam melakukan klasifikasi teks terkait kesehatan mental. Evaluasi dilakukan menggunakan lima metrik utama, yaitu Accuracy, F1-Score, Precision, Recall, dan AUC *(Area Under Curve).* Berdasarkan hasil yang diperoleh, DistilBERT menunjukkan performa terbaik dengan akurasi 88.91%, F1-score 88.90%, precision 88.94%, recall 88.91%, dan AUC 92.60%. Hal ini menunjukkan bahwa meskipun memiliki arsitektur yang lebih ringan dibandingkan BERT, DistilBERT mampu menghasilkan kinerja yang lebih optimal dalam mendeteksi kategori teks dengan efisiensi komputasi yang lebih tinggi. Sementara itu, BERT juga menunjukkan performa yang cukup baik dengan akurasi 85.19% dan AUC 90.99%, menandakan model ini masih mampu menangkap konteks semantik yang kompleks secara efektif. XLNet menempati posisi ketiga dengan nilai akurasi 84.04%, diikuti oleh RoBERTa yang memiliki performa paling rendah di antara keempat model dengan akurasi 75.91% dan AUC 81.20%. Secara keseluruhan, hasil ini mengindikasikan bahwa DistilBERT merupakan model Transformer paling andal dan efisien dalam penelitian ini, karena mampu memberikan keseimbangan terbaik antara akurasi, presisi, dan generalisasi model terhadap data teks yang dianalisis, dan kemampuan model dalam melakukan klasifikasi tanpa indikasi *overfitting.*

Tabel 2. Hasil Klasifikasi Menggunakan Model Transformer

Model	Accuracy	F1-Score	Precission	Recall	AUC
BERT	85.19%	85.18%	85.25%	85.19%	90.99%
DistilBERT	**88.91%**	**88.90%**	**88.94%**	**88.91%**	**92.60%**
RoBERTa	75.91%	75.82%	76.36%	75.91%	81.20%
XLNet	84.04%	84.03%	84.07%	84.04%	90.08%

Gambar 7 menampilkan kurva Receiver Operating Characteristic (ROC) untuk empat model Transformer, yaitu BERT, DistilBERT, RoBERTa, dan XLNet. Kurva ROC digunakan untuk mengevaluasi kemampuan model dalam membedakan antara kelas positif dan negatif, dengan sumbu vertikal menunjukkan *True Positive Rate (TPR)* dan sumbu horizontal menunjukkan *False Positive Rate (FPR)*. Nilai *Area Under the Curve (AUC)* menjadi indikator utama kinerja model; semakin mendekati 1, semakin baik kemampuan klasifikasi model tersebut. Berdasarkan grafik ROC, DistilBERT memperoleh nilai AUC tertinggi sebesar 0.9260, diikuti oleh BERT dengan 0.9099**,** XLNet dengan 0.9008, dan RoBERTa dengan 0.8120. Hasil ini menunjukkan bahwa model DistilBERT memiliki kemampuan paling baik dalam melakukan klasifikasi data dengan keseimbangan optimal antara sensitivitas dan spesifisitas. Perbedaan nilai AUC antar model mengindikasikan variasi kemampuan generalisasi masing-masing model terhadap data uji, yang dapat disebabkan oleh perbedaan strategi pelatihan dan kompleksitas arsitektur pada setiap model Transformer.

