CS 445 - Natural Language Processing

Project 3

Instructor: Reyyan Yeniterzi

Engincan Varan 25050

06.01.2021

1. Introduction

In this project our goal is to develop a text classification system for Turkish news articles. The topics (labels) are as follows:

Label	# in Train Data	# in Test Data
Turkiye	1630	421
Dunya	1606	395
Spor	1583	384
Video	1582	408
Yazarlar	1599	392

As an observation, we see that the data is fairly balanced, which means we do not have any dominant label that could affect our results. There are 8000 training documents given and we want to predict the 2000 test documents. Another important thing is that in the preprocessing part there is no stemming applied. Therefore, the results are affected by it. A good improvement would be to implement a stemming algorithm for our data.

Mainly, there will be 3 approaches implemented:

1. Naive Bayes for Text Classification (multinomial classification)
2. Logistic Regression for Text Classification (multinomial classification)
3. Convolutional Neural Networks for Text Classification

Besides training models and classification methods, another aim is to identify the effects of preprocessing, and hyper-parameter tuning. For the Naive Bayes and Logistic Regression classifiers we tried different combinations of data such as untouched documents versus preprocessed documents or different types of weighting algorithms (TF vs. TF-IDF). Also, in order to the effects of stopwords we tried both inclusion and exclusion of the stopwords for Turkish. The stopwords are:

```
'ise', 'her', 'hepsi', 'için', 'veya', 'mı', 've', 'de', 'belki', 'defa', 'şey', 'tüm',
'birşey', 'niçin', 'biri', 'az', 'o', 'mu', 'niye', 'hem', 'hep', 'nereye', 'yani', 'en',
'diye', 'nerede', 'biz', 'kim', 'ya', 'acaba', 'daha', 'sanki', 'hiç', 'çünkü', 'çok', 'bu',
```

'birkaç', 'aslında', 'mü', 'eğer', 'neden', 'gibi', 'ama', 'ne', 'bazı', 'ki', 'nasıl', 'nerde', 'siz', 'ile', 'şu', 'da', 'kez'

2. Text Classification with Naive Bayes

For Naive Bayes approach, we used multinomial classification and divided our training data into validation(0.1) and training(0.9). After doing the fine-tuning we get the test results with all of the training data used to train the model.

After the test results, we fine-tuned the models with the GridSearchCV algorithm and tried different hyperparameters. In Naive Bayes Classifier case, the fine-tuned parameter is the “alpha” value. Here the alpha values are used for smoothing the probabilities. The range for the alpha values are:

```
[1.0, 0.1, 0.01, 0.001, 0.0001]
```

