BINAR

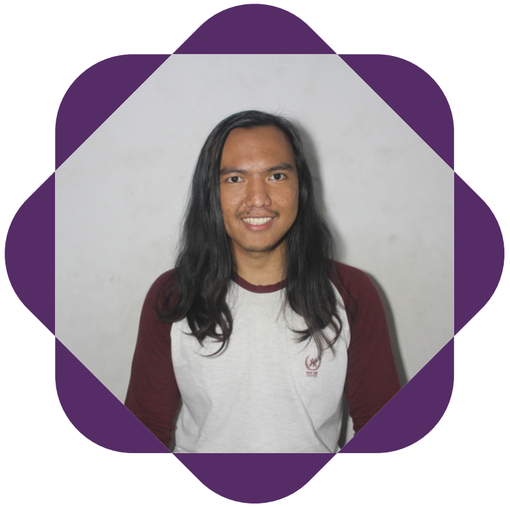
SENTIMENT ANALYSIS OF TWEET NETIZEN +62 USING NEURAL NETWORK AND LSTM

GROUP 4

MIKA HAVENNIA SIRAIT | RIDHO ALFAYET UMAR | SALWA INAS SHABIRA

THE TEAM

MIKA



RIDHO

SALWA

OUTLINE

- Introduction
- Goals
- Process
- EDA
- Deployment results

Introduction

TWITTER

Twitter is an online social media and social networking service owned and operated by American company Twitter, Inc., on which users send and respond publicly or privately texts, images and videos known as "tweets". Registered users can tweet, like, 'retweet' tweets and direct message (DM), while unregistered users only have the ability to view public tweets. Users interact with Twitter through browser or mobile [1].

[1] https://en.wikipedia.org/wiki/Twitter

SENTIMENT

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs [2].

Positive

Positive sentences are sentences that contain constructive, non-demeaning expressions and can contain recommendations, praise, or motivation.

Negative

Negative sentences are sentences that contain expressions that are not constructive, demeaning and can contain judges, insult or dislike.

Neutral

Neutral sentences are sentences that contain expressions that have no specific intention, either praising or insulting

[2] https://monkeylearn.com/sentiment-analysis/

GOALS

1. Analyze sentiment on Twitter sentences (EDA)
2. Determining sentiment of Twitter sentences using Neural Networks and LSTM from deployment use API

MODELS THAT WE USE

Neural Network

Neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.

LSTM

LSTM stands for Long Short Term Memory is one among the modification of Recurrent Neural Network. LSTM come to complite the weaknes of RNN that it's uncapable to makes prediction based on past information stored for a long time.

PROCESS

STEP 1
Identify Problems

STEP 2
Data Preparation

- Checking missing values
- Checking duplicate values
- Cleansing

STEP 3
Data Exploration

- Most used word
- Counts of tweet based on Label

STEP 4
Make Training Models

Make training models:
Neural Networks and LSTM

STEP 5
Evaluation Models

Doing testing on the models to see whether the models is good or not

ABOUT DATA

Data Source	Data Sentimen Twitter https://drive.google.com/file/d/1RCHGfn9JJyyReAh8PIIoF8Ch0H3miP0u/view

Data Size	ROWS : 11000 COLUMNS : 2

Data Variables	Kalimat = String Label = String

WHAT'S IN NEURAL NETWORK AND LSTM DO

Preparing Dataset

STEP 1

In the proses preparing data, first step is import library process is carried out to call the package or library used. Library that we use is:

- Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

STEP2

The process checking missing value is to indicates how many parameter/variable which has null value. This data has no missing value.

STEP3

The process checking duplicate value is to indicates how many parameter/variable which has duplicate value. This data has 67 duplicate value.

	Kalimat	Label
0	warung ini dimiliki oleh pengusaha pabrik tahu...	positive
1	mohon ulama lurus dan k212 mmbri hujjah partai...	neutral
2	lokasi strategis di jalan sumatera bandung . t...	positive
3	betapa bahagia nya diri ini saat unboxing pake...	positive
4	duh . jadi mahasiswa jangan sombong dong . kas...	negative
...	...	...
10928	f - demokrat dorong upaya kemandirian energi n...	neutral
10929	tidak bosan	positive
10930	enak rasa masakan nya apalagi kepiting yang me...	positive
10931	pagi pagi di tol pasteur sudah macet parah , b...	negative
10932	meskipun sering belanja ke yogya di riau junct...	positive

10933 rows × 2 columns

Data of sentiment tweet that we use to analyze has 10933 rows and 2 columns

WHAT'S IN NEURAL NETWORK AND LSTM DO

Preparing Dataset

Checking Missing Values

```
In [4]: df.isnull().sum()

Out[4]: Kalimat    0
        Label      0
        dtype: int64
```

Checking Duplicate Value

```
In [5]: df.duplicated().sum()

Out[5]: 67
```

```
In [6]: df = df.drop_duplicates(ignore_index=True)
        df.duplicated().sum()

Out[6]: 0
```

WHAT'S IN NEURAL NETWORK AND LSTM DO

Cleansing

STEP 1

In the proses normalization/cleansing, first step is import library process is carried out to call the package or library used. Library that we use is:

- NLTK is a leading platform for building Python programs to work with human language data, and text processing libraries for classification, tokenization, and stemming.

STEP2

Before the cleaning process, we use the Alay Dictionary, which contains many words that can be replaced with real words.

The dictionary has 15170 rows and 2 columns are Original and Arti.

