ÜSKÜDAR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY ZEKÂ MÜHENDİSLİĞİ YÜKSEK LİSANS PROGRAMI

YAPAY ZEKÂ UYGULAMALARI

2022-2023

**FİNAL**

ÖĞRENCİ NUMARASI: **xxx**

ÖĞRENCİ ADI-SOYADI: xxx

PROJE: **Bir Mobil Uygulamanın Kullanıcı Yorumları Üzerine, Text Vectorization ve GRU Ağı İle Duygu Analizi**

# **ÖZET**

Günümüzde birçok kurumsal firma, müşteri ve kullanıcıları için özel mobil uygulamaları hazırlamakta ve bunları mobil uygulama mağazalarından (Apple Store, Google Play Store vb) onların kullanımlarına sunmaktadır.

Bu uygulamaların asıl amacı kurumun bazı hizmetlerine erişim amaçlı olsa da diğer önemli bir faydası, kullanıcılar için önemli bir iletişim aracı olmasıdır. Kullanıcılar bu uygulamalar vasıtası ile hem ilgili kurum hakkında hem de uygulamalar hakkında doğrudan yorumlarını yazabilmektedir ve izinleri doğrusunda yayınlamaktadır. Uygulama mağazalarının yorumlar için sağladığı diğer bir imkân ise bu yorumları, çeşitli şekillerde puanlayabilme imkanıdır. Genelde 1-5 arası puan verebilmektedirler.

Bu projemizde amacımız, ülkemizin önemli bir Telekom operatörüne ait, Mobil Hizmet Kanalı uygulamasının uygulama mağazalarındaki yorum ve puanlarının kullanılarak duygu analizi yapmaktır. Bu sayede kullanıcıların yorumlarının olumlu/olumsuz olarak değerlendirmesi sağlanacaktır.

Çalışmada kullanılan yöntem, yorum verilerinin noktalama ve her tür “a-z, 0-9” dışındaki karakterlerden ayrıştırılması, Türkçe diline uygun olarak kelimelere bölünmesi ve bir kelime sözlüğünün oluşturulması, ardından bir GRU yapay sinir ağı kurularak modelin burada eğitilmesi olarak belirlenmiştir. Eğitilen model test verisi ile değerlendirilmiş ve doğrulanmıştır.

Bu model geliştirildikten sonra yapılan değerlendirmeler göstermiştik ki model yorumları yüksek bir başarı ile olumlu/olumsuz olarak sınıflandırmıştır. Daha sonraki çalışmalarla eğitim verisinin çeşitlendirilmesi ve oluşturulan kelime sözlüğünün zenginleştirilmesi sağlanırsa, bu başarının daha da arttırılabileceği görülmüştür.

**Anahtar kelimeler:** yapay zekâ, duygu analizi, NLP, GRU, Text Vector

İÇİNDEKİLER

[**ÖZET** 1](#_Toc137057085)

[1.GİRİŞ 3](#_Toc137057086)

[PROBLEMİN TANIMI 3](#_Toc137057087)

[2. MATERYALLER VE YÖNTEMLER 4](#_Toc137057088)

[KULLANILAN VERİ 4](#_Toc137057089)

[KULLANILAN YÖNTEM 4](#_Toc137057090)

[Text Preprocessing 5](#_Toc137057091)

[Text Normalization & Word Embedding 6](#_Toc137057092)

[Build Model 6](#_Toc137057093)

[Initilalise & Train Model 7](#_Toc137057094)

[Test & Evaluate Model 8](#_Toc137057095)

[Using Model 8](#_Toc137057096)

[3. ARAÇLAR 9](#_Toc137057097)

[4. KODLAR 9](#_Toc137057098)

[5. SONUÇLAR 9](#_Toc137057099)

[7. KAYNAKLAR 10](#_Toc137057100)

[8. KONTROL LİSTESİ 10](#_Toc137057101)

[EK-A – Değişik eğitim parametreleri ile alınan sonuçlar 11](#_Toc137057102)

