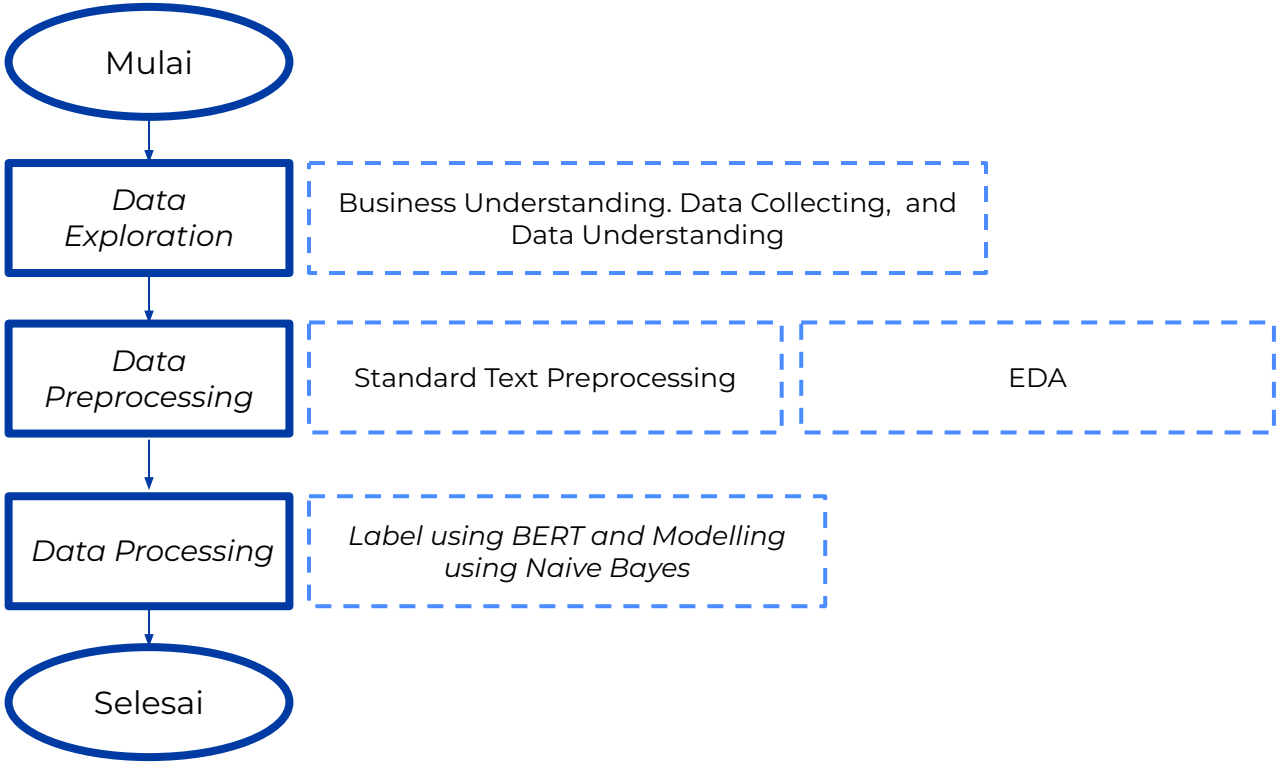
Proyek Akhir Sanbercode

Mochamad Rizal Prasetyo

Table of Contents

Alur Pengerjaan



Latar Belakang

berawal dari keresahan penulis karena kualitas udara jakarta akhir akhir ini sangat buruk sampai - sampai menduduki peringkat satu dunia, sehingga penulis ingin mengetahui bagaimana respon masyarakat terhadap berita ini melalui twitter dan juga beberapa media dengan mengambil data dari headline google search dengan teks 'kualitas udara jakarta'. tweet diambil mulai dari tanggal 20 juni hingga 24 juni 2022 serta google search diambil pada tanggal 25 juni pada laman pertama pencarian google. Penulis ingin mencari apa saja yang beriringan dengan kualitas udara jakarta dari pandangan masyarakat.

Kualitas Udara Jakarta Hari Ini Terburuk Ke-2 di Dunia

Eva Safitri - detikNews

Sabtu, 25 Jun 2022 08:12 WIB

Data Collecting

```
query = "kualitas udara jakarta -is:retweet lang:id"
start_time = '2022-06-20T00:00:00Z'
end_time = '2022-06-24T00:00:00Z'
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 741 entries, 0 to 740
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   tweets  741 non-null    object
dtypes: object(1)
memory usage: 5.9+ KB
```