STEP3

The cleansing process are:

- Changed uppercase word to lowercase word
- Stemmed text
- Replacement uncommon word to common word
- Removed stopwords except "tidak"
  - "tidak" is combined with the word after it
- Remove unnecessary word, link web, and punctuation
- Remove emoticon

Exploratory Data Analysis

MOST USED WORD

In the overall data it was found that there were words that were most used, are:

1. Makan
2. Enak
3. Restoran
4. Bandung
5. Suka
6. Bagus
7. Menu
8. Orang
9. Pesan
10. Nikmat

Exploratory Data Analysis

MOST USED WORD

In the Positive data it was found that there were words that were most used, are:

1. Makan
2. Enak
3. Restoran
4. Bandung
5. Nyaman
6. Bagus
7. Menu
8. Pas
9. Pesan
10. Nikmat
11. Suasana

Exploratory Data Analysis

MOST USED WORD

In the Negative data it was found that there were words that were most used, are:

1. Makan
2. Mahal
3. Pesan
4. Layan
5. Restoran
6. Orang
7. Kecewa
8. Lihat
9. Suka
10. Pakai
11. Jokowi

Exploratory Data Analysis

MOST USED WORD

In the Neutral data it was found that there were words that were most used, are:

1. Demokrat
2. Kepala
3. Daerah
4. Pilihan
5. Demokrasi
6. Partai
7. Indonesia
8. Perjuangan
9. Anies
10. Baswedan
11. Gubernur

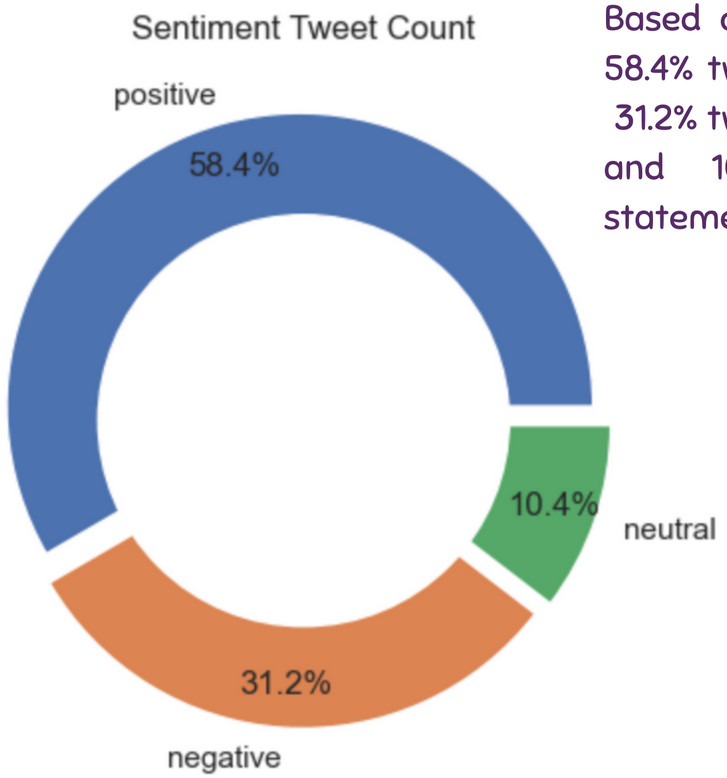
Exploratory Data Analysis

TOP 10 MOST USED WORD

	word	count
0	makan	6457
1	enak	3622
2	harga	2016
3	bandung	1748
4	menu	1718
5	restoran	1289
6	layan	1264
7	pilih	1093
8	jalan	986
9	suasana	986

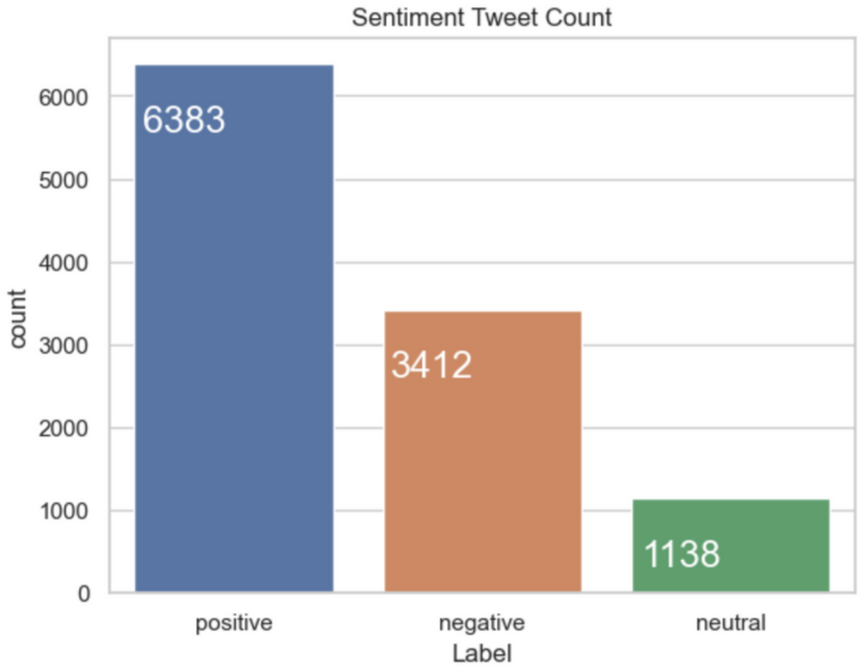
Exploratory Data Analysis

COUNT OF SENTIMENT LABEL



Graph 1.
Donut Chart of Sentimen Tweet Count

Based on Graph 1, we know that 58.4% tweet is positive statement, 31.2% tweet is negative statement, and 10.4% tweet is neutral statement



