Mengungkap Opini Publik: Pendekatan *BERT-based-caused* untuk Analisis Sentimen pada Komentar Film

[1]Andi Aljabar
[2]Binti Mamluatul Karomah
[1,2]Universitas Nahdlatul Ulama Indonesia

Abstrak

BERT (Bidirectional Encoder Representations from Transformers) adalah model bahasa yang revolusioner dalam pemrosesan bahasa alami, mengandalkan encoder untuk menghasilkan representasi kontekstual dari teks input. Melalui pendekatan tokenisasi, embedding, dan attention mechanisms pada setiap layer transformer, BERT memungkinkan pemahaman hubungan antar kata secara mendalam dan bidireksional. Keunikan BERT terletak pada kemampuannya untuk memproses konteks dari kedua arah, menciptakan representasi vektor yang kaya makna. Model ini telah menjadi pionir dalam transfer learning di NLP, memungkinkan pemanfaatan representasi umum pada tugas-tugas khusus setelah proses pelatihan. Dengan demikian, BERT mengubah paradigma pemrosesan bahasa alami, membuka pintu untuk aplikasi yang lebih canggih seperti klasifikasi teks, analisis sentimen, dan pemahaman bahasa yang lebih kontekstual. Berdasarkan pendekatan itu maka penulis melakukan penelitian ini untuk melakukan analisis sentiment pada komentar-komentar film yang ada pada situs *imdb*. Adapun hasil yang diproleh berdasarkan 5.000 baris data komentar, ini menunjukkan rata-rata *accuracy* sebesar 96%, *val accuracy* sebesar 89%, *loss* sebesar 10%, dan *val loss* sebesar 37%.

Kata Kunci: *BERT, NLP, Transformer*

Abstract

BERT (Bidirectional Encoder Representations from Transformers) is a revolutionary language model in natural language processing that relies on an encoder to generate contextual representations of input text. Through tokenization, embedding, and attention mechanisms at each transformer layer, BERT enables a deep and bidirectional understanding of relationships between words. BERT's uniqueness lies in its ability to process context from both directions, creating vector representations rich in meaning. This model has pioneered transfer learning in NLP, allowing the utilization of general representations on specific tasks after the training process. Consequently, BERT has transformed the paradigm of natural language processing, opening doors to more sophisticated applications such as text classification, sentiment analysis, and contextual language understanding. Based on this approach, the author conducted research to perform sentiment analysis on movie comments available on the IMDb website. The results obtained from analyzing 5,000 lines of comment data indicate an average accuracy of 96%, a validation accuracy of 89%, a training loss of 10%, and a validation loss of 37%.

Keywords: *BERT, NLP, Transformer*

1. Pendahuluan

Di era digital saat ini, bioskop bukan satu-satunya tempat untuk berbicara tentang film. Internet, terutama situs komentar film, telah berubah menjad agora digital tanpa henti. Di sana, penonton tidak hanya menikmati cerita layar lebar, tetapi mereka juga aktif mengatakan apa yang mereka rasakan dan pikirkan. Komentar film ini, yang berkisar dari kegembiraan atas plot twist yang memukau hingga kritik pedas terhadap akting yang kaku, membentuk mozaik opini publik yang menanti untuk diuraikan (Delbrouck et al., 2020; Wang et al., 2020; Yuan et al., 2021). Tidak mudah untuk menemukan kesimpulan yang tersembunyi di balik alur suatu cerita, apalagi dengan metode tradisional. Meskipun demikian, metode tradisional untuk analisis sentimen seringkali terbentur pada keterbatasan pemahaman tentang aspek bahasa dan konteks (Zhang et al., 2022). Misalnya, ulasan negatif dapat berasal dari kesalahpahaman, sementara kata-kata pujian dapat mengandung ironi. Dalam hal ini, pendekatan berbasis BERT (*Bidirectional Encoder Representations from Transformers*) merupakan inovasi baru. Model bahasa ciptaan yang dikembangkan oleh *AI* buatan *Google* bernama BERT yang mampu menganalisis kata dan memahami hubungan antar kata dan konteks kalimat(Alaparthi & Mishra, n.d.; Boukabous & Azizi, 2022; Deepa & Tamilarasi, 2021). Kemampuan inilah yang memungkinkan BERT lebih tepatnya *bert-based-caused* untuk berfungsi sebagai media untuk mengungkapkan opini publik di balik komentar-komentar pada film. Tujuan penelitian ini adalah bagaimana penerapan metode *transformer* dengan model *bert-based-caused* pada komentar-komentar film mampu memprediksi komentar-komentar tersebut apakah bersifat positif atau negative.

