

# **ANALISIS *TWEET* AKUN @ELONMUSK**

## **LAPORAN ANALISIS DATA**

disusun untuk memenuhi tugas besar II mata kuliah KU1102 Pengenalan Komputasi yang  
dibimbing oleh Bapak Muhammad Ogin Hasanuddin, S.T., M.T.



**oleh:**

Sulthan Antar Aulia Yunus	(16522057)
Yasmin Fathanah Zakiyyah	(16522087)
Geraldo Valentino De Jesus	(16522227)
Lydia Gracia	(19622217)

**SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA**

**INSTITUT TEKNOLOGI BANDUNG**

**BANDUNG**

**2022**

## KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa. Atas rahmat dan hidayah-Nya, penulis dapat menyelesaikan laporan ini dengan tepat waktu.

Dokumen ini merupakan laporan yang disusun untuk memenuhi tugas besar II mata kuliah KU1102 Pengenalan Komputasi. Selain itu, laporan ini bertujuan menambah wawasan terkait pengolahan data dengan bantuan *library* pandas dan visualisasi data dengan bantuan *library* matplotlib pada python.

Kami mengucapkan terima kasih kepada Bapak Muhammad Ogin Hasanuddin, S.T., M.T. selaku dosen pengampu mata kuliah KU1102 Pengenalan Komputasi. Ucapan terima kasih juga kami sampaikan kepada semua pihak yang telah membantu dalam penyelesaian laporan ini.

Kami menyadari laporan ini masih jauh dari kata sempurna. Oleh sebab itu, apabila terdapat kesalahan dalam penulisan atau pun adanya ketidaksesuaian materi yang kami angkat pada laporan ini, kami mohon maaf. Saran dan kritik yang membangun kami harapkan demi kesempurnaan laporan ini.

Bandung, 20 November 2022  
Penyusun

Kelompok 10

# DAFTAR ISI

<b>KATA PENGANTAR</b>	<b>i</b>
<b>DAFTAR ISI</b>	<b>ii</b>
<b>BAB 1</b>	<b>1</b>
<b>PENDAHULUAN</b>	<b>1</b>
1.1 Latar Belakang	1
1.2 Tujuan	1
1.3 Metode Analisis	1
<b>BAB 2</b>	<b>2</b>
<b>SUMBER DATA</b>	<b>2</b>
2.1 Deskripsi Data	2
2.2 Karakteristik Data	2
<b>BAB 3</b>	<b>5</b>
<b>ANALISIS DATA</b>	<b>5</b>
3.1 Data Cleansing	5
3.2 Statistik Data	5
3.1.1 Penjelasan Sampel Data	5
3.1.2 DataFrame	5
3.1.3 Kesimpulan Sementara	7
3.2 Korelasi Data	8
3.3 Visualisasi Data	10
<b>BAB 4</b>	<b>20</b>
<b>PENUTUP</b>	<b>20</b>
4.1 Kesimpulan	20
4.2 Lesson Learned	20
<b>Lampiran</b>	<b>21</b>
Pembagian Tugas dalam Kelompok	21

# **BAB 1**

## **PENDAHULUAN**

### **1.1 Latar Belakang**

Elon Musk merupakan seorang figur yang sedikit banyak berperan dalam berbagai bidang saat ini. Salah satu isu terbaru dilatarbelakangi pembelian media sosial Twitter oleh Elon. Banyak kebijakan baru dibawah kepemimpinan Elon, salah satunya adalah pemutusan hubungan kerja sejumlah besar karyawan yang bertanggung jawab mengelola Twitter. Keputusan ini dipandang dapat berefek pada perusahaan-perusahaan lain mengikuti jejak yang sama. Dengan demikian, lapangan kerja untuk orang-orang yang mendalami teknologi, seperti *programmer*, *web developer*, *UI/UX designer*, dikhawatirkan akan berkurang. Oleh karena itu, cuitan elon adalah salah satu bahan yang menarik untuk dianalisis.

Data yang dikumpulkan menampilkan beberapa parameter popularitas cuitan dari Elon Musk. Disediakan juga gambaran intensi cuitan hasil analisis sebuah AI. Data-data tersebut akan dianalisis untuk melihat kecenderungan konten cuitan Elon Musk serta pertumbuhan popularitasnya.

### **1.2 Tujuan**

Melalui analisa ini, diharapkan beberapa poin berikut dapat terjawab,

1. Frekuensi cuitan Elon Musk
2. Cuitan Elon Musk dengan *engagement* tertinggi
3. Pertumbuhan popularitas cuitan Elon Musk
4. Kecenderungan sentimen cuitan populer Elon Musk

### **1.3 Metode Analisis**

Metode analisis yang dilakukan untuk menjawab tujuan yang diharapkan adalah metode *descriptive analytics* dengan menggunakan beberapa *library* pada *python* seperti *pandas*, *matplotlib*, dan lain-lain.

## BAB 2

### SUMBER DATA

#### 2.1 Deskripsi Data

Data yang digunakan pada laporan ini adalah data yang memuat daftar *tweet* yang diunggah oleh akun Twitter @elonmusk. Data ini memiliki format csv dan didapatkan dari <https://www.kaggle.com/datasets/alexhugger/elon-tweets-wsentimentclassified-via-roberta> pada *website* Kaggle.

Hal pertama yang dilakukan adalah mengunduh file data pada tautan tersebut. File tersebut bernama “ElonTweets(Sentiment).csv”. Setelah mengunduh file tersebut, langkah berikutnya adalah melakukan loading data pada file dengan cara mengetikkan kode program berikut:

```
import pandas as pd
import matplotlib.pyplot as plt
from wordcloud import WordCloud, ImageColorGenerator
df = pd.read_csv('ElonTweets(Sentiment).csv')
```

Setelah kode program tersebut dijalankan, maka isi dari file “ElonTweets(Sentiment).csv” akan disimpan pada variabel df. Variabel df inilah yang kemudian akan digunakan untuk mengetahui serta menganalisis berbagai informasi yang terkait dengan data cuitan akun @elonmusk.

#### 2.2 Karakteristik Data

Data ini berukuran 4,3 MB. Selain itu, data ini memiliki 16 kolom/atribut dan 17.602 baris/record data dengan berbagai jenis atribut dari kategorikal hingga kuantitatif. Rincian atribut-atribut tersebut dapat dilihat pada tabel berikut.

