ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 11 (2024) No. 5, pp. 457-470
DOI:10.33168/JLISS.2024.0526

Comparative Evaluation of CNN-LSTM Model for Emotion Detection in Indonesian Text

Sandro Owen, Wella

Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Indonesia

Abstract. This study developed an emotion detection model for Indonesian text using a dataset from prior research. The data was refined through multiple pre-processing steps before applying CNN-LSTM machine learning techniques. Comparative analysis indicated the model achieved 58% accuracy, lower than baseline methods. The results imply need for larger annotated corpora, improved text normalization, and integration with state-of-the-art deep learning approaches to enhance performance for Indonesian emotion detection.

Keywords: CNN, emotion detection model, LSTM, machine learning, pre-processing

1. Introduction

Sentiment analysis is one of the most active areas of research (Bata, 2019). Sentient analysis is aimed at detecting, analysing, and evaluating human thoughts about a particular thing based on writing, facial expressions, speech, music, movements, and so on. According to the Merriam-Webster dictionary, sentiment is an attitude, thought, or judgment driven by a feeling. Emotion mining is the study of an emotion (e.g., happy, sad, angry) that is drawn from an input. There are many studies that have been done in the field of emotional detection with various types of input, one of which comes from text (Institute of Electrical and Electronics Engineers, 2017). Emotion detection from text is a relatively more complex task compared to sentiment detection (Institute of Electrical and Electronics Engineers, 2017). Currently, the system of emotional detection based on Indonesian text is beginning to work (Cahyaningtyas et al, 2017), (Saputri et al, 2018), (Vania et al, 2014), (Ahmad et al, 2019), but there are obstacles in the development of the emotional textual detection model of Indonesia which is included in the category of 'under-resource language' (Bata, 2019). This is due to the lack of available corpus or source of text data of Indonesia that can be used for emotion detection (Bata et al, 2015). Some researchers have provided the datasets/corpus that they use in their research, but the amount of data available is still insufficient (Institute of Electrical and Electronics Engineers, 2017). Moreover, there is no standard dataset to use for emotional detection models (Institute of Electrical and Electronics Engineers, 2017). Therefore, the researchers are bound to use dataset on existing emotional sensing models or create a new dataset and label the data manually where this takes a lot of time to do (Institute of Electrical and Electronics Engineers, 2017). In addition to this, there is a scarcity of extensive and high-quality datasets that especially target informal Indonesian text, such as social media postings and chat chats. This hampers the capacity of models to comprehend the subtleties of casual conversation (Costa-jussà et al, 2022), (Mahmud et al, 2023). Existing pre-processing methods may not effectively address the intricacies of informal Indonesian language, including slang, acronyms, and emojis. The presence of these features can have a substantial impact on the sentiment and emotional tone of the text (Darshan et al, 2024), (Sundaram et al, 2023). Conventional emotion recognition methods may not attain the highest level of accuracy when applied to informal Indonesian text, particularly when dealing with intricate emotions or datasets that have an uneven distribution of emotions, with certain emotions occurring more frequently than others (Hung & Alias, 2023), (Yulianti et al, 2021).

Currently, there are several methods or techniques used to develop a model of emotional detection based on Indonesian text (Saputri et al, 2018), (Vania et al, 2014), (Ahmad et al, 2019). The development of an emotional model using YouTube comments as a source of data and word embedding with CNN methods resulted in an accuracy of 76.2% (Saputri et al, 2018). Furthermore, the development of a model for emotion detection using the Latent Dirichlet Allocation (LDA) method and the conversion of emoji symbols (emoji) to an Indonesia-language tweet yielded the highest accuration of 50.8% (Vania et al, 2014). Development of an Emotional Detection model using the Naïve Bayes algorithm and the dataset developed by Julian Bata using the hashtag (#) on a tweet as an annotation resulted with an accurateness of 82% (Bata et al, 2015). A study was also conducted to build a data set of emotional detection of Indonesian text collected from Indonesians tweet, in addition to the study also suggested some techniques that can be used on the data set in order to conduct emotional classification where the results showed that the use of Logistic Regression algorithm and a combination of features on the dataset gives the best accuracy of 69.73% (Ahmad et al, 2019).

Therefore, based on the above exposure, this study will focus on building a model of emotional detection using the datasets used on (Ahmad et al, 2019) by making the best accuracy of such research as the baseline in this study.