Graph 2.
Bar Chart of Sentimen Tweet Count

Based on Graph 2 we know that 6,383 tweet is positive statement, 3,412 tweet is negative statement, and 1,138 tweet is neutral statement

WHAT'S IN NEURAL NETWORK AND LSTM DO

Feature Extraction TF-IDF

STEP 1

first step in the process Feature Extraction use TF-IDF method is import library that :

- Pickle : library that is used for weighting each word that is owned by using the verctorization method, the weighting results are stored and read into or from a file with .pkl format.
- Scikit-Learn (Sklearn) : an open source data analysis library, and the gold standard for Machine Learning (ML) in the Python ecosystem. Key concepts and features include: Algorithmic decision-making methods, including: Classification: identifying and categorizing data based on patterns.

STEP 2

In the feature extraction step we take only column "Kalimat Bersih" that output from cleansing process. That process use TF-IDF method in this case we use TfidfVectorizer() function to stored weighting each word. After that, we save into .pickle format file.

WHAT'S IN NEURAL NETWORK AND LSTM DO

Preparing Train and Test Dataset

Preparing train and test dataset with 20% data is data test and 80% data is data train

Training Neural Network

Experiments

DEFAULT	PARAMETERS					
	hidden_layer_sizes	max_iter	activation	solver	alpha	learning_rate
Value	(100,)	200	relu	adam	0.0001	constant

WHAT'S IN NEURAL NETWORK AND LSTM DO

Training Neural Network

Experiments

1	PARAMETERS					
	hidden_layer_sizes	max_iter	activation	solver	alpha	learning_rate
Value	(5,2),(100,50)	50, 100, 150	tanh, relu	sgd, adam	0.0001, 0.05	constant, adaptive

Best Parameters :

'activation': 'tanh', 'alpha': 0.05, 'hidden_layer_sizes': (5, 2), 'learning_rate': 'adaptive', 'max_iter': 50, 'solver': 'adam'

WHAT'S IN NEURAL NETWORK AND LSTM DO

Training Neural Network

Experiments

2	PARAMETERS					
	hidden_layer_sizes	max_iter	activation	solver	alpha	learning_rate
Value	(5,2),(100,50)	200	tanh, relu	sgd, adam	0.0001, 0.05	constant, adaptive

Best Parameters :

'activation': 'relu', 'alpha': 0.05, 'hidden_layer_sizes': (100,50), 'learning_rate': 'adaptive', 'max_iter': 200, 'solver': 'adam'

WHAT'S IN NEURAL NETWORK AND LSTM DO

Training Neural Network

Experiments

3	PARAMETERS					
	hidden_layer_sizes	max_iter	activation	solver	alpha	learning_rate
Value	(5,2),(100,50)	200	identity, logistic, tanh, relu	sgd, Adam, lbfgs	0.0001, 0.05	constant, adaptive, invscaling

Best Parameters :

'activation': 'logistic', 'alpha': 0.0001 , 'hidden_layer_sizes': (100,50), 'learning_rate': 'invscaling', 'max_iter': 200, 'solver': 'lbfgs'

