BAB II
LANDASAN TEORI

2.1 Penelitian Terkait

Dalam penelitian ini akan menggunakan sepuluh tinjauan studi yang nantinya dapat mendukung penelitian. Berikut ini merupakan tinjauan studi yang digunakan sebagai berikut :

Tabel 2.1 Ringkasan Tinjauan Pustaka

No	Judul Penelitian	Metode	Data	Hasil
1	Algoritma LSTM Untuk Sentimen Klasifikasi dengan Word2vec Pada Media Online.	LSTM, LSTM-CNN, CNN-LSTM	Dataset diambil dari media sosial Detik Finance pada bulan Desember 2017 sampai bulan Desember 2018 yang berjumlah 1.200 data.	Pada penelitian ini dilakukan sebuah prepocessing, dimana proses ini dilakukan untuk membersihkan data teks yang tidak terstruktur tahapan yang dilakukan diantaranya adalah sebagai berikut: Casefolding, Filtering, Tokenization,Sentence Conversion. Hasil pengujian memperlihatkan bahwa metode LSTM, LSTM-CNN, CNN-LSTM memiliki hasil akurasi sebesar, 62%, 65% dan 74%.
2	Analisis Sentimen Data Saran Mahasiswa Terhadap Kinerja Departemen Perguruan Tinggi Menggunakan	CNN	Dataset yang digunakan dari penilaian performa layanan unit departemen	Penelitian ini menggunakan NLTK (Natural Language ToolKit)sebagai library untuk pengolahan kata. Tahapan-tahapan

No	Judul Penelitian	Metode	Data	Hasil
	Convolutional Neural Network		di Politeknik Caltex Riau yang didapatkan dari BP3M PCR dengan jumlah 1.500 data	preprocessing yang diterapkan pada penelitian ini yaitu Case Folding , Stopword Removal, Emoticon Removal, Tokenization. Hasil pengujian akurasi menggunakan metode CNN adalah *Recall* 97%, *Precision* 98% dan *F1-score* 98%.
3	Analisis Sentimen Customer Terhadap Produk Indihome dan First Media Menggunakan Algoritma *Convolutional Neural Network*.	CNN	Jumlah data 13.689 diambil dari *Twitter* dengan tiga label yaitu positif, negatif, dan netral.	Hasil akurasi yang didapatkan, memperoleh akurasi tertinggi sebesar 98% untuk provider IndiHome dan 91% untuk provider First Media.
4	Analisis Sentimen Pada Media Sosial *Twitter* Terhadap Kebijakan Pemberlakuan Pembatasan Kegiatan Masyarakat Berbasis *Deep Learning*.	LSTM	data *Twitter* mulai tanggal 15 Agustus 2021 sampai dengan 24 September 2021 dengan data berjumlah 37.756 *tweet*	Pada penelitian ini ada 5 tahapan preprocessing yang dilakukan yaitu menghilangkan bagian-bagian dari twitter yang tidak merepresentasikan sentimen misalkan mentions, hastag, RT, link, angka, enter, dan tanda baca, kemudian mengubah ke bentuk lowercase, melakukan tokenisasi, melakukan stopword removal,

No	Judul Penelitian	Metode	Data	Hasil
				dan stemming. Hasil penelitian ini dengan algoritma LSTM memperoleh akurasi 87%.
5	Analisis Sentimen Terhadap Pengguna Gojek Menggunakan Metode *K-Nearset Neighbors*.	KNN	Dataset diambil dari *Twitter* berjumlah 1.409 *tweet*	Pada penelitian ini ada 5 tahapan preprocessing yang dilakukan yaitu Cleaning, Case Folding, Tokenizing, Stopword, Stemming. Hasil pengujian metode KNN menggunakan *confusion matrix* mendapatkan tingkat akurasi sebesar 79,43% dengan nilai k=15.
6	Analisis Sentimen Terhadap Review Aplikasi Layanan E-Commerce Menggunakan Metode *Convolutional Neural Network*	CNN	Objek sentimen analisis yang diteliti tentang aplikasi Shopee yang ada di *Google Play*	Hasil menganalisa sentimen kedalam tiga kategori yaitu positif, negatif, dan netral dengan akurasi yang dicapai paling tinggi sebesar 86,6%.
7	Penggunaan Metode GloVe untuk Ekspansi Fitur pada Analisis Sentimen *Twitter* dengan *Naive Bayes* dan *Support Vector*	SVM, *Naive Bayes*	Dataset didapat dengan menggunakan API *Twitter* yang sudah tersedia dengan sebanyak	Hasil pengujian dengan Metode GloVe berhasil diimplementasikan sehingga menghasilkan 3

