EGE ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

YAPAY ZEKA YÖNTEMLERİ ve UYGULAMALARI

2019-2020 BAHAR YARIYILI

PROJE 2 - RAPOR

DOĞAL DİL İŞLEME – DUYGU ANALİZİ

05160000581 CEREN ERDOĞAN

05160000539 AYTUĞ SEVĞİ

İÇİNDEKİLER

1.PROBLEM TANIMI

www.kaggle.com sitesinde bulunan "Turkish Movie Sentiment Analysis" veri setini kullanarak insanların bir film sitesinde, izlediği film hakkında yaptığı yorumların hangilerinin 'olumlu' hangilerinin 'olumsuz' bir yorum olduğunu doğal dil işleme metodu aracılığıyla bulmaktır.

2.ARAŞTIRMA (ÖN ÇALIŞMA)

Son yıllarda yapay zeka konusundaki gelişmelerin yanında doğal dil işleme tabanlı geliştirilen uygulamaların önemi ve sayısı artmakta. Bu konudaki gelişmeler birçok işte fayda sağlamakta ve insan gücünü ortadan kaldırmaktadır. Daha hızlı ve daha doğru sonuçlar elde etmek için kullanılan yöntemler geliştirilmeye devam etmektedir. Doğal dil işleme ile geliştirilen uygulamalar kullanım alanları bu şekilde karşımıza çıkmakta:

- Spam Detection (İçeriklere göre sınıflandırma)
- Sentiment Analysis (Metin içeriğinde olumlu/olumsuz söylem olup olmadığı çıkarılır. Bir ürün/film vb. yorumların analiz edilmesinde kullanılır.)
- Part-Of-Speech (POS) Tagging (Metin etiketleme (Fiil, Nesne, Sıfat, Bağlaç vb.))
- Named Entity Recognation (NER) (Varlık ismi tanımlama kişi, yer, zaman, tarih, sayı tanımlamalarının yapılması)
- Coreference Resolution (Metinde aynı şeyden(kişi/nesne) birden fazla kullanılmasının tespiti)
- Word Sense Disambiguation (WDS) (Çok anlamlı, eş sesli kelimelerin, cümle içerisinde hangi anlamda kullanıldığının anlaşılması)
- Parsing: (Metni parçalara bölme işlemi)
- Machine Translation (MT) (Bir dilden başka bir dile çevirme)
- Information Extraction (Metnin içerisinden bilgi çıkarma)

Seçtiğimiz konu nedeniyle duygu analizi üzerine gideceğiz. Duygu analizi insanların duygularını, görüşlerini değerlendiren, bir konu hakkındaki davranışlarını yazılı dil üzerinden analiz eden çalışma alanıdır (Liu, 2012). Duygu Analizi, temel olarak bir metin işleme işlemi olup, metindeki ifade edilmek istenen duyguyu belirlemeyi amaçlar. Bir metnin ifade ettiği olumlu, olumsuz veya birden fazla duygunun belirlenmesi için kullanılır. Fikir madenciliği olarak da duyabileceğimiz duygu analizi örnek olarak bir filme gelen yorumların taşımış olduğu fikir, duygu ve düşünceyi semantik olarak ortaya çıkarmak için yapılan çalışmalardır.

Bir metindeki duygu analizi, yani o metnin duygusallığı genel olarak ikili sınıflandırmalarla gösterilir (Olumlu/Olumsuz, Pozitif/Negatif). Duygusal kutuplara (ikili sınıflandırma) ayırma işlemi ilk duygu analizi çalışmalarında sıkça görülmüştür. Daha sonra geliştirilen uygulamalarda metni duygu kategorilerine ayırarak daha detaylı sınıflandırma yapılmıştır. Biz sadece metnin olumlu/olumsuz olup olmadığına karar veren bir algoritma geliştirip ikili sınıflandırma işlemini gerçekleştireceğiz.

3.KULLANILAN ORTAM, YÖNTEM, KÜTÜPHANE

Projeyi kaggle notebook ortamında geliştirdik. Deneysel olarak bazı sınıflandırıcılar denendi. Bu sınıflandırıcılar; RandomForest, LogisticRegression, KNeighborsClassifier olmak üzere 3 tanedir. TfIdfVectorizer kütüphanesi ile genel olarak metinde geçen kelimelere bağlı olarak belirlenen text verisinde bir kelimenin nadirliğinin analizi yapıldı. Train_test_split kütüphanesi ile veri seti "train" ve "test" olmak üzere parçalandı. Classification_report kütüphanesi ile en optimal modelin raporunu ekrana yazdırdık.

