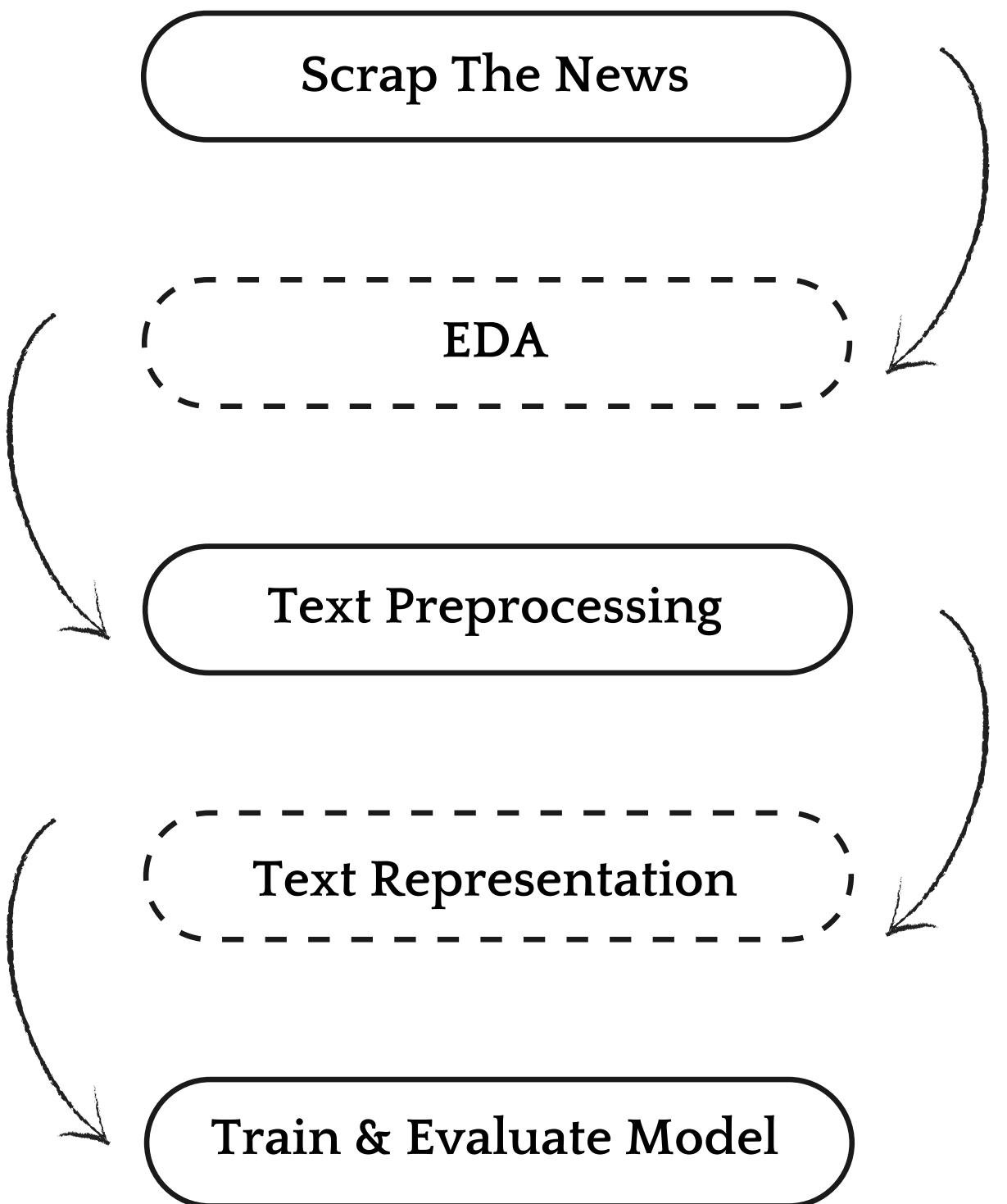


NEWS CATEGORY CLASSIFICATION



Methodology

Methodology



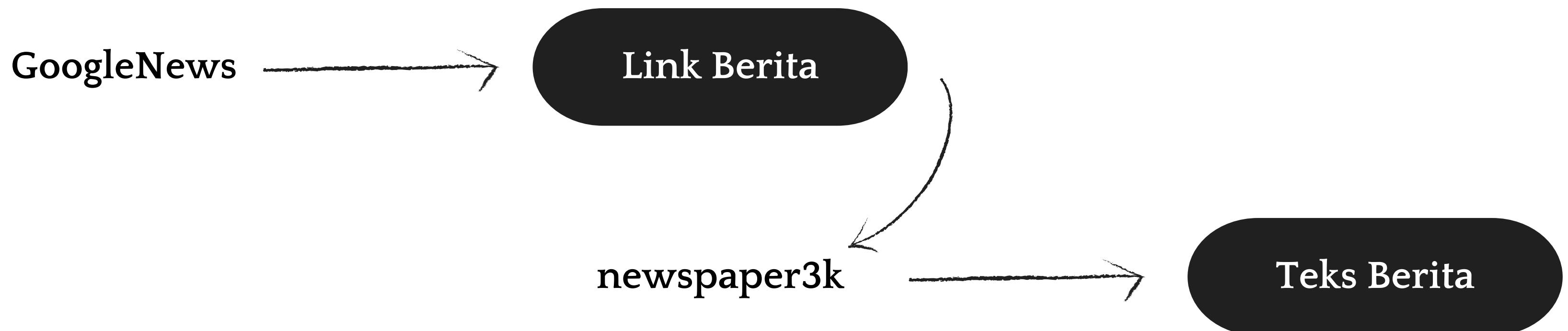
Pada project kali ini kita akan melakukan text classification khususnya untuk teks berita. Teks berita ini akan kita dapatkan dengan melakukan scraping dari 4 buah news outlet di Indonesia, yaitu **kompas**, **pikiran-rakyat**, **tribunnews**, dan **merdeka**.

Selanjutnya, setelah scrap datanya, selanjutnya kita akan mengeksplor datanya, melakukan text preprocessing, text representation menggunakan skipgram dan TF-IDF, serta membuat model prediksi menggunakan RandomForest dan SVM.

Scrap The News

Scrap The News

Untuk mengscraping berita dari 4 news outlet tersebut, kita akan menggunakan library GoogleNews dan newspaper3k. library GoogleNews sendiri akan kita gunakan untuk memperoleh link beritanya dan newspaper3k digunakan untuk mengestrak text beritanya.



Scrap The News

Pertama-tama kita perlu set dulu agar nanti google akan searchnya berita yang berbahasa indonesia dan diterbitkan di indonesia.

```
google = GoogleNews(lang="id", region="id")
google.set_time_range('9/01/2023','10/31/2023')
```

Selanjutnya kita akan buat searchTermList dan CategoryList. searchTermList ini akan berisi keyword untuk mencari berita. Disini kita akan cari 3 kategori (olahraga, politik, hiburan) jadi searchTermList ini akan terdiri dari nama news outlet dan kategorinya. CategoryList sendiri adalah kategori yang correspond dengan searchTermList dan akan digunakan untuk pelabelan.

```
searchTermList = ['kompas.com/sports', 'nasional.kompas.com', 'kompas.com/hype', 'pikiran-rakyat.com/nasional', 'pikiran-rakyat.com/olahraga', 'pikiran-rakyat.com/entertainment', 'tribunnews.
categoryList = ['olahraga', 'politik', 'hiburan', 'politik', 'olahraga', 'hiburan', 'politik', 'olahraga', 'hiburan', 'politik', 'olahraga', 'hiburan']
```