WHAT'S IN NEURAL NETWORK AND LSTM DO

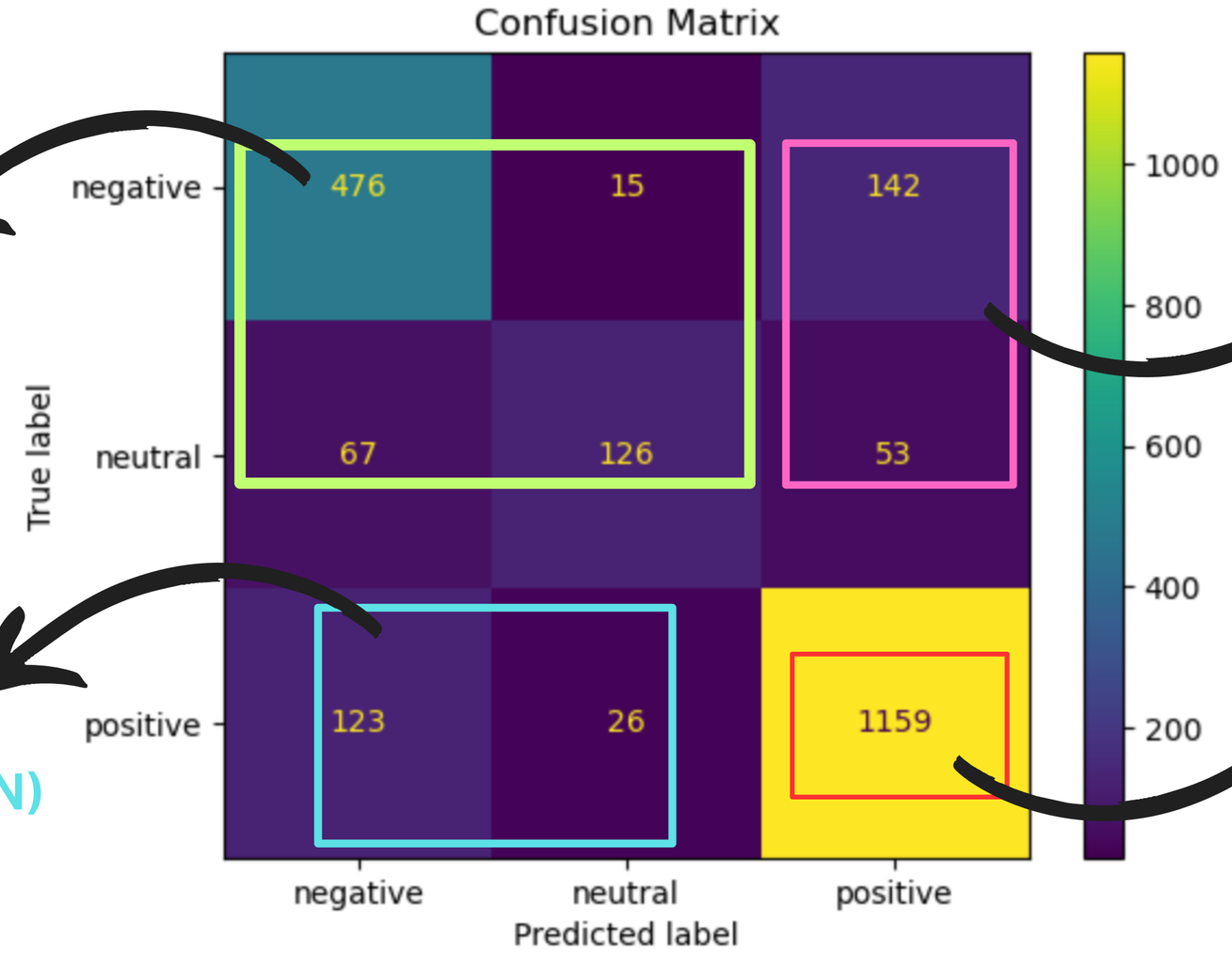
Cross Validation Neural Network

Experiments	VALUES SCORE (AVG)			
	Accuracy	F1	Precision	Recall
Default	80.7828 %	74.8147 %	72.9948 %	77.3799 %
1	84.3041 %	79.1506 %	78.1269 %	80.5476 %
2	82.7037 %	77.5774 %	76.7548 %	78.5882 %
3	83.7647 %	77.9838 %	77.9565 %	78.0559 %

WHAT'S IN NEURAL NETWORK AND LSTM DO

Confession Matrix

POSITIF



True Negative (TN)
(476+15+67+126)

False Positive (FP)
(142+53)

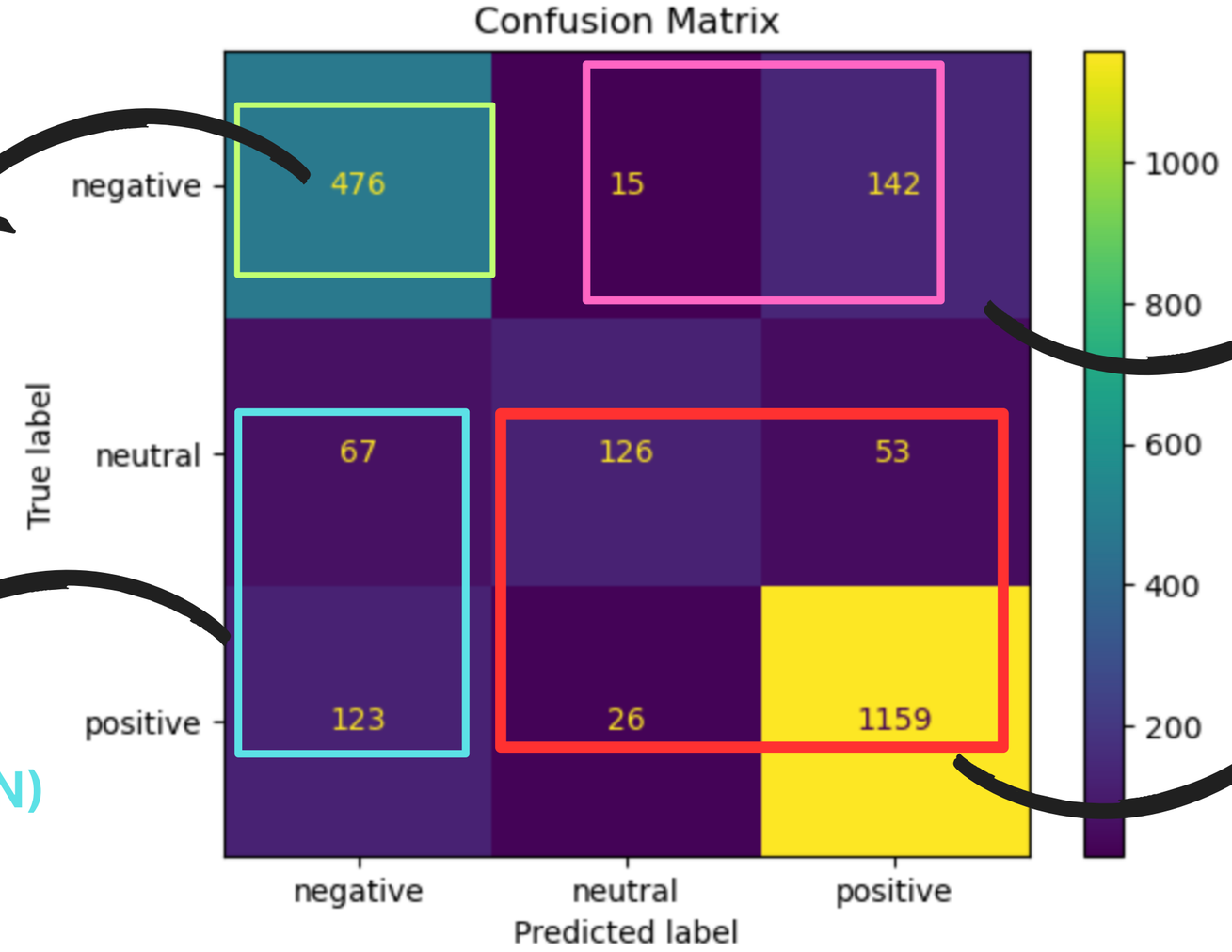
False Negative (FN)
(123+26)

