

Analisis Sentimen Berita Komoditas Emas menggunakan *Long Short Term Memory*

Presented by **Group 2**

Our Team



Aditya Agral S

2043201082



May Yulianti

2043201093



Lovinki Fitra Ananda

2043201112

Latar Belakang

Dataset yang digunakan berasal dari kumpulan data berita untuk pasar komoditas yang telah dianotasi secara manual oleh tiga orang ahli subjek, dengan lebih dari 10.000 judul berita yang tersebar dalam berbagai dimensi dan kelas. Data ini berasal dari berbagai sumber berita selama periode lebih dari 20 tahun (2000-2021).

Dataset ini dapat digunakan untuk mengembangkan model *machine learning* yang memahami berita serta dapat memberikan model peramalan harga jangka pendek dan jangka panjang. Selain itu, dataset ini dapat berguna bagi peneliti yang bekerja di bidang analisis berita untuk komoditas dan untuk mengevaluasi model pada masalah klasifikasi.

Sumber Data

Sentiment Analysis of Commodity News (Gold) -
<https://www.kaggle.com/>

Data yang digunakan berjumlah 8.126 yang terdiri dari kelas positif dan negatif

Metode

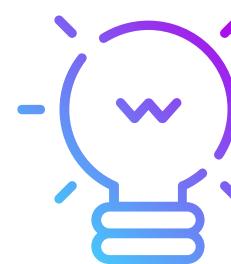
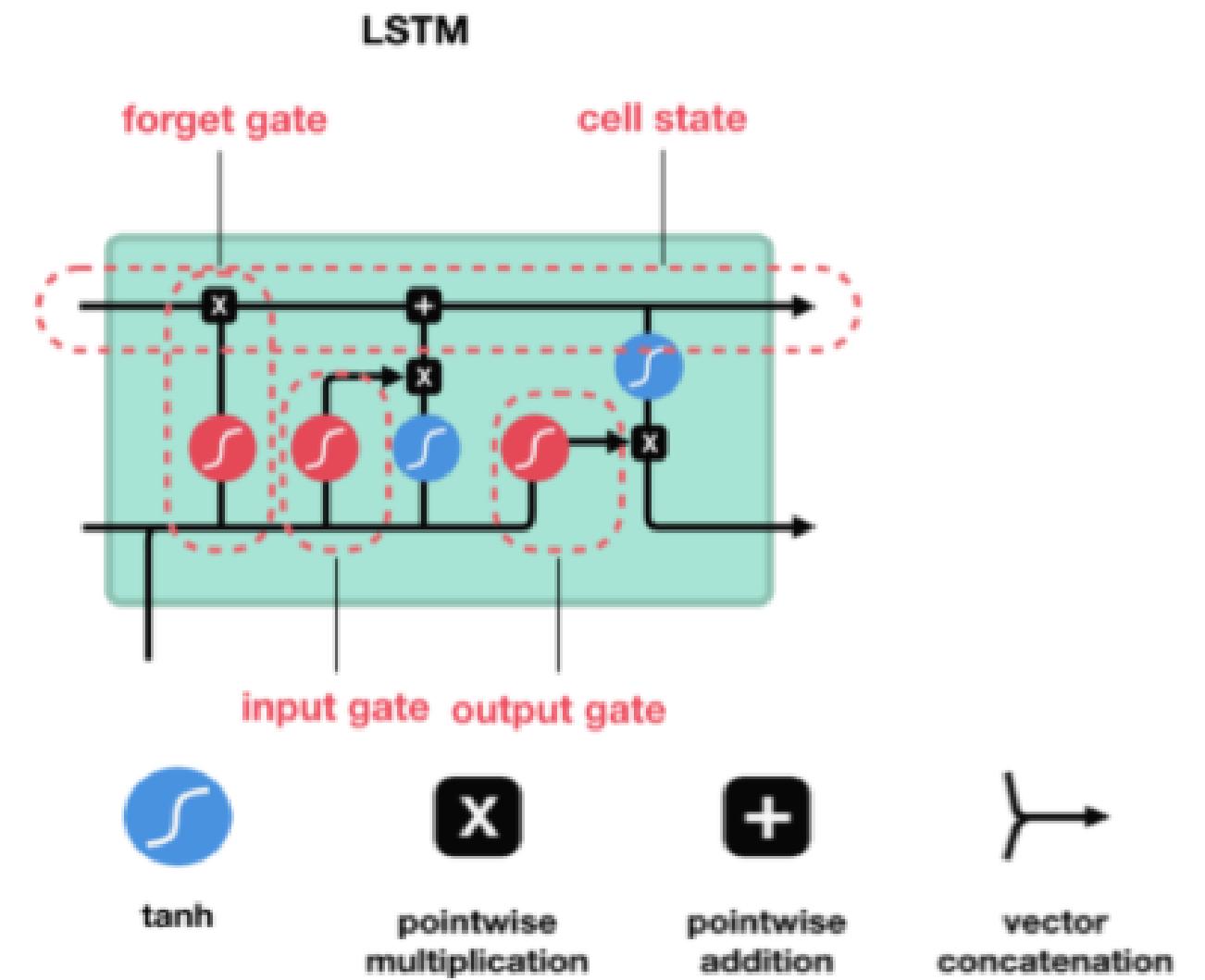
Analisis Sentimen

