

THESIS

**PERBANDINGAN PENGARUH TEKS VEKTORISASI TERHADAP
KINERJA KLASIFIKASI ULASAN PENGGUNA WHATSAPP DI
PLAYSTORE DAN APPSTORE**



STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

Oleh:

Annida Rizki Luthfi Astuti

NIM : 23206051021

**PROGRAM STUDI INFORMATIKA
PROGRAM MAGISTER FAKULTAS SAINS DAN
TEKNOLOGI
UIN SUNAN KALIJAGA**

**YOGYAKARTA
2025**

HALAMAN PENGESAHAN



KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

PENGESAHAN TUGAS AKHIR

Nomor : B-1247/Un.02/DST/PP.00.9/06/2025

Tugas Akhir dengan judul : "Perbandingan Pengaruh Teks Vektorisasi Terhadap Kinerja Klasifikasi Kebutuhan Pengguna WhatsApp di Playstore dan AppStore"

yang dipersiapkan dan disusun oleh:

Nama : ANNIDA RIZKI LUTHFI ASTUTI, S.Kom
Nomor Induk Mahasiswa : 23206051021
Telah diujikan pada : Jumat, 13 Juni 2025
Nilai ujian Tugas Akhir : A-

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR



Valid ID: 6862304f9e7ab

Ketua Sidang

Dr. Agus Mulyanto, S.Si., M.Kom., ASEAN Eng.
SIGNED



Valid ID: 68580a991427f

Penguji I

Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom.,
IPM., ASEAN Eng.
SIGNED



Valid ID: 685ca4e42ad27

Penguji II

Ir. Maria Ulfah Siregar, S.Kom., MIT., Ph.D.
SIGNED



Valid ID: 68625037db10b

Yogyakarta, 13 Juni 2025
UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama : Annida Rizki Luthfi Astuti
NIM : 23206051021
Jenjang : Magister
Program Studi : Informatika

Menyatakan bahwa naskah tesis ini secara keseluruhan adalah hasil penelitian/karya saya sendiri, kecuali pada bagian-bagian yang dirujuk sumbernya.

Yogyakarta, 13 Juni 2025

Saya yang menyatakan,



Annida Rizki Luthfi Astuti
NIM: 23206051021

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

PERNYATAAN BEBAS PLAGIASI

Yang bertanda tangan di bawah ini:

Nama : Annida Rizki Luthfi Astuti
NIM : 23206051021
Jenjang : Magister
Program Studi : Informatika

Menyatakan bahwa naskah tesis ini secara keseluruhan benar-benar bebas dari plagiasi. Jika di kemudian hari terbukti melakukan plagiasi, maka saya siap ditindak sesuai ketentuan hukum yang berlaku

Yogyakarta, 13 Juni 2025

Saya yang menyatakan,



Annida Rizki Luthfi Astuti
NIM: 23206051021

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

NOTA DINAS PEMBIMBING

NOTA DINAS PEMBIMBING

Kepada Yth,
Dekan Fakultas Sains dan Teknologi
UIN Sunan Kalijaga
Yogyakarta

Assalamu'alaikum wr.wb

Setelah melakukan bimbingan, arahan, dan koreksi terhadap penulisan tesis yang berjudul:

**PERBANDINGAN PENGARUH TEKS VEKTORISASI TERHADAP KINERJA
KLASIFIKASI KEBUTUHAN PENGGUNA WHATSAPP DI PLAYSTORE DAN
APPSTORE**

Yang ditulis oleh:

Nama : Annida Rizki Luthfi Astuti
NIM : 23206051021
Jenjang : Magister
Program Studi : Informatika

Saya berpendapat bahwa tesis tersebut sudah dapat diajukan kepada Magister Informatika UIN Sunan Kalijaga untuk diujikan dalam rangka memperoleh gelar Magister Informatika.

Wassalamu'alaikum wr. wb

Yogyakarta, 02 Juni 2025

Pembimbing,



Dr. Agus Mulyanto, S.Si., M.Kom., ASEAN Eng.

ABSTRAK

Meningkatnya popularitas aplikasi pesan instan WhatsApp memerlukan adanya pemahaman lebih mendalam untuk pengembangan aplikasinya melalui ulasan pengguna. Ulasan yang tersedia di platform distribusi aplikasi seperti Playstore dan Appstore mengandung informasi bernilai namun bersifat tidak terstruktur, dikarenakan terdapat singkatan, ejaan tidak baku, serta penggunaan Bahasa informal yang menyulitkan proses analisis. Tantangan ini mendorong perlunya penerapan pendekatan teknologi berbasis *Natural Language Processing* (NLP), khususnya pada tahap representasi teks. Salah satu pendekatan penting dalam NLP adalah vektorisasi teks yang mengubah teks menjadi format numerik agar dapat diproses oleh algoritma *machine learning*. Namun, masih terbatasnya penelitian yang secara khusus membandingkan efektivitas metode vektorisasi dalam klasifikasi kebutuhan pengguna menjadikan isu ini relevan untuk dikaji lebih lanjut. Penelitian ini bertujuan untuk membandingkan kinerja tiga metode vektorisasi teks, yaitu Word2Vec, GloVe, dan FastText dalam tugas klasifikasi kebutuhan pengguna berdasarkan ulasan aplikasi WhatsApp. Dataset yang digunakan diambil dari huggingface berupa hasil scraping dari ulasan di Playstore dan Appstore yang telah melalui proses labeling secara manual. Setelah tahap pre-processing, data diolah menggunakan teknik balancing SMOTE untuk mengatasi ketidakseimbangan kelas. Selanjutnya, dilakukan klasifikasi menggunakan algoritma Random Forest dengan evaluasi performa model melalui confusion matrix, akurasi, presisi, recall, dan f1-score. Hasil penelitian menunjukkan bahwa metode vektorisasi GloVe memberikan kinerja tertinggi dengan akurasi sebesar 95.05% di susul oleh Word2Vec 93.75% dan FastText 93.73%. Temuan ini mengindikasikan bahwa pemilihan metode representasi teks memiliki pengaruh signifikansi terhadap efektivitas klasifikasi kebutuhan pengguna. GloVe unggul dalam menangkap relasi semantic global antarkata, sementara Word2Vec dan FastText masing-masing menunjukkan kekuatan pada pemrosesan konteks lokal dan kata-kata tidak baku. Penelitian ini memberikan kontribusi dalam memahami gambaran terkait perbandingan model teks vektorisasi Word2Vec, GloVe, dan FastText, serta penggunaan algoritma klasifikasi *Random Forest* dan Teknik *imbalancing data* SMOTE dalam melakukan klasifikasi kebutuhan pengguna.

MOTTO

“Sesungguhnya bersama kesulitan ada kemudahan”

(Q.S. Al-Insyirah:5)

Ada yang cepat, ada yang lambat.