Tabel 2.1 Rincian Atribut

No.	Nama Atribut	Makna Atribut	Jenis Atribut	Karakteristik
1.	<i>Datetime</i>	Waktu saat cuitan diluncurkan	Kuantitatif-Diskrit	Integer
2.	<i>Tweet id</i>	Id Cuitan	Kuantitatif-Diskrit	Integer
3.	<i>Text</i>	Isi cuitan	Kategorikal-Nominal	String
4.	<i>Username</i>	Nama akun yang digunakan	Kategorikal-Nominal	String
5.	<i>location</i>	Lokasi saat cuitan diluncurkan	Kategorikal-Nominal	String
6.	<i>reply_count</i>	Jumlah pengguna yang membalas/mengomentari	Kuantitatif-Diskrit	Integer
7.	<i>retweet_count</i>	Jumlah pengguna yang menyebarkan cuitan	Kuantitatif-Diskrit	Integer

8.	<i>like_count</i>	Jumlah pengguna yang menyukai cuitan	Kuantitatif-Diskrit	Integer
9.	<i>language</i>	Bahasa yang digunakan pada cuitan	Kategorikal-Nominal	String
10.	<i>Twitter_Access_Point</i>	Jenis perangkat yang digunakan	Kategorikal-Nominal	String
11.	<i>Follower_Count</i>	Jumlah pengikut	Kuantitatif-Diskrit	Integer
12.	<i>Friends_Count</i>	Jumlah teman	Kuantitatif-Diskrit	Integer
13.	<i>verified</i>	Status akun	Kategorikal-Biner	Binary(True/False)
14.	<i>Date</i>	Tanggal cuitan	Kuantitatif-Diskrit	Integer
15.	<i>mentions</i>	Akun yang ditandai dalam cuitan	Kategorikal-Nominal	String
16.	<i>sentiment</i>	Sentimen yang terkandung dalam cuitan	Kategorikal-Ordinal	String(Negatif, Netral, Positif)

Untuk mengetahui kolom yang ada pada data, kita dapat menggunakan kode sederhana untuk menampilkan data pada DataFrame dengan indeks 0 seperti berikut.

```
df[:0]
```

Baris kode tersebut akan menampilkan data pada baris ke-1 atau indeks ke-0 yang umumnya merupakan label/kolom.

Pada file, atribut *sentiment* memiliki dua nilai yaitu kategori dan nilai *sentiment*. Untuk memudahkan beberapa proses pada analisis data, dua atribut ditambahkan pada dataframe sebagai variabel sementara, yaitu atribut *sentimentDef* yang mengandung kategori dari *sentiment* dan atribut *sentimentNum* yang mengandung nilai dari *sentiment*, dengan cara mengetikkan program sebagai berikut:

```
df[['def', 'num']] = df['sentiment'].str.split(',', expand=True)
df['sentimentDef'] = df.loc[:, 'def'].map(lambda x:
x.lstrip("'").rstrip("'"))
df['sentimentDef']
df['sentimentNum'] = pd.to_numeric(df.loc[:, 'num'].map(lambda x:
x.lstrip("'").rstrip("'")))
df['sentimentNum']
```

Kode program tersebut ketika dijalankan akan menghasilkan kolom baru pada *DataFrame* dengan rincian sebagai berikut:

Tabel 2.2 Rincian Atribut Tambahan

No.	Nama Atribut	Makna Atribut	Jenis Atribut	Karakteristik Data
1.	<i>sentimentDef</i>	Sentimen pada cuitan	Kategorikal-Ordinal	Negatif, Netral, Positif
2.	<i>sentimentNum</i>	Sentimen pada cuitan	Kategorikal-Ordinal	-1 hingga 1

## BAB 3

### ANALISIS DATA

#### 3.1 Data Cleansing

Berdasarkan data yang didapatkan, ditemukan kekotoran data yaitu atribut *location* dengan terdapat 100% data yang memiliki nilai atribut *location* kosong. Untuk membersihkan kekotoran data tersebut, maka diketikkan *code* berikut:

```
df.drop('location', inplace=True, axis=1)
```

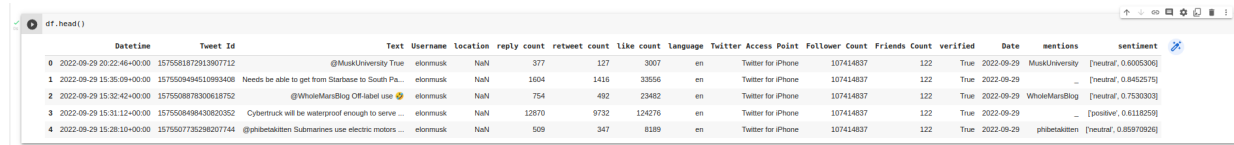
#### 3.2 Statistik Data

##### 3.1.1 Penjelasan Sampel Data

Dalam data yang diambil, terdapat berbagai macam jenis atribut serta karakteristiknya. Setidaknya terdapat lima atribut kuantitatif (enam dengan *Friend\_Count* tetapi memiliki banyak data yang kosong) yang dapat kita hitung secara statistik. Seluruh karakteristik sampel data kuantitatif berbentuk diskrit karena menggambarkan jumlah *user*.

Berikut contoh sampel data pada file:

```
df.head()
```



	Datetime	Tweet Id	Text	Username	Location	reply count	retweet count	like count	language	Twitter Access Point	Follower Count	Friends Count	verified	Date	mentions	sentiment
0	2022-09-29 20:22:46+00:00	1575581872913907712	@MuskUniversity True	elormusk	NaN	377	127	3007	en	Twitter for iPhone	107414837	122	True	2022-09-29	MuskUniversity	[neutral', 0.6005306]
1	2022-09-29 15:35:09+00:00	1575509494510993408	Needs be able to get from Starbase to South Pa...	elormusk	NaN	1604	1416	3356	en	Twitter for iPhone	107414837	122	True	2022-09-29		[neutral', 0.8452575]
2	2022-09-29 15:32:42+00:00	1575508878300618752	@WholeMarsBlog Off-label use 🍷	elormusk	NaN	754	492	23492	en	Twitter for iPhone	107414837	122	True	2022-09-29	WholeMarsBlog	[neutral', 0.7530303]
3	2022-09-29 15:31:12+00:00	1575508498430820352	Cybertruck will be waterproof enough to serve ...	elormusk	NaN	12870	9732	124276	en	Twitter for iPhone	107414837	122	True	2022-09-29		[positive', 0.6118258]
4	2022-09-29 15:28:10+00:00	1575507735298207744	@ghibetaktiten Submarines use electric motors ...	elormusk	NaN	509	347	6189	en	Twitter for iPhone	107414837	122	True	2022-09-29	ghibetaktiten	[neutral', 0.85870926]

##### 3.1.2 DataFrame

Dengan menggunakan contoh baris program berikut.

```
df.describe()
print(df['reply count'].quantile(q=0.1))
print(df['reply count'].quantile(q=0.9))
```

Dapat diketahui untuk semua kolom bertipe data kuantitatif, kita dapat melihat beberapa parameter statistiknya. Selain itu, pada baris kode 2 dan 3 bertujuan untuk menampilkan *percentile* sebesar 10% dan 90% secara manual.

