

YouTube Yorumlarının Duygu Analizi
Yapay Sinir Ağları ile Türkçe
Sınıflandırma

Gereksinim Analiz
Raporu

Ders adı: Yapay Zeka Teknikleri

Hazırlayan: Furkan Yıldırım

İçindekiler

Tablosu



Projenin İçeriği

01.

Projenin Tanımı ve Amacı

Youtubeden eğitim içerikli videoların yorumları alınarak yapay Ağı ile duygu durumlarını ortaya çıkarmak

02.

Yorumların Ve Meta Verilerinin Örneği

Projede kullanılan verilerin örneğini burada bulabilirsiniz.

03.

Yöntemlerin ve Sonuçların Açıklaması

Proje içerisinde kullandığım yöntemlerin, yaptığım işlemlerin ve oluşturmuş olduğum sistemlerin açıklamaları yer almaktadır.

04.

Sonuç ve Yorumlar

Geliştirmiş olduğum sistemler sonucunda elimizde ne olduğu ve yorumların analiz edildiği kısmı buradadır.

05.

Kaynaklar

Proje genelinde kullandığım kaynakların detaylı açıklamaları bu bölümde yer almaktadır.

Tanım:

Bu proje, YouTube platformunda yayınlanan eğitim içeriklerinin altına yapılan Türkçe yorumların duygu analizini yapmayı hedeflemektedir. Çalışmada pozitif, negatif ve nötr olmak üzere üç sınıfa ayrılan yorumlar, Yapay Sinir Ağları (YSA) yöntemi kullanılarak sınıflandırılacaktır. Projenin kapsamı, hem veri toplama ve işleme sürecini hem de sınıflandırma algoritmasının geliştirilmesini içermektedir.

Amaç:

Projenin temel amacı, kullanıcıların eğitim videoları hakkındaki düşünce ve hislerini anlamaktır. Bu bağlamda:

- Yorumların pozitif mi, negatif mi yoksa nötr mü olduğunu belirlemek,
- Eğitim videolarına verilen geri dönüşlerin daha iyi analiz edilmesine olanak sağlamak,
- Türkçe doğal dil işleme (NLP) tekniklerinin kullanımını yaygınlaştırmak ve bu alandaki uygulamalara katkıda bulunmak, hedeflenmiştir.

Bu analiz, içerik üreticilerine geri bildirim mekanizması sağlamanın yanı sıra, eğitim materyallerinin kalitesini artırmak için kullanılabilir.

Hedef:

Projenin sonunda:

- Eğitim videolarına ait yorumların %80 ve üzeri doğruluk oranıyla pozitif, negatif ve nötr olarak sınıflandırılması,
- Yapay Sinir Ağları yöntemiyle yüksek performans gösteren bir model geliştirilmesi,
- Yorumların analizi için uygulanabilir bir arayüz tasarlanması beklenmektedir.

Bu sonuçlar, eğitim içeriklerinin kullanıcılar üzerindeki etkisinin daha objektif bir şekilde değerlendirilmesini sağlayacaktır.

