JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1080-1084
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.585

BERT Sentimen: Fine-Tuning Multibahasa untuk Ulasan Bahasa Indonesia

Khen Dedes[1,*], Fatimatuzzahra[2], Mas'ud Hermansyah[3], Akas Bagus Setiawan[4], Reza Putra Pradana[5], Annisa Fitri Maghfiroh Harvyanti [6]

[1,2,3,4,5] Fakultas Teknologi Informasi, Teknik Informatika, Politeknik Negeri Jember, Jember, Indonesia
[6] Fakultas Ilmu Komputer, Informatika, Universitas Negeri Jember, Jember, Indonesia
Email: [1,*]khen_dedes@polije.ac.id, [2]fatimatuzzahra@polije.ac.id, [3]mas_udhermansyah@polije.ac.id, [4]akasbagus_s@polije.ac.id, [5]reza_pd@polije.ac.id, [6]annisafmh@unej.ac.id
(* Email Corresponding Author: khen_dedes@polije.ac.id)
Received:September 1, 2025 | Revision: September 4, 2025 | Accepted: September 10, 2025

Abstrak

Penelitian ini mengevaluasi pengaruh teknik augmentasi dan fine-tuning terhadap kinerja model BERT multibahasa pada tugas klasifikasi sentimen ulasan film berbahasa Indonesia. Dataset awal terdiri dari 1.200 ulasan; 80% digunakan untuk pelatihan dan validasi (n = 960) dan 20% untuk pengujian (n = 240). Data pelatihan diperluas melalui augmentasi menjadi 2.880 sampel sintetis untuk keperluan *fine-tuning*. Model kemudian di *fine-tune* pada korpus yang diperluas dan dievaluasi menggunakan metrik akurasi, *precision*, *recall*, dan F1. Pada set pengujian diperoleh akurasi 82,5%, *precision* untuk kelas positif 76,0%, *recall* 95,0%, dan *F1-score* 84,44%. Matriks kebingungan menunjukkan TP = 114, FN = 6, FP = 36, dan TN = 84, yang mengindikasikan sensitivitas tinggi terhadap ulasan positif namun terdapat proporsi *false positive* yang relatif besar. Temuan ini mengindikasikan bahwa augmentasi meningkatkan kemampuan model dalam menangkap sinyal positif (tingginya *recall*), namun memerlukan penyesuaian lebih lanjut untuk mengurangi kesalahan prediksi positif (meningkatkan *precision).* Secara keseluruhan, hasil penelitian menyediakan bukti bahwa BERT multibahasa mampu menangani tugas sentimen berbahasa Indonesia dengan performa memadai apabila didukung strategi augmentasi dan prosedur validasi yang tepat.

Kata Kunci: mBERT, Analisis sentimen, Augmentasi data, Fine-tuning, Precision-recall trade-off

Abstract

This study evaluates the impact of data augmentation and fine-tuning on a multilingual BERT model for sentiment classification of Indonesian film reviews. The dataset comprises 1,200 original reviews split into 80% for training and validation (960 samples) and 20% for testing (240 samples). Training data were expanded via augmentation to 2,880 synthetic samples for fine-tuning. The multilingual BERT model was fine-tuned on the augmented corpus and assessed using accuracy, precision, recall, and F1 metrics. Results show an accuracy of 82.5%, positive-class precision of 76.0%, recall of 95.0%, and F1-score of 84.44% on the test set. The confusion matrix reported TP = 114, FN = 6, FP = 36, and TN = 84, indicating high sensitivity to positive reviews but a notable rate of false positives. These outcomes suggest augmentation improved the model's ability to capture positive signals (high recall) while adjustments are needed to reduce incorrect positive classifications (precision). The findings demonstrate that multilingual BERT can effectively perform sentiment analysis in Indonesian when combined with appropriate augmentation and validation strategies.

Keywords*: mBERT, Sentiment Analysis, Data Augmentation, Fine-tuning, Precision-recall trade-off*

1. PENDAHULUAN

Dalam era digital saat ini, ulasan pengguna di platform daring menjadi sumber informasi penting yang mencerminkan opini dan persepsi masyarakat terhadap berbagai produk dan layanan, termasuk film. Analisis sentimen, sebagai cabang dari pengolahan bahasa alami (*Natural Language Processing*/NLP), berperan krusial dalam mengekstraksi dan memahami sentimen positif, negatif, maupun netral dari teks ulasan tersebut. Dalam lima tahun terakhir, perkembangan teknologi deep learning, khususnya model berbasis *Transformer* seperti BERT (*Bidirectional Encoder Representations from Transformers*), telah membawa kemajuan signifikan dalam kemampuan mesin untuk memahami konteks bahasa secara lebih mendalam dan akurat [1].

