Perbandingan Word2vec dan *FastText* Pada *Long Short-Term Memory* Dalam Analisis *Semantic Shift*

SKRIPSI

OLEH:

MARDIATUL HASANAH

228160009

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MEDAN AREA
MEDAN
2026

Perbandingan *Word2vec* dan *FastText* Pada *Long Short-Term Memory* Dalam Analisis *Semantic Shift*

SKRIPSI

Diajukan sebagai Salah Satu Syarat untuk Memperoleh
Gelar Sarjana di Fakultas Teknik
Universitas Medan Area

Oleh:

MARDIATUL HASANAH

228160009

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MEDAN AREA
MEDAN
2026

HALAMAN PENGESAHAN

Judul Skripsi	: Perbandingan *Word2vec* dan *FastText* Pada *Long Short-Term Memory* Dalam Analisis *Semantic Shift*
Nama	: Mardiatul Hasanah
NPM	: 228160009
Fakultas	: Teknik

Disetujui Oleh
Komisi Pembimbing

Dr. Arnes Sembiring, ST, M.Kom
Dosen Pembimbing

Dr. Eng. [illegible]riatno.ST, MT.
Dekan Fakultas Teknik

[illegible], S.Kom., M.Kom
Ka. Prodi

Tanggal Lulus: Kamis, 12 Maret 2026

HALAMAN PERNYATAAN

Saya menyatakan bahwa skripsi yang saya susun, sebagai syarat memperoleh gelar serjana merupakan hasil karya tulis saya sendiri. Adapun bagian-bagian tertentu dalam penulisan skripsi ini yang saya kutip dari hasil karya orang lain telah dituliskan sumbernya secara jelas sesuai dengan norma, kaidah, dan etika penulisan ilmiah.

Saya bersedia menerima sanksi pencabutan gelar akademik yang saya peroleh dan sanksi-sanksi lainnya dengan peraturan yang berlaku, apabila di kemudian hari ditemukan adanya plagiat dalam skripsi ini.

Medan, (23 Mei 2026)

Mardiatul Hasanah
228160009

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR/SKRIPSI/TESIS UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Medan Area, saya yang bertanda tangan di bawah ini:

Nama	: Mardiatul Hasanah
NPM	: 228160009
Program Studi	: Teknik Informatika
Fakultas	: Teknik
Jenis Karya	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Medan Area **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas karya ilmiah saya yang berjudul :

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Medan Area berhak menyimpan, mengalihmedia/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan Choose an item. saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di: Medan
Pada tanggal: 23 Mei 2026
Yang menyatakan

(Mardiatul Hasanah)

RIWAYAT HIDUP

Penulis bernama Mardiatul Hasanah, lahir di Medan pada tanggal 13 Oktober 2003. Penulis menyelesaikan pendidikan sekolah dasar di SD Hangtuah 2 Titipapan dan lulus pada tahun 2015, kemudian melanjutkan ke MTs Ar-Raudhatul Hasanah, Medan dan lulus pada tahun 2018. Setelah itu penulis menyelesaikan pendidikan menengah atas di MA Ar-Raudhatul Hasanah, Medan dan lulus pada tahun 2021. Pada tahun 2022 penulis diterima sebagai mahasiswa di Program Studi Teknik Informatika, Fakultas Teknik, Universitas Medan Area. Selama menempuh pendidikan, penulis aktif dalam berbagai kegiatan akademik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana S. Kom.

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul **"Perbandingan *Word2vec* dan *FastText* Pada *Long Short-Term Memory* Dalam Analisis *Semantic Shift*"** dengan baik, sebagai syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Teknik Informatika, Universitas Medan Area. Dalam proses penyusunan skripsi ini, penulis menyadari bahwa terdapat banyak pihak yang telah memberikan bantuan, dukungan, serta bimbingan, dengan itu penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Bapak Dr. Eng. Supriatno, ST, MT. selaku Dekan Fakultas Teknik Universitas Medan Area.
2. Bapak Rizki Muliono, S.Kom., M.Kom selaku Ketua Program Studi Teknik Informatika Universitas Medan Area.
3. Bapak Dr. Arnes Sembiring, ST, M.Kom selaku Dosen Pembimbing yang telah memberikan arahan, bimbingan, dan motivasi selama proses penyusunan skripsi ini.
4. Bapak Dr. Hartono, S.Kom, M.Kom selaku Dosen Pembimbing Akademik
5. Kedua orang tua tercinta yang telah memberikan kasih sayang, doa, dan dukungan yang tidak ternilai selama hidupnya, serta keluarga yang selalu memberikan semangat dan dukungan.
6. Seluruh dosen Program Studi Teknik Informatika Universitas Medan Area, khususnya para panitia selama proses kegiatan skripsi, yang telah memberikan ilmu pengetahuan selama masa perkuliahan.
7. Teman-teman seperjuangan skripsi dan untuk semua pihak yang tidak dapat disebutkan satu per satu, yang telah memberikan dukungan, motivasi, dan kebersamaan selama proses penyusunan.

Penulis sudah berupaya untuk menulis dan menyusun skripsi ini secara sistematis dan sebaik mungkin, Namun, penulis tetap menyadari bahwa skripsi ini masih memiliki keterbatasann.Oleh karena itu, penulis berharap saran yang membangun dan semoga skripsi ini dapat memberikan manfaat.

Medan, 23 Mei 2026
Penulis

(Mardiatul Hasanah)

ABSTRAK

Perkembangan media sosial menyebabkan penggunaan bahasa menjadi semakin dinamis sehingga banyak kata memiliki makna dan polaritas sentimen yang berbeda tergantung pada konteks kalimat. Dalam kajian linguistik, kondisi ini berkaitan dengan *semantic shift*, khususnya *contextual polarity shift*. Kondisi tersebut menjadi tantangan dalam analisis sentimen berbasis *machine learning* karena model perlu memahami perubahan makna kata berdasarkan konteks penggunaannya. Penelitian ini bertujuan membandingkan performa *Word2Vec* dan *FastText* pada model *Long Short-Term Memory* (LSTM) dalam analisis *semantic shift* pada teks media sosial berbahasa Indonesia. Dataset penelitian dibangun melalui integrasi beberapa dataset Kaggle, yaitu *INA Tweets PPKM Dataset*, *Indonesian Hate Speech Dataset*, dan SMSA. Data melalui tahap prapemrosesan berupa *cleaning*, normalisasi, penghapusan data duplikat, serta penyaringan kalimat pendek. Penelitian ini juga menambahkan variasi kalimat yang merepresentasikan *contextual polarity shift* pada data pelatihan. Proses klasifikasi dilakukan menggunakan dua model arsitektur LSTM dengan *embedding Word2Vec* dan *FastText* . Evaluasi model menggunakan *accuracy, precision, recall, Macro F1-score, confusion matrix,* dan *Shift Robustness Score.* Hasil penelitian menunjukkan bahwa model LSTM dengan *FastText* memperoleh performa terbaik dengan *accuracy* sebesar 0.8945 dan *Macro F1-score* sebesar 0.8909, lebih tinggi dibandingkan *Word2Vec*. *FastText* juga menghasilkan *Shift Robustness Score* sebesar 1.0098, sedangkan *Word2Ve*c sebesar 1.0026. Hasil tersebut menunjukkan bahwa *FastText* lebih efektif dalam memahami *semantic shift* pada teks media sosial berbahasa Indonesia. Model terbaik kemudian digunakan untuk menganalisis opini masyarakat terhadap isu 17+8 Tuntutan Rakyat pada data komentar TikTok dan diimplementasikan dalam aplikasi *Mobile*.

Kata Kunci: Analisis sentimen*, Semantic Shift, LSTM, Word2Vec, FastText* .

ABSTRACT

Social media has led to increasingly dynamic language use, where words may express different meanings and sentiment polarities depending on context. In linguistics, this phenomenon is associated with semantic shift, particularly contextual polarity shift. This creates challenges for machine learning-based sentiment analysis because models must capture contextual variations in word meaning. This study compares Word2Vec and FastText embeddings within a Long Short-Term Memory (LSTM) model for semantic shift analysis on Indonesian social media text. The dataset was constructed by integrating several Kaggle datasets, including the INA Tweets PPKM Dataset, Indonesian Hate Speech Dataset, and SMSA. Data preprocessing involved cleaning, normalization, duplicate removal, and filtering short sentences. Additionally, synthetic variations representing contextual polarity shift were added to the training data. The classification process employed two LSTM architectures using Word2Vec and FastText embeddings. Model evaluation used accuracy, precision, recall, Macro F1-score, confusion matrix, and Shift Robustness Score. Results show that LSTM with FastText achieved the best performance, with an accuracy of 0.8945 and Macro F1-score of 0.8909, outperforming Word2Vec. FastText also achieved a Shift Robustness Score of 1.0098 compared to 1.0026 for Word2Vec. These findings indicate that FastText is more effective in capturing semantic shift in Indonesian social media text. The best model was then used to analyze public opinion on the 17+8 People's Demands issue from TikTok comments and implemented in a mobile applications.

***Keywords*:** *Sentiment Analysis, Semantic Shift, LSTM, Word2Vec, FastText .*

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN

Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Media sosial menghasilkan aliran informasi dalam jumlah besar setiap hari berupa opini, komentar, dan respons masyarakat terhadap berbagai isu yang sedang berkembang. Data teks yang dihasilkan bersifat tidak terstruktur, bervolume besar, dan terus bertambah secara *real-time* sehingga analisis secara manual menjadi kurang efisien. Kondisi tersebut menjadikan analisis sentimen sebagai salah satu pendekatan penting untuk mengelompokkan opini masyarakat ke dalam kategori positif, netral, dan negatif secara otomatis (Darmansyah et al., 2024).

Analisis sentimen dapat membantu memahami kecenderungan opini publik terhadap suatu isu sehingga informasi yang diperoleh menjadi lebih mudah dianalisis secara cepat dan sistematis (Mao et al., 2024). Media sosial juga menjadi sarana utama masyarakat dalam menyampaikan pandangan terhadap berbagai peristiwa sosial, politik, maupun hiburan. Bahasa yang digunakan pada media sosial umumnya bersifat informal, dinamis, dan dipengaruhi oleh tren komunikasi *digital*.

Kondisi tersebut menyebabkan satu kata dapat memiliki polaritas sentimen yang berbeda tergantung pada konteks kalimat yang digunakan. Fenomena perubahan makna berdasarkan konteks penggunaan kata dalam komunikasi digital dapat dikategorikan sebagai *Semantic Shift*, khususnya dalam bentuk perubahan polaritas sentimen (*contextual polarity shift*) pada analisis sentimen. Pada beberapa kondisi, kata yang sama dapat digunakan dalam konteks positif, netral, maupun

negatif sehingga model analisis sentimen perlu memahami hubungan konteks antar kata dalam suatu kalimat agar dapat menghasilkan klasifikasi yang lebih akurat (Baes et al., 2024).

Untuk menangani permasalahan tersebut, pendekatan *Natural Language Processing* (NLP) berbasis *deep learning* semakin banyak digunakan karena mampu memahami hubungan kontekstual antar kata secara lebih baik dibandingkan metode tradisional. Salah satu model yang banyak diterapkan dalam analisis sentimen adalah LSTM, yang merupakan pengembangan dari *Recurrent Neural Network* (RNN). Model ini dirancang untuk mempelajari hubungan sekuensial pada data teks sehingga mampu mempertahankan informasi penting dalam urutan kata dan memahami konteks kalimat secara lebih efektif (Rizky et al., 2024).

Kemampuan tersebut membuat LSTM dinilai sesuai untuk menangani variasi bahasa informal pada media sosial yang sering mengandung ambiguitas konteks sentimen (Siddiqui et al., 2025). Dalam implementasinya, model LSTM memerlukan representasi kata dalam bentuk numerik atau *embedding* agar hubungan antar kata dapat dipelajari oleh model. *Embedding* memungkinkan setiap kata direpresentasikan sebagai vektor yang mengandung informasi hubungan konteks antar kata dalam teks (Olakangil et al., 2023).

