CMPE 251 Course Project

Project Title: Zaytung - Cumhuriyet Fake News Detector

Group Name: Placeholder

Students:

1. Hasan Kemik 116207076
2. Ali Çağan Keskin 116200079

Abstract

The main goal of our project is to understand whether a news is fake or real. To make this real, we need to make sure that the algorithm we've been using can understand the Turkish ironies within the news. We've collected irony containing news from a website called Zaytung, and real, direct news from Cumhuriyet.

We've used knn, logistic regression and neural networks with different databases to understand the mechanism of machine learning with text classification.

At the end of the project, we believe that we've developed some knowledge about big data and machine learning processes.

Introduction

At the beginning, we've started to build a web scraper for Zaytung. After we've collected data from Zaytung, we've implemented some methods for clearing and taking the useful data for our project. Then, same steps have been completed for the Cumhuriyet also.

As a second step of our project, we've stored our cleaned data and flagged data to a pickle file to avoid any data loss or removing the need of collecting data every time we've initialized the program.

After the data collecting, clearing and saving processes are successfully completed, we've moved on the next step which is determining machine learning techniques and implementing them.

We're mainly focusing on Knn which is K-Nearest Neighbors Method in this project, and as a comparing factor we are using logistic regression and neural network algorithms.

Methods that were used for data acquisition and processing

Data engineering Part

(a) Data retrieval

To retrieve the data we needed, we've used python's lxml and request libraries with the xPath values. We've manually entered the Zaytung's page values within an array called page_links.

```
# requesting page as an html file, and parsing it's contents.

page_links = {"sporindex.asp","indexsnm.asp","indexblog.asp","indexkitap.asp","indexgame.asp","digerleri.asp"}

# initializing our dataframe with columns as "Header" for news head line, "Data" for first some rows,
# Flag for stating it's data property.

df = pd.DataFrame([['','','']],columns= ('Header', 'Data','Flag'))

for pages in page_links:
    df = pd.concat([df,controlPages(pages)])
```

