TÜRKİYE CUMHURİYETİ
YILDIZ TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

TÜRKÇE İÇİN AYNI ANLAMLI CÜMLE ÜRETİMİ

19011613 — Metin BİNBİR
19011620 — Sercan AKSOY

BİLGİSAYAR PROJESİ

Danışman
Doç. Dr. Mehmet Fatih AMASYALI

Haziran, 2021

TEŞEKKÜR

Daha önce yapay sinir ağları ile çalışma yapmadığımızı bilmesine rağmen çalışma şevkimizi değerlendirerek bize güvenen, bizi tamamlamaya götürecek şekilde yol gösteren ve süreç boyunca yardımlarını hiç esirgemeyen Doç. Dr. Mehmet Fatih Amasyalı'ya, projenin gidişatından her zaman haberi olan ve test aşamasında etiketlemeye yardım eden Alaaddin Göktuğ Ayar, Göktuğ Köksal ve Selcan Yağmur Atak'a, yapay sinir ağları araştırmalarında kaynak gösterdiği için üniversite arkadaşlarımız Aslı Şeyda Özdemir ve Şafak Bilici'ye en içten şekilde teşekkür etmeyi kendimize borç biliriz.

Metin BİNBİR
Sercan AKSOY

İÇİNDEKİLER

KISALTMA LİSTESİ

BLEU	Bilingual Evaluation Understudy Score
BoW	Bag of Words
C4	Colossal Clean Crawled Corpus
DDİ	Doğal Dil İşleme
MC4	Multi-lingual Colossal Clean Crawled Corpus
MT5	Multi-lingual Text-to-Text Transfer Transformer
QQP	Quora Question Pairs
RNN	Recurrent Neural Network
RVAE	Recurrent Variational Autoencoder
Seq2Seq	Sequence to Sequence
T5	Text-to-Text Transfer Transformer

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

TÜRKÇE İÇİN AYNI ANLAMLI CÜMLE ÜRETİMİ

Metin BİNBİR
Sercan AKSOY

Bilgisayar Mühendisliği Bölümü
Bilgisayar Projesi

Danışman: Doç. Dr. Mehmet Fatih AMASYALI

Projenin konusu olan Türkçe için eş anlamlı cümle üretimi, genel olarak paragraf saflaştırma ve özetlemede, makine çevirisinde ve akademik makalelerde cümle tekrarını azaltmak, akla gelmeyen eş anlamlı sözcükler ile paragrafları tamamlamakta ve bu alandaki veri setlerini genişletmede kullanılmıştır.

Gelişen teknolojiler neticesinde insanlar makinelerle anlaşabilmek adına dillerini onlara öğretmeye başlamıştır. Belirlenen işlerde gerekli veri setleri özellikle Türkçe için eksik kalmıştır. Ana dilimiz olan Türkçe kullanılarak makineler ile iletişim kurabilmek ve belirlenen görevlerin gerçeklenmesi adına proje kapsamında dil modellemeler, modellerin test edilmesi ve neticesinde eş anlamlı cümle üretimi yapılmıştır.

Elde edilen modelin kullanıcı tarafından verilen Türkçe bir cümleyi herhangi bir format beklemeksizin, anlamlandırması ve biçimsel benzerlikten mümkün olduğunca uzak durarak anlamsal bütünlüğü koruması beklenmiştir. Bu amaç uğrunda eğitilen model cümle içerisindeki ögelerin yerlerini değiştirebilir, bazı sözcükler yerine eş anlamlılarını getirebilir hatta özneler yerine zamirler kullanabilir şekilde veri seti hazırlanmıştır.

Anahtar Kelimeler: Doğal Dil İşleme, Derin Öğrenme, Aynı Anlamlı Cümle Üretimi, Makine çevirisi, BoW, Seq2Seq, T5, MT5, Transformer, BLEU, C4.

ABSTRACT

PARAPHRASE GENERATION FOR THE LANGUAGE TURKISH

Metin BİNBİR
Sercan AKSOY

Department of Computer Engineering
Computer Project

Advisor: Assoc. Prof. Dr. Mehmet Fatih AMASYALI

The main purpose of the project is generating paraphrased sentences in Turkish, which is used for purificating and summarizing text, machine translation and to reduce repetition in academic articles.

As a result of improvements on technologies, people have been trying to communicate with the machines in their language. In specified tasks, there are lots of gaps for the dataset in Turkish language. For the Turkish language, which is our native language, we have augmented new data for the purpose of training a new model, for testing it and also for generating paraphrased sentences.

We are expecting from the obtained model to be able to generate new sentences in Turkish which are semantically same but syntactically different from the input sentences. For this purpose, model may change the order of elements in sentence, may change words with their synonyms and may replace subjects with pronouns. Dataset was prepared for these situations.

Keywords: Natural Language Processing, Deep Learning, Paraphrase Generation, Machine Translation, BoW, Seq2Seq, T5, MT5, Transformer, BLEU, C4.

1
Giriş

Aynı anlamlı cümle üretimi görevi, yeni bir veri seti üretilmesinde ya da genişletilmesinde, metin özetlemede, alınan bir cümle için farklı çeviri seçenekleri sunmada kullanılabilir. Aynı anlamlı cümle üretimi çalışması çoğu doğal dil işleme alanında olduğu gibi oldukça büyük veri kümelerine ihtiyaç duyar.

Proje kapsamında, özellikle Türkçe için ihtiyaç duyulan veri kümesine katkıda bulunarak otomatik elde edilmiş veri kümeleri ile aralarındaki farklar değerlendirilmiştir.

Karşılaştırmalar neticesinde, üretilen cümlelerin veri kümeleri arasında boyut farkından ziyade hedef olarak belirlenen cümlelerin niteliklerine bağlı olduğu tespit edilmiş ve ana dili Türkçe olan insanlar tarafından değerlendirmeye tutulmuştur.

Aynı anlama cümle üretimi projesiyle hedeflenenler sıralandığı gibidir; bu görevi Türkçe için en iyi çalıştırabilecek modelleri araştırmak, doğal dil işleme projesinin temeli olan dil modelleme hakkında tecrübe kazanmak, Türkçe için kullanılabilecek veri kümesini geliştirmek, geliştirilen veri kümesinin başarısını ölçmek ve kıyaslamak, muadillerine göre başarılı olduğunu düşünülen aynı anlamlı cümle üretimi görevini bir başka konu için veri kümesi genişletmede kullanabilmek.

Raporun devamında:

- Ön inceleme kısmında başkaları tarafından gerçekleştirilmiş benzer projeleri ve bu çalışmanın katkıları,
- Fizibilite kısmında projenin gerçekleştirilebilmesi için teknik, yasal, ekonomik ve zamansal kısıtlamalar,
- Sistem analizi kısmında projenin gerçekleşmesi için uygun yolları,
- Sistem tasarımı kısmında üç başlıkta incelenmek üzere yazılımın mimarisi ve kullanılan algoritmalar hakkında açıklamalar, veri tabanı özellikleri ve etkileri,

ayrıca projenin çalıştırılmasında girdilerin ve çıktıların nasıl yönetildiği,

- Uygulama kısmında projenin hedefleri doğrultusunda üzerine çalışılmış modellerin bilgileri,
- Deneysel Sonuçlar kısmında eğitilen modelin girdilere karşı ürettiği yeni cümleleri ayrıca uç noktalar ile sınanması,
- Performans Analizi kısmında çalışmanın başarısı ve bunun nasıl ölçüldüğü yer almaktadır.

2
Ön İnceleme

Projenin konusunda örnek çalışmalar araştırıldığında Türkçe dili için yapılmış bir projeye rastlanamamıştır. Bu sebeple mimarileri kıyaslamak için İngilizce dili için yapılmış çalışmalar incelenmiştir. Modelin eğitilebildiği aynı zamanda tarafımızca anlaşılabilen 3 çalışma bulunmuştur. Temel Seq2Seq[1](Sequence to Sequence), BoW[2](Bag of Words) ve T5 (Text-To-Text-Transfer-Transformer) olarak bilinen bu modeller aynı veri seti QQP (Quora Question Pairs) ile eğitilmiş ve yine aynı cümleler ile test edilmiştir. Çıktıların incelenmesi sonucu, en güncel yayımlanmış (2019) çalışma olan T5'in daha başarılı olduğuna karar verilmiş ve bu model ile çalışmaya başlanmıştır.

T5'in Türkçe kullanımını sağlayan MT5(Multi-lingual Text-To-Text-Transfer-Transformer) modeli ile gerçeklenmiş bir proje, çalışmalar esnasında kullanıma sunulmuştur[3]. Bu proje çeşitli girdiler ile sınanmış ve üretilen cümlelerin anlamsal bozukluklar içerdiği, ayrıca referans cümle ile ortak kelime sayısının fazla olduğu tespit edilmiştir. Örnek olarak belirlenen bu çalışmada eğitim kümesi makine çevirisi yöntemi ile hazırlanmış olup Türkçe için uygun olmayan kelimeler içermekteydi. Bu eksikliğin giderilmesi adına danışmanımız tarafından sağlanmış, manuel olarak oluşturulmuş veri seti yine manuel olarak genişletilmiştir. Böylece aynı mimari kullanılsa dahi insan kaynaklı değerlendirmede başarı farkı ortaya çıkarılmıştır.

