

SKRIPSI

**PERBANDINGAN REPRESENTASI TEKS MENGGUNAKAN TF-IDF,
WORD2VEC, DAN SENTENCE-BERT UNTUK KLASIFIKASI
PERTANYAAN PADA CHATBOT SEJARAH KALIMANTAN TENGAH**



Disusun Oleh:

Dectrixie Theodore Mayumi Susanto

NIM 223020503140

JURUSAN TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS PALANGKA RAYA

2026

**PERBANDINGAN REPRESENTASI TEKS MENGGUNAKAN TF-IDF, WORD2VEC, DAN
SENTENCE-BERT UNTUK KLASIFIKASI PERTANYAAN PADA CHATBOT SEJARAH
KALIMANTAN TENGAH**

SKRIPSI

Sebagai salah satu syarat untuk menyelesaikan Program Strata-1 pada Jurusan Teknik Informatika Fakultas
Teknik Universitas Palangka Raya

Oleh

DECTRIXIE THEODORE MAYUMI SUSANTO

223020503140

Telah dipertahankan di depan tim penguji, pada :

Hari/Tanggal : Selasa, 01 September 2026

Waktu : 11:00-12:30 WIB

Tempat : Ruang Ujian 1

1. SEPTIAN GEGES, S.KOM., M.KOM.

NIP. 199109172020121004

..... (Ketua Penguji)

2. NOVA NOOR KAMALA SARI, S.T., M.Kom

NIP. 198904072015042004

..... (Pembimbing Utama/Pertama)

3. RESSA PRISKILA, S.T., M.T.

NIP. 199403012019032016

..... (Pembimbing Pendamping/Kedua)

4. DEVI KAROLITA, S.Kom., M.Kom., Ph.D

NIP. 198212062006042001

..... (Anggota)

Mengetahui :

Jurusan / Program Studi Teknik Informatika
Fakultas Teknik Universitas Palangka Raya
Ketua Jurusan,

ARISTA LESTARI, S.Kom., M.Cs., Ph.D.
NIP. 19800322 200501 2 004

HALAMAN PERNYATAAN

Saya menyatakan bahwa skripsi ini adalah benar-benar hasil karya saya sendiri, bukan merupakan plagiasi, dan belum pernah diajukan untuk memperoleh gelar sarjana di perguruan tinggi mana pun. Saya juga mengakui bahwa terdapat kutipan dari karya ilmiah orang lain yang dijadikan rujukan, yang telah saya cantumkan secara jelas dan dicantumkan sumbernya sesuai dengan etika penulisan karya ilmiah. Apabila di kemudian hari terbukti bahwa skripsi ini bukan karya asli saya atau terdapat unsur plagiarisme, saya bersedia menerima sanksi akademik sesuai ketentuan yang berlaku.

Palangka Raya, 14 September 2026



Dectrixie Theodore Mayumi Susanto

223020503140

RIWAYAT PENULIS

Data Diri

Nama : Dectrixie Theodore Mayumi Susanto
NIM : 223020503140
Fakultas : Teknik
Jurusan/Program Studi : Teknik Informatika
Jenjang : Strata 1 (S-1)
Jenis Kelamin : Perempuan
Tempat, Tanggal Lahir : Palangka Raya, 1 Desember 2004
Agama : Kristen Protestan
Status Dalam Keluarga : Anak Kandung
Anak Ke : 1
Alamat : Jl. Mendawai no. 035
No. Telepon/HP : 081346011204



Data Orang Tua

Nama Ayah : Yudi Susanto, M.Pd
Pekerjaan Ayah : Dosen
Nama Ibu : Sumiatie, M.Pd
Pekerjaan Ibu : Dosen
Alamat Orang Tua : Jalan Mendawai no. 035
No. HP Orang Tua : 081251166408

Riwayat Pendidikan

SD : SD Katolik Santo Yohanes Don Bosco (Tahun Lulus 2016)
SMP : SMP Negeri 3 Palangka Raya (Tahun Lulus 2019)
SMA : SMA Negeri 5 Palangka Raya (Tahun Lulus 2022)