Metode *embedding* yang umum digunakan adalah *Word2Vec* dan *FastText* . *Word2Vec* membangun representasi kata berdasarkan hubungan konteks antar kata dalam data pelatihan, sedangkan *FastText* memanfaatkan informasi *subword* sehingga lebih mampu mengenali variasi bentuk kata pada bahasa informal media sosial (Al-Tarawneh et al., 2024). Pendekatan tersebut membuat *FastText* dinilai

lebih adaptif dalam menangani variasi penulisan kata yang sering muncul pada media sosial (Karakaya & Kilimci, 2024).

Beberapa penelitian sebelumnya telah menerapkan model LSTM dengan berbagai metode *embedding* untuk klasifikasi sentimen. Namun, sebagian besar penelitian masih berfokus pada performa klasifikasi secara umum dan belum secara khusus mengevaluasi kemampuan model dalam menangani *Semantic Shift* pada teks media sosial berbahasa Indonesia (Bellar et al., 2024). Selain itu, penelitian sebelumnya belum banyak yang membahas kombinasi dan perbandingan *Word2Vec* dan *FastText* pada arsitektur LSTM secara spesifik pada konteks bahasa Indonesia yang informal dan dinamis (Ladayya et al., 2025).

Berdasarkan permasalahan tersebut, penelitian ini menggunakan dataset hasil integrasi beberapa dataset publik berbahasa yang diperkaya dengan variasi *Semantic Shift* pada data *training*. Penelitian ini bertujuan untuk membandingkan performa model LSTM dengan *Word2Vec* dan LSTM dengan *FastText* dalam klasifikasi sentimen teks media sosial berbahasa Indonesia, serta mengevaluasi kemampuan kedua metode *embedding* tersebut dalam menangani *Semantic Shift* melalui evaluasi berbasis subset *Semantic Shift* dan *Shift Robustness Score*. Model terbaik yang diperoleh kemudian diterapkan untuk menganalisis opini masyarakat terhadap isu 17+8 Tuntutan Rakyat pada data komentar TikTok dan diimplementasikan ke dalam Aplikasi *Mobile*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah bagaimana perbandingan performa *Word2Vec* dan *FastText*

pada model LSTM dalam menangani *Semantic Shift* pada teks media sosial berbahasa Indonesia?

1.3 Batasan Masalah

Agar penelitian lebih terarah dan fokus pada tujuan yang ingin dicapai, maka ditetapkan beberapa batasan masalah sebagai berikut:

1. Dataset yang digunakan merupakan dataset teks berbahasa Indonesia hasil integrasi *INA Tweets PPKM Dataset*, *Indonesian Hate Speech Dataset*, dan SMSA yang diperkaya dengan variasi *Semantic Shift* pada data *training*.
2. Penelitian ini difokuskan pada *Semantic Shift* yang dibatasi pada perubahan polaritas sentimen berdasarkan konteks penggunaan kata (*contextual polarity shift*).
3. Model klasifikasi sentimen yang digunakan dalam penelitian ini adalah LSTM.
4. Metode *embedding* yang dibandingkan dalam penelitian ini terbatas pada *Word2Vec* dan *FastText* .
5. Klasifikasi sentimen dibatasi pada tiga kategori, yaitu negatif, netral, dan positif.
6. Evaluasi performa model dilakukan menggunakan *accuracy*, *precision*, *recall*, *Macro F1-score*, *confusion matrix*, serta *Shift Robustness Score* untuk mengukur kemampuan model dalam menangani *Semantic Shift*.
7. Implementasi model dilakukan pada data komentar TikTok terkait isu 17+8 Tuntutan Rakyat sebagai data uji nyata.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dijelaskan sebelumnya, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Menganalisis kemampuan model LSTM dalam memahami fenomena *Semantic Shift* pada teks media sosial berbahasa Indonesia.
2. Membandingkan performa *Word2Vec* dan *FastText* sebagai metode *embedding* pada model LSTM.
3. Mengevaluasi kemampuan model dalam menangani *Semantic Shift* menggunakan evaluasi berbasis subset *Semantic Shift* dan *Shift Robustness Score*.
4. Menerapkan model terbaik untuk menganalisis opini masyarakat terhadap isu 17+8 Tuntutan Rakyat pada *platform* TikTok.

1.5 Manfaat Penelitian

Penelitian ini diharapkan memberikan kontribusi baik secara teoretis maupun praktis sebagai berikut:

1. Manfaat Teoretis

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan kajian NLP, khususnya pada bidang analisis sentimen dan representasi kata berbasis *deep learning*. Selain itu, penelitian ini juga diharapkan dapat memperkaya kajian mengenai perbandingan metode *embedding* pada arsitektur LSTM dalam menangani *Semantic Shift* pada teks media sosial berbahasa Indonesia.

2. Manfaat Praktis

Secara praktis, hasil penelitian ini diharapkan dapat menjadi referensi dalam pemilihan metode *embedding* yang lebih efektif untuk analisis sentimen pada teks media sosial berbahasa Indonesia. Selain itu, penelitian ini juga dapat menjadi acuan bagi pengembang sistem NLP dalam merancang model analisis sentimen yang lebih adaptif terhadap variasi konteks sentimen pada komunikasi *digital*.

BAB II

TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen merupakan pendekatan komputasional dalam NLP yang digunakan untuk mengidentifikasi dan mengelompokkan opini publik berdasarkan ekspresi emosi dalam teks. Metode ini memungkinkan pemetaan pandangan masyarakat secara otomatis dan lebih cepat dibandingkan metode survei tradisional (Nip & Berthelier, 2024). Perkembangan analisis sentimen semakin pesat seiring meningkatnya volume data teks yang dihasilkan pengguna internet, khususnya pada media sosial, sehingga metode ini banyak digunakan untuk memahami persepsi publik terhadap berbagai isu melalui pemrosesan data dalam jumlah besar (Mao et al., 2024).

Media sosial menjadi sumber data utama dalam analisis sentimen karena mampu merepresentasikan opini publik *real-time* (Lai et al., 2023). Secara umum, analisis sentimen digunakan untuk mengklasifikasikan teks ke dalam beberapa kategori polaritas, seperti positif, negatif, dan netral. Pendekatan ini telah banyak diterapkan dalam berbagai bidang, termasuk analisis pasar, evaluasi layanan, serta pemantauan opini publik (Nip & Berthelier, 2024).

Namun, karakteristik bahasa pada media sosial yang cenderung informal menjadi tantangan dalam proses analisis teks. Bahasa *digital* sering mengandung slang, singkatan, emoji, *hashtag*, serta variasi penulisan yang dapat memengaruhi pemahaman makna pada teks (Rodríguez-Ibánez dkk.,2023). Selain itu, data media

sosial bersifat dinamis dan kontekstual sehingga makna kata dapat berubah seiring perkembangan sosial dan waktu (Baes et al., 2024) . Oleh karena itu, model analisis sentimen perlu mampu memahami konteks penggunaan kata serta variasi makna yang muncul dalam komunikasi *digital* (Darmansyah et al., 2024).

2.2 *Semantic Shift* dalam konteks analisis sentimen

Pemilihan kosakata dapat merepresentasikan latar belakang sosial, usia, maupun komunitas pengguna (Evizariza et al., 2025). Karakteristik bahasa *digital* yang spontan dan ekspresif sering melibatkan penggunaan singkatan, variasi ejaan, serta campur kode sehingga meningkatkan kompleksitas dalam pengolahan teks otomatis (Rodríguez-Ibánez et al., 2023). Dalam komunikasi *digital,* penggunaan slang, emoji, dan berbagai bentuk ekspresi kreatif lainnya semakin memperkaya sekaligus memperumit interpretasi makna dalam teks *digital* (Hilberts et al., 2025).

Selain menyampaikan makna asli, kata juga dapat mengandung emosi, humor, maupun ironi yang menyebabkan polaritas sentimennya berubah sesuai dengan konteks penggunaannya (Birjali et al., 2021). Fenomena ketika suatu kata memiliki polaritas sentimen yang berbeda tergantung pada konteks kalimat disebut *Semantic Shift* dalam konteks analisis sentimen, yang secara spesifik mengarah pada perubahan polaritas berbasis konteks (*contextual polarity shift*) (Periti et al., 2025).

Untuk menangani fenomena tersebut perlu menggunakan pendekatan model sekuensial seperti *Long Short-Term Memory* (LSTM) yang mampu menangkap hubungan konteks antar kata dalam kalimat sehingga dapat membantu mengenali perubahan polaritas sentimen berdasarkan konteks penggunaannya (Saputra et al., 2023). Secara linguistik, *Semantic Shift* dapat berupa perluasan, penyempitan,

penguatan, pelemahan, maupun penciptaan makna baru yang berkembang dengan cepat dalam komunikasi *digital* (Savitri & Dewi, 2023). Tabel 2.1 dibawah ini menunjukkan beberapa contoh kata slang yang mengalami perubahan makna dalam komunikasi media sosial.

Tabel 2. 1 contoh *Semantic Shift*

No	Kata	Makna Asli	Makna Bergeser	Contoh dalam Dataset	Label
1	Receh	Uang logam kecil / tidak penting	Hal sederhana yang lucu atau menghibur	receh tapi bikin ketawa	Positif
			Hal yang tidak menarik / tidak bermakna	receh banget ga jelas	Negatif
2	Savage	Liar / brutal	Keren, berani, tegas	komentarnya savage tapi keren	Positif
			Kasar / menyakitkan	savage tapi nyakitin	Negatif
3	Baper	Terbawa perasaan	Tersentuh secara emosional	baper karena tersentuh	Positif
			Terlalu sensitif / berlebihan	baper berlebihan	Negatif
4	Gaspol	Gas kendaraan	Semangat penuh / totalitas	gaspol demi perubahan	Positif
			Bertindak tanpa arah / berlebihan	gaspol tapi kacau	Negatif
5	Relate	Berhubungan / terhubung	Merasa sama dengan pengalaman	relate banget sama aku	Positif
			Tidak sesuai / tidak relevan	tidak relate dengan kenyataan	Negatif

Sumber: Hasil konstruksi dataset penelitian berbasis pendekatan *Semantic Shift*.

Dalam dataset penelitian, setiap kata muncul dalam berbagai konteks dengan label sentimen yang berbeda, sehingga mencerminkan fenomena *Semantic Shift* berbasis konteks. Hal ini sejalan dengan teori yang menyatakan bahwa makna kata dalam komunikasi *digital* bersifat dinamis dan dipengaruhi oleh konteks sosial serta ekspresi pengguna (Periti et al., 2025).

2.3 Metode *Embedding*

Embedding merupakan teknik representasi kata dalam bentuk vektor berdimensi rendah yang mampu menangkap hubungan semantik antar kata. Representasi tersebut memungkinkan berbagai tugas dalam NLP, seperti klasifikasi sentimen, pengukuran kesamaan teks, dan pencarian informasi, diproses oleh sistem NLP secara lebih akurat. Model *Embedding* seperti *Word2Vec*, GloVe, dan *FastText* memetakan kata dengan makna serupa ke dalam ruang vektor yang berdekatan sehingga lebih efektif dibandingkan pendekatan berbasis frekuensi semata (C. Zhang et al., 2025).

Namun, performa metode *embedding* dapat berbeda tergantung karakteristik dataset, bahasa yang digunakan, serta variasi konteks dalam data penelitian (Alkaabi et al., 2025). Penggunaan *Word2Vec* dan *FastText* dalam berbagai penelitian terbukti mampu meningkatkan *accuracy* model dibandingkan pendekatan tanpa *Embedding* (Siti Khomsah et al., 2022). Selain itu, *FastText* menunjukkan performa yang lebih baik dalam menangani kata langka dan variasi morfologis, sedangkan *Word2Vec* lebih efektif dalam menangkap pola kemunculan kata berdasarkan hubungan konteks (Adam Rachman et al., 2025).

