BAB II
LANDASAN TEORI

2.1 Tinjauan Pustaka

Tinjauan pustaka pada penelitian-penelitian sebelumnya dalam mendukung penelitian yang sedang dilakukan. Berikut ini adalah penelitian sebelumnya terkait dengan penelitian yang akan di lakukan oleh penulis:

Tabel 2. 1. Tinjauan Pustaka

No	Penulis	Judul	Masalah Penelitian	Variabel Terkait	Hasil
1	Muhammad David Hilmawan (2022)	Deteksi Sarkasme Pada Judul Berita Berbahasa Inggris Menggunakan Algoritme Bidirectional LSTM.	Membuat model klasifikasi dengan algoritme *Bidirectional Long Short-Term Memory* (BiLSTM) untuk mendeteksi gaya Bahasa sarkasme pada judul berita berbahasa inggris dikarenakan judul berita menggunakan kata baku dan tidak ada salah pengejaan kata, menjadikan judul berita sebuah dataset yang tepat untuk di lakukan deteksi sarkasme.	*Bidirectional Long Short-Term Memory* (BiLSTM), *Long Short-Term Memory* (LSTM), *Glove Word Embeddings*, Deteksi Sarkasme.	Didapatkan akurasi validasi dari model sebesar 82,55%, *precision* validasi sebesar 82,36%, *recall* validasi sebesar 79,53% dan *f1 score* validasi sebesar 80,92%
2	Rizka Dwi Wulandari Santosa dkk (2021)	Implementasi Algoritma Long Short-Term Memory (LSTM) Untuk Mendeteksi Penggunaan	Arsitektur lstm diimplementasikan untuk mendeteksi penggunaan kalimat abusive pada teks Bahasa Indonesia.	*Bidirectional Long Short-Term Memory* (BiLSTM), *Long Short-Term Memory* (LSTM)	Pada hasil pengujian LSTM menghasilkan akurasi F1 Score 0.8181%, sedangkan BiLSTM

No	Penulis	Judul	Masalah Penelitian	Variabel Terkait	Hasil
		Kalimat *Abusive* Pada Teks Bahasa Indonesia.			menghasilkan akurasi F1 Score 0.8423%.
3	Aini dan Aristiawan (2019)	Implementasi Algoritma Long Short-Term Memory (LSTM) Untuk Mendeteksi Ujaran Kebencian (Hate Speech) Pada Kasus Pilpres 2019.	Melakukan penelitian tentang mendeteksi ujaran kebencian atau *Hate Speeech* pada kasus Pemilihan Presiden (Pilpres) 2019	Long Short-Term Memory (LSTM)	Hasil uji coba terbaik dengan 190 kalimat memiliki nilai *recall* 0.7021, yang artinya dari keseluruhan data uji ujaran kebencian, 70.21% benar dideteksi sebagai ujaran kebencian, sedangkan sisanya 29.79% salah dideteksi sebagai bukan ujaran kebencian.
4	Dedi Tri Hermanto dkk (2021)	Algoritma LSTM-CNN untuk sentimen klasifikasi dengan Word2vec pada media Online	Topik ekonomi adalah bahasan yang menarik untuk dilakukan penelitian karena memiliki dampak langsung kepada masyarakat Indonesia, Penelitian ini bertujuan untuk melakukan pengklasifikasian judul berita berbahasa Indonesia berdasarkan sentimen positif, negatif dengan menggunakan	Long Short-Term Memory (LSTM)	Berdasarkan hasil pengujian memperlihatkan bahwa metode LSTM, LSTM-CNN, CNN-LSTM memiliki hasil akurasi sebesar, 62%, 65% dan 74%.