Palangka Raya, 14 September 2026

Dectrixie Theodore Mayumi Susanto

NIM. 223020503140

HALAMAN PERSEMBAHAN

Puji dan syukur penulis panjatkan kepada Tuhan Yesus Kristus, atas segala kasih karunia, penyertaan, dan kekuatan yang diberikan, sehingga penulis dapat menyelesaikan skripsi dengan baik. Penulis menyadari bahwa dalam penyusunan skripsi ini, penulis tidak bekerja sendiri, melainkan mendapat banyak bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan yang berbahagia ini, penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Para dosen pembimbing dan penguji yang telah meluangkan waktu, tenaga, dan pikiran untuk memberikan bimbingan, arahan, serta masukan yang sangat berharga demi kesempurnaan skripsi ini, yaitu Ibu Nova Noor Kamala Sari, S.T., M.Kom. selaku Dosen Pembimbing Utama, Ibu Ressa Priskila, S.T., M.T. selaku Dosen Pembimbing Kedua, Bapak Septian Geges, S.Kom., M.Kom. selaku Ketua Penguji, dan Ibu Devi Karolita, S.Kom., M.Kom., Ph.D. selaku Dosen Penguji.
2. Papah dan Mamah tercinta, terima kasih atas setiap pengorbanan, kasih sayang tanpa syarat, serta dukungan moril maupun materil yang selalu mengalir tanpa henti. Terima kasih telah selalu percaya pada kemampuan saya, bersabar menghadapi segala keluh kesah dan emosi saya, serta menjadi tempat pulang paling nyaman. Skripsi ini adalah persembahan kecil untuk kalian, sebagai bukti bahwa didikan dan pengorbanan kalian tidak pernah sia-sia.
3. Untuk Jevon (adik pertama) dan Oreen (adik kedua) kesayangan Cici. Terima kasih sudah menjadi adik-adik yang luar biasa. Terima kasih atas segala canda tawa yang selalu berhasil mengusir penat di tengah pusingnya mengerjakan skripsi. Terima kasih sudah sabar mendengarkan cerita, keluh kesah, hingga ocehan Cici, serta selalu hadir memberikan semangat dan menjadi warna paling indah dalam hidup Cici. Kalian adalah alasan Cici untuk selalu tersenyum dan berjuang lebih keras.

4. Teman-teman "Koda" (Melda, Milky, dan Novita) yang luar biasa. Terima kasih karena selalu bersama dari awal semester hingga akhir, dan saling menguatkan. Tanpa kalian, mungkin saya tidak akan bisa survive melewati dinamika perkuliahan ini.
5. Nina, Jopa, Nisa, Gavin, dan seluruh teman-teman SIPNWARPS lainnya. Terima kasih atas kebersamaan dan dukungan selama masa perkuliahan yang membuat perjalanan ini jauh lebih berwarna.
6. Teman-teman Veloura. Walaupun kita baru bertemu dan komunitas ini baru terbentuk tahun ini, terima kasih banyak sudah menjadi penyegar di tengah penatnya pengerjaan skripsi. Terima kasih sudah menjadi tempat terbaik untuk melampiaskan penat dan healing lewat dance bersama kalian.
7. Untuk diriku sendiri. Terima kasih karena meskipun aslinya sangat pemalas dan sering menunda pekerjaan, pada akhirnya tetap berhasil mengalahkan rasa malas dan menyelesaikan skripsi ini sampai titik ini. *Proud of you!*
8. Seluruh pihak yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan dukungan, bantuan, dan doa sehingga skripsi ini dapat terselesaikan.

KATA PENGANTAR

Segala puji dan syukur penulis haturkan ke hadirat Tuhan Yesus Kristus, atas limpahan rahmat, hikmat, dan penyertaan-Nya, sehingga penulis dapat merampungkan skripsi berjudul "Perbandingan Representasi Teks Menggunakan TF-IDF, Word2Vec, dan Sentence-BERT untuk Klasifikasi Pertanyaan pada Chatbot Sejarah Kalimantan Tengah" dengan lancar.

Penyusunan skripsi ini ditujukan untuk memenuhi salah satu syarat kelulusan Strata-1 pada Jurusan Teknik Informatika, Fakultas Teknik, Universitas Palangka Raya. Fokus utama dari penelitian ini adalah melakukan komparasi terhadap tiga metode representasi teks, yaitu TF-IDF, Word2Vec, dan Sentence-BERT. Tujuannya adalah untuk mengidentifikasi dan memperoleh metode yang paling optimal dalam mengklasifikasikan pertanyaan pengguna pada Chatbot Sejarah Kalimantan Tengah, sehingga sistem dapat memberikan respons yang akurat dan relevan.

