Nama : Agnes Leady Octaviana NIM : 1301160063

Latihan

1. Ubah format penyimpanan data ke CSV
2. Coba buatkan feature berikut (save dan upload feature), lalu laporkan pengaruhnya terhadap akurasi klasifikasi: a. Tanpa proses normalisation b. Tanpa proses lemmatisation c. Tanpa menghilangkan stopwords
3. Coba buat tfidf dengan nilai "max_features" yang berbeda-beda (lebih besar dan lebih kecil dari 300), lalu laporkan pengaruhnya terhadap akurasi klasifikasi.
4. Coba dengan beberapa algoritma klasifikasi yang berbeda (minimal 2 algoritma), carilah parameter terbaik (jelaskan nilai2 parameter yang telah dicoba untuk tiap jenis algoritma).
5. Jika anda ingin menggunakan teks bahasa Indonesia, bagian mana saja yang perlu dilakukan penyesuaian?
6. Opsional: Gunakan word embedding (e.g word2vec, GloVe).

Jawab :

1.
```
#mengubah file menjadi csv
In [60]: df.to_csv("Data/X_train.csv", sep=';',index=False, encoding='utf-8')
In [61]: df.to_csv("Data/X_test.csv", sep=';',index=False, encoding='utf-8')
In [66]: df.to_csv("Data/y_train.csv", sep=';',index=False, encoding='utf-8')
In [67]: df.to_csv("Data/y_test.csv", sep=';',index=False, encoding='utf-8')
In [68]: df.to_csv("Data/df.csv", sep=';',index=False, encoding='utf-8')
In [69]: df.to_csv("Data/features_train.csv", sep=';',index=False, encoding='utf-8')
In [70]: df.to_csv("Data/labels_train.csv", sep=';',index=False, encoding='utf-8')
In [71]: df.to_csv("Data/features_test.csv", sep=';',index=False, encoding='utf-8')
In [73]: df.to_csv("Data/labels_test.csv", sep=';',index=False, encoding='utf-8')
In [72]: df.to_csv("Data/tfidf.csv", sep=';',index=False, encoding='utf-8')
```

2. Dampak yang akan diberikan terhadap akurasi, jika:
    a. Tanpa Proses Normalisation
        - Maka nilai akurasi akan menjadi lebih rendah jika dibandingkan dengan data yang menggunakan normalisasi, karena dengan adanya normalisasi makan data menjadi layak untuk di olah karena akan memperbaiki kata kata yang mungkin saja memiliki kesalahan ketika diketik.

Before Normalisation	After Normalisation
'Japan narrowly escapes recession\n\nJapan\'s economy teetered on the brink of a technical recession in the three months to Septembe r, figures show.\n\nRevised figures indicated growth of just 0.1% - and a similar-sized contraction in the previous quarter. On an a nnual basis, the data suggests annual growth of just 0.2%, suggest ing a much more hesitant recovery than had previously been though t. A common technical definition of a recession is two successive quarters of negative growth.\n\nThe government was keen to play do wn the worrying implications of the data. "I maintain the view tha t Japan\'s economy remains in a minor adjustment phase in an upwar	"Japan narrowly escapes recession Japan's economy teetered on the brink of a technical recession in the three months to September, f igures show. Revised figures indicated growth of just 0.1% - and a similar-sized contraction in the previous quarter. On an annual basis, the data suggests annual growth of just 0.2%, suggesting a much more hesitant recovery than had previously been thought. A co mmon technical definition of a recession is two successive quarter s of negative growth. The government was keen to play down the wo rrying implications of the data. I maintain the view that Japan's

b. Tanpa Proses Lemmatisasi
    - Maka akan ada kata yang dihitung berkali-kali meski maknanya sama dan akan berdampak pada nilai akurasi yang sangat kecil dibandingkan dengan yang melakukan lematisasi pada pre-processing.