Tak masalah, asalkan tidak diam di tempat.

Sebab perjalanan masih panjang.

Selagi tetap berjalan, suatu hari kamu akan sampai.

(Ustadzah Halimah Alaydrus)



HALAMAN PERSEMBAHAN

Tesis ini penulis persembahkan untuk:

Almamater Tercinta

Prodi Magister Informatika

Fakultas Sains dan Teknologi

Universitas Islam Negeri Sunan Kalijaga Yogyakarta



STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

KATA PENGANTAR

Assalamu 'alaikum wr.wb

Dengan mengucapkan puji dan syukur kepada Allah SWT yang telah memberikan rahmat, nikmat, dan hidayah-Nya, serta sholawat beserta salam saya curahkan kepada Nabi Muhammad SAW, saya dapat menyelesaikan tesis ini dengan baik. Tesis ini berjudul “Perbandingan Pengaruh Teks Vektorisasi Terhadap Kinerja Klasifikasi Kebutuhan Pengguna WhatsApp di Playstore dan Appstore”. Penulisan tesis ini disusun sebagai pengajuan untuk memenuhi salah satu syarat kelulusan guna memperoleh gelar Magister Informatika Fakultas Sains dan Teknologi di Universitas Islam Negeri Sunan Kalijaga Yogyakarta. Penulis menyadari bahwa dalam menyelesaikan ini tidak terlepas dari bantuan bimbingan dari berbagai pihak. Oleh karena itu, penulis mengucapkan terimakasih kepada:

1. Keluarga penulis Bapak Sugiharto, Ibu Dwi Astuti, adik Muhammad Taufiq, serta keluarga besar di Yogyakarta. Terima kasih atas segala jasa, iringan doa yang tidak pernah terputus sampai sekarang dan selalu menyemangati dalam mengiringi setiap perjalanan menyelesaikan pendidikan sampai jenjang ini, semoga Allah SWT selalu memberkahi kehidupan kita.
2. Prof. Noorhaidi Hasan. Selaku Rektor di Universitas Islam Negeri Sunan Kalijaga Yogyakarta yang telah memberikan kesempatan untuk menempuh studi.

3. Prof. Dr. Dra. Hj. Khurul Wardati, M.Si. Selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sunan Kalijaga Yogyakarta yang telah memberikan kesempatan untuk menempuh Pendidikan di Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta.
4. Dr. Ir. Sumarsono, S.T., M.Kom. selaku Kepala Prodi Magister Informatika Universitas Islam Negeri Sunan Kalijaga Yogyakarta dan Ir. Muhammad Didik Rohmad Wahyudi, S.T., M.T. Selaku Sekretaris Jurusan Prodi Informatika Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
5. Bapak Dr. Agus Mulyanto, S.Si., M.Kom., ASEAN Eng. Selaku pembimbing yang telah banyak berkenan meluangkan waktu, tenaga, serta ilmunya untuk memberikan bimbingan kepada penulis dalam menyelesaikan tesis ini. Terima kasih penulis ucapkan dan penghargaan setinggi-tingginya yang dengan penuh kesabaran dan kearifan telah memberikan arahan dan semangat di sela-sela kesibukannya.
6. Bapak dan Ibu Dosen, Staff Tata Usaha, dan tenaga lain Prodi Magister Informatika Universitas Islam Sunan Kalijaga Yogyakarta, terkhusus yang memberikan materi kuliah. Terima kasih penulis ucapkan karena sudah memberikan banyak ilmu pengetahuan sehingga penulis dapat menyelesaikan perkuliahan dan penelitian yang menjadi tesis ini.
7. Teman-teman Ciblon yaitu Rahmatun Nazila, Muh. Nur Aslam, Zishwa Muhammad Jauhar Nafis, Muhammad Fahson Hakim, dan Supardi. Terima kasih sudah menjadi bagian perjalanan hidup selama masa perkuliahan

8. Teman-teman Komunitas Milenial Bergerak, Yola Feby Anzani dan Siti Aryani Nabila Anwar yang sudah membantu memberikan dukungan selama perkuliahan.
9. Teman-teman asrama Pondok Pesantren Inggris Inovasi Bangsa, Fadilla Aisya Putri Virgimena dan Siti Asmariyani yang selalu menyemangati dan menemani penulis dalam menyelesaikan tesis ini.
10. Ala Asmaul Husna, Siti Muslichah, dan teman-teman mahasiswa Prodi Magister Informatika dan Program Beasiswa Santri Berprestasi 2023 Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
11. Annida Rizki Luthfi Astuti. *Especially* apresiasi buat diri saya sendiri karena telah bertanggung jawab untuk menyelesaikan apa yang telah dimulai. Terima kasih karena tidak pantang menyerah dan terus berusaha menyelesaikan perkuliahan ini, serta senantiasa menikmati setiap proses perjalanan *part* kedua yang bisa dibilang tidak mudah. Terima kasih tetap bertahan sampai akhir.
12. Ucapan terima kasih penulis sampaikan kepada semua pihak yang tidak mungkin penulis sebutkan satu demi satu, yang telah banyak memberikan dukungan selama perkuliahan sampai penyelesaian tesis ini.

Semoga Allah SWT melimpahkan Rahmat dan Karunia-Nya kepada kita semua. Penulis berharap semoga tesis ini bermanfaat bagi penulis dan pembaca lainnya.

Wassalamu'alaikum wr. wb.

Yogyakarta, 13 Juni 2025

Penyusun



Annida Rizki Luthfi Astuti

NIM. 23206051021



DAFTAR ISI

THESIS.....	I
HALAMAN PENGESAHAN	II
PERNYATAAN KEASLIAN	III
PERNYATAAN BEBAS PLAGIASI.....	IV
NOTA DINAS PEMBIMBING	V
ABSTRAK.....	VI
MOTTO	VII
HALAMAN PERSEMBAHAN	VIII
KATA PENGANTAR	IX
DAFTAR ISI	XIII
DAFTAR TABEL	XVI
DAFTAR GAMBAR	XVII
DAFTAR SINGKATAN	XVIII
BAB I PENDAHULUAN	1
A. LATAR BELAKANG MASALAH	1
B. RUMUSAN MASALAH.....	8
C. TUJUAN PENELITIAN.....	9
D. MANFAAT PENELITIAN	9
E. BATASAN MASALAH	10
BAB II KAJIAN PUSTAKA DAN LANDASAN TEORI.....	11
A. KAJIAN PUSTAKA	11
B. LANDASAN TEORI.....	21
1. <i>Machine learning</i>	21

