Wrong News Detection with NLP

Edanur Anlayan
030716004

INTRODUCTION

Due to the situation of our world; The use of social networks and the tendency towards technology has increased enormously. With the introduction of technology into our lives, the use of newspapers has decreased significantly and gradually began to be distanced from televisions. The pandemic process triggered this too much and internet application using for all information to be reached increased. However, all kinds of news that started to spread on the internet started to lose their reliability. The spread speed of a news on social media is quite high. As fake news spread rapidly, it is thought to be true and people are misled.

In this project, this issue has been addressed and various studies have been carried out to understand the truth and wrongness of the news spread on social media. News from reliable sources and false news from false sources have been assigned as a source for other news to come. Based on these data, the truth or falsity of the news has been tried to be proven.

LITERATURE REVIEW

Along with artificial intelligence, which is one of the benefits of advancing technology, many applications have been moved to computers and phones to make our lives easier. The concept of natural language processing (NLP) is also a concept that entered our lives with artificial intelligence. It is a technology developed for the computer to perceive the language we speak daily and to perform operations on it. Today, there are many projects developed with NLP. If we need to give examples for these projects, we can mention the following applications:

- Text Classification and Categorization
- Named Entity Recognition (NER)
- Language Generation and Multi-document Summarization
- Character Recognition
- Semantic Parsing and Question Answering

STRUCTURE OF THE SOLUTION PROPOSED

First, I applied the process of pulling my news, which I will categorize as true and false, from Twitter. I took 1000 pieces of data from the 'trthaber' twitter account, which I determined as reliable, and 1000 from the 'zaytung' twitter account that publishes false news. So, i took a total 2000 data. I made it lowercase all of letters. Then, using the TF-IDF method, I found the frequencies of the words in the sentence by calculating the weights and made it ready for processing on them. **TF-IDF method**; It is a method used to calculate term weights in a document and provides a value according to the frequency of use. Later, I gave the vectors resulting from the TF-IDF to the logistic regression algorithm. **Logistic regression algorithm**; An algorithm that learns a model using the given data and then generates predictions using the test set given with this model. In this project, it was used to predict the accuracy of the news. Along with the comparison of the predicted values given by the algorithm and the values we expect; I got confusion matrix, precision score, recall score, f score and accuracy values.

HOW TO USE THE SOFTWARE ? WHAT ARE THE REQUIREMENTS TO RUN ?

Since I wrote the program in python language, I wrote my codes on the anaconda navigation program where found jupyter notebook. I have installed several libraries necessary to run the project.
Snscrape library : The library required to extract data from Twitter.

```
: !pip install snscrape
```

Sklearn library: A library that is required to use machine learning algorithms and preprocessing. The cross_val_score function separates the data set, while crossing is also used.

```
: !pip install sklearn
```

Later, libraries and functions were imported for the program to work.

```
: import snscrape.modules.twitter as sns
  import pandas as pd
  import numpy as np
  from sklearn.model_selection import train_test_split,cross_val_score
  from sklearn import preprocessing
  from sklearn.feature_extraction.text import TfidfVectorizer
  from sklearn.linear_model import LogisticRegression
  from sklearn.metrics import precision_score, recall_score, fbeta_score, confusion_matrix
```

In order for the program to run, the python compiler and the library and functions I mentioned above must be installed.

