CSE454 Data Mining

Project Presentation

Ahmet Tuğkan Ayhan
Student

Context

1. Introduction
2. Data Analyzing
3. Preprocessing
4. Model Building
    a. Multinomial Naïve Bayes
    b. Multinomial Logistic Regression
    c. Support Vector Machine
5. Conclusion

Introduction

What kind of Data Modelling am I gonne use?

- A classification model (sentiment analysis)

How many inputs and outputs are there?

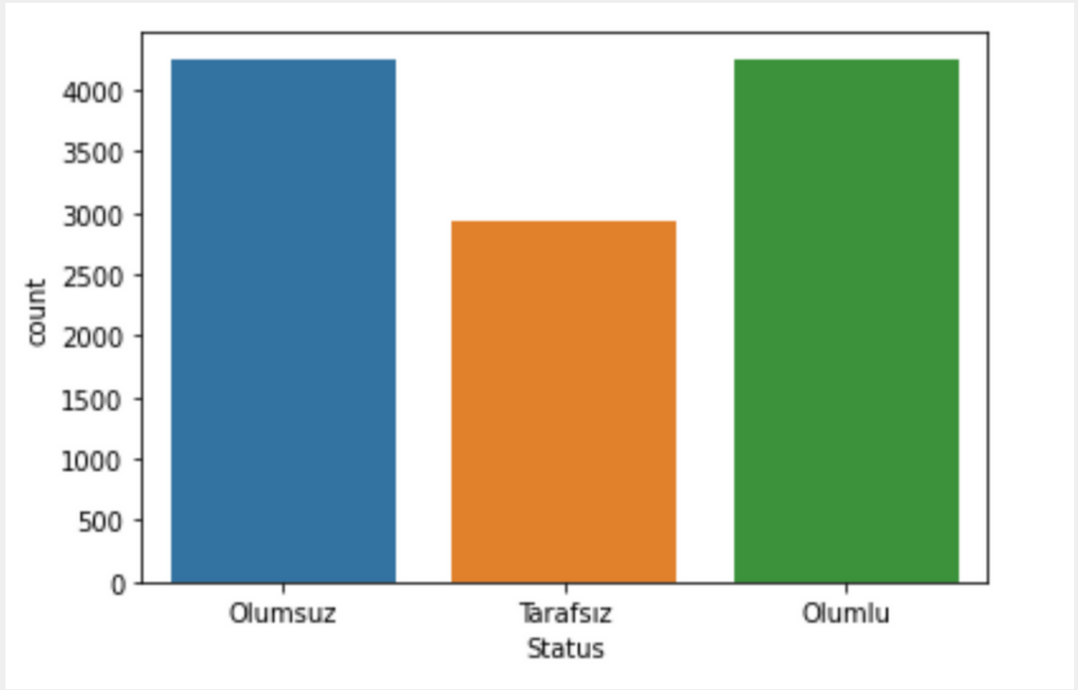
- 1 input and 3 outputs

Input	Output
ürün güzel ancak bazı problemleri var	Tarafsız
güzel bir ürün işime yaradı	Olumlu
almayın kesinlikle tavsiye etmiyorum	Olumsuz

⟩⟩

Data Analysis

Plot View



Word Cloud

Preprocessing

What are the processes?

- Removing null values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11429 entries, 0 to 11428
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Review  11426 non-null  object
 1   Status  11429 non-null  object
dtypes: object(2)
memory usage: 178.7+ KB
```

(Before removing null values)

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11426 entries, 0 to 11428
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Review  11426 non-null  object
 1   Status  11426 non-null  object
dtypes: object(2)
memory usage: 267.8+ KB
```

(After removing null values)

Preprocessing - Cont'd

What are the processes?

