**EGE ÜNİVERSİTESİ**

**MÜHENDİSLİK FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**YAPAY ZEKA YÖNTEMLERİ ve UYGULAMALARI**

**2019-2020 BAHAR YARIYILI**

**PROJE 2**

**YAPAY ZEKA / MAKİNE ÖĞRENMESİ / DERİN ÖĞRENME / ENİYİLEME PROJESİ**

Fatih Cüre

05170000113

1. **Problemin Tanımı**

Doğal Dil İşleme (Natural Language Processing) Alanında Bir Örnek

Ürün yorumlarının olumlu/olumsuz olduğunun bulunması.

1. **Ön çalışma**

Ön çalışma esnasında *duygu analizi nedir?*, *nasıl yapılır?*, *örnekleri nelerdir?*, *Türkçe’de duygu analizi nasıl yapılır?* gibi sorular üzerine durulmuştur. Çalışma içerisinde yararlanılan kaynaklar şöyledir:

**2.1.a. Videolar**

* [Doğal Dil İşleme (Natural Language Processing) (Veri Bilimi Serisi 40)](https://www.youtube.com/watch?v=a9I6MXgn8B4) <https://www.youtube.com/watch?v=a9I6MXgn8B4>

**2.1.b. Yazılı kaynaklar**

* **Doğal Dil İşleme Rehberi**<https://medium.com/t%C3%BCrkiye/do%C4%9Fal-dil-i%CC%87%C5%9Fleme-rehberi-2c4c41260f74>
* **Doğal Dil İşleme Nedir ve Uygulama Alanları Nelerdir?**  
  <https://www.veribilimiokulu.com/dogal-dil-isleme-nedir-ve-uygulama-alanlari-nelerdir/>
* **Doğal Dil İşleme (Natural Language Process)**  
  <https://medium.com/@fahrettinf/5-1-do%C4%9Fal-dil-i%CC%87%C5%9Fleme-natural-language-process-a92079c3d271>
* **Uygulamalı Türkçe Doğal Dil İşleme Evreleri**  
  <http://byoyo.cmpe.boun.edu.tr/sunumlar/gulseneryigit-byoyo18.pdf>
* [**Keras ile Duygu Analizi**](https://derindelimavi.blogspot.com/2017/10/keras-ile-duygu-analizi.html)  
   <https://derindelimavi.blogspot.com/2017/10/keras-ile-duygu-analizi.html>
* **Detection of Cyberbullying in Turkish -**  
   <https://www.kaggle.com/tbrknt/detection-of-cyberbullying-in-turkish>
* **Metin Madenciliği ile Metin Sınıflandırma – 2** <http://www.csharpnedir.com/articles/read/?id=782&title=Metin%20Madencili%C4%9Fi%20ile%20Metin%20S%C4%B1n%C4%B1fland%C4%B1rma%20-%202>

# TÜRKÇE METİN İŞLEMEDE İLK ADIMLAR <http://www.veridefteri.com/2017/11/20/turkce-metin-islemede-ilk-adimlar/>

# Python NLTK sentiment analysis <https://www.kaggle.com/ngyptr/python-nltk-sentiment-analysis>

* **Türkçe Sıfat Sözlüğü ile Duygu Analizi**  
  <https://www.researchgate.net/publication/330351991_Turkce_Sifat_Sozlugu_ile_Duygu_Analizi>
* **How to label text for sentiment analysis — good practices**  
  <https://towardsdatascience.com/how-to-label-text-for-sentiment-analysis-good-practises-2dce9e470708>
* **Unsupervised Sentiment Analysis**  
  <https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483>
* [**ITU Turkish Natural Language Processing Pipeli**](http://tools.nlp.itu.edu.tr/index.jsp)**ne**  
  <http://tools.nlp.itu.edu.tr/>
* **How to Prepare Text Data for Machine Learning with scikit-learn**  
  <https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/>
* [**Doğal Dil İşleme (NLP) ile Sentiment (Duygu) Analizi Tespiti**](https://emrahmete.wordpress.com/2018/11/25/dogal-dil-isleme-nlp-ile-sentiment-duygu-analizi-tespiti/)  
  https://emrahmete.wordpress.com/2018/11/25/dogal-dil-isleme-nlp-ile-sentiment-duygu-analizi-tespiti/
* **Doğal Dil İşlemenin Yakın Tarihi**  
  <https://medium.com/@basakbuluz/do%C4%9Fal-dil-i%CC%87%C5%9Flemenin-yak%C4%B1n-tarihi-6f8f62f6fbea>