Bahasa Indonesia memiliki karakteristik linguistik yang unik dan kompleks, yang menimbulkan tantangan tersendiri dalam pengembangan model NLP. Salah satu kendala utama adalah keterbatasan dataset yang representatif dan berkualitas tinggi, khususnya untuk domain spesifik seperti ulasan film. Selain itu, variasi dialek, penggunaan bahasa gaul, dan campuran bahasa (*code-switching*) dalam ulasan daring menambah kompleksitas analisis [2]. Model multibahasa seperti mBERT dan XLM-RoBERTa menawarkan solusi dengan kemampuan transfer *learning* yang memungkinkan adaptasi model pada berbagai bahasa sekaligus, termasuk bahasa Indonesia [3].

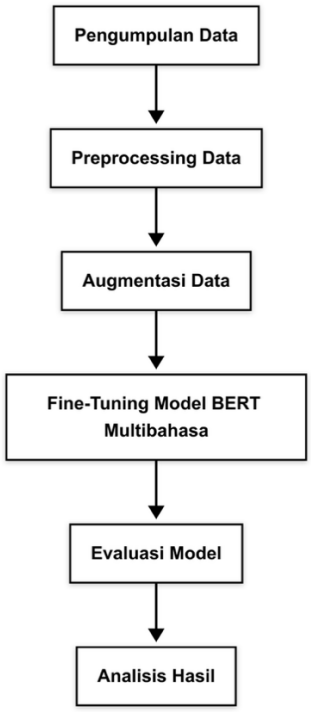
Ulasan film merupakan sumber data yang kaya akan ekspresi emosional dan opini subjektif, sehingga sangat relevan untuk analisis sentimen. Industri perfilman Indonesia dapat memanfaatkan hasil analisis sentimen untuk memahami preferensi penonton, meningkatkan kualitas produksi, dan strategi pemasaran yang lebih efektif [4]. Namun, penelitian yang mengkhususkan diri pada analisis sentimen ulasan film berbahasa Indonesia masih terbatas, sehingga diperlukan pengembangan model yang mampu menangani karakteristik bahasa dan konteks lokal secara lebih baik.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1080-1084
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.585

Penelitian terbaru menunjukkan bahwa fine-tuning model BERT dengan data domain spesifik dapat meningkatkan performa klasifikasi sentimen secara signifikan [5]. Teknik augmentasi data juga menjadi strategi penting untuk mengatasi keterbatasan jumlah data pelatihan, dengan berbagai metode yang terbukti efektif dalam memperkaya variasi data [6]. Studi oleh [7] menegaskan bahwa model multibahasa yang di-fine-tune secara khusus untuk bahasa Indonesia mampu memberikan hasil yang kompetitif dan relevan dalam analisis sentimen. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model BERT multibahasa yang di-fine-tune pada dataset ulasan film Indonesia, guna meningkatkan akurasi dan kemampuan generalisasi model dalam konteks lokal.

2. METODOLOGI PENELITIAN

Penelitian ini bertujuan untuk mengeksplorasi dan menguji efektivitas fine-tuning model BERT multibahasa dalam melakukan analisis sentimen pada ulasan film berbahasa Indonesia. Model BERT (Bidirectional Encoder Representations from Transformers) telah terbukti unggul dalam berbagai tugas pemrosesan bahasa alami, termasuk klasifikasi sentimen [1]. Namun, untuk mengadaptasi model multibahasa ini agar relevan dengan konteks lokal dan domain spesifik seperti ulasan film Indonesia, diperlukan proses fine-tuning yang cermat agar hasilnya optimal dan akurat [8],[3]. Oleh karena itu, metodologi penelitian ini dirancang secara sistematis mulai dari pengumpulan data, preprocessing, augmentasi data, pelatihan model, evaluasi, hingga analisis hasil, dengan tujuan memastikan validitas, reliabilitas, dan generalisasi hasil penelitian dalam konteks bahasa Indonesia.