3
Fizibilite

3.1 Teknik Fizibilite

3.1.1 Yazılım Fizibilitesi

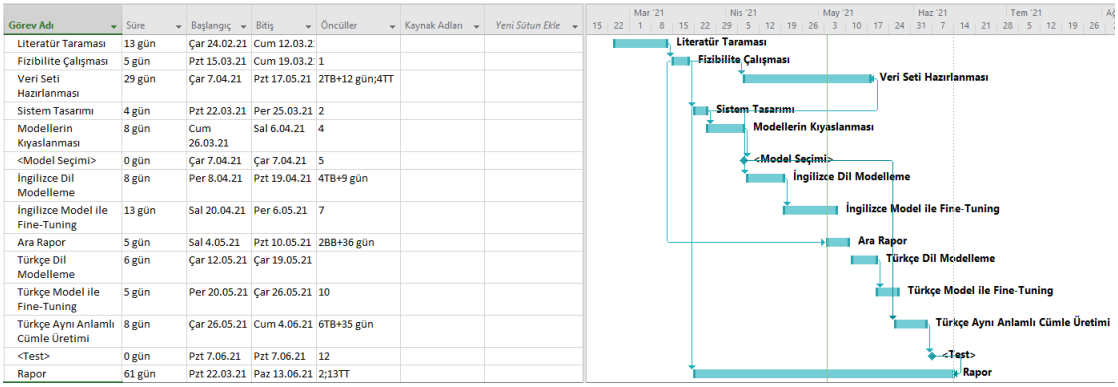
Proje, aynı konuda birçok örnek çalışma bulundurduğu için Python programlama dili ile gerçeklenecektir. Doğal dil işleme alanında kullanılan en büyük platformlardan biri olan Hugging Face aracılığıyla eğitilmiş T5[4] modelleri üzerinden İngilizce dilinde testler yapılması beklenmektedir. Transformers[5] kütüphanesi içinden yine eğitilmiş T5Tokenizer kullanılması beklenmektedir.

Türkçe aynı anlamlı cümle üretimi için ise T5'in Türkçe uyumlu versiyonu olan MT5[6] hazır eğitilmiş dil modelinin mt5-base versiyonu ve parçalıyıcısı olan mt5-base versiyonu kullanılmıştır. Eğitim aşaması için torch[7] ve pytorch-lightning kütüphaneleri kullanılmıştır. Girdilerin işlenmesi amacıyla sentencepiece ve nltk kütüphaneleri kullanılmıştır. Proje başarısının BLEU[8] skoruyla ölçülmesi için de datasets[9] kütüphanesi kullanılmıştır.

3.1.2 Donanım Fizibilitesi

Projenin temeli büyük veri setlerinden faydalanmaktadır ve veri setini işlemek için çok hızlı işlemcilere, ayrıca çok fazla kapasitesi olan hafızalara ihtiyaç duyar. Google Colab bu olanakları belli bir ücret karşılığında sağladığı için internet bağlantısı olan orta düzey bir bilgisayar yeterli olacaktır.

3.2 İş Gücü ve Zaman Fizibilitesi



Görev Adı	Süre	Başlangıç	Bitiş	Öncüller	Kaynak Adları	Yeni Sütun Ekle
Literatür Taraması	13 gün	Çar 24.02.21	Cum 12.03.2			
Fizibilite Çalışması	5 gün	Pzt 15.03.21	Cum 19.03.2	1		
Veri Seti Hazırlanması	29 gün	Çar 7.04.21	Pzt 17.05.21	2TB+12 gün;4TT		
Sistem Tasarımı	4 gün	Pzt 22.03.21	Per 25.03.21	2		
Modellerin Kıyaslanması	8 gün	Cum 26.03.21	Sal 6.04.21	4		
<Model Seçimi>	0 gün	Çar 7.04.21	Çar 7.04.21	5		
İngilizce Dil Modelleme	8 gün	Per 8.04.21	Pzt 19.04.21	4TB+9 gün		
İngilizce Model ile Fine-Tuning	13 gün	Sal 20.04.21	Per 6.05.21	7		
Ara Rapor	5 gün	Sal 4.05.21	Pzt 10.05.21	2BB+36 gün		
Türkçe Dil Modelleme	6 gün	Çar 12.05.21	Çar 19.05.21			
Türkçe Model ile Fine-Tuning	5 gün	Per 20.05.21	Çar 26.05.21	10		
Türkçe Aynı Anlamlı Cümle Üretimi	8 gün	Çar 26.05.21	Cum 4.06.21	6TB+35 gün		
<Test>	0 gün	Pzt 7.06.21	Pzt 7.06.21	12		
Rapor	61 gün	Pzt 22.03.21	Paz 13.06.21	2;13TT		

Şekil 3.1 Proje İş-Zaman Çizelgesi

Projenin gerçeklenebilmesi için kullanılabilecek yöntemleri incelemek kapsamında ilk işlem literatür taramasıdır. Proje hakkında bilgi edinildikten sonra projenin devamlılığına karar verecek olan fizibilite çalışması ile yapılabilirlik test edilir. Fizibilite çalışmasında bu alanda yeterli veri seti bulunmadığı farkedilerek kendi veri setimizi oluşturmak için yeterli süre ataması yapılmalıdır. Çalıştırılacak algoritma seçimi için proje konusunda daha önce yapılmış çalışmalar test edilir, değerlendirme sonuçlarına göre model seçilir. Bu aşamada projenin alt seviyeleri hakkında tecrübe kazanmak için İngilizce dil modelleme üzerine örnekler geliştirilir. Benzer işlemler Türkçe için de uygulanarak proje tamamlanma aşamasına getirilir. Son ana kadar elde edilen bütün çıktılar raporlama aşamasına dahil edilir.

3.3 Yasal Fizibilite

Yapılan proje mevcut yasa ve yönetmeliklere uygun olmakla beraber herhangi bir patent vb. korunmuş hakkı ihlal etmemektedir.

3.4 Ekonomik Fizibilite

- Geliştirme ortamı, JetBrains PyCharm Python IDE'si JetBrains tarafından öğrencilere ücretsiz olarak sağlanmaktadır. Bir diğer ortam olan Google Colab uygulamasının ücretsiz sürümündeki kullanım kotası dolabildiğinden Google Colab Pro için 4 aylık kullanım esnasında 320 TL'lik bir masraf öngörülmektedir.

- Derin öğrenme için lazım olan tüm Python kütüphaneleri ücretsiz olarak kullanılabilmektedir.

- Sahip olduğumuz kişisel bilgisayarlar bazı kütüphanelerin eski versiyonlarında gerekli fonksiyonları çalıştıramasa da projede yeteri düzeyde ilerleme katetme donanımına sahiptir.
- Modelin eğitimi için gereken cümle çiftlerini içeren veri setinin doldurulması için 5 işçiye asgari ücret mukabilinde 1 ay ödeme yapılması uygun görülmüştür.

Görev Tanımı	**Kullanıcı/Adet**	**Birim Fiyat (TL)**	**Toplam Maliyet (TL)**
Google Colab Pro	4	80	320
Microsoft Project Plan 1	1	72	72
JetBrains PyCharm IDE	2	0	0
Lenovo Legion 5	2	9.300	18.600
21.5 Inch HDMI LCD Ekran	1	560	560
20 Inch HDMI LCD Ekran	1	480	480
Google Drive Depolama Alanı	1	40	40
Mühendis Maaşları	2	10.000	20.000
Cümle Yazarları	5	2.825,90	14.129,50
TOPLAM MALİYET			**54.201,50**

Tablo 3.1 Ekonomik Fizibilite Tablosu

4
Sistem Analizi

Proje konusunun gerçekleştirilebilmesi için yöntemler araştırıldığında RVAE[10], temel Seq2Seq, BoW ve T5 modelleri ile yapılmış çalışmalar denenmiştir. RVAE mimarisi ile ilgili çalışmalarda hali hazırda eğitilmiş modeller kullanıldığı için araştırmaya uygun olmayacağı, temel Seq2Seq ve BoW için çalışmaların T5 modeli ile yapılmışların yanında sayılarının az olması, BoW mimarisinin dil modelinden bağımsız gerçeklenmesinin çalışmaya fazladan zorluk katacağı düşünülerek 2019'da yayımlanmış, bulunabilen en güncel model olan T5 mimarisi seçilmiştir.

Proje kapsamında danışman hocamız tarafından yönlendirilmeler neticesinde gereklilikler Hugging Face adlı doğal dil işleme ile ilgili en güncel modellerin ve veri kümelerinin bulunduğu web sitesinden, GitHub adlı açık kaynak kodlu çalışmaların bulunduğu web sitesinden, NLP Yardımlaşma Platformu adlı bu alanda çalışan insanların toplandığı ve yardımlaştığı bir Discord uygulaması grubundan ve Medium sitesindeki bloglardan faydalanılmıştır.

Eğitilen İngilizce dil modelinin başarısını ölçmek için duygu analizi görevindeki tahminin ne kadar isabetli olduğunu ölçmek için F1 skoru, İngilizce aynı anlamlı cümle üretiminin başarısını ölçmek için insan değerlendirmesi kullanılmıştır.

Türkçe aynı anlamlı cümle üretiminin başarısını ölçmek için BLEU[8] skoru ve insan değerlendirmesi kullanılmıştır.

5
Sistem Tasarımı

5.1 Veri Seti Tasarımı

T5 modeli, dil ve görev gözetmeksizin verilen bir metni bir başka metne çevirmeye çalışır. Bu hususta verilen metnin dönüştürülecek metni hangi görevle elde edeceği, tasarlanan veri setindeki ikili cümleler arası ilişkiye bağlıdır.

5.1.1 İngilizce Dil Modelleme

Google tarafından eğitilmiş olan T5 modeli aynı makalede sunulan C4(Colossal Clean Crawled Corpus) veri seti ile eğitilmiştir. Bu veri seti, daha öncesinde kullanılan çoğu veri setinden daha büyük boyutlu (yaklaşık olarak 750 GB) ve kirli veriden daha çok arındırılmış hâldedir, bunun sebepleri [4] ise aşağıda listelenmiştir:

- Sadece belirli noktalama işaretleri (ünlem işareti, soru işareti, nokta ve tırnak işareti) ile biten cümleler seçilmiştir.
- 5'ten az sayıda cümle içeren metinler ve 3'ten az sayıda kelime içeren cümleler elenmiştir.
- "List of Dirty, Naughty, Obscene or Otherwise Bad Words" içerisinde yer alan kelimeler barındıran cümleler veri setine dahil edilmemiştir.
- Javascript kodlarını çalıştırma izni isteyen tüm internet siteleri elenmiştir.
- Lorem Ipsum (Anlamsız Yazılar) içeren bütün sayfalardan uzak durulmuştur.
- Süslü parantez karakterleri ("{" ve "}") program kodunun çalışmasında problem yaratabileceği için bu karakterleri bulunduran tüm sayfalardan kaçınılmıştır.
- İçerik tekrarına düşmemek için veri setinde mevcut olan 3 cümleyi içeren internet sitelerindeki içerikler veri setine eklenmemiştir.

