Platinum Level Challenge

Tweet Sentiment Analysis

Sentiment API Building and Data Analysis

By :

Aldimeola Alfarisy

Raafiandy Ghani

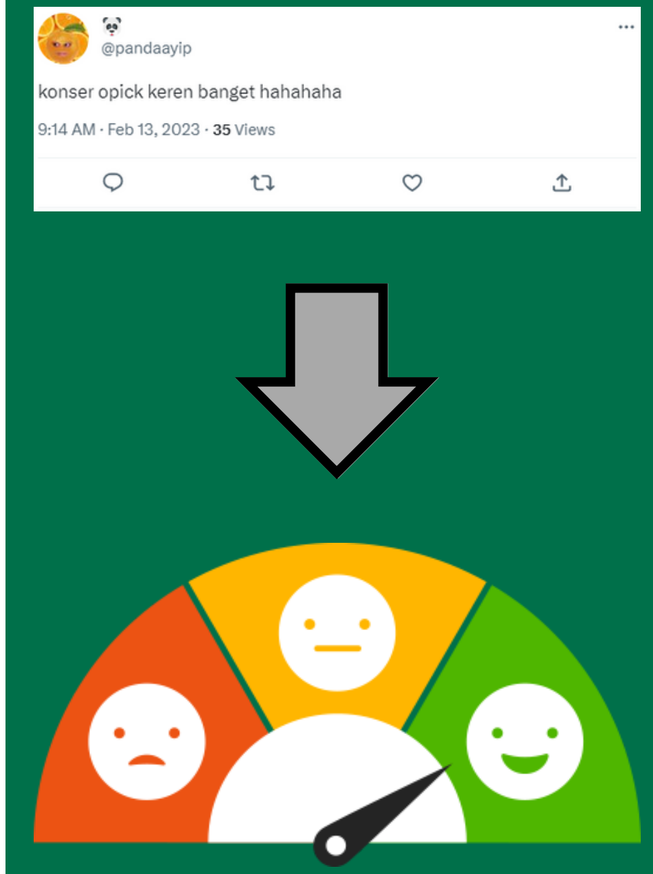
PRELIMINARY

BACKGROUND

Sentiment is an opinion or view of something. Meanwhile, sentiment analysis itself is the process of analyzing any digital text to determine whether the emotional tone of the message is positive, negative, or neutral, including text on social media (twitter).

The application of sentiment analysis is not limited to social media. Many companies apply sentiment analysis to improve their products or services based on specific customer reviews.

So, we try to analyze the sentiments in the existing tweets and create a system (API) that can classify each tweet sentiment. We use deep learning methods (RNN and LSTM) in this sentiment analysis.



OBJECTIVES

- Identify the distribution of each positive, neutral, and negative tweet sentiments
- Get the best performing model used to predict sentiment
- Creating an engine/API that can classify the given sentiments

RESEARCH METHODS

DATA PREPARATION

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11000 entries, 0 to 10999
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Tweet   11000 non-null  object
 1   Label   11000 non-null  object
dtypes: object(2)
memory usage: 172.0+ KB
```

```
# Duplicated data check

print('There are {} duplicated data'.format(df.duplicated().sum()))
```

```
There are 67 duplicated data
```

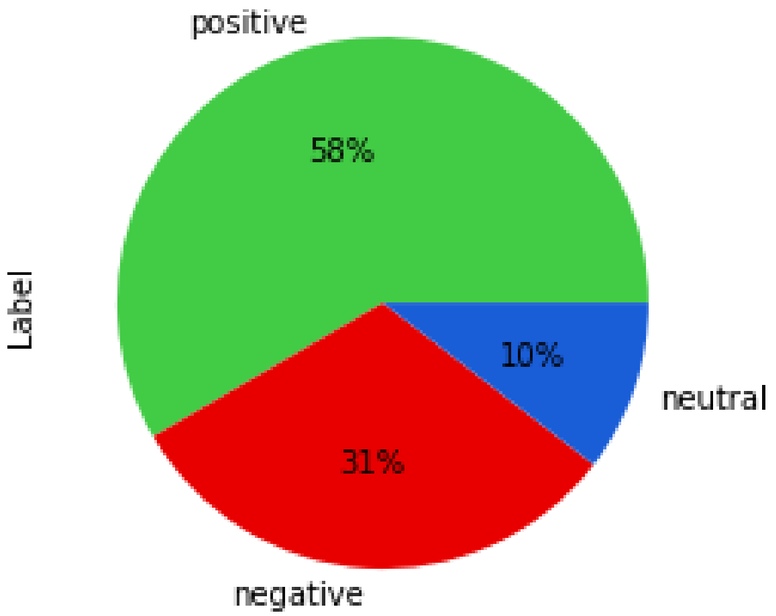
```
# Drop duplicated data