**2.2. Neler Öğrendim?**

* Doğal dil işlemenin; doğal dillerin bilgisayar ortamında, makinelerin anlayabileceği ya da çıkarım yapabileceği ortam sağlanacak ortam oluşturularak verilerin anlam kazandırılması olduğu,
* Doğal dil işlemenin dil çevirisi, belge özetleme, rapor çıkartma, soru cevaplama -kişisel asistan gibi-, etiketleme, metin sınıflandırma ve daha bir çok konuda kullanıldığını,
* Duygu analizinin; makine ortamında verilen yazılı ya da fazladan birkaç katman ekleyerek sözlü olarak verilen girdilere karşılık olarak girdinin oluşturduğu duygu izleniminin yorumlanması olduğunu,
* N-gram algortimasının sıralı bir veride ki N adet sıralı veriyi elde etmede kullanıldığını,
* LSTM’in uzun vadeli bağlılıklarda kullanılabilecek bir çeşit RNN olduğunu ve RNN’den farklı olarak bir önceki bilginin bir sonraki zincir içerisine aktarmak ve aktarım esnasında bilgi ekleme/çıkarma gibi işelemler olduğunu,
* Naive Bayes algoritmasının, Bayes Teoremine dayanan ve bir verinin diğer bütün verilere olan yakınlığın (benzerliğinin) hepsinin hesaplanarak en yüksek oranda yakınlık verenin tahmin cevabı olduğunu,
* DA esnasında çeşitli yaklaşımlar olduğunu ve bunlardan birisinin anlam içeren, diğerinin anlam içermeyen yaklaşım olduğunu,

öğrendim.

* DDİ için, Derin Öğrenme modellerinin Makine Öğrenmesi yöntemine göre daha az etkili ve daha az işçilik –benzerliklerin DÖ modeli tarafından bulunması, MÖ modelinde ileri seviyelerde çoğu zaman DÖ’ye göre daha fazla elle uygulamak gerektirdiği gibi- gerektirdiği.

1. **Kullanılan Ortam, Yöntem ve Kütüphaneler**

**3.1. Kullanılan Ortam**

Proje geliştirim sürecinde programlama dili olarak Python 3.8.1 sürümü, Editör olarak Visual Studio Code 1.45 sürümü kullanıldı.

**3.2. Yöntem**

Naive Bayes ile LSTM yöntemleri tahminlemede, N-gram algoritması kelime grupları elde etmede kullanıldı.

**3.3. Kütüphaneler**

|  |  |  |
| --- | --- | --- |
| Kütüphane | Sürüm | Kaynak Adresi |
| nltk | **3.5** | <https://www.nltk.org/> |
| yaml | **3.12** | <https://pyyaml.org/wiki/PyYAMLDocumentation> |
| Turkish-deascifier | **N/A** | <https://github.com/emres/turkish-deasciifier> |
| TurkishStemmer | **1.3** | <https://github.com/otuncelli/turkish-stemmer-python> |
| Turkish-pos-tagger | **N/A** | <https://github.com/onuryilmaz/turkish-pos-tagger> |

Python ile gelen pickle, math, os, io, re, time ve random kütüphanelerinden yararlanıldı.

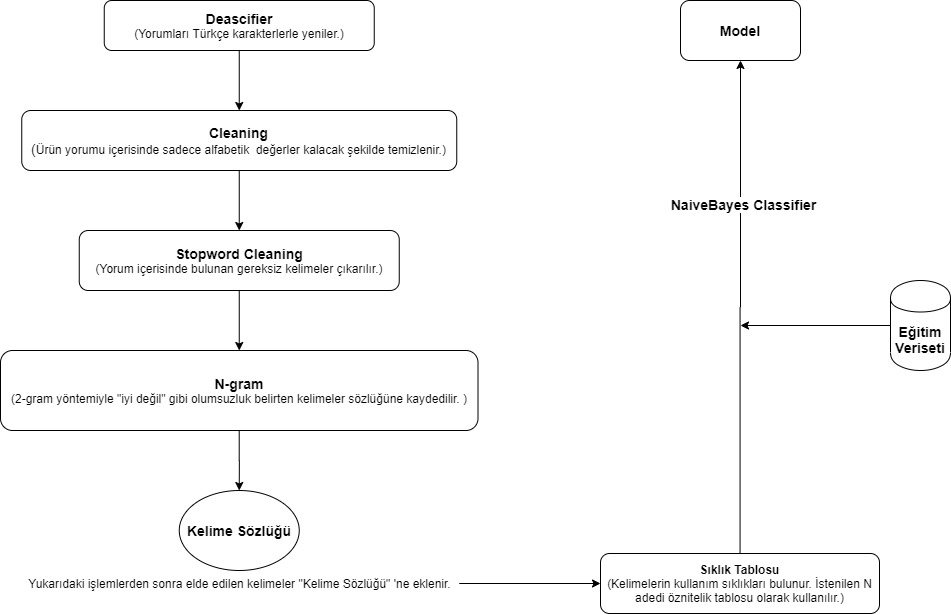
1. **Önerilen (Geliştirilen/Kullanılan) Yöntem**

Proje geliştiriminde iki farklı yöntem kullanılarak karşılaştırılmasının yapılması amaçlandı.

