YILDIZ TEKNİK ÜNİVERSİTESİ
ELEKTRİK-ELEKTRONİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

DOĞAL DİL İŞLEME
DÖNEM PROJESİ

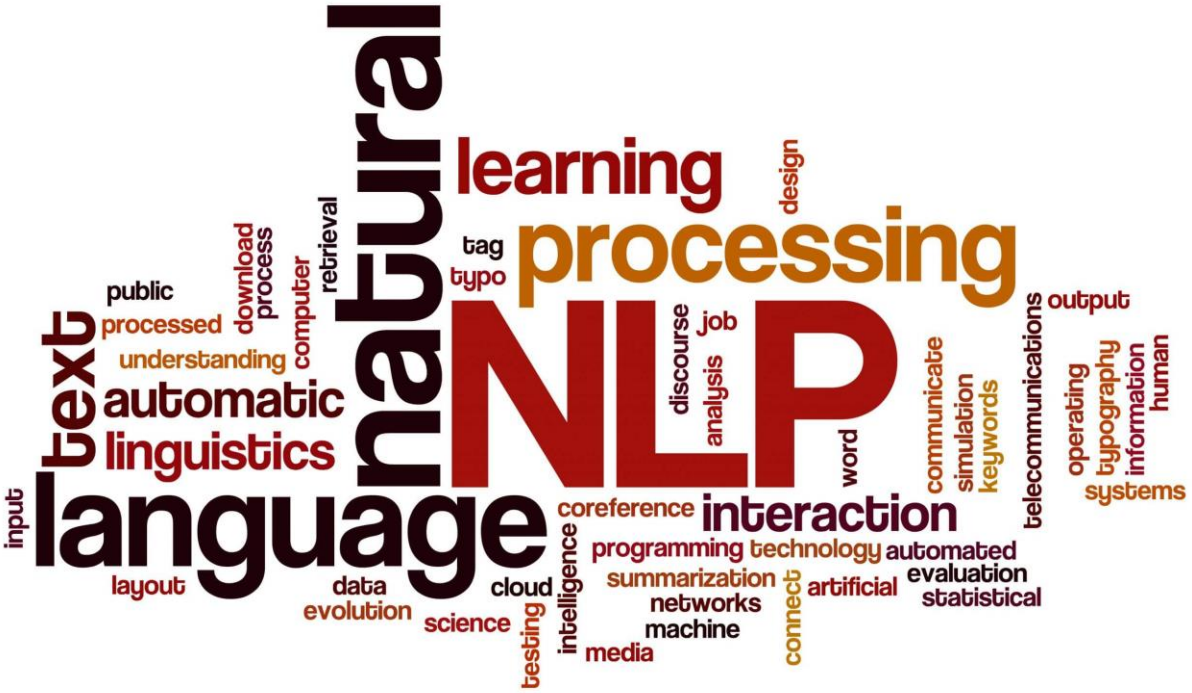
Öğr. Üyesi: Prof. Dr. Banu DİRİ

Hazırlayan: 19011065 Elif Mertoğlu

18011092 Arda Kaşıkçı

Proje Konusu

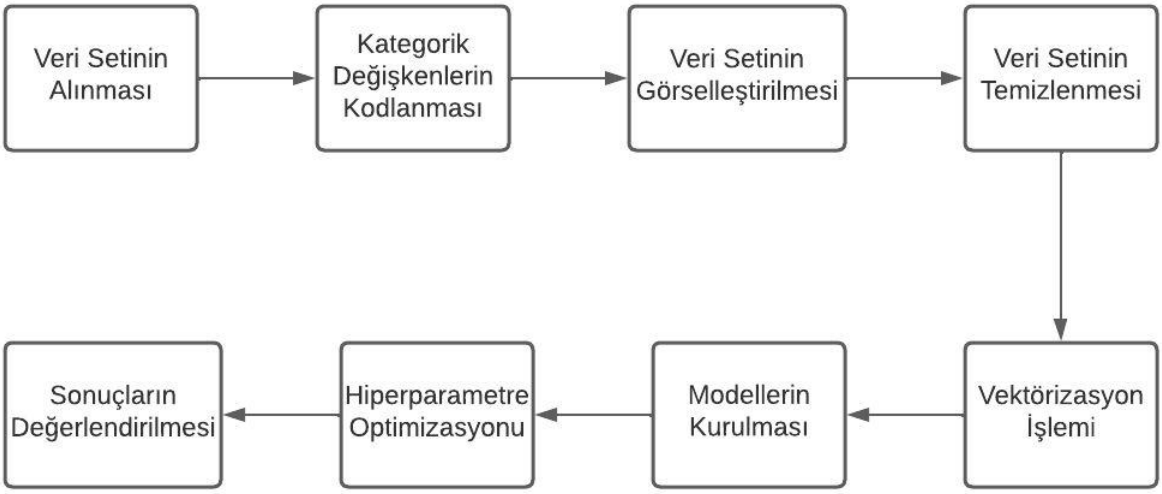
Bu projede, haber içeriklerinden haberin kategorisi (örneğin: iş, bilim, spor, vb.) tespit etme amacıyla doğal dil işleme tekniklerini kullanılmıştır. Bu tahminler yapılırken, Doğal Dil İşleme yöntemleri kullanarak haberlerin metinlerinden anahtar kelimeler çıkartılmış ve bu anahtar kelimeler ışığında haberlerin kategorilerini tespit edilmiştir. Bu çalışma ile birlikte haberlerin kategorilerini tespit etme konusunda otomatikleştirilmiş bir yöntem sunmak ve aynı zamanda haberlerin içeriklerinin anlaşılmasına yardımcı olan bir sistem gerçeklemek amaçlanmıştır.



Projenin gerçeklenmesi için öncelikle internette var olan haber veri setleri incelenmiştir. Veri seti seçilirken dengeli olmasına özen gösterilmiştir. Modelin eğitim aşamasında 6 farklı makine öğrenmesi modeli kullanılmış olup modellerin başarımları başarı ölçme yöntemleriyle hesaplanmış ve en başarılı model, haberlerin kategorisini tahmin etme için kullanılmıştır. Kelimelerin vektörize edilmesi işleminde ise literatür araştırması yapılmış ve var olan vektörize etme yöntemleri değerlendirilmiştir. Bu adımda kodun implementasyonu gerçekleştirilirken 3 farklı vektörize işlemi denenmiş ve yine en başarılı