df = df.drop_duplicates()
print('There are {} duplicated data'.format(df.duplicated().sum()))
print('Duplicated data already dropped')
```

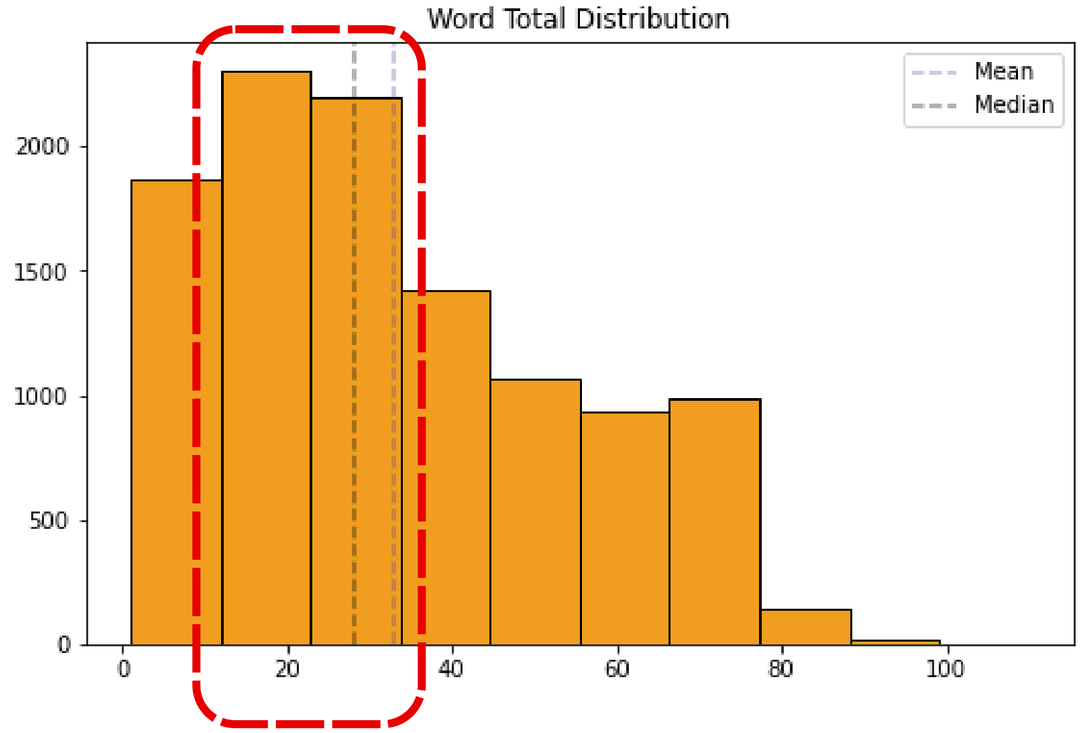
```
There are 0 duplicated data
Duplicated data already dropped
```

- The dataset consists of 11,000 rows and 2 columns containing tweets in Indonesian and sentiment labels for each tweet
- There are no missing values in the dataset
- There are 67 duplicate data and have been removed (total data becomes 10933 data)

RESEARCH METHODS

EXPLORATORY DATA ANALYSIS





Positive Sentiments = 6383 tweets

Negative Sentiments = 3412 tweets

Neutral Sentiments = 1138 tweets

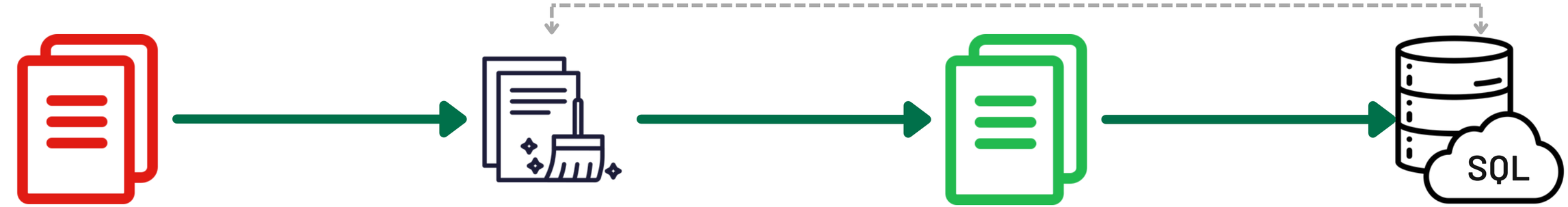
Majority = 20 - 30 words

Mean = 33 words

Median = 28 words

RESEARCH METHODS

TEXT NORMALIZATION



Original Data

Text Processing

- Text cleansing using RegEx library
- Replace non standard words with the help of 'kamusalay' dictionary from database

Clean Data

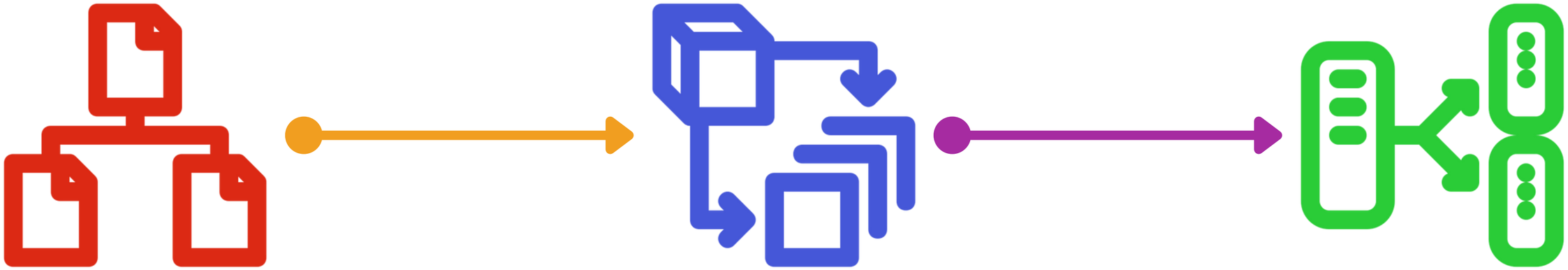
Data Store

TWEET COMPARISON