The classification reports are as follows:

```
Classification Report for Naive Bayesian (Test)
Vectorizer Type: TF / Include Stopwords: yes / Data Type: normal
                  precision    recall  f1-score   support

           dunya       0.80      0.85      0.82       395
            spor       0.89      0.95      0.92       384
         turkiye       0.58      0.77      0.66       421
           video       0.92      0.20      0.32       408
        yazarlar       0.72      0.96      0.83       392

        accuracy                           0.74      2000
       macro avg       0.78      0.75      0.71      2000
    weighted avg       0.78      0.74      0.71      2000
```

```
Classification Report for Naive Bayesian Fine-Tuned (Test)
Vectorizer Type: TF / Include Stopwords: yes / Data Type: normal
              precision    recall  f1-score   support

       dunya       0.81      0.85      0.83       395
        spor       0.91      0.95      0.93       384
     turkiye       0.67      0.65      0.66       421
       video       0.71      0.52      0.60       408
    yazarlar       0.78      0.96      0.86       392

    accuracy                           0.78      2000
   macro avg       0.78      0.79      0.78      2000
weighted avg       0.77      0.78      0.77      2000

Best alpha value: 0.1
```

```
Classification Report for Naive Bayesian (Test)
Vectorizer Type: TF-IDF / Include Stopwords: yes / Data Type:
normal
              precision    recall  f1-score   support

       dunya       0.80      0.76      0.78       395
        spor       0.89      0.94      0.92       384
     turkiye       0.59      0.67      0.63       421
       video       0.95      0.20      0.34       408
    yazarlar       0.58      0.98      0.73       392

    accuracy                           0.71      2000
   macro avg       0.77      0.71      0.68      2000
weighted avg       0.76      0.71      0.67      2000
```

```
Classification Report for Naive Bayesian Fine-Tuned (Test)
Vectorizer Type: TF-IDF / Include Stopwords: yes / Data Type:
normal
              precision    recall  f1-score   support

       dunya       0.82      0.84      0.83       395
        spor       0.91      0.95      0.93       384
     turkiye       0.64      0.67      0.66       421
       video       0.76      0.47      0.58       408
    yazarlar       0.76      0.96      0.85       392

    accuracy                           0.78      2000
   macro avg       0.78      0.78      0.77      2000
weighted avg       0.78      0.78      0.77      2000

Best alpha value: 0.01
```

```
Classification Report for Naive Bayesian (Test)
Vectorizer Type: TF / Include Stopwords: no / Data Type: normal
              precision    recall  f1-score   support

       dunya       0.79      0.86      0.82       395
        spor       0.89      0.96      0.92       384
     turkiye       0.58      0.77      0.66       421
       video       0.91      0.21      0.34       408
    yazarlar       0.75      0.95      0.84       392

    accuracy                           0.75      2000
   macro avg       0.78      0.75      0.72      2000
weighted avg       0.78      0.75      0.71      2000
```

```
Classification Report for Naive Bayesian Fine-Tuned (Test)
Vectorizer Type: TF / Include Stopwords: no / Data Type: normal
              precision    recall  f1-score   support

       dunya       0.81      0.84      0.83       395
        spor       0.91      0.95      0.93       384
     turkiye       0.67      0.66      0.66       421
       video       0.70      0.51      0.59       408
    yazarlar       0.79      0.95      0.86       392

    accuracy                           0.78      2000
   macro avg       0.78      0.78      0.78      2000
weighted avg       0.77      0.78      0.77      2000

Best alpha value: 0.1
```

```
Classification Report for Naive Bayesian (Test)
Vectorizer Type: TF-IDF / Include Stopwords: no / Data Type:
normal
              precision    recall  f1-score   support

       dunya       0.80      0.78      0.79       395
        spor       0.90      0.95      0.92       384
     turkiye       0.60      0.69      0.64       421
       video       0.94      0.22      0.36       408
    yazarlar       0.61      0.97      0.75       392

    accuracy                           0.72      2000
   macro avg       0.77      0.72      0.69      2000
weighted avg       0.77      0.72      0.69      2000
```

```
Classification Report for Naive Bayesian Fine-Tuned (Test)
Vectorizer Type: TF-IDF / Include Stopwords: no / Data Type:
normal
              precision    recall  f1-score   support

       dunya       0.82      0.85      0.83       395
        spor       0.91      0.95      0.93       384
     turkiye       0.64      0.67      0.66       421
       video       0.76      0.46      0.58       408
    yazarlar       0.77      0.96      0.85       392

    accuracy                           0.78      2000
   macro avg       0.78      0.78      0.77      2000
weighted avg       0.77      0.78      0.77      2000

Best alpha value: 0.01
```

```
Classification Report for Naive Bayesian (Test)
Vectorizer Type: TF / Include Stopwords: yes / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.80      0.85      0.82       395
        spor       0.89      0.96      0.92       384
     turkiye       0.58      0.78      0.67       421
       video       0.93      0.22      0.36       408
    yazarlar       0.75      0.96      0.84       392

    accuracy                           0.75      2000
   macro avg       0.79      0.75      0.72      2000
weighted avg       0.79      0.75      0.72      2000
```

```
Classification Report for Naive Bayesian Fine-Tuned (Test)
Vectorizer Type: TF / Include Stopwords: yes / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.80      0.84      0.82       395
        spor       0.91      0.95      0.93       384
     turkiye       0.67      0.65      0.66       421
       video       0.71      0.53      0.61       408
    yazarlar       0.79      0.95      0.86       392

    accuracy                           0.78      2000
   macro avg       0.78      0.79      0.78      2000
weighted avg       0.77      0.78      0.77      2000

Best alpha value: 0.1
```

```
Classification Report for Naive Bayesian (Test)
Vectorizer Type: TF-IDF / Include Stopwords: yes / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.81      0.77      0.79       395
        spor       0.90      0.94      0.92       384
     turkiye       0.59      0.68      0.64       421
       video       0.94      0.23      0.37       408
    yazarlar       0.61      0.98      0.75       392

    accuracy                           0.72      2000
   macro avg       0.77      0.72      0.69      2000
weighted avg       0.77      0.72      0.69      2000
```

```
Classification Report for Naive Bayesian Fine-Tuned (Test)
Vectorizer Type: TF-IDF / Include Stopwords: yes / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.82      0.84      0.83       395
        spor       0.91      0.95      0.93       384
     turkiye       0.64      0.67      0.66       421
       video       0.76      0.48      0.59       408
    yazarlar       0.77      0.96      0.85       392

    accuracy                           0.78      2000
   macro avg       0.78      0.78      0.77      2000
weighted avg       0.78      0.78      0.77      2000

Best alpha value: 0.01
```

```
Classification Report for Naive Bayesian (Test)
Vectorizer Type: TF / Include Stopwords: no / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.79      0.85      0.82       395
        spor       0.89      0.96      0.92       384
     turkiye       0.58      0.78      0.67       421
       video       0.93      0.23      0.36       408
    yazarlar       0.76      0.96      0.84       392

    accuracy                           0.75      2000
   macro avg       0.79      0.75      0.72      2000
weighted avg       0.79      0.75      0.72      2000
```

```
Classification Report for Naive Bayesian Fine-Tuned (Test)
Vectorizer Type: TF / Include Stopwords: no / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.80      0.84      0.82       395
        spor       0.91      0.95      0.93       384
     turkiye       0.67      0.65      0.66       421
       video       0.71      0.53      0.61       408
    yazarlar       0.79      0.95      0.86       392

    accuracy                           0.78      2000
   macro avg       0.78      0.79      0.78      2000
weighted avg       0.77      0.78      0.77      2000
```

```
Best alpha value: 0.1
```

```
Classification Report for Naive Bayesian (Test)
Vectorizer Type: TF-IDF / Include Stopwords: no / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.80      0.78      0.79       395
        spor       0.90      0.94      0.92       384
     turkiye       0.60      0.70      0.64       421
       video       0.94      0.24      0.38       408
    yazarlar       0.61      0.97      0.75       392

    accuracy                           0.72      2000
   macro avg       0.77      0.73      0.70      2000
weighted avg       0.77      0.72      0.69      2000
```

```
Classification Report for Naive Bayesian Fine-Tuned (Test)
Vectorizer Type: TF-IDF / Include Stopwords: no / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.82      0.84      0.83       395
        spor       0.91      0.95      0.93       384
     turkiye       0.64      0.68      0.66       421
       video       0.76      0.48      0.59       408
    yazarlar       0.77      0.96      0.86       392

    accuracy                           0.78      2000
   macro avg       0.78      0.78      0.77      2000
weighted avg       0.78      0.78      0.77      2000

Best alpha value: 0.01
```