Yorumların ve Meta Verilerinin Örneği

| | Yorum ID | Yorum Metni | Beğeni Sayısı | Yayınlanma Tarihi | 5 |
|----|------------------|--|---------------|-------------------|---------|
| 4 | Ugw7mplCpVPj... | Teşekkürler efendim, sizinle tanıştıktan sonra benim için ilham alıyorsunuz aslında doğru yönde ilerliyorumşum gibi hissediyorum, ne kadar zor iş yaptığınızı ve toplumumuza ne kadar zor iş yaptığınızı ... | 0 | 2022-06-20T10:... | Pozitif |
| 5 | UgzLalUUy8nx2... | Bu diziye başladığınız için teşekkürler brü r a | 0 | 2022-06-20T10:... | Nötr |
| 6 | UgyrUwlQnpSP... | dokunadı | 0 | 2022-06-17T17:... | Nötr |
| 7 | UgwJgRWhPCFJ... | Efendim zamanına gelecek | 0 | 2022-06-15T20:... | Nötr |
| 8 | UgxyJSqYgNpq... | | 1 | 2022-06-13T07:... | Nötr |
| 9 | UgwLIBhP-... | Efendim e -posta kimliği | 0 | 2022-06-05T06:... | Nötr |
| 10 | UgzgeEYfjsb5R... | Bence bence harika bir insan tutkuyla tutku ile zihniyetle, düşünme tarzı kesinlikle düz ileni brlove Sırbrby Ben senin öğrenciyim | 0 | 2022-06-03T18:... | Nötr |
| 11 | Ugy6s5nCtjv03... | Sen harika öğretmensin kardeşim | 0 | 2022-06-01T18:... | Nötr |
| 12 | UgxDebywEn8G... | Çok ilham verici efendim bryour öğretme yolu çok iyi, bunu daha önce hiç görmedim | 0 | 2022-06-01T04:... | Nötr |
| 13 | Ugz72nQfj4X-3... | Sen çılgın bir adamsın harika bir hikaye anlatıcısın, senden öğrenilecek çok yerlere gideceksin | 0 | 2022-05-31T19:... | Nötr |
| 14 | UgwA85oqcn4... | Gerçekten efendim harika bir insan olacaksınız Bu tüm dilekler kalbimin dibinden senin için var | 0 | 2022-05-31T18:... | Pozitif |
| 15 | Ugy7Xfz4e2iZ... | Nitish kesinlikle bir gün büyük bir isim olacaksın | 0 | 2022-05-31T18:... | Nötr |
| 16 | UgwhMSLM-... | Buna çok ikna oldum doğrusun | 0 | 2022-05-31T18:... | Nötr |
| 17 | UgwZ7AVWrvz... | Salam Efendim Efendim Kalküta'dan Love | 0 | 2022-05-31T15:... | Nötr |
| 18 | UgwK23_nFLF7r... | | 0 | 2022-05-31T14:... | Nötr |
| 19 | Ugwa1R6pzBc... | Paylaştığınız için çok teşekkür ederim efendim evet sen harika bir öğretmensin ve evet ben öğrenciyim | 0 | 2022-05-31T14:... | Pozitif |
| 20 | UgzxmJCroQ_u... | Yakında en büyük kişilik olacaksın ve size çok başarılı olacaksın | 0 | 2022-05-31T12:... | Negatif |
| 21 | UgxLcP6HGmQ... | Elbette başarılı olacaksın biliyoruz | 0 | 2022-05-31T11:... | Negatif |
| 22 | UgzQWsf4Ba1Z... | Çok iyi video | 0 | 2022-05-31T11:... | Nötr |
| 23 | UoznRit9wtc-... | merhaba efendim youtube pe kasam se itna kachra hai ki pucho matt apke kanal tak sair hi dosya eksik Hai Kabhi Kuch kaam nahi kar raha ama kasam se aapke kanal pe AB ab tak aisi problem nahi hai... | 0 | 2022-05-31T10:... | Pozitif |

Yöntemler ve Sonuçların Açıklanması

Veri Toplama:

Veri toplama aşamasında Kaggle web sitesi kullanılarak belirlenen anahtar kelimelerle eğitim videolarına ait yorumlar ve meta veriler çekilmiştir. Çekilen veriler şu bilgileri içermektedir:

- snippet_topLevelComment_id: Kullanıcı yorumu id'si
- snippet_topLevelComment_snippet_videoId: Kullanıcı yorumunun bulunduğu video id'si
- snippet_topLevelComment_snippet_textDisplay: Kullanıcı Yorumunun görüntülenicek kısmı
- snippet_topLevelComment_snippet_textOriginal: Orijinal kullanıcı yorumu
- snippet_topLevelComment_snippet_viewerRating: Kullanıcı yorumu görüntülenme sayısı
- snippet_topLevelComment_snippet_likeCount: Kullanıcı yorumu beğeni sayısı
- snippet_topLevelComment_snippet_publishedAt: Kullanıcı yorumunun yayınlanma tarihi
- snippet_topLevelComment_snippet_updatedAt: Kullanıcı yorumunun güncellenme tarihi
- snippet_canReply: Kullanıcı yorumu tekrarlanabilir mi?
- snippet_totalReplyCount: Kullanıcı yorumu total yeniden yayınlanma sayısı
- snippet_isPublic: Kullanıcı yorumu herkese açık mı?
- translated_tr: Kullanıcı yorumu Türkçe diline çevrilmiş hali.