Sentiment Analysis adalah sebuah analisis yang biasanya digunakan untuk mendapatkan sebuah informasi sentimen dari kalimat yang terdapat dari teks tersebut.

LSTM

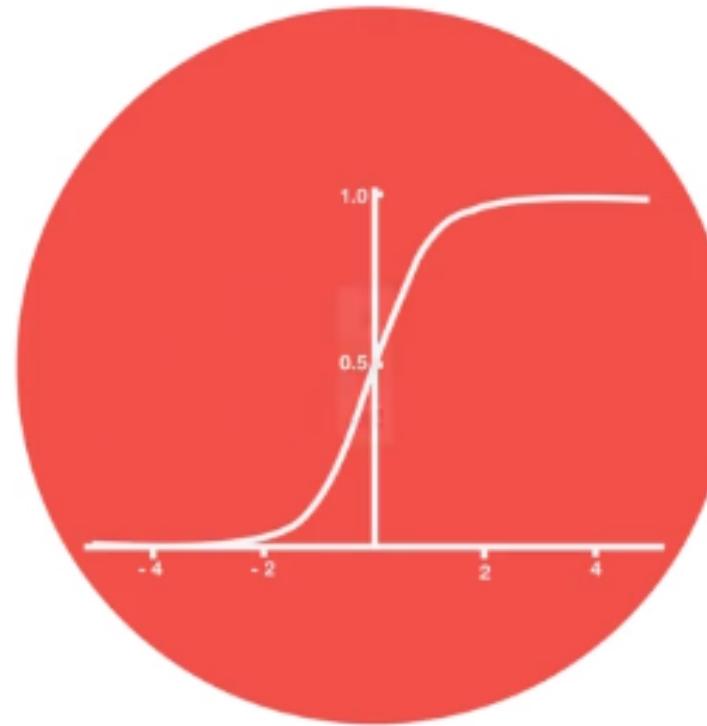
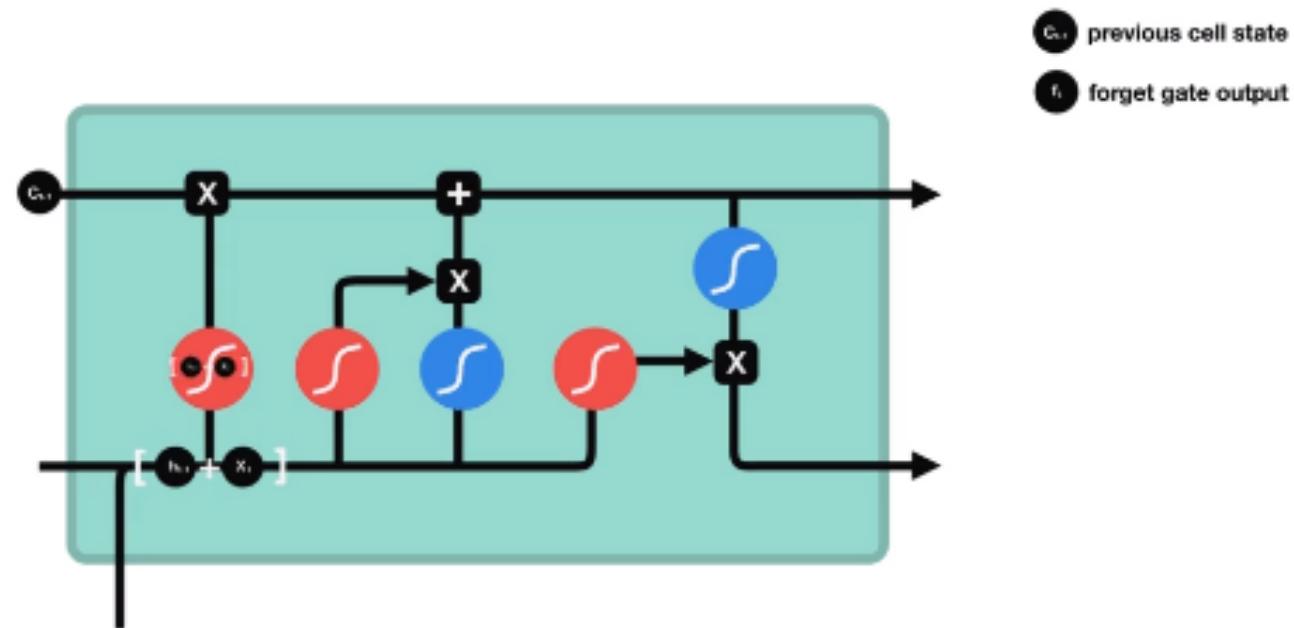
Long Short-Term Memory (LSTM) adalah jenis arsitektur jaringan saraf rekurensial (RNN) yang dirancang untuk mengatasi masalah vanishing gradient dalam pembelajaran jangka panjang. LSTM memiliki unit memori khusus yang memungkinkannya untuk menyimpan informasi dalam jangka waktu yang lama dan mempertahankan atau menghapus informasi ini sesuai kebutuhan. Dengan menggunakan gate seperti forget gate, input gate, dan output gate, LSTM dapat mengontrol aliran informasi dalam dan keluar dari unit memori, memungkinkan pembelajaran dari data urutan yang panjang dan kompleks.

