Wilbert Winardi / 2702238716

Data (deskripsi) akan dibersihkan dengan cara menghapus seluruh tanda baca, link, dan angka agar tidak terjadi bias. Selain itu juga akan membuat seluruh huruf menjadi huruf kecil.

Setelah itu, deskripsi akan divectorize menggunakan TFIDF. Sedangkan label sentimen akan diencode menjadi data berupa angka (0,1,2).

Lalu, data train test akan displit menjadi 0.8 : 0.2. Terakhir, akan dilakukan resampling data untuk mengatasi ketidakseimbangan data training.

In [4]:
```
import pandas as pd
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
import re
from imblearn.over_sampling import SMOTE
```

Importing Data

In [5]:
```
df = pd.read_csv('/content/news_sentiment_analysis.csv')
df
```

Out[5]:

	Source	Author	Title	Description	URL	P
0	stgnews	Bridger Palmer	Pine View High teacher wins Best in State awar...	ST. GEORGE — Kaitlyn Larson, a first-year teac...	https://www.stgeorgeutah.com/news/archive/2024...	12T23:
1	Zimbabwe Mail	Staff Reporter	Businesses Face Financial Strain Amid Liquidit...	Harare, Zimbabwe – Local businesses are grappl...	https://www.thezimbabwemail.com/business/busin...	12T22:
2	4-traders	NaN	Musk donates to super pac working to elect Tru...	(marketscreener.com) Billionaire Elon Musk has...	https://www.marketscreener.com/business-leader...	12T22:
3	4-traders	NaN	US FTC issues warning to franchisors over unfa...	(marketscreener.com) A U.S. trade regulator on...	https://www.marketscreener.com/quote/stock/MCD...	12T22:
4	PLANET	NaN	Rooftop solar's dark side	4.5 million households in the U.S. have solar ...	https://www.npr.org/2024/07/12/1197961036/roof...	12T22:
...	...	...	...	...	...	
3495	etfdailynews	MarketBeat News	Arrow Electronics, Inc. (NYSE:ARW) Shares Purc...	QRG Capital Management Inc. increased its stak...	https://www.etfdailynews.com/2024/07/18/arrow-...	18T14:
3496	etfdailynews	MarketBeat News	3,120 Shares in NICE Ltd. (NASDAQ:NICE) Bought...	QRG Capital Management Inc. bought a new posit...	https://www.etfdailynews.com/2024/07/18/3120-s...	18T14:
3497	etfdailynews	MarketBeat News	QRG Capital Management Inc. Has $857,000 Stock...	QRG Capital Management Inc. boosted its stake ...	https://www.etfdailynews.com/2024/07/18/qrg-ca...	18T14:

	Source	Author	Title	Description	URL	P
3498	finanznachrichten	NaN	Biotechnology Market: Surging Investments and ...	WESTFORD, Mass., July 18, 2024 /PRNewswire/ --...	https://www.finanznachrichten.de/nachrichten-2...	18T14:
3499	etfdailynews	MarketBeat News	QRG Capital Management Inc. Sells 1,665 Shares...	QRG Capital Management Inc. reduced its holdin...	https://www.etfdailynews.com/2024/07/18/qrg-ca...	18T14:

3500 rows × 8 columns

Data Info

In [6]:
```
df.info()
# Untuk mengecek data type dan apakah ada data yang null pada setiap fitur
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3500 entries, 0 to 3499
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Source        3500 non-null   object
 1   Author        2512 non-null   object
 2   Title         3500 non-null   object
 3   Description   3500 non-null   object
 4   URL           3500 non-null   object
 5   Published At  3500 non-null   object
 6   Sentiment     3500 non-null   object
 7   Type          3500 non-null   object
dtypes: object(8)
memory usage: 218.9+ KB
```

