Google Colaboratory jupyter py scikit learn seaborn

UNIVERSITAS JANABADRA

ANALISIS SENTIMEN

MASYARAKAT INDONESIA PADA MEDIA SOSIAL TWITTER TERHADAP PROGRAM PRAKERJA DENGAN METODE NAIVE BAYES CLASIFFIER (BERNOULI)

pandas Sastrawi NLP Bahasa Indonesia NLTK Natural Language Toolkit Bag of Words (BOW)

Data Privacy Infographics

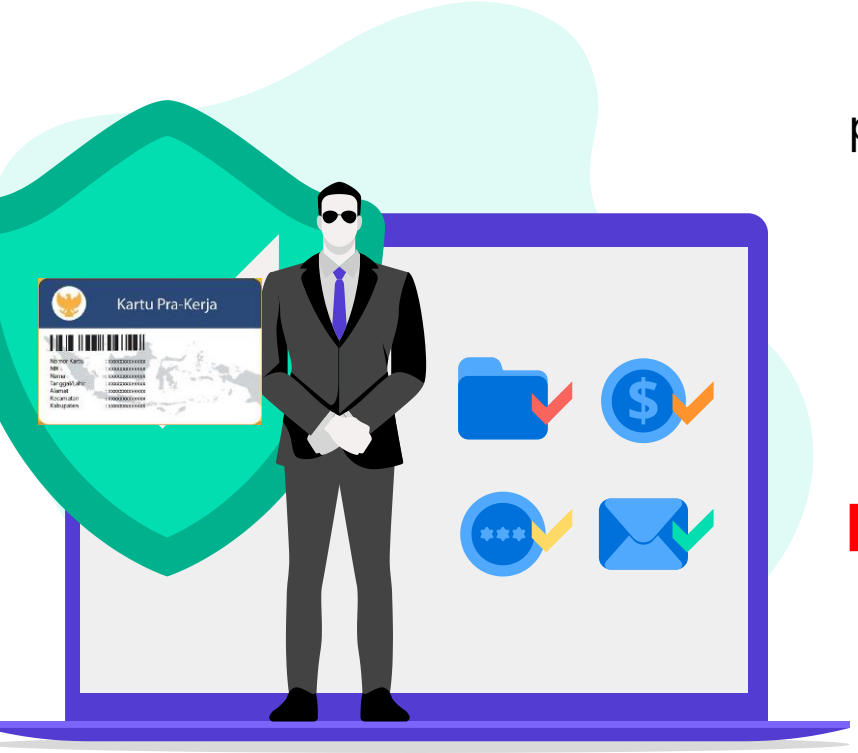
Taofik Krisdiyanto

Mahasiswa

Jemmy Edwing Borong, S.Kom., M.Eng.

Dosen Pembimbing

Background



Program pengembangan kompetensi kerja dan kewirausahaan yang ditujukan untuk pencari kerja, pekerja/buruh yang terkena pemutusan hubungan kerja atau pekerja/buruh yang membutuhkan peningkatan kompetensi, termasuk pelaku usaha mikro dan kecil

sejak bulan April 2020
insentif terlambat cair & akses website lemah karena membludak & salah sasaran