Tarafımızca oluşturulmuş İngilizce dil modeli gerçeklenirken "National Library of Medicine" sitesinde bulunan bazı makalelerin özet kısımları kullanılmıştır. Dil modelleme esnasında veri seti büyüklüğünün öneminin görülebilmesi adına kullanılan veri setleri 1/5 oranında iki parçaya ayrılmıştır.

5.1.2 Duygu Analizi Görevinde İnce Ayar İşlemi

Modellenen dilin mantıklı sonuçlar üretebileceğinden emin olmak için sonucu doğru ya da yanlış olarak verebilecek bir görev belirlenmesi daha sağlıklı gözlemlere olanak tanıyacaktır. Bu yüzden ince ayar işlemi için görev olarak duygu analizi, yani metinleri pozitif veya negatif olarak sınıflandıran bir kod kullanılmıştır. Bu görevde kullanılan veri setinde[11] cümleler, karşılığında pozitif veya negatif olarak etiketlenmiş hâlde bulunmaktadır.

```
Pozitif Cümle Örneği: This is the definite Lars von Trier Movie,
my favorite, I rank it higher than "Breaking the waves" or the
latest "Dancer in the Dark"... I simply love the beauty of the
picture...the framing is so original; acting is wonderful, A
MUST SEE.

Negatif Cümle Örneği: Ned aKelly is such an important story
to Australians but this movie is awful. It's an Australian
story yet it seems like it was set in America. Also Ned was an
Australian yet he has an Irish accent...it is the worst film I
have seen in a long time
```

Şekil 5.1 Duygu Analizi Veri Seti Örneği

5.1.3 Aynı Anlamlı Cümle Üretimi (İngilizce)

Projenin ilk aşamalarında model kıyaslama aşaması için eğitilmiş dil modelleri üzerinde QQP veri setinin bir kısmı (90.000 soru çifti) kullanılmıştır. Bu veri seti, Quora web sayfası üzerinde sorulmuş soruların ikili cümleler hâlinde eşlenmesi ve bu soru cümlelerinin 1 ve 0 (aynı anlamlı ya da aynı anlamlı değil) ile işaretlenmesiyle oluşur [12]. Projenin amacı aynı anlamlı cümle üretimi olduğu için veri setinde sadece 1 ile etiketlenmiş ikili soru cümleleri kullanılmıştır.

Cümle 1	**When should I take BCAAs?**
Cümle 1	When is the best time to take BCAA?
Cümle 2	**What are the qualities of a good mother?**
Cümle 2	What makes a good mum?
Cümle 3	**How I increase my focus in study?**
Cümle 3	How do we increase concentration?

Tablo 5.1 QQP Veri Setindeki Bazı Örnekler

5.1.4 Aynı Anlamlı Cümle Üretimi (Türkçe)

Türkçe dilinde aynı anlamlı cümle üretiminin ince ayar aşamasında eğitim veri setindeki cümleler txt uzantılı train ve validation dosyalarına Şekil 5.2'deki formatta verilmiştir.

$$< s > G === E < /s >$$

G: İlk cümle
E: İlk cümlenin eş anlamlı cümlesi

```
<s>Amerika'da hangi dili konuşuyorlar? === Amerika'da insanlar hangi dili konuşuyor? </s>
<s>Başka bir gezegenden misin? === Uzaylı mısın? </s>
<s>Perdeyi açabilir miyim? === Perdeyi açmamın bir sakıncası var mı? </s>
<s>Onlara ne sordun? === Onlara sorduğun neydi? </s>
<s>Sana yardım edeceğimi de nereden çıkardın? === Size yardım edeceğimi düşündüren nedir? </s>
<s>Ne öğrenmek istiyorsun? === Ne bilmek istiyorsun? </s>
```

Şekil 5.2 Veri Seti Formatı

Proje kapsamında modellerin başarısı veri setlerinin büyüklükleri ve içeriklerine göre test edildiği için Tablo 5.2'de Türkçe veri setlerinin kısaca açıklamaları verilmiştir.

Veri Seti ID	**Açıklama**	**Cümle Çiftleri Sayısı**
DH-S	Manuel olarak ekleme yapılmış forum sitesi başlıklarındaki soru çiftleri	30994
DH	DH-S veri setinin tarafımızca ekleme yapılmış hâli	50994
TQP	Çok sayıdaki dilde aynı anlama gelen cümlelerin çevirileri	51620
MT	QQP'nin Türkçe çevirisi	40200
BIG	DH, MT ve TQP veri setlerinin birleştirilmiş hâli	142814

Tablo 5.2 Eğitimde Kullanılan Veri Setleri

1. Makine Çevirisiyle Oluşturulmuş Veri Setleri

   Makine çevirisiyle bir cümleyi birden çok dile çevirdikten sonra istenen dile tekrar çeviri yaparak aynı anlamlı cümle üretimi yapılabilmektedir fakat bu yöntem ile her zaman başarılı cümle çiftleri elde edilemeyebilir.

   (a) MT: QQP veri setinde yer alan 40200 adet soru çiftinin makine çevirisiyle Türkçe diline çevrilmiş hâlidir.

   (b) TQP: TaPaCo[13] veri setinde 73 dilde bulunan soru cümlelerinden Türkçe olarak 51620 çift bulunmaktadır.

2. Manuel Oluşturulmuş Veri Seti
   Bir kısmı danışman hocamızca sağlanan ve tarafımızca 20000 adet ekleme yapılan Türkçe veri setinde ise yaklaşık 51000 adet aynı anlamlı soru çifti bulunmaktadır. Sorular ise Türkiye'deki ünlü forum web sitesi olan forum.donanimhaber.com adresindeki konu başlıklarının bir araya getirilmesiyle oluşturulmuştur. Soruların aynı anlama gelen çiftleri elle doldurulmuştur.

   (a) DH-S: Danışman hocamızca sağlanmıştır ve 30994 adet soru çifti içermektedir.

   (b) DH: DH-S veri setinin tarafımızca 20000 adet soru çifti eklenmiş hâlidir.

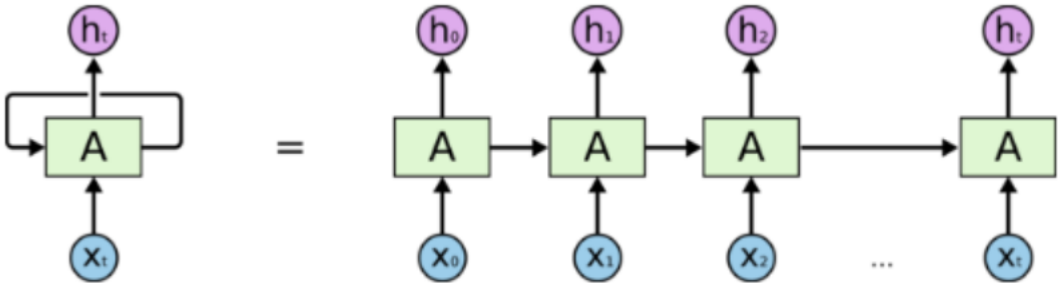
Cümle 1	**Yıldız savaşları izlemeli miyim?**
Cümle 1	Yıldız savaşları filmini seyretmemi önerir misiniz?
Cümle 2	**Vernik seçimi nasıl olmalı?**
Cümle 2	Vernik tercih ederken nelere dikkat edilmelidir?
Cümle 3	**Avrupaya vize kalkarsa gider misiniz?**
Cümle 3	Avrupaya gidiş vizesiz olsa seyahate çıkar mısınız?

Tablo 5.3 Manuel Oluşturulmuş Türkçe Veri Setindeki Bazı Örnekler

5.2 Yazılım Tasarımı

Dil modelleme veya makine çevirisi gibi görevlerde genel olarak Recurrent Neural Networks (RNN) (Şekil 5.3) ve Transformer mimarileri kullanılır. RNN mimarisi, sıralı yapılardan dolayı eğitim esnasında paralelleştirmeye olanak tanımaz, bu yüzden uzun süreli bellek bağımlılıklarını öğrenmek konusunda probleme yol açar. RNN'lerde bulunan bellek sayısının fazla olması, performans olarak daha iyi etki yaratır fakat uzun girdilerin öğrenilmesi sırasında toplu işlem yapmak sınırlandırılmış olur. RNN ile örneğin bir cümledeki son kelimenin hücresine erişmek için kelime kelime ilerlemek gerekmektedir. Yapay sinir ağı uzun bir yapıya sahipse hatırlamak birkaç adım

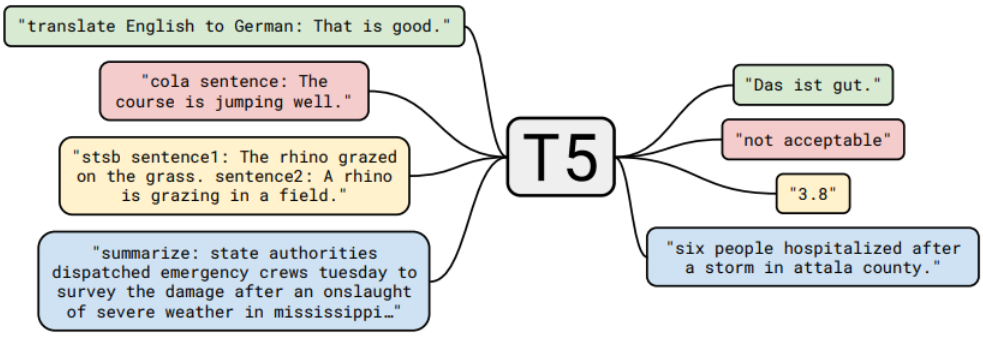
sürebilir, her maskelenmiş durum bir önceki maskelenmiş duruma bağlıdır. Bu sebeple paralelleştirme etki gösteremez. Girdilerin çok uzun olduğu durumlarda, RNN mimarisi uzak konumların içeriklerini art arda unutma eğilimi göstermeye yatkın olabilir.