Long Short Term Memory - Theory

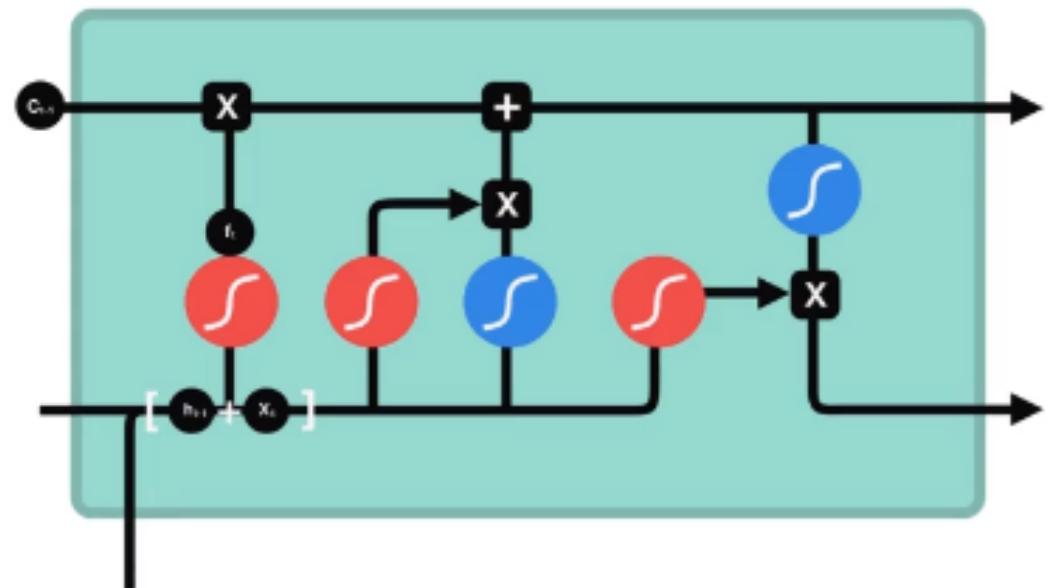


- Konsep inti LSTM adalah **cell state** dan berbagai gerbangnya.
Cell state bertindak sebagai jalan raya transportasi yang mentransfer informasi relatif sepanjang rantai urutan.