All accuracy results:

```
Vectorizer Type: TF       Include Stopwords: yes      Data Type: normal                    -->      0.742
Vectorizer Type: TF       Include Stopwords: yes      Data Type: normal(fineTuned)         -->      0.781
Vectorizer Type: TF-IDF   Include Stopwords: yes      Data Type: normal                    -->      0.7055
Vectorizer Type: TF-IDF   Include Stopwords: yes      Data Type: normal(fineTuned)         -->      0.7765
Vectorizer Type: TF       Include Stopwords: no       Data Type: normal                    -->      0.7465
Vectorizer Type: TF       Include Stopwords: no       Data Type: normal(fineTuned)         -->      0.7795
Vectorizer Type: TF-IDF   Include Stopwords: no       Data Type: normal                    -->      0.717
Vectorizer Type: TF-IDF   Include Stopwords: no       Data Type: normal(fineTuned)         -->      0.7755
Vectorizer Type: TF       Include Stopwords: yes      Data Type: preprocessed              -->      0.749
Vectorizer Type: TF       Include Stopwords: yes      Data Type: preprocessed(fineTuned)   -->      0.7805
Vectorizer Type: TF-IDF   Include Stopwords: yes      Data Type: preprocessed              -->      0.717
Vectorizer Type: TF-IDF   Include Stopwords: yes      Data Type: preprocessed(fineTuned)  -->      0.776
Vectorizer Type: TF       Include Stopwords: no Data Type: preprocessed                    -->      0.75
Vectorizer Type: TF       Include Stopwords: no Data Type: preprocessed(fineTuned)         -->      0.781
Vectorizer Type: TF-IDF   Include Stopwords: no Data Type: preprocessed                    -->      0.7205
Vectorizer Type: TF-IDF   Include Stopwords: no Data Type: preprocessed(fineTuned)         -->      0.777
```