4.GELİŞTİRİLEN/KULLANILAN YÖNTEM

İzlenilen Yöntem:

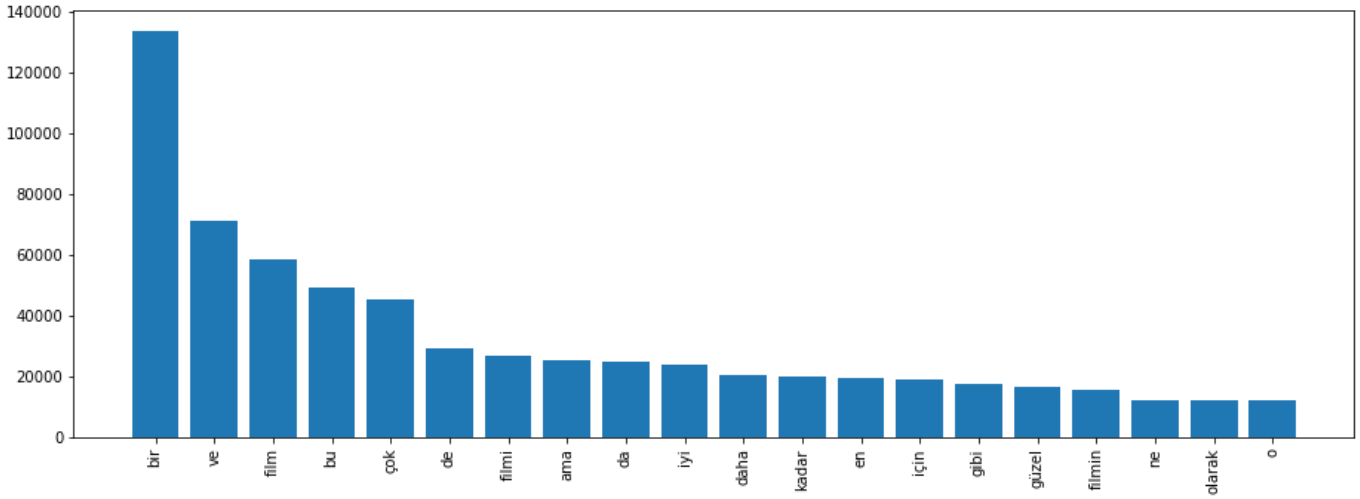
- Veri seti ile dataframe oluşturuldu.
- Veri seti içerisinde ki "\n,\r,\t", noktalama işaretleri ve gereksiz boşluklar temizlenip daha temiz bir veri seti haline getirildi.
- TfidfVectorizer() kütüphanesi ile veri setindeki yorumlar fitlendi. Buna göre belirli bir yorum içinde geçen kelimelerin nadiren kullanılan bir kelime olup olmadığının analizi yapıldı.
- "Na" değer içeren satırlar veri setinden silindi.
- Veri setinde ki Score sütunu String'ten float'a dönüştürüldü.
- Score'ların grafiği çizdirildi.
- Yorumlarda en çok geçen kelimelerin grafiği çizdirildi.
- Veri seti "train" ve "test" olmak üzere ayrıldı.
- Accuracy hesabı için "accuracy_summary" fonksiyon yazıldı.
- Modeller oluşturulup feature ve model içi parametrelere göre pipeline'a parametre olarak gönderildi. "nfeature_accuracy_checker" fonksiyonu ile bu işlem yapılırken bu fonksiyon kendi içinde "accuracy_summary" i çağırarak pipeline.fit() ile model eğitildi. Buna göre score ekrana yazıldı.
- En optimal model ve parametre değerleri için "classification_report" ile ekrana bu modelin raporu yazdırıldı.

5.DENEYSEL ÇALIŞMALAR

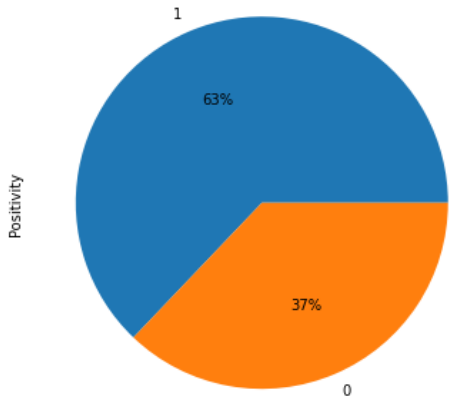
5.1.Veri seti; yapılan yorum, film adı, üyenin filme verdiği puan (0-5) şeklinde 3 adet öznitelik içermektedir. Örnek sayısı 82456 tanedir. 7722 adet filme gelen Türkçe dilinde yazılmış yorumlar pozitif ve negatif anlam içermektedir. Kullanılan veri seti linki: https://www.kaggle.com/mustfkeskin/turkish-movie-sentiment-analysis-dataset

5.2.Veri Analizi/Grafikler

Veri setinde en çok kullanılan ilk 20 kelime ve kaç defa kullanıldıkları grafikte verilmiştir.



Yorumların pozitif veya negatiflik oranları grafikte verilmiştir.