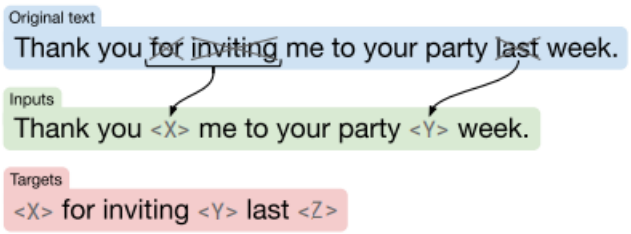
Şekil 5.3 RNN

Transformer mimarisi, tekrarlamayı (recurrent) ortadan kaldırıp bunun yerine girdiler ve çıktılar arasındaki ilişkiyi kurmak için öz dikkat (self attention) mekanizması kullanır. Self attention, bir cümlede yer alan herhangi bir sözcüğün diğer sözcüklerle olan ilişkisini belirler. Kodlayıcı(encoder)-kod çözücü(decoder) prensibine çok benzer olarak çalışmaktadır. Bu prensipten farklı olarak kodlayıcıya gelen girdi word-embedding adı verilen nümerik hâldeki kelime vektörlerine dönüştürülmektedir. Bu vektörler de ilk başta self attention mekanizmasına girer. Örnek olarak "On sene sonra bir ilkokula gittim, orada gördüklerim beni şaşırttı." cümlesinde yer alan her kelimenin, cümledeki diğer kelimelerle olan ilişkisini çıkarıp her ilişkiye bir skor belirlenir. Bu cümlede yer alan "ilkokul" ve "orada" sözcükleri arasında kurulan ilişkinin skorunun, diğer ilişkilere göre daha yüksek bir skora sahip olduğu çıkarımına ulaşılır.

T5 modeli, dil ve görev fark etmeksizin metin halinde verilen girdiyi metin halinde bir çıktıya çevirir [4]. Transformers kütüphanesi, içerisinde farklı boyutlarda veri setleri ile eğitilmiş dil modelleri (t5-base, t5-small, t5-large vb.) bulundurur. Bu modeller, belirli birkaç görevi ön ek olarak kabul edebilmekte ve ince ayar işlemine gerek olmadan çıktı verebilmektedir(Şekil 5.4).
T5 modeli eğitilirken ikili veri setine ihtiyaç duyar, verilen her iki girdi için rastgele ve farklı noktları maskeler (Şekil 5.5). Böylece karşılıklı iki girdide birbirlerinin içerikleri bulunur. Bu sayede girdilerin ikili hâlleri hangi görev için tasarlandıysa bunu gerçeklemesi çok daha kolay olur.



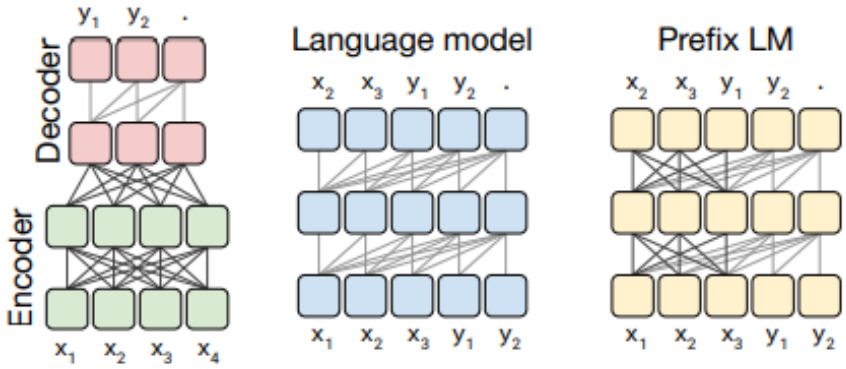
Şekil 5.4 T5 Çalışma Sistemi



Şekil 5.5 T5 Maskelenmiş Cümle Örneği

T5 modeli genel olarak kodlayıcı-kod çözücü prensibi(Şekil 5.6) ile çalışır, girdi olarak verilen cümlelerin parçacıklarına ayrılmasını ise yine Transformers kütüphanesinde bulunan, daha önceden eğitilmiş parçalayıcılar ile yapılabilmektedir. Bu noktada dil modellerinde olduğu gibi farklı boyutlarda kelime hazinesine sahip parçalayıcılar mevcuttur ve kullanılacak görevin içerdiği kelimeleri tanımalarına göre başarı oranları değişir.

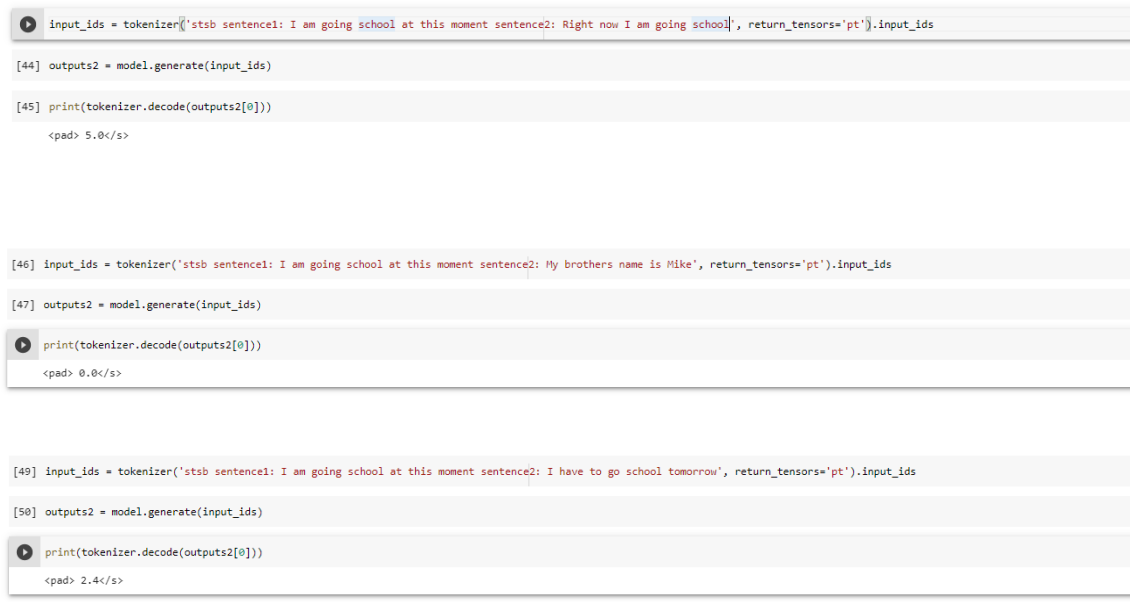
Şekil 5.7, Şekil 5.8, Şekil 5.9 ve Şekil 5.10'de yer alan örneklerden yola çıkarak T5 modelinin nasıl çalıştığının incelenmesinin ardından, tarafımızca İngilizce dil modeli oluşturulmuştur. Ardından bu modele, duygu analizi görevi için ince ayar yapılmış ve skorlar elde edilmiştir. Dil modelleme esnasında kullanılan argümanlar Şekil 5.11'da gösterilmiştir.



Şekil 5.6 T5 Encoder-Decoder

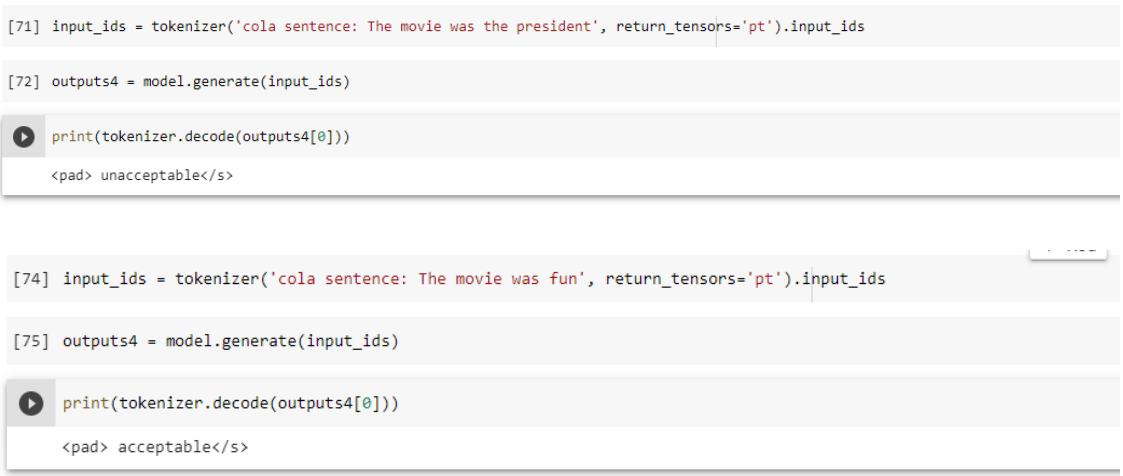
```
!pip install --upgrade transformers

[1] from transformers import T5ForConditionalGeneration,T5Tokenizer

[2] import numpy as np

[3] model =   T5ForConditionalGeneration.from_pretrained('t5-base')

[4] !pip install sentencepiece
    Requirement already satisfied: sentencepiece in /usr/local/lib/python3.7/dist-packages (0.1.95)

[5] tokenizer = T5Tokenizer.from_pretrained('t5-base')

[19] input_ids = tokenizer("translate English to German: The house is wonderful.", return_tensors="pt").input_ids

[20] outputs = model.generate(input_ids)

[21] outputs
    tensor([[    0,   644,  4598,   229, 19250,     5,     1]])

[22] print(tokenizer.decode(outputs[0]))
    <pad> Das Haus ist wunderbar.</s>
```