No	Judul Penelitian	Metode	Data	Hasil
	Machine		16.597 *tweet*	korpus yang digunakan saat ekspansi fitur. Sehingga Peningkatan performa terbaik diperoleh pada Top 5 similarity dengan menggunakan korpus Indonews+Tweet dengan akurasi 83.23% untuk SVM dan 77.86% untuk *Naive Bayes.*
8	Analisis Sentimen *Twitter* Menilai Opini Terhadap Perusahaan Publik Menggunakan Algoritma Deep Neural Network.	DNN	Dataset diambil dari *Twitter* dengan jumlah 5.504 *tweet*	Model tersusun dengan 3 hidden layer dengan susunan node tiap layer pada model tersebut yaitu 128, 256, 128 node dan menggunakan *learning rate* sebesar 0.005, model mampu menghasilkan nilai akurasi mencapai 88.72%.
9	Analisis Sentimen untuk Pengukuran Tingkat Depresi Pengguna *Twitter* Menggunakan *Deep Learning*	CNN	Dataset diambil dari *Twitter* dengan jumlah 3.069 *tweet*	Analisis Sentimen Pilkada Di Tengah Pandemi Covid-19 Menggunakan *Convolutional Neural Network*
10	Analisis Sentimen Pilkada Di Tengah	CNN	500 *tweet* diperoleh dari	Pada penelitian ini ada 5 tahapan

No	Judul Penelitian	Metode	Data	Hasil
	Pandemi Covid-19 Menggunakan *Convolutional Neural Network*		*Twitter* API menggunakan *library tweepy*, lalu diberi label ke dalam 2 kelas	preprocessing yang dilakukan yaitu Penghapusan Karakter, Case Folding, Tokenization, Stopwords Removal, Stemming. Hasil dari penelitian menunjukkan bahwa, metode CNN dengan dataset pilkada ditengah pandemi mendapatkan akurasi tertinggi sebesar 90% dengan 4 layer *convolutional* dan 100 *epoch*. Didapatkan pula bahwa, semakin banyak *epoch* yang digunakan dalam model, akurasi cenderung meningkat
11	Penerapan Metode Recurrent Neural Network Model Gated Recurrent Unit Untuk Prediksi Harga Cryptocurrency	Gated Recurrent Unit (GRU)	Pada penelitian ini menggunakan metode GRU untuk memprediksi harga cryptocurrency, yaitu bitcoin dan ethereum dari tahun 2018	Salah satu tahap preprocessing yang dilakukan adalah windowing berdasarkan nilai window size yang ditentukan agar menjadi data sequence. Berdasarkan hasil

No	Judul Penelitian	Metode	Data	Hasil
			sampai 2021	pengujian, dengan menggunakan nilai window size sebanyak 2, sistem mendapatkan hasil error yang paling kecil. Perhitungan akurasi prediksi untuk 1, 6, dan 12 bulan berikutnya pada data uji bitcoin masing-masing sebesar 90.26%, 77.74%, dan 75.98%, sedangkan pada data uji ethereum masing-masing sebesar 90.15%, 76,88%, dan 66.09%. Dapat dikategorikan sistem prediksi harga cryptocurrency ini tergolong sangat baik untuk memprediksi 1 bulan berikutnya dan dikategorikan cukup untuk memprediksi 6 dan 12 bulan berikutnya.

Berdasarkan hasil literatur yang ada, penulis mengambil kesimpulan yaitu:

1. *Preprocessing* memiliki peran sangat penting dalam klasifikasi data berupa teks sebelum ke model.
2. CNN dapat mengklasifikasi data berupa gambar, teks, dan audio.

3. CNN dapat menghasilkan akurasi cukup baik dengan jumlah data yang besar.
4. CNN menunjukkan bekerja sangat baik dalam melakukan proses klasifikasi data berupa teks.
5. Parameter pada model CNN dengan *drop out* dapat mempengaruhi akurasi dan *loss*, selain itu dapat mencegah terjadinya *overfitting*.