##### 1. Atribut *Follower\_Count*

- Rata-rata : 107.414.348,25
- Standar Deviasi : 1028,13
- Percentile
  - 10% : 107.413.099,0



- 2) 25% : 107.413.409,0
- 3) 50% : 107.413.722,0
- 4) 75% : 107.415.375,0
- 5) 90% : 107.415.611,0

d) Ekstremum

Minimum : 107.412.895,00

Maksimum : 107.415.763,00

2. Atribut *reply\_count*

- a) Rata-rata : 1.555,81
- b) Standar Deviasi : 5.917,44

c) *Percentile*

- 1) 10% : 26,00
- 2) 25% : 64,00
- 3) 50% : 220,00
- 4) 75% : 766,00
- 5) 90% : 2.817,49

d) Ekstremum

Minimum : 0,00

Maksimum : 194.791,00

3. Atribut *like\_count*

- a) Rata-rata : 31.542
- b) Standar Deviasi : 104.472,4

c) *Percentile*

- 1) 10% : 480,00
- 2) 25% : 1.246,00
- 3) 50% : 4.533,50
- 4) 75% : 17.607,50
- 5) 90% : 71.930,69

d) Ekstremum

Minimum : 2,00

Maksimum : 4.799.760,00

4. Atribut *retweet\_count*

- a) Rata-rata : 2.913,79
- b) Standar Deviasi : 12.432,19
- c) *Percentile*
  - 1) 10% : 25,00
  - 2) 25% : 67,00
  - 3) 50% : 262,00
  - 4) 75% : 1.169,00
  - 5) 90% : 5.622,49
- d) Ekstremum
  - Minimum : 0,00
  - Maksimum : 686.210,00

5. Atribut *sentimentNum*

- a) Rata-rata : 0,55
- b) Standar Deviasi : 0,47
- c) *Percentile*
  - 1) 10% : -0,51
  - 2) 25% : 0,53
  - 3) 50% : 0,66
  - 4) 75% : 0,82
  - 5) 90% : 0,92
- d) Ekstremum
  - Minimum : -0,97
  - Maksimum : 0,99

### 3.1.3 Kesimpulan Sementara

Berdasarkan statistik data yang telah didapatkan, kita dapat menyimpulkan untuk masing-masing kolom sebagai berikut.

Khusus pada 4 parameter pertama yaitu *Follower\_Count*, *reply\_count*, *like\_count*, dan *retweet\_count* cenderung memiliki standar deviasi yang sangat besar yang menandakan bahwa data cenderung menyimpang jauh dari reratanya. Hal ini bisa saja disebabkan karena ukuran masing-masing datum dan datanya sangat besar sehingga menghasilkan

variasi data yang sangat banyak. Selain itu, hal tersebut bisa dilihat dari jumlah data yang berkisar sekitar 17.602 data masing-masing kolom dengan range hingga jutaan. Dari segi persentil, empat kolom tersebut memiliki kecenderungan yang hampir sama terutama pada persentil ke-90 (90%). Hal ini ditunjukkan oleh jarak yang sangat jauh antara persentil ke-90 dan sebelumnya. Artinya, nilai tersebut lebih besar dari 90% data lainnya. Hasil ini menggambarkan bahwa nilai tersebut sangat jarang karena hanya terdapat 10% saja dari total data.

Di sisi lain, kolom *sentimentNum* menggambarkan sentimen pada cuitan dengan berbagai parameter seperti kata dan emoji yang kemudian ditampilkan dalam *range* nilai antara -1 hingga 1. Hal ini menggambarkan sentimen negatif, netral, hingga positif secara berurutan. Berdasarkan analisis statistik yang telah dilakukan, standar deviasi *sentimentNum* cukup kecil yaitu 0,47 sehingga mayoritas datanya sangat dekat dengan rerata yaitu 0,55. Kesimpulannya, karena rerata *sentimentNum* positif, dapat kita simpulkan bahwa mayoritas cuitan memiliki *feedback* yang positif antara *user* lain dan pencuit (Elon Musk).

Berdasarkan tujuan yang ingin dicapai kita dapat menentukan cuitan terpopuler yang dicuitkan oleh Elon Musk dengan menggabungkan 3 parameter sekaligus yaitu *reply\_count*, *like\_count*, dan *retweet count*.

Melalui baris program berikut.

```
df.nlargest(1, ["retweet count", "like count", "reply count"])
```

Pada baris kode tersebut, kita mencari satu baris data dengan penjumlahan tiga nilai kolom yang menghasilkan nilai terbesar. Sehingga kita dapatkan cuitan dengan nilai parameter terbesar yaitu sebagai berikut.

	Datetime	Tweet Id	Text	Username	location	reply count	retweet count	like count	language	Twitter Access Point	Follower Count
1768	2022-04-28 00:56:58+00:00	1519480761749016577	Next I'm buying Coca-Cola to put the cocaine b...	elonmusk	NaN	194791	686210	4799760	en	Twitter for iPhone	107415017

Pada gambar tersebut didapatkan bahwa cuitan dengan penjumlahan terbesar dari tiga kolom terdapat pada baris data ke-1769 (1768 merupakan indeks yang dimulai dari 0) dengan isi cuitan yang diluncurkan pada tanggal 28 April 2022 dengan isi teks :

```
Next I'm buying Coca-Cola to put the cocaine back in
```

## 3.2 Korelasi Data

### 3.2.1 Penyajian Korelasi

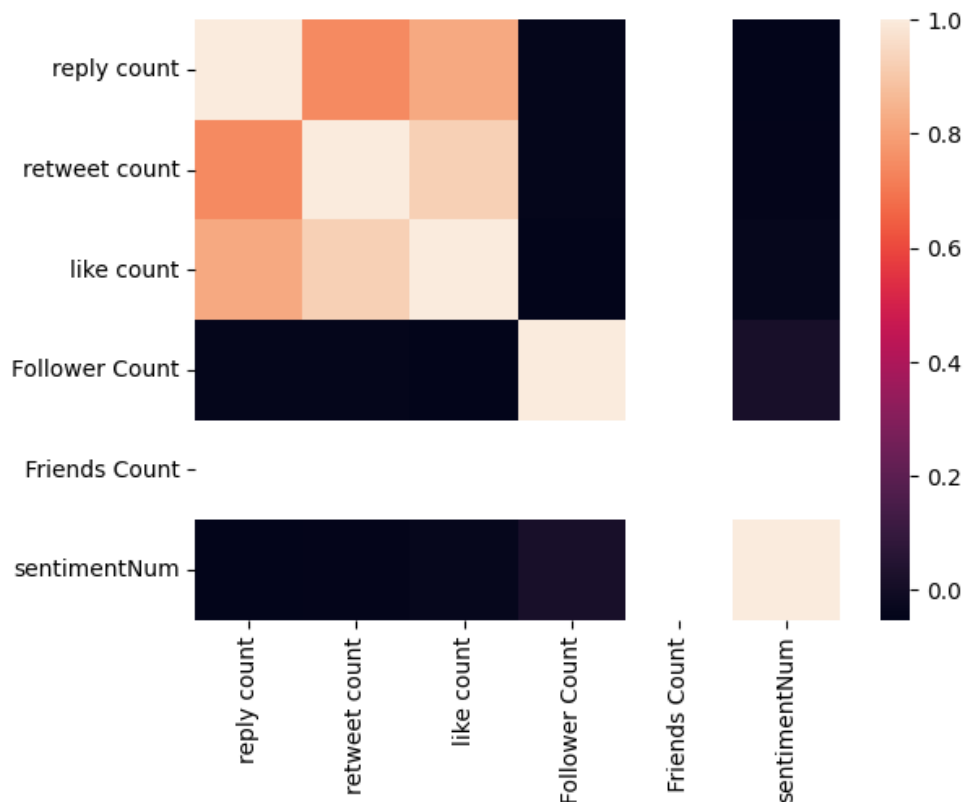
Korelasi disajikan dengan tabel korelasi yang menampilkan angka korelasi serta dengan *heatmap* menggunakan bantuan *library* seaborn.

```
df[["reply count", "retweet count", "like count", "Follower Count",  
"Friends Count", "sentimentNum"]].corr()
```