Data diambil dengan keyword **kualitas udara jakarta** dari 20 juni 2022 hingga 24 juni 2022 menggunakan library tweepy

```
text= "kualitas udara jakarta"
url = 'https://google.com/search?q=' + text
request_result=requests.get( url )
soup = bs4.BeautifulSoup(request_result.text,
                         "html.parser")
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9 entries, 0 to 8
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    9 non-null      object
dtypes: object(1)
memory usage: 200.0+ bytes
```

Data diambil dengan keyword **kualitas udara jakarta** laman pertama pencarian google

Data Collecting

	tweets
0	@MetinBigwin__ok @aniesbaswedan @DKIJakarta Lo...
1	Kualitas Udara Jakarta Tempati Peringkat 1 Ter...
2	Penjelasan Kadis Lingkungan Hidup soal Kualita...
3	Penjelasan Kadis Lingkungan Hidup soal Kualita...
4	@detikcom Anies berush keras perbaiki kualitas...
...	...
736	Kualitas udara pagi ini di Jakarta 😭 diantara ...
737	kualitas udara kota jakarta buruk, sebaiknya p...
738	Kualitas udara ibu kota masuk kategori tidak s...
739	Masih ga ada gunung ya terlihat. Jadi jangan h...
740	Libur lebaran kemarin kualitas udara jakarta n...

Data diambil dengan keyword **kualitas udara jakarta** dari 20 juni 2022 hingga 24 juni 2022 menggunakan library tweepy

:

	text
0	Indeks Kualitas Udara (AQI) Jakarta dan Polusi...
1	Kualitas Udara Jakarta Hari Ini Tidak Sehat, B...
2	Kualitas Udara Jakarta Hari Ini Terburuk Ke-2 ...
3	Laporan Pemantauan Kualitas Udara
4	BMKG Jelaskan Penyebab Kualitas Udara Jakarta ...
5	Kualitas Udara Jakarta Terburuk, Ini Cara Meli...
6	HEADLINE: Kualitas Udara di Jakarta Terburuk S...
7	Lagi, Kualitas Udara Jakarta Terburuk di Dunia...
8	Kualitas Udara Jakarta Memburuk, Ini Dampaknya...

Data diambil dengan keyword **kualitas udara jakarta** laman pertama pencarian google

Data Preprocessing

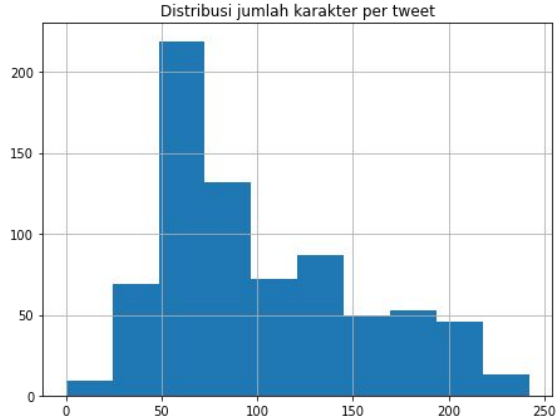
```python
def filtering_text(text):
    # mengubah review menjadi huruf kecil
    text = text.lower()
    # menghilangkan mention, link, hastag
    text = ' '.join(re.sub("([@#][A-Za-z0-9]+)|(\w+:\/\/\S+)"," ", text).split())
    #menghilangkan karakter byte (b')
    text = re.sub(r'(b\'{1,2})',"", text)
    # menghilangkan yang bukan huruf
    text = re.sub('[^a-zA-Z]', ' ', text)
    # menghilangkan digit angka
    text = re.sub(r'\d+', '', text)
    #menghilangkan tanda baca
    text = text.translate(str.maketrans("","",string.punctuation))
    # menghilangkan whitespace berlebih
    text = re.sub(r'\s+', ' ', text).strip()
    return text
```

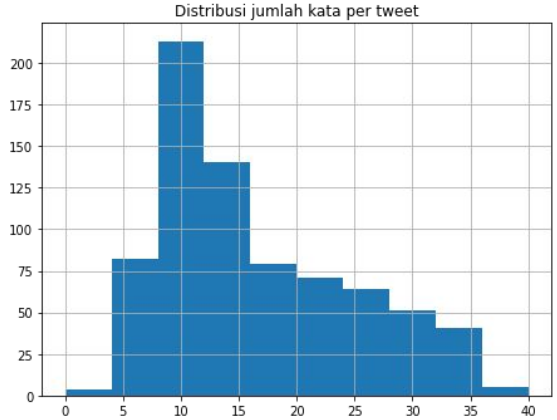
	text	clean_tweet
0	@MetinBigwin__ok @aniesbaswedan @DKIJakarta Lo...	ok loe mau kualitas udara jakarta bersih ikuti...
1	Kualitas Udara Jakarta Tempati Peringkat 1 Ter...	kualitas udara jakarta tempati peringkat terbu...
2	Penjelasan Kadis Lingkungan Hidup soal Kualita...	penjelasan kadis lingkungan hidup soal kualita...
3	Penjelasan Kadis Lingkungan Hidup soal Kualita...	penjelasan kadis lingkungan hidup soal kualita...
4	@detikcom Anies berush keras perbaiki kualitas...	anies berush keras perbaiki kualitas udara jak...