2.4 *Word2Vec*

Word2Vec merupakan metode *embedding* berbasis prinsip distribusional yang merepresentasikan kata ke dalam bentuk vektor sehingga kata dengan konteks serupa memiliki kedekatan semantik dalam ruang vektor. Model ini mempelajari hubungan antar kata melalui pelatihan jaringan saraf sederhana untuk menghasilkan representasi kata berdimensi rendah. *Word2Vec* memiliki dua arsitektur utama,

yaitu *Continuous Bag-of-Words (CBOW)* yang memprediksi kata target berdasarkan kata konteks di sekitarnya dan *Skip-gram* yang memprediksi kata konteks berdasarkan kata target. Melalui pendekatan tersebut, Word2Vec mampu menghasilkan representasi vektor statis yang mencerminkan keterkaitan makna antar kata berdasarkan kemunculan konteks yang serupa (Alkaabi et al., 2025).

Melalui pendekatan ini, model dapat mempelajari hubungan semantik antar kata berdasarkan pola kemunculan dalam korpus teks (Cebeci et al., 2025). Pada penelitian sebelumnya analisis sentimen berbasis LSTM dengan *Word2Vec* menghasilkan *acurracy* dan *F1-score* yang lebih tinggi dibandingkan beberapa metode *Embedding* statis lainnya (Ladayya et al., 2025). Selain itu, *Word2Vec* juga digunakan pada sistem *content-based recommendation* dan *text retrieval* karena kemampuannya dalam menangkap hubungan semantik antar kata (Raja Azian et al., 2025).

2.5 *FastText*

FastText merupakan metode *embedding* yang membangun representasi kata menggunakan *subword* berbasis *character n-gram*. Berbeda dengan *Word2Vec* yang merepresentasikan kata sebagai satu unit utuh, *FastText* membentuk representasi kata dari bagian-bagian karakter sehingga lebih mampu menangani variasi morfologis, perubahan bentuk kata, serta variasi ejaan pada teks (Dirfas & Nastiti, 2024). Pendekatan ini memungkinkan *FastText* mengenali hubungan antar kata meskipun terdapat penggunaan slang, singkatan, maupun kata yang jarang muncul dalam data pelatihan. Selain itu, pemanfaatan struktur karakter membuat

FastText tetap mampu membentuk representasi vektor untuk kata yang tidak terdapat pada korpus pelatihan (Poetra et al., 2022).

Karakteristik tersebut menjadikan *FastText* lebih adaptif dalam memproses bahasa informal pada media sosial yang memiliki variasi penulisan dan konteks yang dinamis (Evizariza et al., 2025). Oleh karena itu, *FastText* banyak diterapkan pada berbagai tugas NLP, termasuk analisis sentimen pada teks media sosial (Memiş et al., 2024). Selain itu *FastText* juga membantu meningkatkan representasi kata yang tidak ditemukan dalam data pelatihan (Mars, 2022).

2.6 LSTM

LSTM merupakan pengembangan dari RNN yang dirancang untuk mengatasi permasalahan *vanishing gradient* melalui mekanisme *gating*, sehingga model mampu mempertahankan informasi penting dalam urutan teks yang panjang. Dibandingkan RNN konvensional, LSTM memiliki stabilitas pelatihan yang lebih baik serta mampu memodelkan hubungan temporal yang kompleks pada data sekuensial (Qixuan, 2024). Kemampuan tersebut menjadikan LSTM banyak digunakan pada berbagai tugas NLP, termasuk analisis sentimen, karena mampu memahami hubungan kontekstual antar kata dengan lebih baik, terutama pada bahasa informal media sosial yang memiliki variasi makna dan penggunaan kata yang dinamis.

Dalam analisis sentimen pada media sosial, integrasi teknik *embedding* dengan model sekuensial seperti LSTM dapat meningkatkan performa klasifikasi. (Ladayya et al., 2025). Penggunaan *embedding* seperti *Word2Vec* dan *FastText* juga dapat memperkuat kemampuan model dalam menangkap hubungan semantik antar

kata sehingga konteks kalimat dapat dipahami secara lebih akurat (Poetra et al., 2022). Mekanisme utama LSTM terdiri dari tiga gerbang: *forget gate*, *input gate*, dan *output gate*, yang berfungsi mengatur aliran informasi dalam *cell state* (el haddaoui dkk.,2022)

a. Rumus *forget gate*

$$f_t = \delta\left(W_f x_t + U_f h_t + b_f\right) \quad 2.1$$

Forget gate berfungsi untuk menentukan seberapa besar informasi dari cell state sebelumnya (c_{t-1}) yang dipertahankan atau dihapus. Nilai sigmoid dengan rentang 0 hingga 1 digunakan untuk mengontrol tingkat retensi memori.

b. Rumus *Input gate*

$$i_t = \delta(W_i x_i + U_i h_t + b_i) \quad 2.2$$

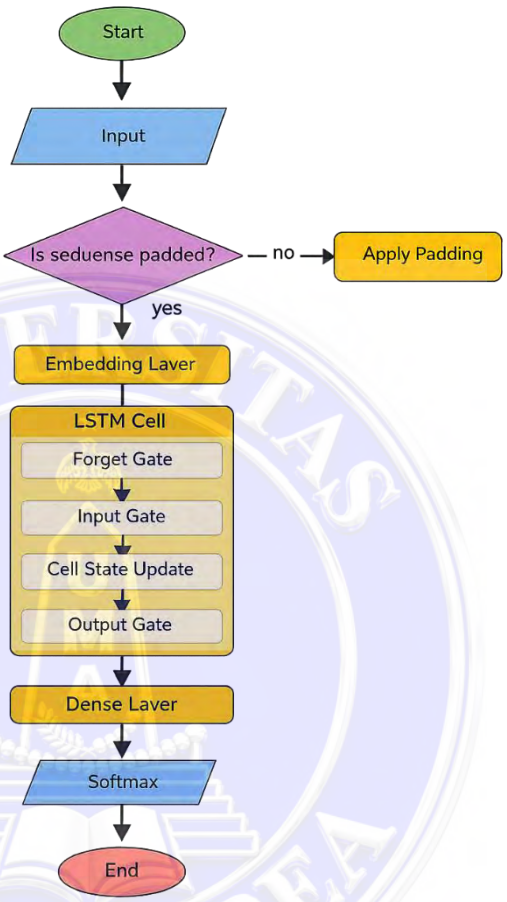
Input gate berfungsi untuk mengatur jumlah informasi baru yang akan ditambahkan ke dalam cell state. Jika nilai mendekati 1, maka informasi baru akan ditambahkan secara signifikan, sedangkan jika mendekati 0 maka informasi baru cenderung diabaikan.

c. Rumus *Output gate*

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_0) \quad 2.3$$

Output gate berfungsi untuk menentukan bagian dari cell state yang akan dikeluarkan sebagai hidden state pada setiap langkah waktu. Integrasi ketiga gerbang tersebut memungkinkan LSTM mempertahankan dependensi jangka panjang serta mengurangi permasalahan vanishing gradient pada data teks berurutan. Oleh karena itu, LSTM dinilai efektif dalam memahami konteks kalimat pada analisis sentimen, termasuk pada teks media sosial yang mengandung

Semantic Shift. Selain itu, efisiensi komputasi yang dimiliki LSTM membuat model ini masih relevan digunakan pada berbagai penelitian NLP berbasis klasifikasi teks (Krichen & Mihoub, 2025).



Gambar 2. 1 *flowchart LSTM*

2.7 Pra-Pemrosesan Teks

Pra-pemrosesan teks merupakan tahap awal yang penting dalam proses NLP yang bertujuan mengubah data teks mentah menjadi format yang lebih terstruktur sehingga dapat diproses oleh model *machine learning* (Siino et al., 2024). Pada analisis sentimen, data yang berasal dari media sosial umumnya mengandung berbagai bentuk noise seperti URL, *mention*, simbol, tanda baca berlebihan, emoji,

singkatan, serta variasi penulisan kata yang tidak baku. Oleh karena itu, pra-pemrosesan diperlukan untuk membersihkan dan menormalisasi teks agar dapat diproses secara lebih efektif oleh algoritma NLP (Abiola et al., 2025).

Secara umum, tahapan pra-pemrosesan teks meliputi *case folding*, *filtering*, atau penghapusan karakter yang tidak diperlukan, normalisasi teks, tokenisasi, serta pembersihan data duplikat (Penggalih et al., 2025). Proses tersebut membantu mengurangi kompleksitas data serta meningkatkan kualitas representasi teks yang digunakan dalam pelatihan model. Dengan demikian, tahap pra-pemrosesan menjadi salah satu komponen penting dalam proses NLP karena berpengaruh terhadap performa dan *accuracy* model analisis sentimen (Sim et al., 2023).

2.8 Tokenisasi dan *Padding* Teks

Tokenisasi merupakan proses memecah teks menjadi unit-unit kecil yang disebut token, yang dapat berupa kata, subkata, atau karakter. Proses ini penting karena model komputasional tidak dapat memproses teks secara langsung tanpa terlebih dahulu mengubahnya menjadi representasi yang dapat dianalisis secara numerik (Wangchuk et al., 2025). Token yang dihasilkan kemudian digunakan sebagai dasar dalam pembentukan fitur maupun representasi vektor kata dalam model NLP (Y. Zhang et al., 2025).

Dalam implementasi model *deep learning* seperti LSTM, setiap urutan token harus memiliki panjang yang sama agar dapat diproses secara *batch* oleh jaringan saraf. Oleh karena itu digunakan teknik *padding*, yaitu penambahan token kosong pada urutan teks yang lebih pendek sehingga semua *input* memiliki panjang yang seragam (Beneš, 2021). Kombinasi tokenisasi dan *padding* memungkinkan teks

mentah diubah menjadi representasi numerik yang konsisten sehingga dapat diproses secara efektif oleh model sekuensial seperti LSTM. Proses ini juga membantu mempertahankan urutan kata dalam kalimat sehingga informasi konteks tetap dapat dipahami oleh model selama proses pelatihan (Fernando et al., 2025).

2.9 Evaluasi Klasifikasi

Evaluasi klasifikasi digunakan untuk mengukur kemampuan model dalam mengelompokkan teks ke dalam tiga kelas sentimen yaitu negatif, netral, dan positif. Metrik evaluasi yang digunakan dalam penelitian ini meliputi *accuracy*, *precision*, *recall*, dan *F1-score (macro average).*

1. *Accuracy* mengukur seberapa banyak prediksi yang dilakukan model sesuai dengan label sebenarnya dibandingkan dengan seluruh data yang diuji.

$$Accuracy = \frac{\text{Jumlah Prediksi Benar}}{\text{Total Prediksi}} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.4)$$

Jika terdapat 100 kalimat dan model berhasil memprediksi 87 kalimat dengan benar, maka nilai *accuracy* model adalah 0.87. Nilai accuracy berada pada rentang 0 hingga 1, dengan nilai yang lebih tinggi menunjukkan performa klasifikasi model yang lebih baik.

2. *Precision* digunakan untuk menilai ketepatan model dalam memprediksi suatu kelas tertentu.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2.5)$$

Jika model memprediksi 100 data sebagai sentimen positif, tetapi hanya 85 di antaranya yang benar, maka nilai *precision* adalah 0.85.

3. *Recall* digunakan untuk mengukur kemampuan model dalam mendeteksi seluruh data yang benar-benar termasuk dalam suatu kelas.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (2.6)$$

Misalnya terdapat 100 kalimat positif, namun model hanya berhasil mendeteksi 80 kalimat, maka nilai *recall* adalah 0.80. Nilai *recall* yang tinggi menunjukkan bahwa model mampu menemukan sebagian besar data aktual pada kelas tersebut.