2. Tinjauan Pustaka

Bert merupakan model salah satu model kontekstual yang dirancang diatas transformer model, Oleh karena itu bert mampu menangkap dan mengklasifikasikan suatu kata dalam kontks dan level tertentu sesuai dengan struktur model yang dirancang (Geni et al., 2023). Tokenisasi juga merupakan salah proses penting yang dilakukan untuk memecah teks atau kalimat menjadi bagian yang lebih kecil yang disebut "sub-word" atau "token-pieces" (Alaparthi & Mishra, n.d.; Deepa & Tamilarasi, 2021; Yüksel et al., 2019). Kata-kata yang tidak umum atau baru yang mungkin tidak ada dalam kosakata awal dapat diatasi dengan lebih fleksibel dengan bantuan ini. Tokenisasi pada BERT juga melibatkan token awal dan token khusus seperti [CLS] (Klasifikasi) dan [SEP] (Separator), yang ditambahkan untuk tujuan tertentu. Proses ini dilakukan pada tingkat kata dan karakter, memungkinkan model untuk memahami konteks dan hubungan antar-karakter. Metode ini membantu BERT mengatasi kata-kata yang sulit atau morfologi yang berbeda sambil mempertahankan struktur informasi yang penting.

Tabel 1. Perbandingan Peneltian.

Jurnal	Metode	Akurasi
Model for Sentiment Analysis of Micro-blogs Integrating Topic Model and BERT	T-Bert	90.81%
Multi-class sentiment analysis of urdu text using multilingual BERT	M-Bert	81.49%
Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models	Indo-Bert	83.50%

Berdasarkan keterangan yang diperoleh dari tabel 1, ada beberapa model bert yang digunakan, pada penelitian Model for Sentiment Analysis of Micro-blogs Integrating Topic Model and BERT dengan menggunakan metode T-Bert dan dataset dari social media menghasilkan akurasi 90.81% (Palani et al., 2021). Penelitian selanjutnya yaitu Multi-class sentiment analysis of urdu text using multilingual BERT dengan menggunakan metode multilingual Bert (M-Bert) dengan menggunakan dataset yang berbahasa urdu dengan akurasi 81.49% (Khan et al., 2022). Selanjutnya pada penelitian Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models dengan dataset dari twitter menggunakan metode Indo-Bert dan akurasi 83.50% (Geni et al., 2023). Berdasarkan ke-3 rujukan utama tersebut penulis berinisiatif untuk melakukan riset sentiment analisis terkait komentar-komentar pada film menggunakan model *bert-based-caused.*

2.1. Tokenize

Dengan menggunakan pendekatan WordPiece dalam BERT, tokenisasi adalah proses penting yang dilakukan pada tingkat kata dan karakter yang bertujuan untuk memecah teks atau kalimat menjadi bagian yang lebih kecil yang disebut "sub-word" atau "token-pieces" (Alaparthi & Mishra, n.d.; Deepa & Tamilarasi, 2021; Yüksel et al., 2019). Kata-kata yang tidak umum atau baru yang mungkin tidak ada dalam kosakata awal dapat diatasi dengan lebih fleksibel dengan bantuan ini.

