DATA MINING

“Text Classification”

Dosen Pengampu

Amalia Anjani Arifiyanti, S.Kom, M.Kom

Disusun Oleh:

Muhammad Nizar Zulmi (18082010013)

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”

JAWA TIMUR

2020

1. **Judul**

   EPS 8 Text Classification

2. **Tujuan**

   Mempelajari dasar Text Classification

3. **Teori Dasar**

   Text classification atau text categorization merupakan salah satu bentuk implementasi text mining. Pada bagian awal akan digunakan pemrosesan NLP dan membentuknya ke dalam bentuk terstruktur (dalam bentuk vectorized) sehingga dapat diolah dengan menggunakan teknik klasifikasi pada task data mining. Proses klasifikasi sama seperti proses klasifikasi yang telah dipelajari pada modul klasifikasi.

4. **Tools & Software**
   a. **Anaconda**
   b. **Phython**
   c. **Google collab**
   d. **Spyder**
   e. **Jupyter Notebook**
5. **Langkah Kerja**
   a. Pertama buka Anaconda Navigator
   b. Setelah anaconda navigator terbuka kita akan memilih ingin menggunakan ide apa
   c. Terdapat beberapa macam pilihan IDE kita pilih jupyter Notebook
   d. Lalu jendela browser akan otomatis terbuka,pilih new kemudian pilih Pyhthon 3 dan lakukan proses coding sesuai dengan Tugas yang diberikan
   e. Tidak lupa mempersiapkan file csv bernama xiaomi.csv

6. Latihan Mahasiswa

1. Jelaskan perubahan pada setiap praproses

a. Source Code

- **Data Awal**

```
In [55]: df=pd.read_csv('E:/UPN JATIM/1. PERKULIAHAN/SEMESTER 6/Data Mining ( Bu Amel )/Week 11/xiaomi2019.csv', sep=';', encoding='ISO-88
         df.head()
```

Out[55]:

	tweet	sentiment
0	pake hp xiaomi bisa kan	Positive
1	Xiaomi yi action kamera bagus juga buat video bisa di jadikan pertimbangan nihhh	Positive
2	Ya Allah jauhkanlah aku dari godaan clickbait line today dan berita-berita di browser xiaomi	Positive
3	hpmu opo se? kok sawangane jernih koyok iph â□□ Xiaomi euy:' https://lap78.ask.fm/igoto/45DKECPW7B667HQMHN2IG6NM7SOD5OAUS7QPPAN4D7ROV45V2Q24OJAMGBFM2RRQK2272FYJJNWDWXQVYYU5Y25C...	Positive
4	numpang nanya itu hapenya xiaomi bukan ya? Kalo iya tipe apa? Jernih kameranya mirip iphone	Positive

- **Data setelah dilakukan filtering text**

```
In [6]: def clean_text(text):
            # mengubah semua karakter hurud menjadi huruf kecil
            text = text.lower()
            # menghilangkan nama akun
            text = re.sub('@[^\s]+','',text)
            # menghilangkan punctuation
            text = re.sub('[%s]' % re.escape(string.punctuation),'',text)
            # menghilangkan angka
            text = re.sub('\w*\d\w*','',text)
            # menghilangkan url
            text = re.sub(r'\w+:\/{2}[\d\w-]+(\.[\d\w-]+)(?:(?:\/[^\s/]))*','',text)
            text = re.sub(r'(https?:\/\/)?([\da-z\.-]+)\.([a-z\.]{2,6})([\/\w\.-])\/?\S','',text)
            # menghilangkan hastag
            text = re.sub('#[^\s]+','',text)
            # menghilangkan huruf tunggal
            text = re.sub(r'\b[a-zA-Z]\b','',text)
            return text

        clean = lambda x: clean_text(x)

        dfx = pd.DataFrame(df.tweet.apply(clean))
        dfx
```

Out[6]:

	tweet
0	pake hp xiaomi bisa kan
1	xiaomi yi action kamera bagus juga buat video bisa di jadikan pertimbangan nihhh
2	ya allah jauhkanlah aku dari godaan clickbait line today dan beritaberita di browser xiaomi
3	hpmu opo se kok sawangane jernih koyok iph â□□ xiaomi euy â□¦
4	numpang nanya itu hapenya xiaomi bukan ya kalo iya tipe apa jernih kameranya mirip iphone
...	...
96	xiaomi bagus ka hehe
97	baru tahu kalo xiaomi bisa screen record wkwkwk
98	bgst mau nanya disini ada yg xiaomi mi susah cari sinyal ga semenjak update ke patch juni
99	hp aku gitu dulu xiaomi tapi ga sampe ke konter ku cari di google alhamdulillah bisa itu siapa tau bootlop nder kalo bootlop gausah ke konter sayang uangnya
100	mau kalo ada xiaomi note yg ga finger print soalnya jarang bgt