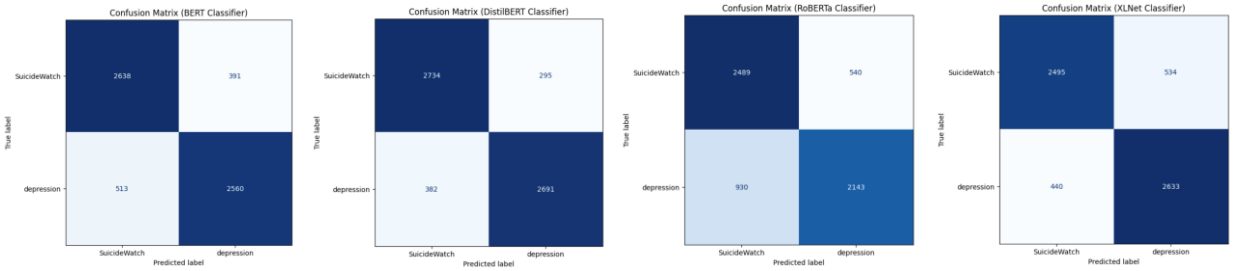


Gambar 7. Grafik ROC Pada Model Transformer

Gambar 8 menampilkan Confusion Matrix dari empat model berbasis Transformer**,** yaitu BERT**,** DistilBERT**,** RoBERTa**,** dan XLNet**,** yang digunakan untuk mengklasifikasikan dua label utama, yaitu *SuicideWatch* dan *Depression*. Confusion Matrix ini menggambarkan performa masing-masing model dalam membedakan kedua kelas tersebut berdasarkan jumlah prediksi yang benar (*True Positive* dan *True Negative*) serta kesalahan prediksi (*False Positive* dan *False Negative*). Dari hasil visualisasi tersebut, model DistilBERT menunjukkan performa paling baik dengan jumlah prediksi benar yang paling tinggi untuk kedua kelas, yaitu 2.734 data benar terklasifikasi sebagai *SuicideWatch* dan 2.691 data benar sebagai

Depression, serta jumlah kesalahan prediksi yang relatif kecil dibandingkan model lainnya. Model BERT juga memperlihatkan hasil yang kompetitif dengan distribusi prediksi yang seimbang, di mana 2.638 data benar diklasifikasikan sebagai *SuicideWatch* dan 2.560 data benar sebagai *Depression*. Sementara itu, model RoBERTa menunjukkan hasil yang lebih rendah dengan tingkat kesalahan prediksi yang lebih tinggi, khususnya pada kelas *Depression* (930 data salah prediksi). Model XLNet menampilkan hasil yang cukup stabil, dengan kemampuan klasifikasi yang seimbang antara kedua kelas, meskipun tingkat kesalahannya sedikit lebih tinggi dibandingkan DistilBERT dan BERT.

Secara keseluruhan, hasil pada Gambar 9 menunjukkan bahwa DistilBERT memiliki kinerja klasifikasi terbaik di antara keempat model Transformer, diikuti oleh BERT, XLNet, dan RoBERTa. Hal ini mengindikasikan bahwa efisiensi arsitektur DistilBERT yang lebih ringan tidak mengurangi akurasi klasifikasinya secara signifikan, bahkan mampu mengungguli model Transformer lainnya dalam tugas klasifikasi teks terkait deteksi *kesehatan mental.*

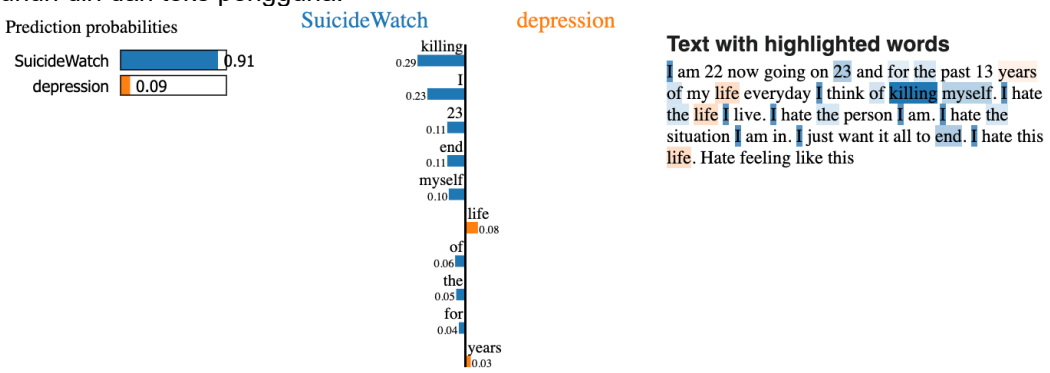


Gambar 8. Grafik *Confusion Matrix* Pada Model Transformer

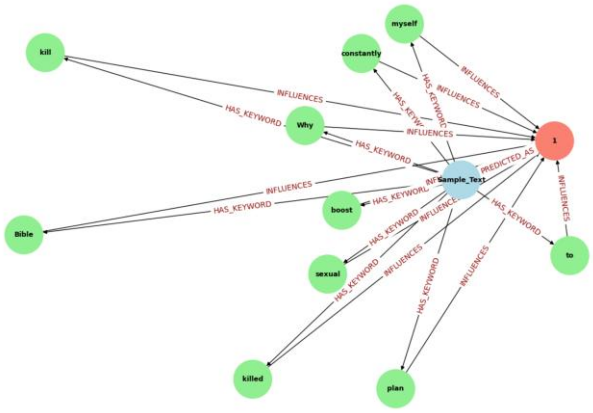
Gambar 9 menunjukkan hasil visualisasi proses interpretabilitas model menggunakan LIME untuk kategori *Suicide Watch*. Visualisasi ini bertujuan untuk menjelaskan bagaimana model Transformer memberikan keputusan klasifikasi terhadap sebuah teks yang mengandung indikasi pikiran bunuh diri. Pada bagian kiri gambar, grafik prediction probabilities menunjukkan bahwa model memprediksi teks tersebut sebagai *SuicideWatch* dengan probabilitas sebesar 0.91, sedangkan probabilitas untuk kategori *Depression* hanya sebesar 0.09**.**

