ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

ANALYSIS OF USER COMPLAINTS FOR TELECOMMUNICATION BRANDS ON X (TWITTER) USING INDOBERT AND DEEP LEARNING

Valianda Farradillah Hakim[1], Dwiza Riana[2]

[1,2]Computer Science, Nusa Mandiri University, Indonesia

email: 14210191@nusamandiri.ac.id[1], dwiza@nusamandiri.ac.id[2]

Abstract

Tweeting on different official accounts is what users of Twitter (X) do most frequently. These tweets ranging from compliments to critiques. One of the official accounts that gets a lot of tweets from its customers is Telkomsel, an Indonesian telecom company. This study aims to find the maximum accuracy that can be obtained by combining CNN and Bi-LSTM algorithms with IndoBERT embeddings. A considerable accuracy level above 90% is demonstrated by the study, with CNN obtaining the greatest accuracy of 99% at a learning rate of 6*10^-5, along with scores of 98%, 97%, and 97% for precision, recall, and F1 correspondingly.

Keywords : Twitter, IndoBERT, CNN, Bi-LSTM, Telkomsel

Received: 16-03-2024 | **Revised:** 30-06-2024 | **Accepted:** 03-07-2024
DOI: https://doi.org/10.23887/janapati.v13i2.76497

INTRODUCTION

X (Twitter) gains increasing popularity. With 556 million members as of early 2023, Twitter was ranked 14th in the world [1], with 24 million of those users being active in Indonesia [2]. This platform is frequently used for exchanging tweets, photos, videos, and other types of interaction.

Good and bad things are happening with the rise of X (Twitter). Posting text as tweets is one of the most common activities among users, particularly on the official accounts of well-known brands. These tweets range from positive to negative. Users frequently use emoticons and other symbols, such as question marks and exclamation points, to convey their feelings. Some people even write entire sentences in capital letters to express strong emotions. These statements are frequently directed at particular brand social media accounts and can range from praise to disappointment depending on the situation.

In this study, two deep learning algorithms, Bidirectional Long Short-Term Memory, or Bi-LSTM, and Convolutional Neural Network, or CNN—are added to the Bahasa version of Bidirectional Encoder Representations from Transformers (BERT) to enable text mining of informal writing. The study focuses on unstructured text data that was taken from tweets about well-known Indonesian telecom companies. Among these carriers, Telkomsel stands out as having the largest subscriber percentage in Indonesia as of 2023 40.27%, according to a Bahasa Internet Service Providers Association (APJII) study with 8,510 participants in 38 provinces. With 174.5 million customers overall in 2022, 68% of them used mobile data [3][4].

The use of IndoBERT embedding transform, Bi-LSTM and CNN algorithms for deep learning in this study is justified by their respective strengths. IndoBERT, which was introduced in 2020, serves as the Indonesian version of BERT, making it a timely choice for text mining applications. Bi-LSTM was chosen because it processes sentences bidirectionally and effectively captures information from both sides [5]. In addition, CNN is selected due to its robust feature extraction capabilities, which enables the identification of complex patterns and automatic learning of important text features [6][7].

This study aims to find the maximum accuracy while utilizing IndoBERT transformation embeddings along with Bi-LSTM and CNN deep learning algorithms for data modeling. Study conducted prior to the 2024 Indonesian elections revealed accuracy of 0.7950 (p1) and 0.7800 (p2) with a dataset of 1,000 tweets [8]. Another study that used IndoBERT as an embedding method to study national capital displacement used IndoBERT and chi-square, showing 94% accuracy without cross-validation on a dataset of 5,892 tweets [9]. There are many applications for

using the IndoBERT embedding: opinion mining, sentiment analysis, etc

Related Works

IndoBERT has been used in a number of study in a variety of scientific fields. Study on public complaints, for example, uses various pre-processing approaches and IndoBERT fine-tuning models to evaluate the effect of text input. These studies also use algorithms like MLP, LSTM, Bi-LSTM, CNN, and GRU to compare the outcomes with raw text data [10]. Furthermore, fine-tuning and twitter data pre-processing approaches have been used to apply IndoBERT to the detection of depression [11] and the detection of cyberbullying on X (Twitter) and Instagram [12].

METHOD

IndoBERT

IndoBERT is a language model particularly trained using Huggingface, based on Bidirectional Encoder Representations from Transformers (BERT) and customized for Bahasa Indonesia [13]. With 74 million, 55 million, and 90 million data points, respectively, from three primary datasets—the Bahasa Wikipedia, news items from Tempo, Kompas, and Liputan 6—and the Bahasa corpus website—IndoBERT uses more than 220 million training parameters [14]. The concept uses a transformer mechanism to look at how words relate to one another in a phrase or text. In general, this system consists of two parts: a decoder that makes predictions and an encoder that reads input text [11].