No	Penulis	Judul	Masalah Penelitian	Variabel Terkait	Hasil
			metode LSTM, LSTM-CNN, CNN-LSTM		
5	Muh dan Azhari (2019)	Sentiment Analysis of Novel review Using Long Short-Term Memory Method	melakukan pengklasifikasian terhadap review novel berbahasa Indonesia berdasarkan sentimen positif, netral dan negatif dengan menggunakan metode Long Short-Term Memory (LSTM)	Long Short-Term Memory (LSTM)	Berdasarkan hasil pengujian memperlihatkan bahwa metode Long Short-Term Memory memiliki hasil akurasi yang lebih baik dibandingkan dengan metode naïve bayes dengan nilai akurasi 72.85%, precision 73%, recall 72%, dan f-measure 72%
6	Auliya Rahman Isnain dkk (2019)	Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection	mendeteksi ujaran kebencian atau bukan ujaran kebencian tweet berbahasa Indonesia dengan mengunakan metode Bidirectional Long Short Term Memory dan metode ekstrasi fitur word2vec dengan arsitektur Continuous bag-of-word (CBOW)	LSTM, BiLSTM, Word2vec.	Penggunaan word2vec dan metode Bidirectional Long Short Term Memory menghasilkan tingkat akurasi 94,66%, dengan masing-masing nilai presisi 99,08%, recall 93,74% dan F-measure 96,29%

Muhammad David Hilmawan (2022) Pada penelitian ini akan dibuat model klasifikasi untuk memprediksi sarkasme pada judul berita berbahasa inggris

dikarenakan judul berita menggunakan kata baku dan tidak ada salah pengejaan kata, menjadikan judul berita sebuah *dataset* yang tepat untuk dilakukan deteksi sarkasme, Algoritme *Bidirectional Long Short-Term Memory* (BiLSTM) yang merupakan salah satu algoritme *deep learning* digunakan pada penelitian untuk membuat model klasifikasi. Model ini lalu dibandingkan dengan model algoritme *Long Short-Term Memory* (LSTM) untuk memvalidasi keunggulan dari algoritme BiLSTM daripada algoritme LSTM dasar. Didapatkan akurasi validasi dari model sebesar 82,55%, *precision* validasi sebesar 82,36%, *recall* validasi sebesar 79,53%, dan *f1 score* validasi sebesar 80,92%.

Rizka Dwi Wulandari Santosa (2021) Arsitektur LSTM diimplementasikan untuk mendeteksi penggunaan kalimat *abusive* pada teks bahasa indonesia. *Dataset* yang digunakan pada penelitian mengalami ketidakseimbangan jumlah data pada setiap kelas sehingga dilakukan penambahan data untuk mengetahui pengaruh penambahan jumlah data terhadap hasil performansi arsitektur. Tahapan pengerjaan dalam penelitian ini dimulai dari pembangunan *dataset*, pra-pemrosesan data, pembuatan model pendeteksi kalimat *abusive,* pelatihan dan pengujian. Pengujian dilakukan terhadap arsitektur LSTM dan didapatkan hasil bahwa arsitektur ini hanya dapat memprediksi terhadap kelas mayoritas sehingga dilakukan penambahan penggunaan arsitektur yaitu *Bidirectional LSTM* (BiLSTM). Hasil uji coba menunjukkan BiLSTM lebih baik dalam mengklasifikasikan kalimat karena terdapat *forward* dan *backward layer* yang membuat proses pembelajaran model lebih kompleks dalam mengenal konteks kalimat dan hal ini akan meningkatkan keakuratan hasil klasifikasi pada setiap label. Pada LSTM hanya menghasilkan nilai

F1 Score untuk kelas mayoritas saja sebesar 0.812 sedangkan pada BiLSTM sudah dapat menghasilkan nilai F1 Score untuk semua kelas.

Aini dan Aristiawan (2019) arsitektur dari RNN yang biasa digunakan pada masalah *deep learning* yaitu *Long Short-Term Memory* (LSTM) akan diimplementasikan untuk mendeteksi ujaran kebencian (*hate speech*) berkaitan dengan Pemilihan Presiden (Pilpres) 2019. Tahapan pengerjaan dalam penelitian ini dimulai dari studi kepustakaan, pengumpulan data, pra-pemrosesan data, pembuatan model *word2vec*, perancangan model pendeteksi ujaran kebencian, pelatihan model, dan pengujian model. Model ini dilatih dengan menggunakan 950 kalimat dan diuji dengan 190 kalimat dari *dataset* yang bersumber dari media sosial Facebook. Hasil uji coba terbaik dengan 190 kalimat memiliki nilai *recall* 0.7021, yang artinya dari keseluruhan data uji ujaran kebencian, 70.21% benar dideteksi sebagai ujaran kebencian, sedangkan sisanya 29.79% salah dideteksi sebagai bukan ujaran kebencian.