101 rows × 1 columns

- **Dilakukan tahapan filtering dengan punctuation**

In [7]:
```
def Punctuation(string):
    #punctuation marks

    punctuations = '''!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~'''

    #traverse the given string and if punctuation
    #marks occur replace it with null
    for x in string.lower():
        if x in punctuations:
            string = string.replace(x, "")

    #print string without punctuation
    return(string)
cleanPunc = lambda x: Punctuation(x)

dfx = pd.DataFrame(dfx.tweet.apply(cleanPunc))
dfx.head()
```

Out[7]:

	tweet
0	pake hp xiaomi bisa kan
1	xiaomi yi action kamera bagus juga buat video bisa di jadikan pertimbangan nihhh
2	ya allah jauhkanlah aku dari godaan clickbait line today dan beritaberita di browser xiaomi
3	hpmu opo se kok sawangane jernih koyok iph â□□ xiaomi euy â□¦
4	numpang nanya itu hapenya xiaomi bukan ya kalo iya tipe apa jernih kameranya mirip iphone

- **Dilakukan tahapan stemming dengan *stopword removal***

In [8]:
```
#D. Penghapusan Stopword
def get_stopword(stopwordsfile):
    stopwords=[]
    file_stopwords =  open(stopwordsfile,'r')
    row = file_stopwords.readline()
    while row:
        word = row.strip ()
        stopwords.append(word)
        row = file_stopwords.readline()
        file_stopwords.close()
        return stopwords
stop_words_indo= get_stopword('E:/UPN JATIM/1. PERKULIAHAN/SEMESTER 6/Data Mining ( Bu Amel )/Week 11/stopwordsindo.txt')
```

In [9]:
```
def stopwords(text):
    tokens = word_tokenize(text)
    filtered=[]

    for w in tokens:
        if w not in stop_words_indo:
            filtered.append(w)

    hasil =' '.join(filtered)
    return hasil
st=lambda x: stopwords(x)

dfx=pd.DataFrame(dfx.tweet.apply(st))
dfx.head()
```

Out[10]:

	tweet
0	pake hp xiaomi bisa kan
1	xiaomi yi action kamera bagus juga buat video bisa di jadi timbang nihhh
2	ya allah jauh aku dari goda clickbait line today dan beritaberita di browser xiaomi
3	hpmu opo se kok sawangane jernih koyok iph xiaomi euy
4	numpang nanya itu hapenya xiaomi bukan ya kalo iya tipe apa jernih kamera mirip iphone

- **Penjelasan Tahap Perubahan Data**

Pada saat dilakukan import data Csv yaitu data xiaomi2019.csv bisa dilihat data tersebut memiliki unque karakter yang tidak dibutuhkan lalu dilakukan clean data dengan menggunakan *clean_text* dan setelah dilakukan filter karakter yang tidak dibutuhkan maka data yang dihasilkan menjadi lebih tertata,kemudian dilanjutkan pada tahapan punctuation dimana fungsi dari punctuation sendiri adalah untuk menghapus symbol yang tidak dibutuhkan pada data tersebut ,lalu untuk tahapan terakhir adalah menggunakan stemming ,dimana stemming sendiri digunakan untuk membersihkan karakter yang tidak dapat dikenali oleh program.

2. Evaluasi model pada contoh hanya akurasi, silahkan buat confusion matrixnya dan hitung dengan metric yang lain misalnya precision, recall, ROC/AUC, dsb.