Purpose

01

Persepsi masyarakat terhadap program prakerja

02

Klasifikasi pandangan masyarakat terhadap program prakerja

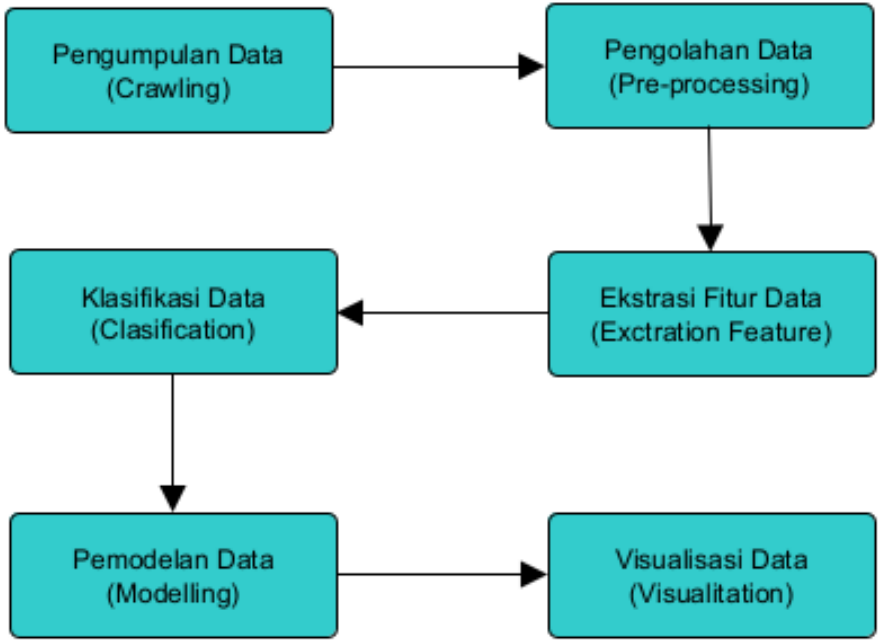
03

Tugas akhir mata kuliah PPL

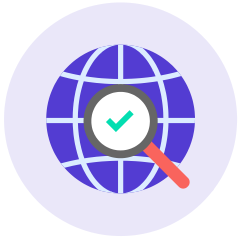
kartu prakerja

Application & Package

Flowchart System



Crawling



```
[ ]  #Memulai Crawling dari twitter
     tweet_fulltext = []
     searchquery = 'prakerja'
     for tweets in tweepy.Cursor(api.search, q=searchquery, count=5000, lang='id',
                                 tweet_mode='extended').items():
       if tweets.full_text not in tweet_fulltext:
         tweet_fulltext.append(tweets.full_text)
     df = pd.DataFrame()
     df['hashtags'] = tweet_fulltext
     print(df.shape)
```

	hashtags
0	@izonemfs kemarin dapet prakerja
1	@TokopediaCare Halo Kak,\nNama : Muhamad Surya...
2	@Chepyzaenal68 Untuk melakukan pendaftaran aku...
3	@bukalapak the best lah bukalapak dan prakerja...
4	@bukalapak emang bukalapak paling bisa di anda...

Preprocessing (Case Folding)

pandas

RegEx
Regular Expression

```
#mengubah huruf kapital dalam data menjadi huruf kecil
df['lower']=df['hashtags'].str.lower()
df.head(5)
```

	Unnamed: 0	hashtags	lower
0	0	@izonemfs kemarin dapet prakerja	@izonemfs kemarin dapet prakerja
1	1	@TokopediaCare Halo Kak,\nNama : Muhamad Surya...	@tokopediacare halo kak,\nnama : muhamad surya...
2	2	@Chepyzaenal68 Untuk melakukan pendaftaran aku...	@chepyzaenal68 untuk melakukan pendaftaran aku...
3	3	@bukalapak the best lah bukalapak dan prakerja...	@bukalapak the best lah bukalapak dan prakerja...
4	4	@bukalapak emang bukalapak paling bisa di anda...	@bukalapak emang bukalapak paling bisa di anda...

Preprocessing (Cleansing)

```
df['clean']=df['lower'].str.replace('(https?://[\w\.\/]*)', '') #hapus mengandung http
df['clean']=df['clean'].str.replace('(http\S+)', '') #hapus mengandung https
df['clean']=df['clean'].str.replace('(?:&(?:lt|nbsp|amp|gt);)', '') #hapus lt,nbsp,gt,amp
df['clean']=df['clean'].str.replace('(@|#)\w+', '') #hapus @ dan #
df['clean']=df['clean'].str.replace('[^A-Za-z0-9\s\-\/]', '') #menghapus selain huruf, spasi dan strip
df['clean']=df['clean'].str.replace('(\-|\/)', ' ') #menghapus -
df['clean']=df['clean'].str.replace('\n', ' ') # hapus enter
df['clean']=df['clean'].str.replace('\t', ' ') # hapus enter
df['clean']=df['clean'].str.replace('\d+', ' ') # hapus angka
df['clean']=df['clean'].str.replace('prakerja', ' ') # hapus kata prakerja
df['clean']=df['clean'].str.replace('\s{2,}', ' ') # hapus spasi lebih dari 2
df['clean']=df['clean'].str.replace('^rt.*', '') #Remove RT Tweets
#df['clean']=df['clean'].str.replace('<.*?>', '') #karakter html
df['clean']=df['clean'].str.replace("\.\.", " ") #hapus .
df['clean']=df['clean'].str.replace('\b[a-zA-Z]\b', '') #huruf mandiri
df.dropna(subset=['clean'], inplace=True) #Remove Empty cell
df = df.drop_duplicates() #Remove Duplicate Tweet
df.head(5)
```