[VARSAYILAN PARAMETRELER 11](#_Toc137057103)

[EMBEDDING SIZE ETKİSİ 12](#_Toc137057104)

# 1.GİRİŞ

Günümüzde birçok kurum, üretim ve hizmetlerine ait geri bildirimler almak için yatırımlar yapmaktadırlar. Bu yatırımların bir kısmı kurumun kendi bünyesindeki hizmet kanallarına olmaktadır ancak sosyal medya ve dijital eko sistemin gelişmesi ile birlikte, müşteriler ve kullanıcılar geri bildirimlerini çok farklı mecralara taşımış durumdadır. Dolayısı ile kurumlar, dijital eko sistemlerinde yer alan bütün mecralardaki verileri toplayarak, ürün ve hizmetlerine yön vermektedir.

Birçok kurumsal firma, müşteri ve kullanıcıları için özel mobil uygulamaları hazırlamakta ve bunları mobil uygulama mağazalarından (Apple Store, Google Play Store vb) onların kullanımlarına sunmaktadır.

Bu uygulamaların asıl amacı kurumun bazı hizmetlerine erişim amaçlı olsa da diğer önemli bir faydası, kullanıcılar için önemli bir iletişim aracı olmasıdır. Kullanıcılar bu uygulamalar vasıtası ile hem ilgili kurum hakkında hem de uygulamalar hakkında doğrudan yorumlarını yazabilmektedir ve izinleri doğrusunda yayınlamaktadır. Uygulama mağazalarının yorumlar için sağladığı diğer bir imkân ise bu yorumları, çeşitli şekillerde puanlayabilme imkanıdır. Genelde 1-5 arası puan verebilmektedirler.

Veri toplanan mecralar çok değişik başka puanlama ve şikâyet/öneri/teşekkür sistemi geliştirmiş olsalar da sonuçta bütün bu veri, veri ön işleme adımlarından geçirilerek modelde kullanılacak basit ve temiz “yorum” ifadelerine dönüştürülebilir.

Özetle uygulanacak yöntem, verinin toplandığı kaynak sistemden bağımsız olacaktır.

Bu projemizde amacımız, ülkemizin önemli bir Telekom operatörüne ait, Mobil Hizmet Kanalı uygulamasının uygulama mağazalarındaki yorum ve puanlarının kullanılarak otomatik duygu analizi yapmaktır. Bu sayede kullanıcıların yorumlarının olumlu/olumsuz olarak değerlendirmesi sağlanacaktır.

## PROBLEMİN TANIMI

Her ne kadar bu puanlama ve yorumlar, kullanıcı memnuniyeti hakkında bir fikir verse de yorumlar için kullanılan konuşma dili, yazma dili ve bunlardaki bozukluklar ve formel olmayan ifadeler, durumun doğru yorumlanmasını zorlaştırmaktadır. Birçok durumda yapılan yorum ile varsa puanlama sistemi arasındaki tutarsızlıklar da sorun yaratmaktadır. Çok geniş bir dijital eko sistemden alınan raporların insan eli ve gözü ile değerlendirilmesi ise verimli olmaktan çok uzaktır.

Geliştirilen model ve otomatik çalışma sayesinde kullanıcı yorumları üzerindeki insan faktöründen doğan değerlendirme hataları ve verimsizliği giderilmiş olacaktır. Kurumlar dijital eko sistemlerinden topladıkları birçok değerlendirme yorumlarını bu modeli kullanarak tek bir yatırımla sağlıklı geri bildirimlere dönüştürebileceklerdir.

# 2. MATERYALLER VE YÖNTEMLER

## KULLANILAN VERİ

Proje için ülkemizdeki büyük bir Telekom kurumunun, Mobil Online İşlem merkezi uygulamasının, Apple Store ve Google Play Store üzerindeki kullanıcı yorumları verisi kullanılmıştır.

2020 yılından itibaren rastgele seçilmiş yaklaşık 19.920 Türkçe dilindeki yorum kullanılmıştır.

