Pengaruh Teknik Augmentasi Dalam Klasifikasi Berita Berbahasa Inggris Menggunakan Algoritma BERT

Proposal Tugas Akhir

Diajukan Untuk Memenuhi

Persyaratan Guna Meraih Gelar Sarjana

Informatika Universitas Muhammadiyah Malang

Muhammad Daffa

202110370311047

Bidang Minat

Data Science

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS MUHAMMADIYAH MALANG

2025

LEMBAR PERSETUJUAN

Pengaruh Teknik Augmentasi Dalam Klasifikasi Berita Berbahasa Inggris MenggunakanAlgoritmaBERT

TUGAS AKHIR

Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1
Informatika Universitas Muhammadiyah Malang

Menyetujui,
Malang, 21 Juli 2025

Dosen Pembimbing 1	Dosen Pembimbing 2
Christian Sri Kusuma Aditya S.Kom., M.Kom	**Ir. Yufis Azhar S.Kom., M.Kom.**
NIP. 180327021991PNS.	**NIP. 10814100544PNS.**

LEMBAR PENGESAHAN

Pengaruh Teknik Augmentasi Dalam Klasifikasi Berita Berbahasa Inggris MenggunakanAlgoritmaBERT

TUGAS AKHIR

Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1
InformatikaUniversitas Muhammadiyah Malang

Disusun Oleh :

MUHAMMAD DAFFA
202110370311047

Tugas Akhir ini telah diuji dan dinyatakan lulus melalui sidang majelis penguji pada tanggal 21 Juli 2025

Menyetujui,

Dosen Penguji 1	Dosen Penguji 2
Setio Basuki MT., Ph.D.	**Vinna Rahmayanti S S.Si., M.Si**
NIP. 10809070477PNS.	**NIP. 180306071990PNS.**

Mengetahui,
Ketua Jurusan Informatika

Dr. Cahrul Wasis Wicaksono S.kom. M.Cs.
NIP. 10814100541PNS.

LEMBAR PERNYATAAN

Yang bertanda tangan dibawah ini :

NAMA : **MUHAMMAD DAFFA**

NIM : **202110370311047**

FAK./JUR. : **Informatika**

Dengan ini saya menyatakan bahwa Tugas Akhir dengan judul **"Pengaruh Teknik Augmentasi Dalam Klasifikasi Berita Berbahasa Inggris Menggunakan Algoritma BERT"** beserta seluruh isinya adalah karya saya sendiri dan bukan merupakan karya tulis orang lain, baik sebagian maupun seluruhnya, kecuali dalam bentuk kutipan yang telah disebutkan sumbernya.

Demikian surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila kemudian ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya ini, atau ada klaim dari pihak lain terhadap keaslian karya saya ini maka saya siap menanggung segala bentuk resiko/sanksi yang berlaku.

Mengetahui,
Dosen Pembimbing

Christian Sri Kusuma Aditya S.Kom., M.Kom

Malang, 21 Juli 2025
Yang Membuat Pernyataan

MUHAMMAD DAFFA

ABSTRAK

Penelitian ini membahas pengaruh teknik augmentasi data terhadap performa model klasifikasi berita berbasis BERT. Lima skenario diujikan, yaitu tanpa augmentasi, serta dengan tiga teknik augmentasi individual (Synonym Replacement, Back Translation, dan Random Swap), serta kombinasi ketiganya. Evaluasi dilakukan menggunakan metrik klasifikasi dan analisis cosine similarity. Hasil menunjukkan bahwa teknik Synonym Replacement memberikan peningkatan kinerja terbaik dengan f1-score 0.984 dan eval loss terendah 0.0489. Namun, beberapa teknik augmentasi menghasilkan teks yang terlalu mirip dengan data asli (cosine similarity > 0.99), sehingga kurang efektif memperluas keragaman data. Penelitian ini menyimpulkan bahwa efektivitas augmentasi sangat bergantung pada kualitas dan variasi semantik dari data baru yang dihasilkan.

Kata kunci: Augmentasi Data, Back Translation, BERT, Klasifikasi Berita, Random Swap, Synonym Replacement

ABSTRACT

This study investigates the impact of data augmentation techniques on the performance of a BERT-based news classification model. Five experimental scenarios were evaluated: one without augmentation, three with individual augmentation techniques (Synonym Replacement, Back Translation, and Random Swap), and one combining all three methods. The model performance was assessed using classification metrics and cosine similarity analysis. Results show that Synonym Replacement provided the best improvement, achieving the highest F1-score of 0.984 and the lowest evaluation loss of 0.0489. However, several techniques generated augmented texts that were overly similar to the original data (cosine similarity > 0.99), limiting their effectiveness in increasing data diversity. This study concludes that the success of data augmentation largely depends on the semantic variety and quality of the newly generated data, which should enrich the training set without introducing noise.