After that we've concatenate the pandas data frame to store the data's we've collected. To collect the data, we've requested every pages' html file and parsed it with xPath values to receive this for every page:

```
In [4]: resultList
```

```
Out[4]: [' Şifremi Unuttum ',
 'Üye Ol',
 'Uzun uzun yaz',
 'BLOG',
 'Uzatmalı Sevgiliden, Şirket Çekilişinde Çıkan Patrona... Hedef Kitleye Göre Yılbaşı Hediyesi Seçme Rehberi',
 'FOTOHABER',
 "Metoroloji, sağanak yağışların AKP'li belediyelerin olduğu yerlerde ''Allah'tan gelen afet'', CHP'li belediyelerin
olduğu yerlerde ise ''altyapı eksiklikleri'' şeklinde kendini göstereceğini duyurdu...",
 "Erdoğan: ''Kendi sıçra köşklerinde siyaset yapanlar, milletin ne düşündüğünü bilemezler...''",
 'SPOR',
 "Türk Futbolu'nda ''Cumhurbaşkanlığı VAR Sistemi'' Dönemi: Erdoğan Görüntüleri Bizzat İzleyerek Son Kararı Verecek.
..",
 'devamı... ',
 'VİDEOHABER',
 "TRT'ye haddi Paris'ten bildirildi...",
 'HALKIN SESİ',
 "Yetim ve öksüzler için kışlık mont ihalesini tasarruf gerekçesiyle iptal eden Gaziantep Şahinbey Belediyesi'nin re
klam harcamalarını artırdığı ortaya çıktı...",
 'ASTROLOJİ',
 'devam... ',
```

```
In [5]: theXpath2 = '//font/text()'
resultList2 = tree.xpath(theXpath2)
resultList2
```

```
 'Fotoğrafçılar, fotoğrafçılığa özenenler, bir heves makine alıp 15. gün saydığı paraya yüreği yananlar! Toplanın he
le, anlatacaklarımız var. İster hobi, ister meslek niyetine olsun; eğer fotoğrafçılığı kafaya taktıysanız 5 adet çet
in level sizi bekliyor! ',
 'Havalar ısındı. "Vücudum su tutuyor, regl oldum" gibi bahanelerle fazlalıkları kapatamayacağınız aylar geldi çattı
. Kış rehavetinden kurtulup, hafiften tombalak olduğunuzu kabul etmenin ve rejime girmenin zamanı. Silkelenip kendin
ize gelmenin ve hayalini kurduğunuz göbeği açık, saçma sapan renklerdeki kıyafetleri rahatça giymeniz için çok değer
li rejim tavsiyelerimizi paylaşıyoruz...\xa0',
 'World Football Week, 6 - 11 Kasım 2018 tarihleri aralığında Antalya Belek'te futbol keyfinin sınırsız bir şekilde
yaşanacağı Dünya'nın her tarafından 100'den fazla Amatör takımın katılacağı bir futbol turnuvasıdır.',
 "Bu hafta 8 Mart Dünya Kadınlar Günü'nüzü kutlarken, kısmetse bir sonraki 8 Mart'ı da tek parça halinde görmeniz ko
nusunda faydalı olabilecek\xa0bazı pratik önerileri sizin için derledik. Herkese şimdiden bol şanslar...\r\n\r\n",
 'Fazla kilo illet bir şeydir. Hele ki kadınsan… Giyecek şey bulamazsın, her yediğin zehir zıkkım olur, oranı buranı
saklayacağım diye bin bir kılığa girersin. Herkesin bayıldığı yaz mevsimi senin için kâbusa döner. Havalar ısındıkça
gereken yerleri kamufle etmek giderek zorlaşır. "Üşüyorum" diye poponu örtmek için hırkayla dolaşmaya kalksan bu sef
er de kilomu saklayacağım derken isilik olursun...',
 "Bu hafta, Klasik Türk musikisiyle ilişkisi\xa0ayda yılda bir gidilen\xa0meyhanede\xa0sözlerinin yarısını uydurarak
eşlik ettiği 3-5 şarkı ve ''Çile Bülbülüm'' başladığında pusuya yatıp tam yerinde ''Allah!'' diye höykürmekten ibare
t olanlar için dev bir hizmetle karşınıdayız.",
 'Nasıl ki insanoğlu çeşit çeşitse kediler de öyle. Ortalama 30 cm kadar boyu ve türlü türlü huyu olan bu hayvanlard
an bir ya da bir kaçıyla evinizi paylaşmayı düşünüyorsanız hangi çeşidinin sizin karakterinize daha uygun olduğunu b
```

```
In [ ]:
```

For Zaytung pages “//a/text()” xPath value has been used for news headers and “//font/text()” value for news summary.

Same method has been applied to the web page Cumhuriyet with the xPath value “//span/text()”.

```
In [7]: resultList
Out[7]: ['Aboneler',
         'İletişim',
         'Apple',
         'Android',
         '6,08',
         '5,37',
         '214,17',
         'Yazarlar',
         'Gündem',
         'Siyaset',
         'Türkiye',
         'Yerel Yönetim',
         'Dünya',
         'Ekonomi',
         'Kültür',
         'Çizerler',
         'Spor',
         'Yaşam',
         'Teknoloji',
```

```
In [7]: resultList
         'Emre Kongar',
         'Çölaşan'a FETÖ'cülük iddiası?',
         '\r\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t',
         '\r\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t',
         'Mustafa Balbay',
         'Bozkurt Güvenç: Bir kütüphane kapandı!',
         '\r\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t',
         '\r\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t',
         'Orhan Bursalı',
         ''Genetiği değiştirilmiş' bebek... Mısır, domates vb'den sonra',
         '1',
         "\r\n\t\t\t\t\t\t\t\t\t\r\n\t\t\t\t\t\t\t\tFenerbahçe ve Galatasaray'ın rakipleri belli oluyor\r\n\t\t\t\t\t\t\t\t\r\n
        \t\t\t\t\t\t",
         "UEFA Avrupa Ligi'nde bugün yapılacak son 32 turu kura çekimiyle Fenerbahçe ve Galatasaray'ın rakipleri belli olacak
        .",
         '\r\n\t\t\t\t\t\t\t\t\r\n\t\t\t\t\t\t\t\tÜnlü oyuncuların menajerinden transfer itirafı\r\n\t\t\t\t\t\t\t\r\n\t\t\
        t\t\t\t\t',
         'Fenerbahçeli birçok futbolcunun menajerliğini yapan Batur Altıparmak, transferlerle ilgili flaş açıklamalar yaptı.'
        ,
         "\r\n\t\t\t\t\t\t\t\t\r\n\t\t\t\t\t\t\t\tMetin Tekin: Artık en iyi kadro Beşiktaş'tadır diyemiyoruz\r\n\t\t\t\t\t\t\
```