RUNTIME EXAMPLES

```
: #belirlenen twitter haber sayfasından haber çekme işleminin gerçekleştirildiği fonksiyon :
```

```
: def getNews(username,c_of_news):
      array_news = []
      for num_of_tw,tweet in enumerate(sns.TwitterSearchScraper("from:@"+username).get_items()):
          if num_of_tw>c_of_news:
              break
          regex = tweet.content.find("http")
          array_news.append(tweet.content[:regex])
      array_news = pd.DataFrame(array_news,columns=["news"])
      return array_news
```

```
#verilere preprocessing işlemi uygulanarak; metin ve sayı dışındaki karakterleri silip,harflerin tamamını küçülten fonksiyon.
```

```
def prep(array_news):
    array_news["news"] = array_news["news"].apply(lambda x:" ".join(x.lower() for x in x.split()))
    array_news["news"] = array_news["news"].str.replace("[^\w\s]","")
    array_news["news"] = array_news["news"].str.replace("[\d]","")
    return array_news
```

```
#çekilen doğru ve yanlış haberler için, alt alta sıralanacak ve doğru haberler için 1 değeri, yanlış haberler için 0 değeri
#yazacak fonksiyon.
```

```
def join_news(true_news,wrong_news):
    true_news["value"] = 1
    wrong_news["value"] = 0
    news = pd.concat([true_news,wrong_news],axis=0,ignore_index=True)
    return news
```

```
#verisetini logistic regression algoritmasına vererek öğreten ve gerçek ve tahmin değerlerini kullanarak istenilen değerleri
#hesaplayan fonksiyon.
```

```
def logistic_regression(traininput_ofdata,trainoutput_ofdata,testinput_ofdata,testoutput_ofdata):
    log_reg = LogisticRegression()
    log_reg_mdl = log_reg.fit(traininput_ofdata,trainoutput_ofdata)
    testdata_predict=log_reg_mdl.predict(testinput_ofdata)
    con_m=confusion_matrix(testoutput_ofdata, testdata_predict)
    pre_score=precision_score(testoutput_ofdata, testdata_predict)
    rec_score=recall_score(testoutput_ofdata, testdata_predict)
    f_score=fbeta_score(testoutput_ofdata, testdata_predict, beta=1)
    accuracy = cross_val_score(log_reg_mdl,testinput_ofdata,testoutput_ofdata,cv=10).mean()
    return accuracy,con_m,pre_score,rec_score,f_score,log_reg_mdl
```

```
#tf-idf yöntemiyle veriler sayısal değerlere çevrilir ve bundan bir vektör oluşturuluyor.
```

```
def tf_idf(traininput_ofdata,testinput_ofdata):
    tf_idf_vectorizer = TfidfVectorizer()
    tf_idf_vectorizer.fit(traininput_ofdata)
    tf_idf_traininput = tf_idf_vectorizer.transform(traininput_ofdata)
    tf_idf_testinput = tf_idf_vectorizer.transform(testinput_ofdata)
    return tf_idf_traininput,tf_idf_testinput
```

```
#verinin çekilip, kendini tekrarlamaması için kaydedilip onun kullanılması için gerekli kodlar.
```

```
#array_news=getNews("trthaber",1000)
```

```
#array_news_wrong=getNews("zaytung",1000)
```

```
#array_news=prep(array_news)
#array_news_wrong=prep(array_news_wrong)
```

```
#news=join_news(array_news,array_news_wrong)
```

```
#news.to_csv("haberler.csv",encoding = "utf-16",index=False)
```

```
news=pd.read_csv("haberler.csv",encoding="utf-16")
```

```
news.head()
```

	news	value
0	dünya genelinde koronavirüs tespit edilen kişi...	1
1	ton demiri at arabalarına yükleyerek çaldılar...	1
2	türkçenin derinliklerine dalınca gözlerime on ...	1
3	ak partide yılın ilk myk toplantısında gündem ...	1
4	gözler yılının aralık ayı enflasyon oranında ...	1

```
news.tail()
```

	news	value
1997	yargıda önemli reform tutukluluk kararı çıkmas...	0
1998	fotohaber türkiye merakla o soruşturmanın son...	0
1999	videohaber seydioğlu baklavaları tam diğer ba...	0
2000	görevden alınan merkez bankası başkanı murat ç...	0
2001	canon d için en iyi monteyi yapacak photoshop ...	0

```
]: #veriler train ve test olarak ayrılır. Train:%70,Test=%30

]: traininput_ofdata,testinput_ofdata,trainoutput_ofdata,testoutput_ofdata=train_test_split(news["news"],news["value"],test_size=0.3

]: #sayısal değerleri katagorik değerlere çeviren kod.

]: encoder = preprocessing.LabelEncoder()
   trainoutput_ofdata_2 = encoder.fit_transform(trainoutput_ofdata)
   testoutput_ofdata_2 = encoder.fit_transform(testoutput_ofdata)

]: #tf-idf fonksiyonu ile vektör oluşturuldu.

]: tf_idf_traininput,tf_idf_testinput = tf_idf(traininput_ofdata,testinput_ofdata)

]: tf_idf_traininput.toarray()

]: array([[0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.],
          ...,
          [0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.]])

]: tf_idf_testinput.toarray()

]: array([[0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.],
          ...,
          [0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.]])

]: #logistic regression fonksiyonu ile istenilen değerler elde edildi.

]: pre_score,rec_score,f_score,log_reg_mdl=logistic_regression(tf_idf_traininput,trainoutput_ofdata_2,tf_idf_testinput,testoutput_of

]: accuracy

]: 0.883551912568306

]: con_m

]: array([[291,  23],
          [ 15, 272]], dtype=int64)

]: pre_score

]: 0.9220338983050848

]: rec_score

]: 0.9477351916376306

]: f_score

]: 0.9347079037800688
```

```
#farklı haberler çekilerek makineye tahmin etmesi için verilir.
```

```
new_news=getNews("Haberturk",10)
```

```
new_news
```