sonuç veren yöntem seçilmiştir. Haber içeriklerinde yer alan kelimelerin köklerinin bulunması adımında var olan yöntemler araştırılmış ve yöntemler uygulamalı olarak karşılaştırılmıştır. Bunun sonucunda kelimenin bulunduğu yerdeki semantik özelliklerine ve kelimenin dil bilimsel özelliklerine bakan yöntem seçilmiştir.

Yukarıda belirtilenler ‘Kullanılan Yöntemler’ bölümünde daha detaylı olarak işlenecektir.

Sistemin Blok Şeması



Veri Seti Hakkında Bilgilendirme

Kullanılan veri seti linki:
https://github.com/mhjabreel/CharCnn_Keras/tree/master/data/ag_news_csv

Haber konusu sınıflandırma veri seti, içerik bakımından kapsayıcı en büyük 4 sınıf seçilerek oluşturulmuştur. Bu sınıflar Business, Sports, Sci/Tech ve World olarak belirlenmiştir. Veri seti sırasıyla iş sektörü ve ekonomi hakkında, spor türleri hakkında, bilim hakkında ve global olaylar hakkında haberler içermektedir. Her sınıf için 30.000 örnek içerir. Toplam örnek sayısı 120.000'dir. Bu veri setinde 1. Kolon sınıf etiketi, 2. Kolon haber başlığı, 3. Kolon haber açıklamasını içermektedir.

Model oluştururken verilerin bütün olarak işlenmesini sağlamak için projemizde haber başlığı ve açıklaması kolonları birleştirilerek kullanılmıştır. Veri setindeki sınıfların dağılımı eşit olduğundan dengeli bir eğitim ve test ortamı sağlanmıştır. Veri setinin ilk kolonundan sınıf etiketleri yerine sınıfların isimleri kullanılmış ve sonrasında label_encoder fonksiyonu yardımı ile etiketler kod üzerinden sınıflandırılmıştır.