2. Literature Review

2.1.Previous Research

The research conducted by Saputri et al. (2018) and Savigny & Purwarianti (2017) provides evidence of the efficacy of machine learning algorithms, such as Linear Regression, in detecting emotions in Indonesian text. Nevertheless, these methods may not adequately capture the intricate connections present in informal language when compared to deep learning frameworks. Ahmad et al. (2019) and Pathak (2020) conducted research on social media data using advanced deep learning models such as CNN and LSTM. Their findings showed promising levels of accuracy. Nevertheless, these research examine the algorithms separately. Our study examines the synergistic capabilities of a CNN-LSTM model in detecting emotions in colloquial Indonesian text. Sari et al. (2020) investigate the influence of word embedding approaches, such as Word2Vec and GloVe, on the accuracy of text classification. The objective of our study is to investigate the efficacy of the CNN-LSTM model, both with and without word embeddings, in order to assess their impact on the detection of emotions in informal Indonesian language. Although ensemble learning with various algorithms has been shown to attain excellent accuracy in several studies (e.g., Saputri et al., 2018; Ahmad et al., 2019), it is important to note that this strategy can be computationally demanding. The emphasis of our study is a single, meticulously crafted CNN-LSTM model, with the goal of achieving equivalent accuracy while ensuring efficiency. Comparing accuracy scores directly between research is difficult because of the differences in datasets and evaluation methods. Our research on a CNN-LSTM model for informal Indonesian text addresses a gap in the present exploration of deep learning architectures for this specific domain and language. This research seeks to enhance the current level of emotion identification for Indonesian social media data by utilising a CNN-LSTM architecture together with customised pre-processing algorithms specifically designed for informal writing.

2.2.LSTM (Long Short Term Memory)

Long Short Term Memory (LSTM) is a class of RNN that can manage memory for each input using memory cells and gate units (Wildan et al, 2018). LSTM has an architecture of remembering and forgetting the results which will be processed back into input (Siami-Namini et al, 2019). Apart from that, LSTM can maintain errors that occur when carrying out back-propagation so that it does not allow errors to increase (Xiong et al, 2023).

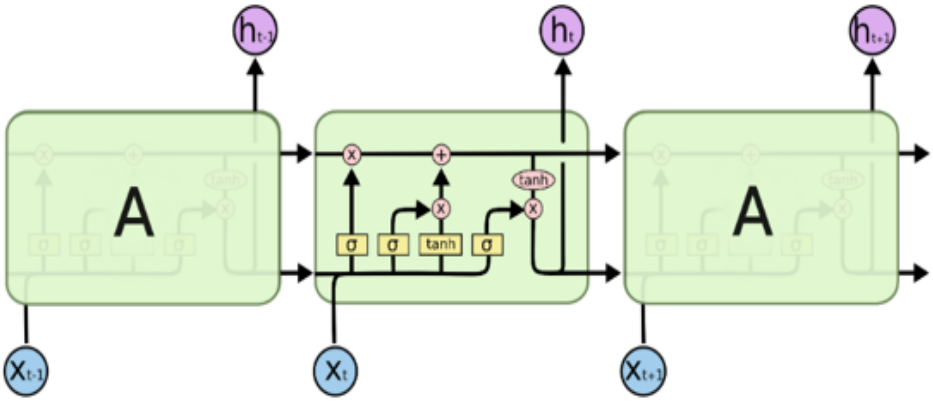

Fig.1: Flowchart memory cells LSTM

Figure 1 illustrates a workflow detailing the memory cells within each LSTM neuron. Each input to the neurons undergoes four activation function processes, collectively termed gate units. These units include forget gates, input gates, cell gates, and output gates (Wildan et al., 2018). The procedures followed in LSTM are outlined as per Olah (2020):

1. Specifies the information to be removed from the cell state. Each input will be processed to choose which ones will be stored or discarded in memory cells. This decision is made by the sigmoid

layer called the forget gate layer. The decision result will be a number between 0 and 1 where 1 states that all input will be saved and 0 states that all input will be discarded.

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

2. Determines what new information we will store in the cell state. In this step there are 2 parts, first the input gate layer determines which value will be updated. Next, the tanh layer will create a vector of new candidate values.

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

3. Updates the old cell state C_{t-1} to become the new cell gate C_t. This value is obtained from the combined value of the forget gate and input gate.

$$C_t = f_t * c_{t-1} + i_t * C_t \quad (3)$$

4. Determine the results (output) that will be issued. The output gates will be based on cell state but in a filtered version using sigmoid activation. First we will run the sigmoid layer which will determine the part of the cell state that will be output. Then, we will change the cell state value to between -1 and 1 with tanh and finally we will multiply the two results to produce the value that will be output.

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \quad (4)$$

3. Research Methodologies

The research object in this study is the dataset used in previous research (Saputri et al, 2019). Table 1 below shows that the dataset consists of 2 columns, namely the tweet column and the label column. This dataset has 4401 tweets with 5 different labels, namely happy, love, fear, sadness, and anger. There are 1101 anger labels, 1017 happy labels, 997 sadness labels, 649 fear labels, and 637 love labels. Usernames, URLs, and sensitive numbers such as telephone numbers and invoice numbers have been changed to certain formats. The process of annotating or labelling the dataset was carried out using the Indonesian emotion dictionary as a guide (Shaver et al, 2001).