5.4.Train-Test Verilerinin Oluşturulması

sklearn.model_selection - eğitim _test_split kütüphanesi yardımıyla 0.25 test, 0.75 eğitim olmak üzere veri seti parçalandı. Buna göre test ve eğitim veri sayıları şöyledir;

```
Train set has total 62420 entries with 37.04% negative, 62.96% positive
```

```
Test set has total 20807 entries with 37.36% negative, 62.64% positive
```

5.5.Verileri Sınıflandırılması

Deneysel çalışma için kullandığımız sınıflandırma yöntemlerinin parametre değerleri ve çıkan sonuç şu şekildedir.

Features (10000,50000) için sonuçlar:

KNeighborsClassifier(n_neighbors=2) ve feature=10000 için doğruluk değeri max. 63.28% olmuştur.
KNeighborsClassifier(n_neighbors=3) ve feature=10000 için doğruluk değeri max. 63.88% olmuştur.
RandomForestClassifier(class_weight='balanced') ve feature=40000 için doğruluk değeri max. 77.79% olmuştur.
LogisticRegression(c=0.01) ve feature=10000 için doğruluk değeri max. 65.48% olmuştur.
LogisticRegression(c=0.05) ve feature=10000 için doğruluk değeri max. 75.28% olmuştur.
LogisticRegression(c=0.25) ve feature=20000 için doğruluk değeri max. 78.85% olmuştur.
LogisticRegression(c=0.5) ve feature=30000 için doğruluk değeri max. 79.62% olmuştur.
LogisticRegression(c=1) ve feature=30000 için doğruluk değeri max. 79.97% olmuştur.

Sonuçlardan da görüldüğü üzere **LogisticRegression() modeli için c=1 ve feature= 30000** değeri için max accuracy score bulunmuştur.

En optimal model ve parametre değerleri (LogisticRegression) için modelin raporu:

	precision	recall	f1-score	support
negative	0.72	0.64	0.68	7774
positive	0.80	0.85	0.82	13033
accuracy			0.77	20807
macro avg	0.76	0.74	0.75	20807
weighted avg	0.77	0.77	0.77	20807

5.KAYNAKÇA

- https://www.whoson.com/customer-service/top-ten-benefits-of-sentiment-analysis/
- https://medium.com/algorithms-data-structures/do%C4%9Fal-dil-i%CC%87%C5%9Fleme-nlp-nin-uygulamalarda-kullan%C4%B1m-alanlar%C4%B1-792e4aac87bd
- https://www.lexalytics.com/technology/sentiment-analysis
- https://emrahmete.wordpress.com/2018/11/25/dogal-dil-isleme-nlp-ile-sentiment-duygu-analizi-tespiti/
- https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
- https://monkeylearn.com/sentiment-analysis/
- https://www.analyticsinsight.net/benefits-of-sentiment-analysis-for-businesses/
- https://www.kaggle.com/kerneler/starter-turkish-movie-sentiment-36ef5175-0/notebook
- https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b
- https://towardsdatascience.com/latent-semantic-analysis-sentiment-classification-with-python-5f657346f6a3

EK 1

Feature değerlerini (10000,50000) arasında yaparak en optimal feature değerini sınıflar için bulduk.

KNeighbors yöntemiyle aldığımız en yüksek sonuç 63.88% oldu. Bu score'u arttırabilmek için RandomForest yöntemini denedik. Bu yöntemle alınan en yüksek sonuç xxx oldu. LogisticRegression bizim için en güzel sonucu veren yöntem oldu. C=1 ve feature=30000 değerleri için 79.97% sonucunu aldık. Eğer KNeighbors ile yapıp başka bir sınıflandırma yöntemi denemeseydik oluşturulan modelin sonucu çok daha düşük olacaktı. Başlangıca göre sonuç değeri 16.19% arttı.

EK 2

Günümüzde bir filmi izlemeden önce çoğu kişi yorumlarına bakmaktadır. Boş zaman değerlendirmek için film izlemeyi seçen kişiler filmin zaman kaybı olmasını istemez. Yorumlara bakılırken film hakkında önemli bir detayın görülmesi ya da filmin sonu ile ilgili bilgilerin öğrenilmesi neredeyse kaçınılmazdır. Film sitelerini yöneten (admin) kişiler, yorum yapan üyelere eğer film senaryosu hakkında detaylı bilgi vereceklerse 'spoiler' butonu eklemiştir. Bu butonun amacı, filmin kötü ya da iyi olduğu dışındaki yorumları gizlemektir. Filmi bitiren bir üye artık tüm yorumları okuyabilir hale gelir ve buğulu metinler 'spoiler' butonuna basıldığı takdirde görünür hale gelir. Fakat tüm kullanıcılar bu butonu kullanmaz ve yine kaçınılmaz bir şekilde büyük bir merakla filme başlamadan yorumları okuyacak olan izleyici filmin sonunu yorumlarda görür. Sonu bilinen bir filmi izlemek kişinin kendi tercihidir fakat çoğu insan filmi izlemekten vazgeçer.