Bi-LSTM

Introduced by Schuster and Paliwa [15], Bidirectional Long Short-Term Memory (Bi-LSTM) is a form of Recurrent Neural Network (RNN) that is an improved version of RNNs that overcomes its drawbacks. Because Bi-LSTM uses memory cells with input, output, and forget gates [16], it may analyze input sequences by using data from both previous and subsequent iterations. This processing makes use of networked interconnecting layers.

CNN

The Convolutional Neural Network (CNN) is a method for extracting features from data using convolutional architecture [17]. CNNs are able to identify long-term relationships in text by using many convolutional layers in conjunction with max-pooling layers, which minimize the amount of parameters and expedite training in comparison to other models [18]. Here is the CNN architecture divided into three main layers :

1.Convolutional Layer: Using a combination of linear and non-linear processes, including convolution and activation functions, this main component extracts features [19].

2. Pooling Layer: The outputs from several places are combined into singular values by these nonlinear components. For smaller input data, this consolidation increases the sensitivity and accuracy of feature translation. This step also minimizes memory usage and maximizes statistical performance by gradually reducing input dimensions. It is essential to the lowering of feature dimension [20].

3.Fully Connected Layer: This layer, which carries out the Fully Connected (FC) strategy, is positioned at the conclusion of each CNN design. CNN classification tasks are made easier by the connections between every neuron and every other neuron in the layer above [21].

Research Flow

This study follows the research flow shown in Figure 1: Data collection, Label Encoding, Training and Validation Data, IndoBERT Embedding, Network (Bi-LSTM and CNN), Model (Bi-LSTM and CNN) with Test Data, Evaluation, and Comparison.

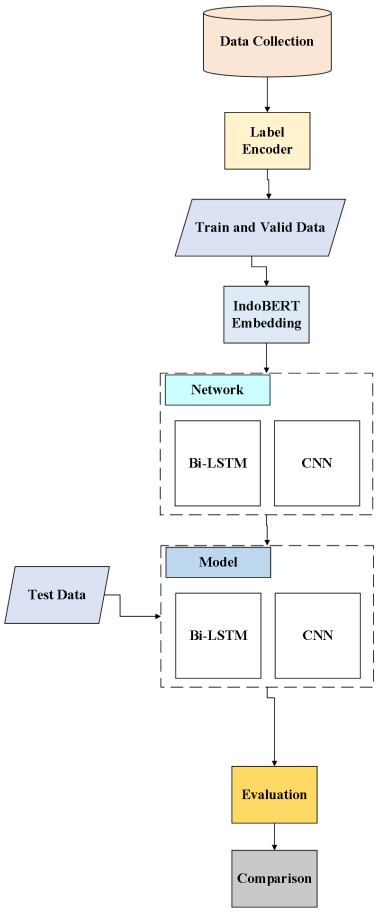

Figure 1 Research Flow

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

The libraries used in this study are tensorflow, pandas, numpy, keras and transformer to read data and perform modeling of IndoBERT with Bi-LSTM and CNN, pyplot, graphviz and seaborn to display the results of the Bi-LSTM and CNN model plots and there is also sklearn for the evaluation metrics of the train, valid and test data.and display the results of the evaluation metrics from the data.

Data Collection

The dataset for this study is text-based or tweets because the text data is taken from X (Twitter) with the object of study on Telkomsel's official social media account (@Telkomsel) with a total of 4,550 tweet starting from January 01, 2023 to March 30, 2023.

The text data is taken using one of the libraries in Python programming to retrieve tweet data from X (Twitter) is SNSCRAPE 0.7.0 and the data is collected in Microsoft Excel software in .csv format. This category attribute has five categories used in study, there are Credit (Pulsa in Bahasa), Signal, Data Package, Card and Application.

The following is information from the dataset in the table 1.

Table 1 Information of Dataset

Attributes	Info	Type
Date	Date and time of tweet text data	Date
Text	Content of user tweet text data	Char
Category	Categories of tweet text data from users	Char

Split Data

A total of 60% (2,730 tweets), 24% (1,092 tweets), and 16% (728 tweets) that will be used are train, validation, and test data. Python 3.10 is the programming language utilized in this study, while Visual Studio Code is the integrated development environment. The indobenchmark indobert-lite-base-p1 tokenizer is used with IndoBERT.

Implementation of Methods

Figure 2 shows how IndoBERT is implemented using Bi-LSTM and CNN. It features components like the Input Layer, IndoBERT as the Embedding, Conv1D/Bidirectional as the Network, Global Max Pooling as the Pooling layer, Dense, Dropout, and Output. As shown in Figure 3, the Input Layer's that are specific include Input, Token Embeddings, Segment Embeddings, and Position Embeddings.