	reply count	retweet count	like count	Follower Count	Friends Count	sentimentNum
reply count	1.000000	0.737151	0.821201	-0.057401	NaN	-0.039163
retweet count	0.737151	1.000000	0.918919	-0.028244	NaN	-0.046045
like count	0.821201	0.918919	1.000000	-0.052614	NaN	-0.032862
Follower Count	-0.057401	-0.028244	-0.052614	1.000000	NaN	0.021959
Friends Count	NaN	NaN	NaN	NaN	NaN	NaN
sentimentNum	-0.039163	-0.046045	-0.032862	0.021959	NaN	1.000000

```
import seaborn as sns
```

```
sns.heatmap(df[["reply count", "retweet count", "like count", "Follower  
Count", "Friends Count", "sentimentNum"]].corr())
```



### 3.2.2 Korelasi Masing-Masing Atribut

#### 1. Korelasi atribut “reply count” dengan atribut-atribut lain

Atribut “reply count” memiliki relasi positif dengan atribut “retweet count” (dengan angka korelasi 0.737151) dan atribut “like count” (dengan angka korelasi 0.821201) yang berarti pertumbuhan reply sejalan dengan jumlah like dan retweet.

Atribut “reply count” memiliki angka korelasi mendekati 0 dengan atribut “Follower Count” (-0.057401) dan “sentimentNum” (-0.039163). Ini berarti pertumbuhan reply tidak mempengaruhi jumlah follower secara signifikan, serta isi tweet (sentimen) tidak mempengaruhi jumlah reply yang diperoleh.

## 2. Atribut “retweet count” dengan atribut-atribut lain

Atribut “retweet count” memiliki relasi positif dengan atribut “like count” (dengan angka korelasi 0.918919 yang berarti pertumbuhan retweet sejalan dengan jumlah like).

Atribut “retweet count” memiliki angka korelasi mendekati 0 dengan atribut “Follower Count” (-0.028244) dan “sentimentNum” (-0.046045). Ini berarti pertumbuhan retweet tidak mempengaruhi jumlah follower secara signifikan, serta isi tweet (sentimen) tidak mempengaruhi jumlah retweet yang diperoleh.

## 3. Atribut “like count” dengan atribut-atribut lain

Atribut “like count” memiliki angka korelasi mendekati 0 dengan atribut “Follower Count” (-0.052614) dan “sentimentNum” (-0.032862). Ini berarti pertumbuhan like tidak mempengaruhi jumlah follower secara signifikan, serta isi tweet (sentimen) tidak mempengaruhi jumlah like yang diperoleh.

## 4. Atribut “Follower Count” dengan atribut-atribut lain

Atribut “Follower Count” memiliki angka korelasi mendekati 0 dengan atribut “sentimentNum” (0.021959). Ini berarti isi tweet (sentimen) tidak mempengaruhi jumlah follower yang diperoleh.

## 5. Atribut “Friends Count” dengan atribut-atribut lain

Atribut “Friends Count” memiliki jumlah yang tetap sehingga atribut tersebut tidak memiliki korelasi dengan atribut-atribut lain (NaN)

### 3.2.3 Kesimpulan Sementara

Atribut-atribut *engagement* (like, reply, dan retweet) cenderung berbanding lurus satu sama lain. Pengecualian pada jumlah followers yang memiliki pertumbuhan yang lebih lambat. Jumlah friends tetap sehingga tidak memiliki korelasi dengan atribut lain. Sentimen secara garis besar tidak berpengaruh signifikan terhadap jumlah *engagement*.

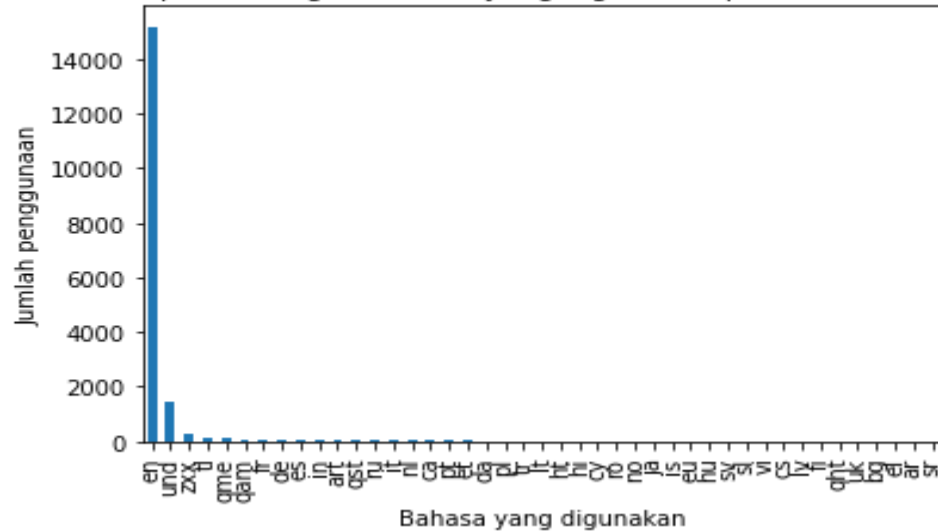
## 3.3 Visualisasi Data

### 1. Perbandingan Kategori

#### a. Perbandingan bahasa yang digunakan

```
df["language"].value_counts().plot(kind = "bar", title="Bar chart
perbandingan bahasa yang digunakan pada tweet @elonmusk")
plt.xlabel('Bahasa yang digunakan')
plt.ylabel('Jumlah penggunaan')
```