True Positive (TP)
(1159)

WHAT'S IN NEURAL NETWORK AND LSTM DO

NEGATIVE

Confession Matrix



True Negative (TN)
(476)

False Positive (FP)
(142+15)

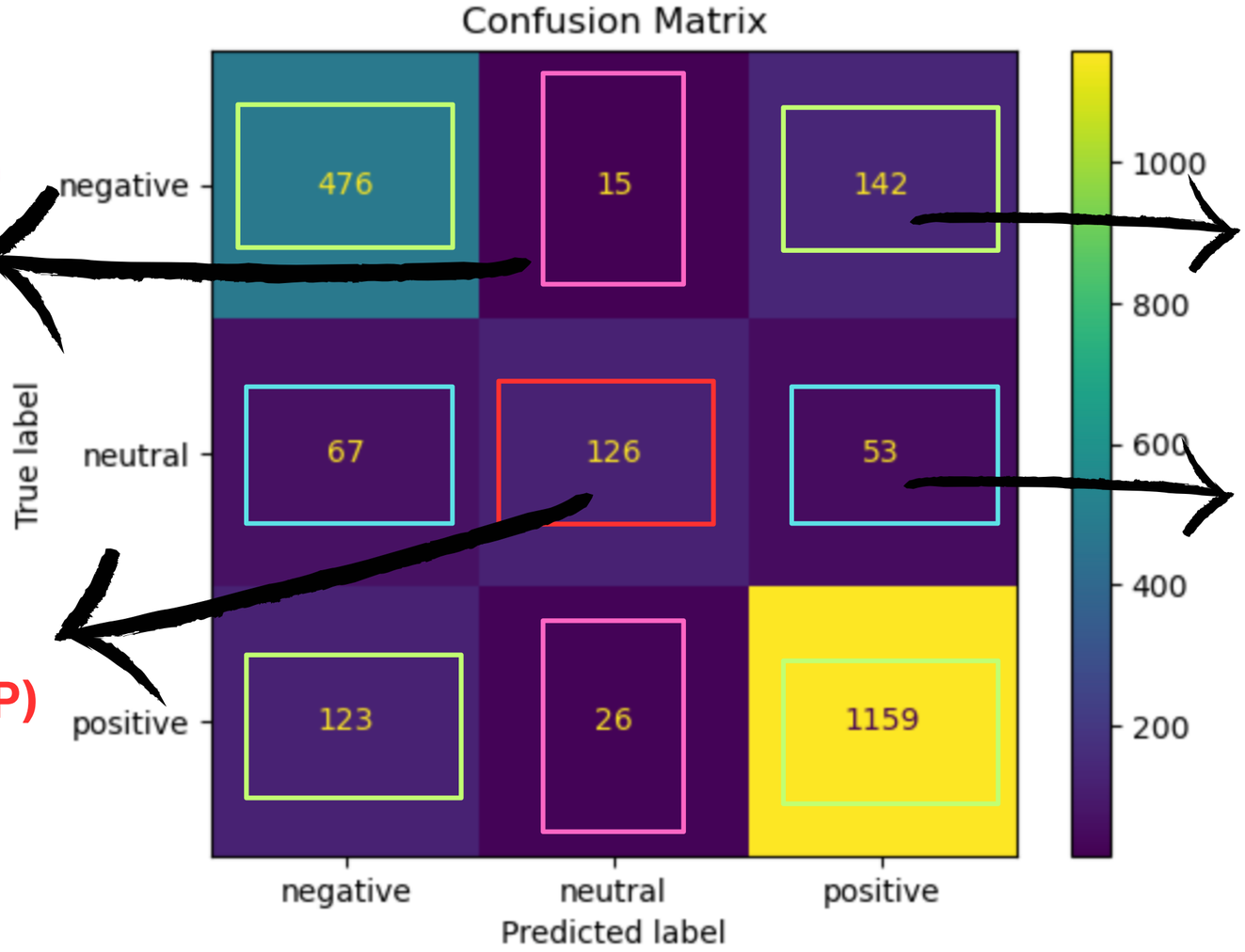
True Positive (TP)
(1159+126+53+26)

False Negative (FN)
(123+67)

WHAT'S IN NEURAL NETWORK AND LSTM DO

NEUTRAL

Confession Matrix



False Positive (FP)
(15+26)

True Negative (TN)
(476+142+123+1159)

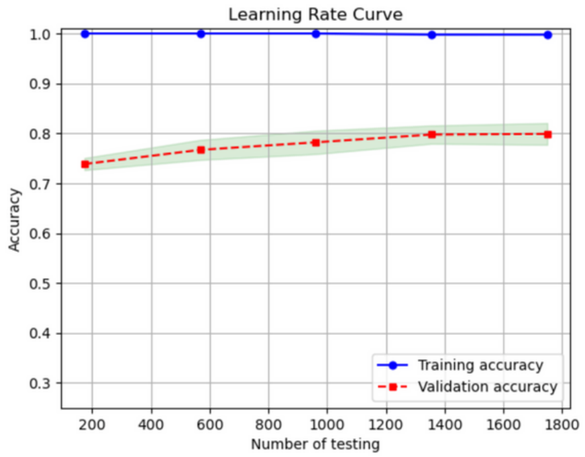
False Negative (FN)
(67+53)