The most challenging part in data collection was caused by multi-paged categories in Zaytung such as sports and archive. Besides these 2 categories, all of the news was under the same html file, which can retrieve by a single request. For sports and archive parts, we need to parse the last page information from the html file of the first page. After parsing the last page information, according to website link we’ve retrieved other pages.

(b) Data preparation

For Zaytung:

As you can see, there are lots of bug words like “devamı…” or some junks caused by parsing such like “\n”, “\r” , “\t” or “\xa0”. In order to remove them, we’ve created specialized methods.

For Cumhuriyet:

To remove the bug words such as “Aboneler”, “İletişim” etc. and junks “\n”, “\r” , “\t” or “\xa0” specialized methods for Cumhuriyet has been created.

After junk parts have been cleared, the sentences for each header and summary parts of the news are saved into a pandas data frame with their flag value.

```
In [10]: df.head()
```

Out[10]:

	Header	Data
0	[yeni, başlayanlar, henüz, batmayanlar, İçin,...	[yeri, açmak, istiyorsunuz, bunun, bir, işiniz...
1	[İkna, sanatı, İnsanlardan, İstediğinizi, a...	[hafta, genel, olarak, kimseyle, sevgili, olas...
2	[yalnızlığına, bahane, arayanlar, İçin, hangi...	[onlar, çevrenizde, onlar, hafta, öğle, vakti,...
3	[İşsizleri, tanıma, koruma, hayata, bağlama, ...	[şey, filmlerdeki, kusursuz, işlemedi, hayatım...
4	[kimse, sevemez, seni, benim, kadar]	['soğuk, falan, bunlar, manipülasyon, diyerek,...

For the next step, we've imported "Natural Language Toolkit" to our program and then we took the stop words list for Turkish. Then we've added some words we've acquired from the inspection of the data we've collected and removed some pre-listed words from the list.

```
# importing stopwords turkish dictionary from nltk corpus
# and removing and adding our own selected words for cleaning process.
from nltk.corpus import stopwords
stop_words = set(stopwords.words('turkish'))
stop_words.add("bir")
stop_words.add("bu")
stop_words.add("onlar")
stop_words.add("bunlar")
stop_words.add("bir şey")
stop_words.add("damga")
stop_words.add("vurmak")
stop_words.add("kırmak")
stop_words.add("sana")
stop_words.add("sen")
stop_words.add("seninle")
stop_words.add("onunla")
stop_words.add("dolar")
stop_words.add("euro")
stop_words.add("pound")
stop_words.add("son")
stop_words.add("hafta")
stop_words.add("sabah")
stop_words.add("akşam")
stop_words.add("iyi")
stop_words.add("de")
stop_words.add("dan")
stop_words.add("gelen")
stop_words.add("gelenler")
stop_words.add("başkaban")
stop_words.add("cumhurbaşkanı")
stop_words.add("benim")
stop_words.add("senin")
stop_words.add("ben")
stop_words.remove("nasıl")
stop_words.remove("nerde")
stop_words.remove("niçin")
stop_words.remove("ya")
stop_words.remove("acaba")
stop_words.remove("birşey")
stop_words.remove("defa")
stop_words.remove("sanki")
stop_words.remove("şey")
stop_words.add("sonra")
stop_words.add("önce")
stop_words.add("gün")
stop_words.add("hafta")
stop_words.add("saat")
stop_words.add("yıl")
```