2.	<i>Natural Language Processing (NLP)</i>	22
3.	<i>Text Vektorization</i>	23
4.	<i>Teknik Sampling</i>	26
5.	<i>Klasifikasi</i>	28
6.	<i>Evaluasi Kinerja</i>	29
BAB III METODE PENELITIAN		31
A.	ALAT DAN BAHAN	31
B.	DATA PENELITIAN	31
C.	TAHAPAN PENELITIAN	33
1.	<i>Pre-Processing</i>	35
a.	Data Cleaning	35
1).	Mapping Label	35
2).	Remove NaN	36
b.	Text Pre-Processing	37
1).	Regular Expression	37
2).	Case Folding	38
3).	Stopwords	39
4).	Lemmatization	40
5).	Tokenization	41
6).	Label Encoder	42
c.	Teks Vektorisasi	43
1).	Word2Vec	43
2).	GloVe	45
3).	FastText	47
2.	<i>Imbalancing data</i>	49
3.	<i>Training dan Klasifikasi</i>	52
4.	<i>Pengukuran Evaluasi</i>	54

BAB IV HASIL DAN PEMBAHASAN	56
A. HASIL	56
1. <i>Word2Vec</i>	56
2. <i>GloVe</i>	59
3. <i>FastText</i>	62
B. PEMBAHASAN.....	65
BAB V PENUTUP	71
A. KESIMPULAN	71
B. SARAN	72
DAFTAR PUSTAKA	74
DAFTAR RIWAYAT HIDUP	84

DAFTAR TABEL

Tabel 2.1 Daftar Literature	15
Tabel 3.1 Dataset Sebelum dan Sesudah Regular Expression.....	38
Tabel 3.2 Dataset Sebelum dan Sesudah Case Folding.....	38
Tabel 3.3 Dataset Sebelum dan Sesudah Stopwords	39
Tabel 3.4 Dataset Sebelum dan Sesudah Lemmatization.....	40
Tabel 3.5 Dataset Sebelum dan Sesudah Tokenization.....	42
Tabel 4.1 Hasil Akurasi, Presisi, Recall, dan F1-Score pada Word2Vec	56
Tabel 4.2 Hasil Akurasi, Presisi, Recall, dan F1-Score pada GloVe ..	59
Tabel 4.3 Hasil Akurasi, Presisi, Recall, dan F1-Score pada FastText62	
Tabel 4.4 Hasil Perbandingan Akurasi, Presisi, Recall, dan F1-Score pada Word2Vec, GloVe, dan FastText.....	67

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

DAFTAR GAMBAR

Gambar 1 Dataset Playstore	32
Gambar 2 Dataset Appstore.....	32
Gambar 3 Dataset Gabungan Playstore dan Appstore	33
Gambar 4 Flowchart Penelitian	34
Gambar 5 Hasil Mapping Label	36
Gambar 6 Dataset After Remove NaN	37
Gambar 7 Label Encoder Dataset.....	43
Gambar 8 Hasil Vektor Word2Vec pada Dataset.....	45
Gambar 9 Hasil Vektor GloVe pada Dataset.....	47
Gambar 10 Hasil Vektor FastText pada Dataset	49
Gambar 11 Distribusi Dataset Sebelum dan Sesudah SMOTE	50
Gambar 12 Visualisasi Data sebelum SMOTE dan setelah SMOTE dengan Teknik GloVe pada Label Performance.....	52
Gambar 13 Hasil Rata-Rata Confusion Matrix Word2Vec.....	58
Gambar 14 Hasil Rata-Rata Confusion Matrix GloVe.....	61
Gambar 15 Hasil Rata-Rata Confusion Matrix FastText	64
Gambar 16 Grafik Distribusi Akurasi	66
Gambar 17 Grafik Perbandingan Teks Vektorisasi Word2Vec, GloVe, dan FastText	67

DAFTAR SINGKATAN

AI	: Artificial Intelligence
BERT	: Bidirectional Encoder Representations from Transformers
BoW	: Bag of Words
CNN	: Convolutional Neural Network
FR	: Functional Requirements
GloVe	: Global Vectors
LSTM	: Long Short – Term Memory
NFR	: Non-Functional Requirements
NLP	: Natural Language Processing
SMOTE	: Synthetic Minority Oversampling Technique
SVM	: Support Vector Machine
TF-IDF	: Term Frequency – Inverse Document Frequency

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

BAB I

PENDAHULUAN

A. Latar Belakang Masalah

Perkembangan teknologi informasi yang sangat pesat telah mendorong pertumbuhan signifikan pada aplikasi media sosial dan layanan pesan instan. Salah satu aplikasi yang menonjol dalam kategori ini adalah WhatsApp. Pada Maret 2025, WhatsApp memiliki pengguna mencapai tiga miliar pengguna aktif di seluruh dunia (Ceci, 2025). WhatsApp digunakan secara luas oleh Masyarakat dari berbagai usia dan latar belakang. Aplikasi ini banyak digunakan karena kemudahan pada fitur yang ditawarkannya dan dalam melakukan percakapan private ataupun grup (Saepulrohman et al., 2021). Seiring meningkatnya jumlah pengguna dan kompleksitas kebutuhan mereka, penting bagi pengembang perangkat lunak untuk memperoleh pemahaman yang lebih dalam mengenai kebutuhan dan preferensi pengguna secara akurat. Salah satu sumber data yang berpotensi memberikan informasi berharga mengenai kebutuhan pengguna tersebut adalah ulasan pengguna yang tersedia di platform distribusi aplikasi seperti Google PlayStore dan Apple AppStore.

Setiap aplikasi memiliki keunggulan dan kekurangan yang dapat memicu berbagai tanggapan dari pengguna termasuk kepuasan dan kekecewaan yang umumnya diungkapkan melalui ulasan (Riski et al., 2023). Ulasan ini mencerminkan pengalaman langsung pengguna serta harapan dan keluhan terhadap fitur dan performa penggunaan aplikasi WhatsApp. Namun, terdapat tantangan utama dalam analisis kebutuhan

pengguna berbasis teks adalah representasi data teks yang tidak terstruktur dan beragam bentuknya. Pada ulasan pengguna kebanyakan berisikan kalimat yang tidak memiliki arti seperti singkatan, ejaan, gaya penulisan yang bervariasi, serta bahasa yang tidak formal (Nurwanda et al., 2024). Adanya tantangan ini, membutuhkan pendekatan teknologi yang cermat untuk dapat memproses dan menganalisis data secara otomatis.