pandas

RegEx
Regular Expression

	Unnamed: 0	hashtags	lower	clean
0	0	@izonemfs kemarin dapet prakerja	@izonemfs kemarin dapet prakerja	kemarin dapet
1	1	@TokopediaCare Halo Kak,\nNama : Muhamad Surya...	@tokopediacare halo kak,\nnama : muhamad surya...	halo kak nama muhamad suryadi syahdoa e mail ...
2	2	@Chepyzaenal68 Untuk melakukan pendaftaran aku...	@chepyzaenal68 untuk melakukan pendaftaran aku...	untuk melakukan pendaftaran akun kartu silaka...
3	3	@bukalapak the best lah bukalapak dan prakerja...	@bukalapak the best lah bukalapak dan prakerja...	the best lah bukalapak dan dalam membantu mas...
4	4	@bukalapak emang bukalapak paling bisa di anda...	@bukalapak emang bukalapak paling bisa di anda...	emang bukalapak paling bisa di andalkan kalo ...

Preprocessing (Tokenizing)

```
import string
import re #regex library
import nltk
nltk.download('punkt')
# impor word_tokenize & FreqDist dari NLTK
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist

# NLTK kata tokenize
def word_tokenize_wrapper(text):
    return word_tokenize(text)

df['tokenize'] = df['clean'].apply(word_tokenize_wrapper)

print('Tokenizing Result : \n')
df.head()
```

RegEx Regular Expression

NLTK

		hashtags	lower	clean	tokenize
0	0	@izonemfs kemarin dapet prakerja	@izonemfs kemarin dapet prakerja	kemarin dapet	[kemarin, dapet]
1	1	@TokopediaCare Halo Kak,\nNama : Muhamad Surya...	@tokopediacare halo kak,\nnama : muhamad surya...	halo kak nama muhamad suryadi syahdoa e mail ...	[halo, kak, nama, muhamad, suryadi, syahdoa, e...
2	2	@Chepyzaenal68 Untuk melakukan pendaftaran aku...	@chepyzaenal68 untuk melakukan pendaftaran aku...	untuk melakukan pendaftaran akun kartu silaka...	[untuk, melakukan, pendaftaran, akun, kartu, s...
3	3	@bukalapak the best lah bukalapak dan prakerja...	@bukalapak the best lah bukalapak dan prakerja...	the best lah bukalapak dan dalam membantu mas...	[the, best, lah, bukalapak, dan, dalam, memban...
4	4	@bukalapak emang bukalapak paling bisa di anda...	@bukalapak emang bukalapak paling bisa di anda...	emang bukalapak paling bisa di andalkan kalo ...	[emang, bukalapak, paling, bisa, di, andalkan,...

Preprocessing (Filtering)

```
from nltk.corpus import stopwords
import nltk
nltk.download('stopwords')
# ---------------------- dapatkan stopword dari stopword NLTK -------------------------------
# dapatkan stopword indonesia
list_stopwords = stopwords.words('indonesian')
print(len(list_stopwords))
# ---------------------------- tambahkan stopword secara manual  ---------------------------------
# tambahkan stopword tambahan
list_stopwords.extend(['yg', 'dg', 'rt', 'dgn', 'ny', 'klo',
                       'kalo', 'amp', 'biar', 'bikin', 'bilang',
                       'gak', 'ga', 'krn', 'nya', 'nih', 'sih',
                       'sih', 'tau', 'tdk', 'tuh', 'utk', 'ya',
                       'jd', 'jgn', 'sdh', 'aja', 'tp','mmg',
                       'nyg', 'hehe', 'pen', 'nan', 'loh', 'nah',
                       'yah','co','ya','aja','the', 'best'])
len(list_stopwords)
# ---------------------- tambahkan stopword dari file txt ------------------------------------
# baca txt stopword menggunakan panda
txt_stopword = pd.read_csv("gdrive/My Drive/PPL/prakerja/ID-Stopwords-master/id.stopwords.02.01.2016.txt",
                           names= ["stopwords"], header = None)
# konversi string stopword ke daftar dan tambahkan stopword tambahan
list_stopwords.extend(txt_stopword["stopwords"][0].split())
len(list_stopwords)
# ----------------------------------------------------------------------------------------
# konversi daftar ke kamus
list_stopwords = set(list_stopwords)
#hapus stopword pada token daftar
def stopwords_removal(words):
    return [word for word in words if word not in list_stopwords]
df['stopword'] = df['tokenize'].apply(stopwords_removal)
df.head(5)
```