Penulis menyadari sepenuhnya bahwa masih terdapat berbagai keterbatasan dalam penelitian maupun penulisan skripsi ini. Oleh karena itu, kritik dan saran yang bersifat konstruktif sangat penulis harapkan demi penyempurnaan karya ini di masa mendatang. Semoga hasil penelitian ini dapat memberikan kontribusi yang berarti, khususnya dalam pengembangan teknologi pemrosesan bahasa alami (Natural Language Processing) untuk pelestarian sejarah daerah, serta bermanfaat bagi para pembaca.

Palangka Raya, 14 September 2026



Dectrixie Theodore Mayumi Susanto

223020503140

**PERBANDINGAN REPRESENTASI TEKS MENGGUNAKAN TF-IDF,
WORD2VEC, DAN SENTENCE-BERT UNTUK KLASIFIKASI
PERTANYAAN PADA CHATBOT SEJARAH KALIMANTAN TENGAH**

Dectrixie Theodore Mayumi Susanto (223020503140)

Jurusan Teknik Informatika Fakultas Teknik Universitas Palangka Raya

Kampus Tanjung Nyaho, Jl. Yos Sudarso, Palangka Raya 73112

Email: dectrixiesusanto@gmail.com

ABSTRAK

Pengembangan chatbot edukasi sejarah lokal memerlukan pemahaman intent pengguna yang akurat. Penelitian ini bertujuan membandingkan kinerja tiga metode representasi teks, yaitu TF-IDF, Word2Vec, dan Sentence-BERT (S-BERT), dalam klasifikasi pertanyaan dan temu kembali jawaban pada chatbot Sejarah Kalimantan Tengah. Dataset yang digunakan terdiri dari 809 pasangan pertanyaan-jawaban yang divalidasi reliabilitasnya menggunakan Cohen's Kappa ($\kappa = 0,9318$). Evaluasi dilakukan terhadap model klasifikasi Support Vector Machine (SVM) dan mekanisme retrieval menggunakan Cosine Similarity, mencakup metrik klasifikasi, Uji McNemar, ketahanan semantik (typo, parafrase, noise), dan efisiensi komputasi. Hasil penelitian menunjukkan bahwa TF-IDF dan S-BERT memiliki kinerja klasifikasi yang tinggi dan setara secara statistik (p -value = 0,8383), sementara Word2Vec gagal akibat hilangnya struktur sintaksis pada mekanisme average pooling. Pada pengujian arsitektur sistem, penambahan filter SVM pada pipeline dua tahap justru memicu error propagation yang menurunkan akurasi retrieval secara drastis saat menghadapi input tidak baku. S-BERT membuktikan ketahanan semantik (robustness) tertinggi terhadap variasi linguistik, meskipun membutuhkan waktu komputasi lebih lama yang masih dalam batas toleransi interaksi manusia-komputer. Penelitian ini merekomendasikan penggunaan S-BERT dengan arsitektur Single-Stage Pipeline (pencarian Cosine Similarity langsung tanpa filter SVM) sebagai konfigurasi optimal untuk menjamin keandalan sistem dalam menghadapi dinamika pertanyaan pengguna yang alami.

Kata Kunci: Chatbot Sejarah, Representasi Teks, Sentence-BERT, Klasifikasi Pertanyaan, Cosine Similarity

**COMPARISON OF TEXT REPRESENTATION USING TF-IDF,
WORD2VEC, AND SENTENCE-BERT FOR QUESTION CLASSIFICATION
IN CENTRAL KALIMANTAN HISTORY CHATBOT**

Dectrixie Theodore Mayumi Susanto (223020503140)