Dedi Tri Hermanto dkk (2021) Media online banyak menghasilkan berbagai macam berita, baik ekonomi, politik, kesehatan, olahraga atau ilmu pengetahuan. Ekonomi memiliki dampak langsung kepada warga negara, perusahaan, bahkan pasar tradisional tergantung pada kondisi ekonomi di suatu negara. Topik ekonomi adalah bahasan yang menarik untuk dilakukan penelitian karena memiliki dampak langsung kepada masyarakat Indonesia. Namun, masih sedikit penelitian yang menerapkan metode *deep learning* yaitu *Long Short-Term Memory* dan *CNN* untuk analisis sentimen pada artikel *finance* di Indonesia. Penelitian ini bertujuan untuk melakukan pengklasifikasian judul berita berbahasa Indonesia berdasarkan sentimen positif, negatif dengan menggunakan metode LSTM, LSTM-CNN, CNN-

LSTM. *Dataset* yang digunakan adalah data judul artikel berbahasa Indonesia yang diambil dari situs Detik *Finance*. Berdasarkan hasil pengujian memperlihatkan bahwa metode LSTM, LSTM-CNN, CNN-LSTM memiliki hasil akurasi sebesar, 62%, 65% dan 74%.

Muh dan Azhari (2019) Berkembang pesatnya internet dan media sosial serta besarnya jumlah data teks, telah menjadi subjek penelitian yang penting dalam memperoleh informasi dari data teks tersebut. Penelitian ini bertujuan untuk melakukan pengklasifikasian terhadap *review* novel berbahasa Indonesia berdasarkan sentimen positif, netral dan negatif dengan menggunakan metode *Long Short-Term Memory* (LSTM). Dataset yang digunakan adalah data *review* novel berbahasa Indonesia yang diambil dari situs *goodreads.com*. Dalam proses pengujian, metode LSTM akan dibandingkan dengan metode *Naïve Bayes* berdasarkan perhitungan dari nilai akurasi, *precision, recall, f-measure*. Berdasarkan hasil pengujian memperlihatkan bahwa metode Long Short-Term Memory memiliki hasil akurasi yang lebih baik dibandingkan dengan metode *naïve bayes* dengan nilai akurasi 72.85%, *precision* 73%, *recall* 72%, dan *f-measure* 72% dibandingkan dengan hasil akurasi metode *Naïve Bayes* dengan nilai akurasi 67.88%, *precision* 69%, *recall* 68%, dan *fmeasure* 68%.

Auliya Rahman Isnain dkk (2019) Ujaran kebencian merupakan komunikasi yang meremehkan seseorang atau kelompok berdasarkan karakteristik seperti (ras, etnis, jenis kelamin, kewarganegaraan, agama dan oragnisasi). *Twitter* salah satu media sosial yang digunakan seseorang untuk mengutarakan perasaan dan opini melalui *tweet*, termasuk *tweet* yang megandung ujaran kebencian. Penelitian ini bertujuan untuk mendeteksi ujaran kebencian atau bukan ujaran kebencian *tweet*

berbahasa Indonesia dengan mengunakan metode *Bidirectional Long Short Term Memory* dan metode ekstrasi fitur *word2vec* dengan arsitektur *Continuous bag-of-word* (CBOW). Untuk pengujian metode BiLSTM dengan perhitungan nilai akurasi, *presisi, recall, dan F-measure*. Penggunaan *word2vec* dan metode *Bidirectional Long Short Term Memory* dengan arsitektur CBOW, dengan epoch 10, learning rate 0.001 dan jumlah neuron 200 pada layer tersembunyi, menghasilkan tingkat akurasi 94,66%, dengan masing-masing nilai presisi 99,08%, *recall* 93,74% dan *F-measure* 96,29%. Sedangkan untuk *Bidirectional Long Short Term Memory* dengan tiga layer memiliki akurasi 96,93%. Penambahan satu layer pada BiLSTM meningkat 2,27%.

2.2 Twitter

Twitter merupakan salah satu media sosial yang hingga saat ini masih sangat digunakan oleh masyarakat Indonesia untuk berbagi informasi atau bahkan untuk memberikan sebuah opini terhadap suatu trending topik (Isnain et al., 2021).