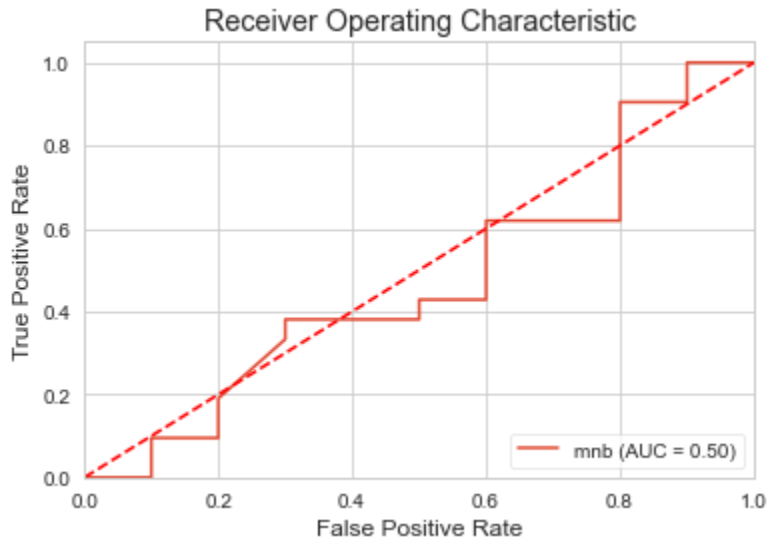
a. Source code dan Hasil

```
In [37]: #Confusion Matrix
         print("Confussion_Matrix")
         confusion_matrix(y_test,y_pred)

         Confussion_Matrix

Out[37]: array([[ 0, 10],
                [ 0, 21]], dtype=int64)
```

```
mnb_roc_auc = roc_auc_score(y_test, y_pred)
mnb_fpr, mnb_tpr, mnb_threesholds = roc_curve(y_test, mnb.predict_proba(x_test_vect)[:,1])
plt.figure()
plt.plot(mnb_fpr, mnb_tpr, label='mnb (AUC = %0.2f)' % mnb_roc_auc)
plt.plot([0,1],[0,1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()
```



- **Penjelasan**
  Pada tahap diatas dilakukan perhitungan dari confusion matrix nya dan juga dilakukan perhitungan menggunakan matrix Reciever Operating Characteristic (ROC) sehingga untuk confusion matrix menghasilkan hasil [0,10] dan [0,21]

3. Silahkan dicoba jika menggunakan cross-validation, dan hitung evaluasinya.

   a. **Source Code dan Hasil**

In [54]:
```
print ("Cross Val Akurasi")
scores_accuracy = cross_val_score(mnb, X_test_vect, y_pred, cv=5, scoring = "accuracy")
print(scores_accuracy)
print("Rata-rata nilai akurasi: %0.2f (+/- %0.2f)" % (scores_accuracy.mean(), scores_accuracy.std()))

print ("==========================")
print ("Corss Val Precission Macro")
scores_precision = cross_val_score(mnb, X_test_vect, y_pred, cv=5, scoring = "precision_macro")
print(scores_precision)
print("Rata-rata nilai precision macro: %0.2f (+/- %0.2f)" % (scores_precision.mean(), scores_precision.std()))

print ("==========================")
print ("Recall Macro")

scores_recall = cross_val_score(mnb, X_test_vect, y_pred, cv=10, scoring = "recall_macro")
print(scores_recall)
print("Rata-rata nilai recall macro: %0.2f (+/- %0.2f)" % (scores_recall.mean(), scores_recall.std()))

print ("==========================")
print ("F1 Macro")
scores_f1 = cross_val_score(mnb, X_test_vect, y_pred, cv=5, scoring = "f1_macro")
print(scores_f1)
print("Rata-rata nilai f1 macro: %0.2f (+/- %0.2f)" % (scores_f1.mean(), scores_f1.std()))
```

```
Cross Val Akurasi
[1. 1. 1. 1. 1.]
Rata-rata nilai akurasi: 1.00 (+/- 0.00)
==========================
Corss Val Precission Macro
[1. 1. 1. 1. 1.]
Rata-rata nilai precision macro: 1.00 (+/- 0.00)
==========================
Recall Macro
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Rata-rata nilai recall macro: 1.00 (+/- 0.00)
==========================
F1 Macro
[1. 1. 1. 1. 1.]
Rata-rata nilai f1 macro: 1.00 (+/- 0.00)
```

- **Penjelasan**

  Pada data diatas dilakukan cross validation dan menghasilkan hasil evaluasi sebagai berikut

- Untuk Nilai Akurasi Mendapatkan Nilai 1.00
- Untuk Nilai Precission Macro 1.00
- Untuk Nilai Rcall Macro 1.00
- Untuk Nilai F1 Macro 1.00