BAB II

TINJAUAN LITERATUR

2.1 Penelitian Terdahulu

Penelitian terdahulu berperan penting dalam memperjelas kerangka berpikir dan menjadi acuan dalam pengembangan metode serta analisis hasil. Dengan meninjau penelitian sebelumnya, peneliti dapat mengidentifikasi kelebihan dan kekurangan pendekatan yang telah digunakan. Oleh karena itu, dalam landasan teori ini, dicantumkan hasil penelitian yang relevan untuk mendukung pemilihan metode dan evaluasi model yang digunakan. Berikut adalah beberapa penelitian terdahulu yang berkaitan dengan penelitian ini dalam table 2.1.

Tabel 2. 1 Penelitian Terdahulu

No	Judul	Metode	Hasil
1	Meningkatkan Akurasi Long-Short Term Memory (LSTM) pada Analisis Sentimen Vaksin Covid-19 di Tweet dengan Glove	Long-Short Term Memory (LSTM), Glove	Hasil yang didapatkan LSTM ditambah Glove juga mampu meningkatkan precision menjadi 89%, dari yang awalnya hanya 82% apabila hanya menggunakan LSTM saja, F1-Score dari metode LSTM ditambah Glove mendapatkan hasil tertinggi di antara ketiga metode yang dijalankan. Nanum mengalami penurunan pada recall menjadi 87%, dibandingkan dengan

			metode LSTM yang dimana menghasilkan recall sebesar 91%
2	Deteksi Depresi pengguna Twitter Indonesia menggunakan LSTM-RNN	LSTM-RNN	Penelitian yang menggunakan metode LSTM-RNN menghasilkan nilai presisi, recall, dan F1-Score masing-masing sebesar 86%, dengan akurasi keseluruhan juga mencapai 86%. Sistem deteksi ujaran depresi ini diharapkan mampu mendukung upaya analisis kondisi depresi di kalangan masyarakat melalui platform media sosial.

2.2 Analisis Sentimen

Analisis sentimen adalah proses mengumpulkan, mengolah, dan menganalisis opini, pemikiran, serta kesan orang terhadap berbagai topik, produk, layanan, atau merek. Melalui teknik analitik teks, data dikumpulkan dari berbagai sumber seperti internet dan platform media sosial untuk mengungkapkan emosi yang tercermin dalam tulisan pengguna. Analisis sentimen membantu perusahaan memahami persepsi konsumen secara otomatis, baik dari ulasan, interaksi media sosial, maupun sumber lainnya, sehingga dapat mendukung pengambilan keputusan strategis. Dalam praktiknya, analisis sentimen mengklasifikasikan teks berdasarkan emosi, membedakan antara kalimat subjektif dan objektif, serta mengidentifikasi pendapat eksplisit maupun implisit. Tingkatan analisis ini meliputi level pesan (message level), kalimat (sentence level), dan aspek tertentu (aspect level),

memungkinkan perusahaan mendapatkan masukan, saran, kritik, atau bahkan ujaran negatif secara lebih mendalam dan efisien [16].

Analisis sentimen terbagi menjadi dua tahap utama, yakni polaritas dan subjektivitas. Polaritas berfungsi untuk menentukan intensitas emosi yang terkandung dalam teks, dengan skala nilai mulai dari -1 yang menandai sentimen negatif, 0 untuk netral, dan +1 yang menunjukkan sentimen positif. Di sisi lain, subjektivitas menilai tingkat keberadaan opini versus fakta dalam teks, menggunakan rentang nilai dari 0 hingga 1, di mana 0 mencerminkan opini murni dan 1 menunjukkan fakta murni [17].

2.3 Media Sosial X

Media sosial merupakan sebuah platform atau sarana yang memberikan kesempatan luas bagi masyarakat dalam mengekspresikan berbagai bentuk pemikiran mereka, mulai dari aspirasi, ide-ide kreatif, hingga kritik terhadap berbagai hal yang terjadi di lingkungan sekitar. Melalui media sosial, setiap individu dapat berpartisipasi dalam komunikasi publik, berbagi informasi, ataupun menyuarakan pendapat mereka secara terbuka. Media sosial sendiri sangatlah mudah dijangkau, dikarenakan dapat dilakukan melalui berbagai perangkat elektronik seperti telepon genggam, laptop, maupun komputer, yang terhubung dengan jaringan internet yang memadai[1].