NLTK

	Unnamed: 0	hashtags	lower	clean	tokenize	freq_dist_token	stopword
0	0	@izonemfs kemarin dapet prakerja	@izonemfs kemarin dapet prakerja	kemarin dapet	[kemarin, dapet]	{'kemarin': 1, 'dapet': 1}	[kemarin, dapet]
1	1	@TokopediaCare Halo Kak,\nNama : Muhamad Surya...	@tokopediacare halo kak,\nnama : muhamad surya...	halo kak nama muhamad suryadi syahdoa e mail ...	[halo, kak, nama, muhamad, suryadi, syahdoa, e...	{'halo': 1, 'kak': 1, 'nama': 1, 'muhamad': 1,...	[halo, kak, nama, muhamad, suryadi, syahdoa, e...
2	2	@Chepyzaenal68 Untuk melakukan pendaftaran aku...	@chepyzaenal68 untuk melakukan pendaftaran aku...	untuk melakukan pendaftaran akun kartu silaka...	[untuk, melakukan, pendaftaran, akun, kartu, s...	{'untuk': 1, 'melakukan': 1, 'pendaftaran': 1,...	[pendaftaran, akun, kartu, silakan, kunjungi, ...
3	3	@bukalapak the best lah bukalapak dan prakerja...	@bukalapak the best lah bukalapak dan prakerja...	the best lah bukalapak dan dalam membantu mas...	[the, best, lah, bukalapak, dan, dalam, memban...	{'the': 1, 'best': 1, 'lah': 1, 'bukalapak': 1...	[bukalapak, membantu, masyarakat]
4	4	@bukalapak emang bukalapak paling bisa di anda...	@bukalapak emang bukalapak paling bisa di anda...	emang bukalapak paling bisa di andalkan kalo ...	[emang, bukalapak, paling, bisa, di, andalkan,...	{'emang': 1, 'bukalapak': 1, 'paling': 1, 'bis...	[emang, bukalapak, andalkan, penambahan, ilmu,...

Preprocessing (Stemming)

```
#import paket Sastrawi
!pip install Sastrawi
!pip install swifter
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
import swifter
#buat stemmer
factory = StemmerFactory()
stemmer = factory.create_stemmer()

# stemmer
def stemmed_wrapper(term):
    return stemmer.stem(term)

term_dict = {}

for document in df['stopword']:
    for term in document:
        if term not in term_dict:
            term_dict[term] =''

print(len(term_dict))
print("------------------------")

for term in term_dict:
    term_dict[term] = stemmed_wrapper(term)
    print(term,":",term_dict[term])

print(term_dict)
print("------------------------")


# terapkan istilah bertangkai ke kerangka data
def get_stemmed_term(document):
    return [term_dict[term] for term in document]

df['stemming'] = df['stopword'].swifter.apply(get_stemmed_term)
print(df.head())
```

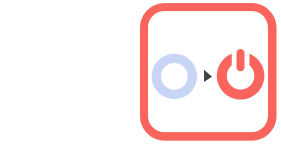
Sastrawi
NLP Bahasa Indonesia

swifter

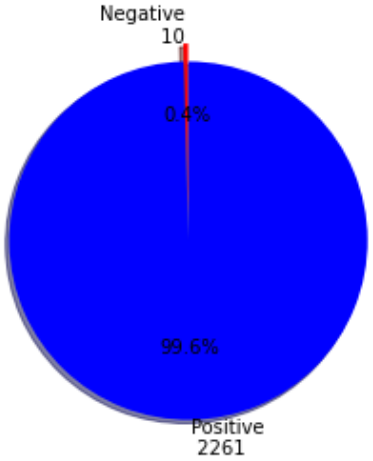
```
gabung : gabung
kompetisi : kompetisi
banyaknya : banyak
pengangguran : anggur
akibat : akibat
diperbanyak : banyak
kuotanya : kuota
kejelasan : jelas
himbara : himbara
```