Scrap The News

Selanjutnya kita akan search beritanya berdasarkan searchTermList dan melakukan pelabelan berdasarkan category list. Nantinya 1 search term akan terdiri dari 1 dataframe berisi 10 berita. Sehingga nanti dataframe gabungannya akan terdiri dari 120 berita (4 news outlet dengan masing-masing 3 kategori)

```
df_list = []

for searchTerm, category in zip(searchTermList, categoryList):
    result_list = []
    google.search(searchTerm)

    search = google.page_at(1)
    result_list.append(search)

    df = pd.concat([pd.DataFrame(result) for result in result_list], ignore_index=True)
    df['category'] = category
    df_list.append(df)
```

	title	media	date	datetime	desc	link	img	category
0	Alasan Rivan Nurmukti Absen pada Livoli 2023 ~...	Kompas.com	2 hari lalu	NaN	https://www.kompas.com/sports/read/2023/11/02/...	data:image/gif;base64,R0lGODlhAQABAAAP///...		olahraga
1	Presiden FIFA Resmi Umumkan Arab Saudi Tuan Ru...	Kompas.com	3 hari lalu	NaN	https://www.kompas.com/sports/read/2023/11/01/...	data:image/gif;base64,R0lGODlhAQABAAAP///...		olahraga
2	Klasemen Liga 1: Borneo FC ke Puncak, Persib N...	Kompas.com	6 hari lalu	NaN	https://www.kompas.com/sports/read/2023/10/28/...	data:image/gif;base64,R0lGODlhAQABAAAP///...		olahraga
3	Sejarah Rifda Irfanaluthfi Tembus Olimpiade 20...	Kompas.com	bulan lalu	NaN	https://www.kompas.com/sports/read/2023/10/03/...	data:image/gif;base64,R0lGODlhAQABAAAP///...		olahraga
4	Klasemen Medali Asian Games 2022: Indonesia di...	KOMPAS.com	bulan lalu	NaN	https://www.kompas.com/sports/read/2023/10/03/...	data:image/gif;base64,R0lGODlhAQABAAAP///...		olahraga
...
115	Selain Raisa, Ini Deretan Artis yang Nonton Ko...	Merdeka.com	bulan lalu	NaN	https://www.merdeka.com/jateng/selain-raisa-in...	data:image/gif;base64,R0lGODlhAQABAAAP///...		hiburan
116	Diah Permatasari Jalan-Jalan sama Anaknya yang...	Merdeka.com	bulan lalu	NaN	https://www.merdeka.com/arts/diah-permatasari...	data:image/gif;base64,R0lGODlhAQABAAAP///...		hiburan
117	Artis Nindy Ayunda Datangi Mabes Polri, Siap D...	Merdeka.com	bulan lalu	NaN	https://www.merdeka.com/peristiwa/artis-nindy...	data:image/gif;base64,R0lGODlhAQABAAAP///...		hiburan
118	Tinggi & Ganteng, ini Kriteria Pacar Marco Put...	Merdeka.com	bulan lalu	NaN	https://www.merdeka.com/arts/tinggi-amp-gante...	data:image/gif;base64,R0lGODlhAQABAAAP///...		hiburan
119	Profil Rebecca Klopper, Artis Cantik yang seda...	Merdeka.com	bulan lalu	NaN	https://www.merdeka.com/peristiwa/profil-rebec...	data:image/gif;base64,R0lGODlhAQABAAAP///...		hiburan

Scrap The News

The next step, kita akan cek links untuk memastikan bahwa kita punya masing-masing 30 berita untuk setiap media dan 40 berita untuk setiap kategori.

```
df_all['category'].value_counts()
```

```
olahraga    40
politik     40
hiburan    40
Name: category, dtype: int64
```

next, kita samakan nama medianya sesuai dengan nama media yang ada di linknya dan kita cek jumlah news untuk tiap media. Ternyata ada 1 media yang tidak kita inginkan

```
def update_media(row):
    if 'kompas.com' in row['link']:
        return 'Kompas'
    elif 'tribunnews.com' in row['link']:
        return 'Tribun News'
    elif 'pikiran-rakyat.com' in row['link']:
        return 'Pikiran Rakyat'
    elif 'merdeka.com' in row['link']:
        return 'Merdeka'
    else:
        return row['media']

df_all['media'] = df_all.apply(update_media, axis=1)
```

```
df_all['media'].value_counts()
```

```
Kompas        30
Tribun News   30
Merdeka       30
Pikiran Rakyat 29
Radar Cirebon TV 1
Name: media, dtype: int64
```