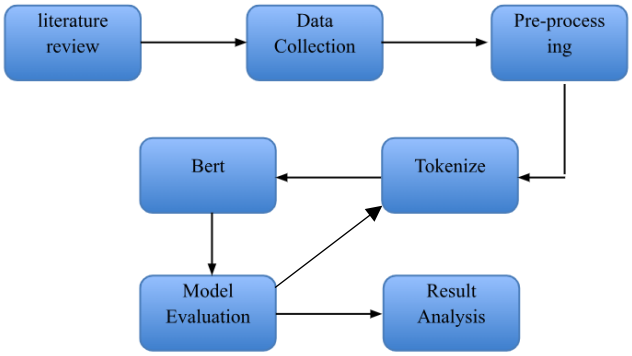
Tokenisasi pada BERT juga melibatkan token awal dan token khusus seperti [CLS] (Klasifikasi) dan [SEP] (Separator), yang ditambahkan untuk tujuan tertentu. Proses ini dilakukan pada tingkat kata dan karakter, memungkinkan model untuk memahami konteks dan hubungan antar-karakter. Metode ini membantu BERT mengatasi kata-kata yang sulit atau morfologi yang berbeda sambil mempertahankan struktur informasi yang penting (Chiorrini et al., 2021; Hutama & Suhartono, 2022; Karayiğit et al., 2022).

2.2. Encoder

Untuk mengubah input teks menjadi representasi vektor yang penuh dengan informasi, encoder pada BERT (Bidirectional Encoder Representations from Transformers) adalah komponen arsitektur model (Tesfagergish et al., 2022). Dalam konteks transformasi bahasa, encoder BERT terdiri dari beberapa lapisan transformasi. Proses encoding BERT berbeda, terutama karena dilakukan secara mendalam dan bidireksional. Cara kerjanya adalah dengan dengan mengubah kata-kata dalam sebuah kalimat menjadi vektor numerik yang menunjukkan hubungan dan konteks antar kata secara rinci. Dalam konteks model transformer, encoder bertanggung jawab untuk melakukan operasi perhatian, yang memungkinkan model untuk mempertimbangkan seluruh konteks kalimat. Dalam encoder BERT, setiap layer terdiri dari dua sublayer utama: *multi-head self-attention* dan *fully connected feedforward networks* (Boukabous & Azizi, 2022). *Multi-head self-attention* memungkinkan model untuk memberikan bobot yang tepat pada kata-kata yang relevan dalam kalimat, sementara *fully connected feedforward networks* bertanggung jawab untuk menghasilkan representasi yang lebih abstrak dan kaya informasi.

3. Riset Metodologi

Untuk memulai penelitian ini, pertama-tama literatur review secara menyeluruh, tujuannya adalah untuk mendapatkan pemahaman yang lebih baik tentang bidang analisis sentimen saat ini. Data kemudian dikumpulkan untuk model pendukung. Selanjutnya, untuk memastikan bahwa tujuan penelitian tercapai dan memasukkan berbagai sumber dan sentimen, kumpulan data terdiversifikasi yang terdiri dari sentimen berlabel dikumpulkan. Sebelum pengumpulan data, prosedur pra-pemrosesan seperti tokenisasi dan pengurangan kebisingan dilakukan. Pemilihan model Bert yang tepat untuk memulai proses pelatihan didahului dengan pembagian kumpulan data menjadi kumpulan pelatihan, validasi, dan pengujian. Setelah itu, model disesuaikan sesuai kebutuhan. Untuk mengevaluasi performa model, seperti akurasi dan kerugian, metrik evaluasi dibuat. Untuk lebih jelasnya dapat dilihat pada gambar 1 tentang metodologi penelitian.



Gambar 1 Metodologi Penelitian

3.1. Data Collecting

Data dikumpulkan menggunakan metode *web scraping* pada website *imbd.* Proses *collectiong* data dilakukan secara acak dengan tujuan menghindari subjektifitas pengambilan data.