- Removing stopwords
- Removing punctuation marks
- Replacing Turkish characters with English characters

ürünü hiç ama hiç beğenmedim
kimse tavsiye etmiyorum.
Ürünün içi plastik
dolu.

urunu hic hic begenmedim
kimse tavsiye etmiyorum urunun
ici plastik dolu

Model Building

Multinomial Naïve Bayes - With Normal Data

- **Test Size :** %25
- **Vectorizer :** Count Vectorizer
- **Random State :** 1

```
MultinomialNB Accuracy  :  0.7018669778296382
MultinomialNB Precision :  0.7089524386586405
MultinomialNB Recall    :  0.7018669778296382
MultinomialNB F1        :  0.6509991166986239
```

Example Reviews	Predict
urun guzel ancak icinde suzgeci oldugunda kapak tam olarak oturmuyor	Olumsuz
guzel bir urun isime yaradi	Olumlu
almayin tavsiye etmiyorum	Olumsuz

Model Building - Cont'd

Multinomial Naïve Bayes - With Oversampled Minority Data (SMOTE)

- **Test Size :** %15
- **Vectorizer :** TF-ID
- **Random State :** 42
- **k-neighbors :** 2
- **Sampling Strategy :** minority

```
MultinomialNB_SMOTE Accuracy  :  0.7222870478413069
MultinomialNB_SMOTE Precision :  0.7156001171477675
MultinomialNB_SMOTE Recall    :  0.7222870478413069
MultinomialNB_SMOTE F1        :  0.7180479732535984
```

Example Reviews	Predict
urun guzel ancak icinde suzgeci oldugunda kapak tam olarak oturmuyor	Tarafsız
guzel bir urun isime yaradi	Olumlu
almayin tavsiye etmiyorum	Olumsuz

Model Building - Cont'd

Multinomial Naïve Bayes - With Undersampled Majority Data (RandomUndersampler)

- **Test Size :** %15
- **Vectorizer :** TF-ID
- **Random State :** 1
- **Sampling Strategy :** majority

```
MultinomialNB_RU Accuracy  :  0.7147024504084014
MultinomialNB_RU Precision :  0.7140080170487434
MultinomialNB_RU Recall    :  0.7147024504084014
MultinomialNB_RU F1        :  0.662260163803594
```

Example Reviews	Predict
urun guzel ancak icinde suzgeci oldugunda kapak tam olarak oturmuyor	Olumlu
guzel bir urun isime yaradi	Olumlu
almayin tavsiye etmiyorum	Olumsuz

Model Building - Cont'd

Multinomial Logistic Regression

- **Test Size :** %15
- **Vectorizer :** TF-ID
- **Random State :** 0
- **Multi Class :** multinomial
- **Max Iteration :** 11500

```
Multinomial Logistic Regression Accuracy  :  0.7158693115519253
Multinomial Logistic Regression Precision :  0.7024994468316285
Multinomial Logistic Regression Recall    :  0.7158693115519253
Multinomial Logistic Regression F1        :  0.7032094322079269
```

Example Reviews	Predict
urun guzel ancak icinde suzgeci oldugunda kapak tam olarak oturmuyor	Tarafsız
guzel bir urun isime yaradi	Olumlu
almayin tavsiye etmiyorum	Olumsuz

Model Building - Cont'd

Support Vector Machine

- **Test Size :** %25
- **Vectorizer :** TF-ID
- **Random State :** 2

```
SVC Accuracy  :  0.7292882147024504
SVC Precision :  0.7193960566233893
SVC Recall    :  0.7292882147024504
SVC F1        :  0.7174452278476063
```

Example Reviews	Predict
urun guzel ancak icinde suzgeci oldugunda kapak tam olarak oturmuyor	Tarafsız
guzel bir urun isime yaradi	Olumlu
almayin tavsiye etmiyorum	Olumsuz

Conclusion

As we seen from the model scores, it is important to preprocess dataset before we start modeling. Score table for the models is like this:

- Best accuracy score achieved with **Multinomial Logistic Regression (%72.1)**
- Best precision score achieved with **Resampled Multinomial Naïve Bayes (%71.7)**
- Best recall score achieved with **Multinomial Logistic Regression (%72.1)**
- Best F1 score achieved with **Resampled Multinomial Naïve Bayes (%71.5)**

- Worst accuracy score achieved with **Normal Multinomial Naïve Bayes (%70.1)**
- Worst precision score achieved with **Support Vector Machine (%69.9)**
- Worst recall score achieved with **Normal Multinomial Naïve Bayes (%70.1)**
- Worst F1 score achieved with **Normal Multinomial Naïve Bayes (%65.0)**

Thank You

References

https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall
https://dataaspirant.com/svm-kernels/#t-1608054630732
https://www.kaggle.com/questions-and-answers/49890