Scrap The News

Disini kita cek, berita mana yang asalnya dari media yang tidak kita inginkan. Lalu kita carikan berita penggantinya dengan kategori yang sama manually

```
df_check = df_all[df_all['media'] == 'Radar Cirebon TV'][['link', 'category']]  
df_check
```

link	category
------	----------

40	https://www.radarcirebon.tv/2023/11/04/resmi-v... olahraga
----	--

```
from pandas.core.dtypes.missing import notnull  
new_row = {  
    'title': 'PON Papua Ditunda, Tim Atletik Jabar Tetap Jalankan Arahan KONI',  
    'media': 'Pikiran Rakyat',  
    'date': np.nan,  
    'datetime': np.nan,  
    'desc': np.nan,  
    'link': 'https://www.pikiran-rakyat.com/olahraga/pr-01376956/pon-papua-ditunda-tim-atletik-jabar-tetap-jalankan-arahan-koni',  
    'img': np.nan,  
    'category': 'olahraga'  
}  
  
df_all = df_all.append(new_row, ignore_index=True)
```

Scrap The News

Sekarang kita sort datanya berdasarkan category lalu berdasarkan media dan kita drop semua kolom selain title, media, link, dan category.

```
df_all = df_all.sort_values(by=['category', 'media'], ascending=True)
df_all = df_all.reset_index(drop=True)
df_all
```

Selanjutnya kita estrak berita dari semua link berita menggunakan library newspaper3k:

```
def get_text(url):
    article = Article(url, language='id')
    try:
        article.download()
        article.parse()
        return article.text
    except:
        print(f"Failed to download and parse {url}")
        return None

df_all['text'] = df_all['link'].apply(get_text)
```

sayangnya, dari 4 media, kita hanya berhasil estrak berita dari media Kompas saja. Oleh karena itu, untuk 3 media yang lain kita akan kurangi jadi 5 berita olahraga, 5 politik, dan 5 hiburan. Lalu kita estrak manual melalui link yang ada.

Scrap The News

Berikut merupakan 5 row pertama dari data final yang sudah kita scrap. Total berita yang kita miliki kini hanya 75 karena hanya media kompas saja yang berhasil kita scrap, sisa 3 media yang lain kita estrak sendiri sebanyak 15 berita per media. (30 Kompas, 15 tribun, 15 pikiran rakyat, 15 merdeka)

	title	media	link	text	category
0	Jungkook Rilis Album Solo Perdana, Golden Hala...	Kompas	https://www.kompas.com/hype/read/2023/11/03/09...	KOMPAS.com - Jungkook BTS merilis album solo p...	Hiburan
1	Tak Mau Disebut Serakah, Inara Jelaskan Alasan...	Kompas	https://www.kompas.com/hype/read/2023/11/03/08...	JAKARTA, KOMPAS.com- Istri penyanyi Virgoun, I...	Hiburan
2	Pernah Jadi Pacar Prilly Latuconsina, Kiki TBA...	Kompas	https://www.kompas.com/hype/read/2023/11/01/14...	JAKARTA, KOMPAS.com - Penyanyi Teuku Ryzki ata...	Hiburan
3	Polisi Ungkap Fakta Baru di Balik Kematian Mat...	Kompas	https://www.kompas.com/hype/read/2023/11/01/14...	KOMPAS.com - Tim penyidik kepolisian LA menjel...	Hiburan
4	Matthew Perry Ditemukan Meninggal di Bak Jaccu...	Kompas	https://www.kompas.com/hype/read/2023/10/29/17...	KOMPAS.com - Aktor Matthew Perry ditemukan men...	Hiburan
...
70	Kompak, 3 Parpol Lokal Aceh Dukung Anies - Gus...	Tribun News	https://www.tribunnews.com/mata-lokal-memilih/...	Tiga partai politik lokal Aceh mendeklarasikan...	Politik
71	PDIP Bongkar Sederet Elite Politik yang Diduga...	Tribun News	https://www.tribunnews.com/mata-lokal-memilih/...	Ketua DPP PDIP, Djarot Saiful Hidayat membocor...	Politik
72	PAN Tegaskan Presiden Jokowi Tak Pernah Campur...	Tribun News	https://www.tribunnews.com/mata-lokal-memilih/...	Sekretaris Jenderal (Sekjen) Partai Amanat Nas...	Politik
73	Gibran Buka Suara soal Tudingan Sengaja Pilih ...	Tribun News	https://www.tribunnews.com/mata-lokal-memilih/...	Wali Kota Solo, Gibran Rakabuming Raka, buka s...	Politik
74	TPN Bentuk Tim Pemenangan Muda Ganjar-Mahfud, ...	Tribun News	https://www.tribunnews.com/mata-lokal-memilih/...	Ketua Tim Pemenangan Nasional (TPN) Ganjar-Mah...	Politik