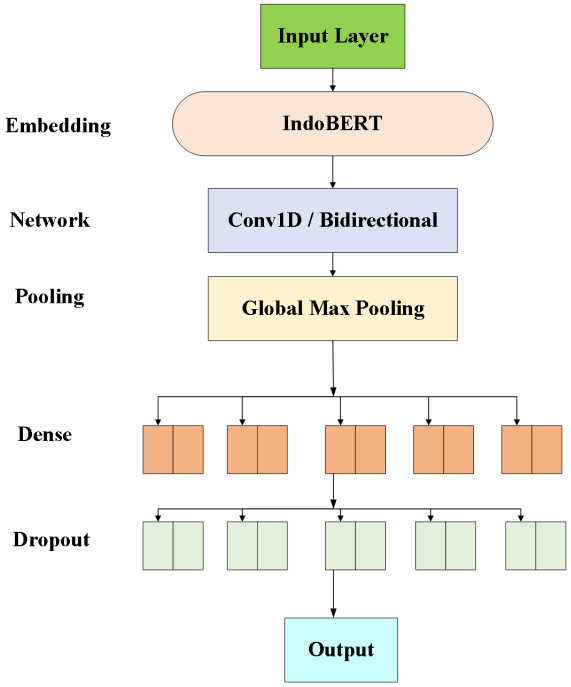

Figure 2 The Proposed Method Architecture

This architectures has seven processes, namely the Input Layer (for the full explanation is on Figure 3), Embedding which uses the transformer method, namely IndoBERT, Network which uses deep learning algorithms like CNN (Conv1D) and Bi-LSTM (Bidirectional), Pooling which uses global max pooling, dense, dropout and the last one is Output, also used various types of tuning hyperparameters in table 2.

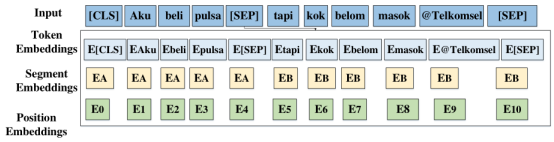

Figure 3. Input Layer

Table 2. Hyperparameter Tuning

Parameter	Number
Batch Size	16
Learning Rate	$2*10^{-5}$, $3*10^{-5}$, $4*10^{-5}$ and $6*10^{-5}$
Epoch	5
Loss	*Sparse Categorical Cross-entropy*
Optimizer	*Adam*
Activation	*Softmax*
Pool	*Global Max Pooling*
Bi-LSTM	64
Dropout	(0,1)
Dense	-
Conv1D (CNN)	(64,5)

In order to achieve a compromise between memory economy and precise evaluation metric values during model execution, a batch size of 16 was used. The dataset's multiclass character led to the selection of sparse categorical cross-entropy as the loss function [22]. Because of its better performance and quicker calculation time than other optimizers, as seen in related research like illness prediction [23], the optimizer Adam was selected. The choice of hyperparameters was in line with earlier studies on categorizing casual text on various subjects [10].

This input layer has four processes there are Input, Token, Segment and Position processes, this input will be exemplified by one informal text data tweet data with the sentence "Aku beli pulsa tapi kok belom masok @Telkomsel" in the input there are also two additional words, CLS which is a class and SEP which is a separator from the data, then there are Tokens that mark the previous input of each word, then it will be processed in Segment with symbols A and B which have been encoded into vectors and the last one is Position which means the relative position proximity in the input sequence.

RESULT AND DISCUSSION

Dataset

This dataset is collected on X (Twitter) social media from Telkomsel official account @Telkomsel using one of the libraries in Python, SNSCRAPE 0.7.0 with a total of 4,550 data starting from January 01 to March 30, 2023. The data length maximum is 1000.

The data results will be labelled into five categories, namely credit, signal, data package, card and application. Here are the detail of the five categories and the number of each categories in Table 3.

Table 3. Five Categories of Dataset

Categories	Total
Credit	693
Signal	2.324
Data Package	1.183
Card	167
Application	183

The highest total of data from the five data categories is the **Signal** category with a value of 2,324 data and the lowest total of data from the five categories is the **Card** category with a value of 167 data. This data labeling is carried out by determining the top five types of complaints on X (Twitter) and also validated by an expert who works at Telkomsel. Here are the data examples from the dataset (the language of the dataset are in Indonesian) in Table 4.

Table 4. Example of Dataset

idx	Text	Category
1	@Telkomsel Mending kuota gak usah dibagi2 @Telkomsel, kaya gak ikhlas banget bikin paket kuot"	Data Package
2	@Telkomsel semakin kesini semakin lama dan ilang2an sinyal telkom. Semakin jauh tertinggal dibelakang. Lokasi mau di tengah kota ataupun ujung tetap sama. Perbaiki	Signal
3	@Telkomsel min saya pesan kartu perdana via web telkomsel kok belum di kirim?Ã½ ya dari tgl 29. padahal pake gosend instant?	Card
4	@Telkomsel min saya isi ulang pulsa melalui my Telkomsel tapi belum masuk nih, gmn min? Saya isi 25rb bayar pake dana udh kepotong di sana pulsa ya belum masuk masukÃ°Å¸Ëœâ€™"""	Credit
5	Kenapa sih pas kerja di rumah @Telkomsel sinyalnya ilang mulu da	Signal
6	@Telkomsel halo kak, ini gimana ya, saya udh kena masa tenggang, dan harus isi pulsa. Dan masa aktif sesuai dgn nominal pulsa yg dibeli. Masa saya harus beli pulsa terus agar tetap aktif? Bukannya jangka nya panjang ya masa aktif gakcuma per 5 hari dst."	Credit
7	@Telkomsel kapan dimunculin lagi daily checkin nya di app myTelkomsel ☹️	Application
8	@Telkomsel Halo min, bisa kah saya blokir nomor telepon untuk digunakan kembali di kemudian hari? karena HP saya hilang	Card