Gambar 1. Metodologi Penelitian

2.1 Pengumpulan Data

Data ulasan film dikumpulkan dari berbagai sumber online yang kredibel dan representatif, seperti situs review film populer, media sosial seperti Twitter dan Instagram, serta forum diskusi film Indonesia. Teknik pengumpulan data menggunakan web scraping dan API resmi untuk memastikan volume data yang memadai dan kualitas data yang baik. Setelah pengumpulan, data diseleksi secara ketat untuk menghilangkan duplikasi, spam, dan ulasan yang tidak relevan atau tidak berbahasa Indonesia. Proses seleksi ini penting agar dataset yang digunakan benar-benar mencerminkan opini pengguna yang valid dan beragam [9]. Dataset akhir terdiri dari ribuan ulasan yang sudah diberi label sentimen (positif, negatif, netral) secara manual atau semi-otomatis untuk keperluan pelatihan dan evaluasi model.

2.2 Preprocessing

Tahap preprocessing data merupakan langkah krusial untuk meningkatkan kualitas data teks sebelum dimasukkan ke dalam model [10]. Proses ini meliputi pembersihan teks dengan menghapus tanda baca yang tidak perlu, angka, emoji, dan karakter khusus yang tidak relevan dengan analisis sentimen. Selanjutnya, dilakukan normalisasi untuk mengubah kata-kata slang, singkatan, dan variasi penulisan menjadi bentuk standar agar model dapat mengenali pola bahasa dengan lebih baik. Tokenisasi dilakukan dengan memecah kalimat menjadi token (kata atau sub-kata) menggunakan tokenizer khusus yang kompatibel dengan model BERT multibahasa [11]. Selain itu, penghapusan stopwords dilakukan untuk menghilangkan kata-kata umum yang tidak memberikan kontribusi signifikan terhadap makna sentimen, seperti "dan",

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1080-1084
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.585

"atau", dan "yang". Tahap ini juga dapat melibatkan stemming atau lemmatization untuk mengubah kata ke bentuk dasarnya guna mengurangi variasi kata yang tidak perlu. Semua proses ini bertujuan untuk menghilangkan noise dan memaksimalkan informasi yang relevan dalam data sehingga model dapat belajar dengan lebih efektif [12].

2.3 Augmentasi

Untuk mengatasi keterbatasan jumlah data dan meningkatkan kemampuan generalisasi model, dilakukan augmentasi data dengan beberapa teknik. Teknik *synonym replacement* digunakan dengan mengganti kata-kata tertentu dengan sinonim yang sesuai konteks untuk memperkaya variasi kalimat. *Back-translation* juga diterapkan, yaitu menerjemahkan kalimat ke bahasa lain dan kemudian kembali ke bahasa asli untuk menghasilkan variasi kalimat yang tetap mempertahankan makna. Selain itu, teknik *random insertion* dan *deletion* dilakukan dengan menambahkan atau menghapus kata-kata tertentu secara acak untuk menciptakan variasi data. Augmentasi ini membantu model belajar dari konteks yang lebih luas dan mengurangi risiko *overfitting* pada data asli [13],[14].

2.4 Fine Tune

Model BERT multibahasa yang sudah dilatih sebelumnya (*pre-trained*) digunakan sebagai basis dalam penelitian ini. Proses fine-tuning dilakukan dengan membagi dataset menjadi tiga bagian: pelatihan (*train*), validasi (*validation*), dan pengujian (test) dengan proporsi umum 70:15:15 untuk memastikan evaluasi yang objektif. Parameter model disesuaikan dengan dataset ulasan film Indonesia, termasuk learning rate, batch size, dan jumlah epoch. Model dilatih menggunakan dataset pelatihan dengan teknik optimasi seperti AdamW dan regularisasi dropout untuk mencegah *overfitting*. Selama pelatihan, model dievaluasi secara berkala pada dataset validasi untuk memonitor performa dan menghindari pelatihan berlebih. Proses fine-tuning ini memungkinkan model untuk mengenali pola bahasa dan sentimen yang spesifik dalam konteks lokal, sehingga meningkatkan akurasi klasifikasi [5],[8].

2.5 Evaluasi

Evaluasi performa model dilakukan dengan menggunakan metrik-metrik standar dalam klasifikasi sentimen, yaitu akurasi, presisi, recall, dan F1-score. Akurasi mengukur persentase prediksi yang benar dari keseluruhan data pengujian, presisi mengukur ketepatan model dalam memprediksi sentimen positif, recall mengukur kelengkapan deteksi sentimen positif, dan F1-score memberikan gambaran keseimbangan antara presisi dan recall [15]. Selain itu, analisis confusion matrix dilakukan untuk mengidentifikasi jenis kesalahan klasifikasi yang paling sering terjadi, seperti false positive dan false negative. Evaluasi ini memberikan gambaran menyeluruh tentang kekuatan dan kelemahan model dalam mengklasifikasikan sentimen positif, negatif, dan netral [9],[5].