clean	tokenize	freq_dist_token	stopword	stemming
kemarin dapet	['kemarin', 'dapet']	<FreqDist with 2 samples and 2 outcomes>	['kemarin', 'dapet']	['kemarin', 'dapet']
halo kak nama muhamad suryadi syahdoa e mail ...	['halo', 'kak', 'nama', 'muhamad', 'suryadi', ...	<FreqDist with 22 samples and 22 outcomes>	['halo', 'kak', 'nama', 'muhamad', 'suryadi', ...	['halo', 'kak', 'nama', 'muhamad', 'suryadi', ...
untuk melakukan pendaftaran akun kartu silaka...	['untuk', 'melakukan', 'pendaftaran', 'akun', ...	<FreqDist with 15 samples and 16 outcomes>	['pendaftaran', 'akun', 'kartu', 'silakan', 'k...	['daftar', 'akun', 'kartu', 'sila', 'kunjung',...
the best lah bukalapak dan dalam membantu mas...	['the', 'best', 'lah', 'bukalapak', 'dan', 'da...	<FreqDist with 8 samples and 8 outcomes>	['bukalapak', 'membantu', 'masyarakat']	['bukalapak', 'bantu', 'masyarakat']
emang bukalapak paling bisa di andalkan kalo ...	['emang', 'bukalapak', 'paling', 'bisa', 'di',...	<FreqDist with 11 samples and 12 outcomes>	['emang', 'bukalapak', 'andalkan', 'penambahan...	['emang', 'bukalapak', 'andal', 'tambah', 'ilm...

Data Feature Extraction (LM)



matplotlib



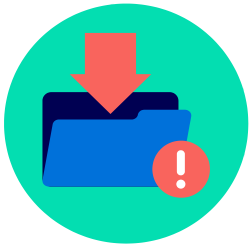

```
# load positive word
positive = pd.read_csv('gdrive/My Drive/PPL/prakerja/ID-OpinionWords-master/positive.txt', header=None)
positive = positive[0].values.tolist()
positive = '|'.join(positive)
# load negative word
negative = pd.read_csv('gdrive/My Drive/PPL/prakerja/ID-OpinionWords-master/negative.txt', header=None)
negative = negative[0].values.tolist()
negative = '|'.join(negative)
```

```
df['positive'] = [len(re.findall(positive, i.lower())) / len(i.split()) for i in df.stemming]
df['negative'] = [len(re.findall(negative, i.lower())) / len(i.split()) for i in df.stemming]
```

```
df['sentimen'] = ['positive' if df.iloc[i].positive >= df.iloc[i].negative else 'negative' for i in range(df.shape[0])]
```

clean	tokenize	freq_dist_token	stopword	stemming	positive	negative	sentimen
kemarin dapet	['kemarin', 'dapet']	<FreqDist with 2 samples and 2 outcomes>	['kemarin', 'dapet']	['kemarin', 'dapet']	1.000000	0.000000	positive
halo kak nama muhamad suryadi syahdoa e mail ...	['halo', 'kak', 'nama', 'muhamad', 'suryadi', ...	<FreqDist with 22 samples and 22 outcomes>	['halo', 'kak', 'nama', 'muhamad', 'suryadi', ...	['halo', 'kak', 'nama', 'muhamad', 'suryadi', ...	0.800000	0.050000	positive
untuk melakukan endaftaran akun kartu silaka...	['untuk', 'melakukan', 'pendaftaran', 'akun', ...	<FreqDist with 15 samples and 16 outcomes>	['pendaftaran', 'akun', 'kartu', 'silakan', 'k...	['daftar', 'akun', 'kartu', 'sila', 'kunjung',...	0.900000	0.000000	positive
the best lah bukalapak dan dalam membantu mas...	['the', 'best', 'lah', 'bukalapak', 'dan', 'da...	<FreqDist with 8 samples and 8 outcomes>	['bukalapak', 'membantu', 'masyarakat']	['bukalapak', 'bantu', 'masyarakat']	2.666667	0.333333	positive
emang bukalapak paling bisa di andalkan kalo ...	['emang', 'bukalapak', 'paling', 'bisa', 'di',...	<FreqDist with 11 samples and 12 outcomes>	['emang', 'bukalapak', 'andalkan', 'penambahan...	['emang', 'bukalapak', 'andal', 'tambah', 'ilm...	1.833333	0.333333	positive