4. *F1-score* merupakan rata-rata harmonik antara *precision* dan *recall*, yang digunakan untuk menyeimbangkan kedua metrik tersebut.

$$\mathrm{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall} \quad (2.7)$$

Dalam penelitian ini digunakan *Macro F1-score,* yaitu rata-rata nilai *F1-score* dari seluruh kelas tanpa mempertimbangkan proporsi jumlah data pada masing-masing kelas. Penggunaan *Macro F1-score* bertujuan untuk memastikan bahwa performa model seimbang pada seluruh kelas sentimen. Nilai Macro F1-score berada pada rentang 0 hingga 1, dengan nilai yang lebih tinggi menunjukkan performa klasifikasi yang semakin baik pada seluruh kelas sentimen.

2.10 *Confusion Matrix*

Untuk menganalisis kesalahan klasifikasi yang dihasilkan oleh model, digunakan *confusion matrix*, yaitu matriks yang membandingkan label prediksi dengan label aktual pada data uji. *Confusion matrix* memberikan informasi mengenai distribusi kesalahan klasifikasi pada setiap kelas sentimen. Komponen utama dalam *confusion matrix* meliputi:

a. *True Positive* (TP): model berhasil memprediksi data ke kelas yang benar.

b. *True Negative* (TN): model berhasil mengidentifikasi data yang bukan termasuk ke dalam suatu kelas tertentu.

c. *False Positive* (FP): model salah memprediksi suatu kelas, tetapi label aktualnya berbeda.

d. *False Negative* (FN): model gagal mendeteksi kelas yang sebenarnya.

Melalui *confusion matrix* dapat diketahui pola kesalahan klasifikasi serta kemampuan model dalam membedakan setiap kelas sentimen.

2.11 Evaluasi *Semantic Shift*

Evaluasi dilakukan untuk mengukur kemampuan model dalam menangani *Semantic Shift* dengan membandingkan performa pada subset data yang mengandung kata yang berpotensi mengalami perubahan makna dan keseluruhan data uji. Perbandingan dilakukan menggunakan Macro F1-score serta *Shift Robustness Score* untuk melihat kestabilan performa model pada variasi konteks penggunaan kata. *Shift Robustness Score* dihitung sebagai perbandingan antara Macro F1-score pada subset Semantic Shift dan Macro F1-score keseluruhan data uji.

$$ShiftRobustness = \frac{F1_{subset\ shift}}{F1_{overall}} \quad (2.8)$$

Nilai *Shift Robustness Score* mendekati 1 menunjukkan bahwa performa model tetap stabil pada data yang mengandung *Semantic Shift*. Nilai di bawah 1 menunjukkan adanya penurunan performa model pada *subset Semantic Shift*, sedangkan nilai di atas 1 menunjukkan bahwa performa model pada subset tersebut

lebih baik dibandingkan performa pada keseluruhan data uji. Interpretasi nilai *Shift Robustness Score* ditunjukkan pada tabel berikut.

Tabel 2. 2 Interpretasi nilai *Shift Robustness Score*

Nilai	Arti
> 1	Performa meningkat
= 1	Stabil
< 1	Performa turun

2.12 Penelitian Terdahulu

Beberapa penelitian terdahulu telah memanfaatkan arsitektur *deep learning* dan metode *embedding* untuk memahami konteks serta hubungan semantik antar kata pada analisis sentimen teks media sosial.

Tabel 2. 3 Penelitian Terdahulu

No	Peneliti	Arsitektur	*Embedding*	Fokus Penelitian	Hasil
1	Memiş et al. (2024)	BiLSTM	*FastText*	Analisis sentimen media sosial berbasis bahasa informal	Accuracy 0.91 dan F1-score 0.90
2	Ladayya et al. (2025)	LSTM	*Word2Vec*	Perbandingan *embedding* pada analisis sentimen Twitter berbahasa Indonesia	Accuracy 0.89 dan Macro F1-score 0.88
3	Dirfas & Nastiti (2024)	Deep Learning	*FastText*	Analisis teks media sosial menggunakan *FastText* berbasis subword	Accuracy 0.87
4	Alasmari et al. (2024)	LSTM	Word *Embedding*	Analisis sentimen pada teks dengan konteks emosional kompleks	Accuracy 0.93
5	Evizariza et al. (2025)	NLP Model	*FastText*	Pemrosesan bahasa informal dan slang media sosial	F1-score 0.89

Berdasarkan penelitian terdahulu, penggunaan arsitektur LSTM dan metode *embedding* terbukti mampu meningkatkan performa analisis sentimen pada teks media sosial. *FastText* menunjukkan keunggulan dalam menangani variasi bahasa informal dan bentuk kata yang beragam, sedangkan *Word2Vec* efektif dalam

mempelajari hubungan semantik berdasarkan konteks kemunculan kata. Namun, penelitian yang secara khusus membandingkan *Word2Vec* dan *FastText* pada model LSTM untuk menangani *Semantic Shift* pada teks media sosial berbahasa Indonesia masih terbatas. Oleh karena itu, penelitian ini dilakukan untuk membandingkan kedua metode *embedding* tersebut dalam analisis sentimen berbasis konteks kalimat.

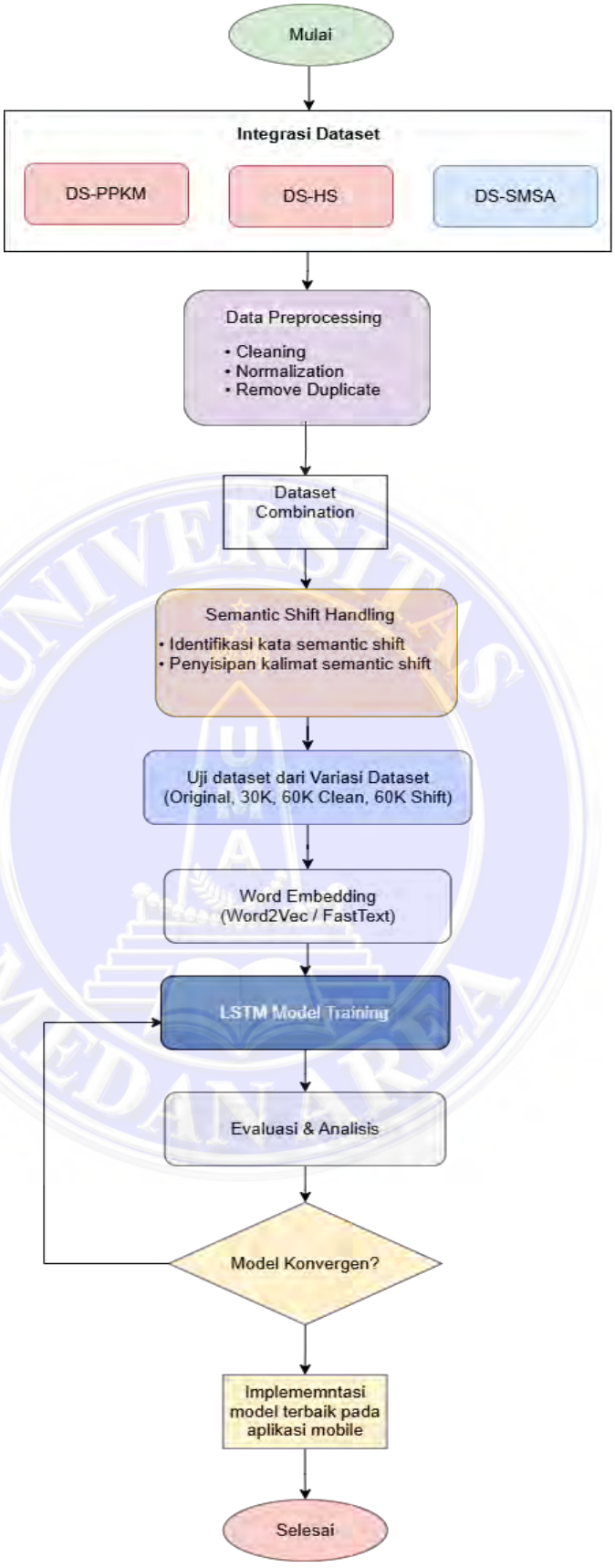
BAB III

METODOLOGI PENELITIAN

3.1 Rancangan Penelitian

Penelitian ini bertujuan untuk menganalisis kemampuan model LSTM dalam memahami fenomena *Semantic Shift* pada teks media sosial berbahasa Indonesia. Dua metode *embedding*, yaitu *Word2Vec* dan *FastText* akan dibandingkan dalam arsitektur model yang sama serta dikombinasikan dengan fitur berbasis leksikon untuk mendeteksi indikasi *Semantic Shift*. Eksperimen dilakukan dengan konfigurasi yang sama pada kedua metode *embedding*, masing-masing diuji sebanyak lima kali untuk mengukur konsistensi performa.

Evaluasi menggunakan metrik *accuracy* dan *Macro F1-score*, serta analisis tambahan melalui perbandingan performa pada data yang mengandung *Semantic Shift*. Hasil terbaik dari masing-masing metode kemudian dibandingkan untuk menentukan metode yang paling efektif dalam menangani *Semantic Shift* pada teks media sosial. Alur keseluruhan proses penelitian ditunjukkan pada Gambar 3.1.



Gambar 3. 1 *flowchart* penelitian

3.2 Dataset Penelitian

Dataset yang digunakan dalam penelitian ini merupakan hasil integrasi beberapa dataset dari *platform* Kaggle. Penggabungan dataset dilakukan untuk meningkatkan jumlah data pelatihan serta memperluas variasi konteks bahasa yang dipelajari oleh model. Dataset berasal dari tiga sumber utama, yaitu:

1. *INA Tweets PPKM Dataset*
2. *Indonesian Hate Speech Dataset*
3. SMSA

Pada dataset *Indonesian Hate Speech*, hanya data yang memiliki label *HS_Strong* atau *Abusive* yang digunakan dalam penelitian ini dan seluruhnya dikategorikan sebagai sentimen negatif.

Tabel 3. 1 Sumber Dataset Penelitian

Kode Dataset	**Sumber**	**Fitur Data**	**Kontribusi Label**
DS-PPKM	Kaggle	Tweet dan Sentiment	Positif, Netral, Negatif
DS-HS	Kaggle	Tweet dan Hate Speech	Negatif
DS-SMSA	Kaggle	Text dan Label	Positif, Netral, Negatif

Ketiga dataset tersebut kemudian digabungkan untuk membentuk dataset yang lebih besar dan lebih beragam secara linguistik. Proses penggabungan juga diikuti dengan penghapusan data duplikat serta penyaringan teks yang terlalu pendek agar setiap data memiliki konteks yang memadai. Selain dataset pelatihan yang diperoleh dari Kaggle, penelitian ini juga menggunakan data kasus berupa komentar media sosial yang diambil melalui proses *scraping* menggunakan *platform* Apify.

Data tersebut berasal dari komentar TikTok yang berkaitan dengan isu 17+8 tuntutan rakyat demo 2025. Dataset ini tidak digunakan dalam proses pelatihan model, melainkan digunakan sebagai data uji nyata untuk mengukur kemampuan model dalam melakukan klasifikasi sentimen pada data dunia nyata.

3.2.1 Pembentukan Dataset

Dataset dirancang untuk merepresentasikan fenomena *Semantic Shift* pada bahasa media sosial. Tahap awal dilakukan dengan mengidentifikasi kata yang berpotensi memiliki ambiguitas makna, seperti receh, *savage*, baper, gaspol, dan *relate*. Tahapan pembentukan dataset dimulai dengan membagi dataset menjadi data *training*, *validation*, dan *test* menggunakan stratifikasi label. Selanjutnya dilakukan *balancing* pada data *training* untuk menjaga distribusi kelas tetap seimbang.