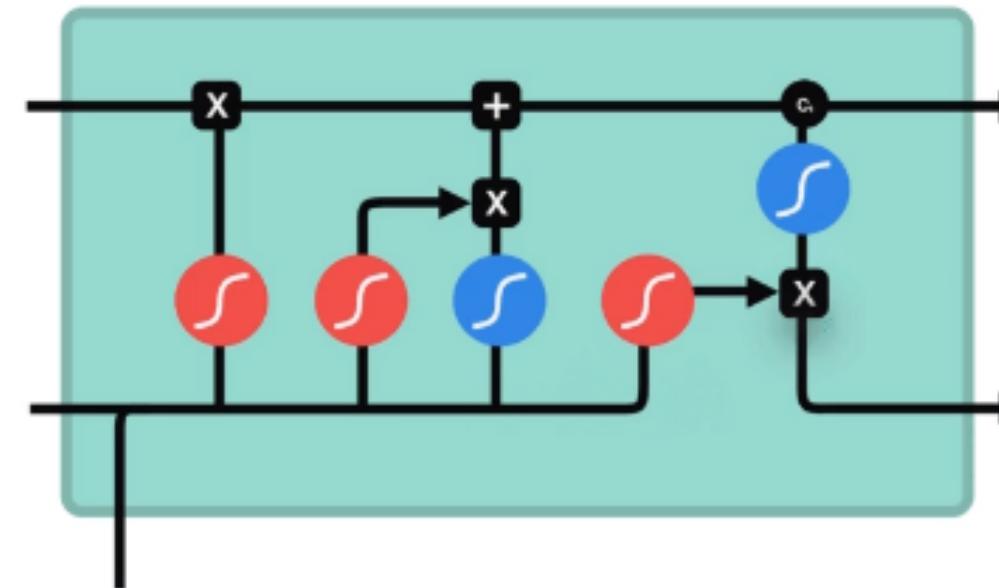
Gerbang Pertama: *Forget Gate*



Gerbang Kedua: *Input Gate*

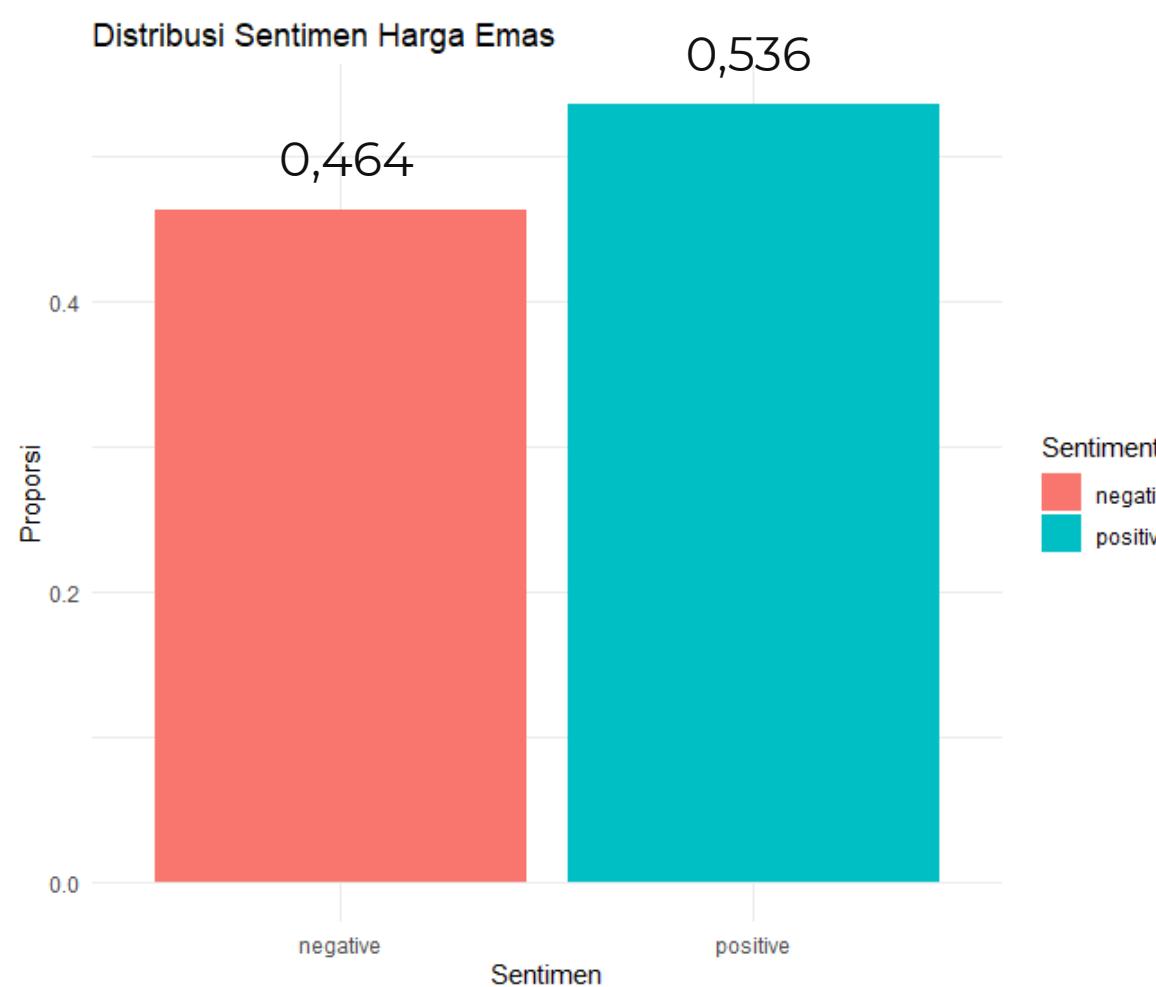


Gerbang Ketiga: *Output Gate*



Karakteristik Data

Distribusi Sentimen



Wordcloud Positif

dec prices lower silver low falls dollar
gold futures ounce

Wordcloud Negatif

dollar high gains ounce
futures gold silver higher prices gain

Data Teks Pre-Processing

Case Folding

- Menghilangkan tanda baca & simbol
- Menghilangkan angka
- Menghilangkan kata-kata yang kurang formal
- Menggantikan huruf kapital menjadi huruf kecil
- Menghilangkan spasi yang berlebih
- Metransformasi kata-kata yang kurang formal atau baku
- Menghilangkan singkatan

News
april gold down 20 cents to settle at \$1,116.10/oz
gold suffers third straight daily decline
Gold futures edge up after two-session decline
Gold snaps three-day rally as Trump, lawmakers reach debt-ceiling deal
Dec. gold climbs \$9.40, or 0.7%, to settle at \$1,356.90/oz

News_clean
april gold down cents to settle at oz
gold suffers third straight daily decline
gold futures edge up after two session decline
gold snaps three day rally as trump lawmakers reach debt ceiling deal
dec gold climbs or to settle at oz

Tokenizing

- Menentukan jumlah kata yang ingin digunakan berdasarkan jumlah frekuensi kemunculan yang paling banyak (**2.699 Kata**)
- Menghitung jumlah kata dalam sebuah teks (**19 Kata**)

Label Encoding

Tahap ini merupakan pengubahan data string menjadi datanumerik dengan fungsi mutate agar data yang ada dapat dilakukan analisis dengan metode LSTM.

News_Clean	label
april gold down cents to settle at oz	0
gold suffers third straight daily decline	0
gold futures edge up after two session decline	1
gold snaps three day rally as trump lawmakers reach debt ceiling deal	0
dec gold climbs or to settle at oz	1

Model LSTM

Input Layer

- input_dim: Parameter ini bisa diisi dengan jumlah kosakata dalam data teks.
- input_length: Parameter diisi dengan jumlah kosakata yang paling panjang yang terdapat pada satu kalimat.
- output_dim: Ini adalah ukuran ruang vektor di mana kata-kata akan disematkan. Ini mendefinisikan ukuran vektor output dari lapisan ini untuk setiap kata.

