BAB III
METODE

3.1 Jenis Penelitian

Jenis penelitian yang digunakan adalah penelitian kuantitatif. Pendekatan kuantitatif digunakan untuk mengetahui perbandingan performa dua model analisis sentimen yaitu LSTM dan IndoBERT terhadap opini publik mengenai program MSIB. Kedua model tersebut diterapkan untuk mengklasifikasikan data sentimen dari media sosial. Hasil klasifikasi kedua model akan dievaluasi menggunakan metrik evaluasi klasifikasi seperti seperti *accuracy, precision, recall* dan *F1-score* untuk menentukan model yang lebih unggul dalam analisis sentimen.

Pemilihan pendekatan kuantitatif didasarkan pada kemampuannya dalam mengukur dan mengevaluasi data sentimen melalui pengumpulan data numerik serta analisis statistik. Karakteristik pendekatan ini sangat sesuai dengan tujuan penelitian yang berfokus pada pengujian performa model secara objektif dan terukur, sehingga hasil analisis dapat digunakan untuk menarik kesimpulan mengenai performa model yang diuji untuk analisis sentimen.

3.2 Desain Penelitian

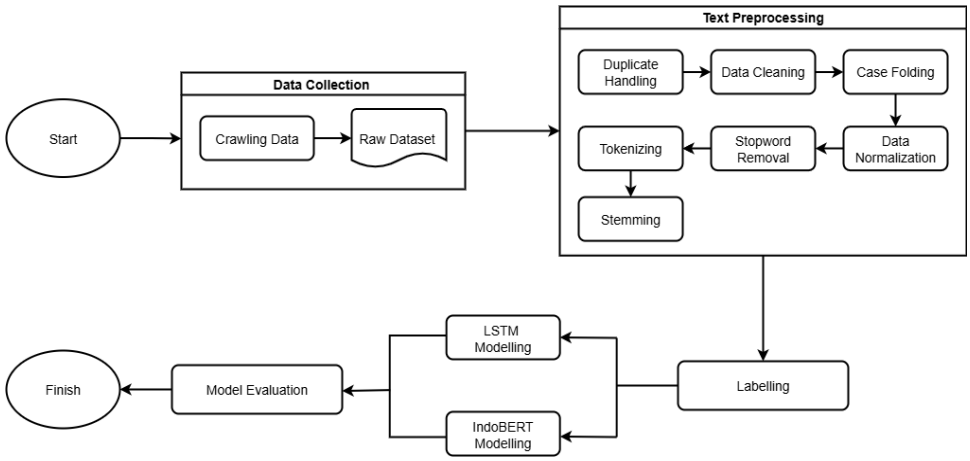
Desain penelitian yang digunakan adalah kuantitatif dengan pendekatan deskriptif komparatif. Penelitian komparatif diterapkan untuk membandingkan performa dua model *deep learning* dalam klasifikasi sentimen. Pendekatan ini dipilih untuk mengukur dan mengevaluasi data sentimen secara objektif dan sistematis. Adapun pendekatan deskriptif digunakan untuk menganalisis dan menyajikan gambaran kecenderungan opini yang berkembang di masyarakat.

Pendekatan deskriptif dalam penelitian ini untuk mengidentifikasi sentimen positif dan negatif yang terdapat dalam komentar pengguna di platform media sosial X. Pendekatan ini memungkinkan untuk mendapatkan gambaran mengenai persepsi publik yang hasilnya akan disajikan dalam bentuk deskripsi kuantitatif. Pengukuran sentimen ini menjadi dasar untuk perbandingan performa model dalam klasifikasi teks.

Alya Sahrani, 2025
KOMPARASI ANALISIS SENTIMEN PUBLIK DENGAN MODEL LONG SHORT-TERM MEMORY (LSTM) DAN INDONESIA BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (INDOBERT) (STUDI KASUS : PROGRAM MSIB)
Universitas Pendidikan Indonesia | repository.upi.edu | perpustakaan.upi.edu

3.3 Prosedur Penelitian

Diagram alir berikut ini menggambarkan secara keseluruhan proses desain penelitian yang dilakukan :



Gambar 3. 1 *Flowchart* Prosedur Penelitian

Penelitian ini terdiri dari beberapa tahapan yang dimulai dengan proses pengumpulan data dari platform media sosial X dengan teknik *crawling*. Data yang diperoleh kemudian melalui tahap *text preprocessing* yang mencakup teknik penanganan duplikasi (*duplicate handling*), pembersihan data (*data cleaning*), *case folding*, normalisasi data (*data normalization*), penghapusan *stopword* (*stopword removal*), tokenisasi (*tokenizing*) dan *stemming*. Proses ini bertujuan untuk membersihkan dan menyiapkan data teks mentah agar siap digunakan dalam analisis.