However, in this study, users who complained about the problems are not shown due to privacy concerns.

Label Encoder

This encoder label will change the category of the data type from character to integer to facilitate the data analysis process in program, here are the results of the encoder label in Table 5.

Table 5 Label Encoder

Text	Encoding
Credit	1
Signal	2
Data Package	3
Card	4
Application	5

IndoBERT Encoding

This encoding will change the text of the data type from character to integer. The purpose of the encoding is to be received and processed by a computer system or learning model. Here are the results of the IndoBERT encoding in Table 6.

Table 6. IndoBERT Encoding

idx	Encode
1	[2, 30478, 6950, 12653, 7835, 1489, 3370, 4077, 30378, 30478, 6950, 30468, 2913, 1489, 9051, 2174, 2999, 1998, 425, 105, 3]
2	[2, 30478, 6950, 977, 13508, 977, 985, 41, 20968, 30378, 5, 5711, 4808, 30470, 977, 1229, 9157, 16855, 30470, 1604, 422, 26, 1172, 626, 2057, 3287, 830, 500, 30470, 12510, 30470, 3]
3	[2, 30478, 6950, 459, 209, 2054, 1805, 6380, 3807, 1138, 6950, 2105, 659, 26, 2785, 30477, 1, 286, 98, 8172, 2883, 30470, 2234, 3125, 23148, 179, 17671, 30477, 3]
4	[2, 30478, 6950, 459, 209, 2313, 2420, 4293, 709, 3717, 6950, 469, 659, 804, 2904, 30468, 12169, 459, 30477, 209, 2313, 1423, 6630, 5759, 3125, 1869, 13131, 30317, 2821, 30365, 26, 2377, 4293, 286, 659, 804, 1, 30458, 3]
5	[2, 2124, 1966, 280, 494, 26, 448, 30478, 6950, 5711, 57, 20968, 20968, 19176, 1299, 3]
6	[2, 30478, 6950, 9302, 1844, 30468, 92, 4255, 286, 30468, 209, 13131, 5381, 890, 27841, 30468, 41, 308, 2313, 4293, 30470, 41, 890, 1786, 786, 4240, 11644, 4293, 741, 7074, 30470, 890, 209, 308, 1320, 4293, 944, 579, 830, 1786, 30477, 8003, 3577, 1107, 1422, 286, 890, 1786, 1489, 25013, 62, 418, 406, 13998, 30470, 3]
7	[2, 30478, 6950, 2854, 369, 749, 8993, 423, 16402, 3910, 1107, 26, 4763, 3717, 6701, 3498, 1796, 30477, 30472, 30464, 3]
8	[2, 30478, 6950, 9302, 459, 30468, 166, 7970, 209, 22335, 1288, 3178, 90, 781, 755, 26, 682, 406, 30477, 211, 2109, 209, 2867, 3]

Device Specifications

The device specifications in this study use a laptop device with detailed specifications ranging from the laptop brand to the operating system used in Table 7.

Table 7. Device Specifications

Device Specifications
Asus TUF Dash F15
NVIDIA GeForce RTX 3050 Ti
11th Gen Intel ® Core ™ i7-11370H @ 3.30 Hz (8 CPUs) ~33Ghz
RAM 40960 MB
Windows 11 Home 64-bit (10.0, Build 22621)

The device specifications in Table 4 are qualified in processing larger amounts of data, because it requires large memory and high processing speed such as large RAM specifications or using 40 GB and its GPU which already uses RTX 3050 Ti, in one studies this GPU is quite capable, especially in running models from deep learning [24].

The Result of IndoBERT with Bi-LSTM and CNN

The results of this study using IndoBERT with Bi-LSTM and CNN in Table 8 and Table 9.