Bar chart perbandingan bahasa yang digunakan pada tweet @elonmusk



#### b. Kata yang sering muncul

- Membersihkan isi kolom “Text” pada *DataFrame*

```
import re
import nltk
nltk.download('words')
words = set(nltk.corpus.words.words())

def cleaner(tweet):
    tweet = re.sub("@[A-Za-z0-9]+", "", tweet)
    tweet = re.sub(r"(?:\@|http?://|https?://|www)\S+", "", tweet)
    tweet = " ".join(tweet.split())
    tweet = tweet.replace("#", "").replace("_", " ")
    tweet = " ".join(w for w in nltk.wordpunct_tokenize(tweet) \
                      if w.lower() in words or not w.isalpha())
    return tweet
df['Text'] = df['Text'].map(lambda x: cleaner(x))
```

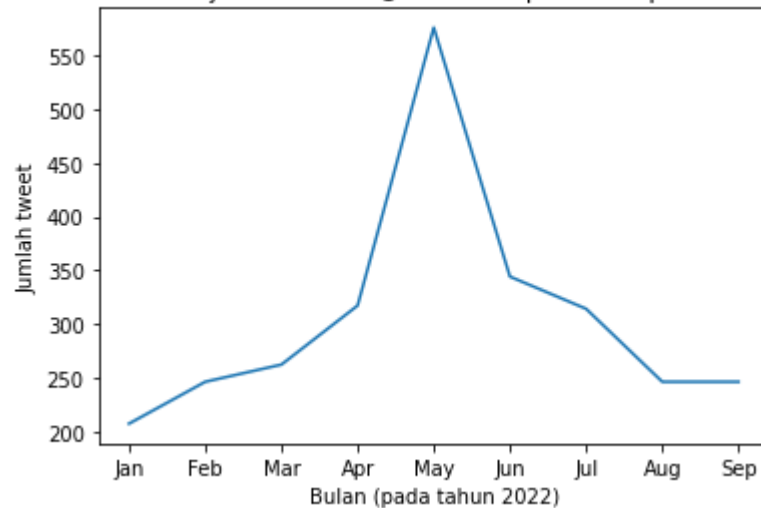
- Menampilkan *WordCloud*

```
from wordcloud import STOPWORDS
text = " ".join(i for i in df.Text)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords,
background_color="white").generate(text)
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
```



```
df_time = pd.to_datetime(df['Datetime'])
df.loc[df_time.dt.strftime('%Y') ==
'2022'].groupby(df_time.dt.strftime('%b'),
sort=False)['Text'].size().iloc[::-1].plot(title="Grafik frekuensi
jumlah tweet @elonmusk per bulan pada tahun 2022")
plt.xlabel('Bulan (pada tahun 2022)')
plt.ylabel('Jumlah tweet')
```

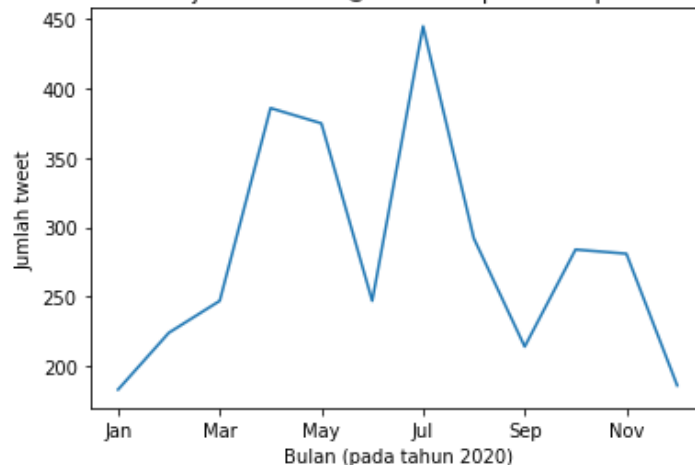
**Grafik frekuensi jumlah tweet @elonmusk per bulan pada tahun 2022**



### 3) Jumlah *tweet* per bulan pada tahun 2020

```
df_time = pd.to_datetime(df['Datetime'])
df.loc[df_time.dt.strftime('%Y') ==
'2020'].groupby(df_time.dt.strftime('%b'),
sort=False)['Text'].size().iloc[::-1].plot(title="Grafik frekuensi
jumlah tweet @elonmusk per bulan pada tahun 2020")
plt.xlabel('Bulan (pada tahun 2020)')
plt.ylabel('Jumlah tweet')
```

**Grafik frekuensi jumlah tweet @elonmusk per bulan pada tahun 2020**





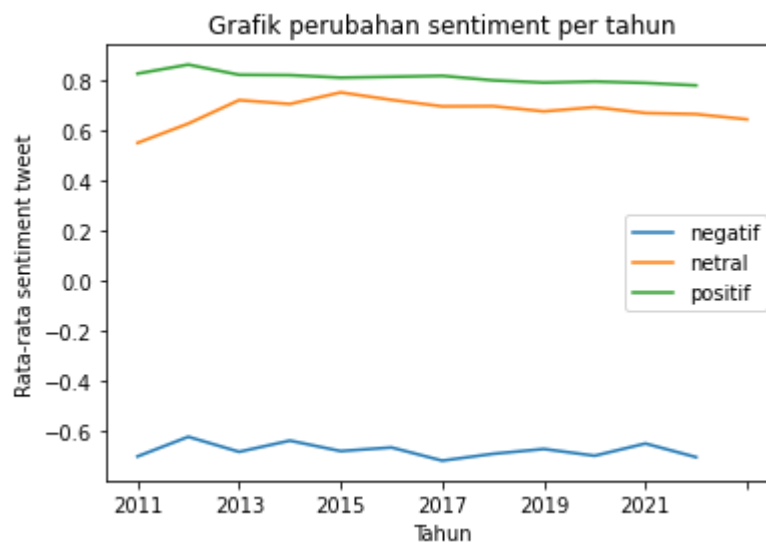
Kesimpulan yang didapatkan adalah jumlah *tweet* yang dibuat oleh akun Twitter @elonmusk semakin meningkat dari tahun ke tahun. Namun, pada grafik tersebut terlihat bahwa jumlah terbanyak bukan pada tahun 2022, melainkan pada tahun 2020. Hal tersebut dapat terjadi karena data *tweet* pada tahun 2022 terakhir pada *DataFrame* adalah pada bulan September 2022 (hanya 9 bulan) sementara pada tahun 2020 data yang diambil selama 12 bulan penuh. Selain itu, terlihat bahwa jumlah *tweet* per bulan maksimal pada tahun 2022 lebih besar daripada tahun 2020. Hal tersebut menandakan bahwa jumlah *tweet* akun @elonmusk semakin meningkat seiring waktu, terutama pada tahun 2020-2022, dan ini sesuai dengan fakta bahwa pemilik akun tersebut, Elon Musk, mulai membeli saham Twitter pada tahun 2022.

Selain naiknya jumlah *tweet*, didapatkan pula informasi bahwa Mei adalah bulan dengan jumlah *tweet* terbanyak yang dibuat akun @elonmusk.