Twitter menyediakan kemudahan kepada pelanggan dalam mengakses data *twitter* melalui *API Twitter*. Dalam pengambilan data, *twitter* membatasi jumlah data *tweet* yang yang akan diambil dalam satu jam. Jenis *API Twitter* yang disediakan oleh *twitter* untuk membantu pengambilan data *tweet* yaitu: *REST API*: *API* yang digunakan untuk memberikan akses dalam menulis dan membaca *twitter,* seperti *post tweet,* mengambil data *tweet* pada sebuah lokasi tertentu, pada *user* tertentu atau topik tertentu. Pada *REST API* membutuhkan otentikasi menggunakan *OAuth* untuk menjalankan *API.*

Streaming API: API yang digunakan untuk mendapatkan data *realtime* dari data *tweet* dengan kata kunci tertentu.

2.3 Preprocessing

Text Preprocessing merupakan suatu proses untuk menyeleksi data teks agar menjadi lebih terstruktur lagi dengan melalui serangkaian tahapan yang meliputi tahapan *case folding, stopword removal, punctuation removal, dan tokenizing*.

2.3.1 Case Folding

Case Folding adalah proses mengubah semua huruf kapital pada tweet menjadi tidak kapital, serta menghilangkan karakter selain huruf a-z dan data numberik. *Case Folding* dilakukan untuk menyamakan semua kata karena kata yang sama tetapi penggunaan huruf kapital yang berbeda dapat mempengaruhi hasil dari penelitian (Alita & Isnain, 2020).

2.3.2 Stopword Removal

Stopword di perkenalkan pertama kali oleh H.P. Luhn pada tahun 1958. *Stopword* adalah kumpulan kata yang sering muncul dalam sebuah dokumen dan dapat sedikit informasi yang biasanya tidak dibutuhkan. Menghapus *stopword* tidak hanya mengurangi kompleksitas komputasi, dapat juga meningkatkan kualitas hasil yang didapatkan. Contoh dari kata-kata dalam *stoplist* Bahasa Indonesia yaitu: “dan”, “seperti”, “jika”, “andai”, “jikalau”, dll (Alita & Isnain, 2020).

2.3.3 Tokenizing and Punctuation Removal

Tokenizing adalah proses pemisahan setiap kata pada suatu dokumen dimana potongan-potongan tersebut dinamakan dengan *token*. Umum nya setiap kata terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses tokenisasi mengandalkan karakter spasi pada dokumen untuk melakukan pemisahan kata. Setelah melalui proses tokenisasi maka kalimat tersebut menjadi

sekumpulan *array* yang setiap selnya berisi kata-kata yang ada pada kalimat tersebut (Alita & Isnain, 2020). *Punctuation removal* adalah tanda baca pada dokumen tidak memiliki arti yang bermakna pada suatu kalimat. Maka lebih baik tanda baca tersebut untuk dihilangkan di dalam dokumen seperti tanda #, /, ?, dll (Alita et al., 2020).

2.4 Deep Learning

Deep Learning atau dikenal juga dengan *Deep Structured Learning* merupakan bagian dari pembelajaran *Machine Learning* yang menggunakan *Artificial Neural Network* (ANN). Singkatnya, *deep learning* adalah metode pembelajaran oleh mesin dengan meniru cara kerja sistem saraf otak manusia. Algoritma *deep learning* salah satunya *yaitu Long Short-Term Memory* (LSTM) (Dwi et al., 2021). jenis *reccurent neural network* (RNN) yang dapat mempelajari dan menghafal ketergantungan pola jangka panjang, teknologi ini mampu mengingat seluruh informasi dari waktu ke waktu.

Recurrent neural network (RNN) memiliki koneksi yang bisa membentuk siklus terarah, siklus tersebut yang memungkinkan *output* dari LSTM untuk diumpankan sebagai input ke *fase* terbaru. Setelah *output* dari LSTM menjadi *input* terbaru, ia dapat mengingat *input* sebelumnya karena kinerja *memory internal*. RNN biasanya digunakan untuk teks gambar, analisis deret waktu, *natural language preprocessing*, pengenalan tulisan tangan dan mesin translasi.