Conclusions for Naive Bayes:

- As expected Naive Bayes algorithm is not that naive and does a pretty good job for text classification. It provides a good baseline for the classification models.

- Hyperparameter tuning plays an important role for text classification with Naive Bayes. The results show that changing the alpha value slightly, affects the results around 4-6 percent.
    - For TF weighting the alpha values are 0.1
    - For TF-IDF weighting the alpha values are 0.01

- The training document size is around 8000. It is not high enough. Therefore, preprocessing or stopword removal does not change the accuracy results as drastically.

- In terms of individual F1 scores:
    - The classifiers are doing a good job with the “spor” class whereas they perform poorly with the “video” class.
    - “turkiye”, “dunya” and “yazarlar” classes vary a lot with our classifiers within the range of average-good.

- In terms of general F1 scores:
    - Both macro and micro averages are close to each other since our data is somewhat balanced.
    - However, the scores are a good baseline for text classification.

3. Text Classification with Logistic Regression

For the Logistic Regression approach, we used multinomial classification and divided our training data into validation(0.1) and training(0.9). After doing the fine-tuning we get the test results with all of the training data used to train the model.

After the test results, we fine-tuned the models with the GridSearchCV algorithm and tried different hyperparameters. In the Logistic Regression Classifier case, the fine-tuned parameter is the "C" value. The range for the C values are:

```
[0.001,0.01,0.1,1,10,100]
```