True Positive (TP)
(126)

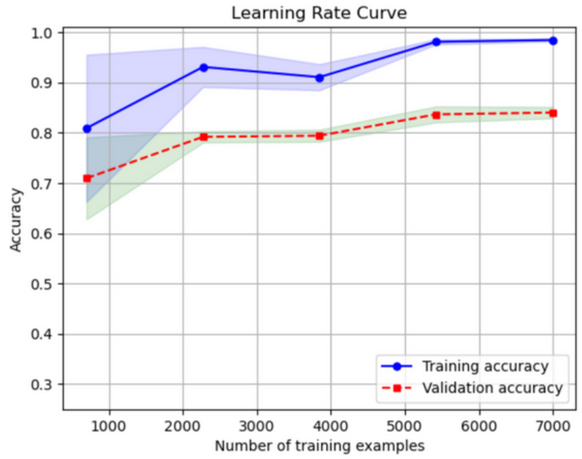
WHAT'S IN NEURAL NETWORK AND LSTM DO

Learning Curve Neural Networ

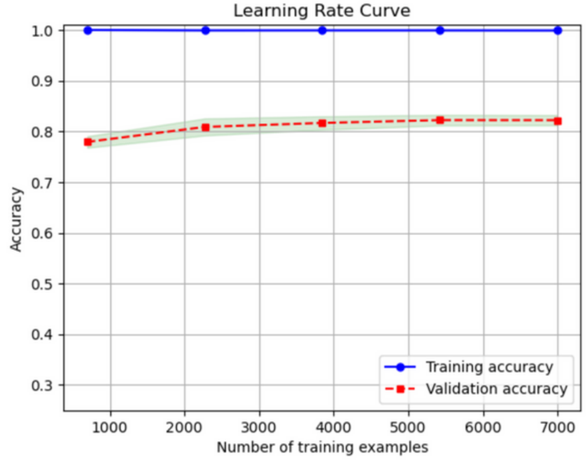
Default



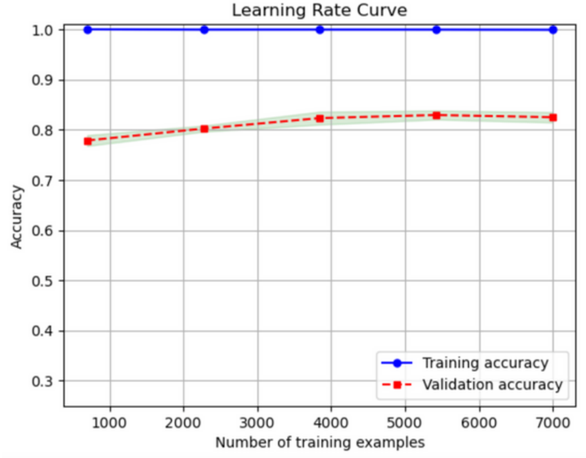
1st



2nd



3rd



Under – Fitting

WHAT'S IN NEURAL NETWORK AND LSTM DO

Feature Extraction LSTM

In Feature extraction we use Tokenizer and pad_sequences modules from Tensorflow. The tokenizer converts text into a sequence of integers or into a vector where the coefficients for each token can be binary based on the number of words based on tf-idf. Pad Sequences converts a list of sequences into a 2D vector/array so that models can be processed

WHAT'S IN NEURAL NETWORK AND LSTM DO

Training LSTM

Hyperparameter	Experiments		
	I	II	III
Max Features	10000	50000	100000
Embed dim	32	16	32
Units	32	16	32
Batch size	64	128	256