3.2. Pre-Processing

Setelah data terkumpul sesuai dengan target 5000 rows + 30% proses selanjutnya yaitu *pre-processing*. Data yang dikumpulkan sejumlah 7632 rows yang selanjutnya dilakukan teknik cleaning data. Proses ini dilakukan karena ketika proses data *collecting* terdapat data yang tidak *complete* atau rusak. Dengan kata lain pre-processing dilakukan agar ketika proses trining, validation dan testing nantinya bisa mendapatkan akurasi yang sesuai dengan harapan.

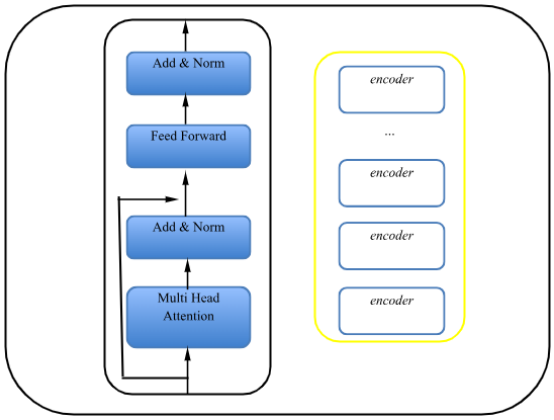
3.3. Tokenize

Tokenize merupakan proses penomoran pada suatu pada metode transformer. Proses ini dilakukan dengan memberikan token atau nomor pada setiap kata. Sebagai contoh seperti pada kata ”*I love this movie*”. Kata tersebut bersifat positif, misalkan kata I diberi token 1, love diberi token 2, this diberi token 3, dan movie diberi token 4. Sehingga masing-masing kata tersebut sudah memeiliki token.

3.4. *Bert-based-cased model*

BERT (Bidirectional Encoder Representations from Transformers) membuat representasi kontekstual dari input teks dengan menggunakan encoder. Proses dimulai dengan tokenisasi teks, yang membentuk token kecil seperti kata, sub-kata, atau karakter. Selanjutnya, proses embedding digunakan untuk mengubah tiap token menjadi vektor kata. Proses ini menggunakan embedding yang telah dilatih sebelumnya oleh model. Kemudian, encoder BERT, yang terdiri dari berbagai layer transformer, bekerja pada token-token tersebut. Untuk membuat representasi kontekstual, setiap lapisan melakukan operasi multi-head self-attention dan full-connected feedforward networks. BERT unik karena dapat memahami konteks secara bidireksional dan menangkap hubungan kontekstual yang lebih kuat dengan memperhatikan token kiri dan kanan. Proses ini berulang pada setiap layer encoder yang ditumpuk, yang menghasilkan pemahaman teks input yang semakin hierarkis. BERT menghasilkan representation yang digabungkan dari token [CLS] (Alaparthi & Mishra, n.d.; Boukabous & Azizi, 2022), yang dapat digunakan untuk klasifikasi dan analisis sentimen. Setelah representasi vektor dibuat, model dapat disesuaikan untuk tugas tertentu. Secara keseluruhan, BERT memiliki pemahaman kontekstual dan bidireksional yang kuat, yang membuatnya sangat baik untuk pemrosesan bahasa alami dengan kemampuan transfer learning yang kuat (Tabinda Kokab et al., 2022).