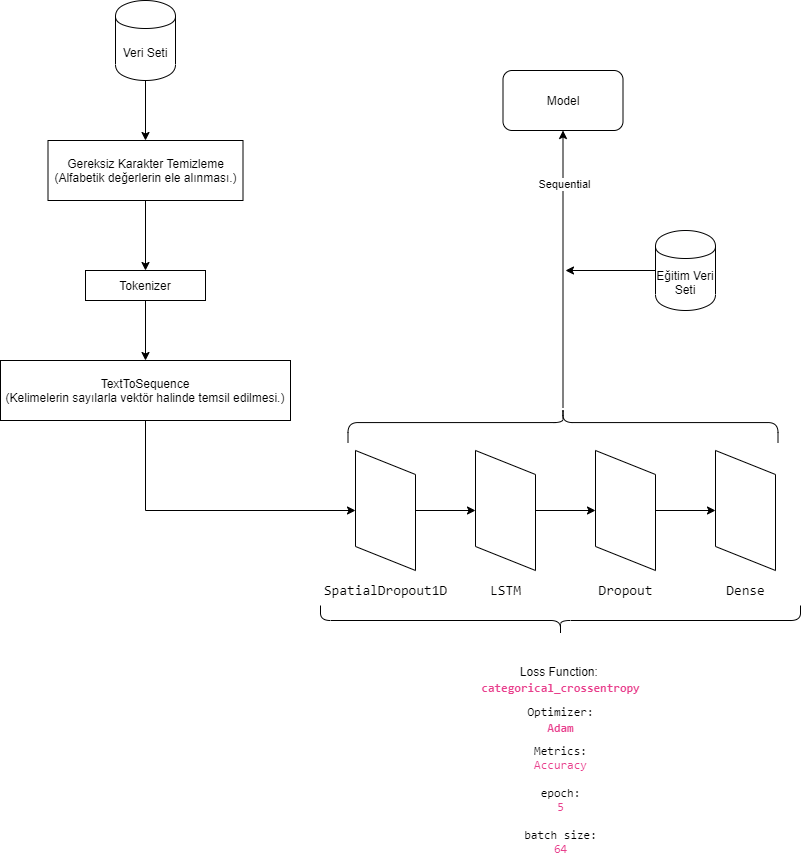
Bunlardan birincisi olan NaiveBayes kullanılırken bir yorumlardan elde edilen kelimelerin köklerinden sıklık tablosu oluşturulup öğrenme veri setiyle birlikte bu tablodan yararlanarak her bir yorumun sıklık tablosundaki her bir elemana karşılık kelimeyi içerip içermediğini gösterecek şekilde birer tablo elde edilmesi, ardından tahmin edilecek olan ürün yorumunun bu tablolar arasında en yüksek olasılıkla benzerlik gösteren (NaiveBayes burada işlevli geri kalan kısımlar el ile yapılıyor.) sonucun elde edilmesi sağlanıyor. Ek olarak tahmin edilecek yorumun içerisinde olumsuz yapan “değil” aranıp öncesinde bulunan (n-gram ile kelime grupları elde ediliyor.) kelime sıfat ise o “değil” kelimesine kadar olan değerlendirme sonucuna göre pozitif ise –2, negatif ise +2 puanlama yapılıyor. Sonuçta elde edilen değer 0’dan büyük ise pozitif değil ise negatif olarak yorumlanıyor. [1]

İkinci yöntem olan LSTM ile veri setindeki yorumlar pozitif ve negatif olmak üzere iki gruba ayrılıp daha sonra Naive Bayes ile ortak işlemler olan kelimelerin anlamsız karakterlerden arındırılması ve anlamsız kelimelerden arındırılması gibi işlemler uygulanıyor. Sonrasında LSTM modeli ile eğitilerek tahminleme sonucu elde ediliyor.

NaiveBayes yöntemi için çalışma düzeni:



LSTM yöntemi için çalışma düzeni:



1. **Deneysel Çalışmalar**

**5.1. Veri Seti**

Proje için iki adet veri seti mevcut. Bunlardan ilki olan ‘reviews’ adlı veri seti internet ortamından hazır elde edildi. Veri setinde pozitif yorumlar reviews.pos, negatif yorumlar review.neg dosyaları içerisinde barındırılıyor. Yorumlar çeşitli e ticaret sitelerinden elde edilerek hazırlanmış. Yorumların tamamı çeşitli ürünlere yapılan yorumlarından oluşmakta ve toplamda olumlu ve olumsuz olmak üzere iki sınıftan oluşmakta. Negatif yorum sayısı 14.881, pozitif yorum sayısı 220.284 olacak şekildedir ancak kullanım esnasında donanım özellikleri karşılamadığı için her ikisinden ilk 1000 yorum örnek alınarak kullanıldı.