WHAT'S IN NEURAL NETWORK AND LSTM DO

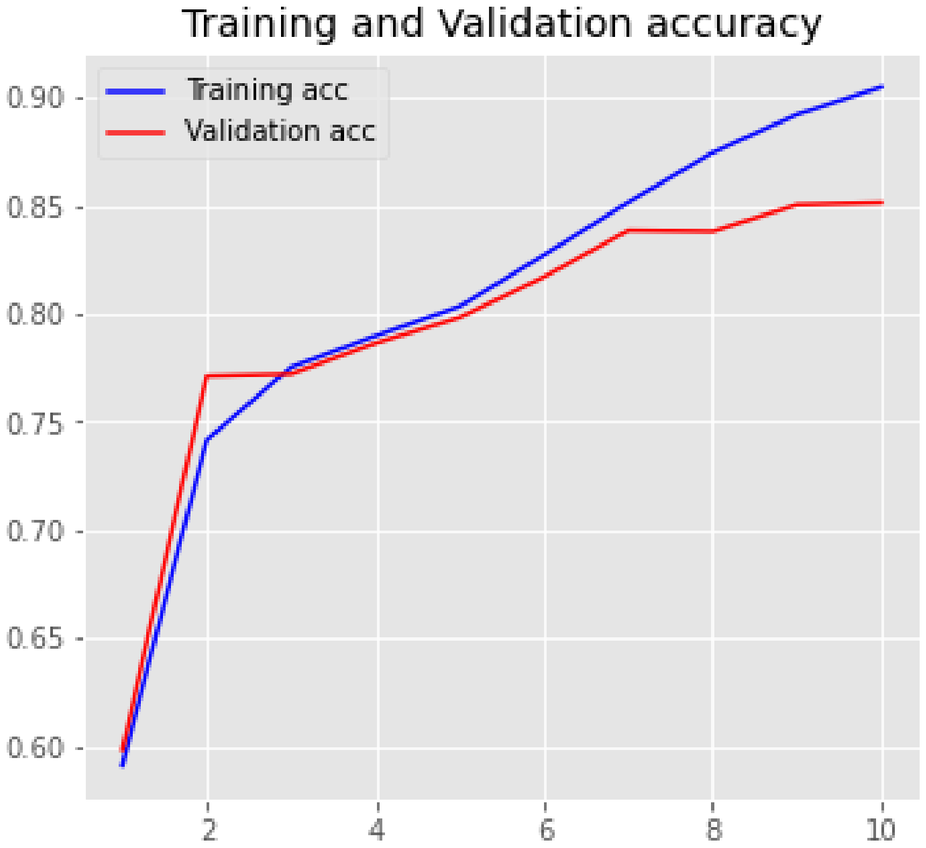
Evaluation Menggunakan Cross Validation

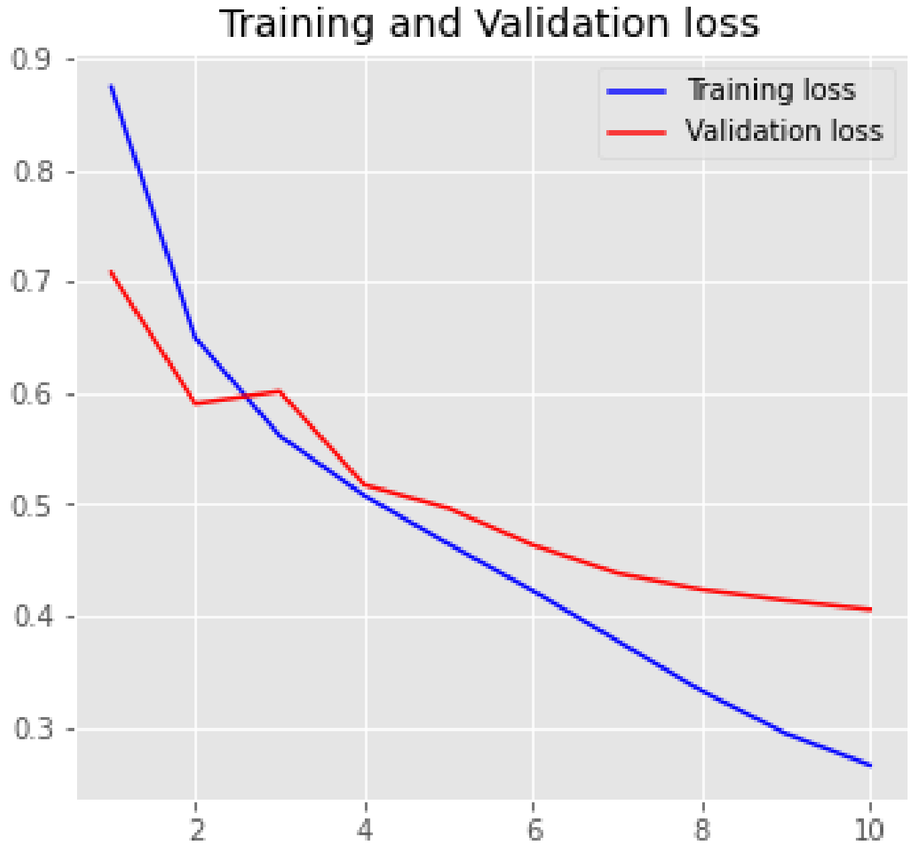
Result Hyperparameter	Experiments		
	I	II	III
AVG accuracy	0.8642	0.8606	0.8434
AVG F1 Score	0.8250	0.8186	0.7825
AVG precision	0.8235	0.8226	0.8137
AVG Recall	0.8271	0.8161	0.7770

Based on the experiments, the best accuracy was obtained in the first experiment with max feature = 1000, embed dim 32, units = 24 and batch zise = 64 which resulted in the best accuracy of 0.8642

WHAT'S IN NEURAL NETWORK AND LSTM DO

Curve LSTM





Under-Fitting

WHAT'S IN NEURAL NETWORK AND LSTM DO

Predict

NEURAL NETWORK


```
{
  "data": {
    "sentiment": "positive",
    "text": "betapa bahagia unboxing paket barang bagus tetap beli"
  },
  "description": "Result of Sentiment Analysis using NN",
  "status_code": 200
}
```

```
{
  "data": {
    "sentiment": "negative",
    "text": "aduh mahasiswa sombong kasih kartu kuning ajar usah politik selesai kuliah politik telat dasar mahasiswa"
  },
  "description": "Result of Sentiment Analysis using NN",
  "status_code": 200
}
```