2.2 Teori Dasar

2.2.1 *Twitter*

Twitter adalah sebuah media sosial dan layanan *microblogging* yang mengizinkan penggunanya untuk mengirimkan pesan *realtime*. Pesan yang berupa teks, gambar dan video ini populer dengan sebutan *tweet*. *Twitter* memberikan akses kepada penggunanya untuk mengirimkan pesan singkat (*tweet*) dengan maksimal 140 karakter menjadi 280 karakter [6]. Dikarenakan keterbatasan jumlah karakter yang dapat ditulis, *tweet* sering mengandung singkatan, bahasa gaul atau kesalahan tata Bahasa [7].

2.2.2 Analisis Sentimen

Analisis sentimen adalah proses menganalisis teks dari berbagai sumber data dengan tujuan untuk memperoleh informasi emosional pada suatu kalimat opini [4]. Informasi yang dikumpulkan dapat berupa pendapat umum tentang produk, layanan, kebijakan, dan lainnya. Analisis sentimen adalah cabang dari *text mining* yang bertujuan untuk menganalisis, memahami, mengolah dan mengekstrak data tekstual berupa opini yang menganalisis pendapat, penilaian, evaluasi, sikap, dan perasaan orang tentang objek seperti produk, layanan, organisasi, individu, topik, peristiwa, topik tertentu [8]. Setelah itu akan dilakukan evaluasi terhadap opini tersebut, yaitu *positive*, *neutral* dan *negative*.

2.2.3 *Preprocessing*

Text Preprocessing merupakan proses pengolahan teks yang bertujuan untuk mengurangi *noise* pada dataset serta mengubah dataset menjadi bentuk yang lebih terstuktur[9]. *Preprocessing* merupakan salah satu langkah penting dalam analisis sentimen. Maka dari itu perlu proses *Preprocessing* untuk menseleksi data yang berguna untuk mengoptimalkan data agar dapat diproses dan mendapatkan hasil

yang lebih baik dalam meningkatkan kinerja klasifikasi [10]. Pemrosesan data mencakup 6 tahapan sebagai berikut :

1. *Cleansing*

   Cleansing membersihkan data *tweet* yang bertujuan untuk menghapus simbol, *username*, angka, kata 'RT', hashtag (#), *Uniform Resource Locator (URL)*, emoji, dan ruang kosong atau *white space*.

2. *Case folding*

   Case Folding adalah proses merubah setiap katakter huruf pada seluruh data *tweet* menjadi huruf kecil atau non-kapital. Hanya huruf "a" sampai "z" saja yang diterima, selain itu kata akan hilang.

3. *Tokenizing*

   Tokenizing adalah Proses pemecahan sebuah *string* data menjadi token. Token adalah memisahkan kalimat yang ada pada dataset menjadi sebuah kata. Proses ini memanfaatkan fungsi dari pustaka *Natural Language Toolkit (NLTK).* Proses tokenisasi bisa dilakukan berdasarkan adanya spasi di sebuah kalimat, bisa juga dilakukan berdasarkan parameter tertentu [11].

4. *Stopword removal*

   Stopword removal adalah proses menghilangkan kata yang tidak merepresentasikan data. pada proses ini kata yang tidak memiliki makna penting untuk melakukan klasifikasi akan dihilangkan [12].

Selain melakukan pembersihan data maka dilakukan *balancing dataset*. Salah satu masalah umum yang ditemukan dalam kumpulan data untuk klasifikasi adalah persebaran data yang tidak seimbang. Persebaran data yang tidak seimbang dapat menyebabkan kurang tepatnya model yang dibuat pada saat *training* data serta algoritma klasifikasi memiliki kinerja yang buruk [13]. Untuk menangani masalah ketidakseimbangan data dapat dilakukan dengan teknik resampling seperti *oversampling* dan *undersampling* [14] seperti yang disajikan pada Gambar 2.1.