Dalam konteks ini, kemajuan di bidang *Artificial Intelligence* (AI), khususnya *Natural Language Processing* (NLP) memberikan peluang besar untuk menganalisis data tekstual yang kompleks dan tidak terstruktur. NLP berkembang dalam berbagai topik penelitian dengan bidang yang fokus pada signifikan aspek desain model (Biswas & De, 2022). NLP merupakan bahasa yang dirancang untuk membuat komputer memahami pernyataan dalam bahasa alami yang digunakan manusia (Jain et al., 2018). Selain itu, NLP dirancang untuk memungkinkan komputer menafsirkan dan menghasilkan Bahasa manusia secara natural. Dengan kemajuan teknologi, NLP bermanfaat sebagai alat penting dalam analisis dan pemahaman data tekstual yang dihasilkan oleh pengguna. Perkembangan NLP membuka peluang signifikan untuk meningkatkan kualitas penelitian analisis kebutuhan pengguna berbasis teks. Teknologi ini telah diterapkan pada berbagai bidang seperti ekstraksi informasi, klasifikasi dokumen, analisis sentiment, dan pencarian informasi.

Pada konteks analisis ulasan pengguna aplikasi, NLP memungkinkan system untuk mengidentifikasi kebutuhan pengguna secara otomatis. Dengan menggunakan NLP, ditujukan untuk mengatasi tantangan dalam pemrosesan bahasa seperti ejaan, emoji, gaya

penulisan, dan kemampuan model dalam menangkap sentiment lebih akurat (Nurwanda et al., 2024). Penggunaan NLP mencakup pengumpulan data ulasan pengguna, preprocessing, teks vektorisasi, penerapan algoritma klasifikasi, dan evaluasi hasil. NLP berfungsi dalam mengekstraksi informasi penting dalam bentuk teks seperti subjek, predikat, objek, semantik, dan lainnya (Nur Oktavia et al., 2024). Kemampuan NLP untuk mengekstraksi pola semantic kompleks dari data yang tidak terstruktur memungkinkan identifikasi kebutuhan pengguna yang lebih akurat bahkan catatan data besar seperti ulasan aplikasi yang berisikan ejaan yang tidak baku, penggunaan emoji, dan gaya penulisan informal.

Proses analisis melibatkan beberapa tahapan untuk mengubah data yang telah diproses menjadi data yang siap diolah. Salah satu tantangan utama dalam analisis data teks ialah mengubah data teks menjadi format yang dapat dipahami oleh algoritma *machine learning*. Penerapan teknik vektorisasi yang tepat dapat digunakan untuk mengubah teks menjadi format numerik yang dapat dipahami oleh algoritma *machine learning*. NLP tidak hanya membantu dalam pengolahan data, tetapi dalam pengambilan keputusan berbasis data juga. Penelitian ini menggunakan data yang berasal dari ulasan pengguna aplikasi WhatsApp di Playstore dan Appstore. Data tersebut digunakan untuk analisis kebutuhan pengguna pada aplikasi WhatsApp. Analisis terhadap ulasan pengguna dapat memberikan pengetahuan terkait kebutuhan dan preferensi pengguna.

Dalam konteks pengembangan perangkat lunak, ulasan pengguna dapat dianalisis untuk mengidentifikasi klasifikasi kebutuhan pengguna. Klasifikasi persyaratan melibatkan pengelompokan sejumlah

persyaratan ke dalam kategori atau kelas berdasarkan sistem klasifikasi (Binkhonain & Zhao, 2023). Klasifikasi diperlukan untuk mengolah data sesuai dengan kelompoknya. Dalam perancangan perangkat lunak, kebutuhan yang diperlukan yaitu *Functional Requirements* (FR) dan *Non-Functional Requirements* (NFR). Persyaratan fungsional berkaitan dengan apa yang sistem lakukan, mendefinisikan fungsi atau tindakan yang diinginkan. Sedangkan kebutuhan non-fungsional menangani bagaimana sistem melaksanakan fungsi dengan standar performa, keamanan, ketersediaan, dan aspek lainnya yang tidak bersifat fungsional (Github, n.d.). Proses klasifikasi kebutuhan pengguna ini tidak hanya berguna untuk analisis teknis, tetapi juga dapat memberikan wawasan strategis bagi tim pengembang dalam pengambilan keputusan berbasis data.

Salah satu tahapan penting dalam proses analisis data teks adalah transformasi data dari bentuk teks mentah menjadi representasi numerik yang dapat dipahami oleh algoritma *machine learning*. Pendekatan *machine learning* yang digunakan melibatkan teks vektorisasi. Teks vektorisasi atau penyematan kata merupakan representasi pemetaan kata atau dokumen dalam korpus ke vektor numerik. Ini merupakan bagian dari pemrosesan *natural language processing* dan bertujuan untuk merepresentasikan kata dan dokumen menggunakan representasi numerik yang tersedia (Abubakar et al., 2022). Transformasi ini menjadi kunci utama karena algoritma *machine learning* tidak dapat memahami teks mentah secara langsung. Oleh karena itu, representasi numerik sangat diperlukan agar model dapat mengenali pola dan hubungan dalam data. Tujuan dari penyematan atau pemetaan kata untuk

merepresentasikan teks dengan pemahaman yang lebih baik terkait konteks (Asudani et al., 2023).

Vektorisasi teks mengubah bentuk dari data teks menjadi bentuk vektor sehingga memudahkan dalam pengolahan data. Pemilihan metode teks vektorisasi yang tidak tepat dapat berdampak langsung terhadap akurasi klasifikasi yang dibangun, sehingga menjadi komponen yang sangat penting untuk dianalisis lebih lanjut. Vektorisasi digunakan dalam banyak hal seperti *text mining*, NLP, pencarian informasi, dalam tugas-tugas seperti klasifikasi, pengelompokan, deteksi topik, pembuatan pertanyaan, rekomendasi, *machine translation*, dan lainnya (Mansour et al., 2022). Selain meningkatkan efisiensi pemrosesan, vektorisasi juga memungkinkan model untuk mengenali semantic dan sintaksis dalam teks. Oleh karena itu, pemilihan metode yang tepat sangat mempengaruhi kualitas hasil dari tugas-tugas NLP yang kompleks. Proses ini dapat menyelesaikan permasalahan dalam pemrosesan Bahasa alami (NLP) dengan membantu memetakan tiap kata dalam kosakata ke nilai bilangan bulat tertentu (Singh & Shashi, 2019). Mengingat pentingnya pemilihan metode vektorisasi, berikut ini beberapa pendekatan populer yang sering digunakan dalam penelitian NLP. Beberapa metode vektorisasi teks yang banyak digunakan dan telah terbukti efektif dalam berbagai penelitian adalah Word2Vec, GloVe (Global Vectors), dan FastText. Ketiga metode ini termasuk model terdistribusi yang memiliki wawasan lebih bermakna ke dalam dokumen sekaligus merepresentasikan isinya (Catak et al., 2024).