***Keywords**: Back Translation, BERT, Data Augmentation, News Classification, Random Swap, Synonym Replacement*

LEMBAR PERSEMBAHAN

Bissmillahirrahmanirrahim..

Puji Syukur kehadirat Allah SWT yang telah melimpahka rahmat serta hidayah-Nya sehingga penulis dapat menyelesaikan Laporan Tugas Akhir ini dengan penuh suka duka dan kesabaran yang luar biasa.

Keberhasilan dalam penulisan Laporan Tugas Akhir ini tentunya tidak lepas dari keterlibatan kepada semua pihak yang telah memberikan dukungan, bimbingan, serta doa selama proses penyusunan. Oleh karena itu, penulis menyampaikan terima kasih kepada :

1. Kedua Orang Tua saya terutama mama saya yang selalu memperjuangkan apa yang saya butuhkan di masa kuliah, memastikan anaknya ini tidak kekurangan dalam hal apapun. Mendukung perjuangan saya dalam mengerjakan skripsi dan mendoakan saya supa selalu kuat menghadapi apapun ujian yang saya hadapi serta, menjadi orang yang sukses kelak.
2. Kepada mama nunuk yang selalu memberikan dukungan dan semangat bahwa saya pasti bisa melewati masa skripsi yang terbilang cukup berat.
3. Saudara saya yaitu dek Defi, kak Elga, kak Icha, mas Wahyu, dan mas Dhamay yang selalu mendukung saya, memberikan saya semangat, membuat saya bahagia disaat saya dalam keadaan sedih maupun saat stress.
4. Bapak Christian Sri Kusuma Aditya, S.Kom., M.Kom., dan bapak Ir. Yufis Azhar, S.Kom., M.Kom selaku dosen pembimbing yang telah dengan sabar dan penuh ketelitian membimbing penulis dalam menyusun tugas akhir ini. Terima kasih atas ilmu, arahan, serta masukan berharga yang diberikan, juga atas dedikasi dan waktu yang telah Bapak luangkan selama proses ini.
5. Bapak/Ibu Dosen Program Studi Informatika Universitas Muhammadiyah Malang, yang telah membekali penulis dengan ilmu dan wawasan berharga sepanjang masa perkuliahan. Terima kasih atas segala dedikasi dan pengajaran yang diberikan.
6. Teman CEMARA, sahabat seperjuangan yang telah menemani sejak awal perkuliahan hingga akhir perjalanan Tugas Akhir ini. Semua momen yang

kita lalui bersama akan selalu saya kenang sebagai bagian berharga dari masa studi yang tak terlupakan.

7. Teman-teman kos bu Eni yang selalu memberikan tawa bahagia, mensupport satu sama lain, menemani di saat masa sulit.
8. Teman-teman kontrakan atas segala semangat dan dorongan yang terus mengalir tanpa henti dalam proses penyelesaian Tugas Akhir ini, saya mengucapkan terima kasih yang tulus. Kehadiran dan kebersamaan yang diberikan telah menjadi sumber kekuatan tersendiri.

Malang, 26 Juli 2025

Penulis

UNIVERSITAS MUHAMMADIYAH MALANG

KATA PENGANTAR

Dengan memanjatkan puji Syukur kehadirat Allah SWT. Atas limpahan rahmat dan hidayah-Nya sehingga peneliti dapat menyelesaikan tugas akhir yang berjudul :

“Pengaruh Teknik Augmentasi Dalam Klasifikasi Berita Berbahasa Inggris Menggunakan Algoritma BERT”

Di dalam Tulisan ini menyajikan pembahasan mengenai pengaruh penerapan teknik augmentasi terhadap kinerja model klasifikasi berita berbahasa Inggris menggunakan algoritma BERT. Pokok-pokok bahasan mencakup proses pengumpulan dan praproses data, penerapan tiga teknik augmentasi (synonym replacement, back translation, dan random swap), pembangunan model klasifikasi berbasis BERT, serta evaluasi performa model pada lima skenario pengujian. Diharapkan penelitian ini dapat memberikan kontribusi dalam upaya meningkatkan kualitas pelatihan model klasifikasi teks melalui augmentasi data.