1. *Oversampling*

   Oversampling adalah teknik pengambilan sampel yang menyeimbangkan kumpulan data dengan mereplikasi kelas minoritas. Keuntungan dari metode ini adalah tidak ada kehilangan data sedangkan kerugian dari

teknik ini dapat menyebabkan pemasangan yang berlebihan dan dapat menyebabkan *overhead* komputasi tambahan.

2. *Undersampling*

   Metode *undersampling* dilakukan menggunakan subset dari kelas mayoritas untuk melatih classifier dengan menghapus kelas mayoritas.

Gambar 2.1 *Ilustrasi (a) Oversampling dan (b) Undersampling*

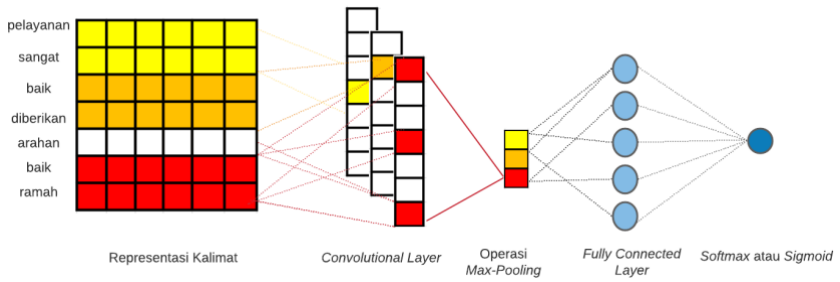
2.2.4 *Deep Learning*

Deep Learning adalah sebuah penerapan jaringan syaraf tiruan yang meniru cara kerja dari *kortex* manusia yang memiliki banyak layer tersembunyi (*hidden layer*) dan termasuk kedalam kajian dari *Machine Learning* di dalam bidang kecerdasan buatan. Dalam implementasinya pada permasalahan dataset yang besar *Deep Learning* memberikan ketepatan pada berbagai penelitian seperti deteksi suatu objek, pengenalan suara, terjemahan bahasa, dan lain-lain. Berbeda dengan teknik pada *machine learning* yang masih tradisional harus mengenali masukan terlebih dahulu [15], *Deep Learning* mampu menganalisa jutaan kemungkinan berdasarkan data latih sebelumnya dan dilakukan dalam waktu singkat. *Deep Learning* di klaim mampu beradaptasi dengan data dalam jumlah besar serta mampu menyelesaikan masalah yang sulit diselesaikan oleh *machine learning* lainnya. Sistem *Deep Learning* juga dapat mempelajari dari fungsi pemetaan yang kompleks dari mulai *input* hingga *output* tanpa konsep dari buatan manusia. *Deep Learning* memiliki beberapa jenis algoritma diantaranya *Convolutional Neural Network*, *Recurrent Neural Network*, *Long Short Term Memory*, dan *Self Organizing Map*.

2.2.5 *Convolutional Neural Network*

Convolutional Neural Network adalah salah satu metode algoritma *deep learning*. CNN juga didefinisikan sebagai algoritma yang biasa digunakan untuk pemproses

data gambar dan teks [7]. Konvolusi didefinisikan sebagai matriks yang berfungsi melakukan klasifikasi dan filter untuk gambar dan teks. Tujuan utama dari konvolusi adalah untuk mengekstrak fitur input, dan *pooling* adalah untuk mengambil sampel matriks konvolusi [8]. *Convolutional Neural Network* memiliki beberapa layer yang digunakan untuk melakukan filter dalam setiap proses. Proses ini dikenal sebagai proses *training*. Pada proses *training* terdapat 3 tahapan yaitu *Convolutional layer, Pooling layer, dan Fully connected layer* [7]. Arsitektur proses *training Convolutional Neural Netwok* dapat di lihat pada Gambar 2.2.