Table 1. Example of tweet in bahasa indonesia

No	Label	Tweet in Bahasa Indonesia
1	Anger	Ku juga pengen ngamuk bacain komen2 netizen yg mahabenar. Ya Lord. Ga abis pikir. Komen2 macem gini yg bikin kasus pelecehan seksual akan tenggelam begitu saja. Tidak ada hukuman untuk pelaku. Lebih sedihnya, rerata yg komen jahat adalah perempuan. Mbaq2, kesehatanmu loh! [URL] *I also want to get angry reading the comments of netizens who are correct. Oh my god. Never mind. Comments like this make cases of sexual harassment just go away. There is no punishment for the perpetrator. What's even sadder is that the average person who comments maliciously is women. Sister, your health! [URL]*
2	Fear	Tadi lagi asik ngobrol sama temen-temen SMA terus terbersit di pikiran sendiri "gimana ya aku nanti di alam kubur teman-temanku ini kan hanya fana" serem banget pikiranku ini *I was having fun chatting with my high school friends and the thought came to myself "what will I do in the grave, my friends are only mortal" my thoughts were really scary.*

No	Label	Tweet in Bahasa Indonesia
3	Happy	Kepingin gudeg mbarek Bu hj. Amad Foto dari google, sengaja, biar teman-teman jg membayangkannya. Berbagi itu indah. *Want to eat gudeg mbarek Bu hj. Amad (restaurant name). Photo from Google, on purpose, so that friends can also imagine it. Sharing is beautiful.*
4	Love	kan kupeluk engkau erat2 hingga tak ada seorang pun yang bisa merebut mu dari pelukan ku happy anniversary syg #6nov #31months *I'll hug you so tightly that no one can snatch you from my arms. Happy anniversary darling #6nov #31months*
5	Sadness	Turut berduka cita buat kawan a cak. Semoga. Keluarga yg di tinggalkan a. Bisa di beri kesabaran &amp meridoi kepergian alm.. #sajete #wani *My condolences to my friend. Hopefully. The family left behind can be given patience & condolences for the passing of the deceased... #sajete #wani*

3.1.Research Methods

This study conducted several phases that can be seen in Figure 2 to develop an emotion detection model based on Indonesian text.

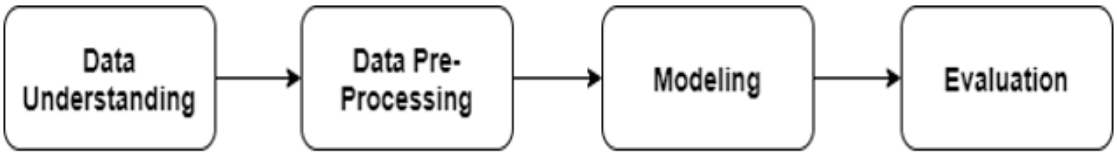

Fig.2: Research Method

Data understanding is the first phase that intended to recognize and analyse dataset that will be used. The dataset used in this study is a dataset that was built by collecting Indonesian tweets from 1st June 2018 to 14th June 2018. This dataset has 4401 tweet with 5 different label (happy, love, fear, sadness, and anger). There are 1101 anger data, 1017 happy data, 997 sadness data, 649 fear data, and 637 love data. Tweets in this dataset are not using formal Indonesian language. There are a lot of tweet that used abbreviations, slang, regional, or foreign languages. Figure 3 shows the number of data for each emotion in this dataset.

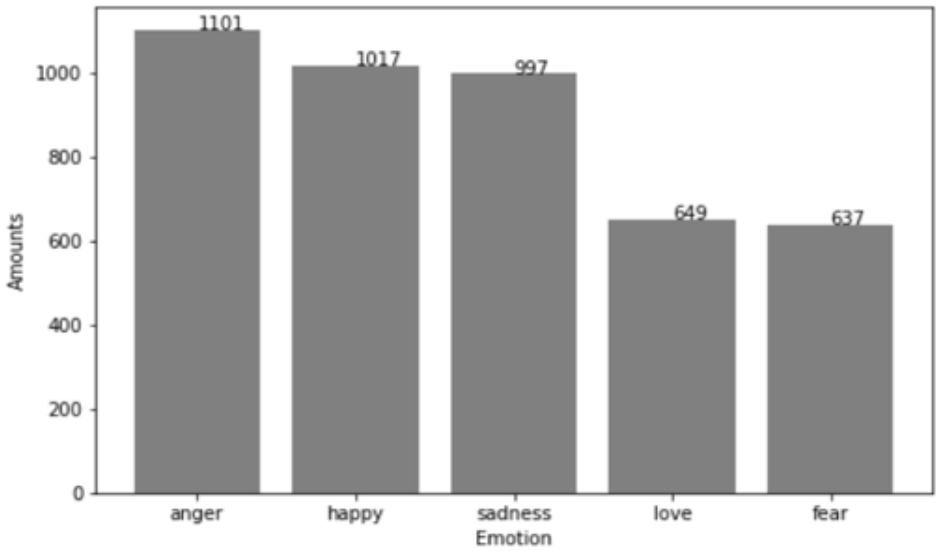

Fig.3: Number of data for each emotion

Data pre-processing is phase where dataset will be prepared first before it being used by the model that will be developed in this study. There are several stage that will be performed on this phase:

a. Data normalization
   In this stage, all of the tweet data will be converted to lowercase (a-z). After that, [USERNAME], [URL], [SENSITIVE-NO], number, white space, and punctuation will be removed from tweet data. Informal literature frequently employs colloquial language, acronyms, and words that are stretched out. The goal of normalisation is to transform these texts into standard Indonesian in order to enhance model comprehension. Glossary of slang terms: Construct a lexicon to establish a correspondence between colloquial terms and their conventional equivalents (e.g., "cepet" to "cepat" - denoting swiftness). Expansion of abbreviations: Create guidelines for the expansion of commonly used abbreviations, such as transforming "gw" into "saya" - I. Elongation management: Formulate a plan (for example, retaining the first letters plus a suffix, transforming "baaaanyak" into "banyak" - meaning many). Casual text may include errors and incorrect spellings. Correction enhances the data quality for the model. Abbreviations, slang, or misspelled words will be change into correct words with Indonesian Typography Dictionaries that been used on previous studies (Nadarzynski et al, 2019), (Dahria, 2008) open source Indonesian Typography Dictionary, and Indonesian Typography Dictionary that created manually by this study. In this stage, the research also train a spell checker on a dataset of informal Indonesian text to capture common misspellings and use contextual information to identify and correct misspelled words, considering the informal nature of the language.
b. Stemming
   Stemming is stage where tweet data that have been normalize before will be change to their root word in order to improve the probability of words that has similar root words counted as one. Stemming will be done using Python's library named Sastrawi where this library is used to change Indonesian words to their root word with Nazief & Andriani algorithm (Ahmad, 2017), (Palasundram et al, 2019).
c. Tokenizing
   Tokenizing is stage where tweet data will be divided to word and each word will be transformed as a token. Each tweet will be converted into token based on dictionary that has been created in order to make it easier for the model to analyze the meaning of a sentence or text based on the word order (Dey, 2020). The dictionary consists of 10,717 words created using dataset in this study as the source.
d. One Hot Encoding
   One hot encoding is one of the technique that use to transform categorical data into integer (Indurkhya, 2020). In this stage, labels in this dataset will be converted using one hot encoding.
e. Split Data
   In this stage, data will be divided into training and validation data where 70% from dataset will be set as training data and 30% will be set as validation data. Data will be randomized beforehand.

Modelling is a phase where model will be built and trained with dataset that has been prepared before in order to meet this study objective. Model will be built using one of deep learning algorithm, CNN with LSTM layer. This algorithm is used because based on previous studies, CNN algorithm with LSTM layer can provide good accuracy and performance for classifying text.

Evaluation is the last phase where the model that has been built will be evaluated. Evaluation was carried out by comparing performance between model in this study and model in previous study that acted as baseline on this study. Performance that will be compared in this study is accuracy where the accuracy in the previous model is 69.73% (Savigny & Purwarianti, 2017).

4. Research and Discussion

The dataset used is a dataset originating from previous research (Saputri et al, 2018). This dataset was built by collecting Indonesian language tweets that were tweeted from June 1 2018 to June 14 2018. Tweets from news accounts, government, and tweets intended for advertising were removed from this dataset. This dataset uses Shaver's basic emotions as a guide for defining emotions or what will later be referred to as labeling a tweet (Shaver et al, 2001). The emotions used in the labeling process in this dataset are happy, love, fear, sadness, and anger. The labeling process was carried out by 2 people (annotators). First, each annotator will be given the dataset that has been collected, then they are asked to separate tweets that do not contain any emotion. After that, each annotator was asked to label each existing tweet with 1 or more emotions because it is possible that a tweet has more than 1 emotion in it. Finally, tweets with more than 1 emotion were removed from this dataset. After carrying out the labeling process and removing unnecessary tweets, this dataset has 4401 tweets with 5 different labels, namely happy, love, fear, sadness, and anger. There are 1101 anger labels, 1017 happy labels, 997 sadness labels, 649 fear labels, and 637 love labels. The tweet data in this dataset does not use standard Indonesian, many tweets use abbreviations, slang, regional languages or foreign languages. Table 2 below are several examples of tweets that use non-standard Indonesian.

Table 2. Examples of tweets using non-standard Indonesian

No	Tweet	Label
#1	Soal jln Jatibaru,polisi tdk bs GERTAK gubernur .Emangny polisi tdk ikut pmbhasan? Jgn berpolitik. Pengaturan wilayah,hak gubernur. Persoalan Tn Abang soal turun temurun.Pelik.Perlu kesabaran. [USERNAME] [USERNAME] [URL]	Angry
#67	btw doa restu orang tua itu emang gilingan ya... pas udah mentok nyari literatur tambahan, nemu aja dong di RUMAH. padahal dari kemarin udah panik kudu mesen dari banyumas apakah ini berkah ramadan [URL]	Happy
#71	[USERNAME] jaringannya mati ya? Tidak bisa dibuka mobile jkn. Saya mau ke puskesmes trus piye mau tunjukkan kartu elektoniknya? #kecewa	Anger
#26	kan kupeluk engkau erat2 hingga tak ada seorang pun yang bisa merebut mu dari pelukan ku happy anniversary syg #6nov #31months	Love
#87	Tiada henti 2ny q mengucap syukut alhmdllh.... Skrg aku bsa.... Uang jadi panitia kmrn alhmdllh trmkah kau bri brlmpah.... Mski tak sbrpa inysllh lma2 mmbukit amn [URL]	Happy