The classification reports are as follows:

```
Classification Report for Logistic Regression (Test)
Vectorizer Type: TF / Include Stopwords: yes / Data Type: normal
              precision    recall  f1-score   support

       dunya       0.83      0.84      0.84       395
        spor       0.94      0.90      0.92       384
     turkiye       0.80      0.69      0.74       421
       video       0.75      0.88      0.81       408
    yazarlar       0.91      0.92      0.91       392

    accuracy                           0.84      2000
   macro avg       0.85      0.85      0.84      2000
weighted avg       0.85      0.84      0.84      2000
```

```
Classification Report for Logistic Regression Fine-Tuned (Test)
Vectorizer Type: TF / Include Stopwords: yes / Data Type: normal
              precision    recall  f1-score   support

       dunya       0.85      0.85      0.85       395
        spor       0.95      0.91      0.93       384
     turkiye       0.83      0.70      0.76       421
       video       0.75      0.89      0.81       408
    yazarlar       0.91      0.92      0.92       392

    accuracy                           0.85      2000
   macro avg       0.86      0.85      0.85      2000
weighted avg       0.86      0.85      0.85      2000

Best C value: 0.1
```

```
Classification Report for Logistic Regression (Test)
Vectorizer Type: TF-IDF / Include Stopwords: yes / Data Type:
normal
              precision    recall  f1-score   support

       dunya       0.82      0.86      0.84       395
        spor       0.94      0.94      0.94       384
     turkiye       0.78      0.68      0.73       421
       video       0.77      0.78      0.78       408
    yazarlar       0.87      0.94      0.91       392

    accuracy                           0.84      2000
   macro avg       0.84      0.84      0.84      2000
weighted avg       0.83      0.84      0.83      2000
```

```
Classification Report for Logistic Regression Fine-Tuned (Test)
Vectorizer Type: TF-IDF / Include Stopwords: yes / Data Type:
normal
              precision    recall  f1-score   support

       dunya       0.81      0.87      0.84       395
        spor       0.94      0.95      0.94       384
     turkiye       0.78      0.71      0.75       421
       video       0.79      0.77      0.78       408
    yazarlar       0.91      0.95      0.93       392

    accuracy                           0.85      2000
   macro avg       0.85      0.85      0.85      2000
weighted avg       0.85      0.85      0.85      2000

Best C value: 100
```

```
Classification Report for Logistic Regression (Test)
Vectorizer Type: TF / Include Stopwords: no / Data Type: normal
              precision    recall  f1-score   support

       dunya       0.83      0.84      0.84       395
        spor       0.94      0.92      0.93       384
     turkiye       0.82      0.68      0.75       421
       video       0.75      0.88      0.81       408
    yazarlar       0.92      0.92      0.92       392

    accuracy                           0.85      2000
   macro avg       0.85      0.85      0.85      2000
weighted avg       0.85      0.85      0.85      2000
```

```
Classification Report for Logistic Regression Fine-Tuned (Test)
Vectorizer Type: TF / Include Stopwords: no / Data Type: normal
              precision    recall  f1-score   support

       dunya       0.83      0.84      0.84       395
        spor       0.94      0.92      0.93       384
     turkiye       0.82      0.68      0.75       421
       video       0.75      0.88      0.81       408
    yazarlar       0.92      0.92      0.92       392

    accuracy                           0.85      2000
   macro avg       0.85      0.85      0.85      2000
weighted avg       0.85      0.85      0.85      2000
```

```
Best C value: 1
```

```
Classification Report for Logistic Regression (Test)
Vectorizer Type: TF-IDF / Include Stopwords: no / Data Type:
normal
              precision    recall  f1-score   support

       dunya       0.82      0.87      0.84       395
        spor       0.93      0.93      0.93       384
     turkiye       0.77      0.69      0.73       421
       video       0.78      0.76      0.77       408
    yazarlar       0.86      0.92      0.89       392

    accuracy                           0.83      2000
   macro avg       0.83      0.84      0.83      2000
weighted avg       0.83      0.83      0.83      2000
```

```
Classification Report for Logistic Regression Fine-Tuned (Test)
Vectorizer Type: TF-IDF / Include Stopwords: no / Data Type:
normal
              precision    recall  f1-score   support

       dunya       0.83      0.87      0.85       395
        spor       0.94      0.94      0.94       384
     turkiye       0.79      0.72      0.75       421
       video       0.78      0.77      0.78       408
    yazarlar       0.91      0.95      0.93       392

    accuracy                           0.85      2000
   macro avg       0.85      0.85      0.85      2000
weighted avg       0.85      0.85      0.85      2000

Best C value: 100
```

```
Classification Report for Logistic Regression (Test)
Vectorizer Type: TF / Include Stopwords: yes / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.83      0.84      0.84       395
        spor       0.94      0.90      0.92       384
     turkiye       0.80      0.69      0.74       421
       video       0.75      0.88      0.81       408
    yazarlar       0.91      0.92      0.91       392

    accuracy                           0.84      2000
   macro avg       0.85      0.85      0.84      2000
weighted avg       0.85      0.84      0.84      2000
```

```
Classification Report for Logistic Regression Fine-Tuned (Test)
Vectorizer Type: TF / Include Stopwords: yes / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.85      0.85      0.85       395
        spor       0.95      0.91      0.93       384
     turkiye       0.83      0.70      0.76       421
       video       0.75      0.89      0.81       408
    yazarlar       0.91      0.92      0.92       392

    accuracy                           0.85      2000
   macro avg       0.86      0.85      0.85      2000
weighted avg       0.86      0.85      0.85      2000

Best C value: 0.1
```

```
Classification Report for Logistic Regression (Test)
Vectorizer Type: TF-IDF / Include Stopwords: yes / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.82      0.86      0.84       395
        spor       0.94      0.94      0.94       384
     turkiye       0.78      0.68      0.73       421
       video       0.77      0.78      0.78       408
    yazarlar       0.87      0.94      0.91       392

    accuracy                           0.84      2000
   macro avg       0.84      0.84      0.84      2000
weighted avg       0.83      0.84      0.83      2000
```

```
Classification Report for Logistic Regression Fine-Tuned (Test)
Vectorizer Type: TF-IDF / Include Stopwords: yes / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.81      0.87      0.84       395
        spor       0.94      0.95      0.94       384
     turkiye       0.78      0.71      0.75       421
       video       0.79      0.77      0.78       408
    yazarlar       0.91      0.95      0.93       392

    accuracy                           0.85      2000
   macro avg       0.85      0.85      0.85      2000
weighted avg       0.85      0.85      0.85      2000

Best C Value: 100
```

```
Classification Report for Logistic Regression (Test)
Vectorizer Type: TF / Include Stopwords: no / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.83      0.84      0.84       395
        spor       0.94      0.92      0.93       384
     turkiye       0.82      0.68      0.75       421
       video       0.75      0.88      0.81       408
    yazarlar       0.92      0.92      0.92       392

    accuracy                           0.85      2000
```

```
Classification Report for Logistic Regression Fine-Tuned (Test)
Vectorizer Type: TF / Include Stopwords: no / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.83      0.84      0.84       395
        spor       0.94      0.92      0.93       384
     turkiye       0.82      0.68      0.75       421
       video       0.75      0.88      0.81       408
    yazarlar       0.92      0.92      0.92       392

    accuracy                           0.85      2000
```

```
   macro avg       0.85      0.85      0.85      2000
weighted avg       0.85      0.85      0.85      2000
```

```
   macro avg       0.85      0.85      0.85      2000
weighted avg       0.85      0.85      0.85      2000

Best C value: 1
```

```
Classification Report for Logistic Regression (Test)
Vectorizer Type: TF-IDF / Include Stopwords: no / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.82      0.87      0.84       395
        spor       0.93      0.93      0.93       384
     turkiye       0.77      0.69      0.73       421
       video       0.78      0.76      0.77       408
    yazarlar       0.86      0.92      0.89       392

    accuracy                           0.83      2000
   macro avg       0.83      0.84      0.83      2000
weighted avg       0.83      0.83      0.83      2000
```

```
Classification Report for Logistic Regression Fine-Tuned (Test)
Vectorizer Type: TF-IDF / Include Stopwords: no / Data Type:
preprocessed
              precision    recall  f1-score   support

       dunya       0.83      0.87      0.85       395
        spor       0.94      0.94      0.94       384
     turkiye       0.79      0.72      0.75       421
       video       0.78      0.77      0.78       408
    yazarlar       0.91      0.95      0.93       392

    accuracy                           0.85      2000
   macro avg       0.85      0.85      0.85      2000
weighted avg       0.85      0.85      0.85      2000

Best C value: 100
```