After clearing stop words and unnecessary words has been completed, we wrote a method to control our data to check there won't be any news with empty header and data in same row.

```
# checking if there's any empty news and data row, and if there's any deleting them.
for i in range(34920):
    if(df.Header[i] == [] and df.Data[i] == []):
        df = df.drop([i])
```

After all of the data process part has completed, our data frame was like this:

```
In [18]: df.head()
```

Out[18]:

	Header	Data	Flag
0	kapital 3. cilt'ten, aşiretler raporu'na... 14...	evet sevgili zaytung kitap okurları. bu hafta ...	1
1	haftanın kitapları: mahir ünsal eriş'den ''öbü...	eveeet sevgili zaytung kitap okurları. yıl son...	1
2	arafta (abd başkanı evladını kaybediyor), kadı...	merhabalar pek muhterem, entelektüel, kitap ku...	1
3	haftanın kitapları: başlangıç (da vinci yok am...	önce anekta, devamında yine de âmin, en son da...	1
4	röportaj: sinem sal - ''türkiye'de süper kahra...	çoğunlukla aşırı kültür kokmadığından ortamlar...	1

```
In [19]: df.tail()
```

Out[19]:

	Header	Data	Flag
108	baba ve 12 yaşındaki oğlunu öldürdü: pişmanım!	kayseri'nin sarıoğlan ilçesinde baba ve 12 yaş...	0
109	aydın'da hastanede yangın	aydın kadın doğum ve çocuk hastalıkları hastan...	0
110	İsveç'te bomba alarmı!	göteborg kent merkezinde, içinde patlayıcı bul...	0
111	'kılıçdaroğlu yapamaz' deyince araya girdi, aç...	İstanbul tv adlı youtube kanalının bir sokak ...	0
112	müslüm filmi 6 milyon izleyiciyi aştı	müslüm gürses'in gerçek yaşam hikayesini sinem...	0

We've exported this data frame as separate files name x.pickle and y.pickle which x contains header and data columns, while y contains only flag column.

Data science Part

(a) Data exploration

First of all, we've imported our x and y pickle files to the program which we're going to observe results of machine learning algorithms.

```
In [29]: x.head()
```

Out[29]:

	Header	Data
0	[yeni, başlayanlar, henüz, batmayanlar, İçin,...	[yeri, açmak, istiyorsunuz, bunun, bir, işiniz...
1	[İkna, sanatı, insanlardan, istediğinizi, a...	[hafta, genel, olarak, kimseyle, sevgili, olas...
2	[yalnızlığına, bahane, arayanlar, İçin, hangi...	[onlar, çevrenizde, onlar, hafta, öğle, vakti,...
3	[İşsizleri, tanıma, koruma, hayata, bağlama, ...	[şey, filmlerdeki, kusursuz, işlemedi, hayatım...
4	[kimse, sevemez, seni, benim, kadar]	['soğuk, falan, bunlar, manipülasyon, diyerek,...

```
In [30]: y.head()
```

```
Out[30]: 0    1
         1    1
         2    1
         3    1
         4    1
         Name: Flag, dtype: object
```

In our data frame object,

"Header" represents the header of the news,

"Data" represents the summary of the news provided by the website,

"Flag" which is the data frame named y, represents the trueness of the news, such as 1: fake, 0: true.