Veri seti örnek düzeni:

757	World	CARACAS, Venezuela (Reuters) - Three Venezuelan government ministers said on Monday President Hugo Chavez had easily survived a referend
758	Sci/Tech	Proposals for identity cards and a population register are opposed by Britain's information watchdog.Watchdog attacks ID card scheme
759	Sci/Tech	Internet-based holiday company Ebookers says second-quarter loses have been cut compared with the same period a year ago.Ebookers sees 'e
760	World	Reuters - Venezuelan President Hugo\Chavez has survived a referendum to recall him, according to\results released by electoral authorities on M
761	World	AFP - Malaysian emergency services rushed to the tightly-guarded US embassy in Kuala Lumpur after a powder which police said could be anthra
762	Sci/Tech	A forthcoming video game aims to educate children about the global fight against hunger.UN creates game to tackle hunger
763	World	CARACAS, Venezuela (Reuters) - Venezuelan President Hugo Chavez has survived a referendum to recall him, according to preliminary results re
764	Business	Reuters - Oil prices jumped to a new record\high near #36;47 on Monday with traders on tenterhooks for the\result of Venezuela's weekend ref
765	Sports	ATHENS -- During yesterday's celebration of the assumption of the Virgin Mary, the Greek orthodox clergy had a stern reminder for the organizer
766	Sports	Manu Ginobili's off-balance shot left his hand just a split-second before the final buzzer, dropping through the basket to give Argentina an 83-82 v
767	Sports	Sports car ace Ron Fellows nearly pulled off what arguably would have been the biggest upset in NASCAR history, finishing second after starting la
768	Sports	The US men's and women's eights pulled off huge victories in yesterday's Olympic rowing heats, each setting world bests to advance directly to S
769	Sports	ATHENS -- The evening began on a down note for the US swimming team, and descended from there. First, world champion Jenny Thompson stru
770	Sports	Tony Azevedo whizzed a last-second shot past Croatian goalkeeper Frano Vican to give the American water polo team a 7-6 victory in its tournar
771	Sports	ATHENS -- Did Japan's Kosuke Kitajima break the rules when he beat world record-holder Brendan Hansen by 17-100ths of a second in yesterday'
772	Sports	ATHENS -- When the Olympic softball schedule was released, Lisa Fernandez grabbed a marker and began counting down the days. quot;This gar
773	Sports	ATHENS -- There's only room for one Cinderella in a boxing ring. That was the sad lesson Andre Berto learned last night as his long road to the Oly
774	World	Heavy fighting erupts in Georgia's breakaway South Ossetia region, shattering a two-day ceasefire.Fighting rages in South Ossetia
775	Sci/Tech	PCS Vision Multimedia streams faster video plus audio channels to Samsung phone.Sprint Puts Streaming Media on Phones
776	Sports	ATHENS (Reuters) - Michael Phelps, one gold won but one of his eight title chances now lost, takes on Ian Thorpe and Pieter van den Hoogenba
777	World	Liechtenstein's Prince Hans-Adam hands over power to his son and invites the whole nation to a garden party.Liechtenstein royals swap power

Kullanılan Yöntemler

Projenin yazılım kısmı, 3 ana başlık altında yazılmıştır. Bunlar sırasıyla veri setinin temizlenmesi, kelimelerin vektörize edilmesi ve modellerin eğitilmesi olarak sıralanabilir.