Tüm veriler saf halinde alınmıştır ve ilgili yöntemsel metotlar ile modelin öğrenmesi için uygun formata getirilmiştir. Bu Format için src/services/comment_service dosyasındaki CommentService classı kullanılmıştır

Örnek Veri Satırı Saf Hali:

→ 1, UgzYqnTw6zHursr0gvB4AaABAg, DRXYOfbksGE, Thank you for this. Please add more questions for sql, Thank you for this. Please add more questions for sql, none, 0, 2022-09-22T20:58:06Z, 2022-09-22T20:58:06Z, True, 0.0, True, Bunun için teşekkürler lütfen SQL için daha fazla soru ekleyin

Veri İşlemleri:

```
def process_all_steps(self, dosya_adi):
```

```
    # Yorumları temizle
```

```
    self.data['yorum_temiz'] =
```

```
self.data['snippet_topLevelComment_snippet_textDisplay'].apply(self.correct_comment)
```

```
    # Yorumları çevir
```

```
    for index, yazi in enumerate(self.data['yorum_temiz']):
```

```
        self.data.at[index, 'snippet_topLevelComment_snippet_textDisplay'] = self.translate_tr(yazi, index)
```

```
        if index % 100 == 0:
```

```
            print(f"[{index}] Çeviri işlemi devam ediyor...")
```

```
    # Gereksiz geçici sütunları kaldır
```

```
    self.data.drop(columns=['yorum_temiz'], inplace=True)
```

```
    self.save_to_csv(f'./datasets/{dosya_adi}_translated.csv', ['snippet_topLevelComment_snippet_textDisplay'])
```

```
# Tokenizasyon
self.tokenize_comments()
self.save_to_csv(f'./datasets/tokenized/{dosya_adi}_tokenized.csv', ['yorum_tokenize'])

# Stopword'leri çıkarma ve stemming
self.data['yorum_no_stopwords'] = self.data['yorum_tokenize'].apply(self.remove_stopwords)
self.data['yorum_stemmed'] = self.data['yorum_no_stopwords'].apply(self.stem_words)
self.data['yorum_islenmis'] = self.data['yorum_stemmed'].apply(lambda tokens: " ".join(tokens))
self.save_to_csv(f'./datasets/processed/{dosya_adi}_processed.csv', ['yorum_islenmis'])

# Duygu analizi
self.analyze_sentiment(f'{dosya_adi}_processed')
```

- Öncelikle yorumları Türkçeye çeviriyor.
- Gereksiz geçici sütunları kaldırıyoruz.
- Tokenizasyon işlemlerine başlıyoruz. Burada yorumların her bir kelimesini ayrı ayrı ayırma işlemini gerçekleştiriyoruz. Örnek bir Girdi: "Bu video çok faydalıydı!"; Çıktı: ['video', 'çok', 'faydalı']
- StepWords kelimeleri de çıkardık. Bunlar genel olarak bağlaçlar olarak bilinirler. Bu kelimeler modeli eğitirken değerlendirilmemesi için çıkarılırlar.