## b. Perubahan *sentiment*

### 1) Rata-rata *sentiment* per tahun

```
neg_df1 = df.loc[df["sentimentDef"] ==
"negative"].groupby(pd.to_datetime(df['Datetime']).dt.strftime('%Y'),
, sort=True)['sentimentNum'].mean()
neu_df2 = df.loc[df["sentimentDef"] ==
"neutral"].groupby(pd.to_datetime(df['Datetime']).dt.strftime('%Y'),
, sort=True)['sentimentNum'].mean()
pos_df3 = df.loc[df["sentimentDef"] ==
"positive"].groupby(pd.to_datetime(df['Datetime']).dt.strftime('%Y'),
, sort=True)['sentimentNum'].mean()
neg_df1.plot(kind='line', label="negatif").legend()
neu_df2.plot(kind='line', label="netral").legend()
pos_df3.plot(kind='line', label="positif").legend()
plt.xlabel('Tahun')
plt.ylabel('Rata-rata sentiment tweet')
plt.title('Grafik perubahan sentiment per tahun')
plt.show()
```

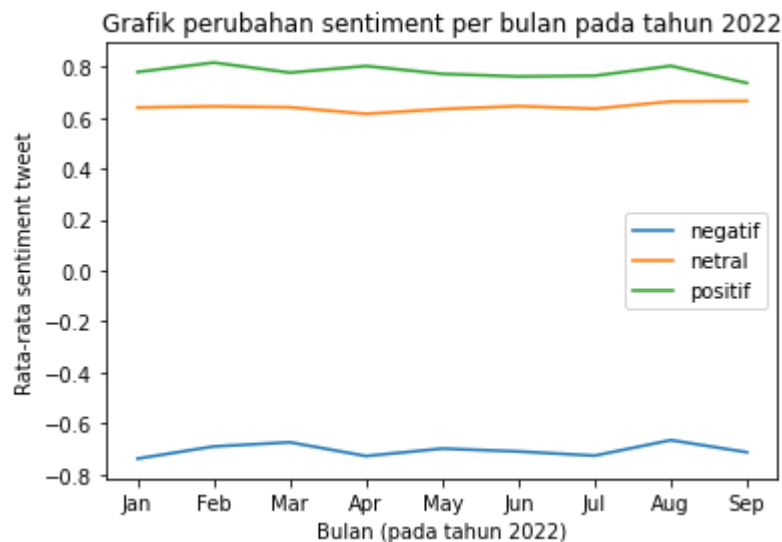


### 2) Rata-rata *sentiment* per bulan pada tahun 2022

```

neg_df1 = df.loc[(df["sentimentDef"] == "negative") &
(df_time.strftime('%Y') ==
'2022')].groupby(pd.to_datetime(df['Datetime']).dt.strftime('%b'),
sort=False)['sentimentNum'].mean().iloc[::-1]
neu_df2 = df.loc[(df["sentimentDef"] == "neutral") &
(df_time.strftime('%Y') ==
'2022')].groupby(pd.to_datetime(df['Datetime']).dt.strftime('%b'),
sort=False)['sentimentNum'].mean().iloc[::-1]
pos_df3 = df.loc[(df["sentimentDef"] == "positive") &
(df_time.strftime('%Y') ==
'2022')].groupby(pd.to_datetime(df['Datetime']).dt.strftime('%b'),
sort=False)['sentimentNum'].mean().iloc[::-1]
neg_df1.plot(kind='line', label="negatif").legend()
neu_df2.plot(kind='line', label="netral").legend()
pos_df3.plot(kind='line', label="positif").legend()
plt.xlabel('Bulan (pada tahun 2022)')
plt.ylabel('Rata-rata sentiment tweet')
plt.title('Grafik perubahan sentiment per bulan pada tahun 2022')
plt.show()

```



Dari kedua grafik tersebut, didapatkan kesimpulan sementara bahwa perubahan nilai sentimen seiring waktu semakin mengarah ke netral.

#### c. Perubahan rata-rata jumlah retweet

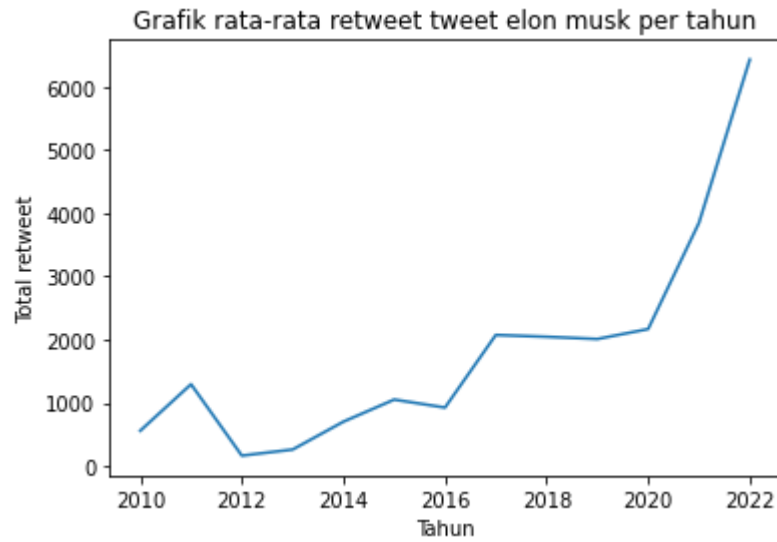
##### 1) Rata-rata *retweet tweet* per tahun

```

df.groupby(pd.to_datetime(df['Datetime']).dt.strftime('%Y'),
sort=True)['retweet count'].mean().plot(title="Grafik rata-rata
retweet tweet elon musk per tahun")

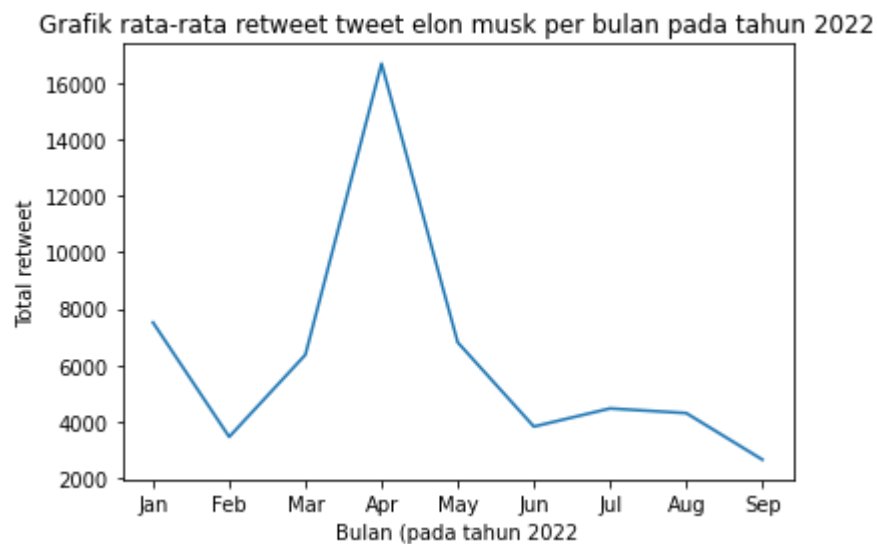
```

```
plt.xlabel('Tahun')
plt.ylabel('Total retweet')
```