Her iki uygulama mağazasındaki 5 yıldız üzerinden yapılmış yorumlar, olumlu/olumsuz şeklinde indirgemek için;

* 1,2 olanlar 0-Olumsuz,
* 3,4,5 olanları 1-Olumlu

Olarak dönüştürülmüştür.

Sonuç olarak bütün veri aşağıdaki yapıda bir CSV dosyası olarak hazırlanmıştır.

TTMOIM\_Review.csv dosyası

Rating : Uygulama mağazasındaki yıldız puanlama sisteminin 0/1 olarak indirgenmiş hali, integer(1)

Review : Kullanıcının yaptığı yorumun metin hali, text(1000)

Örnek:

|  |  |
| --- | --- |
| **Rating** | **Review** |
| 1 | Çok güzel bir uygulama |
| 0 | Çok kötü |

Veri setinin içinde toplam 19.922 yorum bulunmaktadır. Bunların 1847 adeti olumsuz yorum olarak etiketlenmiştir.

## KULLANILAN YÖNTEM

Son yıllarda yapay zeka konusundaki gelişmeleri takiben doğal dil işleme tabanlı geliştirilen yapay zeka uygulamalarının önemi ve sayısı gün geçtikçe artmaktadır. Bu konuda geliştirilen uygulamalar birçok işte insan gücünü ortadan kaldırarak daha hızlı ve daha doğru çalışan alt yapılar kullanmamıza olanak sağlayacaktır. Doğal dil işleme ile geliştirilen yaygın örnek uygulamalar ise şu alanlarda karşımıza çıkmakta.

* Metinleri Sınıflandırma (Spam Detector, Haber Konularının Tespiti)
* Sentiment Analizi (Duygu analizi)
* Yazar Tespiti / Yazar Tanıma
* Makine Çevirisi (Google Translate, Amazon Translate..)
* Chatbot

Duygu analizi (Sentiment analizi) doğal dil işleme teknikleri ile geliştirilen ve karşımıza en sık çıkan problemlerden bir tanesidir. Yazılan bir metnin hangi duygu durumu ile yazıldığına karar vermek için modeller geliştirmemize olanak sağlamaktadır.

Duygu analizi yapabilmek için en genel anlamda aşağıdaki süreç işletilecektir.

## Text Preprocessing

Bu ilk aşamada topladığımız veri setinin içinden aşağıdaki işlemler ile temizlenmesi ve düzenlenmesi sağlanmıştır.

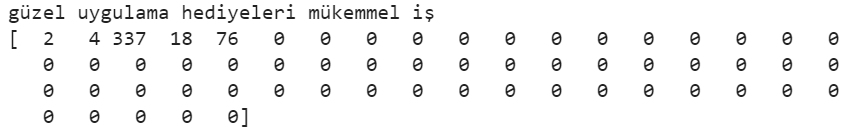
* Veri seti içindeki Türkçe olmayan yorumlar çıkarılmıştır.
* Çok bilinen Türkçe karakter uyumsuzlukları giderilmiştir (ç,ş,ı,ü,ö,ğ, İ) karakterleri yerine basılan ? problemi ancak belirli ölçüde düzenlenmiştir)
* 1000 karakterden daha uzun ve bir anlamı olmayan yorumlar temizlenmiştir. (“\*aldjlksjflksj” gibi ifadeler)
* Puanlama sisteminde 1-5 arasında olmayan değerlere yönelik yorumlar çıkarılmıştır.
* 1-5 arasındaki puanlar “ [1,2] > 0 Olumsuz , [3,4,5]>1 Olumlu “ olacak şekilde yeniden yorumlanmıştır.
* Bütün yorumlar küçük harfe dönüştürülmüştür
* Bütün yorumlar içindeki noktalama işaretleri ve [a-z][0-9] arasındaki karakterlerin dışındakiler PYTHON-REGEX-REPLACE ile kaldırılmıştır.
* Hazırlanan veri setinin %20’si test olarak geri kalanı eğitim verisi olarak belirlenmiştir.