Gambar 2.2 Arsitektur Convolutional Neural Network

Convolutional layer berisi serangkaian filter yang ukurannya tetap yang digunakan untuk mengkonvolusikan data. Output dari convolutional layer adalah feature maps. Berikut ini adalah persamaan operasi convolutional:

$$FM_{a,b}=bias+\sum_{c}^{C}\sum_{d}^{D} Z_{c,d}+X_{a+c-1,b+d-1} \quad \text{..... (2.1)}$$

Pooling layer memastikan bahwa jaringan hanya fokus pada pola yang paling penting serta data dirangkum dengan menggeser jendela melintasi feature maps, kemudian menerapkan beberapa operasi linear atau non linear pada data yang ada pada jendela. Pooling layer memiliki fungsi untuk mengurangi dimensi dari feature maps yang akan digunakan pada layer selanjutnya

$$f_h\left(0,FM_{a,b}\right)=\max\left(0,FM_{a,b}\right)=\begin{cases} FM_{a,b}, & \text{jika } FM_{a,b} \geq 0, \\ 0 & \text{jika } FM_{a,b} < 0, \end{cases} \quad \text{..... (2.2)}$$

Layer terakhir yang digunakan adalah fully-connected layer. layer ini digunakan untuk memahami pola yang dihasilkan dari layer sebelumnya. Neuron pada layer ini memiliki koneksi penuh ke semua aktivasi pada layer sebelumnya. Metode CNN juga menggunakan fungsi aktivasi yang dilakukan

ketika berada di antara convolutional layer dan pooling layer. Aktivasi di antara kedua layer tersebut menggunakan fungsi aktivasi ReLU. Sedangkan untuk fungsi aktivasi output menggunakan softmax. Persamaan fungsi aktivasi ReLU terdapat pada Persamaan 2.2.

Fungsi aktivasi softmax mempunyai tujuan untuk mendapatkan hasil klasifikasi serta menghasilkan nilai yang diinterpretasi sebagai probabilitas yang belum dinormalisasi untuk tiap kelas. Nilai kelas yang dihitung dengan menggunakan fungsi softmax ditunjukan oleh Persamaan 2.3.

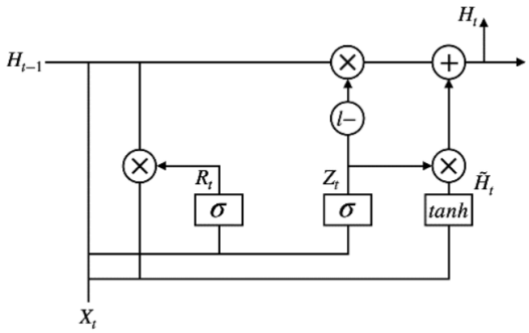
$$y_{ijk} = \frac{e^{x_{ijk}}}{\sum_{t=1}^{D} e^{x_{ijt}}} \quad \text{..... (2.3)}$$

Fungsi terakhir adalah loss function untuk menghitung loss (nilai error) dengan menggunakan categorical cross-entropy. Persamaan 2.4 adalah loss function yang dimaksud.

$$L_{log}(Y, Y_{pred}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1_{y_i \in C_c} \log p_{model}[y_i \in C_c] \quad \text{..... (2.4)}$$

2.2.6 *Gated Recurrent Unit*

GRU adalah salah satu mekanisme dari RNN yang mirip dengan LSTM [1].GRU pertama kali diusulkan Gers, dkk. pada tahun 2014 yang merupakan model sederhana dari LSTM. Ada dua gate pada GRU yaitu forget gate dan input gate yang kemudian diteruskan ke update gate. Dari update gate informasi diteruskan secara selektif ke hidden layer untuk mengurangi masalah gradient saat mengingat informasi[2]. Karena kinerjanya mirip LSTM, GRU cocok digunakan pada penelitian ini dengan karakteristik yang sederhana, parameter yang sedikit, kemampuan menangani overfitting yang lebih baik, dan kecepatan konvergensi yang lebih cepat. Untuk lebih jelasnya pada gambar 2.4. Sedangkan Bi-GRU sendiri GRU yang bekerja dari dua arah[3].



Gambar GRU architecture

Reset gate:

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_t) \quad \text{..... (2.5)}$$

Candidate activation vector

$$\tilde{h}_t = tanh(W_h + [r_t * h_{t-1}, x_t] + b_h) \quad \text{..... (2.6)}$$

Update gate

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \quad \text{..... (2.7)}$$

Candidate activation vector

$$\tilde{h}_t = tanh(W_h + [z_t * h_{t-1}, x_t] + b_h) \quad \text{..... (2.8)}$$