Word2Vec merupakan kumpulan model dan optimasi arsitektur untuk memeriksa kata-kata dari kumpulan data besar (Asudani et al.,

2023). Word2Vec memiliki keterbatasan yaitu ketidakmampuannya menangani kata-kata di luar kosa kata, karena setiap kata diperlakukan sebagai unit yang berbeda (Moitra, 2024). GloVe merupakan metode untuk memeriksa kata-kata yang efisien dengan melakukan pelatihan dan memiliki keunggulan dalam menangkap fitur bahasa secara global dengan menganalisis kemunculan kata-kata di seluruh korpus (Wang et al., 2020). FastText dapat memasukkan informasi subkata ke dalam proses pembelajaran dan merepresentasikan tiap kata sebagai koleksi n-gram tingkat karakter. Hal ini memungkinkan model FastText untuk menghasilkan penyematan kata yang tidak terlihat dengan memanfaatkan penyematan dari subkata mereka.

Word2Vec, GloVe, dan FastText adalah tiga metode vektorisasi teks yang umum digunakan untuk merepresentasikan teks dalam bentuk numerik. Dengan demikian, Word2vec unggul dengan keefektifannya dalam menangkap hubungan semantic antar kata, namun memiliki keterbatasan dalam menangani kata-kata baru. GloVe mengandalkan informasi global frekuensi kata dalam korpus, sedangkan FastText unggul dalam menangani kata baru. Ketiga metode ini memiliki keunggulan dan kelemahan masing-masing tergantung pada karakteristik data dan konteks penggunaan. Dalam konteks ulasan aplikasi WhatsApp yang seringkali mengandung kata tidak baku, singkatan, emotikon, atau ejaan kreatif. Hal ini menimbulkan tantangan tersendiri dalam pemrosesan teks, terutama dalam hal representasi numerik yang akurat. Oleh karena itu, pemilihan metode teks vektorisasi yang mampu menangkap karakteristik pada ulasan pengguna menjadi krusial dan penting untuk dieksplorasi.

Beberapa penelitian sebelumnya telah membandingkan kinerja metode vektorisasi teks dalam berbagai konteks seperti analisis sentiment atau klasifikasi berita. Pada penelitian yang dilakukan oleh Arliyanti dkk (Nurdin et al., 2020) membandingkan kinerja dari *word embedding* Word2Vec, GloVe, dan FastText dan diklasifikasikan dengan algoritma *Convolutional Neural Network*. Dataset yang digunakan yaitu 20 *Newsgroup* dan *Reuters Newswire*. Evaluasi kinerjanya menggunakan F-measure menghasilkan model FastText unggul dengan nilai F-measure sebesar 0.979 untuk dataset 20 *Newsgroup* dan 0.715 untuk *Reuters*. Penelitian lainnya yang dilakukan oleh Siti Khomsah dkk (Khomsah et al., 2022) membandingkan akurasi antara Word2Vec dan FastText pada analisis sentiment ulasan hotel. Penelitian ini menggunakan dataset dari TripAdvisor. Kedua metode yang digunakan yaitu Word2Vec dan FastText menggunakan parameter yang sama. Kemudian digabungkan dengan ensemble learning *Random Forest*, *Extra Tree*, dan *AdaBoost*. Penelitian ini menghasilkan kinerja akurasi *Fasttext-Extra Tree* dan *Fasttext-Random Forest* lebih unggul dengan akurasi mencapai 93%. Namun, dari penelitian-penelitian tersebut tidak secara khusus ada yang membahas terkait klasifikasi kebutuhan pengguna dari ulasan aplikasi WhatsApp pada PlayStore dan AppStore. Hal ini menunjukkan bahwa masih terdapat gap penelitian dalam penggunaan dan evaluasi metode vektorisasi teks untuk klasifikasi kebutuhan pengguna pada ulasan WhatsApp di platform Playstore dan Appstore.

Berdasarkan latar belakang dan gap penelitian yang telah diidentifikasi, penelitian ini bertujuan untuk mengetahui performa terbaik dari tiga metode teks vektorisasi, yaitu Word2Vec, GloVe, dan

FastText terhadap kinerja klasifikasi kebutuhan pengguna pada dataset yang digunakan. Penelitian ini dilakukan karena masih sedikitnya penelitian terkait perbandingan metode teks vektorisasi pada kinerja klasifikasi kebutuhan pengguna. Penelitian sebelumnya banyak menggunakan teks vektorisasi untuk perbandingan analisis sentiment atau klasifikasi berita. Dataset yang digunakan pada penelitian ini yaitu dataset hasil scraping ulasan WhatsApp pada Playstore dan Appstore. Penelitian ini menggunakan pendekatan klasifikasi berbasis algoritma *Random Forest*. Penggunaan *Random Forest* untuk klasifikasi dikarenakan dapat digunakan untuk menangani kumpulan data kompleks dan besar, serta memiliki ketahanan terhadap adanya overfitting (Bratijaya, 2024). Selain itu, penelitian ini juga menggunakan teknik *imbalancing data* untuk mengatasi ketidakseimbangan kelas dalam dataset. Teknik yang digunakan yaitu dengan SMOTE. Adapun evaluasi yang digunakan yaitu *confusion matrix*, *accuracy*, *precision*, *recall*, dan *f1-score*. Dari tiga metode teks vektorisasi, algoritma klasifikasi, dan teknik *imbalancing data* yang digunakan akan dilakukan perbandingan. Dengan pendekatan ini, penelitian diharapkan dapat memberikan kontribusi praktis dan teoritis dalam pemilihan metode representasi teks terbaik untuk klasifikasi kebutuhan pengguna berbasis data ulasan aplikasi.

B. Rumusan Masalah

Berdasarkan Latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Apa jenis model teks vektorisasi yang berpotensi memberikan hasil terbaik terhadap kinerja klasifikasi kebutuhan pengguna pada dataset WhatsApp Playstore dan Appstore?
2. Bagaimana perbedaan hasil klasifikasi kebutuhan pengguna WhatsApp di Playstore dan Appstore berdasarkan model teks vektorisasi (Word2Vec, GloVe, dan FastText)?

C. Tujuan Penelitian

1. Menentukan model teks vektorisasi yang memiliki tingkat kesalahan terendah dan performa terbaik berdasarkan confusion matrix.
2. Menganalisis dan membandingkan performa model teks vektorisasi (Word2Vec, GloVe, dan FastText) dalam klasifikasi kebutuhan pengguna WhatsApp di Appstore dan Playstore berdasarkan grafik perbandingan akurasi, presisi, recall, dan f1-score.

D. Manfaat Penelitian

1. Memberikan wawasan mengenai model teks vektorisasi yang memiliki Tingkat kesalahan terendah dan performa terbaik untuk klasifikasi kebutuhan pengguna WhatsApp di Appstore dan Playstore.
2. Memberikan gambaran terkait perbandingan model teks vektorisasi Word2Vec, GloVe, dan FastText, serta penggunaan algoritma klasifikasi *Random Forest* dan Teknik *imbalancing data* SMOTE dalam melakukan klasifikasi kebutuhan pengguna.