Jurusan Teknik Informatika Fakultas Teknik Universitas Palangka Raya

Kampus Tanjung Nyaho, Jl. Yos Sudarso, Palangka Raya 73112

Email: dectrixiesusanto@gmail.com

ABSTRACT

The development of local history educational chatbots requires accurate user intent comprehension. This study aims to compare the performance of three text representation methods: TF-IDF, Word2Vec, and Sentence-BERT (S-BERT), in question classification and response retrieval for a Central Kalimantan History chatbot. The dataset comprises 809 question-answer pairs validated using Cohen's Kappa ($\kappa = 0.9318$). The evaluation covers Support Vector Machine (SVM) classification models and Cosine Similarity retrieval mechanisms, encompassing classification metrics, McNemar's Test, semantic robustness (typos, paraphrases, noise), and computational efficiency. Results indicate that TF-IDF and S-BERT exhibit statistically equivalent classification performance (p -value = 0.8383), whereas Word2Vec failed due to the loss of syntactic structure in its average pooling mechanism. Regarding system architecture, adding an SVM filter in a two-stage pipeline triggered error propagation, drastically reducing retrieval accuracy when handling non-standard inputs. S-BERT demonstrated the highest semantic robustness against linguistic variations, despite requiring longer computation time which remains within human-computer interaction tolerance limits. This study recommends utilizing S-BERT with a Single-Stage Pipeline architecture (direct Cosine Similarity search without SVM filtering) as the optimal configuration to ensure system reliability in handling natural and non-standard user queries.

Keywords: History Chatbot, Text Representation, Question Classification, Cosine Similarity

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PENGESAHAN.....	ii
RIWAYAT PENULIS	iv
HALAMAN PERSEMBAHAN.....	v
KATA PENGANTAR	vii
ABSTRAK	viii
ABSTRACT.....	ix
DAFTAR ISI.....	x
DAFTAR GAMBAR	xiv
DAFTAR TABEL.....	xv
BAB I PENDAHULUAN	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah.....	4
1.4 Tujuan Penelitian.....	5
1.5 Manfaat Penelitian.....	5
1.6 Sistematika Penulisan.....	5
1.7 Jadwal Kegiatan.....	7
BAB II LANDASAN TEORI	8
2.1 Tinjauan Pustaka	8
2.2 Teori Umum	17
2.2.1 Python	17
2.2.2 Chatbot.....	17

2.2.3	Klasifikasi Pertanyaan.....	18
2.2.4	Cohen's Kappa	18
2.2.5	TF-IDF	20
2.2.6	Word2Vec	22
2.2.7	Sentence-BERT.....	23
2.2.8	Karakteristik Komparatif Representasi Teks	25
2.2.9	Algoritma Support Vector Machine.....	27
2.2.10	Evaluasi Klasifikasi Intent	28
2.2.10	Evaluasi Retrieval Response	31
2.2.11	Evaluasi Efisiensi Komputasi	33
2.2.12	Uji Signifikansi Statistik	33
2.2.13	Cosine Similiarity.....	36
2.2.14	Sejarah Kalimantan Tengah	36
BAB III METODE PENELITIAN.....		38
3.1	Tahapan Penelitian	38
3.2	Data Preparation	41
3.2.1	Teknik Pengumpulan Data.....	41
3.2.2	Perancangan Kategori Intent.....	42
3.2.3	Validasi Anotasi	50
3.3	Preprocessing Data	57
3.4	Pembagian Data.....	57
3.5	Augmentasi Data Latih.....	58
3.6	Ekstraksi Fitur	59
3.7	Pelatihan dan Prediksi Model.....	65
3.8	Evaluasi Klasifikasi Intent.....	66

3.9	Uji Signifikansi Statistik.....	68
3.10	Cosine Similiarity	69
3.11	Evaluasi Retrieval Response.....	70
3.11	Evaluasi Efisiensi Komputasi	71
BAB IV HASIL DAN PEMBAHASAN		72
4.1	Hasil.....	72
4.1.1	Deskripsi Data dan Persiapan Eksperimen	72
4.1.2	Hasil Implementasi Representasi Teks	81
4.1.3	Hasil Evaluasi Klasifikasi	86
4.1.4	Hasil Uji Signifikansi Statistik (McNemar's Test).....	94
4.1.5	Evaluasi Retrieval Response	96
4.1.6	Evaluasi Kualitas Representasi Semantik	98
4.1.6.1	Evaluasi terhadap Variasi Tidak Baku (Typo & Singkatan) ...	99
4.1.6.2	Evaluasi terhadap Variasi Parafrase	105
4.1.6.3	Evaluasi Variasi Panjang Query dan Noise	114
4.1.7	Evaluasi Efisiensi Komputer.....	120
4.1.8	Demo Interaksi	122
4.2	Pembahasan	140
4.2.1	Karakteristik dan Perbandingan Kinerja Klasifikasi.....	140
4.2.2	Implikasi Uji McNemar	142
4.2.3	Efektivitas Retrieval dan Arsitektur Pipeline.....	143
4.2.4	Ketahanan Semantik terhadap Variasi Bahasa.....	145
4.2.5	Efisiensi Komputasi dan Trade-Off Sistem	148
4.2.6	Rekomendasi Arsitektur Sistem Chatbot Sejarah	149
BAB V KESIMPULAN & SARAN		155