75 rows × 5 columns

Scrap The News

Untuk mengakses full code yang digunakan dalam proses mengscraping data, klik logo Google Colab berikut ini. Notebook ini sudah dilengkapi penjelasan untuk setiap cell codenya:

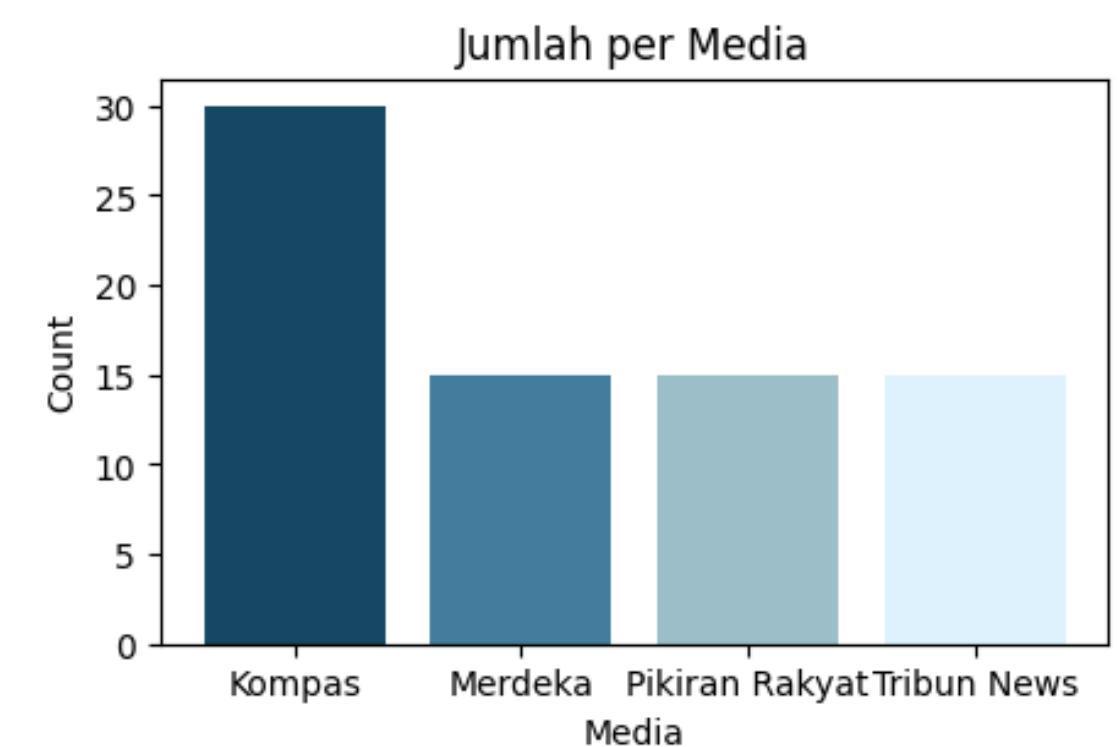
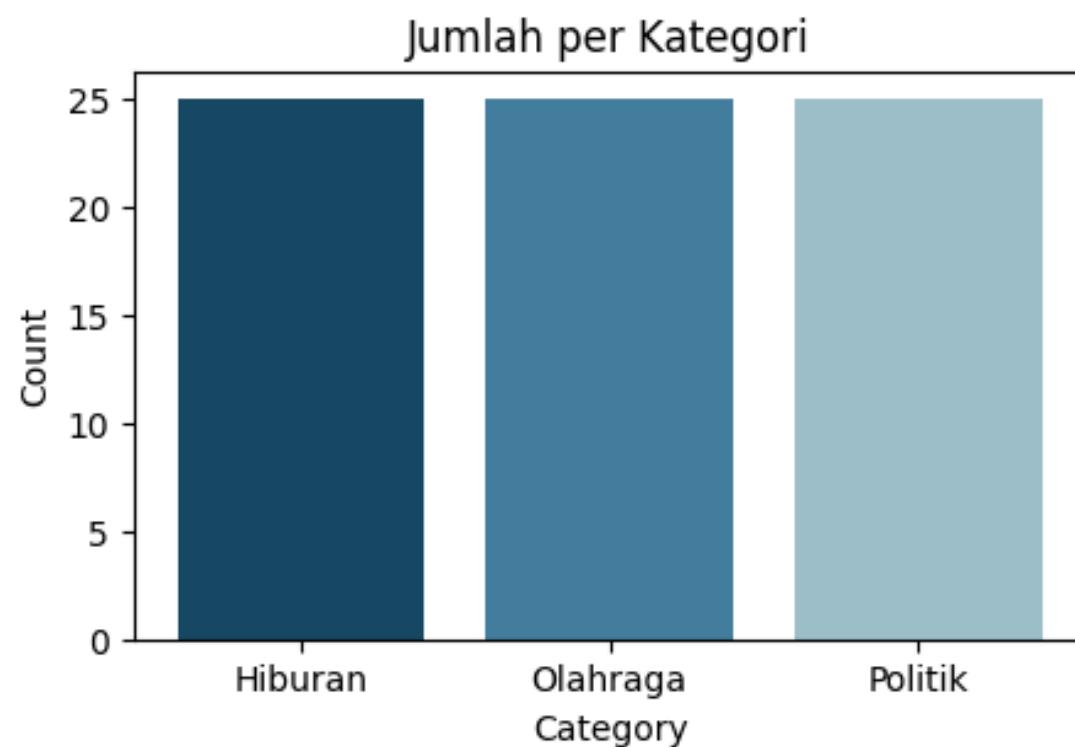


Exploratory Data Analysis (EDA)

EDA

Pada proses EDA, kita bisa mengetahui bahwa dataset kita sudah balance dan sudah tidak ada Null values sehingga data ini sudah baik dan akan kita proses ke tahap berikutnya.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 75 entries, 0 to 74
Data columns (total 5 columns):
 #   Column   Non-Null Count   Dtype  
 ---  --       --       --       --      
 0   title    75 non-null     object  
 1   media    75 non-null     object  
 2   link     75 non-null     object  
 3   text     75 non-null     object  
 4   category 75 non-null    object  
dtypes: object(5)
memory usage: 3.1+ KB
```



Text Preprocessing

Text Preprocessing

Pada tahap ini text cleaning yang akan dilakukan ada 7, yaitu:

1. Case Folding
2. Remove URL (co. <https://blabla.com>, www.blabla.com)
3. Remove Domain (co: Kompas.com, merdeka.com, blabla.co.id)
4. Remove Symbol (co: +/ - & * @ #)
5. Remove Slice, Tab, New space, dll
6. Remove Number
7. Remove Stopwords & extra spaces

Selain itu kita juga akan melakukan text preprocessing berikut:

1. Encoding label
2. Tokenisasi
3. Train test split

Text Preprocessing

1) Case Folding