Salah satu media sosial yang sebelumnya dikenal sebagai twitter merupakan platform media sosial yang memungkinkan penggunanya untuk membagikan pesan singkat atau "tweet" dengan panjang maksimal 280 karakter. Platform ini digunakan secara luas untuk menyampaikan informasi secara cepat, mulai dari berita, opini pribadi, hingga promosi bisnis. Dengan fitur seperti retweet, like, dan hashtag, X menjadi alat komunikasi yang dinamis dan real-time, yang memfasilitasi diskusi publik dan penyebaran tren secara global. Twitter juga sering dimanfaatkan dalam analisis sentimen karena kontennya yang padat, terbuka, dan mencerminkan reaksi pengguna terhadap berbagai isu. Selain sebagai media komunikasi sosial, Twitter juga banyak

dimanfaatkan dalam bidang penelitian, salah satunya dalam analisis sentimen. Hal ini karena Twitter menyediakan data yang terbuka, cepat, dan mencerminkan opini masyarakat secara langsung terhadap suatu peristiwa, produk, atau isu tertentu. Melalui analisis sentimen, data dari Twitter dapat digunakan untuk mengetahui respons emosional pengguna, baik yang bersifat positif, negatif, maupun netral, sehingga sangat berguna dalam pengambilan keputusan berbasis data [1], [2].

2.4 Random OverSampling

Random Oversampling merupakan salah satu teknik resampling yang banyak digunakan dalam pemrosesan data untuk mengatasi masalah ketidakseimbangan kelas (*class imbalance*), terutama pada kasus klasifikasi. Ketidakseimbangan kelas sering kali menjadi kendala serius karena dapat menyebabkan model pembelajaran mesin cenderung bias terhadap kelas mayoritas dan mengabaikan kelas minoritas. Hal ini berdampak pada penurunan performa model, khususnya dalam mengenali pola-pola penting yang terdapat pada kelas dengan jumlah data terbatas.

Teknik *Random Oversampling* bekerja dengan cara memperbanyak data pada kelas minoritas melalui proses penyalinan acak (*random duplication*) hingga jumlah data pada kelas tersebut sebanding dengan kelas mayoritas. Dengan demikian, distribusi dataset menjadi lebih seimbang dan memberikan kesempatan yang sama bagi model untuk mempelajari karakteristik dari kedua kelas. Proses ini pada dasarnya tidak menambah informasi baru, namun mampu memperbaiki representasi data yang digunakan selama pelatihan.

Keseimbangan proporsi data yang dihasilkan melalui Random Oversampling diharapkan dapat meningkatkan kemampuan model dalam mengenali pola dari kelas minoritas secara lebih akurat, sehingga hasil klasifikasi menjadi lebih optimal. Meskipun sederhana, metode ini terbukti efektif untuk meningkatkan akurasi dan metrik evaluasi lainnya pada berbagai penelitian yang melibatkan data tidak seimbang. Dengan demikian, *Random*

Oversampling dapat dianggap sebagai salah satu pendekatan dasar namun penting dalam meningkatkan kualitas model pembelajaran mesin.

2.5 Feature Extraction

GloVe, singkatan dari *Global Vectors for Word Representation*, adalah metode ekstraksi fitur yang mengubah teks menjadi representasi *vektor numerik*. Tujuan utamanya adalah membangun matriks ko-okurensi yang mencatat frekuensi kemunculan kata-kata bersama dalam jendela konteks tertentu. *Word Embedding* GloVe menciptakan matriks ini untuk merepresentasikan keterkaitan antar kata dalam korpus. Dengan memanfaatkan matriks ko-okurensi tersebut, GloVe menghasilkan vektor representasi untuk setiap kata dalam korpus. Pendekatan yang digunakan oleh GloVe adalah faktorisasi matriks global, yang memproses matriks untuk menghasilkan representasi kata yang mencerminkan keberadaan atau ketidakhadiran kata-kata dalam dokumen [18].