Case folding is carried out where the tweet data will be converted into all lowercase letters (a-z). Figures 4 show the case folding and examples of tweets before and after case folding.

```
orig_df['tweet'][1]
```

```
'Sesama cewe lho (kayaknya), harusnya bisa lebih rasain lah yang harus sibuk jaga diri, rasain sakitnya haid, dan pan
iknya pulang malem sendirian. Gimana orang asing? Wajarlah banyak korban yang takut curhat, bukan dibela malah dihuja
t.'
```

```
df['tweet'][1]
```

```
'sesama cewe lho (kayaknya), harusnya bisa lebih rasain lah yang harus sibuk jaga diri, rasain sakitnya haid, dan pan
iknya pulang malem sendirian. gimana orang asing? wajarlah banyak korban yang takut curhat, bukan dibela malah dihuja
t.'
```

Fig.4: Example of tweet before and after case folding

Then, [USERNAME], [URL], [SENSITIVE-NO], numbers, white space and punctuation are removed from the dataset if present. Figures 5 show the script used to delete [USERNAME], [URL], [SENSITIVE-NO], numbers, white space and punctuation and examples of tweets before and after deletion.

```
orig_df['tweet'][0]
```

```
'Soal jln Jatibaru,polisi tdk bs GERTAK gubernur .Emangny polisi tdk ikut pmbhasan? Jgn berpolitik. Pengaturan wilayah,hak gubernur. Persoalan
Tn Abang soal turun temurun.Pelik.Perlu kesabaran. [USERNAME] [USERNAME] [URL]'
```

```
df['tweet'][0]
```

```
'soal jln jatibaru polisi tdk bs gertak gubernur  emangny polisi tdk ikut pmbhasan  jgn berpolitik  pengaturan wilayah hak gubernur  persoalan
tn abang soal turun temurun pelik perlu kesabaran'
```

Fig.5: Example of tweet before and after deletion

The large number of tweets that use abbreviations and slang words can cause words that should have the same meaning to be considered different, therefore it is necessary to normalize them by changing abbreviations or slang words into more standard or uniform words (for example, ak, q, sy , aq, gw, changed to the words I or me) . The process of changing the words contained in the tweet into more standard words is carried out using the Indonesian Typography Dictionary. Figures 6 show the change words using several Indonesian Typography Dictionaries and examples of tweets before and after the changes were made.

```
orig_df['tweet'][0]
```

```
'Soal jln Jatibaru,polisi tdk bs GERTAK gubernur .Emangny polisi tdk ikut pmbhasan? Jgn berpolitik. Pengaturan wilaya
h,hak gubernur. Persoalan Tn Abang soal turun temurun.Pelik.Perlu kesabaran. [USERNAME] [USERNAME] [URL]'
```

```
df['tweet'][0]
```

```
'soal jalan jatibaru polisi tidak bisa gertak gubernur emangny polisi tidak ikut pembahasan jangan berpolitik pengatu
ran wilayah hak gubernur persoalan tn abang soal turun temurun pelik perlu kesabaran'
```

Fig.6: Example of tweet before and after change

Tweets that have been normalized at this stage will be converted into basic words. Stemming is carried out using a Python library called Sastrawi where this library is used to change Indonesian words into basic words using the Nazief & Andriani algorithm (Librian, 2017). Figures 7 show perform stemming with the Sastrawi library and examples of tweets before and after stemming.

```
#Before Stemming
df['tweet'][0]
```

```
'soal jalan jatibaru polisi tidak bisa gertak gubernur emangny polisi tidak ikut pembahasan jangan berpolitik pengaturan wilayah hak gubernur p
ersoalan tn abang soal turun temurun pelik perlu kesabaran'
```

```
#After Stemming
df['tweet'][0]
```

```
'soal jalan jatibaru polisi tidak bisa gertak gubernur emangny polisi tidak ikut bahas jangan politik atur wilayah hak gubernur soal tanah aban
g soal turun turun pelik perlu sabar'
```

Fig.7: Example of tweet before and after stemming

Tweet data that has gone through the normalization and stemming process will be used as a source to form an Indonesian dictionary which will later be used to convert tweets into number form (encoding) based on the index of each token (word). This is done to make it easier for the model to analyze the meaning of a sentence or text based on its word order. Tokenizing is done using the Tensorflow library. Figure 4.10 shows the script used to tokenize and create an Indonesian dictionary using tweet data in the dataset. After the dictionary was formed, it was found that there were 10,717 words in the tweet data after normalization and stemming were carried out. Next, encoding will be carried out on the tweet data using the dictionary that we created previously.