Data Cleaning

In [7]:
```
def cleaning(text):
    text = text.lower()   # Membuat seluruh huruf menjadi huruf kecil
    text = re.sub(r'http\S+', '', text)     # Hapus URL
    text = re.sub(r'[^\w\d\s\-]', '', text) # Hapus tanda baca
    text = re.sub(r'\d+', '', text) # Hapus angka
    return text
```

Extracting Independent and Dependent Variable

In [8]:
```
X = df['Description'].apply(cleaning)   # Independent Variable menggunakan description

tfidf = TfidfVectorizer(ngram_range=(1, 2))   # Menggunakan unigram dan bigram untuk konteks yang lebih baik
X = tfidf.fit_transform(X)   # Vectorizing menggunakan TFIDF untuk mengekstrak informasi pada teks

encoder = LabelEncoder()    # Encoding menggunakan label encoder
y = encoder.fit_transform(df['Sentiment'])  # Dependent Variable menggunakan label sentimen yang sudah diencode
```

In [9]:
```
print(X)
```

```
<Compressed Sparse Row sparse matrix of dtype 'float64'
        with 280022 stored elements and shape (3500, 86726)>
  Coords        Values
  (0, 70254)    0.050771315067249204
  (0, 30943)    0.11008218305750153
  (0, 41093)    0.11561845486670108
  (0, 42564)    0.11561845486670108
  (0, 27895)    0.03412096780577225
  (0, 85863)    0.0523352797482869
  (0, 73054)    0.10310730545713845
  (0, 6487)     0.033241933424306716
  (0, 57214)    0.11561845486670108
  (0, 82246)    0.0985130710312075
  (0, 34250)    0.06426255332826959
  (0, 65639)    0.07280602073013408
  (0, 62016)    0.06829685901761416
  (0, 61934)    0.07166751770794104
  (0, 74251)    0.08847873300561364
  (0, 9072)     0.06829685901761416
  (0, 36252)    0.04559703534533684
  (0, 70611)    0.06442225253277663
  (0, 7449)     0.08287728988175973
  (0, 28449)    0.032096228685869505
  (0, 10819)    0.04383194886877984
  (0, 43121)    0.08351248005219056
  (0, 16956)    0.09234160905157712
  (0, 42566)    0.11561845486670108
  (0, 71447)    0.06559572132652777
  :     :
  (3499, 39794) 0.09477681334181935
  (3499, 27527) 0.08186279889697798
  (3499, 60285) 0.09348569590634936
  (3499, 1316)  0.09411992670729565
  (3499, 66680) 0.09411992670729565
  (3499, 61987) 0.08835149113575086
  (3499, 45621) 0.11081069099344154
  (3499, 62271) 0.11248374119291439
  (3499, 62273) 0.12277556272372163
  (3499, 37477) 0.118671758550596
  (3499, 37271) 0.14297999914292542
  (3499, 66110) 0.14297999914292542
  (3499, 49581) 0.14297999914292542
```

```
  (3499, 66111) 0.14297999914292542
  (3499, 73350) 0.13887619496979978
  (3499, 34819) 0.13887619496979978
  (3499, 57593) 0.14297999914292542
  (3499, 49582) 0.14297999914292542
  (3499, 18168) 0.14297999914292542
  (3499, 71078) 0.14297999914292542
  (3499, 45623) 0.13887619496979978
  (3499, 60803) 0.23273979877412368
  (3499, 60804) 0.23273979877412368
  (3499, 56733) 0.13268817761211818
  (3499, 52892) 0.14827072290575846
[2 1 2 ... 2 1 2]
```

In [12]:
```
print(y)
# negatif = 0; netral = 1; positif = 2
```

```
[2 1 2 ... 2 1 2]
```

Split Data

In [10]:
```
# Split data untuk training dan testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Train : Test
#   0.8 : 0.2
```

Resampling

In [11]:
```
# Resampling menggunakan SMOTE untuk mengatasi ketidakseimbangan data
smote = SMOTE(random_state=42)
X_res, y_res = smote.fit_resample(X_train, y_train)
```