Şekil 5.7 T5 ile İngilizce-Almanca Çeviri



Şekil 5.8 T5 ile İki Örnek Cümle Arasındaki Anlam Benzerliği Tespiti

Şekil 5.9 T5 ile Metin Özetleme



Şekil 5.10 T5 ile Cümlenin Oluşturulma Mümkünlüğü

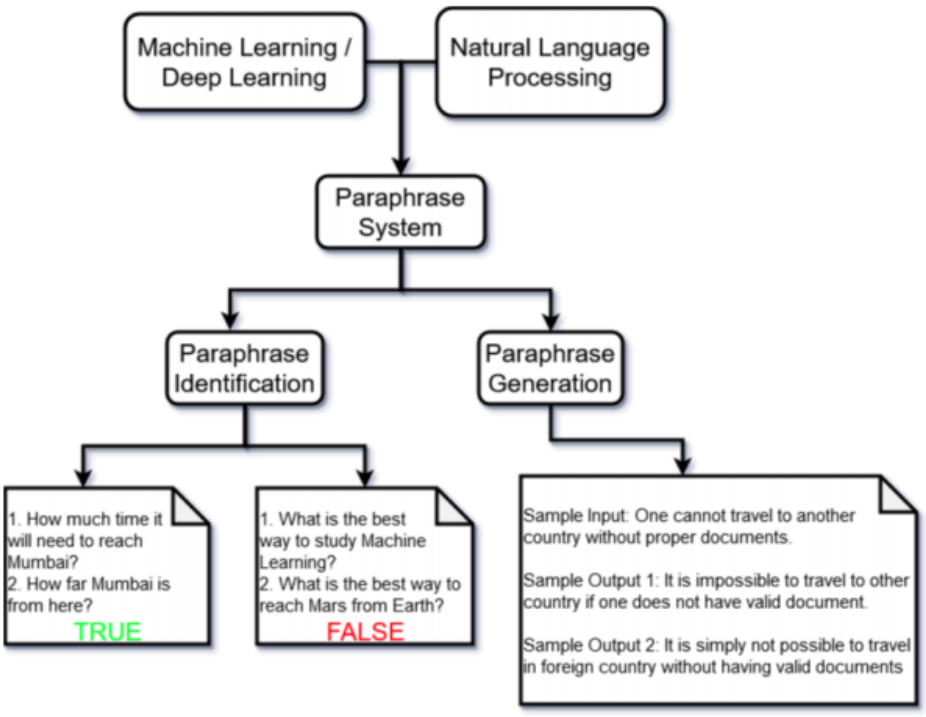
```
args_dict = dict(
    output_dir="", # path to save the checkpoints
    model_name_or_path=hparam.model,
    tokenizer_name_or_path=hparam.model,
    max_input_length=int(hparam.input_length),
    max_output_length=int(hparam.output_length),
    freeze_encoder=False,
    freeze_embeds=False,
    learning_rate=1e-5,
    weight_decay=0.0,
    adam_epsilon=1e-8,
    warmup_steps=0,
    train_batch_size=4,
    eval_batch_size=4,
    num_train_epochs=2,
    gradient_accumulation_steps=1,
    n_gpu=1,
    resume_from_checkpoint=None,
    val_check_interval = 1.0,
    n_val=0,
    val_percent_check= 0,
    n_train=-1,
    n_test=-1,
    early_stop_callback=False,
    fp_16=False, # if you want to enable 16-bit training then install apex and set this to true
    opt_level='O1', # you can find out more on optimisation levels here https://nvidia.github.io/apex/amp.html#opt-levels-and-properties
    max_grad_norm=1.0, # if you enable 16-bit training then set this to a sensible value, 0.5 is a good default
    seed=101,
)
```

Şekil 5.11 T5 Dil Modelleme İçin Kullanılan Argümanlar

Model, eğitilirken ve ince ayar yapılırken aynı optimize etme algoritması olarak "Adam" ve aynı parçalayıcı olarak "T5Tokenizer.from_pretrained('t5-base')" kullanılmıştır. (Şekil 5.12)

```
[ ] args_dict = dict(
        data_dir="", # path for data files
        output_dir="", # path to save the checkpoints
        model_name_or_path='t5-base',
        tokenizer_name_or_path='t5-base',
        max_seq_length=512,
        learning_rate=3e-4,
        weight_decay=0.0,
        adam_epsilon=1e-8,
        warmup_steps=0,
        train_batch_size=8,
        eval_batch_size=8,
        num_train_epochs=2,
        gradient_accumulation_steps=16,
        n_gpu=1,
        early_stop_callback=False,
        fp_16=False, # if you want to enable 16-bit training then install apex and set this to true
        opt_level='O1', # you can find out more on optimisation levels here https://nvidia.github.io/apex/amp.html#opt-levels-and-properties
        max_grad_norm=1.0, # if you enable 16-bit training then set this to a sensible value, 0.5 is a good default
        seed=42,
    )
```

Şekil 5.12 T5 Duygu Analizi İnce Ayar İşleminde Kullandığımız Argümanlar



Şekil 5.13 Aynı Anlamlı Cümle Tespiti ve Aynı Anlamlı Cümle Üretimi

Proje konusu olan Türkçe dilinde aynı anlamlı cümle üretimi için Türkçe kelime hazinesine sahip bir parçalayıcının girdi olarak verilen metni ayıklaması, daha sonra ayrı ayrı elde edilen kelime köklerinin kodlayıcıya verilmesi, yine aynı şekilde Türkçe ve göreve uygun kodlanmış bir başka metnin referans verilmesi gerekmektedir. Dil eğitimi tamamlandıktan sonra ince ayar işlemine geçilecek ve bu sefer hazırlanmış ikili soru cümleleri ile yapay sinir ağının ağırlıkları hesaplanacak, hangi kelimelerin yerine hangi kelimelerin kullanılabileceği, hangi kelimelerin hangi kelimeler ile daha sık geçtiği tespit edilebilecektir.

T5 dil modelinin çok sayıda dildeki versiyonu olan MT5 ile proje, Türkçe dilinde gerçeklenmiştir. MT5 dil modelinin çalışma prensibi, T5 ile bire bir denecek kadar benzerdir. T5'in veri seti C4 iken MT5'in oluşumunda C4'ün çok sayıda dildeki versiyonu olan MC4 (Multi-lingual Colossal Clean Crawled Corpus) veri seti kullanılmıştır. Her dilin kuralları farklı olacağından ve her dilde bulunan veri miktarı aynı olmayacağından parametreler konusunda iki model değişiklik göstermiştir.

Türkçe aynı anlamlı cümle üretimi işleminde önceden eğitilmiş MT5 dil modeli olarak mt5-base, tokenizer olarak da mt5-base kullanılmıştır. Türkçe aynı anlamlı cümle üretimi için kullanılan eğitim fonksiyonunun parametreleri Şekil 5.14'te verilmiştir. Train ve validation veri seti 2/1 oranında ayrılmıştır. Veri setlerinin büyüklüklerine göre her epoch için geçen süre değişmekle birlikte her modelin 3 epoch'luk eğitimi

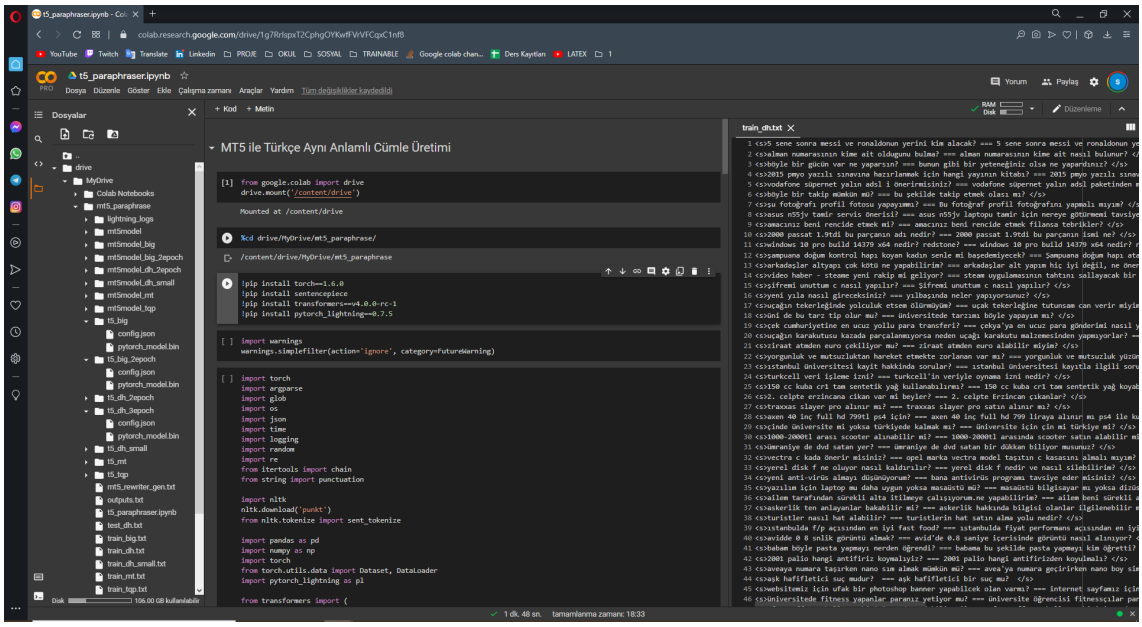
yaklaşık 3-5 saat aralığında sürmüştür. mt5-base modelinin daha büyük olan versiyonu olan mt5-large modeli ile eğitim denenmiştir fakat Google Colab'deki kısıtlamalar dolayısıyla eğitim tamamlanamamıştır. Ek olarak maximum sequence length ve batch size değerleri de değiştirilerek eğitim denenmiştir fakat yine Google Colab'deki kısıtlamalar buna engel olmuştur.

```
args_dict = dict(
    data_dir="",
    output_dir="mt5model_big_3epoch/",
    model_name_or_path="google/mt5-base",
    tokenizer_name_or_path="google/mt5-base",
    max_seq_length=150,
    learning_rate=3e-4,
    weight_decay=0.0,
    adam_epsilon=1e-8,
    warmup_steps=0,
    train_batch_size=4,
    eval_batch_size=4,
    num_train_epochs=3,
    gradient_accumulation_steps=16,
    early_stop_callback=False,
    n_gpu=1,
    fp_16=False,
    opt_level='O1',
    max_grad_norm=1.0,
    seed=42,
)
```