Data[0] is the tweet "about Jalan Jatibaru, the police can't bully the governor, but the police won't take part in the discussion, don't do politics, regulate the governor's right area, regarding Tanah Abang, the problem of going down and down is complicated, you need to be patient" which has been converted into a number based on the index in the dictionary that we created previously. Figure 4.12 shows an example where the number '166' that appears in data[0] is the word "problem" in the dictionary that we created.

```
tokenizer.index_word[166]

'soal'
```

Fig.8: Decode numbers into words

Finally, we need to change the dimension length between the tweets we have so that they are the same. The longest tweet dimension will be used as the universal dimension length. There are 85 words in both tweets, the longest tweet in the dataset we have is 85 words. There are 5 categories of labels (emotions) in this dataset, namely anger, happiness, sadness, fear, and love. The labels of this dataset are non-ordinal categorical, where there is no relationship or order correlation between one category and another. The data will be divided into 2, namely training data and validation data with a composition of 7:3 where 70% is for training data and 30% validation data. Data that will be previously shared will be shuffled first. Apart from that, the data will also be divided according to the composition of the total number of labels. This data splitting or dividing process utilizes the Sklearn library using the train_test_split function with additional parameters, namely random_state = 42, shuffle = True, and stratify = label.

The amount of training data after being divided is 3080 data, while the amount of validation data after being divided is 1321 data. Table 3 shows the amount of each data in each category or label after the data is divided into training data and validation data.

Table 3. Amount of each data in each category

Data	Label	Total Data/Label	Total Data
Data *Training*	*Anger*	770	3080
	Happy	712	
	Sadness	698	
	Fear	454	
	Love	446	
Data *Validation*	*Anger*	331	1321
	Happy	305	
	Sadness	299	
	Fear	195	
	Love	191	

The model was built using deep learning algorithm, CNN with LSTM layer. Table 4 is the list of hyper parameters that used in modelling in this study.

Table 4. Hyper parameter list

Hyper Parameter	Value
Epochs	30
Learning Rate (lr)	0.001
Optimizer	Adam
Loss Function	Categorical Cross Entropy
Activation Function – Hidden	Tanh
Activation Function – Output	Softmax

Embedding Size	200
LSTM size	128
LSTM type	Bidirectional
Input length	85

Model was trained using training data that has been divided before. After training, model will be validated using validation data. Model stopped being trained at epoch-30 This happen because after epoch-6, there was no significant change in model's accuracy and after epoch-21, there was no significant change in model's loss.

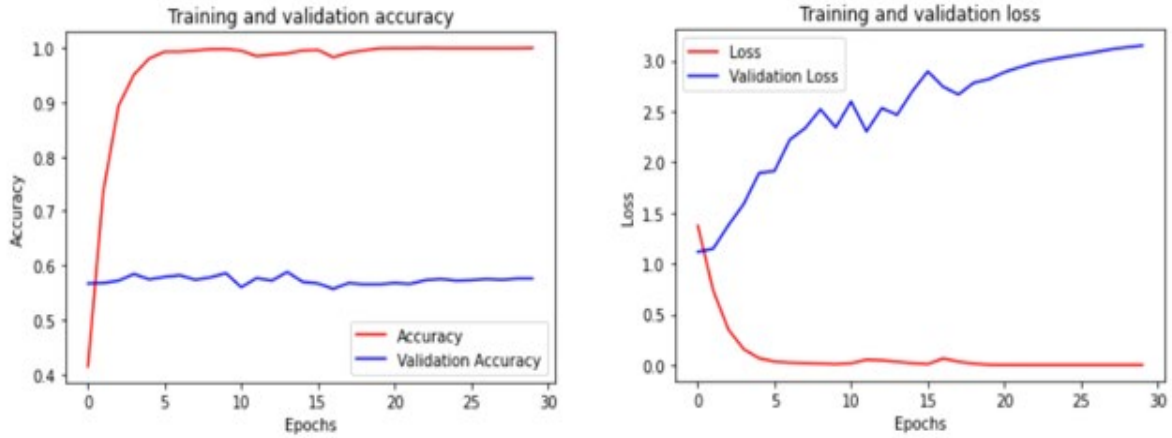

Fig.9: Loss and accuracy of model

Figure 9 shows model's accuracy and loss difference when model uses training & validation data. Accuracy for training data increased drastically in the first to 5th epochs with accuracy almost 100% but for validation data, the accuracy tends to remain and there is no significant change. Other than that, Figure 9 also show the loss difference between training & validation data. Loss in data validation tends to increase while loss in training data decreases to almost 0. This indicates the possibility of overfitting on the training data that used to train the model. It also indicate the possibility of training and validation data have high variance so the model's accuracy on training and validation data is much different. High variance can be happen due to a lot of words that still not formal. There are a lot of abbreviations, slang, regional, or foreign words that still remain and don't get transform due to the lack of source for Indonesian Typography Dictionary. This could affect model performance because words that should have the same meaning are considered as different words because the word form is different.