5.1	Kesimpulan.....	155
5.2	Saran.....	156
DAFTAR PUSTAKA		157
LAMPIRAN.....		164

DAFTAR GAMBAR

Gambar 2. 1 Support Vector Machine	27
Gambar 2. 2 konversi dari nilai X^2 ke p-value	35
Gambar 3. 1 Alur Sistem.....	40
Gambar 4. 1 Sampel Data Preprocessing.....	77
Gambar 4. 2 Hasil implementasi metode TF-IDF.....	82
Gambar 4. 3 Hasil implementasi metode Word2Vec.....	83
Gambar 4. 4 Hasil implementasi metode Sentence-BERT	85
Gambar 4. 5 Pengujian menggunakan Data Validasi (TF-IDF)	87
Gambar 4. 6 Confusion Matrix Model SVM dengan Representasi TF-IDF.....	88
Gambar 4. 7 Hasil klasifikasi SVM terhadap TF-IDF	89
Gambar 4. 8 Pengujian menggunakan Data Validasi (Word2Vec)	90
Gambar 4. 9 Confusion Matrix Model SVM dengan Representasi Word2Vec....	90
Gambar 4. 10 Hasil klasifikasi SVM terhadap Word2Vec	91
Gambar 4. 11 Pengujian menggunakan Data Validasi (S-Bert)	92
Gambar 4. 12 Confusion Matrix Model SVM dengan Representasi Sentence-BERT.....	93
Gambar 4. 13 Hasil Klasifikasi SVM terhadap Sentence-BERT.....	94
Gambar 4. 14 Evaluasi retrieval kombinasi SVM dan cosine similiarity	97
Gambar 4. 15 Evaluasi retrieval yang hanya menggunakan cosine similiarity	98
Gambar 4. 16 Hasil Efisiensi Pada Model Klasifikasi	120
Gambar 4. 17 Hasil Efisiensi Retrieval (SVM + Cosine Similiarity)	121
Gambar 4. 18 Hasil Efisiensi Retrieval (Cosine Similiarity)	122
Gambar 1. Hasil pertanyaan prediksi di terminal.....	164
Gambar 2. Tampilan Chatbot.....	164

DAFTAR TABEL

Tabel 1. 1 Jadwal Kegiatan	7
Tabel 2. 1 Intrepretasi Pelabelan.....	19
Tabel 2. 2 Komponen utama Confusion Matrix.....	28
Tabel 3. 1 Kategori Faktual Dasar	43
Tabel 3. 2 Kategori Definisi dan Deskripsi.....	44
Tabel 3. 3 Kategori Analitis Sejarah	46
Tabel 3. 4 Kategori Kontekstual dan Filosofis.....	47
Tabel 3. 5 Kategori Naratif dan Mekanis.....	49
Tabel 3. 6 Elemen <i>guideline</i>	50
Tabel 3. 7 Elemen <i>guideline</i> masing-masing kategori	51
Tabel 3. 8 Proses anotasi.....	56
Tabel 3. 9 Contoh sampel hasil augmentasi	59
Tabel 3. 10 Contoh Koleksi Dokumen Pertanyaan.....	59
Tabel 3. 11 Contoh TF-IDF	60
Tabel 3. 12 Contoh vector per-kata.....	62
Tabel 3. 13 Perbedaan sebelum dan sesudah memahami konteks.....	63
Tabel 3. 14 Contoh 5 dimensi pertama Sentence Bert	64
Tabel 3. 15 Ilustrasi Nilai TP, FP, dan FN.....	67
Tabel 4. 1 Hasil perhitungan proporsi Cohen's Kappa	73
Tabel 4. 2 pertanyaan yang mewakili masing-masing kategori	75
Tabel 4. 3 Pembagian Data	79
Tabel 4. 4 Perbandingan Distribusi Kelas Sebelum dan Sesudah Augmentasi	81
Tabel 4. 5 Perhitungan Statistik	95
Tabel 4. 6 Variasi pertanyaan Typo	99
Tabel 4. 7 Jenis Parafrase.....	105
Tabel 4. 8 Hasil Parafrase Baku: Aktif ke Pasif, Konversi Kata Kerja ke Kata Benda, dan Perubahan Klausa.....	107
Tabel 4. 9 Hasil Parafrase Baku: Substitusi Frasa Sintaksis, Perubahan Koordinasi, Substitusi Antonim, dan Implikasi Semantik.....	108
Tabel 4. 10 Variasi Panjang Query dan Noise	114