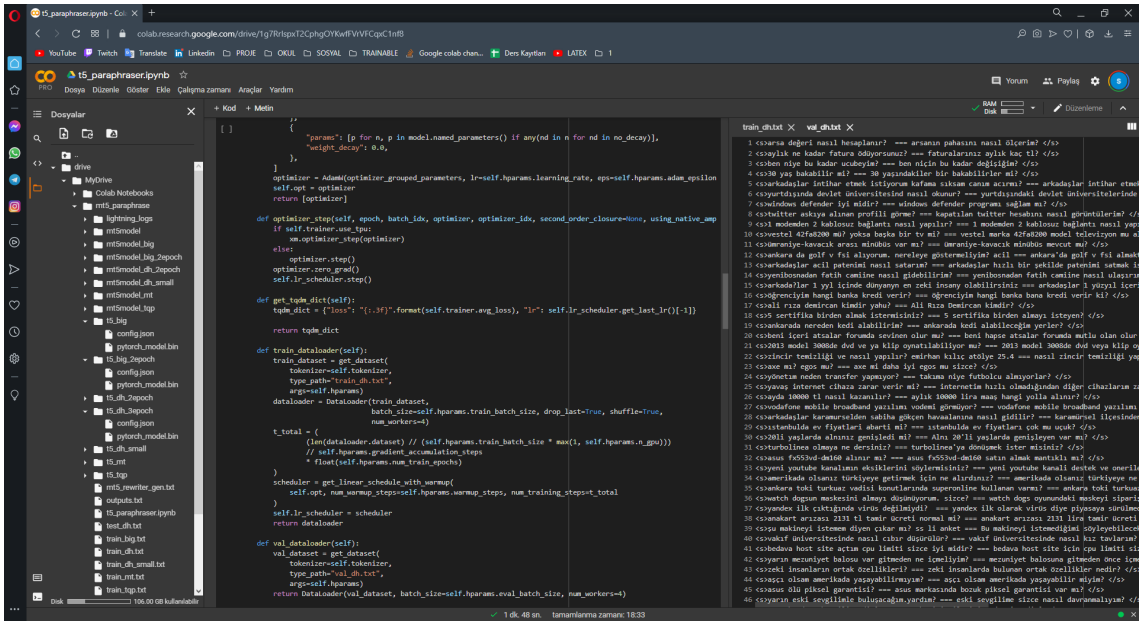
Şekil 5.14 Türkçe Aynı Anlamlı Cümle Üretimi İnce Ayar Fonksiyonundaki Argümanlar

5.3 Girdi-Çıktı Tasarımı

Türkçe aynı anlamlı cümle üretimi projesi, Opera tarayıcısından Google Colab'e girerek çalıştırılmaktadır. Model ve veri seti Google Drive'a kaydedilmektedir ve oradan yüklenmektedir.

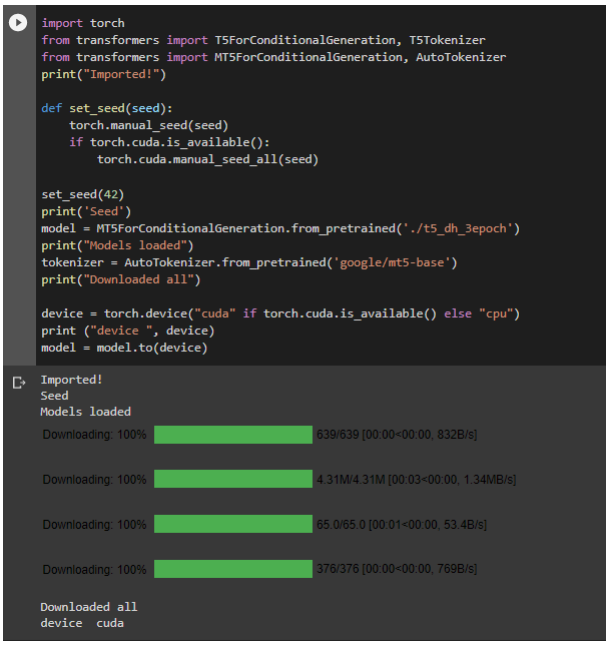


Şekil 5.15 Modellerin, Kodların ve Eğitim Veri Kümesini İçeren Ekran Görüntüsü



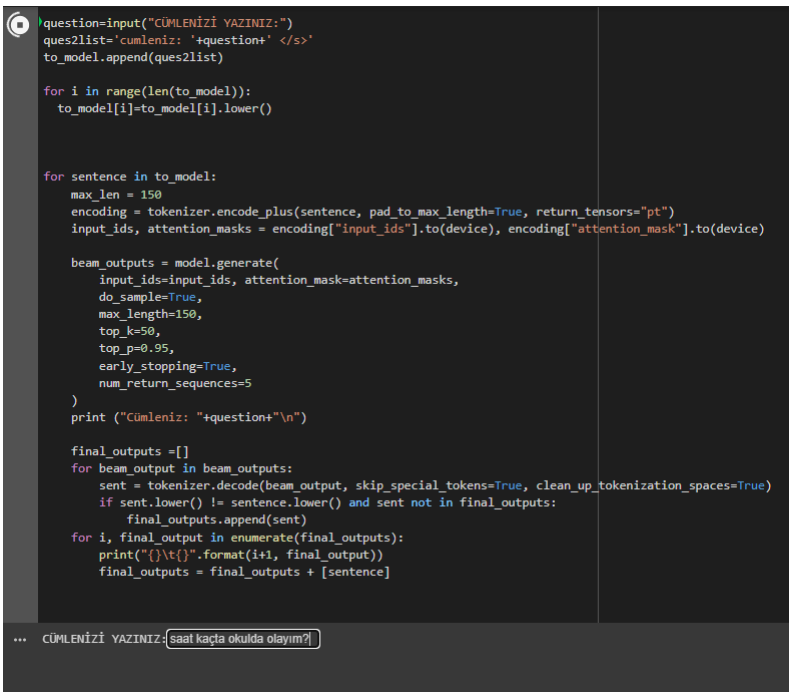
Şekil 5.16 Validation Veri Setini İçeren ve Kodda Dizin Gösterilen Ekran Görüntüsü

Eğitimi tamamlanan model Google Drive'a kaydedilmektedir. Modelin testlerinin görülebilmesi için Google Drive'daki modelin bulunduğu klasör dizini koda eklendikten sonra modelin yüklenmesi ve ardından daha önceden eğitilmiş olan "mt5-base" tokenizer yüklenmektedir.

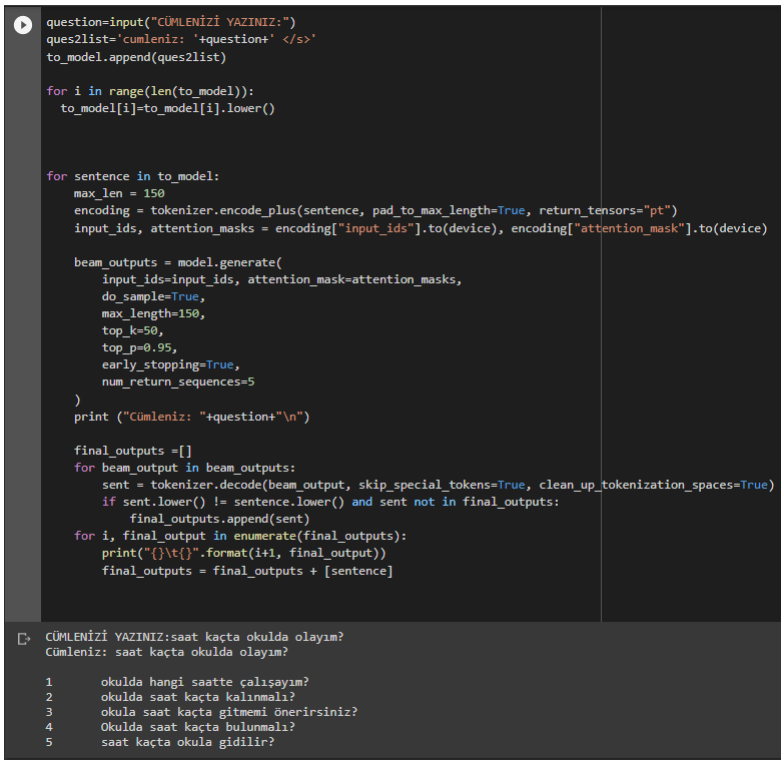


Şekil 5.17 Tarafımızca İnce Ayar Yapılmış Modelin ve Daha Önce Eğitilmiş Tokenizer'ın Yüklenmesi

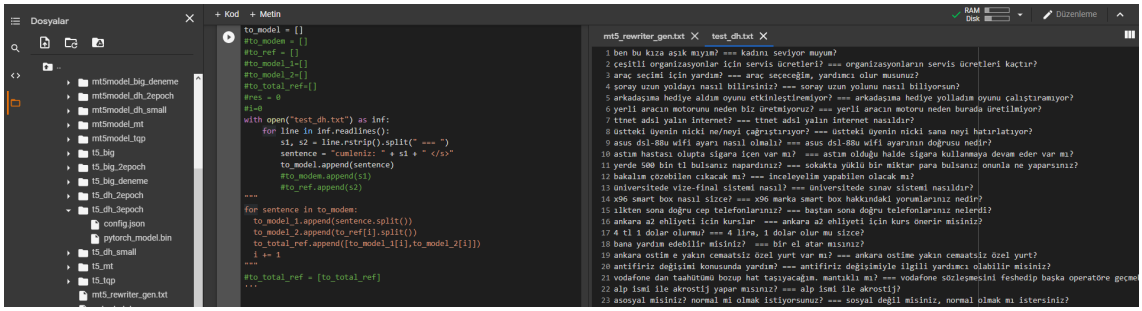
Model ve tokenizer yüklenmesinin ardından kullanıcının konsola gireceği girdi cümle alınarak Şekil 5.18'te generate fonksiyonu aynı anlama gelen cümlelerin üretilmesi sağlanmıştır (Şekil 5.19).