E. Batasan Masalah

Untuk menghindari pembahasan yang meluas, maka perlu adanya pembatasan terhadap penelitian. Batasan masalah pada penelitian ini adalah:

1. Data yang digunakan adalah dataset WhatsApp Appstore dan Playstore.
2. Model teks vektorisasi yang digunakan pada penelitian ini adalah Word2Vec, GloVe, dan FastText.
3. Algoritma klasifikasi *machine learning* yang digunakan adalah *Random Forest*.
4. Jenis pengukuran evaluasi yang digunakan adalah *confusion matrix*, *accuracy*, *precision*, *recall*, dan *f1-score*.



BAB V

PENUTUP

A. Kesimpulan

Berdasarkan hasil evaluasi yang dilakukan, penelitian ini berhasil mencapai tujuan yang telah ditetapkan, dengan kesimpulan sebagai berikut:

1. Model teks vektorisasi yang memiliki Tingkat kesalahan terendah dan performa terbaik adalah GloVe. Pada hasil *confusion matrix* terdapat beberapa label kelas yang memiliki hasil prediksi benar secara akurat dan terdapat satu label kelas yaitu pada label 2 (*functional*) yang memiliki prediksi benar sedikit. Label 2 memiliki kemiripan fitur-fitur dengan beberapa label kelas lainnya sehingga model mengalami kesulitan dalam membedakan fitur-fitur tersebut. GloVe unggul dikarenakan dataset yang digunakan pada penelitian ini didominasi pola repetitif atau berulang yang memungkinkan GloVe memanfaatkan pendekatan *statistic global co-occurrence* secara optimal sehingga menghasilkan representasi vektor kata yang lebih stabil dan informatif untuk kebutuhan klasifikasi.
2. Perbandingan performa model teks vektorisasi dalam klasifikasi kebutuhan pengguna pada ulasan WhatsApp berdasarkan grafik perbandingan akurasi, presisi, recall, dan f1-score menunjukkan hasil adanya perbedaan yang signifikan terhadap performa kinerja ketiga model teks vektorisasi yang digunakan. Pada model GloVe grafik hasil metrik *accuracy*, *precision*, *recall*, dan *f1-score* lebih tinggi dibandingkan Word2Vec dan FastText.

GloVe memiliki kinerja akurasi yang bagus sebesar 95.05% dibandingkan pada akurasi model Word2Vec 93.75% dan FastText 93.73%. Word2Vec dan FastText memiliki perbedaan yang tidak terlalu signifikan.

B. Saran

Berdasarkan hasil analisis yang telah dilakukan, penggunaan model teks vektorisasi, klasifikasi dengan *Random Forest*, dan *imbalancing data* SMOTE memberikan dampak yang cukup signifikan terhadap performa evaluasi. Evaluasi menggunakan confusion matrix juga dapat menunjukkan performa model dalam prediksi data. Selain itu metrik dengan *accuracy*, *precision*, *recall*, dan *f1-score* juga membantu dalam memberikan perbandingan performa kinerja hasil evaluasi antara Word2Vec, GloVe, dan FastText. Oleh karena itu, terdapat beberapa rekomendasi yang dapat menjadi acuan untuk pengembangan penelitian selanjutnya, antara lain:

1. Mengeksplorasi penggunaan teknik *fine-tuning* model untuk membantu memperbaiki kesalahan prediksi dan meningkatkan representasi vektor dari kata-kata atau kalimat sehingga model lebih sensitif terhadap perbedaan antar label.
2. Eksplorasi model teks vektorisasi lain yang lebih baru atau menggabungkan beberapa metode, seperti kombinasi Word2Vec dan GloVe untuk meningkatkan representasi fitur teks dan performa klasifikasi.

3. Eksplorasi algoritma klasifikasi lain seperti XGBoost, SVM, atau deep learning untuk membandingkan efektivitasnya dalam klasifikasi kebutuhan pengguna.
4. Penanganan *imbalancing data* yang lebih variatif selain SMOTE, bisa menggunakan ADASYN atau Random Under Sampling untuk melihat pengaruhnya terhadap performa model.



DAFTAR PUSTAKA

- Abubakar, H. D., Umar, M., & Bakale, M. A. (2022). Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology*, 4(1&2), 27–33. <https://doi.org/10.56471/slujst.v4i.266>
- Adrian, M. G., Prasetyowati, S. S., & Sibaroni, Y. (2023). Effectiveness of Word Embedding GloVe and Word2Vec within News Detection of Indonesian uUsing LSTM. *Jurnal Media Informatika Budidarma*, 7(3), 1180. <https://doi.org/10.30865/mib.v7i3.6411>
- Af'idah, D. I., Dairoh, D., Handayani, S. F., & Pratiwi, R. W. (2021). Pengaruh Parameter Word2Vec terhadap Performa Deep Learning pada Klasifikasi Sentimen. *Jurnal Informatika: Jurnal Pengembangan IT*, 6(3), 156–161. <https://doi.org/10.30591/jpit.v6i3.3016>
- Ahmadi, A., Sharif, S. S., & Banad, Y. M. (2024). A Comparative Study of Sampling Methods with Cross-Validation in the FedHome Framework. *ArXiv Preprint ArXiv:2406.01950*, 1–10. <https://arxiv.org/abs/2406.01950> <https://arxiv.org/pdf/2406.01950>
- Ahsan, M., Gomes, R., & Denton, A. (2018). SMOTE Implementation on Phishing Data to Enhance Cybersecurity. *IEEE International Conference on Electro Information Technology, 2018-May(October)*, 531–536. <https://doi.org/10.1109/EIT.2018.8500086>
- Almeida, F., & Xexéo, G. (2019). *Word Embeddings: A Survey. 1991*.

<http://arxiv.org/abs/1901.09069>

- Amaliah, F., & Dwi Nuryana, I. K. (2022). Perbandingan Akurasi Metode Lexicon Based Dan Naive Bayes Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter. *Journal of Informatics and Computer Science (JINACS)*, 3(03), 384–393. <https://doi.org/10.26740/jinacs.v3n03.p384-393>
- Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: a review. In *Artificial Intelligence Review* (Vol. 56, Issue 9). Springer Netherlands. <https://doi.org/10.1007/s10462-023-10419-1>
- Ayuningtyas, P., & Tantyoko, H. (2024). Perbandingan Metode Word2vec Model Skipgram pada Ulasan Aplikasi Linkaja menggunakan Algoritma Bidirectional LSTM dan Support Vector Machine. *JUSTIN (Jurnal Sistem Dan Teknologi Informasi)*, 12(1), 189–196. <https://doi.org/10.26418/justin.v12i1.72530>
- Binkhonain, M., & Zhao, L. (2023). A machine learning approach for hierarchical classification of software requirements. *Machine Learning with Applications*, 12(September 2022), 100457. <https://doi.org/10.1016/j.mlwa.2023.100457>
- Biswas, R., & De, S. (2022). A Comparative Study on Improving Word Embeddings Beyond Word2Vec and GloVe. *PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing*, 113–118. <https://doi.org/10.1109/PDGC56933.2022.10053200>
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching

- Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.
https://doi.org/10.1162/tacl_a_00051
- Bratijaya, P. (2024). *Unleashing the Power of Random Forest: Why it Outperforms Decision Trees and Expert Rules*.
<https://pratikbarjatya.medium.com/unleashing-the-power-of-random-forest-why-it-outperforms-decision-trees-and-expert-rules-472a9bea1b8a>
- Canedo, E. D., & Mendes, B. C. (2020). Software requirements classification using machine learning algorithms. *Entropy*, 22(9).
<https://doi.org/10.3390/E22091057>
- Catak, T., Durdu, P. O., & Omurca, S. I. (2024). Enhancing Agile Effort Estimation: An NLP Approach for Software Requirements Analysis. *HORA 2024 - 6th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*.
<https://doi.org/10.1109/HORA61326.2024.10550870>
- Ceci, L. (2025). *Jumlah pengguna WhatsApp aktif bulanan di seluruh dunia dari April 2013 hingga Maret 2025*.
<https://www.statista.com/statistics/260819/number-of-monthly-active-whatsapp-users/>
- Dewi, S. (2024). *Arus Jurnal Sains dan Teknologi (AJST) Analisis Sentimen Aplikasi WhatsApp berdasarkan Ulasan di*. 2(2).
- Dewi, T. A., & Mailoa, E. (2023). Perbandingan Implementasi Metode Smote Pada Algoritma Support Vector Machine (Svm) Dalam Analisis Sentimen Opini Masyarakat Tentang Mixue. *Jurnal Indonesia : Manajemen Informatika Dan Komunikasi*, 4(3), 849–

855. <https://doi.org/10.35870/jimik.v4i3.289>
- Dharma, E. M., Gaol, F. L., Warnars, H. L. H. S., & Soewito, B. (2022). the Accuracy Comparison Among Word2Vec, Glove, and Fasttext Towards Convolution Neural Network (Cnn) Text Classification. *Journal of Theoretical and Applied Information Technology*, 100(2), 349–359.
- Diantika, S. (2023). *PENERAPAN TEKNIK RANDOM OVERSAMPLING UNTUK MENGATASI IMBALANCE CLASS DALAM KLASIFIKASI WEBSITE PHISHING*. 7(1), 19–25.
- Držić, D., & Šteflovich, K. (2023). Text Vectorization Techniques Based on Wordnet. *Jazykovedny Casopis*, 74(1), 310–322. <https://doi.org/10.2478/jazcas-2023-0048>
- Ebrahimi, F., Tushev, M., & Mahmoud, A. (2022). Classifying Mobile Applications Using Word Embeddings. *ACM Transactions on Software Engineering and Methodology*, 31(2). <https://doi.org/10.1145/3474827>
- Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, 113(7), 4903–4923. <https://doi.org/10.1007/s10994-022-06296-4>
- Fajriyah, N., Lapatta, N. T., Nugraha, D. W., & Laila, R. (2025). *Implementasi svm dan smote pada analisis sentimen media sosial x terhadap pelantikan agus harimurti yudhoyono*. 10(2), 1359–1370.
- Github. (n.d.). <https://github.com/AleksandarMitrevski/se-requirements-classification/tree/master/0-datasets>
- Halim, A. M., Dwifabri, M., & Nhita, F. (2023). Handling Imbalanced

- Data Sets Using SMOTE and ADASYN to Improve Classification Performance of Ecoli Data Sets. *Building of Informatics, Technology and Science (BITS)*, 5(1), 246–253. <https://doi.org/10.47065/bits.v5i1.3647>
- Halim, S. F. N., & Azmi, U. (2023). Comparative Analysis of Classification and Application of SMOTE Technique in Imbalanced Data on Credit Card Defaults. *Jurnal Sains Dan Seni ITS*, 12(2). <https://doi.org/10.12962/j23373520.v12i2.111833>
- Indrawati, A. (2021). Comparative Analysis of Classification and Application of SMOTE Technique in Imbalanced Data on Credit Card Default Application of Oversampling and Undersampling Combination Techniques Hybrid Oversampling and Undersampling Techniques To Handling Imbalanced. *JIKO (Jurnal Informatika Dan Komputer)*, 4(1), 38–43. <https://doi.org/10.33387/jiko>
- Ismail, E., Gad, W., & Hashem, M. (2023). SMOTE-RUS : COMBINED OVERSAMPLING AND UNDERSAMPLING TECHNIQUE TO CLASSIFY THE IMBALANCED AUTISM SPECTRUM. 23(3), 83–94. <https://doi.org/10.21608/ijicis.2023.216833.1278>
- Jain, A., Kulkarni, G., & Shah, V. (2018). Natural Language Processing. *International Journal of Computer Sciences and Engineering*, 5(10), 112–124. <https://doi.org/10.26438/ijcse/v6i1.161167>
- Khomsah, S., Ramadhani, R. D., & Wijayanto, S. (2022). The Accuracy Comparison Between Word2Vec and FastText On Sentiment Analysis of Hotel Reviews. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 6(3), 352–358. <https://doi.org/10.29207/resti.v6i3.3711>

- Ma'rifah, H., Wibawa, A. P., & Akbar, M. I. (2020). Classification of Scientific Articles with Various Preprocessing Scenarios. *Sains, Aplikasi, Komputasi Dan Teknologi Informasi*, 2(2), 70. <https://doi.org/10.30872/jsakti.v2i2.2681>
- Mahesh, B. (2019). *Machine Learning Algorithms - A Review | Enhanced Reader*. October. <https://doi.org/10.21275/ART20203995>
- Mansour, A., Mohammad, J., & Kravchenko, Y. (2022). Text Vectorization Method Based on Concept Mining Using Clustering Techniques. *2022 6th International Conference on Information Technologies in Engineering Education, Inforino 2022 - Proceedings*, 1–10. <https://doi.org/10.1109/Inforino53888.2022.9782908>
- Moitra, A. (2024). Efficient Unicode Ordinal Values for Text Embedding with FastText and Word2Vec. *19th International Joint Symposium on Artificial Intelligence and Natural Language Processing, ISAI-NLP 2024*, 1–6. <https://doi.org/10.1109/ISAI-NLP64410.2024.10799464>
- Musfiroh, Tholib, A., & Arifin, Z. (2024). Analisis Sentimen Terhadap Ulasan Aplikasi Shopee di Google Play Store Menggunakan Metode TF-IDF dan Long Short-Term Memory (LSTM). 6(2), 371–381. <https://doi.org/10.33650/jeecom.v4i2>
- Nur Oktavia, A., Iqbal, M., Saputra, R. W., Zulfikar, M. I., & Saifudin, A. (2024). Implementation of Natural Language Processing Methods in Analysis Studies. *Jurnal Ilmiah Ilmu Komputer Dan Multimedia*, 2(1), 154–159. <https://jurnalmahasiswa.com/index.php/biikma>