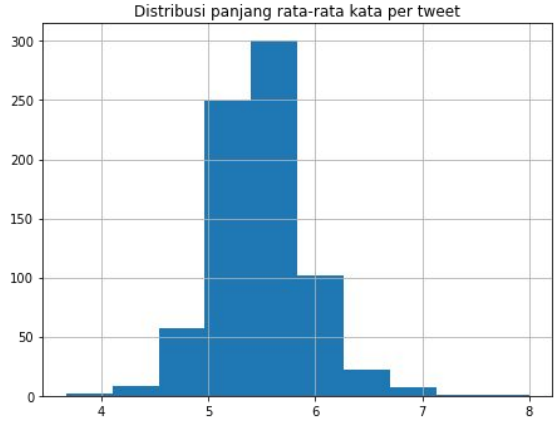
```python
def stop_stem(text):
    #stopword
    with open('kamus.txt') as kamus:
        word = kamus.readlines()
        list_stopword = [line.replace('\n',"") for line in word]
    dictionary = ArrayDictionary(list_stopword)
    stopword = StopWordRemover(dictionary)
    text = stopword.remove(text)
    # stemming
    factory_stemmer = StemmerFactory()
    stemmer = factory_stemmer.create_stemmer()
    text = stemmer.stem(text)
    return text

df['cleaned'] = df['clean_tweet'].apply(stop_stem)
```

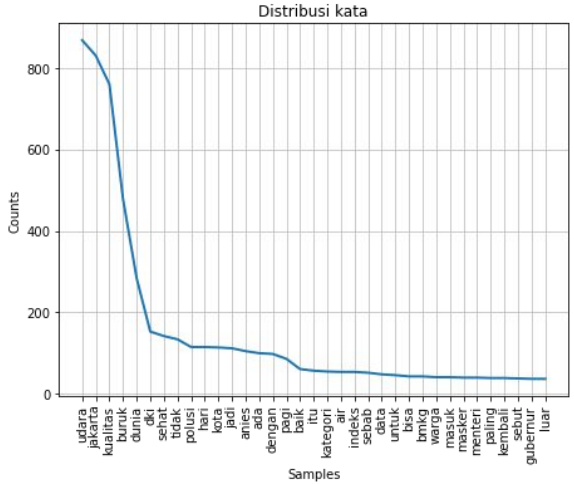
```
Out[15]: 0      ok loe kualitas udara jakarta bersih ikut naek...
         1      kualitas udara jakarta tempat peringkat buruk ...
         2      jelas kad lingkung hidup soal kualitas udara d...
         3      jelas kad lingkung hidup soal kualitas udara d...
         4      anies berush keras baik kualitas udara jakarta...
                                     ...
         745    bmkg jelas sebab kualitas udara jakarta tidak ...
         746    kualitas udara jakarta buruk cara lindung diri...
         747    headline kualitas udara jakarta buruk dunia upaya
         748       kualitas udara jakarta buruk dunia hari kumpar
         749      kualitas udara jakarta buruk dampak untuk sehat
         Name: cleaned, Length: 750, dtype: object
```

Kamus.txt didapat dari kumpulan NLP dalam bahasa indonesia









EDA

'Distribusi Bi-Gram'

(kualitas, udara)	733
(udara, jakarta)	492
(buruk, dunia)	248
(jakarta, buruk)	159
(dki, jakarta)	110
(tidak, sehat)	89
(udara, buruk)	81
(udara, dki)	56
(polusi, udara)	53
(dengan, kualitas)	47
(kategori, tidak)	40
(buruk, kualitas)	39
(jakarta, jadi)	38
(indeks, kualitas)	36
(jakarta, kualitas)	35
(anies, baswedan)	35
(iq, air)	32
(menteri, lhk)	30
(jakarta, pagi)	29
(jakarta, kembali)	28
(ibu, kota)	28
(dunia, kualitas)	27
(hari, akhir)	26
(kota, dengan)	25

dtype: int64

Mayoritas didominasi oleh kata kualitas, udara, jakarta, buruk, sehat.
Ada satu bigram menarik yaitu anies dan baswedan.