Bagian tengah gambar memperlihatkan kontribusi kata-kata penting terhadap keputusan model. Kata seperti *killing*, *I*, *myself*, *end*, dan *23* memiliki bobot kontribusi positif yang tinggi terhadap label *SuicideWatch*, yang ditunjukkan dengan batang berwarna biru. Sebaliknya, kata *life* dan *years* memberikan kontribusi kecil terhadap label *Depression*. Sementara itu, pada bagian kanan gambar ditampilkan teks dengan kata-kata yang disorot sesuai pengaruhnya terhadap prediksi model**.** Warna biru menunjukkan kata yang mendorong model untuk mengklasifikasikan teks sebagai *SuicideWatch*, sedangkan warna oranye menunjukkan kata yang mendukung label *Depression*. Berdasarkan hasil ini, dapat disimpulkan bahwa model mampu mengenali konteks emosional dan ekspresi linguistik yang kuat terkait keinginan untuk mengakhiri hidup, seperti frasa *killing myself* atau *I just want it all to end*.

Secara keseluruhan, visualisasi LIME ini menunjukkan bahwa model tidak hanya memberikan hasil klasifikasi yang akurat, tetapi juga mampu memberikan penjelasan interpretatif yang transparan terhadap keputusan klasifikasinya, sehingga dapat membantu memahami bagaimana model mendeteksi indikasi pikiran bunuh diri dari teks pengguna.



Gambar 9. Hasil Visualisasi Proses LIME untuk Kategori *Suicide Watch*



Gambar 10. *Knowledge Graph* Berdasarkan Proses LIME

Gambar 10 menampilkan knowledge graph yang menggambarkan hubungan antara teks sampel, kata-kata kunci, dan hasil prediksi model. Node Sample Text terhubung dengan node prediksi kelas 1 (depresi) melalui relasi PREDICTED_AS. Kata-kata kunci seperti myself, kill, why, killed, Bible, plan, boost, dan sexual terhubung ke teks melalui relasi HAS_KEYWORD serta ke node prediksi melalui relasi INFLUENCES, yang menunjukkan kontribusinya terhadap hasil klasifikasi. Visualisasi ini membantu menjelaskan faktor-faktor yang memengaruhi keputusan model dan meningkatkan transparansi serta interpretabilitas proses klasifikasi.

4. KESIMPULAN

Penelitian ini mengembangkan kerangka kerja *post-hoc explainability* yang menggabungkan model Transformer, metode XAI berbasis LIME, serta visualisasi *Knowledge Graph* untuk meningkatkan transparansi dalam klasifikasi teks kesehatan mental. Pendekatan ini mengatasi keterbatasn sifat *black-box* pada model Transformer dengan menyediakan penjelasan yang dapat ditelusuri dan divisualisasikan secara sistematis. Hasil eksperimen menunjukan bahwa model DistilBERT memberikan performa terbaik dibandingkan model Transformer lainnya, dengan akurasi sebesar 88.91%**,** *precision* sebesar 88.94%**,** *recall* sebesar 88.91%**,** *F1-score* sebesar 88.90%**,** dan nilai *AUC* tertinggi sebesar 92.60%**.** Nilai metrik yang seimbang ini menunjukkan kemampuan model dalam melakukan klasifikasi secara stabil dan efektif pada dua kategori yaitu *Depression* dan *SuicideWatch.* Visualisasi interpretabilitas menggunakan LIME menunjukkan bahwa model mampu mengidentifikasi kata-kata dengan bobot emosional tinggi seperti *"killing"*, *"myself"*, dan *"end"* sebagai indikator utama dalam prediksi kategori *SuicideWatch*. Selanjutnya, integrasi *Knowledge Graph* memperluas hasil interpretasi dengan merepresentasikan hubungan semantik antar-entitas seperti kata kunci, konteks emosional, dan hasil prediksi dalam bentuk graf relasional. Setiap node dan edge dalam *Knowledge Graph* menggambarkan pengaruh semantik yang memperjelas hubungan antara teks masukan dan keputusan model. Dengan menggabungkan *Transformer, LIME,* dan *Knowledge Graph*, penelitian ini tidak hanya menghasilkan model dengan akurasi tinggi, tetapi juga meningkatkan transparansi, keterlacakan, dan kepercayaan terhadap sistem prediksi kesehatan mental.