2.5 Sarkasme

Dalam Kamus Besar Bahasa Indonesia, sarkasme adalah penggunaan kata-kata pedas untuk menyakiti hati orang lain, berupa cemoohan atau ejekan kasar (Hilmawan., 2022). Kata sarkasme diturunkan dari kata Yunani *sarkasmos* yang

berarti “merobek-robek daging seperti anjing”, “menggigit bibir karena marah”, atau “berbicara dengan kepahitan”. Sarkasme merupakan suatu acuan yang lebih kasar dari ironi dan sinisme dan mengandung kepahitan serta celaan yang getir. Sarkasme dapat bersifat irons, atau tidak, tetapi yang pasti adalah bahwa gaya Bahasa ini selalu akan menyakiti hati dan kurang enak didengar. Gaya Bahasa sarkasme dapat diciri-cirikan sebagai berikut:

1. Maknanya mengandung olok-olok, ejekan, atau sindiran
2. Mengatakan makna yang bertentangan
3. Mengandung kepahitan dan kurang enak didengar
4. Lebih kasar dibandingkan dengan gaya bahasa ironis dan gaya bahasa sinisme.

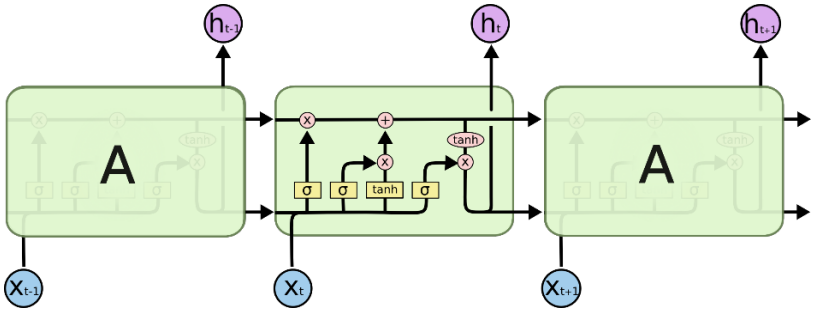
2.6 Word Embeddings

Word embeddings merupakan salah satu metode *deep learning* yang sangat berguna untuk menyusun *representasi* kata pada suatu dokumen menjadi *vektor*. Metode ini dapat menangkap hubungan sintaksis dan semantik antar kata yang ada pada suatu dokumen (Ibrahim et al., 2020). Pada penelitian ini digunakan *GloVe Word Embeddings. GloVe Word Embeddings* dibuat berdasarkan *log bilinear* model dan menggabungkan keuntungan dari metode *local window* dan faktorisasi *matriks global.*

2.7 *Long Short-Term Memory* (LSTM)

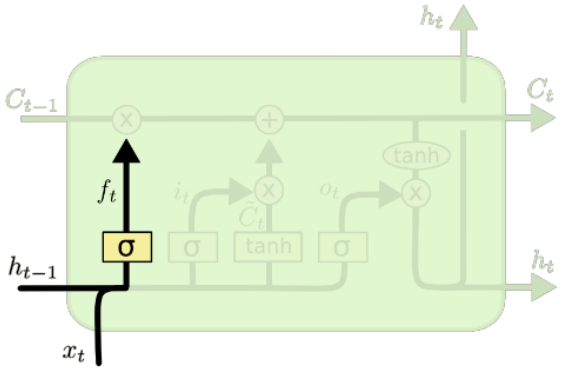
Long Short-Term Memory (LSTM) merupakan evolusi dari *Recurrent Neural Network* (RNN) untuk mengatasi kekurangan dari RNN dengan cara menambahkan interaksi tambahan pada setiap modul. LSTM dapat mempelajari dependensi *long-term* dan mengingat informasi dalam jangka waktu yang cukup lama. Pada tahap

ini dilakukan proses klasifikasi teks berbahsa Indonesia dengan menggunakan salah satu arsitektur RNN yaitu LSTM menggunakan tiga gerbang yaitu gerbang *input*, gerbang *forget*, dan gerbang *output*. Gerbang *input* dan gerbang *output* berfungsi untuk mengatur alur data yang masuk dan keluar dalam *network*, sementara gerbang *forget* akan menghapus atau mengabaikan informasi yang berbobot rendah. Berikut merupakan penjelasan mengenai arsitektur LSTM (Dwi et al., 2021).