Pada tahap berikutnya, dilakukan penambahan data *Semantic Shift* pada data *training*. Variasi konteks dibangun secara terkontrol dengan memanfaatkan kombinasi kata dalam kalimat tanpa menggunakan *template* tetap. Data hasil konstruksi ini kemudian digabungkan ke dalam data *training*. Sebagai tahap akhir, dilakukan pengecekan *overlap* antar *subset* untuk mencegah data *leakage.* Pendekatan ini menghasilkan dataset dengan variasi konteks yang merepresentasikan *Semantic Shift*, sehingga dapat digunakan untuk melatih model dalam mempelajari perbedaan konteks penggunaan kata.

3.2.2 Pembagian Dataset

Pada tahap pembuatan dataset penelitian, data yang telah dikumpulkan dan diproses langsung dibagi menjadi tiga bagian, yaitu data *training*, *validation*, dan *test* menggunakan teknik *stratified split*. Dapat dilihat pada tabel di bawah ini.

Tabel 3. 2 Pembagian Dataset

Subset	**Presentasi**
Training	80%
Validation	10%
Test	10%

Pembagian ini bertujuan untuk menjaga distribusi label tetap seimbang pada setiap *subset* serta menghindari bias selama proses pelatihan dan evaluasi model. Setelah pembagian dataset, dilakukan *balancing* pada data *training* untuk menjaga distribusi sentimen tetap seimbang.

3.3 Pra-Pemrosesan Data

Pra-pemrosesan data merupakan tahap awal sebelum proses pelatihan model. Tahap ini bertujuan membersihkan teks serta menyeragamkan format data agar lebih mudah diproses oleh model *machine learning*. Data teks dari media sosial umumnya mengandung berbagai bentuk *noise*. Oleh karena itu dilakukan proses pembersihan menggunakan fungsi *clean_text* dengan bantuan *regular expression*. Tahapan pra-pemrosesan meliputi:

1. *Case folding*, mengubah seluruh huruf menjadi huruf kecil
2. *Filtering,* menghapus *noise* seperti URL, *mention*, dan karakter non-alfanumerik

3. Normalisasi spasi, menghapus spasi berlebih
4. Setelah proses pembersihan, dilakukan penghapusan data duplikat
5. Penyaringan kalimat dengan panjang yang terlalu pendek

3.4 Representasi Data

Dataset penelitian terdiri dari dua atribut utama, yaitu teks dan label sentimen. Kolom *clean_text* berisi teks yang telah melalui proses pembersihan, sedangkan kolom label berisi kategori sentimen dalam bentuk numerik. Dalam penelitian ini digunakan tiga kategori sentimen.

Tabel 3. 3 Struktur Dataset dan Representasi Label

Komponen	Keterangan
*clean*_text	Teks yang telah melalui proses pembersihan (*cleaning*) dan normalisasi
Label	Kategori sentimen dalam bentuk numerik
0	Sentimen Negatif
1	Sentimen Netral
2	Sentimen Positif

Setiap baris dataset merepresentasikan satu teks beserta label sentimen yang digunakan sebagai data pelatihan dan evaluasi model klasifikasi.

3.5 *Embedding*

Pada penelitian ini digunakan dua metode *embedding*, yaitu *Word2Vec* dan *FastText* . *Embedding* digunakan untuk merepresentasikan teks dalam bentuk vektor numerik yang dapat diproses oleh model. Proses *Embedding* dilakukan

setelah tahap tokenisasi dan *padding,* sehingga setiap kata yang telah diubah menjadi indeks token dapat dipetakan ke dalam vektor numerik. *Embedding* dilatih menggunakan dataset pelatihan agar mampu merepresentasikan hubungan semantik serta variasi konteks penggunaan kata, dalam dataset, termasuk kata yang merepresentasikan *Semantic Shift*.

Representasi kata yang dihasilkan menyesuaikan karakteristik bahasa media sosial dan digunakan untuk membentuk *embedding matrix* yang berisi vektor setiap kata dalam kosakata tokenizer. Matriks tersebut kemudian digunakan sebagai bobot awal pada *Embedding Layer* dalam arsitektur model LSTM sehingga model dapat memanfaatkan informasi semantik yang telah dipelajari selama proses pelatihan.

3.5.1 *FastText*

FastText dilatih menggunakan dataset *training* penelitian dengan memanfaatkan informasi *subword* berupa potongan karakter *n-gram*, sehingga mampu mengenali hubungan antar kata yang memiliki bentuk morfologis serupa. Model *FastText* dilatih menggunakan pustaka Gensim dengan dimensi vektor 300, ukuran jendela konteks 7, serta nilai min_count sebesar 2. Model ini menggunakan arsitektur Skip-gram dengan parameter min_n sebesar 3 dan max_n sebesar 6 untuk membentuk representasi karakter n-gram pada setiap kata.

Setelah proses pelatihan selesai, vektor kata dari model *FastText* digunakan untuk membangun *Embedding matrix* yang kemudian dimasukkan sebagai bobot awal pada *Embedding Layer* dalam model LSTM. Sama seperti pada *Word2Vec*, bobot *Embedding* bersifat *trainable* sehingga dapat diperbarui selama proses pelatihan model. Penggunaan *FastText* diharapkan mampu membantu model

memahami variasi bentuk kata dalam bahasa media sosial serta meningkatkan kemampuan model dalam menangani *Semantic Shift*.

3.5.2 *Word2Vec*

Word2Vec digunakan sebagai metode embedding pembanding dalam penelitian ini. Berbeda dengan *FastText* yang memanfaatkan informasi *subword*, *Word2Vec* merepresentasikan kata sebagai vektor berdasarkan konteks kemunculannya dalam kalimat. Vektor kata yang dihasilkan kemudian digunakan untuk membentuk *embedding matrix* sebagai bobot awal pada *Embedding Layer* model LSTM.

Jika suatu kata ditemukan dalam *Word2Vec,* maka vektor tersebut digunakan sebagai representasi kata. Sebaliknya, jika kata tidak ditemukan, maka vektor diinisialisasi secara acak menggunakan distribusi normal. Bobot pada *Embedding Layer* bersifat *trainable* sehingga masih dapat diperbarui selama proses pelatihan model.

3.6 Arsitektur Model LSTM

Model klasifikasi sentimen pada penelitian ini dibangun menggunakan arsitektur LSTM yang dirancang untuk memproses data teks berbentuk urutan kata. Model menerima dua jenis *input*, pertama berupa urutan token teks hasil tokenisasi dan *padding*, dan kedua fitur leksikon yang merepresentasikan merepresentasikan indikasi *Semantic Shift*. Urutan token teks terlebih dahulu diproses oleh *Embedding Layer* yang mengubah setiap kata menjadi representasi vektor berdimensi 300 menggunakan *Embedding matrix* yang telah dilatih dengan *Word2Vec* atau *FastText* .

Setelah proses *Embedding*, data dilewatkan ke *SpatialDropout1D Layer* untuk mengurangi risiko overfitting dengan menghilangkan sebagian fitur *Embedding* secara acak selama proses pelatihan.

Selanjutnya, data diproses oleh *LSTM Layer* dengan 128 unit untuk mempelajari hubungan sekuensial antar kata dalam kalimat. *Output* dari LSTM kemudian diproses menggunakan *GlobalMaxPooling1D Layer* untuk mengekstraksi fitur paling dominan dari urutan keluaran LSTM. Fitur tersebut kemudian digabungkan dengan fitur leksikon *Semantic Shift* menggunakan *Concatenate Layer*, sehingga model dapat memanfaatkan informasi tambahan terkait kata yang berpotensi mengalami pergeseran makna dalam konteks kalimat.

Hasil penggabungan fitur kemudian diproses oleh *Dense Layer* dengan fungsi aktivasi ReLU untuk mempelajari representasi fitur yang lebih kompleks. Selanjutnya diterapkan *Dropout Layer* untuk mengurangi kemungkinan *overfitting* selama pelatihan. Pada lapisan terakhir digunakan *Output Layer* dengan fungsi aktivasi *Softmax* yang menghasilkan probabilitas untuk tiga kelas sentimen, yaitu negatif, netral, dan positif. Model dilatih menggunakan fungsi *loss Sparse Categorical Crossentropy* dan *optimizer* Adam dengan *learning rate* sebesar 2e-4.

Tabel 3. 4 Konfigurasi Arsitektur Model

Komponen	**Konfigurasi**
Embedding Dimension	300
Max Sequence Length	40
Vocabulary Size	12000
LSTM Units	128
Dense Layer	64

Dropout Rate	0.4
Output Class	3
Loss Function	Sparse Categorical Crossentropy
Optimizer	Adam

3.7 Proses Pelatihan Model

Proses pelatihan model dilakukan setelah tahap representasi kata menggunakan metode *Word2Vec* dan *FastText* . Dataset hasil pembagian sebelumnya digunakan dalam proses pelatihan model. Data pelatihan digunakan untuk melatih model, data *valid*asi digunakan untuk memantau proses pelatihan serta mencegah *overfitting*, sedangkan data pengujian digunakan untuk mengevaluasi performa model secara keseluruhan.

Sebelum pelatihan, teks diubah menjadi urutan token menggunakan Tokenizer, di mana setiap kata dipetakan ke dalam indeks numerik berdasarkan kosakata yang terbentuk dari data pelatihan. Selanjutnya dilakukan *padding sequence* untuk menyeragamkan panjang setiap urutan teks dengan panjang maksimum 40 token. Setelah proses tokenisasi dan padding, setiap kata pada urutan token diubah menjadi vektor menggunakan *Embedding matrix* yang diperoleh dari model *Word2Vec* atau *FastText* , kemudian digunakan sebagai *input* pada model LSTM.

Model dilatih menggunakan *batch size* 32 dengan maksimum 50 *epoch.* Untuk meningkatkan stabilitas pelatihan, digunakan teknik *Reduce Learning Rate on Plateau* untuk menurunkan *learning rate* ketika performa model pada data *valid*asi tidak meningkat. Untuk memperoleh hasil yang lebih stabil, eksperimen pelatihan model dilakukan sebanyak 5 kali percobaan untuk masing-masing metode

Embedding. Setiap percobaan menghasilkan nilai evaluasi yang kemudian dibandingkan untuk menentukan model terbaik yang dipilih berdasarkan nilai *Macro F1-score* tertinggi pada data pengujian. Selanjutnya, hasil terbaik dari *Word2Vec* dan *FastText* dibandingkan untuk menentukan metode *Embedding* yang paling efektif dalam penelitian ini.

3.8 Evaluasi Model

Evaluasi model dilakukan menggunakan beberapa metrik klasifikasi yaitu *accuracy*, *precision*, *recall*, dan *Macro F1-score* untuk mengukur performa model dalam mengklasifikasikan sentimen teks. Selain itu, digunakan *confusion matrix* untuk menganalisis distribusi prediksi serta kesalahan klasifikasi antar kelas sentimen. Selain evaluasi klasifikasi umum, penelitian ini juga melakukan evaluasi tambahan terhadap kalimat yang mengandung kata yang berpotensi mengalami *Semantic Shift*.

Evaluasi ini dilakukan dengan menghitung *Shift Robustness Score*, yaitu perbandingan antara nilai Macro F1-score pada subset data yang mengandung *Semantic Shift* dengan Macro F1-score keseluruhan dataset. Nilai ini digunakan untuk mengukur kemampuan model dalam mempertahankan performa ketika dihadapkan pada variasi konteks makna dalam teks media sosial.

3.9 Penerapan Model Terbaik pada Analisis Isu 17+8

Setelah model terbaik diperoleh dari proses pelatihan dan evaluasi, model tersebut digunakan untuk menganalisis opini masyarakat terhadap isu 17+8 Tuntutan Rakyat pada *platform* TikTok. Data yang digunakan dalam analisis ini

diperoleh melalui proses *scraping* komentar atau unggahan yang berkaitan dengan isu tersebut. Data yang telah dikumpulkan kemudian diproses menggunakan tahapan pra-pemrosesan yang konsisten dengan prosedur pada data pelatihan, meliputi pembersihan teks, normalisasi, serta penghapusan karakter yang tidak relevan.

