# APPENDIX

## CRAWLING CODING

```
import tweepy
import csv
from tweepy import OauthHandler
import string

consumer_key = 'Gq6DQpjxV4tMo3BH2zZmmD51T'
consumer_secret = 'Bg7XsPSmnlWyKeZLtZrsjW34WdULkGcUWHxwACIwa1J4exDlsK'
access_token = '1456889597883912228-CePiSfi4SZWXVGPOkDRLa9n8WDxKJc'
access_secret = 'yTNTjituna3Xa2pEAaatbhgmgt8GRbcW5nh8GiyM68WP5'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)
api = tweepy.API(auth,wait_on_rate_limit=True)

csvFile = open('/content/sample_data/hasilcrawling.csv', 'a', encoding='utf-8')
csvWriter = csv.writer(csvFile)

from google.colab import drive
drive.mount('/content/drive')

for tweet in tweepy.Cursor(api.search,q="vaksin, covid", lang="id", since="2021-11-01").items(10000):
print (tweet.created_at,tweet.author.screen_name,tweet.text)
csvWriter.writerow([tweet.created_at, tweet.author.screen_name, tweet.text])
```

## PRE-PROCESSING

```
!pip install Sastrawi
!pip install nltk

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer

#download the data
import nltk
nltk.download('stopwords')

#using the stopwords
From nltk.corpus import stopwords

#initialize the stopwords
```

```
Stoplist = stopwords.words('indonesian')

import string
import re
from Sastrawi.StopWordRemover.SopWordRemoverFactory import StopWordRemoverFactory
import csv

From google.colab import drive
Drive.mount('/content/drive')

df = pd.read_csv('/content/drive/Mydrive/coba/datacrawling.csv', encoding = 'unicode_escape')

df.head()

Print('Dataset size:',df.shape)
Print('Columns are:',df.columns)

#case folding
df['TWEET'] = df['tweet'].str.lower()
df.head()

def remove_punct(text):
    #text = "".join([char for char in text if char not in string.punctuation])
    text = re.sub(r'[^a-zA-Z0-9]', ' ', str(text))
    text = re.sub(r'\b\w{1,2}\b', '', text) #menghilangkan 2 kata
    text = re.sub(r'\s\s+', ' ', text)
    text = re.sub(r"\d+", "", text)
    return text
df['tweet_clean'] = df['TWEET'].apply(lambda x: remove_punct(x))
df.head()

#tokenization
def tokenization(text):
    text = re.split('\W+', text)
    return text

df['TOKENIZATION'] = df['tweet_clean'].apply(lambda x: tokenization(x.lower()))
df.head()

#filtering stopwords removal
Stopword = nltk.corpus.stopwords.words('indonesian')

def remove_stopwords(text):
    text = [word for word in text if word not in stopword]
    return text

df['STOP_REMOVAL'] = df['TOKENIZATION'].apply(lambda x: remove_stopwords(x))
```

```
df.head()

Stop_removal = df[['STOP_REMOVAL']]

def fit_stopwords(text):
    text = np.array(text)
    text = ' '.join(text)
    return text

df['STOP_REMOVAL'] = df['STOP_REMOVAL'].apply(lambda x: fit_stopwords(x))
df.head()

df.drop('tweet',axis=1)

df.to_csv('/content/drive/MyDrive/dataskripsi/hasilpreprocessing.csv', sep= ',' , encoding='utf-8')
```

## CODING SVM

```
from flask import Flask, render_template, url_for
import numpy as np
import pandas as pd
import csv
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import svm
from sklearn.metrics import accuracy_score

# Packages for visuals
import matplotlib.pyplot as plt
import seaborn as sns; sns.set(font_scale=1.2)

#labelling
import pandas as pd
from textblob import TextBlob

from google.colab import drive
drive.mount('/content/drive')
```

```

df = pd.read_csv('/content/drive/MyDrive/dataskripsi/hasilpreprocessing.csv', encoding = 'unicode_escape')
df['label'] = ''
for i,x in df.tweet_akhir.iteritems():
label = TextBlob(x)
df['label'][i] = label.sentiment.polarity
print("Index: ",i, "label", label.sentiment.polarity)

def polarity_to_label(x):
    if(x >= -1 and x < 0):
      return 'negatif'
    if(x == 0):
      return 'neutral'
    if(x > 0 and x <=1):
      return 'positif'
df.label = df.label.apply(polarity_to_label)
df

df.to_csv('/content/drive/MyDrive/dataskripsi/hasillabelling.csv', sep=',', encoding='utf-8')

# split data
# Split into train and test data
Train_X, test_X, train_X, test_Y=model_selection.train_test_split (df[' tweet_akhir '], df['label'], test_size = 0.1, random_state = 0)
# state = 0 here there is no randomization on the split data which means the order is still the same

df_train = pd.DataFrame()
df_train[' tweet_akhir '] = train_X
df_train['label'] = train_Y

df_test = pd.DataFrame()
df_test[' tweet_akhir '] = test_X
df_test['label'] = test_Y
```

```
df_train

df_test

df_train.to_csv(r"/content/drive/MyDrive/dataskripsi/df_train.csv")
df_test.to_csv(r"/content/drive/MyDrive/dataskripsi/df_test.csv")

#TF-IDF to analyze the relationship between a phrase or sentence and a set of documents

from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vect = TfidfVectorizer(max_features = 5000)
tfidf_vect.fit(df[' tweet_akhir '])
train_X_tfidf = tfidf_vect.transform(df_train['STOP_REMOVAL'])
test_X_tfidf = tfidf_vect.transform(df_test['STOP_REMOVAL'])

tfidf_vect

print(train_X_tfidf)

print(test_X_tfidf)

print(train_X_tfidf.shape)
print(test_X_tfidf.shape)

#the syntax below is used to see the learned vocabulary of the corpus
print(tfidf_vect.vocabulary_)

# training process
from sklearn.svm import SVC

model = SVC(kernel='linear')
model.fit(train_X_tfidf,train_Y)