Ekstraksi fitur memiliki beberapa kelebihan, termasuk kemampuan menangkap hubungan semantik antar kata, menghasilkan vektor dengan dimensi tetap (seperti 50, 100, atau 300 dimensi), serta ketersediaan model pra-latih pada korpus besar. Selain itu, teknik ini memungkinkan pemahaman kontekstual kata yang baik, sehingga dapat merepresentasikan kesamaan semantik antar kata dan memudahkan identifikasi hubungan makna. GloVe juga memfasilitasi pemrosesan pelatihan data dalam jumlah besar [19]. Algoritma GloVe terdiri dari langkah-langkah sebagai berikut [18]:

2.5.1 Mengumpulkan Statistik *Word Co-occurrence*

Ini merupakan langkah pertama dalam proses feature extraction dengan GloVe. Matriks kemunculan bersama (*co-occurrence matrix*) yang dikumpulkan menjadi dasar untuk membentuk representasi kata dalam bentuk vektor. Matriks ini berfungsi sebagai fitur awal yang akan digunakan untuk membangun vektor kata.

2.5.2 Menentukan Soft Constraints untuk Setiap Pasangan Kata

$$w_i^T w_j + b_i + b_j = log\ (X_{ij}) \quad (1)$$

Keterangan:

- w_i : vector untuk kata utama
- w_j : Vector untuk kata konteks
- $b_i\ dan\ b_j$: bias skalar untuk kata utama dan kata konteks.

2.5.3 Menentukan *Cost Function*

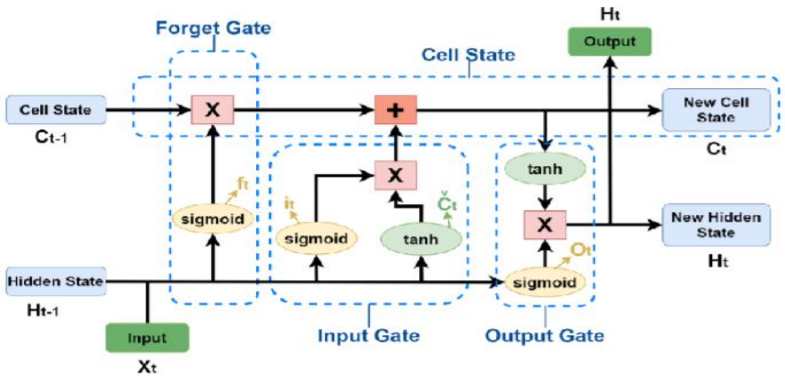
$$J = \sum_{i=1}^{v} \quad \sum_{j=1}^{v} \quad f(X_{ij})(w_i^T w_j + b_i + b_j - logX_{ij}) \quad (2)$$

Keterangan :

- $f(X_{ij})$: Fungsi pembobotan yang membantu mengurangi pengaruh pasangan kata yang sangat umum, sehingga model tidak terlalu berfokus pada pasangan kata yang sering muncul tetapi tidak memberikan informasi yang berarti.

2.6 LSTM

Long Short Term Memory (LSTM) merupakan bagian dari keluarga *Recurrent Neural Network* (RNN), suatu jenis jaringan saraf yang dirancang untuk mengatasi data sekuensial dengan menggunakan bobot internal yang dibagi di seluruh urutan [14]. RNN berfungsi untuk menangkap hubungan temporal antara kata-kata dalam suatu kalimat. Terutama, data tekstual dianggap sebagai deret waktu, di mana urutan kata memegang peran yang sangat penting dalam menentukan makna dari kata dan kalimat tersebut. LSTM dapat dianggap sebagai implementasi khusus dari RNN, dengan adanya koneksi khusus antar node. Komponen khusus dalam struktur LSTM melibatkan input gate, output gate, dan forget gate [20].