Gambar 2. 1 Arsitektur LSTM

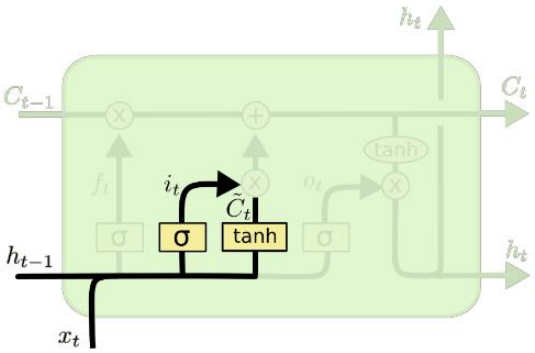
Gambar 2.1 merupakan arsitektur LSTM dimana di bagian bawah terdapat *cell gates* yang berfungsi untuk meregulasi informasi yang akan dikeluarkan ke *cell state* atau unit berikutnya. *Cell state* adalah jalur pada bagian atas untuk mengirimkan informasi ke unit kerja selanjutnya.



Gambar 2. 2 Forget Gate Layer

$$ft = \sigma\,(Wf\,[ht\text{-}1\ ,\ xt] + bf) \tag{1}$$

Pada persamaan (1) terdapat formula untuk menghitung keluaran dari lapisan *Forget Gate* dimana informasi akan dihapus dan informasi yang penting akan diteruskan ke *cell state.*

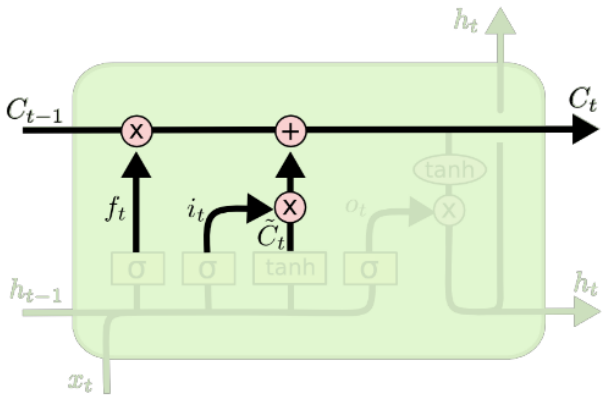


Gambar 2. 3 Input Gate Layer

$$it = \sigma.\,Wi\,[ht\text{-}1\ ,\ xt] + bi) \tag{2}$$

$$\hat{C}t = tanh\,(Wc\,.[ht\text{-}1\ ,\ xt] + bc) \tag{3}$$

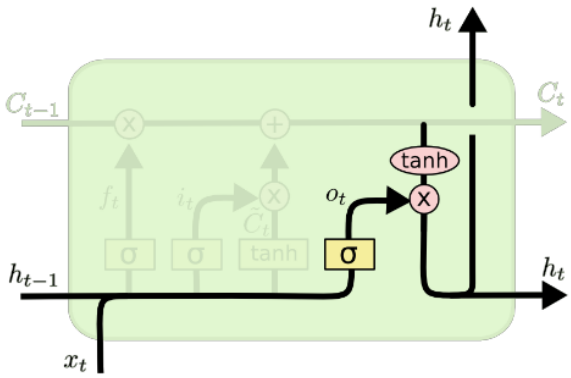
Pada persamaan (2) terdapat formula untuk memasukan nilai yaitu informasi yang akan diarahkan *cell state.*



Gambar 2. 4 Update Gate Layer

$$Ct = ft\ x\ Ct\text{-}1 + it\ x\ \hat{C}t \tag{4}$$

Pada persamaan (4) terdapat formula untuk mengubah nilai pada *cell state* yang didapat dari dua lapisan sebelumnya dari proses penghapusan dan penambahan informasi.



Gambar 2. 5 Output Gate Layer

$$Ot = \sigma\,(Wo\,.[ht\text{-}1\,,\,xt] + bo) \tag{5}$$

$$ht = Ot\ x\ tanh(Ct\,) \tag{6}$$

Pada persamaan (5) akan menghasilkan hasil yang sesuai untuk diarahkan ke *hidden unit* selanjutnya. Dalam perhitungan LSTM dapat dirumuskan sebagai berikut.