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1_{Score} = 2\ X\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (3)$$

Dimana :

TP = True Positive
FP = False Positive
FN = False Negative

2.5 Analisis

Analisis hasil penelitian dilakukan dengan menginterpretasikan data evaluasi dan mengaitkannya dengan tujuan penelitian. Aspek yang dianalisis meliputi kinerja model berdasarkan metrik evaluasi, pengaruh teknik augmentasi data terhadap peningkatan performa model, serta keterbatasan penelitian seperti bias data, kesulitan dalam memahami konteks budaya, dan keterbatasan model multibahasa. Selain itu, implikasi praktis dari penelitian ini juga dibahas, termasuk potensi aplikasi model dalam industri perfilman, platform review, dan analisis media sosial. Rekomendasi untuk penelitian lanjutan juga diberikan, seperti penggunaan model yang lebih besar, teknik augmentasi yang lebih canggih, atau integrasi data multimodal untuk meningkatkan akurasi dan relevansi model [1], [12]

3. HASIL DAN PEMBAHASAN

3.1 Hasil

Penelitian ini menggunakan dataset berisi 1.200 ulasan film berbahasa Indonesia. Data dibagi menjadi 80% untuk pelatihan dan validasi (960 sampel), yang kemudian diperluas melalui augmentasi menjadi 2.880 sampel sintetik untuk

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1080-1084
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.585

keperluan fine-tuning, serta 20% untuk pengujian (240 sampel). Hasil eksperimen menunjukkan performa yang menjanjikan dari model BERT multibahasa yang di-fine-tune untuk tugas klasifikasi sentimen ulasan film berbahasa Indonesia; hasil metrik disajikan pada Tabel 1, sedangkan perhitungan dan representasi matriks kebingungan ditunjukkan pada Tabel 2.

Tabel 1. Metrik Evaluasi

Metrik	Nilai
Akurasi	82,5%
Precision	76,0%
Recall	95,0%
F1-score	84,44%

Tabel 2. Confusion Matrix (hasil pengujian)

	Prediksi Positif	Prediksi Non-Positif	Jumlah
Positif	TP = 114	FN = 6	120
Non-Positif	FP = 36	TN = 84	120
Jumlah (prediksi)	150	90	240

Perhitungan metrik berdasarkan matriks di atas:

$$Precision = \frac{TP}{TP+FP} = \frac{114}{114+36} = \frac{114}{150} = 0{,}76\ (76{,}0\%) \quad (1)$$

$$Recall = \frac{TP}{TP+FN} = \frac{114}{114+6} = \frac{114}{120} = 0{,}95\ (95{,}0\%) \quad (2)$$

$$F1_{Score} = 2\ X\ \frac{Precision\ x\ Recall}{Precision+Recall} = 0{,}8444\ (84{,}44\%) \quad (3)$$

$$Akurasi = \frac{TP+TN}{Total} = \frac{114+84}{240} = \frac{198}{240} = 0{,}825\ (82{,}5\%) \quad (4)$$