- Nurdin, A., Anggo Seno Aji, B., Bustamin, A., & Abidin, Z. (2020). Comparison of Word Embedding Performance of Word2Vec, Glove, and Fasttext on Text Classification. *Jurnal Tekno Kompak*, 14(2), 74. <https://doi.org/10.33365/jtk.v14i2.732>
- Nurwanda, Suarna, N., & Prihartono, W. (2024). Penerapan NLP (Natural Language Processing) dalam Analisis Sentimen Pengguna Telegram di Playstore. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(2), 1841–1846. <https://doi.org/10.36040/jati.v8i2.8469>
- Or, B. (2025). *Improving Requirements Classification with SMOTE-Tomek Preprocessing*. 1–8. <http://arxiv.org/abs/2501.06491>
- Oryza Habibie Rahman, Gunawan Abdillah, & Agus Komarudin. (2021). Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(1), 17–23. <https://doi.org/10.29207/resti.v5i1.2700>
- Padurariu, C., & Breaban, M. E. (2019). Dealing with data imbalance in text classification. *Procedia Computer Science*, 159, 736–745. <https://doi.org/10.1016/j.procs.2019.09.229>
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *AES: Journal of the Audio Engineering Society*, 19(5), 1532–1543.
- Prastyo, A. P., Yunizar, A., & Yusuf, P. (2024). *Sentiment Analysis of LinkedIn User Reviews on the Google Playstore Application with the Decision Tree Method*. 8(1).
- Putra, A. H., & Salam, A. (2025). *A Comparative Performance of SMOTE , ADASYN and Random Oversampling in Machine*

- Learning Models on Prostate Cancer Dataset*. 9(3), 603–610.
- Rajamani, S. K., T, K., P, P. K., & M, S. S. (2023). Ensemble Text Classification with TF-IDF Vectorization for Hate Speech Detection in Social Media. *2023 International Conference on System, Computation, Automation and Networking, ICSCAN 2023, March*. <https://doi.org/10.1109/ICSCAN58655.2023.10395354>
- Rani, D., Kumar, R., & Chauhan, N. (2022). Study and Comparison of Vectorization Techniques Used in Text Classification. *2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022*, 1–6. <https://doi.org/10.1109/ICCCNT54827.2022.9984608>
- Regex*. (2024). <https://www.computerhope.com/jargon/r/regex.htm>
- Rianto, Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data*, 8(1), 1–17. <https://doi.org/10.1186/s40537-021-00413-1>
- Rifaldy, F., Sibaroni, Y., & Prasetyowati, S. S. (2025). *Effectiveness of word2vec and tf-idf in sentiment classification on online investment platforms using support vector machine 1*. 10(2), 863–874.
- Riski, M., Fikry, M., & Yusra. (2023). Klasifikasi Sentimen Ulasan Aplikasi WhatsApp di Play Store Menggunakan Metode K-Nearest Neighbor. *Media Online*, 4(1), 438–444. <https://doi.org/10.30865/klik.v4i1.1050>
- Rollo, F., Bonisoli, G., & Po, L. (2024). A Comparative Analysis of Word Embeddings Techniques for Italian News Categorization. *IEEE Access*, 12(January), 25536–25552. <https://doi.org/10.1109/ACCESS.2024.3367246>

- Saepulrohman, A., Saepudin, S., & Gustian, D. (2021). Analisis Sentimen Kepuasan Pengguna Aplikasi Whatsapp Menggunakan Algoritma Naïve Bayes Dan Support Vector Machine. *Is The Best Accounting Information Systems and Information Technology Business Enterprise This Is Link for OJS Usf@*, 6(2), 91–105. <https://doi.org/10.34010/aisthebest.v6i2.4919>
- Sahamony, N. F., Terttiaavini, T., & Rianto, H. (2024). Analisis Perbandingan Kinerja Model Machine Learning untuk Memprediksi Risiko Stunting pada Pertumbuhan Anak. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(2), 413–422. <https://doi.org/10.57152/malcom.v4i2.1210>
- Shaday, E. N., Engel, V. J. L., & Heryanto, H. (2024). Application of the Bidirectional Long Short-Term Memory Method with Comparison of Word2Vec, GloVe, and FastText for Emotion Classification in Song Lyrics. *Procedia Computer Science*, 245(C), 137–146. <https://doi.org/10.1016/j.procs.2024.10.237>
- Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, 121(March 2023), 102342. <https://doi.org/10.1016/j.is.2023.102342>
- Singh, A. K., & Shashi, M. (2019). Vectorization of Text Documents for Identifying Unifiable News Articles. *International Journal of Advanced Computer Science and Applications*, 10(7), 305–310. <https://doi.org/10.14569/ijacsa.2019.0100742>
- Sudrajat, M. R., & Zakariyah, M. (2024). *Penerapan Natural Language*

Processing dan Machine Learning untuk Prediksi Stres Siswa SMA Berdasarkan Analisis Teks. 6(3).

<https://doi.org/10.47065/bits.v6i3.6180>

Suneera, C. M., & Prakash, J. (2020). Performance Analysis of Machine Learning and Deep Learning Models for Text Classification. *2020 IEEE 17th India Council International Conference, INDICON 2020*. <https://doi.org/10.1109/INDICON49873.2020.9342208>

Supian, A., Revaldo, B. T., Marhadi, N., Rahmaddeni, & Efrizoni, L. (2024). Application of SVM and Word2Vec for Sentiment Analysis of DANA App User Reviews. *Jurnal Ilmiah KOMPUTASI*, 23(4), 319–330.

Wang, C., Nulty, P., & Lillis, D. (2020). A Comparative Study on Word Embeddings in Deep Learning for Text Classification. *ACM International Conference Proceeding Series*, 37–46. <https://doi.org/10.1145/3443279.3443304>

Wibawa, A. P., Guntur, M., Purnama, A., Akbar, M. F., & Dwiyanto, F. A. (2018). *Metode-metode Klasifikasi*. 3(1), 134–138.

Widayat, W. (2021). Analisis Sentimen Movie Review menggunakan Word2Vec dan metode LSTM Deep Learning. *Jurnal Media Informatika Budidarma*, 5(3), 1018. <https://doi.org/10.30865/mib.v5i3.3111>

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information (Switzerland)*, 14(1). <https://doi.org/10.3390/info14010054>