Table 8. Result of IndoBERT with Bi-LSTM

Learning Rate	IndoBERT Bi-LSTM			
	Accuracy	Precision	Recall	F1
$2*10^{-5}$	**99%**	**95%**	**96%**	**95%**
$3*10^{-5}$	98%	94%	93%	93%
$4*10^{-5}$	99%	9%5	96%	95%
$6*10^{-5}$	99%	94%	96%	95%

The highest accuracy obtained in Table 8 using IndoBERT and Bi-LSTM is 99% with three different learning rates, namely $2*10^{-5}$, $4*10^{-5}$ and $6*10^{-5}$, but there are differences especially in the evaluation metrics on precision with a value of 95%, so the highest accuracy value obtained is at a learning rate of $2*10^{-5}$ with the respective values of the evaluation metrics of accuracy, precision, recall and F1 are 99%, 95%, 96% and 95%. The lowest accuracy value obtained using IndoBERT and Bi-LSTM is 98% with a learning rate of $3*10^{-5}$ with precision, recall and F1 values of 94%, 93% and 93% respectively, thus the percentage difference especially in accuracy is only 1%.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

Table 9 Result of IndoBERT with CNN

Learning Rate	IndoBERT CNN			
	Accuracy	Precision	Recall	F1
$2*10^{-5}$	99%	96%	94%	95%
$3*10^{-5}$	99%	97%	97%	97%
$4*10^{-5}$	98%	94%	95%	94%
$6*10^{-5}$	**99%**	**98%**	**97%**	**97%**

The highest accuracy obtained in Table 9 using IndoBERT and CNN is 99% with three different learning rates, namely $2*10^{-5}$, $3*10^{-5}$ and $6*10^{-5}$, but there is a difference in the precision metric with a value of 98%, so this highest accuracy value is obtained using a learning rate of $6*10^{-5}$ with values of accuracy, precision, recall and F1 evaluation metrics of 99%, 98%, 97% and 97%, respectively. The lowest accuracy value obtained using IndoBERT and CNN is 98% with a learning rate of $4*10^{-5}$ with precision, recall and F1 values of 94%, 95% and 94% respectively.

The Comparison of Accuracy and Loss of Epoch

Here is the comparison of accuracy and loss of epoch in Table 10, Table 11, Figure 4 and Figure 5 based on test data with 16% of the total tweet (728 tweet).

Table 10. Accuracy of Epoch

Learning Rate	Epoch	IndoBERT	
		Bi-LSTM	CNN
		Accuracy	
$2*10^{-5}$	1	0,8864	0,8527
	2	0,9861	0,9846
	3	0,993	0,9905
	4	0,9952	0,9927
	5	0,9971	0,9945
$3*10^{-5}$	1	0,9114	0,8897
	2	0,9853	0,9883
	3	0,9941	0,9897
	4	**0,9985**	0,9912
	5	0,996	**0,9949**
$4*10^{-5}$	1	0,8923	0,8557
	2	0,9875	0,9542
	3	0,9945	0,9542
	4	0,9912	0,9619
	5	0,996	0,9846
$6*10^{-5}$	1	0,9018	0,9088
	2	0,9853	0,9846

The highest accuracy obtained in Epoch to Accuracy of IndoBERT and Bi-LSTM is at a learning rate of $3*10^{-5}$, epoch of 4 and accuracy of 0.9985 and the highest accuracy obtained in Epoch with Accuracy of IndoBERT and CNN is at a learning rate of $3*10^{-5}$, epoch of 5 and accuracy of 0.9949.

Table 11 Loss of Epoch

Learning Rate	Epoch	IndoBERT	
		Bi-LSTM	CNN
		Loss	
$2*10^{-5}$	1	0,3218	0,7197
	2	0,0554	0,2213
	3	0,0284	0,1606
	4	0,0188	0,1303
	5	0,0166	0,1126
$3*10^{-5}$	1	0,2657	0,4658
	2	0,0604	**0,1077**
	3	0,0261	0,096
	4	**0,0114**	0,0755
	5	0,0168	0,0545
$4*10^{-5}$	1	0,2987	0,5326
	2	0,0559	0,1892
	3	0,0245	0,1646
	4	0,0314	0,1509
	5	0,0196	0,1247
$6*10^{-5}$	1	0,2999	0,4356
	2	0,0616	0,1737

The lowest loss obtained in Epoch of Loss of IndoBERT and Bi-LSTM is at a learning rate of $3*10^{-5}$, epoch of 4 and loss of 0.0114. The lowest loss obtained in Epoch to Loss of IndoBERT and CNN is at a learning rate of $3*10^{-5}$, epoch of 2 and loss of 0.1077.

Figure 4 is a comparison of accuracy of epoch model. The more epoch value increases in one iteration, the more it increases. Firstly at learning rate of $3*10^{-5}$ using IndoBERT and Bi-LSTM there is a decrease in Epoch 5 whose accuracy drops from 0.9985 at Epoch 4 has decreased by 0.0025 to 0.9960 therefore these results have inconsistencies in terms of model comparison.

Secondly at learning rate of $6*10^{-5}$ using IndoBERT and Bi-LSTM decreased when entering epoch 3 whose accuracy dropped from 0.9853 at epoch 2 decreased by 0.0007 to 0.9846 then returned consistent until epoch 5, this result is also the same as the use of a learning rate of $3*10^{-5}$ which is also inconsistent.