Selanjutnya, data yang telah diproses dimasukkan ke dalam model terbaik hasil evaluasi, yaitu model LSTM dengan representasi kata *FastText* , untuk menghasilkan prediksi sentimen pada setiap teks. Model kemudian mengklasifikasikan setiap teks ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif. Hasil prediksi tersebut selanjutnya dianalisis untuk mengetahui distribusi sentimen masyarakat terhadap isu 17+8 Tuntutan Rakyat. Distribusi sentimen ini memberikan gambaran mengenai kecenderungan opini publik yang muncul dalam diskusi media sosial terkait isu tersebut.

3.10 *Mockup* implementasi Aplikasi *Mobile*

Mockup ini menggambarkan rancangan antarmuka aplikasi *Mobile* berbasis Android dan iOS yang dirancang untuk melakukan analisis sentimen teks secara otomatis. Aplikasi ini berfungsi sebagai media implementasi model klasifikasi sentimen berbasis LSTM dengan metode *embedding Word2Vec* atau *FastText* dalam bentuk sistem yang dapat digunakan langsung oleh pengguna.

Gambar 3. 2 *Mockup* Halaman Utama

Mockup halaman utama menampilkan rancangan antarmuka yang menyediakan kolom *input* "Teks Ulasan" bagi pengguna untuk memasukkan kalimat yang akan dianalisis. Di bawah kolom *input* terdapat tombol "Analisis Sekarang" yang digunakan untuk menjalankan proses klasifikasi sentimen. Pada bagian atas antarmuka juga ditampilkan informasi bahwa sistem analisis didukung oleh teknologi LSTM dan *FastText* sebagai dasar model kecerdasan buatan yang digunakan.

Gambar 3. 3 Halaman statistik

Mockup halaman statistik menampilkan rancangan visualisasi distribusi sentimen dalam bentuk grafik batang. Fitur ini dirancang untuk menampilkan

perbandingan jumlah sentimen positif, netral, dan negatif dari hasil analisis teks. Pada tahap mockup awal, tampilan “Belum ada data” menunjukkan kondisi ketika belum terdapat hasil analisis yang tersimpan.

Gambar 3. 4 Halaman Riwayat

Mockup halaman riwayat menunjukkan rancangan halaman yang berfungsi untuk menyimpan daftar teks yang telah dianalisis sebelumnya. Halaman ini memungkinkan pengguna meninjau kembali hasil analisis yang pernah dilakukan. Pada tahap mockup awal ditampilkan keterangan “Belum ada riwayat” yang menandakan belum terdapat data analisis yang tersimpan.

Gambar 3. 5 Halaman Info

Mockup halaman info menampilkan rancangan halaman yang berisi informasi teknis mengenai aplikasi. Pada halaman ini ditampilkan keterangan mengenai teknologi yang digunakan, yaitu Flutter dan Dart pada sisi frontend serta *Python FastAPI* pada sisi *backend*. Selain itu juga ditampilkan informasi fitur aplikasi seperti confidence score serta proses analisis sentimen berbasis *Deep learning* menggunakan *TensorFlow*/Keras.

BAB V

KESIMPULAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa model LSTM mampu melakukan klasifikasi sentimen pada teks media sosial berbahasa Indonesia dengan baik menggunakan metode *embedding Word2Vec* dan *FastText* . Kedua metode *embedding* mampu membantu model memahami hubungan semantik antar kata pada dataset penelitian yang telah diperkaya dengan variasi *Semantic Shift*. Hasil pengujian menunjukkan bahwa model LSTM + *FastText* memperoleh performa terbaik dibandingkan model LSTM + Word2Vec pada seluruh metrik evaluasi utama.

FastText menghasilkan nilai *accuracy*, *Macro F1-score*, serta robustness yang lebih tinggi dibandingkan Word2Vec. Hasil tersebut menunjukkan bahwa *FastText* lebih efektif dalam membantu model memahami perubahan polaritas sentimen berdasarkan konteks kalimat. Keunggulan *FastText* dipengaruhi oleh penggunaan representasi subword berbasis character n-gram yang memungkinkan model mengenali variasi bentuk kata, slang, singkatan, serta kata yang jarang muncul pada media sosial.

Pendekatan tersebut membuat *FastText* lebih adaptif dalam menangani *Semantic Shift* dibandingkan Word2Vec yang merepresentasikan kata sebagai satu unit utuh. Evaluasi pada subset *Semantic Shift* menunjukkan bahwa kedua model tetap mampu mempertahankan performa ketika menghadapi variasi konteks makna. Namun, *FastText* menunjukkan kemampuan yang lebih baik dalam menjaga

kestabilan performa pada data yang mengandung perubahan polaritas sentimen berbasis konteks.

Model terbaik hasil penelitian kemudian diterapkan pada analisis sentimen isu 17+8 Tuntutan Rakyat menggunakan data komentar TikTok hasil scraping. Hasil analisis menunjukkan bahwa sebagian besar komentar memiliki kecenderungan sentimen netral. Hal tersebut menunjukkan bahwa model yang dibangun mampu diterapkan pada data dunia nyata untuk menganalisis opini masyarakat terhadap isu yang berkembang di media sosial.

Secara keseluruhan, penelitian ini menunjukkan bahwa metode *FastText* merupakan representasi kata yang paling efektif pada arsitektur LSTM dalam menangani *Semantic Shift* pada analisis sentimen teks media sosial berbahasa Indonesia.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat digunakan untuk pengembangan penelitian selanjutnya. Penelitian selanjutnya dapat memperluas jumlah dan variasi kata pada *Semantic Shift* lexicon agar model mampu mengenali lebih banyak perubahan polaritas sentimen yang muncul pada bahasa media sosial. Variasi konteks kalimat juga dapat diperbanyak sehingga model memperoleh representasi konteks yang lebih beragam selama proses pelatihan.

Selain itu, proses konstruksi data *Semantic Shift* dapat dikembangkan dengan menambahkan variasi pola kalimat yang lebih kompleks. Dengan tujuan agar model dapat mempelajari hubungan konteks secara lebih mendalam khususnya pada

penggunaan bahasa informal, slang, dan ekspresi digital yang terus berkembang di media sosial. Penelitian selanjutnya juga dapat menggunakan data dari lebih banyak topik atau isu media sosial sehingga kemampuan generalisasi model dalam memahami *Semantic Shift* dapat diuji pada konteks yang lebih luas.

Pada tahap evaluasi, analisis *Semantic Shift* dapat diperluas dengan menambahkan lebih banyak subset kata ambigu sehingga pengukuran robustness model terhadap perubahan konteks makna menjadi lebih detail. Selain itu, pengembangan aplikasi *Mobile* dapat dilanjutkan dengan menambahkan fitur visualisasi hasil analisis yang lebih interaktif agar hasil klasifikasi sentimen lebih mudah dipahami oleh pengguna.

DAFTAR PUSTAKA

Abiola, H. M., Iyanuoluwa, A., A., A. A., Gadafi, A. M., & Ishaq, A. (2025). Tiktok Through AI Eyes: A Deep Learning Approach to Sentiment Analysis. *Kwaghe International Journal of Engineering and Information Technology*, *2*(2), 57–77. https://doi.org/10.58578/kijeit.v2i2.5485

Adam Rachman, M., Agussalim, A., & Dyar Wahyuni, E. (2025). komparasi performa model klasifikasi emosi dengan *embedding* menggunakan algoritma svm dan random forest. *jati (Jurnal Mahasiswa Teknik Informatika)*, *9*(2), 2872–2878. https://doi.org/10.36040/jati.v9i2.13197

Alasmari, A., Farooqi, N., & Alotaibi, Y. (2024). Sentiment analysis of pilgrims using CNN-LSTM deep learning approach. *PeerJ Computer Science*, *10*, e2584. https://doi.org/10.7717/peerj-cs.2584

Alkaabi, H., Jasim, A. K., & Darroudi, A. (2025). From Static to Contextual: A Survey of *Embedding* Advances in NLP. *PERFECT: Journal of Smart Algorithms*, *2*(2), 57–66. https://doi.org/10.62671/perfect.v2i2.77

Al-Tarawneh, M. A. B., Al-irr, O., Al-Maaitah, K. S., Kanj, H., & Aly, W. H. F. (2024). Enhancing Fake News Detection with Word *Embedding*: A *Machine learning* and Deep Learning Approach. *Computers*, *13*(9), 239. https://doi.org/10.3390/computers13090239

Baes, N., Haslam, N., & Vylomova, E. (2024). A Multidimensional Framework for Evaluating Lexical Semantic Change with Sosial Science Applications. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1390–1415. https://doi.org/10.18653/v1/2024.acl-long.76

Bellar, O., Baina, A., & Ballafkih, M. (2024). Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers. *Mathematics*, *12*(15), 2403. https://doi.org/10.3390/math12152403

Beneš, A. (2021). *Counting extensions of imaginary quadratic fields*. http://arxiv.org/abs/2109.09848

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/J.KNOSYS.2021.107134

Cebeci, U., Simsir, U., & Dogan, O. (2025). Machine Selection for Inventory Tracking with a Continuous Intuitionistic Fuzzy Approach. *Applied Sciences*, *15*(1), 425. https://doi.org/10.3390/app15010425

Darmansyah, M. R., Lubis, A. H., & Kamaruddin, M. I. H. (2024). *The Sentiment Analysis Utilization for Indonesian SMEs*.

Dehghani, M., & Yazdanparast, Z. (2023). *Political Sentiment Analysis of Persian Tweets Using CNN-LSTM Model*.

Dirfas, N. A., & Nastiti, V. R. S. (2024). Perbandingan Kinerja Pre-*Train*ed *Embedding* Terhadap Performa Klasifikasi Sentimen Ulasan Produk Tokopedia Dengan Long Short-Term Memory(LSTM). *Building of Informatics, Technology and Science (BITS)*, *6*(2). https://doi.org/10.47065/bits.v6i2.5634

el haddaoui, B., CHIHEB, R., CHIHEB, R., & EL AFIA, A. (2022). LSTM based models stability in the context of Sentiment Analysis for sosial media.