Tabel 4. 11 Kunci pertanyaan yang tepat.....	123
Tabel 4. 12 hasil pengujian dengan perpaduan kinerja SVM dan cosine similiarity menggunakan baseline TF-IDF.....	126
Tabel 4. 13 hasil pengujian dengan perpaduan kinerja SVM dan cosine similiarity menggunakan baseline Word2Vec.....	128
Tabel 4. 14 hasil pengujian dengan perpaduan kinerja SVM dan cosine similiarity menggunakan baseline S-Bert.....	130
Tabel 4. 15 hasil pengujian dengan perpaduan kinerja SVM dan cosine similiarity secara keseluruhan baseline	133
Tabel 4. 16 hasil pengujian dengan kinerja cosine similiarity menggunakan baseline TF-IDF	134
Tabel 4. 17 hasil pengujian dengan kinerja cosine similiarity menggunakan baseline Word2Vec	136
Tabel 4. 18 hasil pengujian dengan kinerja cosine similiarity menggunakan baseline S-Bert	138
Tabel 4. 19 hasil pengujian dengan kinerja cosine similiarity secara keseluruhan baseline.....	139

DAFTAR PUSTAKA

- Ahmed, A., Mohandas, R., Joorabchi, A., & Hayes, M. J. (2025). Subjective Evaluation of Operator Responses for Mobile Defect Identification in Remanufacturing : Application of NLP and Disagreement Tagging. *Big Data and Cognitive Computing*, 9.
- Ali, M. I., Fardiarsyah, R. D., Shodik, L., Kinanti, F. Z. D., & Pujiono, I. P. (2025). Analisis Komparatif Efisiensi Memori dan Waktu Komputasi pada 8 Algoritma Sorting menggunakan C++. *LogicLink: Journal of Artificial Intelligence and Multimedia in Informatics*, 2(1), 1–17.
- Anugraha, A. A. P., Swanjaya, D., & Sehartian, J. (2025). Analisis Hasil Rekomendasi Cerita Panji Kediri Menggunakan Item-Based Collaborative Filtering. *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, 9(2), 1578–1584.
- Azhar, R. A., & Widodo, W. (2026). Analisis Sentimen Ulasan EA FC 26 di Steam : Perbandingan SVM dan Multinomial Naïve Bayes. *Management of Information System Journal*, 4(3), 238–247. <https://doi.org/10.47065/mis.v2i3.2974>
- Bhagat, R., & Hovy, E. (2013). What Is a Paraphrase? *Computational Linguistics*, 39, 463–472. https://doi.org/10.1162/COLI_a_00166
- Boussaad, L. (2025). Spam detection using sentence-BERT with attention-enhanced GRU and SVM. *Applied Computing and Informatics*, 1–16. <https://doi.org/10.1108/ACI-02-2025-0035>
- Church, K. W. (2016). Emerging Trends Word2Vec. *Natural Language Engineering*, 23(1), 155–162. <https://doi.org/10.1017/S1351324916000334>
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(3), 1986.
- Cortes, C., & Vapnik, V. (1995). *Support-Vector Networks* (L. Saitta (ed.)). Kluwer Academic Publishers.

