Jeffry Gotama – 2201816392

Fred Michael – 2201838380

Dokumentasi Project NLP
Detecting Hoax News Indonesian Society using Binarized Naïve Bayes

1. Library

Dalam mengerjakan project ini, digunakan bahasa python dan library:

a. pandas
library ini digunakan untuk melakukan *read file* pada dataset *hoax-valid.csv*

b. re
library ini menyediakan operasi pencocokan ekpresi regular seperti yang ditemukan di perl

c. nltk
- ✓ nltk.corpus import stopwords
  library ini digunakan untuk men-*download* stopwords yang untuk menghilangkan kata kata yang umum yang tidak berguna untuk pembelajaran (learning) yang memungkinkan untuk mengurangi token dalam dokumen secara signifikan dan dengan demikian dapat mengurangi dimensi fitur
- ✓ nltk.stem.porter import PoerterStemmer
  library yang mengacu pada proses heuristic dimana memotong ujung kata-kata yang sering kali menghilangkan afiks derivasional

d. sklearn
- sklearn.naive_bayes import BernoulliNB
  pada projek ini digunakan library BernoulliNB untuk membuat model dan mencoba untuk melakukan *testing* pada model
- sklearn.model_selection import train_test_split
  dataset dibagi menjadi dua, yaitu *Training Dataset* (80%) dan *Testing Dataset* (20%)
- sklearn.feature_extraction.text import CountVectorizer
  library ini digunakan untuk mengoversi kumpulan dokumen text menjadi matrix token (*matrix of token counts*).
- sklearn.metrics import confusion_matrix
  library ini digunakan untuk membuat *confusion matrix*

e. seaborn
library ini memberikan grafik secara visual dan menyediakan grafik statistil yang menarik dan informatif. Library ini akan digunakan untuk memnberikan *interface* yang menarik pada *confusion matrix*

f. matplotlib (matplotlib.pyplot)
library ini memberikan *output* secara visual dari hasil library seaborn yang dipakai.

2. Algorithms

Pada projek ini akan dipakai algorithma Bernoulli (Binarized) Naive Bayes yang berasal dari library sklearn.naive_bayes.BernoulliNB

Fitting Naive Bayes to the Training set

```
In [143]: # fitting naive bayes to the training set
          from sklearn.naive_bayes import BernoulliNB

In [144]: classifier = BernoulliNB()
          classifier.fit(x_train, y_train)

Out[144]: BernoulliNB(alpha=1.0, binarize=0.0, class_prior=None, fit_prior=True)
```

Bernoulli naïve bayes menggunakan fraksi dari setiap kelas dokumen yang berisi kata-kata (*terms*), saat melakukan klasifikasi (*training model*), bernoulliNB menggunakan informasi biner, yang berarti menghasilkan indicator 1 untuk setiap istilah (*terms*) dalam kosa kata pada *training* dokuments atau indikator 0, yang mengindikasikan tidak adanya istilah (*terms*) dalam *training dokuments.* BernoulliNB mengabaikan jumlah kejadian atau banyak kata yang muncul.

Parameter BernoulliNB:

a. alpha = 1.0
   menambahkan *add 1 Smoothing*

b. binarize = 0.0
   batas untuk pemetaan ke Boolean dari fitur dataset.

c. class_prior = None
   probabilitas kelas sebelumnya, jika ditentukan, probabiliats sebelumnya disesuaikan menurut dataset yang ada

d. fit_prior = True
   apakah akan mempelajari probabilitas kelas sebelumnya, atau tidak.

(* = untuk lebih lengkap yang dimulai dari preprocessing text sampai pembuatan model, dapat ditemukan pada file *Bernoulli_Naive-Bayes.ipynb*)

3. Evaluasi

Berikut hasil *accuracy* dari pelatihan (*training*) *model* dan *testing model*

```
In [194]: classifier.score(x_train, y_train)

Out[194]: 0.6145833333333334

In [195]: # predicting test set results
          y_pred = classifier.predict(x_test)

In [196]: classifier.score(x_test, y_test)

Out[196]: 0.6416666666666667
```

Dapat dilihat pada pelatihan model, didapat accuracy 61.5% dan pada saat *testing model* didapat *accuracy* model 64.2%.

Berikut ini juga merupakan *performance evaluation* (Precision, recall, f1-score) dari hasil *testing model*

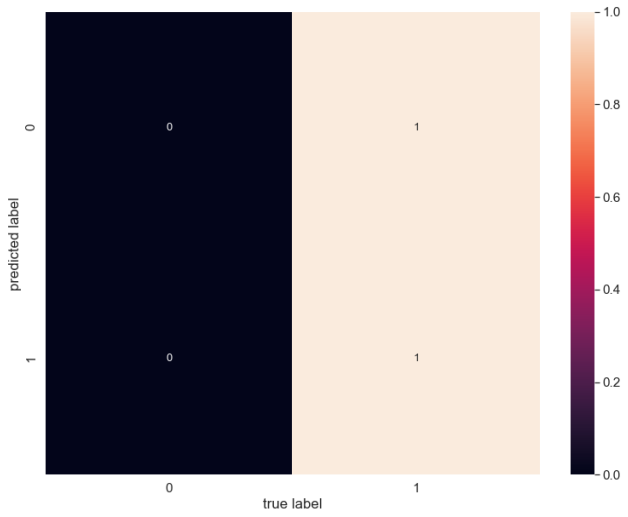
```
In [199]: print(metrics.classification_report(y_test, y_pred, digits=3))
                  precision    recall  f1-score   support

            Hoax      0.000     0.000     0.000        43
           Valid      0.642     1.000     0.782        77

        accuracy                          0.642       120
       macro avg      0.321     0.500     0.391       120
    weighted avg      0.412     0.642     0.502       120
```

4. Screenshot

Confusion Matrix dari model



5. Kesimpulan

Berdasarkan hasil evaluasi yang ada, dapat disimpulkan bahwa model yang didapat dari model BernoulliNB tidak *overfitting* (*standard*) karena accuracy pada *testing model* lebih besar pada saat melatih model serta dihasilkan accuracy model = 64.2%

Model ini masih dapat diperluas (*more training*) dengan cara menambahkan jumlah dataset yang ada (*only* 600 *records*), sehingga model ini masih belum bisa disebut *generalized model*

6. Daftar Pustaka

https://data.mendeley.com/datasets/p3hfgr5j3m/1
https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html
https://nlp.stanford.edu/IR-book/html/htmledition/the-bernoulli-model-1.html
https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html