1- Veri Setinin Temizlenmesi

Veri setinin temizlenmesi adımında ilk olarak haberlerin içeriklerinde yer alan numaralar 'number' ile değiştirilmiştir. Sonrasında metinlerde geçen noktalama işaretleri kaldırılmıştır. Çift boşluklar tokenize işlemi sırasında sorun çıkartmaması için tek boşluğa dönüştürülmüştür. $ ve takibinde number içeren söz öbekleri 'money' kelimesine çevirilmiştir. Elde edilen textlerdeki bütün kelimeler küçük harfe dönüştürülmüştür ve cümleden bağımsız kullanıldığında bir anlam ifade etmeyen stop word'ler veri setinden temizlenmiştir. Sonrasında kelimeler cümlelerdeki boşluklara göre tokenize edilmiştir ve kök bulma işlemine verilmiştir. Ayrılan kelimeler cümlede kullanımlarına göre dilbilimsel açıdan etiketlenmiş ve lemmatize işlemine tabi tutulmuştur.

```python
#Cleaning texts
def cleanText(text):
    text = re.sub("\d+", " number ", text)  # change numbers to word " number "
    text = text.translate(text.maketrans("£.,;:\-/", "$       ", "!#%&'()*+<=>?@[]^_`{|}~"))

    text = re.sub("\s\s+" , " ", text)

    text = text.replace("$ number", "money")
    text = text.replace("number bn", "money")
    text = text.replace("money bn", "money")
    text = text.replace("money money", "money")
    text = text.replace("number number", "number")

    text = to_lower(text)

    # Stopword temizligi
    stop_words = set(stopwords.words('english'))
    words = word_tokenize(text)
    text = [x for x in words if x not in stop_words]
    text = " ".join(text)

    text = lemmatizer(text)

    return text
```

2- Kelimelerin Vektörize Edilmesi

Kelimeler vektörize edilirken literatürde yer alan vektörize yöntemleri araştırılmıştır. Araştırma sonucu 3 yöntem üzerinde durulmuştur. Bunlardan ilki Count Vectorizer'dır. Bu yöntem her kelime için bir vektör oluşturur ve bu vektörün her elemanını kelimeyi içeren metinlerdeki kelime sayısı ile doldurur. Kelime sayısı büyüdükçe oluşacak vektörün boyutu da büyüyeceği için işlemlerin yavaşlamasına sebep olur. Proje kapsamında kullanılan veri seti içerisinde 120 bin haber olduğu düşünüldüğünde kelime sayısının çokluğu nedeniyle bu yöntem tercih edilmemiştir.

Performansı denenen diğer vektörize modeli Word2vec'tir. Word2vec, bir metin içinde yer alan kelimeleri vektörel bir forma dönüştürür ve bu vektörler arasındaki benzerlikler kelime anlamları arasındaki benzerliği yansıtır. Bu model ise uzun çalışması sebebiyle tercih edilmemiştir.

Son olarak tf-idf modeli seçilmiştir. Bu yöntem, bir metin içinde yer alan kelimelerin sıklığını (Term Frequency - TF) ve metinler arasında bu kelimenin nadir olup olmadığını (Inverse Document Frequency - IDF) ölçer ve bu iki değerin çarpımı olarak önemini tahmin eder. Proje kapsamında çalışma zamanı/performans olarak en iyi sonucu veren model olduğu için kelime vektörize işleminde tercih edilmiştir.

3- Modellerin Eğitilmesi

Haberlerin kategorisini belirlemek için makine öğrenmesi modelleri araştırılmış ve 6 adet model aday olarak seçilmiştir. Bu modellerin parametreleri aşağıda verilmiştir. En yüksek başarım KNN modeli ile elde edilmiştir.

```
# Log Reg
print("-------LOGISTIC REGRESSION-------\n")

lg = LogisticRegression(
    fit_intercept=True,
    class_weight=None,
    max_iter=1000,
    random_state=42)

lg.fit(x_train, y_train)
y_pred = lg.predict(x_test)
```

```
# K Neighbous Classifier
print("-------KNN-------")
knn = KNeighborsClassifier(algorithm='auto',
                           leaf_size=10,
                           metric='minkowski',
                           metric_params=None,
                           n_jobs=1,
                           n_neighbors=5,
                           p=2,
                           weights='uniform')
knn.fit(x_train,y_train)
y_pred = knn.predict(x_test)
```

```
# Random Forest
print("-------RANDOM FOREST CLASSIFIER-------")
rf = RandomForestClassifier(n_estimators=100, criterion='entropy', random_state=0, bootstrap=True )

rf.fit(x_train, y_train)
y_pred = rf.predict(x_test)
```