Setelah *preprocessing* selesai, data diberi label sentimen untuk mengklasifikasikan opini menjadi kategori positif atau negatif. Tahap berikutnya adalah pemodelan (*modelling*). Dalam tahapan ini dua model machine learning yaitu LSTM dan IndoBERT digunakan untuk melatih dan menguji data. Tahap akhir adalah evaluasi performa model, yang dilakukan dengan menggunakan metrik seperti akurasi, presisi, *recall* dan *F1-score* untuk menentukan model yang paling optimal. Berikut penjelasan dalam setiap tahapan yang dilakukan dalam penelitian ini :

3.3.1 *Data Collection*

Dalam penelitian ini, data dikumpulkan menggunakan teknik *crawling* data yaitu pengambilan informasi secara otomatis dari internet termasuk media sosial X. Proses ini memanfaatkan URL untuk menemukan dan mengindeks tautan lalu mengekstrak kata-kata dari dokumen di setiap tautan. Penelitian ini menggunakan layanan *tweet-harvest* untuk mencari *tweet* berdasarkan kata kunci tertentu. Data yang telah terkumpul disimpan dalam file .csv, berisi *tweet* relevan sesuai kata kunci dengan rentang waktu yang telah ditentukan.

3.3.2 *Text Preprocessing*

Tahapan selanjutnya yaitu melakukan *preprocessing* pada data yang sudah disimpan dalam bentuk file .csv. *Preprocessing* dilakukan dengan menggunakan bahasa pemrograman python. Hal pertama yang dilakukan yaitu menghilangkan duplikasi pada data. Setelah itu, dilakukan *preprocessing* lanjutan yaitu *duplicate handling*, *data cleaning*, *case folding*, *data normalization*, *stopword removal*, *tokenizing* dan *stemming.*

1. ***Duplicate Handling***

   Pada tahap *duplicate handling* atau penanganan duplikasi, data yang memiliki kemiripan dihapus karena dapat menyebabkan redudansi dan menurunkan keakuratan analisis. Dalam proses ini, *tweet* dengan isi serupa akan dihapus, sehingga hanya satu *tweet* yang dipertahankan. Penghapusan duplikasi dilakukan berdasarkan kesamaan pada kolom *username* dan *text*, sehingga tidak ada *tweet* dengan isi dan pengguna yang sama yang muncul lebih dari sekali dalam dataset.

2. ***Data Cleaning***

   Pembersihan data (*Data Cleaning*) adalah tahap untuk membersihkan data dari komponen-komponen yang dapat mengganggu proses analisis. Pada tahap ini, kalimat-kalimat akan dihapus dari elemen seperti tanda baca, URL, *username*, *retweet*, karakter HTML dan *username* atau *display name* mengandung kata MSIB. Tahap pembersihan data dilakukan dengan memanfaatkan *library regular expression* (*regex*) guna mengenali dan menghapus pola spesifik dalam data.

3. *Case Folding*

Tahapan ini mengubah semua huruf dalam teks menjadi huruf kecil atau huruf besar. Selain mengubah huruf, karakter seperti tanda baca atau spasi dianggap sebagai delimiter yang dapat diabaikan atau dihapus. Tujuan utama dari *case folding* adalah menyeragamkan format teks untuk mencegah kesalahan dalam pemrosesan karena teks yang sama dapat menghasilkan hasil berbeda jika menggunakan kombinasi huruf kapital dan huruf kecil.

4. *Data Normalization*

Data normalization merupakan tahapan dalam *text preprocessing* yang bertujuan untuk mengganti kata-kata informal termasuk singkatan atau serapan dari bahasa asing ke dalam bentuk kata yang lebih umum digunakan. Proses ini dilakukan untuk mempermudah pemahaman dan pengolahan teks, sekaligus mengurangi *noise* yang diakibatkan oleh beragam variasi kata gaul, yang dapat memengaruhi akurasi pemodelan atau analisis teks.

5. *Stopword Removal*

Penghapusan *stopword* (*stopword removal*) merupakan tahapan yang berfokus pada penyaringan kata-kata penting. Proses ini dilakukan dengan mengeliminasi kata-kata yang dianggap tidak relevan, berdasarkan daftar *stopword* (*stoplist*) yang telah ditentukan. *Stoplist* dalam Bahasa Indonesia umumnya mencakup kata-kata seperti "juga", "yang", "ke", "dan", "itu", "apa", serta kata penghubung seperti "maka" dan "tetapi" yang cenderung tidak memberikan informasi signifikan mengenai inti topik. Tujuan utama dari penghapusan ini adalah untuk menghasilkan representasi teks yang lebih ringkas dan informatif, sehingga mempermudah pemahaman oleh model dan meningkatkan efisiensi analisis teks.