anies berush keras perbaiki kualitas udara jakarta yg sumber polusinya dr kendaraan dgn cr menggalakkan transportasi umum perby k bus listrik jalur sepeda revitalisasi ratusan taman kota dls ditunggu action pemegang kekuasaan utk tindkn tegas dr sumber pa brik sekitaran jkt
---
sngat jelas yg dikatakan anies bhwa kualitas udara yg buruk di jakarta sifatnya temporal pnyebabnya bisa berasal dr daerah indu stri di luar jakarta misal banten tangerang atau bekasi jika kualitas udara yg buruk tsb konstan berarti sumbernya mmg dari dlm wilayah jakarta
---
asmara dah jangan berisik ntar di bully loh bos mu kaga memberikan penjelasan berapa lama kualitas udara jakarta yg buruk berla ngsung terus apa hubungannya sama ktp pendatang pada kentut serempak apa nganies dan pendukungnya pada celangap serempak
---
kualitas udara jakarta sering terburuk di dunia anies berkelit begini
---
kualitas udara jakarta buruk penjelasan anies amp gt lt amp amp amp amp amp
---
udara jakarta puncak di bawah pemerintahan pak anies baswedan kualitas udara di jakarta berhasil mencapai pucak puncak terburuk di dunia
---
mari kita simak dan dengarkan pencerahan dari buzzerbalkont asmara yg mana sgt memahami apa yg disampaikan anies ttg buruknya k ualitas udara di jakarta waktu amp t dipersilakan eka agus
---
hut dki jakarta ke djarot minta anies perhatikan kembali kualitas udara di jakarta dapat kado mengagetkan
---
gubernur dki jakarta anies baswedan menyebut buruknya kualitas udara jakarta tanggung jawab bersama
---
soal kualitas udara jakarta terburuk di dunia anies ada suatu peristiwa terjadi
---

EDA

Data Processing

```
import pandas as pd
import numpy as np
import torch
import string
import re
import json
import torch.nn.functional as F
from torch import nn
from transformers import AutoModel
from transformers import BertTokenizer
```

https://huggingface.co/indobenchmark/indobert-base-p1

```
from transformers import BertTokenizer, AutoModel
tokenizer = BertTokenizer.from_pretrained("indobenchmark/indobert-base-p1")
model = AutoModel.from_pretrained("indobenchmark/indobert-base-p1")
```

Pemberian label sentimen dilakukan dengan memanfaatkan pre-trained model indobert-base-p1 yang kemudian di fine-tuned oleh penulis untuk menghasilkan text classification based on sentimen positif, netral, dan negatif.

```
0    ok loe kualitas udara jakarta bersih ikut naek...
1    kualitas udara jakarta tempat peringkat buruk ...
2    jelas kad lingkung hidup soal kualitas udara d...
3    jelas kad lingkung hidup soal kualitas udara d...
4    anies berush keras baik kualitas udara jakarta...
Name: cleaned, dtype: object
```

	text	sentiment
0	ok loe kualitas udara jakarta bersih ikut naek...	negative
1	kualitas udara jakarta tempat peringkat buruk ...	negative
2	jelas kad lingkung hidup soal kualitas udara d...	negative
3	jelas kad lingkung hidup soal kualitas udara d...	negative
4	anies berush keras baik kualitas udara jakarta...	negative

Evaluasi

```
negative    527
neutral     177
positive     46
Name: sentiment, dtype: int64
```

Hasil dari pemberian model menggunakan metode BERT dari 750 data dengan negative sebanyak 527, neutral sebanyak 177, dan positive sebanyak 46.

Kemudian model prediksi dibangun ulang menggunakan metode naive bayes dengan proporsi train 0.8 dan test 0.2.

```
[[82 23  8]
 [12 17  1]
 [ 5  0  2]]
              precision    recall  f1-score   support

    negative       0.83      0.73      0.77       113
     neutral       0.42      0.57      0.49        30
    positive       0.18      0.29      0.22         7

    accuracy                           0.67       150
   macro avg       0.48      0.53      0.49       150
weighted avg       0.72      0.67      0.69       150

nilai akurasinya adalah  0.6733333333333333
```

Akurasi model Naive Bayes:
0.673

Kesimpulan

Sample pencarian anies, baswedan dapat dilihat pada slide EDA sebelumnya. Memang pada akhirnya pemimpin yang disalahkan, akan tetapi menarik dari statement anies baswedan bahwa peristiwa ini bersifat temporal. Akan lebih baik jika pencarian juga dilakukan secara berkala untuk mengetahui apakah benar statement anies bahwa ini bersifat temporal dan polusi berasal dari luar jakarta. Ketika kualitas udara jakarta konstan pada tingkat saat ini maka penyebab utama kejadian ini adalah dari dalam jakarta sendiri.

Dari sisi sentimen analisis, terlihat bahwa masyarakat mayoritas memiliki sentimen negatif terhadap berita ini.

Dari sisi pembangunan model, model naive bayes dinilai buruk dalam melakukan prediksi sentimen dengan skor akurasi 0.673.