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computer*, 10(7), 1895–1923. <https://doi.org/10.1162/089976698300017197>
- Gupta, M., Dheekonda, V., & Masum, M. (2024). International Journal of Information Genie : Enhancing information management in the restaurant industry through AI-powered chatbot. *International Journal of Information Management Data Insights*, 4(2), 100255. <https://doi.org/10.1016/j.jjime.2024.100255>
- Halizah, M. N., Najla, S., Putri, S. A., Febita, N., & Hidayat, M. L. (2026). Perbandingan Kinerja Artificial Neural Network Random Forest dan LightGBM dalam Prediksi Penyakit Jantung Menggunakan Dataset Kaggle. *Jurnal Ilmiah Wahana Pendidikan*, 12(1), 196–209.
- Hamdani, I. M., & Bustamin, S. (2025). *Dasar Pemrograman Python: Algoritma dan Struktur Data* (A. H. Julyaningsih & Penata (eds.)). CV. Science Tech Group.
- Hanum, A. R., Pramukantoro, E. S., & Kartikasari, D. P. (2025). Studi Perbandingan Kinerja Tf-Idf Dan Indobert Untuk Rekomendasi Resep Berdasarkan Ketersediaan Bahan Makanan Berbasis Website. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 9(10), 1–8.
- Hasbi, M. A., Imanda, R., & Fathan Fauzan, M. (2025). Implementasi Chatbot Berbasis Large Language Model Untuk Pencarian Skripsi Mahasiswa Terintegrasi dengan Whatsapp. *Arcitech: Journal of Computer Science and Artificial Intelligence*, 5(1), 148–167. <https://doi.org/10.29240/arcitech.v5i1.13974>

- Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018). Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 2(1), 306–312.
- Jamshidian, M. (2023). Evaluation of Text Transformers for Classifying Sentiment of Reviews by Using TF-IDF , BERT (word embedding), SBERT (sentence embedding) with Support Vector Machine Evaluation. *Computer Engineering Commons*.
- Jayanti, H. D., & Rohman, A. (2026). Cyberbullying Detection in Indonesian TikTok Comments Using IndoBERT with Fairness Evaluation. *Journal of Information Systems and Informatics*, 8(1), 907–927.
- Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features Thorsten. *Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, 1398, 137–142.
- Judijanto, L., Wibowo, G. A., Husnita, L., & Sepriano. (2024). *PENGANTAR ILMU SEJARAH (Teori, Konsep, dan Metodologi dalam Kajian Sejarah)*.
- Lahbari, I., & Alaoui, S. O. El. (2024). Exploring Sentence Embedding Representation For Arabic Question Answering. *International Journal of Computing and Digital Systems*, 15(1), 1229–1241.
- Li, X., & Roth, D. (2002). Learning Question Classifiers. *COLING 2002: The 19th International Conference on Computational Linguistics*, 1.
- Mabaquiao, N. M. (2018). Speech Act Theory: From Austin to Searle. *Augustinian: A Journal for Humanities, Social Sciences, Business, and Education*, 19(1), 35–45.
- Malvin, Dylan, C., & Rangkuti, A. H. (2022). WhatsApp Chatbot Customer Service Using Natural Language Processing and Support Vector Machine. *International Journal of Emerging Technology and Advanced Engineering*, 12(03), 130–136. <https://doi.org/10.46338/ijetae0322>

- Manning, C. D. (2008). *Introduction to Information Retrieval*. Syngress Publishing.
- Masuzzahra, T. R., Umam, K., Mustofa, H., & Handayani, M. R. (2025). HANA : An AI Chatbot for Islamic Jurisprudence on Menstruation using SBERT and TF-IDF. *Journal of Applied Informatics and Computing (JAIC)*, 9(3), 1013–1024.
- Mikolov, T., & Com, T. G. (2014). Distributed Representations of Sentences and Documents. *ArXiv:1405.4053*.
- Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv: 1301.3781*, 1–12.
- Mitera-kielbasa, E., & Zima, K. (2024). Automated Classification of Exchange Information Requirements for Construction Projects Using Word2Vec and SVM. *Infrastructures*, 9(11), 1–14.
- Mutawa, A. M., & Alrumaih, A. (2025). Determining the meter of classical Arabic poetry using deep learning: a performance analysis. *Frontiers in Artificial Intelligence*, 1–14. <https://doi.org/10.3389/frai.2025.1523336>
- Parameswari, S. D., Lubis, M., Suakanto, S., Ramadhan, Y. Z., Amanah, N., & Dila, R. A. (2025). Studi Perbandingan Naive Bayes dan Support Vector Machine (SVM) dalam Analisis Sentimen Pengguna Metaverse. *Jurnal Teknologi Dan Manajemen Industri Terapan (JTMIT)*, 4(3), 1059–1065.
- Pebdika, A., Herdiana, R., Solihudin, D., Pintar, P. I., Bayes, N., Penentuan, U., & Bantuan, P. (2023). *KLASIFIKASI MENGGUNAKAN METODE NAIVE BAYES UNTUK MENENTUKAN CALON PENERIMA PIP*. 7(1), 452–458.
- Powers, D. M. . (2020). Evaluation : From Precision , Recall And F-Measure To ROC , Informedness, Markedness & Correlation. *ArXiv Preprint ArXiv:2010.16061.*, 37–63.
- Qodrin, S. Al, Yusliani, N., & Syahrini, A. (2022). Klasifikasi Pertanyaan

