Reviews_binary_classifier_onevsall_v3_1_final

October 13, 2022

1 Decision Support Model for Prioritizing Software Features with Value Delivery to the GOV.BR App

2 FPO / PPCA-2022 / UNB

2.1 Importing libraries

[323]:
```
#!pip install pydot scikit-plot  graphviz xgboost xgboost #test
```

[324]:
```
from urllib.parse import urlencode
from imblearn.under_sampling import RandomUnderSampler
from imblearn.pipeline import Pipeline
from gensim.models import KeyedVectors
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.corpus import stopwords
from sklearn.dummy import DummyClassifier
from imblearn.over_sampling import SMOTE
from sklearn.metrics import f1_score
from sklearn import metrics
from collections import Counter
import pandas as pd
import numpy as np
import os
import nltk
import string
nltk.download("punkt")
nltk.download("stopwords")
```

```
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

[324]: True

2.2 ## 1 - Prepare Data

Prepare classified data.

[325]:
```
url = "https://raw.githubusercontent.com/vanerven/fpo-app-review/main/data/
↪govbr-fuzzy-ahp-classified.csv"
classified_data_original = pd.read_csv(url, sep=',', encoding='ISO-8859-1')
classified_data_original.dropna(subset = ['Content'], inplace=True)
classified_data = classified_data_original.query("not Feature.isnull() and not␣
↪Heuristic.isnull()", engine="python")
classified_data['Content'].dropna(inplace=True)
classified_data['Content'].str.strip()
classified_data['Feature'].str.strip()
classified_data['Heuristic'].str.strip()
classified_data
```

[325]:

	Content	Score	Feature	Heuristic
0	Não tem 2 minutos que baixei o aplicativo, vou…	1	F1 - User access	C5 - Error prevention
1	Resolve um problema, aparece outro. No app ped…	1	F0 - None feature	C5 - Error prevention
2	O erro é exatamente o que já descrevi, e que v…	1	F1 - User access	C5 - Error prevention
3	Para acessar são duas etapas, senha e código e…	1	F1 - User access	C1 - Visibility of system status
4	Consegui fazer reconhecimento com uma luminári…	1	F1 - User access	C4 - Consistency and standards
..	…	…	…	…
147	Quando meu aparelho está com TODOS aplicativos…	5	F5 - User data	C3 - User control and freedom
148	App de excelência nota 10 pra quem busca prati…	5	F2 - Public services	C7 - Flexibility and efficiency of use
149	aplicativo de reconhecimento de dados pessoais…	5	F5 - User data	C7 - Flexibility and efficiency of use
150	Bem funcional, facilitando o processo de ident…	5	F4 - Digital signature	C4 - Consistency and standards
151	Assinatura eletrônica, facilita meu dia em 100%.	5	F4 - Digital signature	C7 - Flexibility and efficiency of use

[152 rows x 4 columns]

Prepare not classified data.

[326]:
```
url = "https://raw.githubusercontent.com/vanerven/fpo-app-review/main/data/
 ↪govbr-fuzzy-ahp-not-classified.csv"
not_classified_data_original = pd.read_csv(url, sep=',', encoding='ISO-8859-1')
not_classified_data_original.dropna(subset = ['Content'], inplace=True)
not_classified_data = not_classified_data_original.query("Feature.isnull() and␣
 ↪not Score.isnull() or Heuristic.isnull() and not Score.isnull()",␣
 ↪engine="python")
not_classified_data.drop(['Feature', 'Heuristic'],  axis=1).query(' not Content.
 ↪isna() ', engine="python", inplace=True)
not_classified_data['Content'].dropna(inplace=True)
not_classified_data['Content'].str.strip()
not_classified_data
```

[326]:

	Content	Score	Feature	Heuristic
0	Aplicativo péssimo assim como todos os aplicat…	1	NaN	NaN
1	App lixo, reconhecimento fácil não funciona só…	1	NaN	NaN
2	O app é péssimo, qualquer coisa que vc queira …	1	NaN	NaN
3	Não serve praticamente para nada, sem nenhuma …	1	NaN	NaN
4	Interface até intuitiva, mas peca na funcional…	1	NaN	NaN
…	…	…	…	…
10967	Meu celular foi furtado dia 2, mas em poucos m…	5	NaN	NaN
10968	aparentemente fácil de fazer, fiz da minha mãe…	5	NaN	NaN
10969	De 1 para 5 estrelas. Tinha dado 1 estrela, de…	5	NaN	NaN
10970	Fiz minha foto de perfil com um pouco de dific…	5	NaN	NaN
10971	É um pouco difícil de a validação facial. A di…	5	NaN	NaN

[10972 rows x 4 columns]

2.3 2 - Classifiers

---

[327]:
```
def classify(classifier_name, classifier, vectorizer, X_smote_treino,␣
↪y_smote_treino, y_teste, X_tfidf_teste, data_class):
    model = {}

    classifier.fit(X_smote_treino, y_smote_treino)
    predicted_class = classifier.predict(X_tfidf_teste)
    accuracy = f1_score(predicted_class, y_teste)
    print("Algorithm:", classifier_name, "F1-Score:", round(accuracy, 2))
    cr = classification_report(y_teste, predicted_class)
    model = (data_class, accuracy, classifier, vectorizer, classifier_name)
    print(cr)

    return model
```

[328]:
```
def model(data_class, col):
    model = {}

    X = classified_data['Content']
    y = classified_data[col].apply(lambda label: 1 if label.lower() ==␣
↪data_class.lower() else 0)