Veri seti: <https://github.com/fthbrmnby/turkish-text-data>

Diğer veriseti ‘kitapyurdu’ benim tarafımdan kitapyurdu.com sitesinden veri toplayarak oluşturuldu. Veri setinin oluşturulmasında kitapyurdu.com sitesinde bulunan “100 Temel Eser” kategorisinden 5 yıldız verilmiş olan yorumları pozitif olarak değerlendirerek, “Kelepir Kitaplar” kategorisinden 3 ve daha az yıldız verilmiş olan yorumları negatif olarak değerlendirerek toplandı. Negatif yorum sayısı 826, pozitif yorum sayısı 19.799 adettir ancak kullanım esnasında negatiflerin tamamı pozitif yorumların ilk 1128 adeti kullanılmıştır.

Veri seti proje dosyası içerisinde kitapyurdu klasörü adı altında bulunmaktadır. Veri seti üretmek için kullanılan kod parçası “scraper” dosyası altında bulunuyor. (Üretim için BeatifulSoup ve request kütüphaneleri kullanıldı.)

Yapılmış çalışmaların çıktıları “Deneysel Çalışmalar” dosyası içerisinde kıs açıklamalar ile mevcut.

**Ek 1.** Başarım iyileştirme için Naive Bayes’te “değil” kelimesinin verdiği olumsuzluğu göz önünde bulundurmaya çalışıldı. Ayrıca İngilizce karakterlerle yazılmış yorumların kelime sözlüğünde iki ayrı yer tutacağı hesaplanarak Türkçe karakterleştirme uygulandı. POS(part-of-speech) adı verilen kelimenin görevini elde etmek için Türkçe verilerle eğitilmiş hazır modelden yararlanıldı.

**Ek.2**

|  |  |  |  |
| --- | --- | --- | --- |
|  | [Cumali TÜRKMENOGLU Tez Çalışması](http://www.tantug.com/wp-content/uploads/2016/05/YOK-CumaliTurkmenoglu.pdf) | [turkce-duygu-analizi](https://github.com/BaharYilmaz/turkce-duygu-analizi) | **Önerilen Çalışma** |
| **İstatistiksel sınıflandırma** | Naive Bayes | TfidfVectorizer  LogisticRegression  CountVectorizer | Naive Bayes |
| **Derin Öğrenme** | - | - | LSTM |
| **Makine Öğrenmesi** | Karar Ağaçları | - | Hazır model kullanıyor.  (POS Prediction) |
| **Olumsuzluk Durumunun Kontrolü** | Çok çeşitli olarak var. | **-** | ‘değil’ kelimesi için olumsuzluk kontrolü var. |

**Ek.3** Yararlanılan benzer uygulamalar:

<https://towardsdatascience.com/basic-binary-sentiment-analysis-using-nltk-c94ba17ae386>

<https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17>

**Özdeğerlendirme Tablosu**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **İstenen Madde** | **Var** | **Açıklama** | **Tahmini Not** |
| **1** | **Kapak Sayfası, Problemin Tanımı, Kullanılan Ortam, Yöntem ve Kütüphaneler (10)** |  | Yapıldı. | 10 |
| **2** | **Araştırma (10)** |  | Yapıldı. | 10 |
| **3** | **Önerilen Yöntem (10)** |  | Yapıldı. | 10 |
| **4** | **Deneysel Çalışmalar** + **Ek 1: Başarım İyileştirme (10)** |  | Yapıldı. | 10 |
| **5** | **Proje Rapor Biçimi, Organizasyonu, Boyutu, Kalitesi, Kaynakça ve atıflar (10)** |  | Yapıldı. | 10 |
| **6** | **Ek 2 ve Ek 3 (10)** |  | Yapıldı. | 10 |
| **7** | **Ara rapor ve demosu (çalışma sürekliliği ve kalitesinin görülmesi açısından) (10)** |  | Yapıldı. | 10 |
| **8** | **Teams üzerinden Demo, Kodlama gösterimi ve Soru Cevap (20) (Sonradan yapılacağı için yaklaşık değer verebilirsiniz)** |  | Yapılması tahmin ediliyor. | 10 |
| **9** | **Özdeğerlendirme Tablosu (10)** |  | Yapıldı. | 10 |
| **100 üzerinden Toplam Not:** | | | | 100 |