Synthesis Lectures on Human Language Technologies, *5*(1), 1–184. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Elhassan, N., Varone, G., Ahmed, R., Gogate, M., Dashtipour, K., Almoamari, H., El-Affendi, M. A., Al-Tamimi, B. N., Albalwy, F., & Hussain, A. (2023). Arabic Sentiment Analysis Based on Word *Embedding*s and Deep Learning. *Computers*, *12*(6), 126. https://doi.org/10.3390/computers12060126

Evizariza, Palandi, E. H., & Amaliah. (2025). Semantic Study of the Impact of the Use of Slang on the Sosial Identity of Adolescents in Indonesia. *Journal of Hunan University Natural Sciences*, *52*(7). https://doi.org/10.55463/issn.1674-2974.52.7.4

Fernando, R., Proboningrum, Y. D., Supriati, S. D., & Nurmalitasari, N. (2025). NLP Implementation For AI Generated Text Detection (ChatGPT) Using Naive Bayes Method. *J-INTECH*, *13*(02), 292–302. https://doi.org/10.32664/j-intech.v13i02.2026

Hilberts, S., Govers, M., Petelos, E., & Evers, S. (2025). The Impact of Misinformation on Sosial Media in the Context of Natural Disasters: Narrative Review. *JMIR Infodemiology*, *5*, e70413–e70413. https://doi.org/10.2196/70413

Karakaya, O., & Kilimci, Z. H. (2024). An efficient consolidation of *embedding* and deep learning techniques for classifying anticancer peptides: *FastText* +BiLSTM. *PeerJ Computer Science*, *10*, e1831. https://doi.org/10.7717/peerj-cs.1831

Khano, M. N. A. P., Saputro, D. R. S., Sutanto, & Wibowo, A. (2023). sentiment analysis with long-short term memory (lstm) and gated recurrent unit (gru)

algorithms. *Barekeng*, *17*(4), 2235–2242. https://doi.org/10.30598/barekengvol17iss4pp2235-2242

Krichen, M., & Mihoub, A. (2025). Long Short-Term Memory Networks: A Comprehensive Survey. *AI*, *6*(9), 215. https://doi.org/10.3390/ai6090215

Ladayya, F., Rahayu, W., Rohimah, S. R., Saputra, F. R., Maulana, T. A., & Madinah, N. N. (2025a). performance evaluation of *embedding* techniques in twitter sentiment analysis using lstm. *Jurnal Statistika Dan Aplikasinya*, *9*(2), 55–68. https://doi.org/10.21009/JSA.09206

Ladayya, F., Rahayu, W., Rohimah, S. R., Saputra, F. R., Maulana, T. A., & Madinah, N. N. (2025b). performance evaluation of *embedding* techniques in twitter sentiment analysis using lstm. *Jurnal Statistika Dan Aplikasinya*, *9*(2), 55–68. https://doi.org/10.21009/JSA.09206

Lai, S., Hu, X., Xu, H., Ren, Z., & Liu, Z. (2023). *Multimodal Sentiment Analysis: A Survey*. http://arxiv.org/abs/2305.07611

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, *36*(4), 102048. https://doi.org/10.1016/j.jksuci.2024.102048

Mars, M. (2022). From Word *Embedding*s to Pre-*Train*ed Language Models: A State-of-the-Art Walkthrough. *Applied Sciences*, *12*(17), 8805. https://doi.org/10.3390/app12178805

Memiş, E., Akarkamçı (Kaya), H., Yeniad, M., Rahebi, J., & Lopez-Guede, J. M. (2024). Comparative Study for Sentiment Analysis of Financial Tweets with

Deep Learning Methods. *Applied Sciences*, *14*(2), 588. https://doi.org/10.3390/app14020588

Nip, J. Y. M., & Berthelier, B. (2024). Sosial Media Sentiment Analysis. *Encyclopedia*, *4*(4), 1590–1598. https://doi.org/10.3390/encyclopedia4040104

Olakangil, A., Jethwa, K., Wang, C., Li, J., Nguyen, J., Narendra, A., Zhou, Q., Patel, N., Rajaram, A., & Fremd, W. (2023). *Exploring Embeddings for Measuring Text Relatedness: Unveiling Sentiments and Relationships in Online Comments*.

Penggalih, S. A., Mujilahwati, S., & Bettaliyah, A. A. (2025). Twitter Sentiment Analysis to Assess Public Opinion on Jokowi's Performance Over Two Periods using the Recurrent Neural Network (RNN) Method. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, *5*(1), 1860–1869. https://doi.org/10.59934/jaiea.v5i1.1736

Periti, F., Picascia, S., Montanelli, S., Ferrara, A., & Tahmasebi, N. (2025). Studying word meaning evolution through incremental *Semantic Shift* detection. *Language Resources and Evaluation*, *59*(2), 1363–1399. https://doi.org/10.1007/s10579-024-09769-1

Poetra, C. K., Pane, S. F., & Fatonah, N. S. (2022). Meningkatkan Akurasi Long-Short Term Memory (LSTM) pada Analisis Sentimen Vaksin Covid-19 di Twitter dengan Glove. *Jurnal Telematika*, *16*(2), 85–90. https://doi.org/10.61769/telematika.v16i2.400

Qixuan, Y. (2024). *Three-Class Text Sentiment Analysis Based on LSTM*. http://arxiv.org/abs/2412.17347

Raja Azian, Nola Ritha, & Muhamad Radzi Rathomi. (2025). implementasi teknik *embedding* untuk rekomendasi hasil pencarian katalog online menggunakan algoritma *Word2Vec*. *Jurnal Sustainable: Jurnal Hasil Penelitian Dan Industri Terapan*, *12*(2), 37–41. https://doi.org/10.31629/sustainable.v12i2.5956

Rehman, A. U. (2025). Unveiling Public Opinion: A Study of Sentiment Analysis Using LSTM and Traditional Models. *2025 4th International Conference on Communication, Computing and Digital Systems (C-CODE)*, 1–6. https://doi.org/10.1109/C-CODE67372.2025.11204093

Rizky, M. Z. F., Sibaroni, Y., & Prasetiyowati, S. S. (2024). Sentiment Analysis on TikTok App using Long Short-Term Memory (LSTM) with Stochastic Gradient Descent (SGD) Optimization. *jurnal media informatika budidarma*, *8*(3), 1292. https://doi.org/10.30865/mib.v8i3.7699

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023a). A review on sentiment analysis from sosial media *platform*s. In *Expert Systems with Applications* (Vol. 223). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2023.119862

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023b). A review on sentiment analysis from sosial media *platform*s. *Expert Systems with Applications*, *223*, 119862. https://doi.org/10.1016/J.ESWA.2023.119862

Saputra, D., Damayanti, V. S., Mulyati, Y., & Rahmat, W. (2023). Expressions of the use of slang among millennial youth on sosial media and its impact of the extension of Indonesia in society. *BAHASTRA*, *43*(1), 21–40. https://doi.org/10.26555/bs.v43i1.325

Sarıtaş, K., Öz, C. A., & Güngör, T. (2024). A comprehensive analysis of static *word embeddings* for Turkish. *Expert Systems with Applications*, *252*, 124123. https://doi.org/10.1016/J.ESWA.2024.124123

Savitri, P. W., & Dewi, A. A. S. S. S. (2023). Semantic Change on Imitative Slang Used by Indonesian Netizen. *Lingual: Journal of Language and Culture*, *15*(1), 43. https://doi.org/10.24843/LJLC.2023.v15.i01.p06

Shen, Y. (2024). Impact of sosial media on the evolution of English semantics through linguistic analysis. *Forum for Linguistic Studies*, *6*(2). https://doi.org/10.59400/fls.v6i2.1184

Siddiqui, A. G., Sumbul Ghulamani, & Sadam Hussain. (2025). A Comparative Analysis of Text Sentiment Analysis Algorithms using Sosial Media Tweets. *Pakistan Journal of Engineering and Technology*, *8*(3), 10–18. https://doi.org/10.51846/vol8iss3pp10-18

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*(2), 102342. https://doi.org/10.1016/j.is.2023.102342

Sim, J., Huang, X., Horan, M. R., Stewart, C. M., Robison, L. L., Hudson, M. M., Baker, J. N., & Huang, I.-C. (2023). Natural language processing with *machine learning* methods to analyze unstructured patient-reported outcomes derived from electronic health records: A systematic review. *Artificial Intelligence in Medicine*, *146*, 102701. https://doi.org/10.1016/j.artmed.2023.102701

Siti Khomsah, Rima Dias Ramadhani, & Sena Wijaya. (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

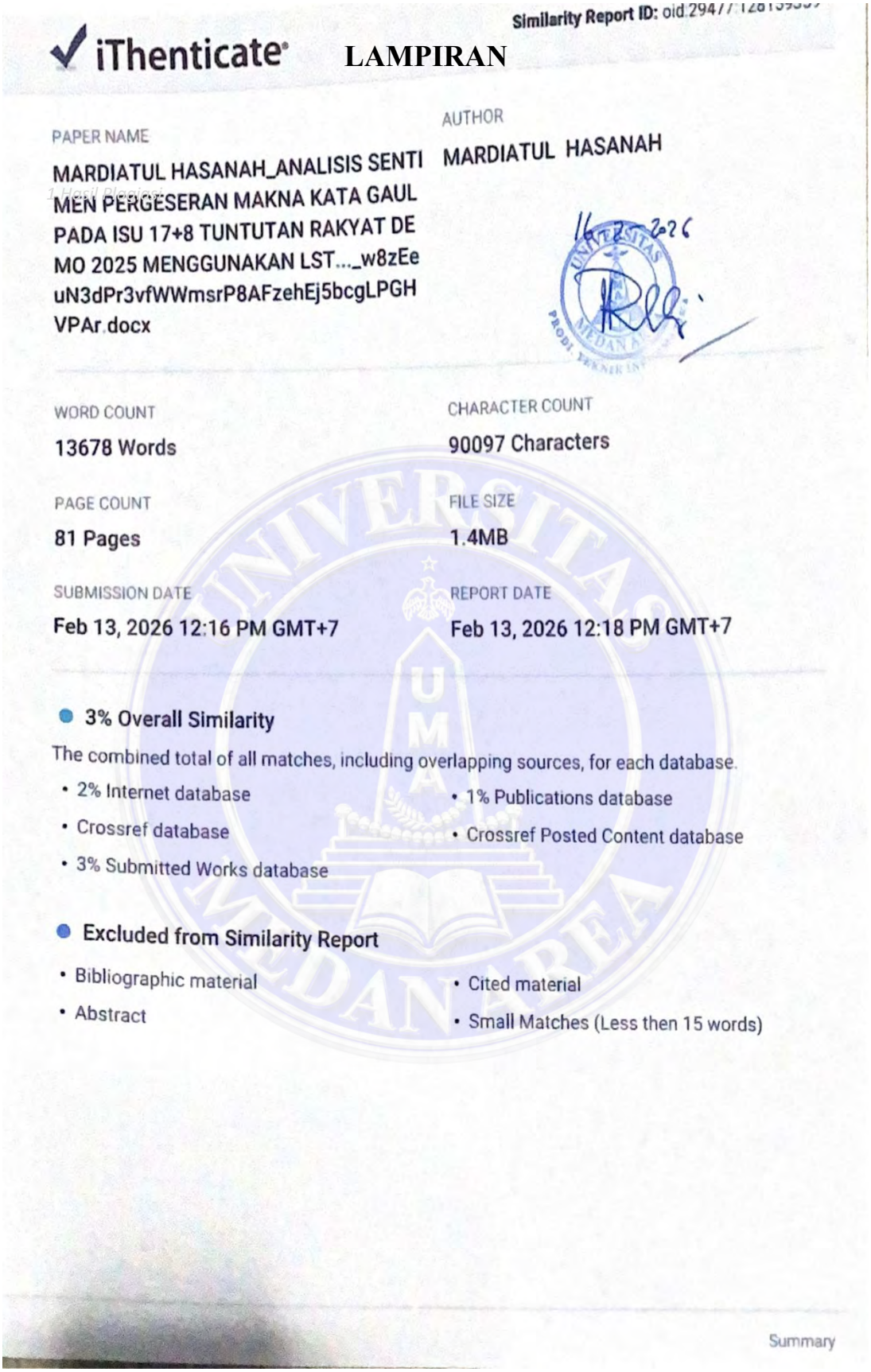
Wang, Z., Wu, J., Wang, Y., Wang, W., Yang, J., Johnson, J., Sastry, N., & De, S. (2024a). *Revealing COVID-19's Sosial Dynamics: Diachronic Semantic Analysis of Vaccine and Symptom Discourse on Twitter*.

Wang, Z., Wu, J., Wang, Y., Wang, W., Yang, J., Johnson, J., Sastry, N., & De, S. (2024b). *Revealing COVID-19's Sosial Dynamics: Diachronic Semantic Analysis of Vaccine and Symptom Discourse on Twitter*. http://arxiv.org/abs/2410.08352

Wangchuk, T., Orcid, [, & Gonsalves, T. (2025). *Comparative Analysis of Tokenization Algorithms for Low-Resource Language Dzongkha*.