Penelitian ini berkontribusi pada pengembangan system kecerdasan buatan yang lebih transparan, dapat dipercaya, dan berorientasi pada manusia *(human- centered AI)* dalam domain kesehatan mental berbasis teks. Untuk penelitian selanjutnya, evaluasi dapat diperluas dengan mengukur aspek *robustness* dan *fairness model,* serta melakukan validasi interpretasi melalui penilaian ahli domain kesehatan mental. Selain itu, eksplorasi terhadap integrasi pengetahuan domain secara lebih mendalam dapat menjadi arah pengembangan untuk meningkatkan kualitas penjelasan model.

REFERENSI

Abdelwahab, Y., Kholief, M., & Sedky, A. A. H. (2022). Justifying Arabic Text Sentiment Analysis Using Explainable AI (XAI): LASIK Surgeries Case Study. *Information (Switzerland)*, *13*(11). https://doi.org/10.3390/info13110536

Aldraimli, M., Soria, D., Parkinson, J., Thomas, E. L., Bell, J. D., Dwek, M. V., & Chaussalet, T. J. (2020). Machine learning prediction of susceptibility to visceral fat associated diseases. *Health and Technology*, *10*(4), 925–944. https://doi.org/10.1007/s12553-020-00446-1

Aljrees, T. (2024). Improving Prediction of Arabic Fake News Using ELMO's Features-Based Tri-Ensemble Model and LIME XAI. *IEEE Access*, *12*, 63066–63076. https://doi.org/10.1109/ACCESS.2024.3392297

Al-Moslmi, T., Gallofre Ocana, M., Opdahl, A. L., & Veres, C. (2020). Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access*, *8*, 32862–32881. https://doi.org/10.1109/ACCESS.2020.2973928

Alshuwaier, F. A., & Alsulaiman, F. A. (2025). Fake News Detection Using Machine Learning and Deep Learning Algorithms: A Comprehensive Review and Future Perspectives. In *Computers* (Vol. 14, Number 9). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/computers14090394

Chen, C., & Shu, K. (2023). Combating Misinformation in the Age of LLMs: Opportunities and Challenges. http://arxiv.org/abs/2311.05656

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). *A Survey of the State of* Explainable *AI for Natural Language Processing*. https://doi.org/10.18653/v1/2020.aacl-main.46

DARPA. (2025, September 10). *XAI: Explainable Artificial Intelligence*. Https://Www.Darpa.Mil/Research/Programs/Explainable-Artificial-Intelligence. https://www.darpa.mil/research/programs/explainable-artificial-intelligence

Google Developers. (2025). *Classification: ROC and AUC*. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#exercise_check_your_understanding

Hameed, S., Nauman, M., Akhtar, N., Fayyaz, M. A. B., & Nawaz, R. (2025). Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1627078

Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers*, *3*, 238–248. https://doi.org/10.1016/j.susoc.2022.03.001

Hedhili, A., & Bouallagui, I. (2024). Hybrid Approach to Explain BERT Model: Sentiment Analysis Case. *International Conference on Agents and Artificial Intelligence*, *3*, 251–259. https://doi.org/10.5220/0012318400003636

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*, *25*(1), 114–146. https://doi.org/10.1177/1094428120971683

Hu, L., Liu, Z., Zhao, Z., Hou, L., Nie, L., & Li, J. (2023). *A Survey of Knowledge Enhanced Pre-trained* Language *Models*. http://arxiv.org/abs/2211.05994

Huntington-Psychological Services. (2024). *The Latest Mental Health Statistics: What the Numbers Say* About *the State of Our Minds in 2024*. https://huntingtonpsych.com/blog/the-latest-mental-health-statistics-what-the-numbers-say-about-the-state-of-our-minds-in-2024

Jang, H., Kim, S., & Yoon, B. (2023). An eXplainable AI (XAI) model for text-based patent novelty analysis. Expert *Systems with Applications*, 1–29. https://ssrn.com/abstract=4341594

Joshi, A., & Jain, N. (2025). Improving Opinion Spam Identification with Sentiment Analysis using Explainable AI and Machine Learning. *International Journal of Engineering Trends and Applications (IJETA)*, *12*(4). www.ijetajournal.org

Kabir, S., Hossain, M. S., & Andersson, K. (2025). A Review of Explainable Artificial Intelligence from the Perspectives of Challenges and Opportunities. In *Algorithms* (Vol. 18, Number 9). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/a18090556

Kau, A., He, X., Nambissan, A., Astudillo, A., Yin, H., & Aryani, A. (2024). *Combining Knowledge Graphs and* Large *Language Models*. http://arxiv.org/abs/2407.06564