Hidden Layer

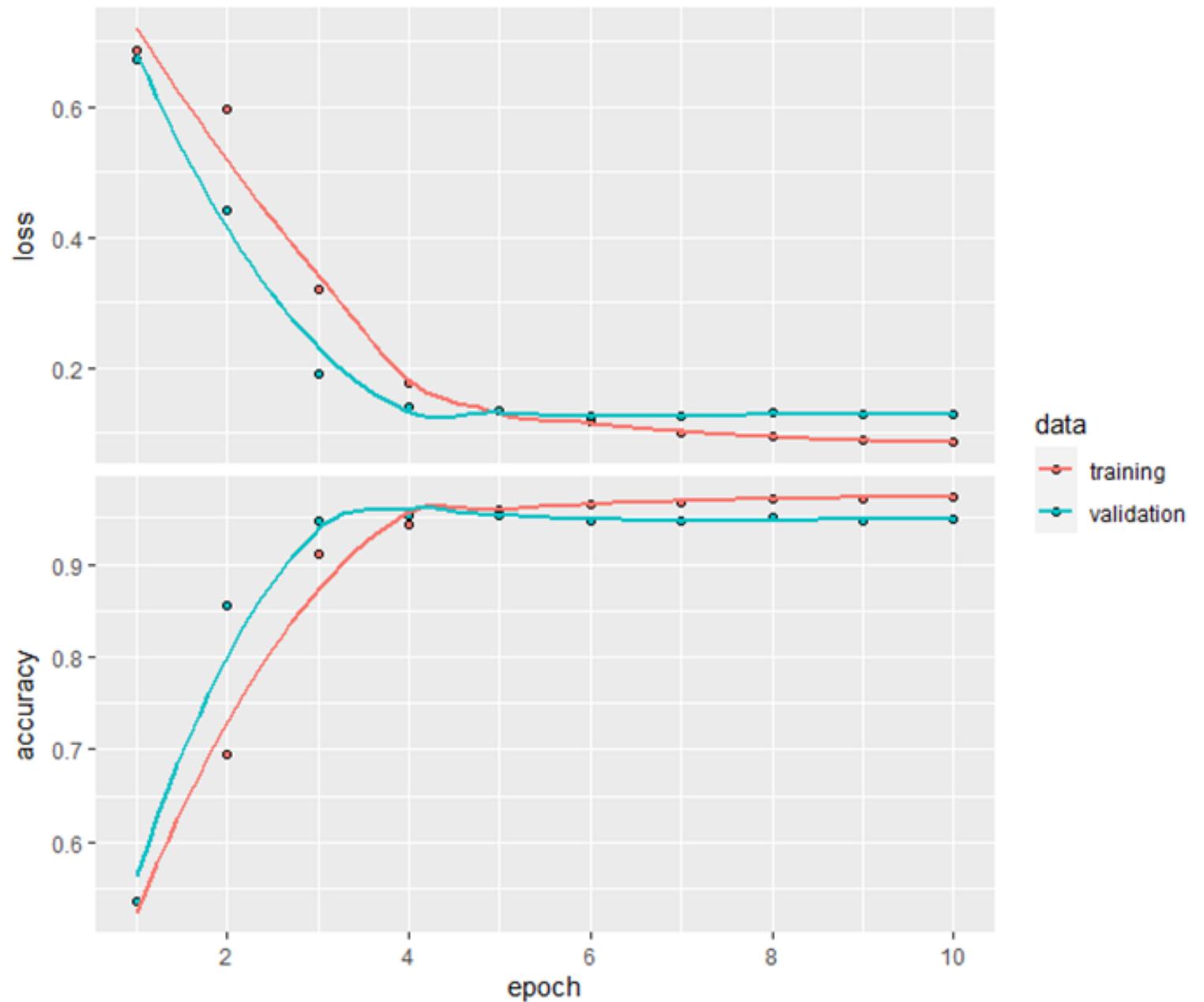
- units: Parameter diisi dengan dengan jumlah memroy cell yang ingin disiapkan untuk mempelajari atau mengingat informasi yang dikeluarkan dari embedding layer.
- dropout: Parameter dropout ditambahkan untuk mengurangi risiko overfit, range nilai dropout antara 0 hingga 1.

Output Layer

- units: Parameter units diisi dengan jumlah target variabel dari data kita.
- activation: Parameter ini berfungsi untuk mengtransformasi nilai dari setiap input, sehingga nilainya nanti tidak terlalu besar dan dapat ditafsirkan menjadi sebuah kelas yang sesuai dengan jumlah target variable kita.

Layer (type)	Output Shape	Param #
input (Embedding)	(None, 19, 32)	48000
lstm (LSTM)	(None, 64)	24832
output (Dense)	(None, 2)	130
Total params:	72962 (285.01 KB)	
Trainable params:	72962 (285.01 KB)	
Non-trainable params:	0 (0.00 Byte)	

Evaluasi Model



Accuracy

0,95

Sensitivity

0,945

Specificity

0,955

Confusion Matrix Heatmap LSTM



Kesimpulan



- Dari 8.126 data yang digunakan, **sentimen positif** terhadap harga emas adalah sebanyak **53,6%** sedangkan sisanya sebanyak **46,4%** merupakan **sentimen negatif** terhadap harga emas.
- Hasil model analisis sentimen menggunakan metode LTSM menghasilkan akurasi model sebesar 0,9502, sensitivity 0,9450, specificity 0,9548, serta AUC-ROC sebesar 0,95. Artinya **model sudah cukup baik dan dapat membedakan kelas positif atau negatif dengan cukup baik.**



Thank you