```

```
# testing process
from sklearn.metrics import accuracy_score

predictions_SVM = model.predict(test_X_tfidf)
test_prediction = pd.DataFrame()
test_prediction[' tweet_akhir '] = test_X
test_prediction['label'] = predictions_SVM
SVM_accuracy = accuracy_score(predictions_SVM, test_Y)*100
SVM_accuracy = round(SVM_accuracy,1)

test_prediction

test_prediction.to_csv(r"/content/drive/MyDrive/dataskripsi/test_prediction.csv”)

SVM_accuracy

#Accuracy, Precision, Recall, f1-score

from sklearn.metrics import classification_report

print ("\nLaporan klasifikasi:")
print (classification_report(test_Y, predictions_SVM))

#Bar plot for commentary analysis of covid vaccine
labels = ['negatif', 'neutral', 'positif']
Category1 = [80, 9227, 674]
Plt.bar(labels, Category1, tick_label=Labels, width=0.5, color['navy', 'c', 'blue'])
plt.xlabel('Kelas Sentimen')
plt.ylabel('Data')
plt.title('Diagram Bar Data Analisis Sentimen')
plt.savefig(r"/content/drive/MyDrive/dataskripsi/bar_data.png")
plt.show()

#Bar plot untuk train set
```

```
labels = ['negatif', 'neutral', 'positif']
Category1 = [72, 8318, 592]
Plt.bar(labels, Category2, tick_label=Labels, width=0.5, color['navy', 'c', 'blue'])
plt.xlabel('Kelas Sentimen')
plt.ylabel('Data')
plt.title('Diagram Bar Data Latih')
plt.savefig(r"/content/drive/MyDrive/dataskripsi/bar_datalatih.png")
plt.show()

#Bar plot untuk test set
labels = ['negatif', 'neutral', 'positif']
Category1 = [8, 909, 82]
Plt.bar(labels, Category3, tick_label=Labels, width=0.5, color['navy', 'c', 'blue'])
plt.xlabel('Kelas Sentimen')
plt.ylabel('Data')
plt.title('Diagram Bar Data Uji')
plt.savefig(r"/content/drive/MyDrive/dataskripsi/bar_datauji.png")
plt.show()

#Bar plot untuk klasifikasi dengan SVM
labels = ['negatif', 'neutral', 'positif']
Category1 = [5, 921, 73]
Plt.bar(labels, Category4, tick_label=Labels, width=0.5, color['navy', 'c', 'blue'])
plt.xlabel('Kelas Sentimen')
plt.ylabel('Data')
plt.title('Diagram Bar pada Klasifikasi dengan Support Vector Machine')
plt.savefig(r"/content/drive/MyDrive/dataskripsi/bar_svm.png")
plt.show()
```

PLAGIARISM CHECK.ORG

# 0.78% PLAGIARISM APPROXIMATELY

# Report #14353461

INTRODUCTION Background The covid vaccine has been implemented in early 2021. But until now there are still many people who have not been vaccinated. some people refuse to be vaccinated because they think that covid doesn't exist, are worried about side effects after vaccines, and don't believe in certain vaccines. Nowadays people are free to express their opinions and opinions on social media. Twitter is one of the social media that is often used to express opinions. Currently, many Twitter users provide opinions or comments about vaccines in Indonesia. But the tweets that appear are still arranged randomly, making it difficult for readers to find out more negative comments or appear on the topic of this covid vaccine. Based on the explanation above, a sentiment analysis research was conducted on Twitter to classify Twitter user comments about the Covid vaccine in Indonesia. To conduct this research, firstly, a data crawling process will be carried out whose function is to retrieve tweet data, after that the data will be pre-processed, then