## Text Normalization & Word Embedding

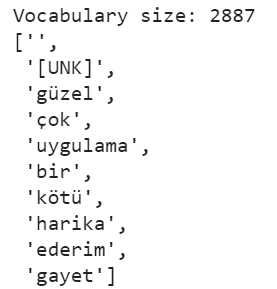
Bu aşamada amaç, veri setinden faydalanarak bir kelime sözlüğü oluşturmak ve bu kelimeleri makinaların anlayabileceği ve üzerinde bir dil modeli oluşturabileceği bir sayısal vektöre dönüştürebilmektir. Kelime temsil (word embedding)’i yapay sinir ağları, olasılık modelleri vb. gibi yöntemler kullanılarak üretilebilir. Bu aşama için Tensorflow.keras kütüphanesine ait **TextVectorization** kullanılmıştır. Geliştirici firma yeni uygulamalarda, çok yaygın olarak kullanılan bir önceki versiyona ait keras.preprocessing. tokenizer kütüphanesini artık tavsiye etmemektedir.

Bu aşamada aşağıdaki adımlar uygulanmıştır.

* Veri setindeki yorumların uzunlukları bulunmuş ve ortalama bir yaklaşımla en uygun kelime uzunluk vektörü (token length) tespit edilmiştir. Bunun için ortalama yorum uzunluğu değerine, standart sapmanın iki katı eklenmiştir.
* En sık tekrar eden kelime listesinin büyüklüğü 10.000 kelime olarak belirlenmiştir.
* İşlem sonucunda her biri 59 uzunluğunda olan TOKEN vektörü elde edilmiştir. Eğer daha küçük ise sonuna 0 ile tamamlanmıştır.



* Yine bu işlem sonucunda en sık kullanılanlar başta yer almak üzere bir kelime sözlüğü de oluşmuştur.



## Build Model

Bu aşamada İşlenmiş ve kelime temsil vektörlerine dönüşmüş veri seti kullanılarak bir yapay sinir ağı sayesinde dil modeli oluşturacağız. Böyle bir dil modeli oluşturmak için genelde iki yaklaşım bulunmaktadır.

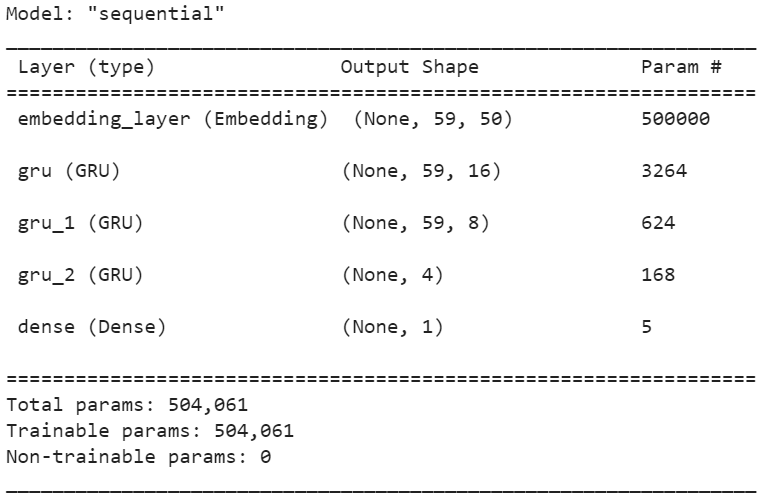
* Sözlük tabanlı yaklaşımlar
* Makine Öğrenmesi – ML (Machine Learning) tabanlı yaklaşımlar

Yapay zekâ alanındaki gelişmelerle makine öğrenmesi ile oluşturulan modellerin daha başarılı sonuçlar verdiği görülmüştür. Projemizde, önceden etiketlenmiş veri ile “**Denetimli Makine Öğrenmesi”** yöntemlerini kullandık.