Peneliti menyadari sepenunya bahwa dalam penulisan tugas akhir ini masih banyak kekurangan dan keterbatasan. Oleh karena itu, peneliti mengharapkan saran yang membangun agar tulisan ini bermanfaat bagi perkembangan ilmu pengetahuan.

Malang, 26 Juli 2025

Muhammad Daffa

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR PUSTAKA

[1] M. D. Ria and A. Budiman, "Perancangan Sistem Informasi Tata Kelola Teknologi Informasi Perpustakaan," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 2, no. 1, pp. 122–133, 2021.

[2] W. Afandi, S. N. Saputro, A. M. Kusumaningrum, H. Adriansyah, M. H. Kafabi, and S. Sudianto, "Klasifikasi Judul Berita Clickbait menggunakan RNN-LSTM," *J. Inform. J. Pengemb. IT*, vol. 7, no. 2, pp. 85–89, 2022, doi: 10.30591/jpit.v7i2.3401.

[3] C. N. Daiman, A. Y. Rahman, and F. Nudiyansyah, "Klasifikasi Teks Berita Breaking News Di Manggarai Menggunakan Long Short Term Memory (Lstm)," *J. Mnemon.*, vol. 7, no. 2, pp. 170–174, 2024.

[4] S. Nur and K. Fithriasari, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K- Nearest," *J. Sains dan Seni*, vol. 5, no. 2, p. 6, 2021.

[5] K. I. Gunawan and J. Santoso, "Multilabel Text Classification Menggunakan SVM dan Doc2Vec Classification Pada Dokumen Berita Bahasa Indonesia," *J. Inf. Syst. Hosp. Technol.*, vol. 3, no. 01, pp. 29–38, 2021, doi: 10.37823/insight.v3i01.126.

[6] F. Sholekhah, A. D. Putri, R. Rahmaddeni, and L. Efrizoni, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbors untuk Klasifikasi Metabolik Sindrom," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 507–514, 2024, doi: 10.57152/malcom.v4i2.1249.

[7] N. M. Farhan and B. Setiaji, "Indonesian Journal of Computer Science," *Indones. J. Comput. Sci.*, vol. 12, no. 2, pp. 284–301, 2023.

[8] I. Athiyyah Rahma and L. Hulliyyatus Suadaa, "Penerapan Text Augmentation Untuk Mengatasi Data Yang Tidak Seimbang Pada Klasifikasi Teks Berbahasa Indonesia Studi Kasus: Deteksi Judul Clickbait Dan Komentar Hate Speech Pada Berita Online," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1329–1340, 2023, doi:

10.25126/jtiik.2023107325.

[9] J. Kapusta, D. Držik, K. Šteflovič, and K. S. Nagy, “Text Data Augmentation Techniques for Word Embeddings in Fake News Classification,” *IEEE Access*, vol. 12, no. February, pp. 31538–31550, 2024, doi: 10.1109/ACCESS.2024.3369918.

[10] F. O. Dayera, Musa Bundaris Palungan, “G-Tech : Jurnal Teknologi Terapan,” *G-Tech J. Teknol. Terap.*, vol. 8, no. 1, pp. 186–195, 2024.

[11] H. S. Anggraheni, M. J. Naufal, and N. Yudistira, “DETEKSI SPAM BERBAHASA INDONESIA BERBASIS TEKS MENGGUNAKAN MODEL BERT TEXT-BASED INDONESIAN SPAM DETECTION USING THE BERT MODEL,” vol. 11, no. 6, pp. 1291–1301, 2024, doi: 10.25126/jtiik.2024118121.

[12] L. Atikah, D. Purwitasari, and N. Suciati, “Deteksi Kejadian Lalu Lintas pada Teks Twitter dengan Pendekatan Klasidikasi Multi-Label Berbasis Deep Learning,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 1, p. 87, 2022, doi: 10.25126/jtiik.2022915206.

[13] N. T. Harnia, F. Meliasanti, and H. Setiawan, “Analisis Framing Berita Perundungan pada Media Online Detik.Com dan Tribunnews.Com sebagai Bahan Ajar Teks Berita di SMP,” *Edukatif J. Ilmu Pendidik.*, vol. 3, no. 5, pp. 3145–3153, 2021, doi: 10.31004/edukatif.v3i5.1240.