	news
0	Kadın kılığına girip 5 milyonluk cipi çaldı!
1	Yargıtay, kocasına hakaret eden kadın ile doğu...
2	AK Parti'de Merkez Yürütme Kurulu (MYK) "refor...
3	#SONDAKİKA | Gabar'da öldürülen teröristlerden...
4	Pınar'ın katiliyle ilgili flaş tespit!
5	Fransa'da çatışma: 16 yaşında bir genç hayatın...
6	THY kargo uçağı zorunlu iniş yaptı
7	Gönüllere dokundu! Sokakta ameliyat
8	Feci yangın! Yaşlı adam hayatını kaybetti
9	İş Sanat Masal Tiyatrosu'nda bugün 'Rapunzel' ...
10	Kesilen narenciye ağaçları başında gözyaşı dök...

```
#çekilen haberler düzenlendi, makine öğrenmesi algoritmasına bu haberler verildi ve tahminleri gözlendi.
```

```
new_preprocessing=prep(new_news)
```

```
new_preprocessing=pd.Series(new_preprocessing["news"])
```

```
vectorizer_2=TfidfVectorizer()
vectorizer_2.fit(traininput_ofdata)
new_news_vector = vectorizer_2.transform(new_preprocessing)
```

```
new_value=log_reg_mdl.predict(new_news_vector)
```

```
new_news["value"]=new_value
```

```
new_news
```

	news	value
0	kadın kılığına girip milyonluk cipi çaldı	1
1	yargıtay kocasına hakaret eden kadın ile doğum...	1
2	ak partide merkez yürütme kurulu myk reform gü...	0
3	sondakika gabarda öldürülen teröristlerden si...	1
4	pınarın katiliyle ilgili flaş tespit	1
5	fransada çatışma yaşında bir genç hayatını ka...	1
6	thy kargo uçağı zorunlu iniş yaptı	1
7	gönüllere dokundu sokakta ameliyat	1
8	feci yangın yaşlı adam hayatını kaybetti	1
9	iş sanat masal tiyatrosunda bugün rapunzel var...	1
10	kesilen narenciye ağaçları başında gözyaşı dök...	1

RESULT AND INTERPRETATION

The Logistic Regression algorithm model, which we trained with the data we collected, predicted the accuracy and falsity of the new data. An accuracy ratio of 0.88 was obtained. This gives us the information that the predictions to be made will be 88% correct. Nevertheless, the fact that the data set consists of a single source and the use of a limited data set such as 2000 data will affect the estimates and predictions cannot always be made with this accuracy.

FUTURE WORK

In order to develop the project; The data collected for the dataset can be extracted from several different sources, different time periods, and the number of data can be increased for each. Thus, the accuracy value obtained can be fully accurate. Another method of Logistic Regression, Count Vektorization, can be used or the algorithm can be changed to try Naive-Bayes and Random-Forest algorithms.

REFERANCES

https://www.veribilimiokulu.com/blog/dogal-dil-isleme-nedir-ve-uygulama-alanlari-nelerdir/

https://medium.com/algorithms-data-structures/tf-idf-term-frequency-inverse-document-frequency-53feb22a17c6

https://www.udemy.com/

https://www.kaggle.com/onurakkse/veri-bilimi-notlar

https://towardsdatascience.com/14-popular-evaluation-metrics-in-machine-learning-33d9826434e4