Hasil hidden gate

$$h_t = (1 - z_t) * h_t + z_t * \tilde{h}_t \quad \text{..... (2.9)}$$

Dimana:

x_t : *input vector*

h_t : *output vector*

$\tilde{h}_t$: *candidate activation vector*

z_t : *update gate vector*

r_t : *reset gate vector*

W, b: *parameter matrices and vector*

Dalam jaringan GRU (Gated Recurrent Unit), terdapat pintu-pintu yang mengontrol aliran informasi di dalam unit GRU. Salah satunya adalah pintu pembaruan z_t, yang mengontrol nilai pembaruan aktivasi. Di sini, W_z dan U_z adalah matriks bobot yang harus dipelajari. Pintu pembaruan ini memengaruhi sejauh mana informasi baru akan dimasukkan ke dalam unit GRU. Selain pintu

pembaruan, terdapat juga aktivasi kandidat c_t. Pintu istirahat r_t memungkinkan unit GRU untuk melupakan keadaan sebelumnya dengan membaca simbol pertama dari suatu urutan masukan.

2.2.7 *Word2Vec*

Word2Vec adalah salah satu metode *embedding word* yang berguna untuk merepresentasikan kata menjadi sebuah *vector* [16]. Word2Vec dapat memiliki 50 sampai dengan 300 dimensi. Word2Vec mulai ramai digunakan dalam bidang *natural language processing* di tahun 2013, karena Word2Vec merupakan *dense vectors* yang dapat merepresentasikan hubungan antar kata dengan lebih baik (dibandingkan dengan TF-IDF), secara semantik maupun sintaksis [16]. Word2Vec memiliki dua model arsitektur yaitu *Skip-Gram* dan *Continous Bag of Words (CBOW)* [17]. Kedua metode ini menggunakan konsep jaringan saraf tiruan yang memetakan kata ke variabel target yang merupakan sebuah kata. Tujuan dalam arsitektur *skip-gram* adalah untuk memprediksi kata yang ada di sekitar *current word*. Sedangkan arsitektur CBOW digunakan untuk memprediksi kata yang ada pada sekitar kata tersebut.

2.2.8 *Confussion Matrix*

Confusion matrix adalah ringkasan hasil prediksi pada masalah klasifikasi. Jumlah prediksi yang benar dan salah dirangkum dengan nilai hitungan hasil akurasi pada konsep data mining dan dipecah oleh masing-masing kelas. Terdapat empat istilah dari hasil klasifikasi dalam I, antara lain :

1. TP (*True Positive*) merupakan data yang bersifat positif dan terdeteksi benar.
2. TN (*True Negative)* merupakan data yang bersifat negatif dan terdeteksi benar.
3. FP (*False Positive*) merupakan data yang bersifat negatif namun terdeteksi sebagai data positif.
4. FN (*False Negative*) merupakan data yang bersifat positif namun terdeteksi sebagai data negatif.

Confusion matrix juga berguna untuk menilai bagaimana kinerja suatu model dibangun. Hasil klasifikasi tidak dapat dilihat hanya dengan satu angka, sehingga keempat istilah TP, FP, TN, dan FN sama pentingnya dalam memberikan

informasi dari temuan. Secara umum, perhitungan yang biasa digunakan dalam c*onfusion matrix* meliputi *precision, recall, F1-score* dan *accurary.*

Precision adalah perhitungan untuk menghasilkan tingkat akurasi antara data yang diminta dengan hasil prediksi sistem. oleh karena itu, *precision* membandingkan prediksi data benar positif dengan hasil prediksi positif keseluruhan. Dapat dilihat pada persamaan berikut

$$Precision = \frac{\mathrm{TP}}{\mathrm{FP+TP}}$$

Recall merupakan tingkat dari keberhasilan suatu sistem dalam menemukan sebuah informasi kembali yang dapat dilihat dari persamaan berikut :

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

F1-score adalah perbandingan berbobot dari rata-rata presisi dan *recall*. *F1- score* dihitung sebagai berikut :

$$\mathrm{F1} = 2X \frac{precission\ x\ recall}{precission+recall}$$

Accuracy adalah rasio prediksi yang benar (positif dan negatif) terhadap keseluruhan data. *Accuracy* dapat dihitung sebagai berikut

$$Accuracy = 100\ X\ \frac{\mathrm{Total\ Klasifikasi\ Benar}}{\mathrm{Total\ Klasifikasi}}$$