## 2) Rata-rata *retweet tweet* per bulan pada tahun 2022

```
df_time = pd.to_datetime(df['Datetime']).dt
df.loc[df_time.strftime('%Y') ==
'2022'].groupby(df_time.strftime('%b'), sort=False)['retweet
count'].mean().iloc[::-1].plot(title="Grafik rata-rata retweet tweet
elon musk per bulan pada tahun 2022")
plt.xlabel('Bulan (pada tahun 2022)')
plt.ylabel('Total retweet')
```

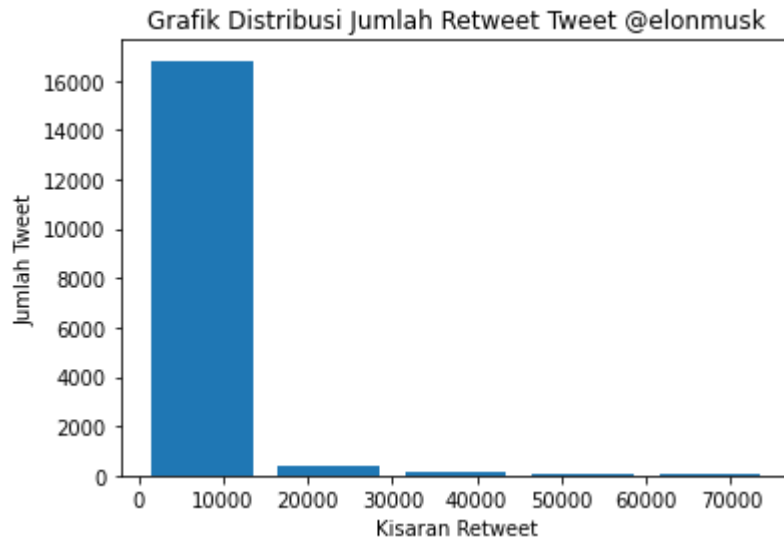


Pada grafik, terlihat bahwa jumlah *retweet* per tahun semakin meningkat dan mencapai jumlah tertinggi pada tahun 2022. Selain itu, jumlah *retweet* per bulan pada tahun 2022 justru semakin menurun, akan tetapi meningkat pada bulan April-Mei, dan ini bersesuaian dengan fakta bahwa Elon Musk mulai membeli Twitter pada bulan tersebut.

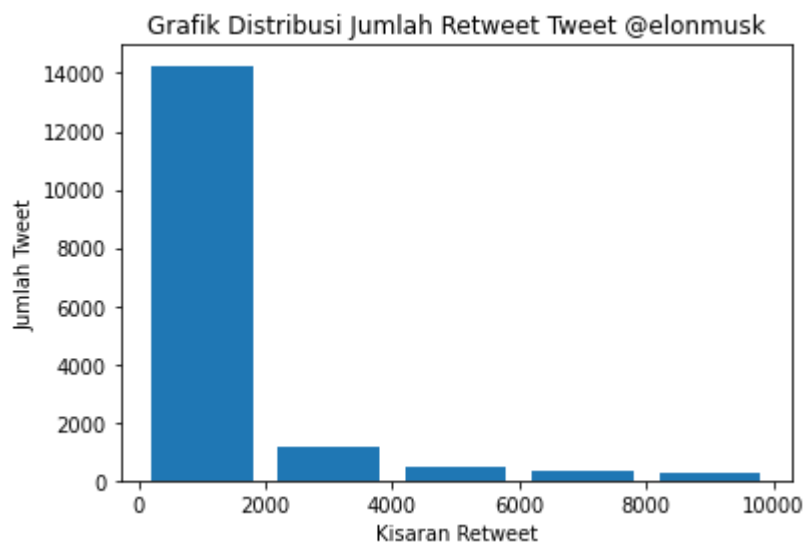
### 3. Penampikan hierarki dan hubungan keseluruhan-bagian

#### a. Distribusi jumlah *retweet*

```
df["retweet count"].plot(kind="hist",bins =[0 ,15000 ,30000 ,45000 ,60000 ,75000] , title="Grafik Distribusi Jumlah Retweet Tweet @elonmusk", rwidth =0.8)  
plt.xlabel('Kisaran Retweet')  
plt.ylabel('Jumlah Tweet')
```



```
df["retweet count"].plot(kind="hist",bins =[0 ,2000 ,4000 ,6000 ,8000 ,10000] , title="Grafik Distribusi Jumlah Retweet Tweet @elonmusk",  
rwidth =0.8)  
plt.xlabel('Kisaran Retweet')  
plt.ylabel('Jumlah Tweet')
```

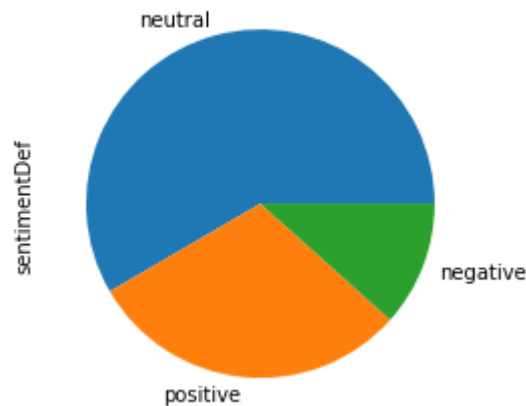


Dari kedua histogram tersebut, terlihat bahwa sebagian besar *tweet* akun @elonmusk memiliki jumlah *retweet* di bawah 2000, namun terdapat pula sebagian *tweet* yang memiliki banyak *retweet* hingga lebih dari 50.000.

b. Perbandingan jumlah tweet per *sentiment*

```
df["sentimentDef"].value_counts().plot(kind = "pie", title="Grafik  
Sebaran Jenis Sentiment Tweet @elonmusk")
```

Grafik Sebaran Jenis Sentiment Tweet @elonmusk

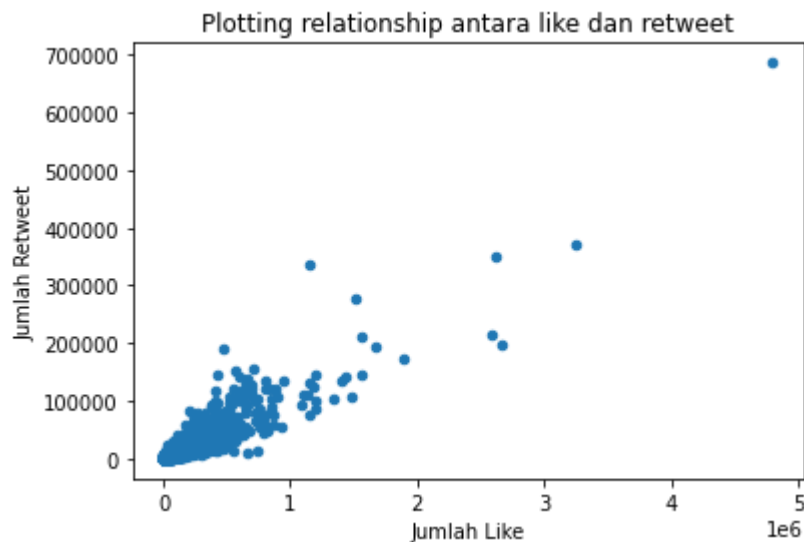


Dari *chart* tersebut, terlihat bahwa sebagian besar (lebih dari setengah) *tweet* akun @elonmusk memiliki suasana yang netral. Sisanya ditempati oleh sentimen negatif dan positif dengan sentimen positif lebih banyak dibanding sentimen negatif.