```
Old tweet: pdip sebut ridwan kamil menang karena berbaju merah

New tweet: partai demokrasi indonesia perjuangan sebut ridwan kamil menang karena berbaju merah
```

RESEARCH METHODS

MACHINE LEARNING PREPARATION



Feature-Label Classification

Separation of features (X) and labels (Y) from the dataset

Feature Extraction

Fetch features and transform text data into vector numbers (Tokenizer and Pad Sequences)

Train-Test Data Split

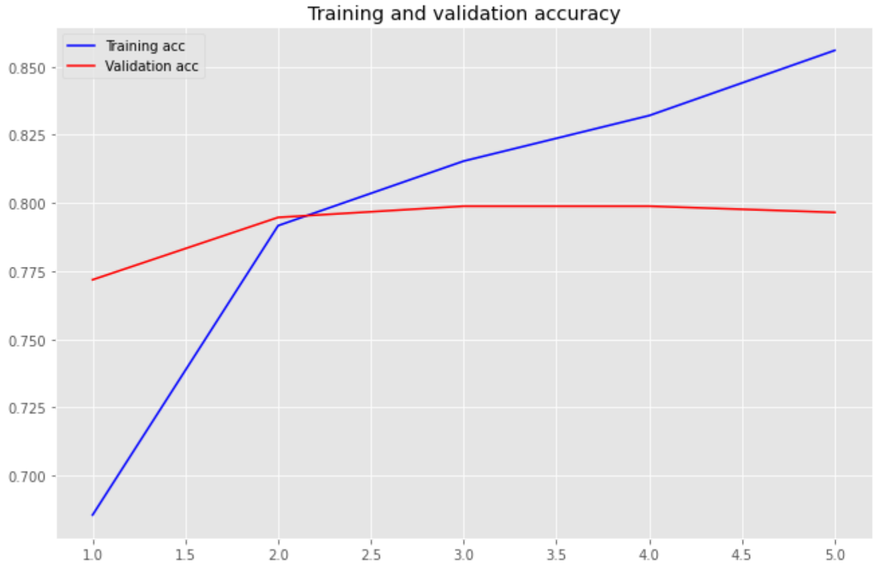
Separation of data into training data (80%) and test data (20%)

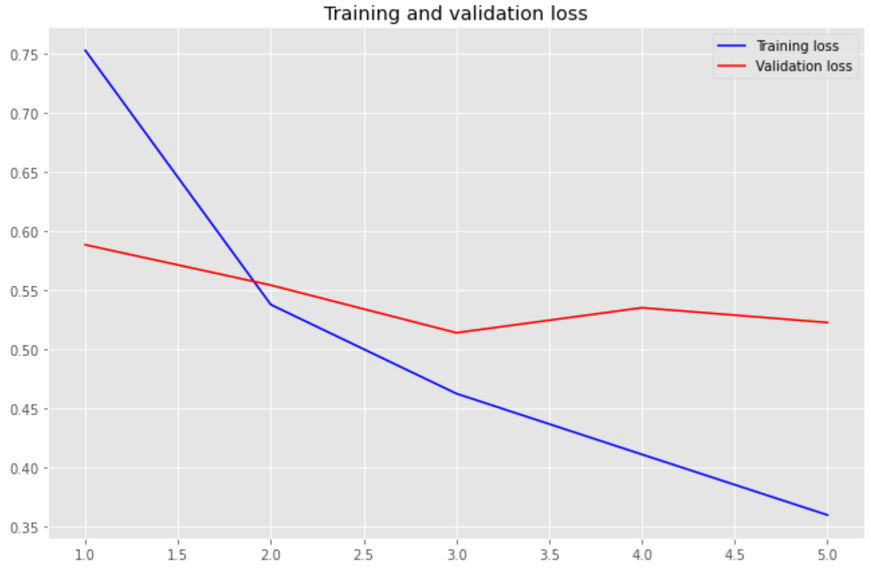
RESEARCH RESULT

MODEL TRAINING & EVALUATION (RNN)

Parameter	
Input Layer	64
Output Layer	3
Activation	Softmax
Learning Rate	0.0005
Epoch	50
Batch Size	32

Regularization	
Dropout	0.8
Early Stopping	
• Monitor	Val_loss
• Mode	min
• Verbose	1
• Patience	2





Cross Validation : 5

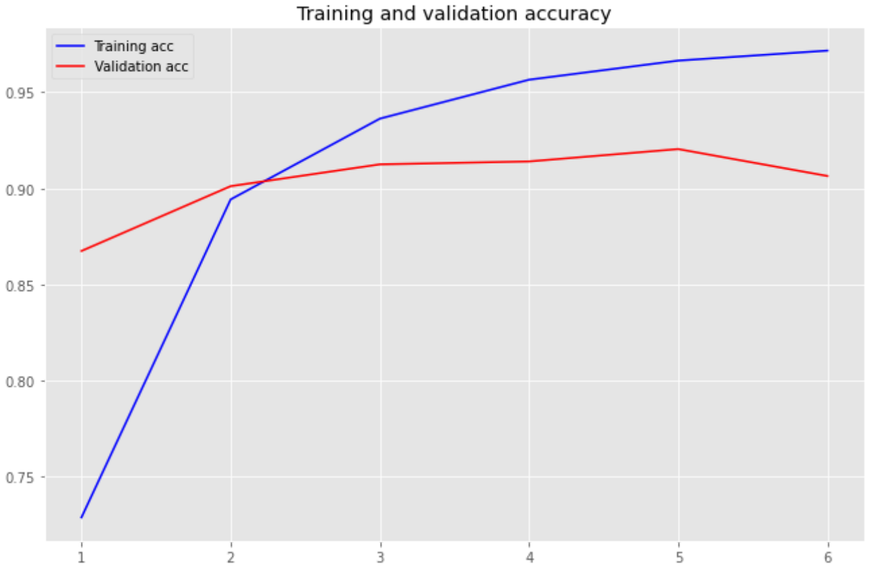
- It can be seen that the performance with minimal loss occurred in the second epoch
- Majority model training stopped after fifth epoch