Makine Öğrenmesi için oluşturduğumuz tercih ettiğimiz RNN Yapay sinir ağı modeli olacak, ancak yinelemeden doğan Vanishing Gradient problemini aşmak için **GRU (Gated Recurrent Units)** ağ mimarisini tercih ettik.

Ağın mimarisindeki algoritma olarak ise **Embedding** parametresini kullandık**.**

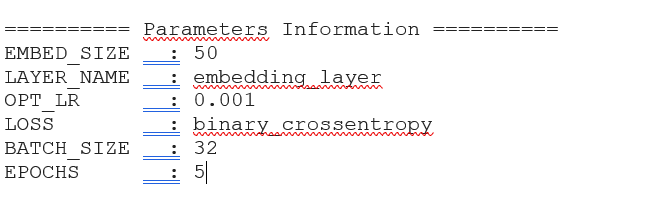
Oluşturduğumuz ağın mimarisi aşağıdaki gibidir.



Mimari de görüldüğü üzere giriş katmanın boyutu bizim tercihimize bağlı olarak token boyutu olan 59 ve giriş parametresi olarak verilen 50 değeri ile 59x50 boyutlu çıktı üretmiştir. Bu değerler daha sonra GRU katmanlarına girdi olarak verilmiş ve son katmana ‘dense’ iletilmiştir. Bu son katmanda sigmoid fonksiyonu ile 1/0 kararı oluşturulmuştur.

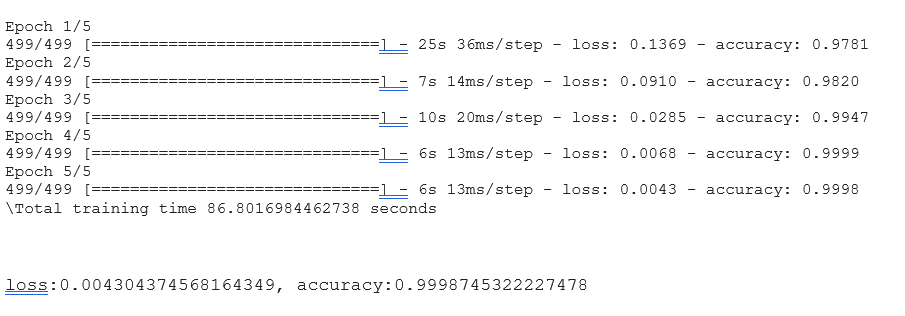
## Initilalise & Train Model

Oluşturduğumuz modelin eğitimi için belirlediğimiz parametreler aşağıdaki şekilde belirlenmiştir.



Modelin eğitiminde takip edilecek metrikler “accuracy” olarak belirlenmiş ve 5 epoch ile 0.999 değerine ulaşılmıştır. Optimizer mode olarak “Adam” ve Learning Rate olarak default değer 0.001 kullanılmıştır.

Eğitim sonuçları ve süresi aşağıdaki gibidir.

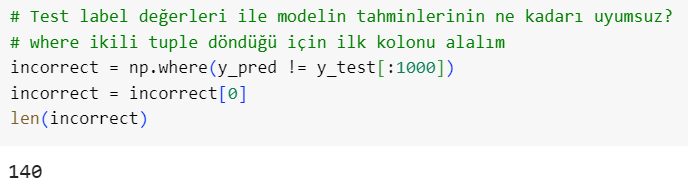


## Test & Evaluate Model

Daha önce veri setimizin içinden %20 olacak şekilde bir test verisi ayırmıştık. Bu aşamada test verisi modele verilerek “adapt” fonksiyonu ile “fit” edilmiş ve ardından etiketlenmiş veri ile “evaluate” edilmiştir. Bu veri için 0.89 doğruluk elde edilmiştir.

