ANALISIS SENTIMENT BERITA ONLINE (SITUS : CNN)

1. Jelaskan Metode Anda untuk preprocessing data! (labelling, case folding, remove punctuation, remove new line, stopword)
   - Labelling

     Untuk labelling, dilakukan dengan polarity yang ada pada package TextBlob. Sebelum dihitung polarity untuk masing-masing baris, dipastikan terlebih dahulu bahwa pada dataset tidak terdapat *missing value.* Setelah didapatkan nilai polarity maka nilai tersebut akan didefinisikan sedemikian rupa dimana apabila nilai polarity = 0 maka akan masuk ke kelas sentimen neutral, -1 < nilai polarity < 0 maka akan masuk ke kelas sentimen negatif dan apabila 0 < nilai polarity < 1 maka akan masuk ke kelas sentimen positif.

```
def polarity_to_label(x):
    if(x >= -1 and x < 0):
        return 'neg'
    if(x == 0):
        return 'neutral'
    if(x > 0 and x <= 1):
        return 'pos'
df.label = df.label.apply(polarity_to_label)
```

   - Case Folding

     Case folding dilakukan dengan mengubah seluruh huruf pada dataset menjadi huruf kecil (*lower case*)

```
#lower text
data=data.lower()
```

   - Remove Punctuation

     Remove punctuation dilakukan untuk menghilangkan tanda-tanda baca pada dataset

```
remove = string.punctuation
translator = str.maketrans(remove, ' '*len(remove))
data = data.translate(translator)
```

   - Remove New Line

     Remove new line dilakukan untuk menghilangkan garis-garis spasi saat teks ada di paragraf baru

```
#remove new line
data = data.replace('\n', ' ')
```

   - Stopword

     Stopword digunakan untuk menghilangkan kata-kata umum yang terdapat pada dataset. Hal ini dilakukan agar saat analisa data yang digunakan lebih valid dan tidak terdapat banyak bias. Untuk analisis sentimen pada data ini, digunakan list stopword dari library Sastrawi.

```
#stopword
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

factory = StopWordRemoverFactory()
stopword = factory.create_stop_word_remover()

review = []
for index, row in df.iterrows():
  review.append(stopword.remove(row["content"]))
```

2. Jelaskan Library Bahasa Indonesia yang dipakai untuk analisis sentimen untuk pembersihan data. Mengapa Anda memilih itu?
   Libary Bahasa Indonesia untuk pembersihan data khusunya untuk stopword yaitu dengan library Sastrawi. Library Sastrawi digunakan karena library tersebut berasal dari Indonesia sehingga diharapkan list stopword lebih banyak dan lebih relevan dibandingkan library lainnya.
3. Jelaskan mengenai algoritma yang dipakai untuk analisa sentimen. Mengapa anda memilih algoritma itu? Berapa akurasinya?
   Untuk analisa sentimen digunakan Naive Bayes Classifier (NBC). Alasan pemilihan NBC adalah NBC telah banyak digunakan dalam text mining karena memiliki kelebihan yaitu algoritma yang sederhana tapi menghasilkan akurasi yang tinggi.
   Hasil akurasi dari model analisis sentimen yang telah dilakukan yaitu 79%

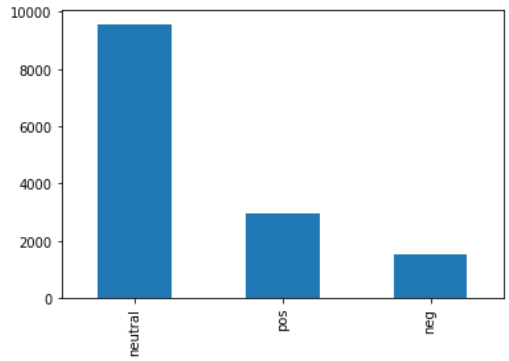
```
#akurasi model
from sklearn.metrics import classification_report
print(classification_report(y_test, pred))
```

```
              precision    recall  f1-score   support

         neg       0.81      0.85      0.83      1000
     neutral       0.74      0.81      0.78      1000
         pos       0.83      0.71      0.77      1000

    accuracy                           0.79      3000
   macro avg       0.80      0.79      0.79      3000
weighted avg       0.80      0.79      0.79      3000
```

4. Buat visualisasi data sederhana dari sentimen analisa yang dihasilkan
   - Bar Plot banyak kelas pada label sentimen



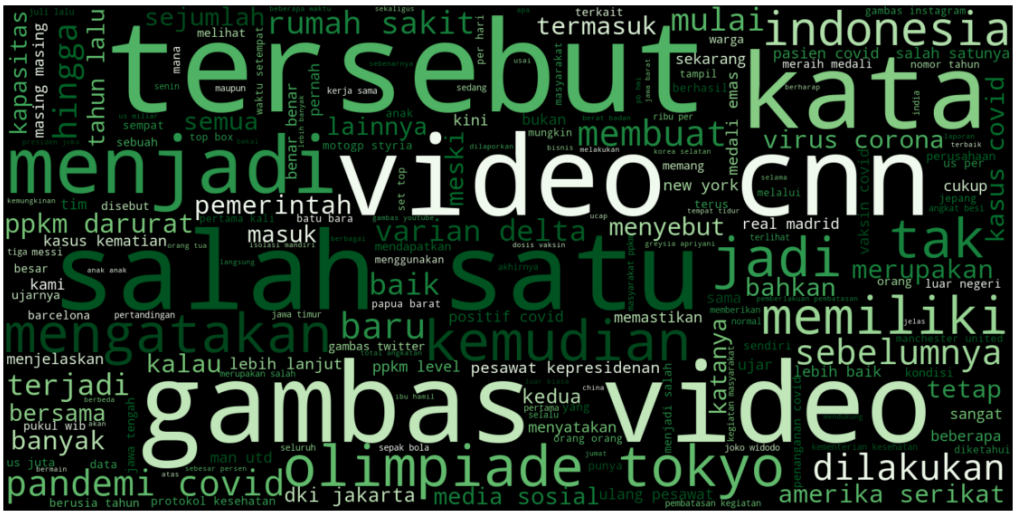
   Pada data asli, diketahui bahwa sentimen terbanyak dari berita adalah sentimen netral sebanyak 9558 berita, dan sentimen yang paling sedikit adalah sentimen negatif sebanyak 1512 berita. Untuk sentimen positif sebanyak 2949 berita.
   Namun demikian, untuk analisis lebih lanjut dilakukan resampling untuk tiap-tiap kelas pada label sentimen berita karena terjadi imbalanced pada dataset asli.

- Wordcloud untuk masing-masing kelas pada label sentimen

Kelas sentimen negatif

Kelas sentimen netral

Kelas sentimen positif



Link Source Code :

https://github.com/Novitadwiutami/Text-Mining-Analisis-Sentimen-Situs-Berita