Yapay Sinir Ağı Modeli ve Optimum Yapı:

Bu çalışmada, duygu analizi yapmak üzere bir Yapay Sinir Ağı (YSA) modeli geliştirilmiştir. Modelin yapısı Keras kütüphanesi kullanılarak Python dilinde oluşturulmuş ve eğitim sürecinde farklı yapılandırmalar denenerek optimum model elde edilmiştir. Modelin detayları aşağıdaki gibidir:

1. Model Yapısı

- **Giriş Katmanı:** Giriş verisi, metin verilerini sayısal verilere dönüştürmek için kullanılan bir **Tokenizer** yardımıyla hazırlanmış ve ardından 4 boyutlu bir dizilim haline getirilmiştir.
- **Gizli Katmanlar:**
 - **İlk Katman:**
 - 512 nöron.
 - **ReLU** (Rectified Linear Unit) aktivasyon fonksiyonu.
 - %50 Dropout uygulaması.
 - **İkinci Katman:**
 - 256 nöron.
 - **ReLU** aktivasyon fonksiyonu.
 - %50 Dropout uygulaması.
- **Çıkış Katmanı:**
 - 3 nöron (pozitif, negatif, nötr sınıfları için).
 - **Softmax** aktivasyon fonksiyonu.
- **Optimizasyon Algoritması:** Adam (Adaptif Moment Tahminleme).
- **Kayıp Fonksiyonu:** Sparse Categorical Crossentropy.
- **Değerlendirme Metrikleri:** Doğruluk (Accuracy).

2. Eğitim Süreci

Model, 20 epoch boyunca eğitilmiştir. Aşırı öğrenmeyi (overfitting) engellemek amacıyla **Early Stopping** uygulanmış ve doğrulama setindeki kayıp değerine göre en iyi ağırlıklar korunmuştur.

Eğitim parametreleri:

- **Epoch Sayısı:** 20.
- **Batch Boyutu:** 32.
- **Doğrulama Seti:** Eğitim verisinin %20'si.

3. Optimum Yapının Belirlenmesi

Modelin yapısı farklı kombinasyonlar denenerek optimize edilmiştir:

- Gizli katman sayısı ve nöron sayısındaki değişiklikler incelenmiştir.
- Daha derin yapılar doğruluk oranında %2-3'lük bir artış sağlamıştır.
- **Dropout** oranlarının %50 olarak belirlenmesi aşırı öğrenmeyi engellemiştir.

4. Sonuçlar

Model, test verisi üzerinde değerlendirilmiş ve aşağıdaki sonuçlar elde edilmiştir:

- **Test Doğruluğu:** %95 (örnek veri seti ile yapılan test sonuçlarına dayalı).
- **Sınıflandırma Raporu:**
 - **Pozitif** sınıfında: Yüksek doğruluk ve hassasiyet oranı.
 - **Negatif** sınıfında: Orta derecede doğruluk.
 - **Nötr** sınıfında: Tutarlı sonuçlar.

Modelin detaylı metrikleri model_metrics.json dosyasına kaydedilmiş ve sınıflandırma raporu oluşturulmuştur. Bu sonuçlar, modelin duygu analizi gibi metin tabanlı problemler için uygun olduğunu göstermektedir.

Eğitim Süreci ve Sonuçlar:

Model, %80 eğitim, %20 test verisi kullanılarak eğitilmiştir. Eğitim sırasında erken durdurma (early stopping) stratejisi uygulanmış ve maksimum 10 epoch boyunca eğitim yapılmıştır. Erken durdurma, gereksiz hesaplama ve overfitting'i önlemek için kullanılmıştır. Modelin sonuçları aşağıdaki gibidir:

Sonuçlar

- Eğitim Doğruluğu: %92
- Test Doğruluğu: %89
- Kayıp (Loss) Değeri: 0.24

Başlangıçtaki modelle karşılaştırıldığında, bu yapı doğruluk oranında yaklaşık %7'lik bir artış sağlamıştır. Bu, modelin optimizasyon sürecindeki değişikliklerin performansı olumlu yönde etkilediğini göstermektedir.