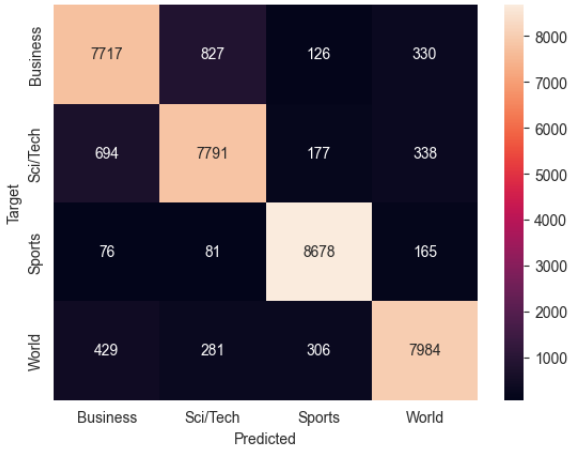
```
# SVC
print("-------SUPPORT VECTOR MACHINE-------")
svc = SVC(kernel="linear")
svc.fit(x_train, y_train)
y_pred = svc.predict(x_test)
```

```
# Decision Tree
print("-------DECISION TREE-------")
dt = DecisionTreeClassifier()
dt.fit(x_train, y_train)
y_pred = dt.predict(x_test)
```

```
# Naive Bayes
print("-------MULTINOMIAL NAIVE BAYES-------")
mnb = MultinomialNB(alpha=0.5, fit_prior=False)
mnb.fit(x_train, y_train)
y_pred = mnb.predict(x_test)

f1 = f1_score(y_test, y_pred, average='macro')
f1_list.append(Model_Acc(f1, mnb))
```

Sistemin Başarımı

1)Logistic Regression

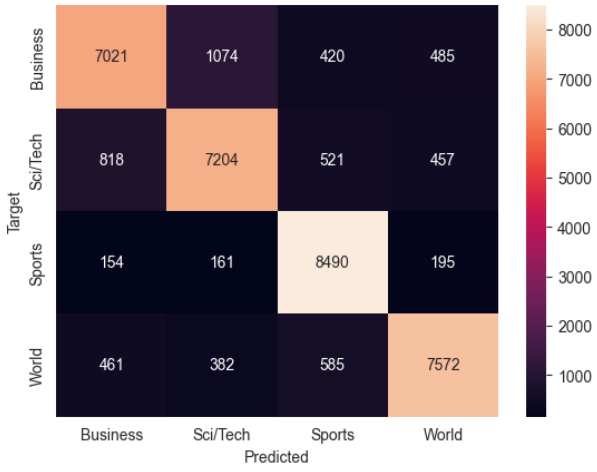



```
-------LOGISTIC REGRESSION-------

              precision    recall  f1-score   support

    Business       0.86      0.87      0.86      8916
    Sci/Tech       0.87      0.87      0.87      8980
      Sports       0.96      0.93      0.95      9287
       World       0.89      0.91      0.90      8817

    accuracy                           0.89     36000
   macro avg       0.89      0.89      0.89     36000
weighted avg       0.89      0.89      0.89     36000
```

Model: LogisticRegression

F1 Score: 0.8933516067151771

2)Random Forest Classifier




```
-------RANDOM FOREST CLASSIFIER-------

              precision    recall  f1-score   support

    Business       0.78      0.83      0.80      8454
    Sci/Tech       0.80      0.82      0.81      8821
      Sports       0.94      0.85      0.89     10016
       World       0.84      0.87      0.86      8709

    accuracy                           0.84     36000
   macro avg       0.84      0.84      0.84     36000
weighted avg       0.85      0.84      0.84     36000
```

Model: RandomForestClassifier

F1 Score: 0.8402724395973087

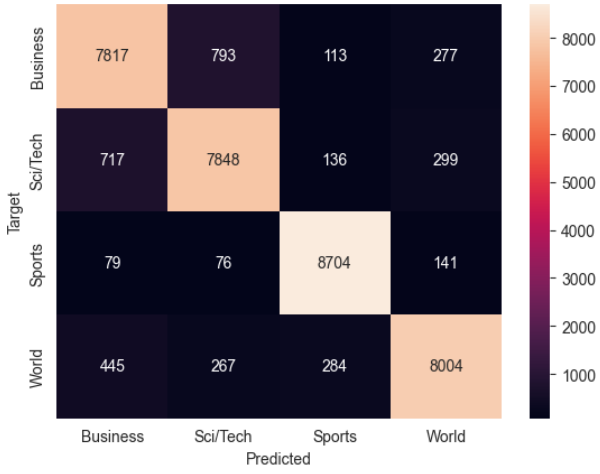
3)Multinomial Naive Bayes


```
-------MULTINOMIAL NAIVE BAYES-------
Confusion Matrix:

              precision    recall  f1-score   support

    Business       0.81      0.85      0.82      8580
    Sci/Tech       0.84      0.83      0.83      9056
      Sports       0.96      0.89      0.92      9692
       World       0.86      0.89      0.87      8672

    accuracy                           0.86     36000
   macro avg       0.86      0.86      0.86     36000
weighted avg       0.87      0.86      0.86     36000
```