After this step, we've concatenated our header and data columns into alldata column to continue with the tf-idf model.

```
In [23]: df
```

```
Out[23]: ['yeni başlayanlar henüz batmayanlar İçin esnaflığın altın kuralı yeri açmak istiyorsunuz bunun bir işiniz yeriniz si
         z patronunuzdan beklerken o sizi işten çıkarmayı düşünüyor aylardır etkilemeye çalıştığınız kişi hala isminizi bilmiy
         or mu herkesin paylaşımları like üstüne like alırken sizinkini sadece ercan ( ) beğeniyor ? o halde yöntemler size gö
         re İkna psikolojisi yöntemlerini kullanarak insanlardan istediğinizi alabileceksiniz',
          'İkna sanatı insanlardan istediğinizi almak İçin etkili İletişim yöntemleri hafta genel olarak kimseyle sevgili olas
         ı gelmeyenler kışta kıyamette uğraşacak diyeüşenenleriçin zodyaktan destek alıyoruz',
          "yalnızlığına bahane arayanlar İçin hangi burçla sevgili olunmaz olunmamalıdır onlar çevrenizde onlar hafta öğle vak
         ti avm'de onlar başvurusu insan kaynaklarıdenen gladyatör arenasında velhasıl kelam onlar da içimizde aramızda İşsizl
         er",
          'İşsizleri tanıma koruma hayata bağlama rehberi şey filmlerdeki kusursuz işlemedi hayatımızda hiçbir zaman göremeyec
         eğiz kaçıp giden fırsatları hayran edici çekim tekniğiyle bizim yönetmen biraz muhafazakar sanırım',
          "kimse sevemez seni benim kadar 'soğuk falan bunlar manipülasyon diyerek düşünce gücüyle ısınabileceğinize da doğalg
         az faturasını zabıtaya şikayet ederekişin içinden sıyrılabileceğinize güvenmiyorsanız acilen doğalgaz yerine yakacak
         başka şey bulmanızgerekiyor zira evet bunlar iyi günleriniz",
          'winter coming doğalgaza alternatif olarak evde bulunabilecek düşük maliyetli yakacaklar evlenilecek mevsimdir esas
         sonbaharda eğlenilir',
          'karşılaştırma sonbahar yaz tabii sonbahar manyak mısın ? ) netflix chill edeyim derken sadece dakikayı ne izlesem "
         harcayanlar belgeselinden dizisine , tıkla izle hizmeti niteliğinde liste…',
          'yapacak İyi şeyiniz yoksa bu yazı bir şekilde atlatmanızı sağlayacak netflix dizileri malum geldi beraberinde aşkla
```

(b) Data modeling

First we tokenized the sentences with nltk word tokenizer. The process allows us to inspect the words as separate objects. After that we have cleared stop words.

```
In [6]: #X_train, X_test, y_train, y_test = train_test_split(df, y, random_state=50, test_size = 0.25)
        X_train, X_test, y_train, y_test = train_test_split(df, y, random_state=50, test_size = 0.25, stratify=y)
        # creating train and test splits according to a random state of 50, with 1/4 test size.
```

df: Every data that we use in project as a combined sentence. (both Zaytung and Cumhuriyet news)

y: Flag values of the news

random_state: To obtain the exactly same randomization in every run.

test_size: The percentage of the data that we use for prediction.

stratify: makes a split so that the proportion of values in the sample produced will be the same as the proportion of values provided to parameter.

Then we use Tf-Idf (frequency-inverse document frequency) to transform our text data to numerical data(double) by calculating the word frequency. And assign these double values to a two-dimensional double array.

After the calculation has completed, machine learning algorithms are applied to the data.

Results that were obtained

Since that all the data obtaining part is finished, we start to use the ML algorithms with different methods. After that, we will choose the most efficent method for the programme.

KNN METHOD (Solution 1)

- 3300 Zaytung, 3300 Cumhuriyet news data.
- 10 Neighbours for KNN.
- Without stratifying.
- Without random_state.
- As we see the accuracy is approximately %52.

```
In [13]: knn = neighbors.KNeighborsClassifier(n_neighbors=10)
         # Learn best parameters
         knn.fit(X_train, y_train)

Out[13]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                    metric_params=None, n_jobs=1, n_neighbors=10, p=2,
                    weights='uniform')

In [14]: y_pred = knn.predict(X_test)
         # Evaluation

In [15]: type(y_pred)

Out[15]: numpy.ndarray

In [17]: accuracy_score(y_test, y_pred)

Out[17]: 0.5145454545454545
```

KNN METHOD (Solution 2)

- 3300 Zaytung, 3300 Cumhuriyet news data.
- 10 Neighbours for KNN.
- With stratifying.
- With random_state.
- As we see the accuracy is approximately %50.