All accuracy results:

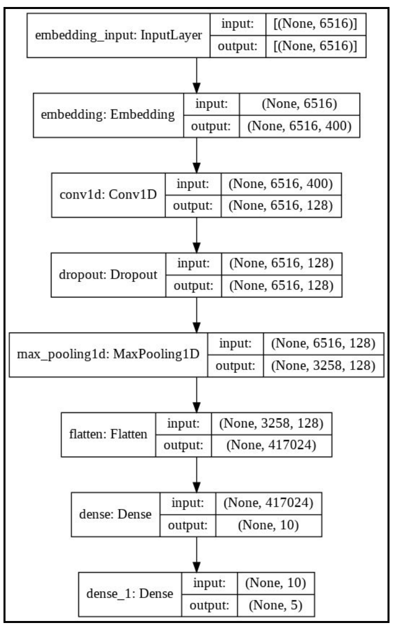
```
Vectorizer Type: TF       Include Stopwords: yes      Data Type: normal                       -->      0.8435
Vectorizer Type: TF       Include Stopwords: yes      Data Type: normal(fineTuned)            -->      0.852
Vectorizer Type: TF-IDF   Include Stopwords: yes      Data Type: normal                       -->      0.8365
Vectorizer Type: TF-IDF   Include Stopwords: yes      Data Type: normal(fineTuned)            -->      0.847
Vectorizer Type: TF       Include Stopwords: no       Data Type: normal                       -->      0.8475
Vectorizer Type: TF       Include Stopwords: no       Data Type: normal(fineTuned)            -->      0.8475
Vectorizer Type: TF-IDF   Include Stopwords: no       Data Type: normal                       -->      0.8325
Vectorizer Type: TF-IDF   Include Stopwords: no       Data Type: normal(fineTuned)            -->      0.8475
Vectorizer Type: TF       Include Stopwords: yes      Data Type: preprocessed                 -->      0.8435
Vectorizer Type: TF       Include Stopwords: yes      Data Type: preprocessed(fineTuned)     -->      0.852
Vectorizer Type: TF-IDF   Include Stopwords: yes      Data Type: preprocessed                 -->      0.8365
Vectorizer Type: TF-IDF   Include Stopwords: yes      Data Type: preprocessed(fineTuned)     -->      0.847
Vectorizer Type: TF       Include Stopwords: no       Data Type: preprocessed                 -->      0.8475
Vectorizer Type: TF       Include Stopwords: no       Data Type: preprocessed(fineTuned)     -->      0.8475
Vectorizer Type: TF-IDF   Include Stopwords: no       Data Type: preprocessed                 -->      0.8325
Vectorizer Type: TF-IDF   Include Stopwords: no       Data Type: preprocessed(fineTuned)     -->      0.8475
```