Third at a learning rate of $6*10^{-5}$ using IndoBERT and CNN decreased when entering epoch 4 whose accuracy dropped from 0.9886 at epoch 3 decreased by 0.0033 to 0.9853 then

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

returned consistently until epoch 5, this result also experienced inconsistency.Lastly at learning rate of $4*10^{-5}$ using IndoBERT and Bi-LSTM decreased when entering epoch 4 whose accuracy dropped from 0.9945 at epoch 3 decreased by 0.0033 to 0.9912 then returned consistently to epoch 5, this result is also the same as the use of a learning rate of $3*10^{-5}$ which is also inconsistent, but when compared to CNN for the two learning rates it is quite consistent.

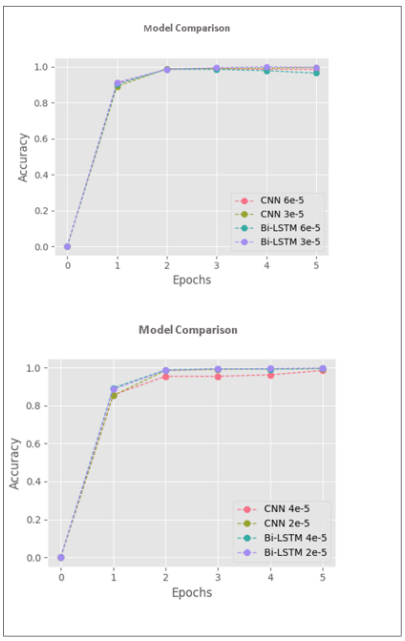

Figure 4 Accuracy of Epoch (based on test data)

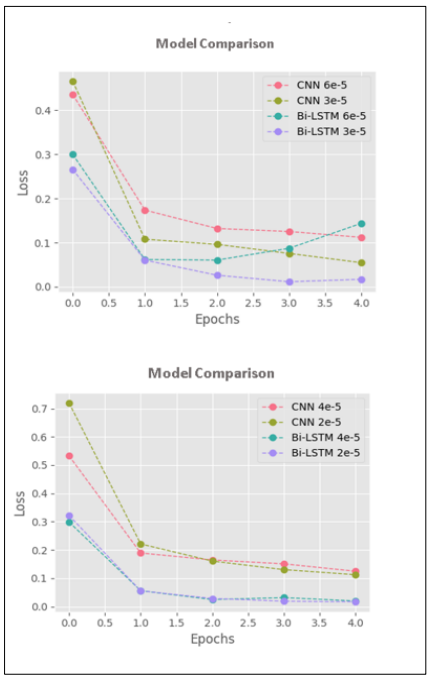

Figure 5 Loss of Epoch (based on test data)

Figure 5 is a comparison of loss of epoch model, in general the more epoch value increases in one iteration, the more the value of the loss obtained generally decreases.Firstly at learning rate of $3*10^{-5}$ using IndoBERT and Bi-LSTM increased when entering epoch 5 where the loss results increased from 0.0114 at epoch 4 increased by 0.0054 to 0.0168.

Secondly at learning rate $6*10^{-5}$ using IndoBERT and Bi-LSTM increased when entering epoch 4 where the loss results increased from 0.0605 at epoch 3 increased by 0.00566 to 0.0872 and at epoch 5 increased again, therefore the results are inconsistent, but in IndoBERT and CNN, the results are consistent.

Third at learning rate of $3*10^{-5}$ using IndoBERT and Bi-LSTM increased when entering epoch 5 where the loss results increased from 0.0114 at epoch 4 increased by 0.0054 to 0.0168. Lastly on the use of a learning rate of $6*10^{-5}$ using IndoBERT and Bi-LSTM increased when entering epoch 4 where the loss results increased from 0.0605 at epoch 3 increased by 0.00566 to 0.0872 and at epoch 5 increased again, therefore the results are inconsistent, but in IndoBERT and CNN. The results are consistent.

Here are the some code to display the results of the epoch comparison with accuracy or loss in Figure 6 that already explained in Figure 4 and Figure 5.

```
history = [history_cnn, history_cnn2, history_bilstm, history_bilstm2]
history = [i.history for i in history]
model_names = ['CNN 4e-5', 'CNN 2e-5', 'Bi-LSTM 4e-5', 'Bi-LSTM 2e-5']

# Set color pallete
import seaborn as sns
qualitative_colors = sns.color_palette("husl", len(history))
```

```
# Accuracy
import matplotlib.pyplot as plt

acc = [i['accuracy'] for i in history]
val_acc = [i['val_accuracy'] for i in history]

plt.figure(figsize=(5, 4))
plt.style.use('ggplot')
for i, a in enumerate(acc):
    plt.plot(range(len(a) + 1), [0] + a,
             linestyle='--',
             marker='o',
             color=qualitative_colors[i],
             linewidth=1,
             label=model_names[i])
plt.legend()
plt.title('Perbandingan Model')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.tight_layout()
plt.show()
```