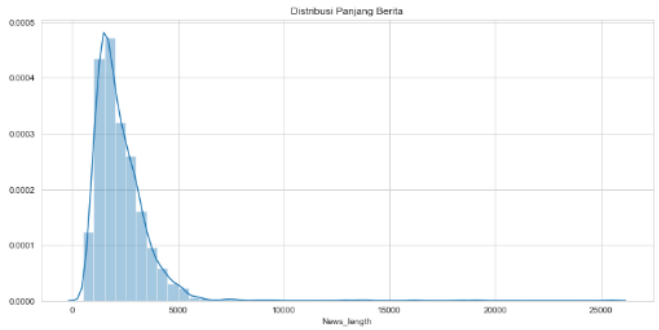
c. Tanpa Menghilangkan Stopwords
    - Maka nilai akurasi akan sangat rendah, karena penggunaan stopword adalah untuk menghapus kata yang memiliki nilai informasi yang rendah pada teks, dan juga sering muncul.

3. a. Untuk max_features kurang dari 300 dengan max_feature = 150
   b. Untuk max_features kurang dari 300 dengan max_feature = 450
4. -
5. Bagian yang memerlukan penyesuaian jika menggunakan teks berbahasa indonesia adalah:
    - Stemming, dan menggunakan sastrawi untuk menjadikan kata yang berimbuhan menjadi kata dasarnya.
    - Menghapus penggunaan possessive pronouns.

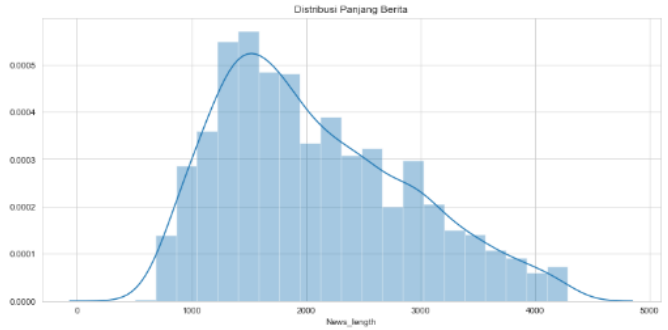
➢ Visualisasi panjang data dengan histogam

```
df['News_length'] = df['Content'].str.len()

plt.figure(figsize=(12.8,6))
sns.distplot(df['News_length']).set_title('Distribusi Panjang Berit
```

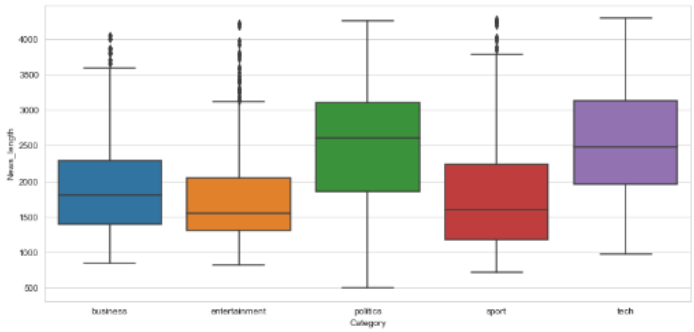


```
plt.figure(figsize=(12.8,6))
sns.distplot(df_95['News_length']).set_title('Distribusi Panjang Be
```



- Dalam percobaan kali ini kita dapat melihat visualisasi dari pendistribusian data dengan estimasi kepadatan dari data.

```
plt.figure(figsize=(12.8,6))
sns.boxplot(data=df_95, x='Category', y='News_length');
```



- Dengan max_features = 300 memberikan nilai akurasi sebesar 94% dan mengetahui performansi base model dengan parameter default.

```
# Classification report
print("Classification report")
print(classification_report(labels_test,classifier_pred))
```

```
Classification report
              precision    recall  f1-score   support

           0       0.92      0.95      0.93        81
           1       0.90      0.96      0.93        49
           2       0.96      0.89      0.92        72
           3       0.99      0.99      0.99        72
           4       0.93      0.92      0.92        60

    accuracy                           0.94       334
   macro avg       0.94      0.94      0.94       334
weighted avg       0.94      0.94      0.94       334
```