Table 5. Classification matrix for each emotions

Emotion	**Total**	**Total Prediction**	**TP**	**FP**	**TN**	**FN**
Anger	331	337	212	125	218	119
Fear	195	195	124	71	124	71
Happy	305	281	169	112	145	136
Love	191	167	116	51	92	75
Sadness	299	341	144	197	186	155

TP = True Positive
FP = False Positive
TN = True Negative
FN = False Negative

Table 5 is classification matrix for each emotion that classified by model using validation data. Based on Figure 3, there are 337 angers, 195 fear, 281 happy, 1677 love, and 341 sadness emotion classified by the model. Figure 9 also show the number of true and false classification performed by the model for each emotion.

	precision	recall	f1-score	support
0	0.63	0.64	0.63	331
1	0.64	0.64	0.64	195
2	0.60	0.55	0.58	305
3	0.69	0.61	0.65	191
4	0.42	0.48	0.45	299
accuracy			0.58	1321
macro avg	0.60	0.58	0.59	1321
weighted avg	0.59	0.58	0.58	1321

Fig.10: Classification matrix summary

Figure 10 is classification matrix summary from Figure 9. Precision is the level of confidence of the model in classifying each type of emotion. Precision is calculated based on the number of true positive ÷ the total predictions (true positive + false positive). Recall is the number of correct emotion classified by the model. Recall is calculated by the number of true positive ÷ (true positive + false negative). Number 0-4 in Figure 9 in a row are anger, fear, happy, love, and sadness. Figure 10 shows that anger and fear have balance score on precision and recall. Love emotion has the highest precision score with 69% that means only 31% data from validation data classified as false positive by the model. In addition, sadness has the lowest score for both precision and recall. The high variance in sadness data may affect the model's performance on this emotion. Figure 10 also shows that 58% is the average accuracy for this model.

Table 6. Multiclass confusion matrix

Emotion	Anger	Fear	Happy	Love	Sadness
Anger	**212**	8	35	5	71
Fear	25	**124**	7	3	36
Happy	35	23	**169**	20	58
Love	10	8	25	**116**	32
Sadness	55	32	45	23	**144**

Table 6 is multiclass confusion matrix from the model using validation data. Table 6 shows that 71 tweets that should have been classified as anger was classified as sadness. The same thing also occurs in other type of emotion. The achieved accuracy for emotion detection aligns with the findings of Yulianti et al, (2021) who reported similar accuracy ranges. However, a closer look reveals limitations, particularly regarding the detection of sadness. As observed, the model struggled with classifying sadness, achieving only 48% accuracy (144 out of 299 tweets) for this emotion. This underperformance compared to other emotions suggests that the model might not adequately capture the nuances of sadness expressed in informal Indonesian text. Furthermore, the study by Savigny & Purwarianti (2017) using a Linear Regression model and combined features achieved an accuracy of 69.73% on a different dataset. While directly comparing results across studies with varying datasets is challenging, it highlights the potential for further exploration of alternative model architectures and feature engineering techniques specifically tailored for sadness detection in informal Indonesian text.

5. Conclusion

The emotion detection model demonstrates initial feasibility but is outperformed by existing techniques, indicating addressing data inconsistencies and limitations as a priority. Applying transformers and semi-supervised methods with linguistic priors tailored for Indonesian language shows promise. Overall, the research contributes valuable insights on challenges involved in advancing emotion analysis for low-resource languages.

The recommendation for future study is:

a. Trying to use another algorithm with the same dataset.
b. Reducing the number of anger, happy, and sadness data by removing some of actual data randomly or make the number of those data same as the average fear and love data.
c. Use more data source especially data with sadness emotion to train the model in order to reduce the variance.
d. Build a larger Indonesian dictionary by using sources other than the dataset used.
e. Use more Indonesian Typography Dictionary or build your own Indonesian Typography Dictionary to change non-standard words that have the same meaning into one word.
f. Using existing pre-trained models such as BERT.

References

Ahmad, A., "Mengenal Artificial Intelligence, Machine Learning, Neural Network, dan Deep Learning," no. October, 2017.

Ahmad, S., Asghar, M. Z., Alotaibi, F. M., and Awan, I., "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," Human-centric Comput. Inf. Sci., vol. 9, no. 1, 2019, doi: 10.1186/s13673-019-0185-6.

Bata, J, "# AkuGalau : Korpus Bahasa Indonesia untuk Deteksi Emosi dari Teks," J. Elektro, vol. 12, no. 2, pp. 103–110, 2019.

Bata, J., Suyoto, and Pranowo, "Leksikon Untuk Deteksi Emosi Dari Teks Bahasa Indonesia," Semin. Nas. Inform. 2015 (semnasIF 2015), vol. 2015, no. November, pp. 195–202, 2015.

Cahyaningtyas, R. M., Kusumaningrum, R., Sutikno, Suhartono, and Riyanto, D. E., "Emotion detection of tweets in Indonesian language using LDA and expression symbol conversion," Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017, vol. 2018-Janua, pp. 253–257, 2017, doi: 10.1109/ICICOS.2017.8276371.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Dahria, M., "Kecerdasan Buatan (Artificial Intelligence)," 2008.