Data Feature Extraction (BOW)



```
bow_model = CountVectorizer(ngram_range=(1,2), min_df=0.0025)
bow_model.fit(df.stemming)
df_vect = pd.DataFrame(bow_model.transform(df.stemming).toarray(), columns=bow_model.get_feature_names())
print(df_vect.shape)
df_vect.head(5)
```

```
df_vect['rate_positive_word'] = [i for i in df['positive']]
df_vect['rate_negative_word'] = [i for i in df['negative']]
df_vect.head()
```

scikit learn

cy

Bag of Words (BOW)

	abis	academy	academy bukalapak	academy dapat	academy nilai	ada	ada gelombang	admin	adu	ah	ahli	ahli sdm	ahli sempat	ahli sertifikasi	airlangga	ajar	akademi	akademi klik	akibat	akibat pandemi	akses	aktif	akun	akun kartu	akun linkaja
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Data Classification

```
from sklearn.naive_bayes import BernoulliNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, classification_report
import seaborn as sn
```

```
x_train, x_test, y_train, y_test = train_test_split(df_vect, df.numbered_sentimen, test_size=0.2,random_state=123)
```

```
test_x = pd.DataFrame(x_test)
test_y = pd.DataFrame(y_test)
test_x.shape[0]
```

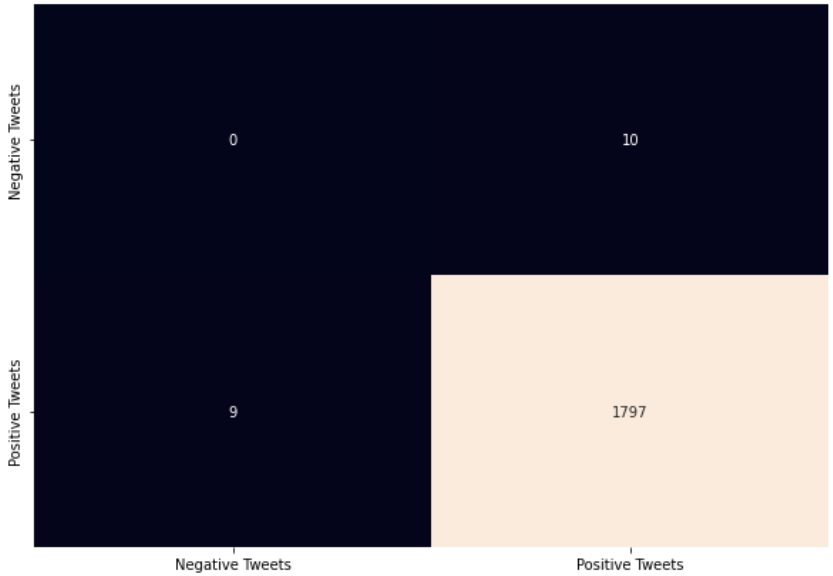
```
#Buat model Bernoulli
model = BernoulliNB()

#Start Training
classifier = model.fit(x_train,y_train)
```

```
Unnamed: 0
2003    1
1976    1
905     1
1215    1
1605    1
Name: numbered_sentimen, dtype: int64
```

	abis	academy	academy bukalapak	academy dapat	academy nilai	ada
2003	0	0	0	0	0	0
1976	0	0	0	0	0	0
905	0	0	0	0	0	0
1215	0	0	0	0	0	0
1605	0	1	0	1	0	0

Data Modeling (CM Train Data)



1816

	precision	recall	f1-score	support
0	0.00	0.00	0.00	10
1	0.99	1.00	0.99	1806
accuracy			0.99	1816
macro avg	0.50	0.50	0.50	1816
weighted avg	0.99	0.99	0.99	1816