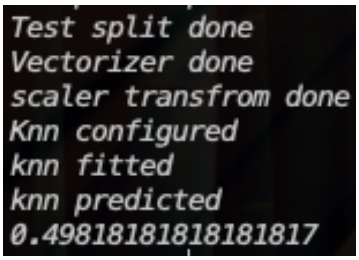
```
In [15]: y_pred = knn.predict(X_test) # applying the learned function into the test part.

In [16]: accuracy_score(y_test, y_pred) # comparing the predicted output with the real values.

Out[16]: 0.4987878787878788
```

KNN METHOD (Solution 3)

- 3300 Zaytung, 3300 Cumhuriyet news data.
- 10 Neighbours for KNN.
- With stratifying.
- With random_state.
- As we see the accuracy is approximately %50.



KNN METHOD (Solution 4)

- 3300 Zaytung, 3300 Cumhuriyet news data.
- 5 Neighbours for KNN.
- Without stratifying.
- Without random_state.
- As we see the accuracy is approximately %51.

```
In [49]: accuracy_score(y_test, y_pred)
Out[49]: 0.5051515151515151
```

flat_list = [item for sublist in data for item in sublist]

KNN METHOD (Solution 5)

- 3300 Zaytung, 11700 Cumhuriyet news data.
- 10 Neighbours for KNN.
- Without stratifying.
- Without random_state.
- As we see the accuracy is approximately %78.

LOGISTIC REGGRESSION METHOD (Solution 6)

- 3300 Zaytung, 3300 Cumhuriyet news data.
- Without stratifying.
- Without random_state.
- As we see the accuracy is approximately %94.

```
In [17]: from sklearn.linear_model import LogisticRegression
         lr = LogisticRegression()
         lr.fit(X_train, y_train)

Out[17]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
                   penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
                   verbose=0, warm_start=False)

In [18]: predictions = lr.predict(X_test)
         from sklearn.metrics import accuracy_score
         accuracy = accuracy_score(y_test, predictions)

In [19]: accuracy

Out[19]: 0.9424242424242424
```

NEURAL NETWORK METHOD (Solution 7)

- 3300 Zaytung, 3300 Cumhuriyet news data.
- With stratifying.
- With random_state.
- As we see the accuracy is approximately %90.

```
In [20]: from sklearn.neural_network import MLPClassifier
         clf = MLPClassifier(solver='lbfgs', alpha=1e-1, hidden_layer_sizes=(10, 10), random_state=1)
         clf.fit(X_train, y_train)

         y_pred = clf.predict(X_test)
         print(accuracy_score(y_test, y_pred))

          0.9018181818181819
```

NEURAL NETWORK METHOD (Solution 8)

- 3300 Zaytung, 3300 Cumhuriyet news data.
- Without stratifying.
- Without random_state.
- As we see the accuracy is approximately %93.

```
In [21]: from sklearn.neural_network import MLPClassifier
         clf = MLPClassifier(solver='lbfgs', alpha=1e-1, hidden_layer_sizes=(10, 10), random_state=1)
         clf.fit(X_train, y_train)

         y_pred = clf.predict(X_test)
         print(accuracy_score(y_test, y_pred))

          0.9284848484848485
```

NEURAL NETWORK METHOD (Solution 9)

- 3300 Zaytung, 11700 Cumhuriyet news data.
- With stratifying.
- With random_state.
- As we see the accuracy is approximately %94.

```
In [21]: from sklearn.neural_network import MLPClassifier
         clf = MLPClassifier(solver='lbfgs', alpha=1e-1, hidden_layer_sizes=(10, 10), random_state=1)
         clf.fit(X_train, y_train)

         y_pred = clf.predict(X_test)
         print(accuracy_score(y_test, y_pred))

          0.9424
```

Conclusion

When we look the results we get, there may be huge differences when we change the ML method. Using stratify and random_state is changing the accuracy between %1-5. Since that the program is based on Natural Language Processing in Turkish and we only have two different flag values (1: fake news, 0: real news) the Neural Network and Logistic Regression methods are more accurate than KNN(K-Nearest Neighbors) method.

KNN takes much longer at evaluation time, especially if we have many data points, and our process verifies the propositions above. The LR and NN is much more faster than KNN.