Kumar, M. (2023). *Mental Health Review*. https://www.kaggle.com/datasets/mritunjay1708/mental-health-review/data

Li, Y., Chan, J., Peko, G., & Sundaram, D. (2024). An explanation framework and method for AI-based text emotion analysis and visualisation. *Decision Support Systems*, *178*. https://doi.org/10.1016/j.dss.2023.114121

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. http://arxiv.org/abs/1907.11692

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., *Holzinger*, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2023). *Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions*. https://doi.org/10.1016/j.inffus.2024.102301

Ma, X., Wang, Z., Ng, P., Nallapati, R., & Xiang, B. (2019). *Universal Text Representation from BERT: An Empirical Study*. http://arxiv.org/abs/1910.07973

Mabokela, K. R., Primus, M., & Celik, T. (2024). Explainable Pre-Trained Language Models for Sentiment Analysis in Low-Resourced Languages. *Big Data and Cognitive Computing*, *8*(11). https://doi.org/10.3390/bdcc8110160

Mamdouh Farghaly, H., & Abd El-Hafeez, T. (2023). A high-quality feature selection method based on frequent and correlated items for text classification. *Soft Computing*, *27*(16), 11259–11274. https://doi.org/10.1007/s00500-023-08587-x

Mustafa, S., & Hama Saeed, M. (2025). Empowering text classification with NLP and explainable AI for enhanced interpretability. *Journal of Electrical Systems and Information Technology*, *12*(1), 81. https://doi.org/10.1186/s43067-025-00273-2

Norval, M., & Wang, Z. (2025). Explainable Artificial Intelligence Techniques for Speech Emotion Recognition: A Focus on XAI Models. *Inteligencia Artificial*, *28*(76), 85–123. https://doi.org/10.4114/intartif.vol28iss76pp85-123

Özkurt, C. (2024). Comparative Analysis of State-of-the-Art Q A Models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 Dataset. *Chaos and Fractals*. https://doi.org/10.69882/adba.chf.2024073

Patwardhan, N., Marrone, S., & Sansone, C. (2023). Transformers in the Real World: A Survey on NLP Applications. *Information (Switzerland)*, *14*(4). https://doi.org/10.3390/info14040242

Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M., & Baz, M. (2022). Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning. *Sensors*, *22*(11). https://doi.org/10.3390/s22114157

Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge Graph Identification. *Springer*.

Rahali, A., & Akhloufi, M. A. (2023). End-to-End Transformer-Based Models in Textual-Based NLP. *AI (Switzerland)*, *4*(1), 54–110. https://doi.org/10.3390/ai4010004

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Advanced Intelligent Systems*, *7*(1). https://doi.org/10.1002/aisy.202400304

Sathyanarayanan, S. (2024). Confusion Matrix-Based Performance Evaluation Metrics. *African Journal of* Biomedical *Research*, 4023–4031. https://doi.org/10.53555/ajbr.v27i4s.4345

Saxena, A., Santhanavijayan, A., Shakya, H. K., Kumar, G., Balusamy, B., & Benedetto, F. (2024). Nested Sentiment Analysis for ESG Impact: Leveraging FinBERT to Predict Market Dynamics Based on Eco-Friendly and Non-Eco-Friendly Product Perceptions with Explainable AI. *Mathematics*, *12*(21). https://doi.org/10.3390/math12213332

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. *Advances in Intelligent Systems and Computing*, *937*, 99–111. https://doi.org/10.1007/978-981-13-7403-6_11

Turton, J., Vinson, D., & Smith, R. E. (2020). *Deriving Contextualised Semantic Features from BERT (and* Other *Transformer Model) Embeddings*. https://doi.org/https://doi.org/10.48550/arXiv.2012.15353

Vimbi, V., Shaffi, N., & Mahmud, M. (2024). Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. In *Brain Informatics* (Vol. 11, Number 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s40708-024-00222-1

Wang, C., Liu, X., & Song, D. (2020). *Language Models are Open Knowledge Graphs*. http://arxiv.org/abs/2010.11967

Yacouby Amazon Alexa, R., & Axman Amazon Alexa, D. (2020). Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 79–91. https://doi.org/10.18653/v1/2020.eval4nlp-1.9

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *XLNet: Generalized* Autoregressive *Pretraining for Language Understanding*. http://arxiv.org/abs/1906.08237

Zhao, A., & Yu, Y. (2021). Knowledge-enabled BERT for aspect-based sentiment analysis. *Elsevier-*Knowledge-*Based Systems*, *227*. https://doi.org/10.1016/j.knosys.2021.107220