Yapacağımız çalışma ile artık yorum okumaya gerek kalmayacak. Duygu analizi ile çözümlenen veriler belirleyeceğimiz yöntemlerle, bir yorumun olumlu/olumsuz olup olmadığını belirleyecek. Böylece bir filmin yüzde kaç iyi/kötü olduğu bulunacak. Veri setinde bulunan her film ayrı gruplandırıp o film için yüzde kaç olumlu/olumsuz yorum yazıldığını görmüş olacağız.

EK 3

Kaggle sitesinde veri setinin veri setinin bulunduğu yerdeki notebook kısmından yararlanıldı.
https://towardsdatascience.com/latent-semantic-analysis-sentiment-classification-with-python-5f657346f6a3
https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b
Yararlandığımız kaynaklardan veri setimiz farklıdır. Tek bir model üzerinden proje geliştirmişler.
Veri setini daha temiz hale getirirken kendi ihtiyaçlarımıza göre olan yöntemler kullandık.
Fakat biz birçok sınıflandırma yöntemi kullandık.
Feature, sınıflandırma yöntemlerinin parametre değerlerini değiştirerek ile bizim veri setimiz için en uygun sonucu bulduk.

Sonuç olarak örneklerde İngilizce yazılmış yorumların analizi eğitimi söz konusu. Bizim veri setimizdeki yorumlar Türkçe olduğu için Türkçe' de bu analizin daha zor olduğunu düşünüyoruz. Fazlasıyla esnek ve herhangi küçük bir ek ile cümlenin anlamı değişebiliyor. Fakat buna rağmen aldığımız sonuçlar çok güzel. Örneğin:

"1-Pozitif, 0-Negatif olmak üzere"

```
y_pred = sentiment_fit.predict(["İzlerken zevk aldım çok güzeldi"])
y_pred
```

```
array([1])
```

+ Code + Markdown

```
y_pred = sentiment_fit.predict(["İzlemenizi tavsiye etmem"])
y_pred
```

```
array([0])
```

```
y_pred = sentiment_fit.predict(["İzlemenizi tavsiye ederim"])
y_pred
```

```
array([1])
```

5.ÖZDEĞERLENDİRME TABLOSU

	İstenen Madde	Var	Açıklama	Tahmini Not
1	**Kapak Sayfası, Problemin Tanımı, Kullanılan Ortam, Yöntem ve Kütüphaneler (10)**	✓	İstenilenler yapılmıştır. Belki Türkçe kelimeleri köküne ayırabilmek için bir kütüphane kullanılabilirdi. Fakat bu işlemin veri setine uygulanınca günler sürdüğünü fark ettik.	9
2	**Araştırma (10)**	✓	Birçok kaynaktan konu hakkında bilgi edindik ve bize uygun olan yöntemleri uyguladık.	10
3	**Önerilen Yöntem (10)**	✓	LogisticRegression yöntemi ile en iyi sonucu aldık.	10
4	**Deneysel Çalışmalar (10)**	✓	3 tane farklı sınıflandırıcı ile score'lar karşılaştırıldı ve her biri için farklı parametre değerleri denendi. Sonuçlar yorumlandı.	10
5	**Proje Rapor Biçimi, Organizasyonu, Boyutu, Kalitesi, (10)**	✓	Raporu özenle hazırladık ve istenilen her maddeyi açıkladık.	10
6	**Ek 1: Başarım İyileştirme (10)**	✓	%13-14 arası deneysel çalışmalar ile başarım iyileştirildi.	10
7	**Ek 2 (10)**	✓	Neden bu konuyu seçtiğimizi ve yaptığımız uygulamanın farklılığını açıkladık.	9
8	**Ek 3 (10)**	✓	Bu kısımda yararlandığımız çalışmalar ve farklılıklarımız belirtildi.	10
9	**Kaynakça ve atıflar (10)**	✓	Kullanılan kaynaklar (stackoverflow hariç) belirtildi.	9
10	**Özdeğerlendirme Tablosu (10)**	✓	Tabloda gerekli açıklamaları yaptığımızı düşünüyoruz.	10
			100 üzerinden Toplam Not:	97

İş bölümü: Ceren Erdoğan, veri setini buldu. Veriyi temizleyip analizini yaptı. Eğitim ve test verilerini oluşturdu. Aytuğ Sevği, farklı sınıflandırıcı yöntemlerini araştırarak bu sınıflandırıcılar ile modeli oluşturdu. En optimum sonuca parametre değerlerini değiştirerek beraber ulaştık. Raporu ve eklerde istenenleri birlikte yaptık.