As a conclusion, according to the data set we've used for the machine learning method and the accuracy we've obtained, we chose the Neural Network machine learning algorithm for the best algorithm to obtain text based results for Turkish.

Project Code-*Describe your machine learning methods*

KNN

```
In [12]: from sklearn import neighbors
         from sklearn.neighbors import KNeighborsClassifier
         knn = neighbors.KNeighborsClassifier(n_neighbors=5) # initializing knn feature with 5 nearest neighbors.
          Learn best parameters
         knn.fit(X_train, y_train)
         y_pred = knn.predict(X_test) # applying the learned function into the test part.
         accuracy_score(y_test, y_pred) # comparing the predicted output with the real values.
         print(accuracy_score)
```

Knn works with analogy. Classification of a data is calculated with it's K nearest neighbors. Using that grouped classification method, KNN learns the data provided, and then predicts on the test data. KNN uses the most time and memory when we compared our machine learning algorithms. We've acquired the worst results with KNN. So we can say that, it isn't good with big data sets, nor text data.

LOGISTIC REGRESSION

```
In [13]: from sklearn.linear_model import LogisticRegression
         lr = LogisticRegression()
         lr.fit(X_train, y_train)
         predictions = lr.predict(X_test)
         from sklearn.metrics import accuracy_score
         accuracy = accuracy_score(y_test, predictions)
         print(accuracy)
```

Logistic Regression is depending on the coordinate system. Algorithm defines a best fit line, parabola, etc. in order to separate data points, in our case fake or not. After a mathematical expression has found, it predicts the test values. We acquire very good accuracy scores with logistic regression, because our text data probably placed where it's closest to x or y axis. That's why we don't think that logistic regression can be a good way to predict text data.

NEURAL NETWORK

```
In [ ]: from sklearn.neural_network import MLPClassifier
clf = MLPClassifier(solver='lbfgs', alpha=1e-1, hidden_layer_sizes=(10, 10), random_state=1)
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
print(accuracy_score(y_test, y_pred))
```

Neural Network method is relying on the connections, links, between the data points, which we think that it might be understand the ironies between the sentences if enough data is given with the linking operation it makes between the words, sentences and flags. So we thought that there might be no need for a tool like Zemberek to process data for detecting ironies.

The results that we obtained is confirming our thoughts were true. So Neural Network is the best machine learning method for our project.

Project Data - *Describe your data*

Data Collection

The data we have collected is News that some writers wrote.

We take fake news from Zaytung. Zaytung is a satirical website which consists of ironic made-up news, had some memorable headlines with more than a ring of truth.

Zaytung website: http://www.zaytung.com/indexblog.asp

And we take the true news from a newspaper called Cumhuriyet. Cumhuriyet is the oldest up-market Turkish daily newspaper. the newspaper has subscribed to a staunchly secular, republican course. In the past closely affiliated with the Kemalist Republican People's Party (CHP), the center-left newspaper turned to a more independent course over time, advocating democracy, social liberal values and free markets.

Cumhuriyet Website: http://www.cumhuriyet.com.tr/

Since that this is a fake news detector program:

- 1 is for Fake News
- 0 is for True News

Discussion

What were the unexpected difficulties in your project?

Writing webscraper for Zaytung and Cumhuriyet was a little bit difficult without Beatiful Soup.

And implementing the KNN algorithms was unexpectedly difficult. After we figured it out implementing KNN, the Logistic Reggression and Neural Network was very easy for us.

How can you improve your results?

We can use all the data(37474). But it would not increase the efficiency of our project. Because the difference of Zaytung data and the Cumhuriyet data will be so big that the ML algorithms might become useless.

Importing the Zemberek Library to detect irony in Turkish text may increase the accuracy of the project.

What can be done for future work?

We can import Zemberek Library to the project.

We can choose the stop_words with more detail.

We can collect data from websites like:

- https://teyit.org/
- https://sputniknews.com/search/?query=fake+news
- https://www.dogrulukpayi.com/

and implement the steps used for Turkish Fake News detector project.