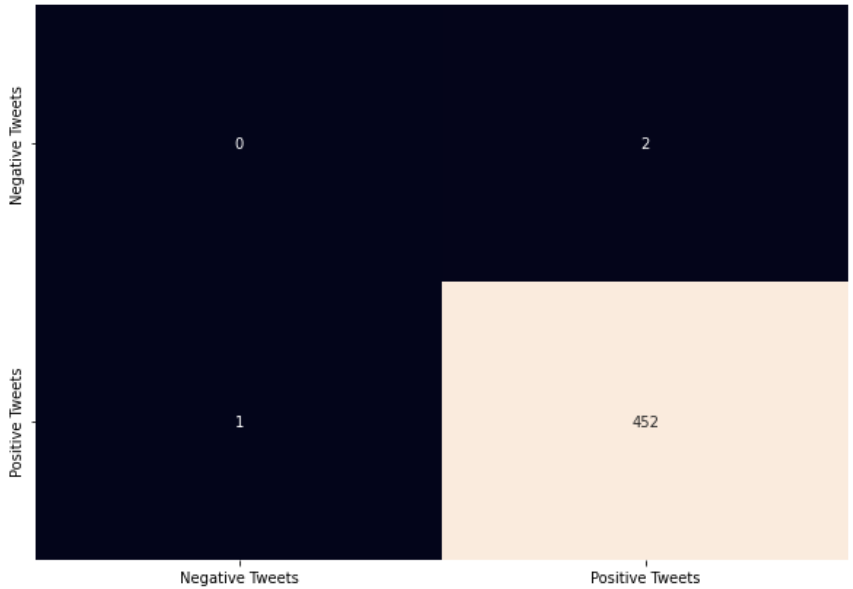
```
#Confusion Matrix Training
cm_train = confusion_matrix(predict,y_train)
print(cm_train)
```

```
[[   0   10]
 [   9 1797]]
```

```
accuracy = cm_train.trace()/cm_train.sum()
print (accuracy)
print(f"{(accuracy*100):3.2f}%")
```

```
0.9895374449339207
98.95%
```

Data Modeling (CM Test Data)



455

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	1.00	1.00	1.00	453
accuracy			0.99	455
macro avg	0.50	0.50	0.50	455
weighted avg	0.99	0.99	0.99	455

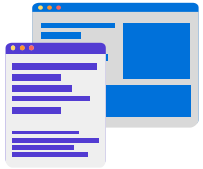
```
#Confusion Matrix Testing
cm = confusion_matrix(predict,y_test)
print(cm)
```

```
[[  0   2]
 [  1 452]]
```

```
predict = classifier.predict(x_test)
classification_report(predict, y_test)
accuracy = cm.trace()/cm.sum()
print (accuracy)
print(f"{(accuracy*100):3.2f}%")
```

```
0.9934065934065934
99.34%
```

AI Sentimen (NBC-Bernoulli)

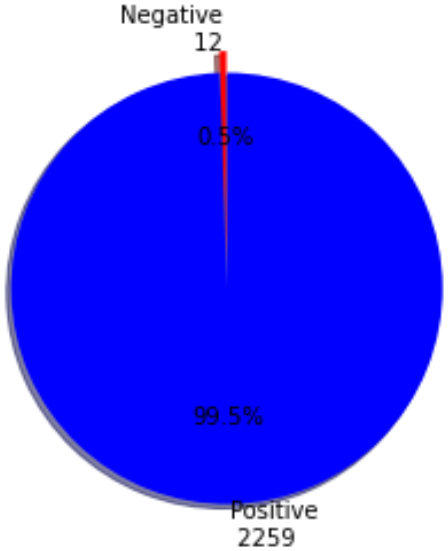


matplotlib

```
#Prediksi Bernoulli dengan BOW(df_vect)
result = model.predict(df_vect)
```

```
print('jumlah positif : ', len([n for n in result if n == 1]))
print('jumlah negatif : ', len([n for n in result if n == 0]))
```