[14] T. H. Lubis and I. Koto, “Diskursus Kebenaran Berita Berdasarkan Undang-Undang Nomor 40 Tahun 1999 Tentang Pers Dan Kode Etik Jurnalistik,” *LEGA LATA J. Ilmu Huk.*, vol. 5, no. 2, pp. 231–250, 2020, doi: 10.30596/dll.v5i2.4169.

[15] T. Ridwansyah, “Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier,” *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 2, no. 5, pp. 178–185, 2022, doi: 10.30865/klik.v2i5.362.

[16] G. Elisabeth, Rahma Salsa Bilah, S. N. Ardini, N. Agustina, and D. A.

Rismayadi, “Klasifikasi Berita Palsu Kenaikan Harga Bahan Bakar Minyak (Bbm) Menggunakan Algoritma Support Vector Machine (Svm),” *Naratif J. Nas. Riset, Apl. dan Tek. Inform.*, vol. 5, no. 2, pp. 104–109, 2023, doi: 10.53580/naratif.v5i2.188.

[17] I. A. DLY, J. Jasril, S. Sanjaya, L. Handayani, and F. Yanto, “Klasifikasi Citra Daging Sapi dan Babi Menggunakan CNN Alexnet dan Augmentasi Data,” *J. Inf. Syst. Res.*, vol. 4, no. 4, pp. 1176–1185, 2023, doi: 10.47065/josh.v4i4.3702.

[18] M. Resa Arif Yudianto, P. Sukmasetya, R. Abul Hasani, and D. Sasongko, “Pengaruh Data Preprocessing terhadap Imbalanced Dataset pada Klasifikasi Citra Sampah menggunakan Algoritma Convolutional Neural Network,” *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1367–1375, 2022, doi: 10.47065/bits.v4i3.2575.

[19] D. Licari and G. Comandè, “ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law,” *CEUR Workshop Proc.*, vol. 3256, 2022.

[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019.

[21] Vidya Chandradev, I Made Agus Dwi Suarjaya, and I Putu Agung Bayupati, “Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT,” *J. Buana Inform.*, vol. 14, no. 02, pp. 107–116, 2023, doi: 10.24002/jbi.v14i02.7244.

[22] R. A. Sunan, H. F. E. K., and C. S. K. Aditya, “Klasifikasi Hoax Berita Politik Menggunakan Algoritma Long Short-Term Memory (LSTM) dengan Penambahan Fitur Embedding Global Vector (GloVe),” *J. Edukasi dan Penelit. Inform.*, vol. 10, no. 2, p. 287, 2024, doi: 10.26418/jp.v10i2.76042.

.

UNIVERSITAS
MUHAMMADIYAH
MALANG

FAKULTAS TEKNIK

INFORMATIKA
informatika.umm.ac.id | informatika@umm.ac.id

FORM CEK PLAGIARISME LAPORAN TUGAS AKHIR

Nama Mahasiswa : **Muhammad Daffa**

NIM : **202110370311047**

Judul TA : **Pengaruh Teknik Augmentasi Dalam Klasifikasi Berita Berbahasa Inggris Menggunakan Algoritma BERT**

Hasil Cek Plagiarisme dengan Turnitin

No.	Komponen Pengecekan	Nilai Maksimal Plagiarisme (%)	Hasil Cek Plagiarisme (%) *
1.	Bab 1 – Pendahuluan	10 %	10 %
2.	Bab 2 – Daftar Pustaka	25 %	9 %
3.	Bab 3 – Analisis dan Perancangan	25 %	9 %
4.	Bab 4 – Implementasi dan Pengujian	15 %	4 %
5.	Bab 5 – Kesimpulan dan Saran	5 %	4 %
6.	Makalah Tugas Akhir	20%	8 %

) Hasil cek plagiarism diisi oleh pemeriksa (staf TU)

***) Maksimal 5 kali (4 Kali sebelum ujian, 1 kali sesudah ujian)**

Mengetahui,

Pemeriksa (Staff TU)

(.....Bertin.....)

Kampus I
Jl. Bandung 1 Malang, Jawa Timur
P: +62 341 551 253 (Hunting)
F: +62 341 460 435

Kampus II
Jl. Bendungan Sutami No.188 Malang, Jawa Timur
P: +62 341 551 149 (Hunting)
F: +62 341 582 060

Kampus III
Jl. Raya Tlogomas No.246 Malang, Jawa Timur
P: +62 341 464 318 (Hunting)
F: +62 341 460 435
E: webmaster@umm.ac.id