- Berbahasa Indonesia Menggunakan Algoritma Support Vector Machine dan Seleksi Fitur Mutual Information. *Jurnal JUPITER*, 14(2), 44–52.
- Rahadiyan, D., Royan, N., Sudrajat, M., Nugraha, B. S., Adityo, V., Alifah, N., Humami, M., & Latief, P. M. Al. (2025). Chatbot for Mental Health Using Tf-Idf and SVM. *IT FOR SOCIETY*, 9(2), 29–32.
- Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 1–14. <https://doi.org/10.1038/s41598-024-56706-x>
- Raschka, S. (2018). *Model Evaluation , Model Selection , and Algorithm Selection in Machine Learning*.
- Raschka, S. (2020). *STAT 451 : Machine Learning Lecture Notes*.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Pro-Cessing (EMNLP-IJCNLP)*, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., & Liu, Y. (2007). Statistical Machine Translation for Query Expansion in Answer Retrieval. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 464–471.
- Singh, S., Kumar, K., & Kumar, B. (2024). Analysis of feature extraction techniques for sentiment analysis of tweets. *Turkish Journal of Engineering*, 8(4), 741–753.
- Singh, V., Pencina, M., Einstein, A. J., Liang, J. X., Berman, D. S., & Slomka, P. (2021). Impact of train / test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific Reports*, 1–8. <https://doi.org/10.1038/s41598-021-93651-5>

- Soebadio, P. D. H. (1978). *Sejarah Daerah Kalimantan Tengah*. PROYEK PENELITIAN PENCATATAN KEBUDAYAAN DAERAH DEPARTEMEN PENDIDIKAN DAN KEBUDAYAAN.
- Sutriawan, Rustad, S., Shidik, G. F., & Pujiono. (2025). Performance Evaluation of Text Embedding Models for Ambiguity Classification in Indonesian News Corpus: A Comparative Study of TF-IDF, Word2Vec, FastText BERT, and GPT. *Ingénierie Des Systèmes d'Information*, 30(6), 1469–1482.
- Syah, M. P., Wardani, A. P., & Idhom, M. (2025). Perbandingan Representasi Teks Tf-Idf Dan Bert Terhadap Akurasi Cosine Similarity Dalam Penilaian Otomatis Jawaban Berbasis Teks. *DSI: Jurnal Data Science Indonesia*, 5(1), 47–59.
- TAMAM, F. F. (2008). *PERTEMPURAN DANAU MARE DI TUMBANG SAMBA KALIMANTAN TENGAH (1945 - 1949)*. BPSNT PONTIANAK.
- Vila, M., Martí, M. A., & Rodríguez, H. (2014). Is This a Paraphrase ? What Kind ? Paraphrase Boundaries and Typology. *Open Journal of Modern Linguistics*, 4, 205–218. <https://doi.org/10.4236/ojml.2014.41016>
- Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2017). *Probability & Statistics for Engineers & Scientists*.
- Wilson, L. (2016). Anderson and Krathwohl Bloom's Taxonomy Revised Understanding the New Version of Bloom's Taxonomy. *Creative Education*, 12(8).
- Witanto, N. K. Y., Swanjaya, D., & Pamungkas, D. (2025). Sistem Rekomendasi Lagu Berbasis Knn Menggunakan Fitur Audio Dari Spotify Api. *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, 9(2), 1516–1522.
- Zhan, Z. (2025). Comparative Analysis of TF-IDF and Word2Vec in Sentiment Analysis : A Case of Food Reviews. *ITM Web of Conferences*, 70.
- Zulkarnain, M. A., Raharjo, M. F., & Olivya, M. (2020). Perancangan Aplikasi

Chatbot Sebagai Media E-Learning Bagi Siswa. *Elektron : Jurnal Ilmiah*, 12(2), 88–95. <https://doi.org/10.30630/eji.12.2.188>