4. *Plotting relationships* area plot

a. Keterkaitan *like* dan *retweet*

```
df.plot(kind="scatter", x="like count", y="retweet count")  
plt.xlabel('Jumlah Like')  
plt.ylabel('Jumlah Retweet')
```

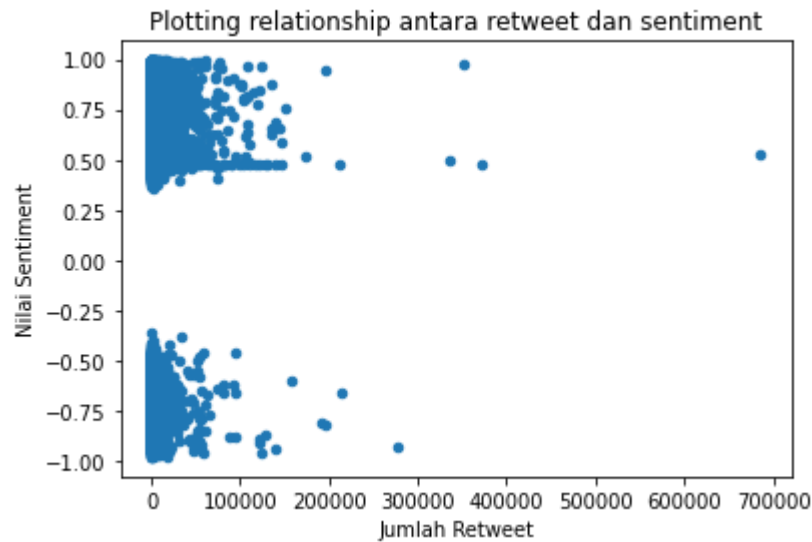


Dari grafik tersebut, terlihat bahwa jumlah *like* dan *retweet* memiliki keterkaitan dan sebagian besar berada di *range like* 0 s.d. 1000000 dan *retweet* 0 s.d. 200000. Namun

terdapat sebagian yang berada di luar *range* tersebut akan tetapi masih sesuai dengan keterkaitan kedua atribut tersebut.

b. Keterkaitan *retweet* dan *sentiment*

```
df.plot(kind="scatter", x="retweet count", y="sentimentNum",  
title="Plotting relationship antara retweet dan sentiment")  
plt.xlabel('Jumlah Retweet')  
plt.ylabel('Nilai Sentiment')
```



Dari grafik tersebut, terlihat bahwa besar nilai *sentiment* dan *retweet* tidak memiliki keterkaitan dan cenderung tersebar merata dari 0.5 hingga 1 serta -1 hingga -0.5.

## BAB 4 PENUTUP

### 4.1 Kesimpulan

- a. Frekuensi jumlah tweet akun twitter @elonmusk semakin meningkat dari tahun ke tahun, dengan jumlah tweet terbanyak terjadi pada tahun 2020. Peningkatan tweet yang terjadi terlihat pada jumlah tweet per bulan pada tahun 2022 melebihi jumlah pada tahun 2020. Peningkatan ini terjadi beriringan dengan sang pemilik, Elon Musk mulai membeli saham twitter pada tahun 2022.
- b. Cuitan terpopuler yang dibuat oleh akun twitter @elonmusk didapat dengan cara mencari jumlah gabungan dari 3 kategori cuitan yaitu *reply\_count*, *like\_count*, dan *retweet count*. Dari analisis didapatkan cuitan terpopuler Elon Musk yaitu cuitan pada baris data ke-1769 dengan isi cuitan yang diluncurkan pada tanggal 28 April 2022 dengan isi teks :

Next I'm buying Coca-Cola to put the cocaine back in

- c. Pertumbuhan popularitas cuitan akun twitter @elonmusk didapat dari nilai korelasi pertumbuhan *reply\_count*, *like\_count*, dan *retweet count* seiring berjalannya waktu. Hasil analisis menunjukkan bahwa pertumbuhan ketiga kategori tersebut mendekati angka 1 yang berarti ketiga kategori memiliki pertumbuhan yang sejalan. Atribut “reply count” memiliki relasi positif dengan atribut “retweet count” (dengan angka korelasi 0.737151) dan atribut “like count” (dengan angka korelasi 0.821201). Melalui grafik analisis juga didapat bahwa rata-rata retweet per tahun semakin meningkat, dengan jumlah tertinggi yaitu pada tahun 2022. Hasil ini menunjukkan bahwa pertumbuhan popularitas cuitan Elon Musk semakin meningkat dari tahun ke tahun.
- d. Kecenderungan *sentiment* cuitan akun twitter @elonmusk semakin mengarah ke netral seiring berjalannya waktu. Melalui analisis data grafik didapat kesimpulan bahwa sebagian besar cuitan Elon Musk bersentimen netral dengan sisanya memiliki sentiment negatif dan positif yang didominasi oleh *sentiment* positif.

### 4.2 Lesson Learned

Dalam menganalisis data yang didapat, dibutuhkan kemampuan untuk mengenal persoalan yang diberikan, menggunakan dekomposisi dan abstraksi dalam pemecahan persoalan, serta menghasilkan artifak komputasi yang relevan. Untuk menganalisis data dengan baik dan efektif, dibutuhkan kolaborasi kelompok dalam memecahkan persoalan komputasi dan kemampuan berkomunikasi dengan berbagai pihak dalam rangka bertukar ide dan berekspres mengenai pemecahan persoalan komputasi.

## Lampiran

Tabel Pembagian Tugas dalam Kelompok

No.	NIM	Nama	Tugas
1.	16522057	Sulthan Antar Aulia Yunus	a. <i>Data cleansing</i> b. Penutup
2.	16522087	Yasmin Fathanah Zakiyyah	a. Deskripsi data b. Visualisasi data
3.	16522227	Geraldo Valentino De Jesus	a. Karakteristik data b. Statistik data
4.	19622217	Lydia Gracia	a. Pendahuluan b. Korelasi data