    X_treino, X_teste, y_treino, y_teste = train_test_split(X, y, test_size=0.
↪2, stratify=y, random_state=42)

    vectorizer = TfidfVectorizer(stop_words=stopwords.words('portuguese'),␣
↪ngram_range = (1,3))
    vectorizer.fit(X_treino)

    X_tfidf_treino = vectorizer.transform(X_treino)
    X_tfidf_teste = vectorizer.transform(X_teste)

    # define pipeline of resampling
    over = SMOTE(sampling_strategy={1: 15200}, k_neighbors=3)
    steps = [('o', over)]
    pipeline = Pipeline(steps=steps)
    X_smote_treino, y_smote_treino = pipeline.fit_resample(X_tfidf_treino,␣
↪y_treino)

    # Classifier LogisticRegression
    classifier = LogisticRegression(max_iter=6000, class_weight="balanced",␣
↪random_state=42, multi_class="ovr")
    model = classify("Logistic Regression", classifier, vectorizer,␣
↪X_smote_treino, y_smote_treino, y_teste, X_tfidf_teste, data_class)

    return model
```

[329]:
```
def warn(*args, **kwargs):
    pass
import warnings
warnings.warn = warn

models_heuristic = {}
print("Heuristics")

heuristic_classes = [
    "C0 - None heuristic",
    "C1 - Visibility of system status",
    "C2 - Match between system and the real world",
    "C3 - User control and freedom",
    "C4 - Consistency and standards",
    "C5 - Error prevention",
    "C6 - Recognition rather than recall",
    "C7 - Flexibility and efficiency of use",
    "C8 - Aesthetic and minimalist design",
    "C9 - Help users recognize, diagnose, and recover from errors",
    "C10 - Help and documentation"
]
for idx in range(len(heuristic_classes)):
    print(heuristic_classes[idx])
    models_heuristic[idx] = model(heuristic_classes[idx], "Heuristic")

models_feature = {}
print("Features")

feature_classes = [
    "F0 - None feature",
    "F1 - User access",
    "F2 - Public services",
    "F3 - Proof of life",
    "F4 - Digital signature",
    "F5 - User data"
    ]

for idx in range(len(feature_classes)):
    print(feature_classes[idx])
    models_feature[idx] = model(feature_classes[idx], "Feature")
```

```
Heuristics
C0 - None heuristic
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        28
```

```
           1       0.00      0.00      0.00         3

    accuracy                           0.90        31
   macro avg       0.45      0.50      0.47        31
weighted avg       0.82      0.90      0.86        31

C1 - Visibility of system status
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.87      1.00      0.93        27
           1       0.00      0.00      0.00         4

    accuracy                           0.87        31
   macro avg       0.44      0.50      0.47        31
weighted avg       0.76      0.87      0.81        31

C2 - Match between system and the real world
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        29
           1       0.00      0.00      0.00         2

    accuracy                           0.94        31
   macro avg       0.47      0.50      0.48        31
weighted avg       0.88      0.94      0.90        31

C3 - User control and freedom
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        29
           1       0.00      0.00      0.00         2

    accuracy                           0.94        31
   macro avg       0.47      0.50      0.48        31
weighted avg       0.88      0.94      0.90        31

C4 - Consistency and standards
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.97      1.00      0.98        30
           1       0.00      0.00      0.00         1

    accuracy                           0.97        31
   macro avg       0.48      0.50      0.49        31
```

```
weighted avg       0.94      0.97      0.95        31

C5 - Error prevention
Algorithm: Logistic Regression F1-Score: 0.5
              precision    recall  f1-score   support

           0       0.81      0.96      0.88        23
           1       0.75      0.38      0.50         8

    accuracy                           0.81        31
   macro avg       0.78      0.67      0.69        31
weighted avg       0.80      0.81      0.78        31

C6 - Recognition rather than recall
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        29
           1       0.00      0.00      0.00         2

    accuracy                           0.94        31
   macro avg       0.47      0.50      0.48        31
weighted avg       0.88      0.94      0.90        31

C7 - Flexibility and efficiency of use
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        29
           1       0.00      0.00      0.00         2

    accuracy                           0.94        31
   macro avg       0.47      0.50      0.48        31
weighted avg       0.88      0.94      0.90        31

C8 - Aesthetic and minimalist design
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.97      1.00      0.98        30
           1       0.00      0.00      0.00         1

    accuracy                           0.97        31
   macro avg       0.48      0.50      0.49        31
weighted avg       0.94      0.97      0.95        31