Berdasarkan contoh yang dipaparkan pada 3.3 dan dianggap kata tersebut bersifat posirif maka selanjutnya adalah melakukan encoder. Misal masing-masing token 1, 2, 3, dan 4 berada pada posisi yang berdekatan [1, 2, 3, 4] maka model bert akan memprediksi bahwa encoding tersebut atau kalimant tersebut bersifat positif. Hal ini diprediksi berdasarkan apa yang telah dipelajari dari model *bert* tersebut. Untuk lebih jelasnya perhatikan gambar 2 *Bert Model.*



Gambar 2 *Bert Model*

3.5. Evaluasi Model

Pada bagian ini penulis mencari model trining dan validation bert terbaik dalam memprediksi komentar-komentar pada film. Maka dipilihlah hasil dengan salah satu variable penilian adalah *epoch = 5* dan beberapa perbandingan rata-rata *acc* dan *loss acc* dengan hasil sebagai berikut

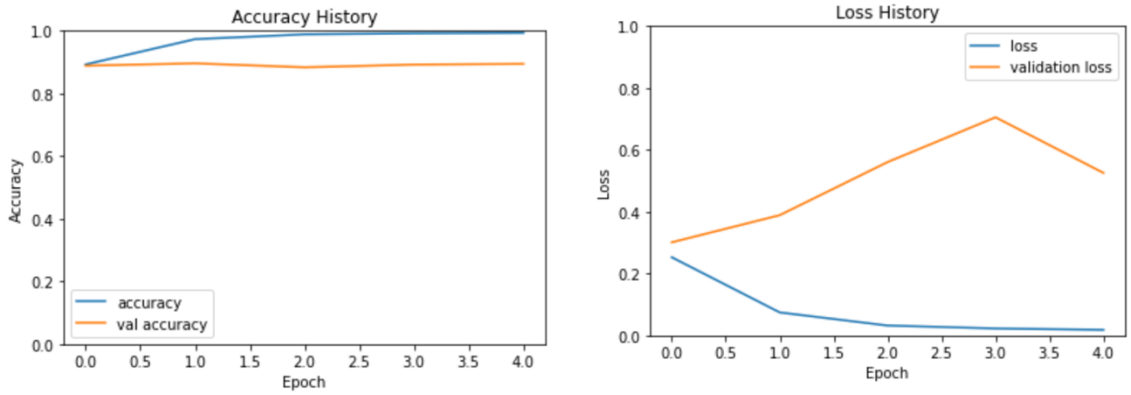
Tabel 2. Perbandingan Evaluasi Model.

Epoch	*acc*	*loss acc*
3	91%	17%
4	93%	14%
5	96%	10%
6	98%	25%
7	99%	37%

Berdasarkan keterngan tabel 2 maka dipilihlah penggunaan *epoch = 5* dimana rata-rata *acc* dan *loss* hampir berbanding lurus dan merupakan kombinasi yang paling ideal dari ke 5 percobaan tersebut.

4. Hasil

Berdasarkan riset dan metodologi penelitian maka dengan menggunakan lebih dari 5.000 baris data, performa bert model pada sentiment analisis pada kasus ini menunjukkan hasil yang cukup baik dengan rata-rata accuracy sebesar 96%, val accuracy sebesar 89%, loss sebesar 10%, dan val loss sebesar 37%. Hasil tersebut diproleh dari jumlah keseluruhan proses per-detiknya dibagi dengan jumlah proses epoch ke 1, 2, 3, 4 dan 5. Untuk lebih jelasnya perhatikan gambar 3



Gambar 3 Hasil Trining dan Valdation

Pada penelitian kali ini terdapat data yang unik. Setelah melakukan trining dan validasi penulis berinisitaif untuk melakukan testing model *bert-based-cased* dengan 5 sample inputan. Dan hasilnya sebagai berikut :

```
I don't like this movie :  Negative
this movie really amizing :  Negative
the title does not match with the movie flow :  Negative
the movie is really scared me :  Negative
I love this movie :  Positive
```

Gambar 4 Hasil 5 data testing

Berdasarkan gambar 4 dijelaskan bahwa terdapat 5 data rows sebagai inputan untuk melakukan testing model, yang unik adalah pada data ”*this movie really amezing*”, yang dinilai sebagai bentuk negative.