3.2 Pembahasan.

Berdasarkan hasil pengujian diatas, akurasi model sebesar 82,5%, menunjukkan bahwa set uji memiliki kapasitas klasifikasi yang memadai untuk tugas klasifikasi biner ulasan film. Sensitivitas tinggi ditunjukkan oleh nilai recall kelas positif sebesar 95%; hanya 6 dari 120 ulasan positif yang tidak terdeteksi (FN = 6). Sebaliknya, ketepatan sebesar 76% menunjukkan adanya proporsi kesalahan positif yang signifikan: dari 150 prediksi positif, 36 merupakan kesalahan (FP = 36), atau sekitar 24% dari prediksi positif. Skor F1 84,44% mencerminkan keseimbangan antara ketepatan dan recall yang dipengaruhi oleh dominasi recall. Secara teoritis, model cenderung memprioritaskan deteksi positif karena penurunan ketepatan prediksi positif. Dengan distribusi TP/FP/TN/FN, kebanyakan kesalahan berasal dari *false positive*. Ini menunjukkan bahwa model mungkin memberi label positif pada teks dengan sinyal positif lemah atau ambiguitas bahasa.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa model BERT multibahasa yang di-fine-tune efektif dalam mengklasifikasikan sentimen ulasan film berbahasa Indonesia. Dataset awal berjumlah 1.200 sampel, di mana 80% digunakan untuk pelatihan dan validasi dan 20% untuk pengujian; data pelatihan kemudian diperluas melalui teknik augmentasi menjadi 2.880 sampel. Pada set pengujian, model mencapai akurasi 82,5%, presisi 76,0%, recall 95,0%, dan F1-score 84,44%, sementara matriks kebingungan melaporkan TP = 114, FN = 6, FP = 36, dan TN = 84. Temuan ini mengindikasikan bahwa augmentasi data berkontribusi signifikan dalam meningkatkan kemampuan model untuk mendeteksi sentimen positif (terlihat dari nilai recall yang tinggi), namun presisi yang lebih rendah menunjukkan adanya proporsi false positive yang perlu diminimalkan. Pendekatan fine-tuning pada model pre-trained BERT multibahasa memberikan keuntungan praktis berupa efisiensi waktu dan sumber daya serta kemampuan adaptasi cepat terhadap bahasa Indonesia tanpa pelatihan dari awal. Di sisi lain, penelitian ini memiliki keterbatasan, antara lain ketimpangan antara presisi dan recall serta kesulitan dalam menangani ulasan netral atau ambigu. Oleh karena itu, pengembangan lebih lanjut direkomendasikan, seperti peningkatan teknik augmentasi, perluasan ukuran dan keragaman dataset, eksplorasi model monolingual khusus Bahasa Indonesia, serta strategi pengurangan false positive (misalnya penalaan threshold, ensemble, atau kalibrasi probabilitas). Secara praktis, model ini berpotensi diaplikasikan pada situs ulasan, media sosial, dan industri perfilman untuk mendukung otomatisasi klasifikasi sentimen dan pengambilan keputusan pemasaran.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1080-1084
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.585

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019, [Online]. Available: http://arxiv.org/abs/1810.04805

[2] E. W. Pamungkas and D. G. P. Putri, "An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia," in *2016 6th International Annual Engineering Seminar (InAES)*, IEEE, Aug. 2016, pp. 28–31. doi: 10.1109/INAES.2016.7821901.

[3] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[4] M. Y. Rizky and Y. Stellarosa, "Preferensi Penonton Terhadap Film Indonesia," *Communicare : Journal of Communication Studies*, vol. 4, no. 1, p. 15, Jan. 2019, doi: 10.37535/101004120172.

[5] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," Feb. 2020, [Online]. Available: http://arxiv.org/abs/1905.05583

[6] M. Bucos and G. Țucudean, "Text Data Augmentation Techniques for Fake News Detection in the Romanian Language," *Applied Sciences*, vol. 13, no. 13, p. 7389, Jun. 2023, doi: 10.3390/app13137389.

[7] D. Nuryadi *et al.*, "FINE TUNING INDOBERT UNTUK ANALISIS SENTIMEN PADA ULASAN PENGGUNA APLIKASI TIKET.COM DI GOOGLE PLAY STORE," 2025.

[8] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is Multilingual BERT?," Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.01502

[9] Y. Zhang, R. Jin, and Z. H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1–4, pp. 43–52, Dec. 2010, doi: 10.1007/s13042-010-0001-0.

[10] K. Dedes, A. B. Putra Utama, A. P. Wibawa, A. N. Afandi, A. N. Handayani, and L. Hernandez, "Neural Machine Translation of Spanish-English Food Recipes Using LSTM," *JOIV : International Journal on Informatics Visualization*, vol. 6, no. 2, p. 290, Jun. 2022, doi: 10.30630/joiv.6.2.804.

[11] J. J. Webster and C. Kit, "Tokenization as the initial phase in NLP," in *Proceedings of the 14th conference on Computational linguistics* -, Morristown, NJ, USA: Association for Computational Linguistics, 1992, p. 1106. doi: 10.3115/992424.992434.

[12] C. Zhou, B. Li, H. Fei, F. Li, C. Teng, and D. Ji, "Revisiting Structured Sentiment Analysis as Latent Dependency Graph Parsing." [Online]. Available: https://github.

[13] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 6381–6387. doi: 10.18653/v1/D19-1670.

[14] M. Fadaee, A. Bisazza, and C. Monz, "Data Augmentation for Low-Resource Neural Machine Translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 567–573. doi: 10.18653/v1/P17-2090.

[15] F. Panjaitan, W. Ce, H. Oktafiandi, G. Kanugrahan, Y. Ramdhani, and V. H. C. Putra, "Evaluation of Machine Learning Models for Sentiment Analysis in the South Sumatra Governor Election Using Data Balancing Techniques," *Journal of Information Systems and Informatics*, vol. 7, no. 1, pp. 461–478, Mar. 2025, doi: 10.51519/journalisi.v7i1.1019.