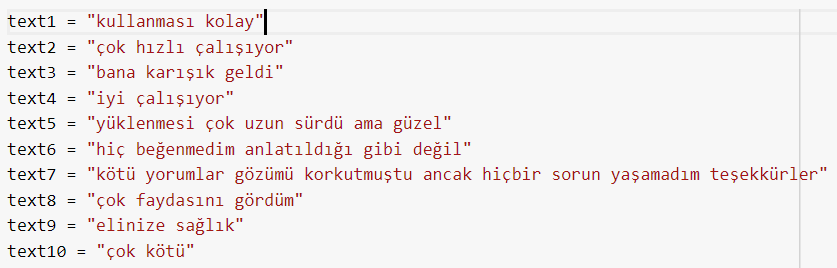


Test verisi içinden seçilen 1000 örnek ile tahmin yapılmış ve bu tahmin sonucunda 140 örneğin etkilenmiş sonuç verisi ile farklı olduğu görülmüştür.



## Using Model

Daha önce hiç kullanılmayan bir veri seti üretmek amacı ile 10 adet yorum ifadesi manuel belirlenmiştir.



Bu veri modele uyarlanarak sonuçları tahmin etmesi istendiğinde aşağıdaki sonuçlar elde edilmiştir.



Bu sonuçları tek tek kontrol edersek sadece son yorumda modelin hatalı olarak olumsuz bir ifadeyi olumlu olarak değerlendirdiğini söyleyebiliriz.

# 3. ARAÇLAR

Proje boyunca kullanılan araçlar:

* Bütün kodlar **Python** dilinde yazılmıştır.
* Text Vektörü elde etmek için *Tensorflow. keras.layers.TextVectorization*
* Model için keras.models ve yapay sinir ağı için *keras.layers (Embedding, GRU ve Dense)* kütüphaneleri kullanılmıştır.
* Geliştirme ortamı olarak **Google Colab** ortamı tercih edilmiştir. Eğitim aşamasında GPU tercihi imkanlar doğrultusunda kullanılabilmiştir. Bazı denemeler CPU bazıları GPU ile yapılmıştır.

# 4. KODLAR

Ekli zip dosyası içinde *“TTMOIM\_sentanalysis.ipynb”* olarak ilave edilmiştir. Aynı sıkıştırılmış dosya içinde veri seti de csv uzantılı olarak bulunmaktadır.

# 5. SONUÇLAR

* Duygu analizi amacı ile yapılan çalışmalarda makine öğrenmesi ve yapay sinir ağı mimarilerinin çok iyi sonuçlar verdiğini söyleyebiliriz.
* Model eğitimi %99 doğruluk ile eğitimi tamamlandıktan sonra sağlıklı tahminlerde bulunabilmektedir. Yeni oluşturduğumuz 10 yorum tahmininde sadece 1 yorumda hata görülmüştür.
* 5 Epoch ile %99 doğruluk yakalandığı için daha fazla arttırmaya ihtiyaç duyulmamıştır.
* learning rate ve batch\_size değişiklikleri model üzerinde çok etkili değişikliklere yol açmamıştır. Örneğin LR=0.001 yerine LR=0.0005 olması veya batch\_size 32 yerine 256.
* Benzer şekilde EMBEDDING\_SIZE=50 ile 100 arasında eğitimde kullanılan parametre sayısı 1mio olsa bile sonuçlar üzerinde etkisi olmamıştır.

Bu denemelere ait sonuçların çıktı dosyaları, raporun EK-A başlığında sunulmuştur.

# 7. KAYNAKLAR

[1] Özekes, Serhat, Ph.D., Üsküdar Üniversitesi, Veri Madenciliğinde İleri Konular Yüksek lisans Ders Notları, 2023.

# 8. KONTROL LİSTESİ

|  |  |
| --- | --- |
|  | EVET /HAYIR |
| Raporunuzu şablonda belirtildiği gibi hazırladınız mı? | EVET |
| Çalışmanızın sonuçlarını (print screen) rapora eklediniz mi? | EVET |
| Rapor dosyanızı şablondaki gibi yeniden adlandırdınız mı? | EVET |
| Raporu sisteme yüklediniz mi? | EVET |
| Kodları sisteme yüklediniz mi? | EVET |

Öğrenilen bilgiler ile yeni fikirler üretmeye araştırmaya ve tartışmaya olanak sağlayan esnek ve serbest final projeleri için teşekkürlerimi sunarım.