Gambar 2. 1 Arsitektur LSTM

Keterangan:

χ_t: Vector Input Saat ini

h_t: Hidden State Saat ini

C_t: Cell state saat ini

LSTM memiliki empat komponen utama yang berperan dalam pemrosesan data, yaitu forget gate, input gate, update gate, dan output gate. Tahapan pertama dalam proses ini adalah forget gate. Forget gate dihitung melalui kombinasi hidden state sebelumnya (h_t) dan input gate saat ini (χt). Hasil dari forget gate akan dikalikan dengan cell state pada timestep sebelumnya (C_{t-1}). Persamaan 1 mencerminkan operasi matematika yang terlibat dalam perhitungan forget gate.hidden state sebelumnya (h_t) dan input gate sekarang (χ_t). Forget gate akan dikalikan dengan cell state timestep sebelumnya (C_{t-1}). Persamaan 1 merupakan operasi matematika forget gate.

$$f_t = \sigma\left(W_f . \left[h_{t-1,} x_t\right] + b_f\right) \qquad (1)$$

Kemudian selanjutnya adalah proses pada input gate. Pada input gateini, terdapat opsi atau pemilihan informasi yang akan diperbaharui untuk ke bagian cell state. Proses pemilihan informasi ini menggunakan fungsi sigmoid juga. Berikut adalah persamaan dari proses input gate.i.

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \qquad (2)$$

$$\mathcal{C}_t = tanW_c . [h_{t-1}, x_t] + b_c \qquad (3)$$

Proses yang ketiga adalah update gate. Pada tahap ini akan menggunakan output dari input gate. Output Dari updategate akan menghasilkan cell state terbaru (Ct). Berikut adalah cara menghitung proses update gate.

$$C_t = f_t * C_{t-1} + i_t * \mathcal{C}_t \qquad (4)$$

Proses yang terakhir yaitu output gate(ot). Pada proses ini akan menghasilkan nilai output dan nilai hidden state(ht). Setelah mendapat hasil dari proses output gate, maka hasil tersebut akan dilanjutkan ke timestep berikutnya

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t \times tan(c_t) \quad (6)$$

2.7 Evaluation Model

Di dunia klasifikasi, akurasi menjadi salah satu ukuran krusial untuk menilai efektivitas sebuah model. Angka akurasi menunjukkan seberapa tepat model tersebut dalam mengklasifikasikan data. Pada penelitian ini, pengukuran akurasi dilakukan melalui teknik *cross-validation*, yang membagi dataset menjadi dua komponen utama: set pelatihan (*training set*) dan set pengujian (*testing set*). Set pelatihan dipakai untuk mengembangkan model, sementara set pengujian berfungsi mengecek kinerjanya. Proses *cross-validation* ini berjalan dalam beberapa epoch atau putaran pelatihan, bertujuan mencegah *overfitting* serta tumpang tindih data pada set pengujian. Setelah pelatihan rampung, model dinilai menggunakan confusion matrix.

Confusion matrix adalah tabel yang mencatat hasil klasifikasi model, baik dari sisi aktual maupun prediksi. Dengan menganalisis data di dalam matriks ini, kita bisa memahami performa model secara mendalam. Confusion matrix untuk klasifikasi bi

Tabel 2. 2 *Confussion Matrix* 3 Kelas

		Predict		
Accuracy		Kelas A	Kelas B	Kelas C
	Kelas A	AA	AB	AC
	Kelas B	BA	BB	BC
	Kelas C	CA	CB	CC

Pada nilai di tabel 2.2, yang mewakili kolom matriks seperti *True Negative* (TN), *False Positive* (FP), *False Negative* (FN), dan *True Positive* (TP), kita dapat menghitung berbagai metrik evaluasi model, termasuk akurasi, presisi, recall, serta F1-Score..

1. Akurasi adalah persentase dari prediksi yang benar dibandingkan dengan total jumlah prediksi yang dibuat oleh model.

$$Accuration = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

2. Recall merupakan gambaran kemampuan model untuk mendeteksi kelas positif yyang diprediksikan dengan benar.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

3. Precision merupakan nilai prediksi dalam mengukur ketepatan antara data yang diminta untuk diberikan oleh model.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

4. F1-Score merupakan gabungan rata-rata pada precision dan recall.

$$F1 - Score = 2.\frac{precision.recall}{precision+recall} \quad (4)$$