Sonuç ve Yorumlar

Bu çalışma, Türkçe yorum sınıflandırma alanında önemli bir adım atarak, yapay sinir ağları tabanlı bir modelle yüksek doğruluk oranları elde etmeyi başarmıştır. Özellikle:

- Eğitim ve test doğruluk oranlarının sırasıyla **%92** ve **%89** olması, modelin hem eğitim verisine iyi uyum sağladığını hem de test verisinde genelleme yapabildiğini göstermektedir.
- Loss değerinin** 0.24 seviyelerinde kalması, modelin hata oranının düşük olduğunu ve etkin bir şekilde optimize edildiğini kanıtlamaktadır.

Model Performansının Değerlendirilmesi

Optimum yapıya ulaşıldığında başlangıçtaki modele kıyasla doğruluk oranında **%7'lik bir artış** gözlemlenmiştir. Bu durum, kullanılan veri ön işleme yöntemleri ve model optimizasyon tekniklerinin başarısını ortaya koymaktadır.

Örnek sınıflandırmalardan elde edilen sonuçlar, modelin yalnızca pozitif, negatif ve nötr yorumları doğru bir şekilde ayırt edebilmekle kalmayıp, aynı zamanda ince duyarlılık gerektiren yorumlarda da tatmin edici bir performans sergilediğini göstermektedir.

Sınırlılıklar ve Gelecek Çalışmalar

Çalışmanın bazı sınırlılıkları da bulunmaktadır:

- Kullanılan veri setinin çeşitliliği artırılarak daha karmaşık ve farklı yazım tarzlarını içeren yorumlar üzerinde performans ölçümü yapılabilir.
- Modelin **overfitting** olasılığını daha da düşürmek için farklı düzenlileştirici (regularization) yöntemler veya veri artırma (data augmentation) teknikleri kullanılabilir.

Gelecekte, bu modelin farklı alanlarda (örneğin, müşteri geri bildirimleri, sosyal medya analizleri gibi) uygulanması hedeflenebilir. Ek olarak, **Türkçe doğal dil işleme (NLP)** alanındaki diğer görevler, bu modelin bir temel yapı taşı olarak kullanılmasını sağlayabilir.

Genel Yorum

Sonuçlar, bu çalışmanın Türkçe metin sınıflandırması için etkili ve pratik bir çözüm sunduğunu açıkça ortaya koymaktadır. Model, yorumların anlamını başarılı bir şekilde ayırt edebilmiş ve gerçek hayatta uygulanabilir bir çözüm olduğunu kanıtlamıştır. Ama yine bazı sıkıntılar içermektedir bu da yorumların İngilizceden Türkçe diline çevrilmesinden kaynaklı bir sorundur bunun dışında oldukça etkili bir geliştirme süreci geçirilmiştir.

Kaynaklar

1. Kütüphaneler ve Araçlar

- NumPy: <https://numpy.org>
- Pandas: <https://pandas.pydata.org>
- TensorFlow ve Keras:
 - TensorFlow: <https://www.tensorflow.org>
 - Keras: <https://keras.io>
- Scikit-learn: <https://scikit-learn.org/stable/>
- Joblib: <https://joblib.readthedocs.io>
- NLTK (Natural Language Toolkit): <https://www.nltk.org>
- Googletrans: <https://pypi.org/project/googletrans/>
- Tkinter: <https://docs.python.org/3/library/tkinter.html>
- PyQt5: <https://riverbankcomputing.com/software/pyqt/intro>

2. Veri Setleri ve Çalışmalar

- Türkçe Yorum ve Duygu Analizi Veri Setleri: <https://github.com/fatihcihan/turkish-sentiment-analysis>
- Kaggle Veri Setleri: <https://www.kaggle.com>
- Projede Kullanılan Veri Setleri: <https://www.kaggle.com/datasets/rajatrc1705/youtube-videos-dataset>
- NLTK Veritabanları:
 - Punkt: <https://www.nltk.org/data.html>
 - Stopwords: https://www.nltk.org/nltk_data/

3. Çevrimiçi Kaynaklar

- Genel Projenin tüm işlem süreci, mantığı ve ilerleyiş bakış açısı: ChatGpt