# EK-A – Değişik eğitim parametreleri ile alınan sonuçlar

## VARSAYILAN PARAMETRELER

========== Parameters Information ==========

EMBED\_SIZE : 50

LAYER\_NAME : embedding\_layer

OPT\_LR : 0.001

LOSS : binary\_crossentropy

BATCH\_SIZE : 32

EPOCHS : 5

Model: "sequential\_6"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

embedding\_layer (Embedding) (None, 63, 50) 500000

gru\_15 (GRU) (None, 63, 16) 3264

gru\_16 (GRU) (None, 63, 8) 624

gru\_17 (GRU) (None, 4) 168

dense\_5 (Dense) (None, 1) 5

=================================================================

Total params: 504,061

Trainable params: 504,061

Non-trainable params: 0

Epoch 1/5

499/499 [==============================] - 25s 36ms/step - loss: 0.1369 - accuracy: 0.9781

Epoch 2/5

499/499 [==============================] - 7s 14ms/step - loss: 0.0910 - accuracy: 0.9820

Epoch 3/5

499/499 [==============================] - 10s 20ms/step - loss: 0.0285 - accuracy: 0.9947

Epoch 4/5

499/499 [==============================] - 6s 13ms/step - loss: 0.0068 - accuracy: 0.9999

Epoch 5/5

499/499 [==============================] - 6s 13ms/step - loss: 0.0043 - accuracy: 0.9998

\Total training time 86.8016984462738 seconds

loss:0.004304374568164349, accuracy:0.9998745322227478

| **Review** |
| --- |
| **1** | kullanması kolay |
| **1** | çok hızlı çalışıyor |
| **1** | bana karışık geldi |
| **1** | iyi çalışıyor |
| **1** | yüklenmesi çok uzun sürdü ama güzel |
| **0** | hiç beğenmedim anlatıldığı gibi değil |
| **1** | kötü yorumlar gözümü korkutmuştu ancak hiçbir ... |
| **1** | çok faydasını gördüm |
| **1** | elinize sağlık |
| **1** | çok kötü |

### EMBEDDING SIZE ETKİSİ

========== Parameters Information ==========

EMBED\_SIZE : 100

LAYER\_NAME : embedding\_layer

OPT\_LR : 0.001

LOSS : binary\_crossentropy

BATCH\_SIZE : 32

EPOCHS : 5

Model: "sequential\_8"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

embedding\_layer (Embedding) (None, 63, 100) 1000000

gru\_21 (GRU) (None, 63, 16) 5664

gru\_22 (GRU) (None, 63, 8) 624

gru\_23 (GRU) (None, 4) 168

dense\_7 (Dense) (None, 1) 5

=================================================================

Total params: 1,006,461

Trainable params: 1,006,461

Non-trainable params: 0

Epoch 1/5

499/499 [==============================] - 25s 36ms/step - loss: 0.1433 - accuracy: 0.9801

Epoch 2/5

499/499 [==============================] - 9s 19ms/step - loss: 0.0912 - accuracy: 0.9820

Epoch 3/5

499/499 [==============================] - 7s 14ms/step - loss: 0.0278 - accuracy: 0.9946

Epoch 4/5

499/499 [==============================] - 8s 15ms/step - loss: 0.0073 - accuracy: 0.9999

Epoch 5/5

499/499 [==============================] - 6s 12ms/step - loss: 0.0042 - accuracy: 1.0000

\Total training time 87.2057535648346 seconds

Review

1 kullanması kolay

1 çok hızlı çalışıyor

0 bana karışık geldi

1 iyi çalışıyor

1 yüklenmesi çok uzun sürdü,ama güzel

0 hiç beğenmedim anlatıldığı gibi değil

1 kötü yorumlar gözümü korkutmuştu ancak hiçbir ...

1 çok faydasını gördüm

1 elinize sağlık

1 **çok kötü**