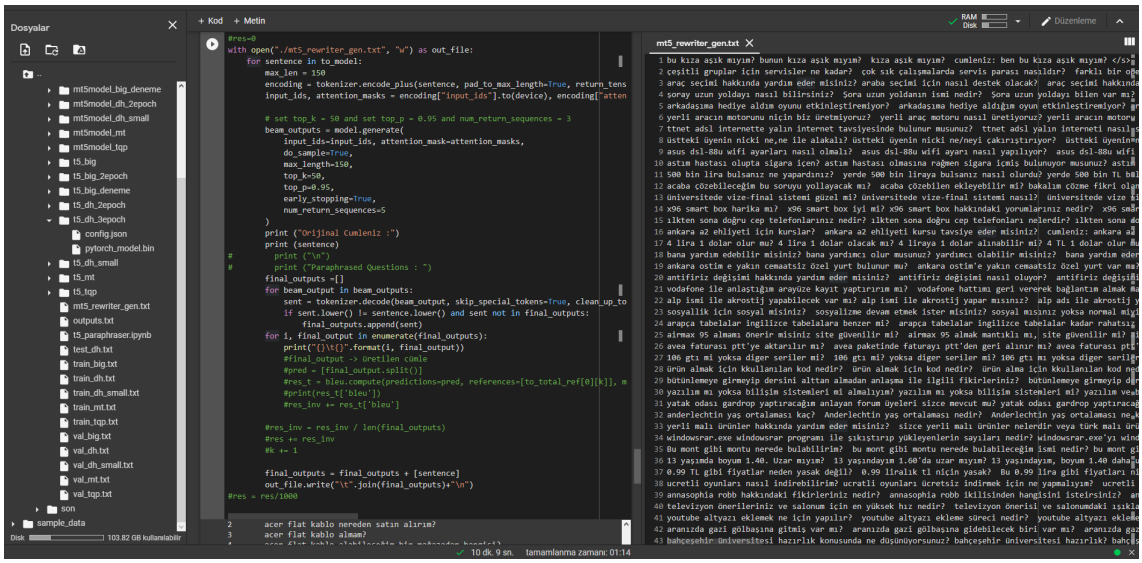
Şekil 5.18 Kullanıcı Tarafından Girdi Cümle Yazılması



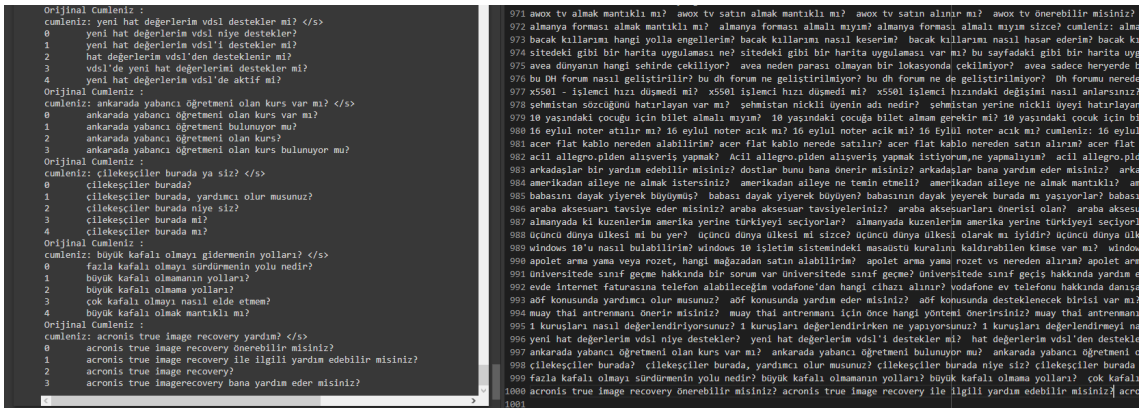
Şekil 5.19 Üretilen Aynı Anlamlı Cümle Çıktıları



Şekil 5.20 Test Cümlelerinin Dosyadan Verilmesi



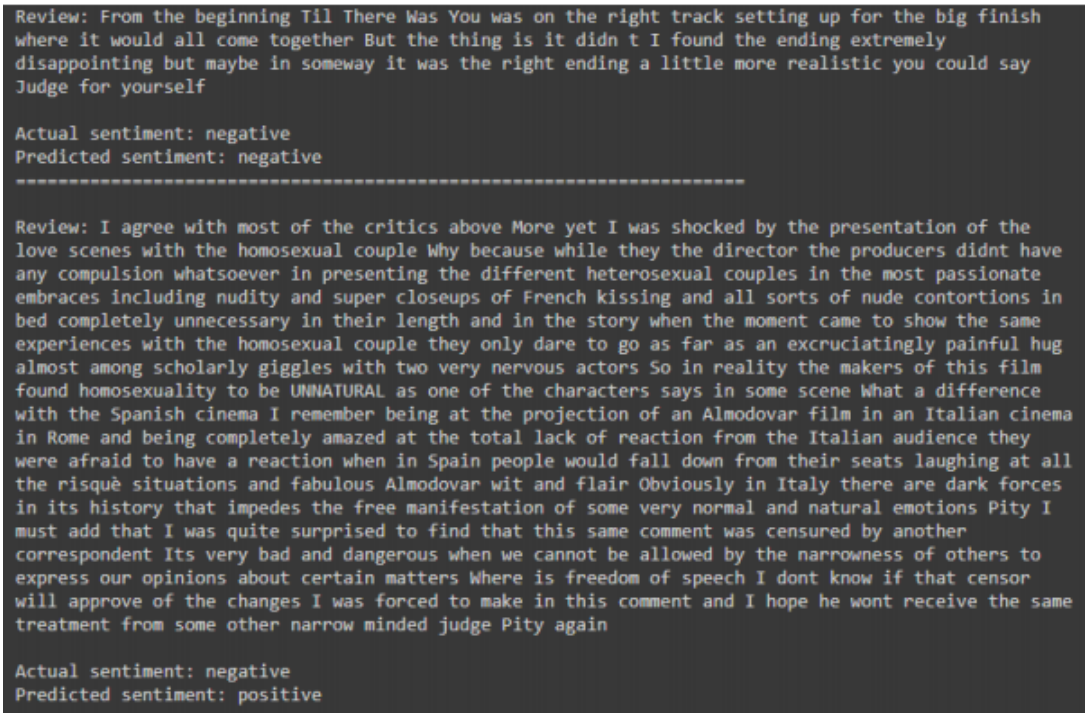
Şekil 5.21 Üretilen Cümlelerin Dosyaya Yazılması



Şekil 5.22 Konsola Yazdırılan ve Dosyaya Yazılan Cümleler

6
Uygulama

T5 ile İngilizce dilinde duygu analizi işleminde büyük veri setine uygulanan ince ayar işlemi sonucu test veri setindeki cümlelerin duygu analizinin sonuç çıktıları Şekil 6.1'de ifade edilmiştir.



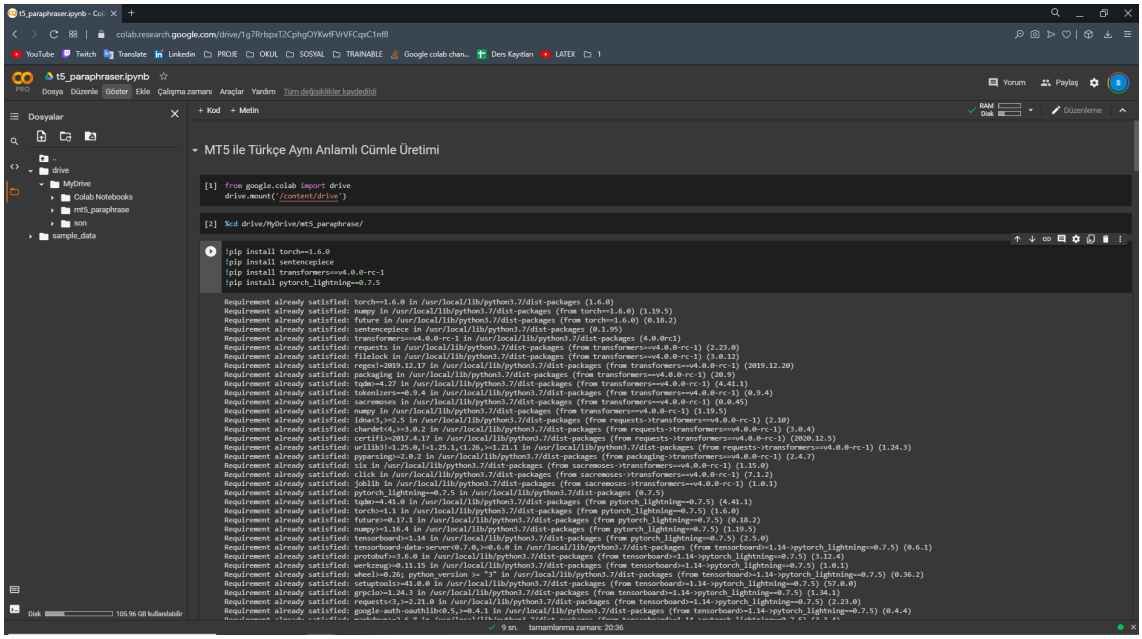
Şekil 6.1 Cümlelere Yapılan Tahminler

T5, BoW ve Seq2Seq ile İngilizce dilinde aynı anlamlı cümle üretimi işleminde verilen cümlelere ait üretilen çıktılar Tablo 6.1'de verilmiştir.