5. Kesimpulan

Secara umum, BERT (Bidirectional Encoder Representations from Transformers) adalah model bahasa yang mengandalkan encoder untuk menghasilkan representasi kontekstual yang kaya dari teks yang dimasukkan. Ini dapat memahami hubungan antar kata secara menyeluruh dan menghasilkan pemahaman hierarkis yang kompleks melalui penggunaan layer transformer yang mendalam dan bidireksional. Berdasarkan hasil pada pembahasan sebelumnya maka disimpulkan bahwa bert mampu melakukan prediksi dan analisis sentiment pada komentar-komentar film.

Daftar Pustaka

Alaparthi, S., & Mishra, M. (n.d.). *Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey*.

Boukabous, M., & Azizi, M. (2022). Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*(2), 1131–1139. https://doi.org/10.11591/ijeecs.v25.i2.pp1131-1139

Chiorrini, A., Diamantini, C., Mircoli, A., & Potena, D. (2021). *Emotion and sentiment analysis of tweets using BERT*. https://code.google.com/archive/p/word2vec/

Deepa, M. D., & Tamilarasi, A. (2021). Bidirectional Encoder Representations from Transformers (BERT) Language Model for Sentiment Analysis task: Review. In *Turkish Journal of Computer and Mathematics Education* (Vol. 12, Issue 7).

Delbrouck, J.-B., Tits, N., Brousmiche, M., & Dupont, S. (2020). *A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis*. https://doi.org/10.18653/v1/2020.challengehml-1.1

Geni, L., Yulianti, E., & Sensuse, D. I. (2023). Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika (JITEKI)*, *9*(3), 746–757. https://doi.org/10.26555/jiteki.v9i3.26490

Hutama, L. B., & Suhartono, D. (2022). Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic. *Informatica (Slovenia)*, *46*(8), 81–90. https://doi.org/10.31449/inf.v46i8.4336

Karayiğit, H., Akdagli, A., & Acı, Ç. İ. (2022). BERT-based Transfer Learning Model for COVID-19 Sentiment Analysis on Turkish Instagram Comments. *Information Technology and Control*, *51*(3), 409–428. https://doi.org/10.5755/j01.itc.51.3.30276

Khan, L., Amjad, A., Ashraf, N., & Chang, H. T. (2022). Multi-class sentiment analysis of urdu text using multilingual BERT. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-09381-9

Palani, S., Rajagopal, P., & Pancholi, S. (2021). *T-BERT -- Model for Sentiment Analysis of Micro-blogs Integrating Topic Model and BERT*. http://arxiv.org/abs/2106.01097

Tabinda Kokab, S., Asghar, S., & Naz, S. (2022). Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, *14*. https://doi.org/10.1016/j.array.2022.100157

Tesfagergish, S. G., Kapočiūtė-Dzikienė, J., & Damaševičius, R. (2022). Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning. *Applied Sciences (Switzerland)*, *12*(17). https://doi.org/10.3390/app12178662

Wang, Z., Wan, Z., & Wan, X. (2020). TransModality: An End2End Fusion Method with Transformer for Multimodal Sentiment Analysis. *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, 2514–2520. https://doi.org/10.1145/3366423.3380000

Yuan, Z., Li, W., Xu, H., & Yu, W. (2021). Transformer-based Feature Reconstruction Network for Robust Multimodal Sentiment Analysis. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, 4400–4407. https://doi.org/10.1145/3474085.3475585

Yüksel, A. E., Türkmen, Y. A., Özgür, A., & Altınel, A. B. (2019). Turkish tweet classification with transformer encoder. *International Conference Recent Advances in Natural Language Processing, RANLP*, *2019-September*, 1380–1387. https://doi.org/10.26615/978-954-452-056-4_158

Zhang, T., Gong, X., & Chen, C. L. P. (2022). BMT-Net: Broad Multitask Transformer Network for Sentiment Analysis. *IEEE Transactions on Cybernetics*, *52*(7), 6232–6243. https://doi.org/10.1109/TCYB.2021.3050508