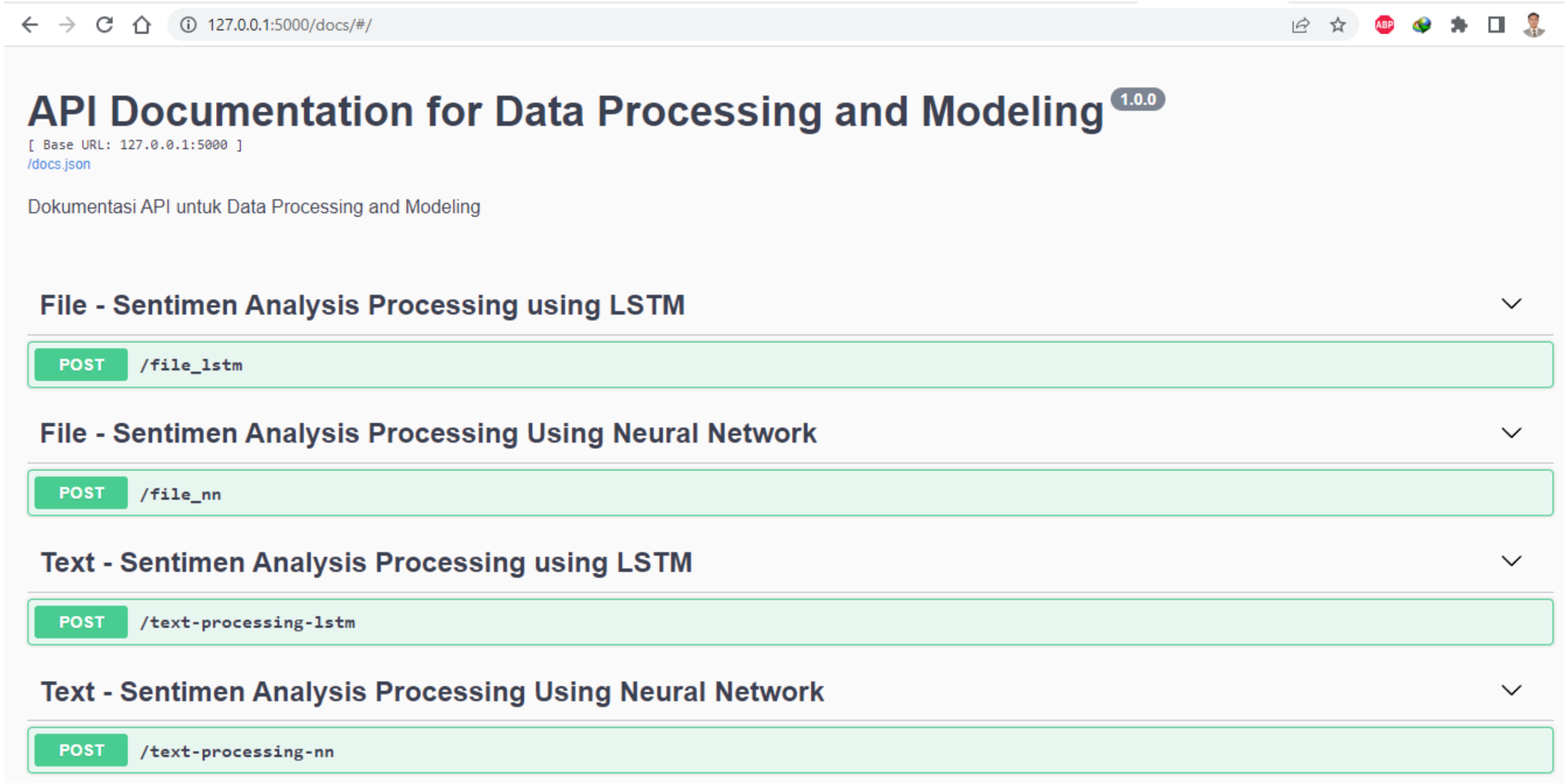

LSTM


```
{
  "data": "betapa bahagia unboxing paket barang bagus tetap beli",
  "description": "Original Teks",
  "sentiment": "negative",
  "status_code": 200
}
```

```
{
  "data": "aduh mahasiswa sombong kasih kartu kuning ajar usah politik selesai kuliah politik telat dasar mahasiswa",
  "description": "Original Teks",
  "sentiment": "negative",
  "status_code": 200
}
```

WHAT'S IN NEURAL NETWORK AND LSTM DO

Deployment Result API



SUMMARY

EDA

- Sentiment that dominate is positive.
- Top 5 the most word used is Makan, enact, harga, Bandung, and menu.
- Topic that twitter users tweet about culinary too dominate.

Neural Network

- By use hyperparameter tunning can make model has better score accuracy, F1, precision and recall value.
- Model using NN are still underfitting. Because the data is not good enough to build a model and analyzing sentiment of sentences.

LSTM

The process that take place in each part of LSTM are:

1. input layer where text data is conditioned to become a 100 dimentional matrix
2. unit layer where the number of neurons is set at 16, the unit layer uses the dropout feature to minimize overfitting
3. the fully connected layer is a place that conditions the results from the unit layer to become input for multilayer. AT this stage softmax.

Underfitting model results can be caused by the use of inappropriate hyperparameter tuning

SUMMARY

Based on result we know that :

- The best model to use is LSTM , becuase of the AVG accuracy score, F1 score, precision score, and recall score from the LSTM model is more better than Neural Network model.
- Hyperparameter of Neural Network with best score are based on result:
  - activation = tanh
  - alpha = 0.05
  - hidden_layer_sizes = (5, 2)
  - learning_rate = adaptive
  - max_iter = 50
  - solver = adam
- Hyperparameter of LSTM with best score are based on result:
  - Max Features = 10000
  - Embed dim = 32
  - Units = 32
  - Batch size =64
- The quantity of hyperparameters does not guarantee the model will be good. The choice of hyperparameters depends on the data and the experience of a data scientist.
- The choices of hyperparameter process by triall and eror.

THANK YOU!

Any questions?