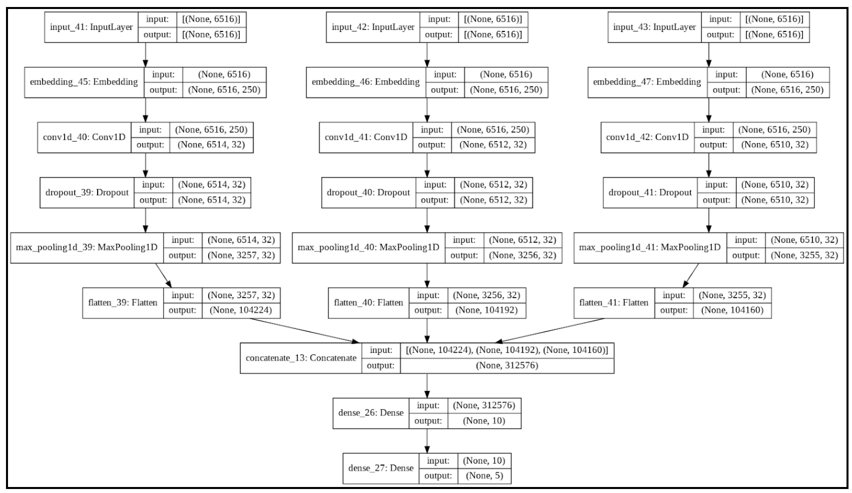
Conclusions for Logistic Regression:

- As expected Logistic Regression classifiers do a much better job than the Naive Bayes classifiers.
- However, the time it requires to give a result is much higher than Naive Bayes since there are more computations and complexity of the classifier is increasing a lot.
- Unlike Naive Bayes classifiers, hyper-parameter tuning does not increase the accuracy of our models that much. There is still 1-2% increase if we fine-tune the model.
- The training document size is around 8000. It is not high enough. Therefore, preprocessing or stopword removal does not change the accuracy results as drastically.
- In terms of F1 scores, the classifiers are performing nearly the same with different types of weighting and data.

4. Text Classification with Word Embeddings & CNN

For Convolutional Neural Network approach to a text classification problem includes utilizing the word embeddings. Therefore, in order to classify our data we need to implement a word embedding to our CNN with either pre-trained or a randomized embedding. Also, we test our CNN model with trainable embeddings and without trainable. The CNN model with trainable word embedding is called CNN-non static models whereas non-trainable model is called CNN-static models. For our CNN model, we tried different types of layering but mainly including CONV1D, GlobalMaxPooling and Dense layers with the “softmax” output layer with 5 different labels. The layers of our model are shown below. In addition, we tried a multichannel model with 3 different layering, which is also shown below, in order to compare the effects of multichanneling.





We tried different architectures for CNN models but the most efficient (in terms of seconds per epoch and accuracy) is the one that is shown above. The tried hyperparameters of the models:

- More CONV1D layers with more GlobalMaxPooling layers followed.
- Different number of filters (128, 256 etc…)
- Different sizes of filters (8, 6, 4 etc…)
- Different pooling sizes (2, 4 etc…)

As pre-trained word embeddings we used:

- https://github.com/akoksal/Turkish-Word2Vec
- Word Embeddings from project 2 for 100k documents
    - With different dimension sizes

- With different word2vec styles (CBOW vs. SKIPGRAM)

See the accuracy results for CNN (layers are shown above) models with different types of word embeddings.