6. *Tokenizing*

Tokenizing atau tokenisasi adalah proses memisahkan teks menjadi kata-kata atau unit kecil yang disebut token. Tahapan ini memudahkan analisis kata-kata dalam teks secara lebih terstruktur.

7. ***Stemming***

Dalam tahapan *stemming*, kata-kata yang memiliki imbuhan diubah menjadi bentuk dasarnya. Tujuan dari *stemming* adalah untuk menemukan bentuk dasar kata dengan menghilangkan imbuhan yang melekat pada kata tersebut.

3.3.3 *Labelling*

Labelling atau pelabelan merupakan proses pemberian nilai sentimen pada suatu teks yang dapat berupa positif atau negatif. Pelabelan data dilakukan secara manual dengan menganalisis kata-kata (*lexical*) yang muncul dalam kalimat. *Lexical* merujuk pada kumpulan kata yang ada dalam kalimat tersebut. Selain itu, konteks kalimat juga diperhatikan agar pelabelan yang dilakukan lebih objektif dan akurat.

3.3.4 *Modelling*

Pada tahap *modelling*, data yang telah diberi label akan diuji menggunakan dua model, yaitu LSTM dan IndoBERT. Pemodelan bertujuan untuk mengukur seberapa akurat model dalam mengklasifikasikan data berdasarkan sentimennya, yang dibagi menjadi dua kategori yaitu positif dan negatif. Penggunaan kedua model ini memungkinkan perbandingan hasil antara model LSTM dan IndoBERT. Tahapan ini diawali dengan membagi dataset yang sudah dilakukan *preprocessing* dan pelabelan menjadi 3 bagian yaitu data pelatihan (*training*), data validasi (*validation*) dan data pengujian (*testing*). Ketiga data tersebut dibagi dalam proporsi 7:2:1 yaitu 70 % untuk pelatihan, 20 % untuk validasi dan 10 % untuk pengujian. Pembagian data ini berlaku untuk kedua model yaitu LSTM dan IndoBERT.

Dalam pemodelan LSTM dan IndoBERT, perlu dilakukan *hyperparameter tuning* untuk mendapatkan performa model yang optimal. Parameter yang dioptimasi pada model LSTM diantaranya jumlah *neurons,* nilai *dropout* dan *learning rate*. Optimasi nilai *dropout* dilakukan untuk mencegah terjadinya *overfitting* pada saat proses pelatihan (*training*). Selain itu, *hyperparameter tuning* juga dilakukan pada model IndoBERT dengan melakukan optimasi pada nilai

dropout dan *learning rate*. Adapun jumlah *epochs* di tiap model berbeda dalam tiap model disesuaikan dengan karakteristik dari masing-masing model.

3.3.5 *Model Evaluation*

Evaluasi dalam penelitian ini menggunakan metrik evaluasi klasifikasi yang dapat diperhatikan dari empat metrik utama yaitu *Accuracy*, *Precision*, *Recall* dan *F1-Score*. Keempat metrik tersebut dihitung berdasarkan jumlah prediksi yang benar dan salah, baik untuk kategori positif maupun negatif, yang meliputi *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)* yang dirangkum dalam *confusion matrix*.

3.4 Populasi dan Sampel

Berdasarkan data yang telah dikumpulkan menggunakan teknik *crawling* melalui media sosial X dengan kata kunci tertentu yang relevan dengan program MSIB dalam rentang waktu 17 Oktober 2023 - 31 Januari 2025. Data yang didapat sebanyak 3593 data yang menjadi populasi dalam penelitian ini. Adapun sampel dalam penelitian ini sebanyak 1041 data yang merupakan hasil dari seluruh tahapan *pre-processing* data. Sampel tersebut kemudian akan digunakan untuk tahapan lebih lanjut untuk pemodelan hingga evaluasi model.

3.5 Instrumen Penelitian

Instrumen penelitian merupakan alat yang digunakan untuk mengumpulkan, mengolah, menganalisis, serta menyajikan data secara sistematis dan objektif. Dalam penelitian ini, instrumen yang digunakan meliputi beberapa komponen utama berbasis perangkat lunak untuk mendukung seluruh proses analisis sentimen berbasis data digital, mulai dari pengumpulan hingga analisis. Untuk pengumpulan data, peneliti memanfaatkan teknik *crawling* data dari media sosial X. Dalam prosesnya, teknik *crawling* ini menggunakan bahasa pemrograman Python sebagai instrumen untuk mengakses dan mengunduh data dari platform secara otomatis.