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

$\tilde{C}_t = \tanh(W_i\,.\,[h_{t\text{-}1}, x_t] + b_i)$

$C_t = f_t C_{t\text{-}1} + i_t \tilde{C}_t$

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

$O_t = \sigma(W_o\,.\,[h_{t\text{-}1}, x_t] + b_o)$

$h_t = O_t \tanh C_t$

dimana

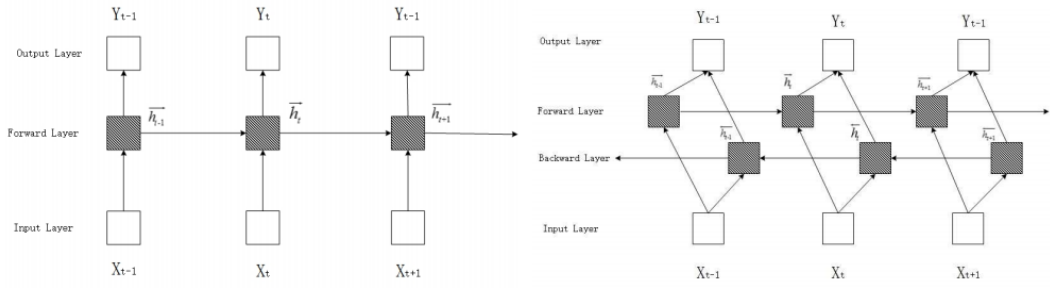
i_t = *input gate*

f_t = *forget gate*
O_t = *output gate*
σ = fungsi sigmoid

x_t = *input* saat ini
h_t = *output* baru
h_{t-1} = *output* sebelumnya
C_t = kondisi sel saat ini
C_t = kondisi sel baru
C_{t-1} = kondisi sel sebelumnya
W_i = bobot matrix *input gate*
b_i = bias *input gate* LSTM
W_f = bobot matrix *forget gate*
b_f = bias *forget gate* LSTM
W_o = bobot matrix *output gate*
b_o = bias *output gate* LSTM

2.8 *Bidirectional Long Short-Term Memory* (BiLSTM)

Algoritme Bidirectional Long Short-Term Memory (BiLSTM) yang merupakan salah satu *algoritme deep learning* digunakan pada penelitian untuk membuat model klasifikasi. Pada dasarnya BiLSTM memiliki struktur yang sama dengan LSTM dengan adanya penambahan *backward layer* dapat melakukan *training* data dua kali, tidak seperti LSTM biasa yang hanya melakukan *training* satu kali dari dataset. Dataset mentah dari proses pertama dimasukan lagi keproses kedua. Ilustrasi dari arsitektur LSTM dan BiLSTM sebagai berikut (Dwi et al., 2021).



Gambar 2. 6 Ilustrasi Arsitektur LSTM dan BiLSTM

(a) LSTM (b) BiLSTM

2.9 Python

Python adalah bahasa pemrograman yang bersifat *open source*. Bahasa pemrograman ini di optimalisasikan untuk *software quality, developer productivity, program portability,* dan *component integration*. *Python* telah digunakan untuk mengembangkan berbagai macam perangkat lunak, *sepeti internet scripting, system programming, user interfaces, product customization, numberic programming* dll. *Python* saat ini telah menduduki posisi 4 atau 5 bahasa pemrograman paling sering digunakan di seluruh dunia (Millman & Aivazis, 2011).

2.10 Bahan Pangan

Fluktuasi harga bahan pangan telah menjadi masalah yang rutin setiap tahun. Kenaikan harga bahan pangan merupakan faktor pemicu terjadinya hingar-bingar salah satu nya di media sosial terutama *twitter,* hal ini menimbulkan keresahan serta pertanyaan di masyarakat. Kenaikan harga bahan pangan ini tentu saja menjadi pukulan bagi masyaraka karena terpenuhinya kebutuhan akan bahan pangan merupakan salah satu hal yang penting dalam kesejahteraan hidup (Sanjaya & Heksaputra, 2020).