Pre_Trained_WE_Type	Train_WE	WE_Dimension Size	Seconds per Epoch	Valid_Acc	Test_Acc
Random	FALSE	100	14	68.00%	66.00%
Random	TRUE	100	53	72.00%	70.90%
Random	FALSE	200	22	60.62%	61.80%
Random	TRUE	200	89	78.87%	79.75%
Random	FALSE	300	31	74.00%	73.60%
Random	TRUE	300	144	81.88%	80.19%
Random	FALSE	400	39	73.50%	74.25%
Random	TRUE	400	181	83.38%	83.20%
GitHub	FALSE	400	37	81.88%	81.09%
GitHub	TRUE	400	175	80.50%	79.00%
CBOW	FALSE	300	29	79.62%	79.19%
CBOW	TRUE	300	140	84.32%	82.67%
CBOW	FALSE	200	21	73.50%	73.19%
CBOW	TRUE	200	96	81.75%	80.94%
CBOW	FALSE	100	14	67.12%	67.77%
CBOW	TRUE	100	56	64.88%	66.00%
SKIPGRAM	FALSE	300	29	84.50%	85.50%
SKIPGRAM	TRUE	300	142	85.37%	85.15%
SKIPGRAM	FALSE	200	21	83.75%	84.45%
SKIPGRAM	TRUE	200	93	85.37%	86.50%
SKIPGRAM	FALSE	100	14	67.75%	66.90%
SKIPGRAM	TRUE	100	52	84.13%	85.79%
Multichannel - Random	FALSE	100	19	62.63%	73.20%
Multichannel - Random	TRUE	100	205	78.62%	78.64%
Multichannel - Random	FALSE	200	35	62.25%	60.00%
Multichannel - Skipgram	FALSE	200	29	84.75%	83.70%

Conclusions for CNN:

- The type of the word embedding is important in our case.
    - Most of the time random initialization of the word embeddings does not result well.
    - Skipgram performs better than Cbow. (as expected from project 2, supports this claim)
    - GitHub version is good but not suitable for our data since our data is not stemmed and preprocessed nicely.

- Preprocess is really important in pre-trained embedding, since while implementing the embedding layers we utilize the weights of the word2vec, which means that the more words are in our embedding the better the results are. Therefore, using pre-trained word embeddings requires good preprocessing. In our case word embeddings from project2 performed mostly better than the github version.

- Dimension size of the embedding changes the accuracy of results. As dimension size increases, the accuracy also increases, however train time also increases.

- Making word embedding trainable results in a good outcome for every case, yet the training time increases a lot.

- CNN models usually overfit our training data pretty quickly, therefore there is always a callback function implemented such as:

```
es_loss = EarlyStopping(monitor='val_loss', patience=4)
es_acc = EarlyStopping(monitor='val_accuracy', patience=4)
```

- CNN models usually overfit our training data pretty quickly, therefore there is always a dropout layer added to the layers.

- Multichannel CNN models (both trainabled and not) often perform better than single channel models, however the training time and parameter size increase rapidly with the multichannel models. So, in terms of efficiency, single channel models are more efficient in terms of time and accuracy.
    - We can always fine-tune multichannel models. In this case, we did not change the model much.

- Overall, CNN models perform better compared to Logistic regression or Naive bayes approach, however preprocessing and stemming is a must. Therefore, complex languages like Turkish require more effort to perform better on text classification with CNN.

Future Plans

1) For text classification tasks, Turkish is really hard to analyze. Therefore, to improve our results for our models (Naive Bayes, Logistic Regression, CNN), we need to implement an analyzer to stem and preprocess the words.
2) Multichannel models perform really well, but in this project they are not hypertuned enough to perform better on our training and test data.
3) We need to increase our training data in order to get better results for CNN. 8k sample size is not enough for CNN to perform higher accuracies, yet it performs good enough.