Darshan, K., Samuel, J., Manjunatha Swamy, C., Koparde, P., & Shivashankara, N. (2024). NLP-Powered Sentiment Analysis on the Twitter. Saudi J Eng Technol, 9(1), 1-11.

Dey, A., "Machine Learning Algorithms: A Review," 2016. Accessed: 29-Apr-2020. [Online]. Available: www.ijcsit.com.

Hung, L. P., & Alias, S. (2023). Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection. Journal of Advanced Computational Intelligence and Intelligent Informatics, 27(1), 84-95.

Indurkhya, N., "Natural language processing," Computing Handbook, Third Edition: Computer Science and Software Engineering, 2014. https://socs.binus.ac.id/2013/06/22/natural-language-processing/ (accessed Apr. 29, 2020).

Institute of Electrical and Electronics Engineers, "Emotion Classification on Youtube," 2017 Int. Conf. Adv. Informatics Concepts, Theory Appl. Proc. Kuta, Bali, Indones. , pp. 1–5, 2017.

Librian, A. (2017). High quality stemmer library for Indonesian Language (Bahasa). GitHub.[daring] Tersedia pada:< https://github. com/sastrawi/sastrawi>[Diakses 18 Apr 2018].

Mahmud, T., Ptaszynski, M., & Masui, F. (2023). Automatic vulgar word extraction method with application to vulgar remark detection in chittagonian dialect of bangla. Applied Sciences, 13(21), 11875.

Nadarzynski, T., Miles, O., Cowie, A., and Ridge, D., "Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study," Digit. Heal., vol. 5, pp. 1–12, 2019, doi: 10.1177/2055207619871808.

Olah, C., "Understanding LSTM Networks -- colah's blog," Aug-2015. http://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed Apr. 29, 2020).

Palasundram, K., Sharef, N. M., Nasharuddin, N. A., Kasmiran, K. A., and Azman, A., "Sequence to sequence model performance for education chatbot," Int. J. Emerg. Technol. Learn., vol. 14, no. 24, pp. 56–68, 2019, doi: 10.3991/ijet.v14i24.12187.

Pathak, S., "Twitter Sentiment analysis using different Algorithms," Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 9, pp. 1023–1026, 2020, doi: 10.22214/ijraset.2020.31647.

Sailunaz, K., Dhaliwal, M., Rokne, J., and Alhajj, R., "Emotion detection from text and speech: a survey," Soc. Netw. Anal. Min., vol. 8, no. 1, 2018, doi: 10.1007/s13278-018-0505-2.

Saputri, M. S., Mahendra, R., and Adriani, M., "Emotion Classification on Indonesian Twitter Dataset," Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018, pp. 90–95, 2019, doi: 10.1109/IALP.2018.8629262.

Sari, W. K., Rini, D. P., & Malik, R. F. (2019). Text Classification Using Long Short-Term Memory with GloVe. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), 5(2), 85-100.

Sari, W. K., Rini, D. P., Malik, R. F., & Azhar, I. S. B. (2020). Multilabel Text Classification in News Articles Using Long-Term Memory with Word2Vec. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 4(2), 276-285.

Savigny, J., & Purwarianti, A. (2017, August). Emotion classification on youtube comments using word embedding. In 2017 international conference on advanced informatics, concepts, theory, and applications (ICAICTA) (pp. 1-5). IEEE.

Shaver, P. R., Murdaya, U., and Fraley, R. C., "Structure of the Indonesian emotion lexicon," Asian J. Soc. Psychol., vol. 4, no. 3, pp. 201–224, 2001, doi: 10.1111/1467-839X.00086.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019, December). The performance of LSTM and BiLSTM in forecasting time series. In 2019 IEEE International conference on big data (Big Data) (pp. 3285-3292). IEEE.

Sundaram, A., Subramaniam, H., Ab Hamid, S. H., & Nor, A. M. (2023). A Systematic Literature Review on Social Media Slang Analytics in Contemporary Discourse. IEEE Access, 11, 132457-132471.

Vania C., Ibrahim, M., and Adriani, M., "Sentiment Lexicon Generation for an Under-Resourced Language," Int. J. Comput. …, vol. 5, no. 1, pp. 59–72, 2014.

Wildan, M., Jondri, and Aditsania, A., "Analisis dan Implementasi Long Short Term Memory Neural Network untuk Prediksi Harga Bitcoin," e-Proceeding Eng., vol. 5, no. 2, pp. 3548–3555, 2018.

Xiong, R., Wang, S., Yu, C., Fernandez, C., Xiao, W., & Jia, J. (2023). A novel nonlinear decreasing step-bacterial foraging optimization algorithm and simulated annealing-back propagation model for long-term battery state of health estimation. Journal of energy storage, 59, 106484.

Yulianti, E., Kurnia, A., Adriani, M., & Duto, Y. S. (2021). Normalisation of Indonesian-English code-mixed text and its effect on emotion classification. International Journal of Advanced Computer Science and Applications, 12(11).