Model: MultinomialNB

F1 Score: 0.8633801827579124

4)Support Vector Machine

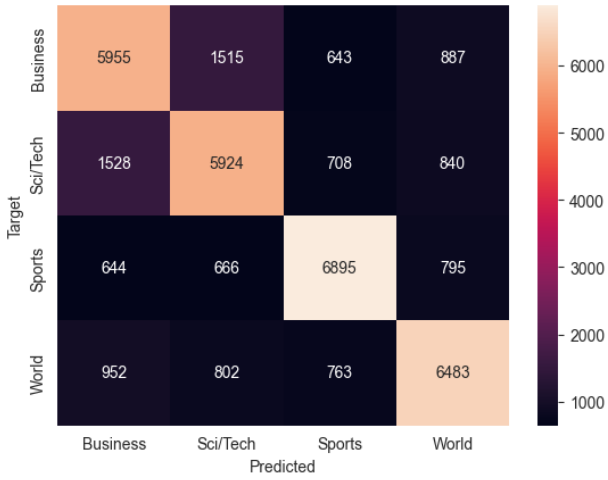



```
-------SUPPORT VECTOR MACHINE-------
Confusion Matrix:

              precision    recall  f1-score   support

    Business       0.87      0.86      0.87      9058
    Sci/Tech       0.87      0.87      0.87      8984
      Sports       0.97      0.94      0.95      9237
       World       0.89      0.92      0.90      8721

    accuracy                           0.90     36000
   macro avg       0.90      0.90      0.90     36000
weighted avg       0.90      0.90      0.90     36000
```

Model: SVM

F1 Score: 0.8991049136515377

5)Decision Tree

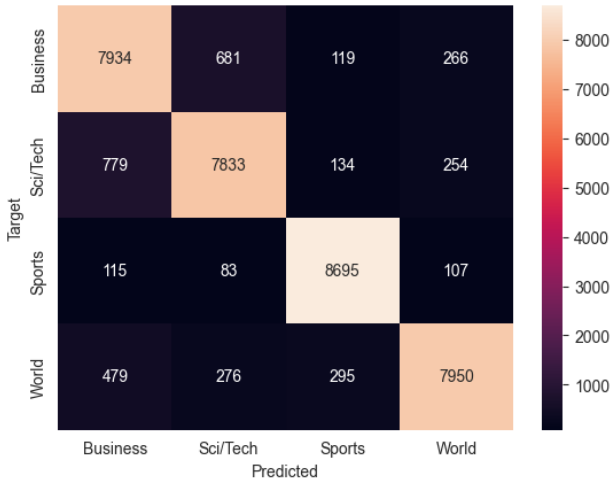



```
-------DECISION TREE-------
Confusion Matrix:

              precision    recall  f1-score   support

    Business       0.66      0.66      0.66      9079
    Sci/Tech       0.66      0.67      0.66      8907
      Sports       0.77      0.77      0.77      9009
       World       0.72      0.72      0.72      9005

    accuracy                           0.70     36000
   macro avg       0.70      0.70      0.70     36000
weighted avg       0.70      0.70      0.70     36000
```

Model: DecisionTreeClassifier

F1 Score: 0.7015694043015751

6)KNN




```
-------KNN-------
Confusion Matrix:

              precision    recall  f1-score   support

    Business       0.88      0.85      0.87      9307
    Sci/Tech       0.87      0.88      0.88      8873
      Sports       0.97      0.94      0.95      9243
       World       0.88      0.93      0.90      8577

    accuracy                           0.90     36000
   macro avg       0.90      0.90      0.90     36000
weighted avg       0.90      0.90      0.90     36000
```

Model: KNeighborsClassifier

F1 Score: 0.9002808729812936