```
dfClean['textLower'] = dfClean['text'].str.lower()
```

2) Remove URL

```
dfClean['textNoUrl']=[re.sub(r'((www\.[^\s]+)|(https?://[^\s]+)|(http?://[^\s]+)| ()', ' ', i)for i in dfClean['textLower']]
```

3) Remove Domain

```
dfClean['textNoDomain']=[re.sub(r'\S*\.\com\S*|\S*\.\co\.id\S*|\S*\.\id\S*', ' ', i)for i in dfClean['textNoUrl']]
```

4) Remove Symbol

```
dfClean['textNoSymbol']=[re.sub(r'[^w\s]', ' ', i)for i in dfClean['textNoDomain']]
```

5) Remove Tab

```
dfClean['textNoTab'] = [re.sub(r'\t|\n|\n\n', ' ', i) for i in dfClean['textNoSymbol']]
```

6) Remove Number

```
dfClean['textNoNumber'] = dfClean['textNoTab']
dfClean.loc[dfClean['category'] != 'Politik', 'textNoNumber'] = dfClean.loc[dfClean['category'] != 'Politik', 'textNoTab'].apply(lambda x: re.sub(r'\d+', ' ', x))
```

Text Preprocessing

7) Remove Stopwords & Extra Spaces

```
dfClean['textClean'] = dfClean['textNoNumber'].apply(lambda x: ' '.join([word for word in x.split() if word not in dfStopword.stopword.values]))
```

Selanjutnya kita drop semua kolom kecuali textClean dan category lalu simpan di dfAllClean dan lanjut kita process kembali

1) Ubah label kedalam bentuk numerik

```
dfAllClean['category'] = dfAllClean['category'].replace({'Hiburan': 0, 'Olahraga': 1, 'Politik': 2})
```

2) Tokenisasi

```
dfAllClean['token'] = dfAllClean['textClean'].apply(word_tokenize)
```

3) Train Test Split (80% Train 20% Test)

```
xTrain, xTest, yTrain, yTest = train_test_split(dfAllClean['token'], dfAllClean['category'], test_size = 0.2, random_state = 13, stratify=dfAllClean['category'])
```

Text Representation

Text Representation

Untuk project ini kita akan menggunakan 2 teknik teks representation yang berbeda. Yang pertama adalah skipgram dengan vector size = 50 dan minimal kemunculan token = 3. Kedua, kita akan gunakan teknik TF-IDF.

1) Skipgram

```
modelSkipgram = gensim.models.Word2Vec(xTrain, min_count = 3, vector_size = 50, window = 5, sg=1)

def sentence_vector(sentence, model):
    return np.mean([model.wv[word] for word in sentence if word in model.wv], axis=0)

xTrain_vector = [sentence_vector(sentence, modelSkipgram) for sentence in xTrain]
xTest_vector = [sentence_vector(sentence, modelSkipgram) for sentence in xTest]
```

Text Representation

2) TF-IDF

```
xTrain_str = [' '.join(words) for words in xTrain]  
xTest_str = [' '.join(words) for words in xTest]
```

```
vectorizer = TfidfVectorizer()  
xTrain_tfidf = vectorizer.fit_transform(xTrain_str)  
xTest_tfidf = vectorizer.transform(xTest_str)
```

```
TFIDF = pd.DataFrame(xTrain_tfidf.toarray(),columns=vectorizer.get_feature_names_out())  
TFIDF.head()
```

	09	10	11	12	13	154	16	17	18	19	...	zhejiang	ziarah	zikir	zodiac	zodiak	zona	zuhri	zulhas	zulkifli	zumba
0	0.0	0.0	0.00000	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	0.0	0.0	0.0	0.0
1	0.0	0.0	0.00000	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	0.0	0.0	0.0	0.0
2	0.0	0.0	0.00000	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	0.0	0.0	0.0	0.0
3	0.0	0.0	0.00000	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	0.0	0.0	0.0	0.0
4	0.0	0.0	0.03865	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	0.0	0.0	0.0	0.0

5 rows × 2838 columns

Train & Evaluate Model

Train & Evaluate Model

Selanjutnya kita akan buat model menggunakan SVM dan Random Forest untuk masing-masing teknik text representation (TF-IDF, Skipgram)

1) SVM - Skipgram