C9 - Help users recognize, diagnose, and recover from errors
Algorithm: Logistic Regression F1-Score: 0.0
```

```
              precision    recall  f1-score   support

           0       0.87      1.00      0.93        27
           1       0.00      0.00      0.00         4

    accuracy                           0.87        31
   macro avg       0.44      0.50      0.47        31
weighted avg       0.76      0.87      0.81        31

C10 - Help and documentation
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.97      1.00      0.98        30
           1       0.00      0.00      0.00         1

    accuracy                           0.97        31
   macro avg       0.48      0.50      0.49        31
weighted avg       0.94      0.97      0.95        31

Features
F0 - None feature
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        29
           1       0.00      0.00      0.00         2

    accuracy                           0.94        31
   macro avg       0.47      0.50      0.48        31
weighted avg       0.88      0.94      0.90        31

F1 - User access
Algorithm: Logistic Regression F1-Score: 0.83
              precision    recall  f1-score   support

           0       0.80      0.36      0.50        11
           1       0.73      0.95      0.83        20

    accuracy                           0.74        31
   macro avg       0.77      0.66      0.66        31
weighted avg       0.76      0.74      0.71        31

F2 - Public services
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        29
```

```
           1       0.00      0.00      0.00         2

    accuracy                           0.94        31
   macro avg       0.47      0.50      0.48        31
weighted avg       0.88      0.94      0.90        31

F3 - Proof of life
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        29
           1       0.00      0.00      0.00         2

    accuracy                           0.94        31
   macro avg       0.47      0.50      0.48        31
weighted avg       0.88      0.94      0.90        31

F4 - Digital signature
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.97      1.00      0.98        30
           1       0.00      0.00      0.00         1

    accuracy                           0.97        31
   macro avg       0.48      0.50      0.49        31
weighted avg       0.94      0.97      0.95        31

F5 - User data
Algorithm: Logistic Regression F1-Score: 0.0
              precision    recall  f1-score   support

           0       0.87      1.00      0.93        27
           1       0.00      0.00      0.00         4

    accuracy                           0.87        31
   macro avg       0.44      0.50      0.47        31
weighted avg       0.76      0.87      0.81        31
```

[330]:
```
predict_data = not_classified_data
results = []
for idx, row in predict_data.iterrows():
    data_feature = 'F0 - Nenhuma funcionalidade'
    data_heuristic = 'C0 - Nenhuma heurística'
    for i in range(1,len(models_feature)):
```

```
        classifier_feature = models_feature[i][2]
        vectorizer_feature = models_feature[i][3]

        classifier_heuristic = models_heuristic[i][2]
        vectorizer_heuristic = models_heuristic[i][3]

        result1 = classifier_feature.predict(vectorizer_feature.
↪transform([row['Content']]))
        result2 = classifier_heuristic.predict(vectorizer_heuristic.
↪transform([row['Content']]))

        if result1[0] == 1:
            data_feature = models_feature[i][0]


        if result2[0] == 1:
            data_heuristic = models_heuristic[i][0]

    results.append([row['Content'], row['Score'], data_feature, data_heuristic])
results_ds = pd.DataFrame(columns=predict_data.columns, data=results)
total_result = classified_data.append(results_ds)

file = 'review-by-feature-and-heuristic.csv'
if(os.path.exists(file) and os.path.isfile(file)):
    os.remove(file)
total_result.to_csv("review-by-feature-and-heuristic.csv", sep=',', index=False)
```

[331]:
```
print(total_result)
```

```
                                                 Content  Score  \
0      Não tem 2 minutos que baixei o aplicativo, vou…      1
1      Resolve um problema, aparece outro. No app ped…      1
2      O erro é exatamente o que já descrevi, e que v…      1
3      Para acessar são duas etapas, senha e código e…      1
4      Consegui fazer reconhecimento com uma luminári…      1
…                                                    …    …
10967  Meu celular foi furtado dia 2, mas em poucos m…      5
10968  aparentemente fácil de fazer, fiz da minha mãe…      5
10969  De 1 para 5 estrelas. Tinha dado 1 estrela, de…      5
10970  Fiz minha foto de perfil com um pouco de dific…      5
10971  É um pouco difícil de a validação facial. A di…      5

                 Feature                        Heuristic
0       F1 - User access            C5 - Error prevention
1      F0 - None feature            C5 - Error prevention
2       F1 - User access            C5 - Error prevention
3       F1 - User access  C1 - Visibility of system status
```

```
4       F1 - User access    C4 - Consistency and standards
…                   …                                 …
10967   F1 - User access           C0 - Nenhuma heurística
10968   F1 - User access           C0 - Nenhuma heurística
10969   F1 - User access           C0 - Nenhuma heurística
10970   F1 - User access           C0 - Nenhuma heurística
10971   F1 - User access           C0 - Nenhuma heurística

[11124 rows x 4 columns]
```