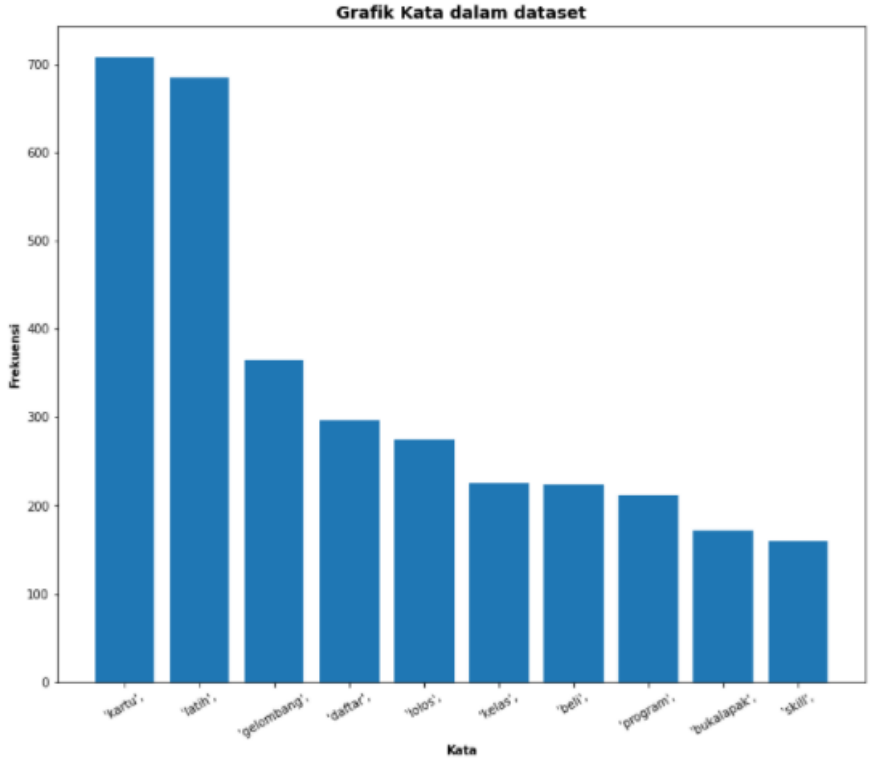
```
jumlah positif :  2259
jumlah negatif :  12
```



Manual vs NBC

Unnamed: 0	hashtags	lower	clean	tokenize	freq_dist_token	stopword	stemming	positive	negative	sentimen	numbered_sentimen	ai_predict	sentiment_ai
0	@izonemfs kemarin dapet prakerja	@izonemfs kemarin dapet prakerja	kemarin dapet	['kemarin', 'dapet']	<FreqDist with 2 samples and 2 outcomes>	['kemarin', 'dapet']	['kemarin', 'dapet']	1.000000	0.000000	positive	1	1	positive
1	@TokopediaCare Halo Kak,\nNama : Muhamad Surya...	@tokopediacare halo kak,\nnama : muhamad surya...	halo kak nama muhamad suryadi syahdoa e mail ...	['halo', 'kak', 'nama', 'muhamad', 'suryadi', ...	<FreqDist with 22 samples and 22 outcomes>	['halo', 'kak', 'nama', 'muhamad', 'suryadi', ...	['halo', 'kak', 'nama', 'muhamad', 'suryadi', ...	0.800000	0.050000	positive	1	1	positive
2	@Chepyzaenal68 Untuk melakukan pendaftaran aku...	@chepyzaenal68 untuk melakukan pendaftaran aku...	untuk melakukan pendaftaran akun kartu silaka...	['untuk', 'melakukan', 'pendaftaran', 'akun', ...	<FreqDist with 15 samples and 16 outcomes>	['pendaftaran', 'akun', 'kartu', 'silakan', 'k...	['daftar', 'akun', 'kartu', 'sila', 'kunjung',...	0.900000	0.000000	positive	1	1	positive
3	@bukalapak the best lah bukalapak dan prakerja...	@bukalapak the best lah bukalapak dan prakerja...	the best lah bukalapak dan dalam membantu mas...	['the', 'best', 'lah', 'bukalapak', 'dan', 'da...	<FreqDist with 8 samples and 8 outcomes>	['bukalapak', 'membantu', 'masyarakat']	['bukalapak', 'bantu', 'masyarakat']	2.666667	0.333333	positive	1	1	positive
4	@bukalapak emang bukalapak paling bisa di anda...	@bukalapak emang bukalapak paling bisa di anda...	emang bukalapak paling bisa di andalkan kalo ...	['emang', 'bukalapak', 'paling', 'bisa', 'di',...	<FreqDist with 11 samples and 12 outcomes>	['emang', 'bukalapak', 'andalkan', 'penambahan...	['emang', 'bukalapak', 'andal', 'tambah', 'ilm...	1.833333	0.333333	positive	1	1	positive

Data Visualization



Data Visualization

Conclusion

Model Accuracy

98,95 % Data Train
99, 34 % Data Test

Sentimen

99,5 % Positive,
0,5 % Negative

Best of Word

Kartu, Latih, Program, Gelombang

Maturwurn