RESEARCH RESULT

MODEL TRAINING & EVALUATION (LSTM)

Parameter	
Input Layer	64
Output Layer	3
Activation	Softmax
Learning Rate	0.0005
Epoch	50
Batch Size	32

Regularization	
Dropout	0.5
Early Stopping	
• Monitor	Val_loss
• Mode	min
• Verbose	1
• Patience	2

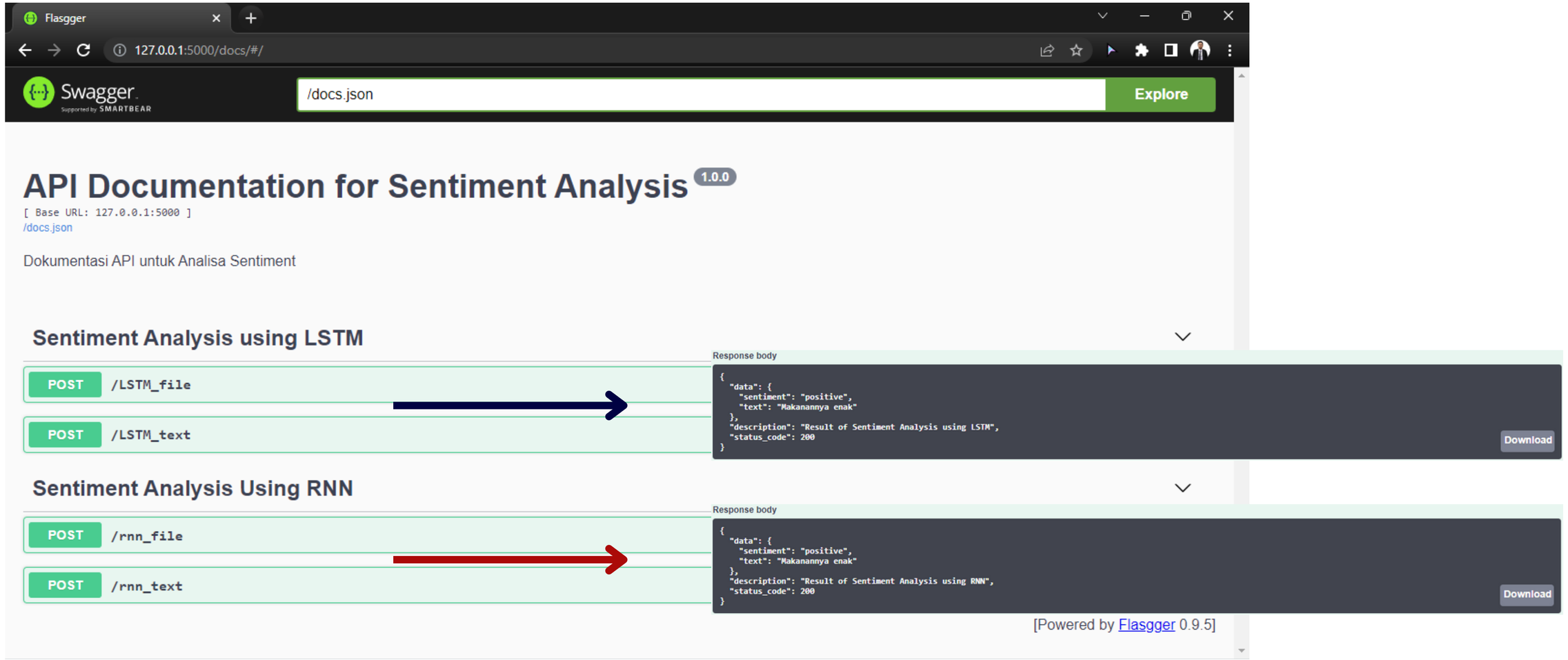


Cross Validation : 5

- It can be seen that the performance with minimal loss occurred in the second epoch
- Majority model training stopped after sixth epoch

RESEARCH RESULT

PREDICT & DEPLOYMENT



RESEARCH RESULT

RESULT

- The dataset consists of positive, neutral, and negative sentiment tweets with a distribution for positive label are 6,383 (58%) tweets, negative label are 3,412 (31%) tweets, and neutral label are 1,138 (10%) tweets
- LSTM model has better performance than RNN model, even though the two models tend to be slightly overfit
- The API created has 2 endpoints for each model (to process text and file data) and can give positive, negative or neutral label based on the sentiment provided (although the results given have errors for some sentiments)

CONCLUSION AND RECOMMENDATION

Deep learning is a powerful method because it can learn complex patterns from data and provide more accurate results than traditional machine learning methods. Besides that, deep learning can also be applied to various kinds of data (e.g. images, sounds, signals, etc.). However, deep learning has a drawback, namely the need for very large data. This will be a problem if the available data is limited because it tends to produce an overfit model. In addition, deep learning is computationally expensive, because deep learning requires powerful hardware to be able to quickly train models.

Finally, our recommendation is that deep learning must be used wisely. If the data you have is still relatively simple and relatively small, it would be wiser to use a simpler algorithm (traditional machine learning). Use deep learning if it's really necessary (when the data we have is very large and very complex).

FOR DOCUMENTATION

GitHub Repository

THANK YOU

CONTACTS :

Aldimeola Alfarisy

Raafiandy Ghani