```
modelSVM = svm.LinearSVC( random_state=13)
modelSVM.fit(xTrain_vector, yTrain)
predictSVM = modelSVM.predict(xTest_vector)
print(classification_report(yTest, predictSVM, target_names=['0','1','2']))
```

	precision	recall	f1-score	support
0	0.75	0.60	0.67	5
1	1.00	0.40	0.57	5
2	0.44	0.80	0.57	5
accuracy			0.60	15
macro avg	0.73	0.60	0.60	15
weighted avg	0.73	0.60	0.60	15

2) Random Forest - Skipgram

```
modelRF = RandomForestClassifier(random_state=13)
modelRF.fit(xTrain_vector, yTrain)
predictRF = modelRF.predict(xTest_vector)
print(classification_report(yTest, predictRF, target_names=['0','1','2']))
```

	precision	recall	f1-score	support
0	0.75	0.60	0.67	5
1	1.00	1.00	1.00	5
2	0.67	0.80	0.73	5
accuracy			0.80	15
macro avg	0.81	0.80	0.80	15
weighted avg	0.81	0.80	0.80	15

Train & Evaluate Model

1) SVM - TF-IDF

```
modelSVMtfidf = svm.LinearSVC( random_state=13)
modelSVMtfidf.fit(xTrain_tfidf, yTrain)
predictSVMtfidf = modelSVMtfidf.predict(xTest_tfidf)
print(classification_report(yTest, predictSVMtfidf, target_names=['0','1','2']))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5
1	1.00	0.80	0.89	5
2	0.83	1.00	0.91	5
accuracy			0.93	15
macro avg	0.94	0.93	0.93	15
weighted avg	0.94	0.93	0.93	15

2) Random Forest - TF-IDF

```
modelRFTfidf = RandomForestClassifier(random_state=13)
modelRFTfidf.fit(xTrain_tfidf, yTrain)
predictRFTfidf = modelRFTfidf.predict(xTest_tfidf)
print(classification_report(yTest, predictRFTfidf, target_names=['0','1','2']))
```

	precision	recall	f1-score	support
0	0.83	1.00	0.91	5
1	1.00	1.00	1.00	5
2	1.00	0.80	0.89	5
accuracy			0.93	15
macro avg	0.94	0.93	0.93	15
weighted avg	0.94	0.93	0.93	15

Klik logo Google Colab berikut untuk mengakses full code mulai dari EDA hingga Evaluate Model:



Summary

Summary

Model	Accuracy
SVM - Skipgram	0.60
Random Forest - Skipgram	0.80
SVM - TF-IDF	0.93
Random Forest - TF-IDF	0.93

Berdasarkan accuracynya dapat kita simpulkan bahwa **model SVM dan RF dengan vector yang dihasilkan dari TF-IDF memiliki accuracy yang paling besar dan keduanya sama-sama besar**, hal ini mungkin dikarenakan TF-IDF memberikan bobot pada setiap kata dalam dokumen berdasarkan frekuensi kemunculannya dalam dokumen tersebut dan di seluruh kumpulan dokumen. Kata-kata yang sering muncul dalam satu dokumen tetapi jarang muncul di dokumen lain akan mendapatkan bobot yang lebih tinggi. Hal ini dapat membantu model dalam mengidentifikasi fitur penting dan membedakan antara kategori. Sementara disisi lain, model Skipgram menghasilkan representasi vektor kata yang berdasarkan konteks di mana kata tersebut muncul, dan mungkin tidak seefektif TF-IDF dalam kasus klasifikasi teks ini.

THANK YOU!