Figure 6 The Comparison Accuracy and Loss of Epoch (programming code)

Discussion

The results of the study were carried out at the time of comparison of loss to epoch as well as comparison of accuracy to epoch, but the results obtained had inconsistent results using

the IndoBERT transform embedding with using the Bi-LSTM and CNN deep learning algorithms, the cause of this inconsistency was because the number of each class of datasets had significantly different results such as for example in the Signal category which amounted to 2.324 data and the Card category which only amounts to 167 data, besides that there is also noise data also caused in categories where some attributes of the text do not match the category or can be included in two categories and also the model that makes predictions produces different predictions every time training data is done.

The use of the IndoBERT method in this study is to read the input of Indonesian text and processed it in Input Layer (Input, Token, Segment and Position) also known as encoder and then generate the predictions also known as decoder.

The use of Bi-LSTM in this study is after processed in IndoBERT, each of the word is inserted into the Bi-LSTM unit one by one according to the order of the index of the sequence. After that, the input data is processed in the dense layer to classify the data based on the output of the convolutional layer.

The use of CNN in this study is extract features from data with a convolution structure by structuring multiple convolutional layers the long term of the text can be retrieved with a max-pooling layer using fewer parameters in Convolutional Layer, then aggregated into a single value for improving the accuracy and minimize memory consumption in Pooling Layer, the last is to classified the data by CNN in Fully Connected Layer.

The advantages of this study are that there is an estimated processing time that does not take a long time, only less than three hours per method and algorithm because the specifications of the device being analyzed are quite capable in terms of analysis data such as GPUs that already use RTX 3050 Ti and also large RAM or 40 GB, besides that the algorithm methods used are quite capable of text classification such as deep learning algorithms Bi-LSTM and CNN as well as IndoBERT transformation embedding, after that the results of the evaluation metrics obtained are above 90% which is good evaluation.

The disadvantage of this study is that the number of learning rates used is small or only uses four numbers there are $2*10^{-5}$, $3*10^{-5}$, $4*10^{-5}$ and $6*10^{-5}$ as well as the use of batch size which only uses one value, namely 16, when compared to previous studies that used two batch sizes, namely 16 and 32 [8], the cause is the time problem will take a long time, then the study results are also not pre-processed first such as removing @, stemming, casefolding, stopwords and other pre-processing and only cleaning up irrelevant hashtag keywords at the time of data collection such as #ZonaUang, #ZonaJajan, and other hashtags.

CONCLUSION

This study has successfully collected informal text data on user complaints on Telkomsel's official X (Twitter) account @Telkomsel with a total of 4,550 complaint data, which was then classified and encoded. Labelled into five categories there are Credit, Signal, Data Package, Card and Application into categories 0,1,2,3 and 4 respectively, with the most complaints in the Signal category with a total of 2,324 complaints or with a percentage of 51.1% and with the least complaints in the Card category with a total of 167 complaints or with a percentage of 3.7%.

The data is divided into train, valid and test data with 60% (2,730 data), 24% (1,092 data) and 16% (728 data) respectively, then embedding with IndoBERT transformation embedding with Bi-LSTM and CNN deep learning algorithms by adding several tuning hyperparameters which then produce models from these deep learning methods and algorithms. The results of this study obtained good accuracy above 90% with the highest accuracy obtained by the deep learning algorithm, namely CNN of 99% at a learning rate of $6*10^{-5}$ with precision, recall and F1 values getting 98%, 97% and 97% respectively.

Future study is expected to compare machine learning algorithms with deep learning such as comparing SVM, Random Forest and Naïve Bayes with LSTM, GRU and MLP. The amount of data analyzed can be increased again to get 10,000 data. The learning rate, batch size and epoch parameters can be added or increased, then can implement it into informal text of other objects such as complaints of Indosat, XL and other operators in Indonesia, as well as implementing BERT with other Bahasa language versions such as IndoLEM, IndoNLU and RoBERTa.

The device specifications can use higher specifications that can speed up the processing time of the algorithms used than this study. In addition, the resulting high accuracy model can also be implemented in units authorized to identify customer complaints in telecommunications operator accounts, especially Telkomsel.

REFERENCES

[1] "Essential Twitter statistics and trends for 2023," 2023. https://datareportal.com/essential-twitter-stats (accessed Jul. 20, 2023).

[2] S. Kemp, "Twitter users in Indonesia in 2023," 2023. https://datareportal.com/reports/digital-2022-indonesia (accessed Mar. 12, 2023).

[3] A. Ahdiyat, "Operator Seluler yang Digunakan Responden (Januari 2023)," 2023. https://databoks.katadata.co.id/datapublish/2023/06/23/ini-operator-seluler-dengan-pengguna-terbanyak-di-indonesia-awal-2023 (accessed Jul. 21, 2023).