Data yang telah dikumpulkan selanjutnya masuk ke tahapan pra-pemrosesan untuk membersihkan dan menyiapkan data. Proses ini melibatkan penggunaan pustaka pemrosesan teks seperti *NLTK*, *Sastrawi*, *Pandas*, dan *Regex* untuk

tahapan *tokenization*, *normalization*, *stopword removal*, dan *stemming*. Setelah melalui tahapan pra-pemrosesan, data kemudian dilabeli secara manual untuk mengidentifikasi sentimen positif atau negatif. Data yang telah berlabel selanjutnya digunakan dalam tahapan pemodelan menggunakan model LSTM dan IndoBERT. Seluruh proses dalam pemodelan dan evaluasi ini dilakukan menggunakan platform analisis data seperti Google Colab, dengan bahasa pemrograman python sebagai instrumen kerja utama.

3.6 Teknik Pengumpulan Data

Proses pengumpulan data dalam penelitian ini dilakukan dengan menggunakan teknik *crawling* data, yaitu pengambilan informasi secara otomatis dari berbagai sumber di internet, termasuk media sosial X. *Crawling* dilakukan dengan memasukkan URL untuk menemukan tautan-tautan di situs web yang kemudian diindeks untuk mengekstrak kata-kata dalam dokumen yang terdapat pada setiap tautan. Untuk mengakses data *tweet* di X, diperlukan *auth-token* yang bisa diperoleh melalui akun X. Dalam penelitian ini, pengambilan data tweet dilakukan dengan menggunakan layanan *tweet-harvest* yang memungkinkan pencarian berdasarkan kata kunci tertentu.

Pada penelitian ini, kata kunci yang digunakan adalah “magang studi independen bersertifikat”, “magang bersertifikat”, “msib”, “magang bersertifikat kampus merdeka”. “magang msib”, “magang bersertifikat msib”, “magang kampus merdeka”, “kebijakan msib”, “#magangbersertifikat” dan “#msib” dengan rentang waktu pengambilan data antara 17 Oktober 2023 hingga 31 Januari 2025. Rentang waktu ini dipilih berdasarkan periode meningkatnya perhatian publik terhadap program MSIB. Setelah data diambil, hasilnya akan disimpan dalam file berformat .csv, yang berisi tweet dengan kata kunci yang relevan serta sesuai dengan rentang waktu yang telah ditentukan.

Setelah data dikumpulkan, langkah selanjutnya adalah evaluasi kualitas data. Evaluasi ini mencakup pemeriksaan awal untuk memahami format, struktur dan jenis data yang diperoleh, serta untuk mengidentifikasi potensi masalah yang ada dalam data. selanjutnya dilakukan *filtering* data untuk menghilangkan *spam*

atau *tweet* yang tidak relevan. Data yang telah dievaluasi akan digunakan pada tahapan berikutnya dalam penelitian.

3.7 Teknik Analisis Data

Teknik analisis data pada penelitian ini menggunakan analisis data kuantitatif deskriptif. Pada penelitian ini, data kuantitatif diolah dengan melakukan tahapan preprocessing data termasuk penanganan duplikasi, pembersihan data, *case folding,* normalisasi data, penghapusan *stopword*, tokenisasi dan *stemming*. pembersihan teks, tokenisasi, penghapusan *stopword* dan normalisasi untuk memastikan data dapat dilakukan proses pemodelan. Selanjutnya, model LSTM dan IndoBERT diimplementasikan untuk melakukan proses pelatihan, validasi, dan pengujian terhadap dataset. Evaluasi untuk kedua model dilakukan dengan menggunakan metrik evaluasi klasifikasi seperti *accuracy, precision, recall* dan *F1-score*. Hasil evaluasi model disajikan dalam bentuk tabel perbandingan metrik untuk melihat perbandingan performa antara kedua model. Berikut merupakan tabel yang menjadi indikator metriks evaluasi dalam mengukur evaluasi model :

Tabel 3. 1 Tabel Indikator

Metrik Evaluasi	Rumus Perhitungan	Keterangan
Accuracy	$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$	Akurasi dianggap baik jika model mampu memprediksi sebagian besar data dengan benar, baik untuk kategori positif maupun negatif
Precision	$Precision = \frac{TP}{(\mathrm{TP} + \mathrm{FP})}$	Presisi dianggap baik jika sebagian besar prediksi positif model adalah benar (*true positive*), dengan jumlah prediksi positif palsu (*false positive*) bernilai kecil.

Metrik Evaluasi	Rumus Perhitungan	Keterangan
Recall	$Recall = \frac{TP}{(\mathrm{TP} + \mathrm{FN})}$	*Recall* akan bernilai tinggi jika model mampu menangkap hampir semua data positif yang sebenarnya (*true positive),* serta nilai *false negative* kecil.
F1-Score	$F1\text{-}Score = 2\ x\ \frac{(Precision \times Recall)}{(Precision + Recall)}$	*F1-Score* akan bernilai tinggi jika nilai precision dan recall sama-sama tinggi.