Zhang, C., Peng, B., Sun, X., Niu, Q., Liu, J., Chen, K., Li, M., Feng, P., Bi, Z., Liu, M., Zhang, Y., Song, X., Fei, C., Yin, C. H., Yan, L. K., He, H., & Wang, T. (2025). *From Word Vectors to Multimodal Embeddings: Techniques, Applications, and Future Directions For Large Language Models*. http://arxiv.org/abs/2411.05036

Zhang, Y., Lin, Z., Tong, C. C., & Ho, S. W. (2025). Enhancing tokenization *accuracy* with dynamic patterns: cumulative logic for segmenting user-generated content in logographic languages. *Journal of Computational Sosial Science*, *8*(3), 80. https://doi.org/10.1007/s42001-025-00406-7



Similarity Report ID: oid:294[illegible]

iThenticate

LAMPIRAN

PAPER NAME

MARDIATUL HASANAH_ANALISIS SENTIMEN PERGESERAN MAKNA KATA GAUL PADA ISU 17+8 TUNTUTAN RAKYAT DEMO 2025 MENGGUNAKAN LST..._w8zEeuN3dPr3vfWWmsrP8AFzehEj5bcgLPGHVPAr.docx

AUTHOR

MARDIATUL HASANAH

1 Word Playground

16 2 2026

WORD COUNT

13678 Words

CHARACTER COUNT

90097 Characters

PAGE COUNT

81 Pages

FILE SIZE

1.4MB

SUBMISSION DATE

Feb 13, 2026 12:16 PM GMT+7

REPORT DATE

Feb 13, 2026 12:18 PM GMT+7

3% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 2% Internet database
- 1% Publications database
- Crossref database
- Crossref Posted Content database
- 3% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Cited material
- Abstract
- Small Matches (Less then 15 words)

Summary

Document Accepted 1/7/26

UNIVERSITAS MEDAN AREA
FAKULTAS TEKNIK
Kampus I : Jalan Kolam Nomor 1 Medan Estate / Jalan Gedung PBSI, Medan 20223
Kampus II : Jalan Sei Serayu Nomor 70 A / Jalan Setia Budi Nomor 79 B, Medan 20112 Telepon : (061) 8225602, 8201994
Fax : (061) 8226331 HP : 0811 607 259 website: www.uma.ac.id Email : univ_medanarea@uma.ac.id

Nomor : 2457/FT/01.10/X/2025 1 Oktober 2025
Lampiran : -
Hal : **Pembimbing Tugas Akhir**

Yth. Pembimbing Tugas Akhir
Dr Arnes Sembiring ST, M. Kom (Sebagai Pembimbing)
di Tempat

Dengan hormat, sehubungan telah dipenuhinya persyaratan untuk memperoleh Tugas Akhir dari mahasiswa atas :

Nama : MARDIATUL HASANAH
NIM : 228160009
Jurusan : TEKNIK INFORMATIKA

Maka dengan hormat kami mengharapkan kesediaan saudara :

Dr Arnes Sembiring ST, M. Kom (Sebagai Pembimbing)

Adapun Tugas Akhir Skripsi berjudul :

Analisis sentimen pergeseran makna kata gaul pada isu 17+8 Tuntutan Rakyat Demo 2025 dengan LSTM

SK Pembimbing ini berlaku selama enam bulan terhitung sejak SK ini diterbitkan. Jika proses pembimbing melebihi batas waktu yang telah ditetapkan, SK ini dapat ditinjau ulang.

Demikian kami sampaikan, atas kesediaan saudara diucapkan terima kasih.

Dekan,

Dr Eng. Supriatno.ST, MT.

UNIVERSITAS MEDAN AREA
FAKULTAS TEKNIK
Kampus I : Jalan Kolam Nomor 1 Medan Estate ☎ (061) 7360168, Medan, 20223
Kampus II : Jalan Setiabudi Nomor 79 / Jalan Sei Serayu Nomor 70 A ☎ (061) 42402994, Medan, 20122
Website: www.teknik.uma.ac.id E-mail: univ_medanarea@uma.ac.id

Nomor : 347/FT.6/01.10/XI/2025 27 November 2025
Lamp : -
Hal : **Penelitian Dan Pengambilan Data Tugas Akhir**

Yth. Wakil Rektor Bidang Mutu Sumber Daya Manusia dan Perekonomian
Jln. Kolam No.1
Di
Medan

Dengan hormat, kami mohon kesediaan bapak kiranya berkenan untuk memberikan izin dan kesempatan kepada mahasiswa kami tersebut dibawah ini :

NO	NAMA	NPM	PRODI
1	Mardiatul Hasanah	228160009	Teknik Informatika

Untuk melaksanakan Penelitian dan Pengambilan Data Tugas Akhir di **Laboratorium Komputer Program Studi Teknik Informatika Fakultas Teknik Universitas Medan Area**.

Perlu kami jelaskan bahwa Pengambilan Data tersebut adalah semata-mata untuk tujuan Ilmiah dan Skripsi, yang merupakan salah satu syarat bagi mahasiswa tersebut untuk mengikuti ujian sarjana pada Fakultas Teknik Universitas Medan Area dan tidak untuk dipublikasikan, dengan judul :

Analisis sentimen pergeseran makna kata gaul pada isu 17+8 Tuntutan Rakyat Demo 2025 dengan *LSTM*.

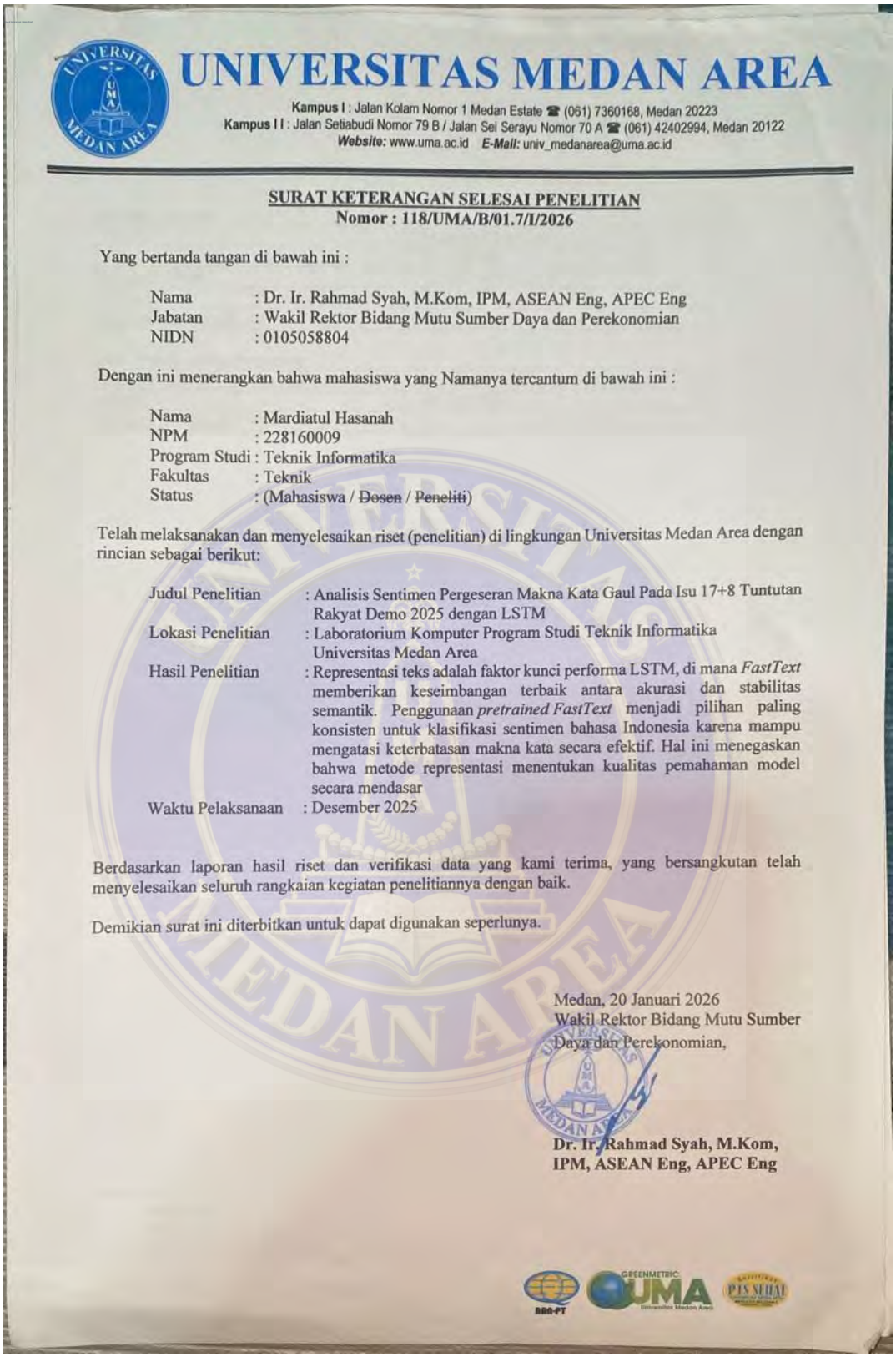
Mohon kiranya tanggal Surat Izin Pengambilan Data Tugas Akhir agar disesuaikan dengan tanggal Terbitnya Surat ini.

Atas perhatian dan kerja sama yang baik diucapkan terima kasih.

Dekan

Dr. Ir. Supriatno, ST, MT

Tembusan :
1. Ka. BPMPP
2. Mahasiswa
3. File



UNIVERSITAS MEDAN AREA

Kampus I : Jalan Kolam Nomor 1 Medan Estate ☎ (061) 7360168, Medan 20223
Kampus II : Jalan Setiabudi Nomor 79 B / Jalan Sei Serayu Nomor 70 A ☎ (061) 42402994, Medan 20122
Website: www.uma.ac.id E-Mail: univ_medanarea@uma.ac.id

SURAT KETERANGAN SELESAI PENELITIAN
Nomor : 118/UMA/B/01.7/I/2026

Yang bertanda tangan di bawah ini :

Nama : Dr. Ir. Rahmad Syah, M.Kom, IPM, ASEAN Eng, APEC Eng
Jabatan : Wakil Rektor Bidang Mutu Sumber Daya dan Perekonomian
NIDN : 0105058804

Dengan ini menerangkan bahwa mahasiswa yang Namanya tercantum di bawah ini :

Nama : Mardiatul Hasanah
NPM : 228160009
Program Studi : Teknik Informatika
Fakultas : Teknik
Status : (Mahasiswa / ~~Dosen~~ / ~~Peneliti~~)

Telah melaksanakan dan menyelesaikan riset (penelitian) di lingkungan Universitas Medan Area dengan rincian sebagai berikut:

Judul Penelitian	: Analisis Sentimen Pergeseran Makna Kata Gaul Pada Isu 17+8 Tuntutan Rakyat Demo 2025 dengan LSTM
Lokasi Penelitian	: Laboratorium Komputer Program Studi Teknik Informatika Universitas Medan Area
Hasil Penelitian	: Representasi teks adalah faktor kunci performa LSTM, di mana *FastText* memberikan keseimbangan terbaik antara akurasi dan stabilitas semantik. Penggunaan *pretrained FastText* menjadi pilihan paling konsisten untuk klasifikasi sentimen bahasa Indonesia karena mampu mengatasi keterbatasan makna kata secara efektif. Hal ini menegaskan bahwa metode representasi menentukan kualitas pemahaman model secara mendasar
Waktu Pelaksanaan	: Desember 2025

Berdasarkan laporan hasil riset dan verifikasi data yang kami terima, yang bersangkutan telah menyelesaikan seluruh rangkaian kegiatan penelitiannya dengan baik.

Demikian surat ini diterbitkan untuk dapat digunakan seperlunya.

Medan, 20 Januari 2026
Wakil Rektor Bidang Mutu Sumber Daya dan Perekonomian,

Dr. Ir. Rahmad Syah, M.Kom, IPM, ASEAN Eng, APEC Eng