[4] D. Novianty and D. Prastya, "Riset Counterpoint: Telkomsel Jadi Operator Seluler Terbesar di Indonesia," 2022. https://www.suara.com/tekno/2022/07/17/160231/riset-counterpoint-telkomsel-jadi-operator-seluler-terbesar-di-indonesia?page=all (accessed Jul. 31, 2023).

[5] K. S. Nugroho, I. Akbar, A. N. Suksmawati, and Istiadi, "Deteksi Depresi Dan Kecemasan Pengguna Twitter Menggunakan Bidirectional Lstm," *arXiv*, no. Ciastech, pp. 287–296, 2023, doi: 10.48550/arXiv.2301.04521.

[6] Y. Widhiyasana, T. Semiawan, I. Gibran, A. Mudzakir, and M. R. Noor, "Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia (Convolutional Long Short-Term Memory Implementation for Indonesian News Classification)," *J. Nas. Tek. Elektro dan Teknol. Inf.* |, vol. 10, no. 4, pp. 354–361, 2021.

[7] I. Ayu Shafirra N, "Klasifikasi Sentimen Ulasan Film Indonesia dengan Konversi Speech-to-Text (STT) Menggunakan CNN," *J. sains dan seni ITS*, vol. 9, no. 1, pp. 2301–9271, 2020.

[8] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 3, pp. 746–757, 2023, doi: 10.26555/jiteki.v9i3.26490.

[9] D. Arista, Y. Sibaroni, and S. Prasetyowati, "Sentiment Analysis on Twitter(X) Related to Relocating the National Capital using the IndoBERT Method using Extraction Features of Chi-Square," *J. Media Inform. Budidarma*, vol. 8, pp. 403–411, 2024, doi: 10.30865/mib.v8i1.7198.

[10] A. Kurniasih and L. P. Manik, "On the Role of Text Preprocessing in BERT Embedding-based DNNs for Classifying Informal Texts," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, pp. 927–934, 2022, doi: 10.14569/IJACSA.2022.01306109.

[11] I. R. Hidayat and W. Maharani, "General Depression Detection Analysis Using IndoBERT Method," *Int. J. Inf. Commun. Technol.*, vol. 8, no. 1, pp. 41–51, 2022, doi: 10.21108/ijoict.v8i1.634.

[12] A. Candra, Wella, and A. Wicaksana, "Bidirectional encoder representations from transformers for cyberbullying text detection in indonesian social media," *Int. J. Innov. Comput. Inf. Control*, vol. 17, no. 5, pp. 1599–1615, 2021, doi: 10.24507/ijicic.17.05.1599.

[13] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th COLING*, 2020.

[14] "IndoBERT," 2020. https://indolem.github.io/IndoBERT/ (accessed May 12, 2023).

[15] M. J. Hamayel and A. Y. Owda, "A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms," *Ai*, vol. 2, no. 4, pp. 477–496, 2021, doi: 10.3390/ai2040030.

[16] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM," *Chaos, Solitons and Fractals*, vol. 140, p. 110212, 2020, doi: 10.1016/j.chaos.2020.110212.

[17] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, 2022, doi: 10.1109/TNNLS.2021.3084827.

[18] M. Umer *et al.*, "Impact of convolutional neural network and FastText embedding on text classification," *Multimed. Tools Appl.*, vol. 82, no. 4, pp. 5569–5585, 2023, doi: 10.1007/s11042-022-13459-x.

[19] R. K. G. D. & K. T. Rikiya Yamashita, Mizuho Nishio, "Convolutional neural networks: an overview and application in radiology https://doi.org/10.1007/s13244-018-0639-9," *Springer*, vol. 195, pp. 21–30, 2018.

[20] A. Zafar *et al.*, "A Comparison of Pooling

Methods for Convolutional Neural Networks," *Appl. Sci.*, vol. 12, no. 17, pp. 1–21, 2022, doi: 10.3390/app12178643.

[21] L. Alzubaidi *et al.*, *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00444-8.

[22] Keras, "Probabilistic losses." https://keras.io/api/losses/probabilistic_losses/ (accessed Jul. 20, 2023).

[23] R. Haque, S. B. Ho, I. Chai, and A. Abdullah, "Parameter and Hyperparameter Optimisation of Deep Neural Network Model for Personalised Predictions of Asthma," *J. Adv. Inf. Technol.*, vol. 13, no. 5, pp. 512–517, 2022, doi: 10.12720/jait.13.5.512-517.

[24] R. Lohith, K. E. Cholachgudda, and R. C. Biradar, "PyTorch Implementation and Assessment of Pre-Trained Convolutional Neural Networks for Tomato Leaf Disease Classification," *2022 IEEE Reg. 10 Symp. TENSYMP 2022*, pp. 1–6, 2022, doi: 10.1109/TENSYMP54529.2022.9864390.