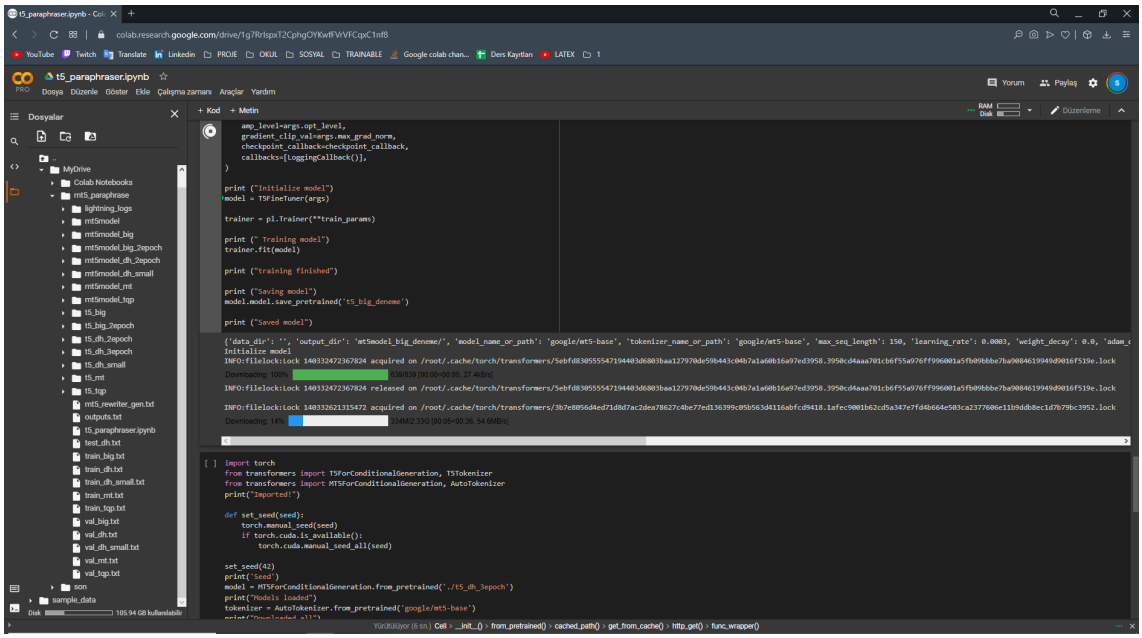
Referans	**Is this the start of an aviation revolution?**
BoW	Was this the start of an aviation revolution?
Seq2Seq	Is the world aviation revolution?
T5	Is the first runway in the world about to go into air-strips?
Referans	**What's your favorite sleeping position?**
BoW	What is your favorite position position?
Seq2Seq	What is your favorite sleeping position?
T5	What are some of your favorite sleeping positions?
Referans	**What's the best career decision you've ever made?**
BoW	What 's the best decision you ever made?
Seq2Seq	What is the best career decision you ever made?
T5	What was the best career decision of your life?
Referans	**Can I make my own hand sanitizer?**
BoW	Can i make make sanitizer?
Seq2Seq	Can I make my own hand sanitizer?
T5	What are some easy ways to make personal hand sanitizer?
Referans	**Do you offer business insurance and protection plans?**
BoW	Do you have insurance and protection coverage for businesses?
Seq2Seq	Do you offer business insurance plans?
T5	Do you sell business insurance? How does it work?
Referans	**What should I do if I have problems with my truck?**
BoW	What should i do if i have problems with my truck?
Seq2Seq	How do I get my truck?
T5	What should we do if my truck is leaking?

Tablo 6.1 İngilizce Aynı Anlamlı Cümle Üretimi Sonuçları

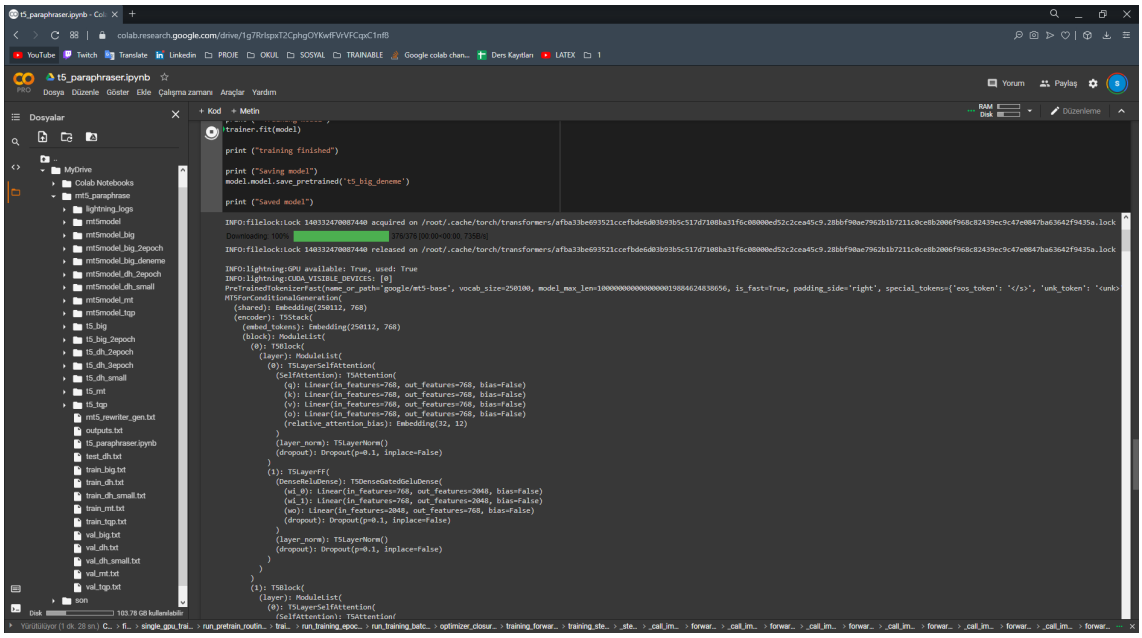
Türkçe dilinde aynı anlamlı cümle üretiminde ise uygulamanın çalışmasına ilişkin ekran çıktıları aşağıdaki şekillerdeki gibidir.



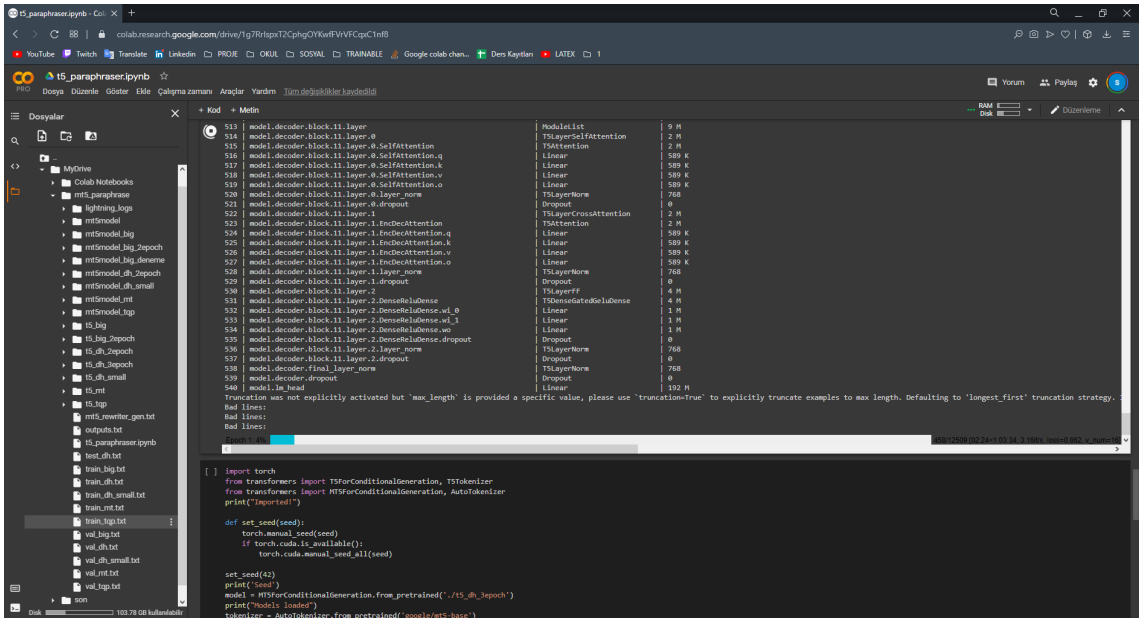
Şekil 6.2 Google Drive Bağlantısı Kurulması



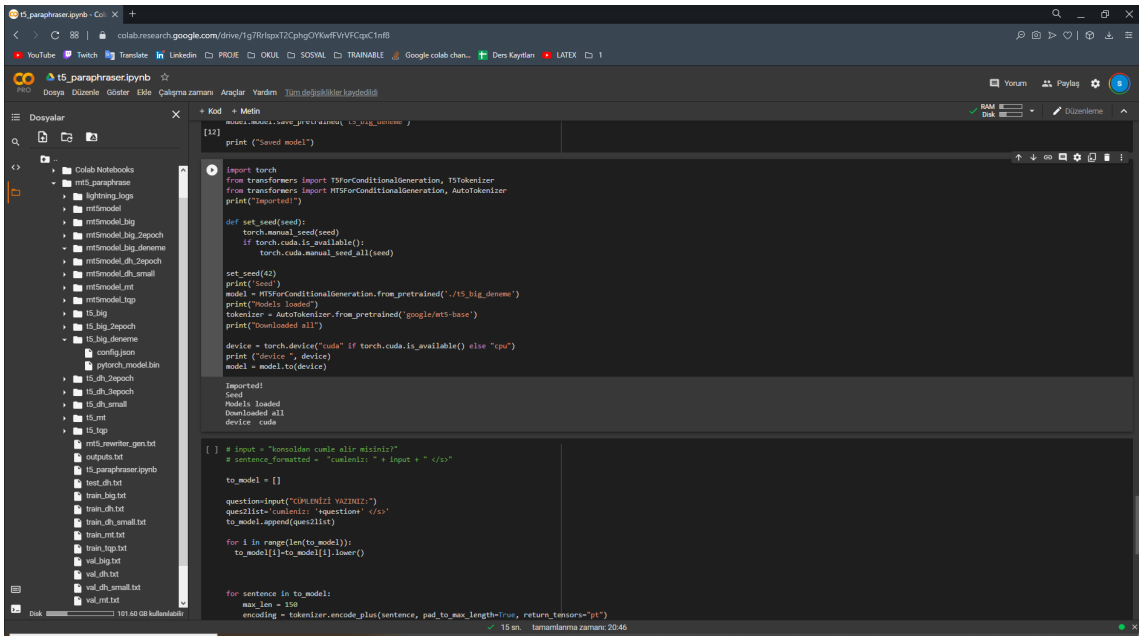
Şekil 6.3 Gerekli Kütüphanelerin Kurulumu



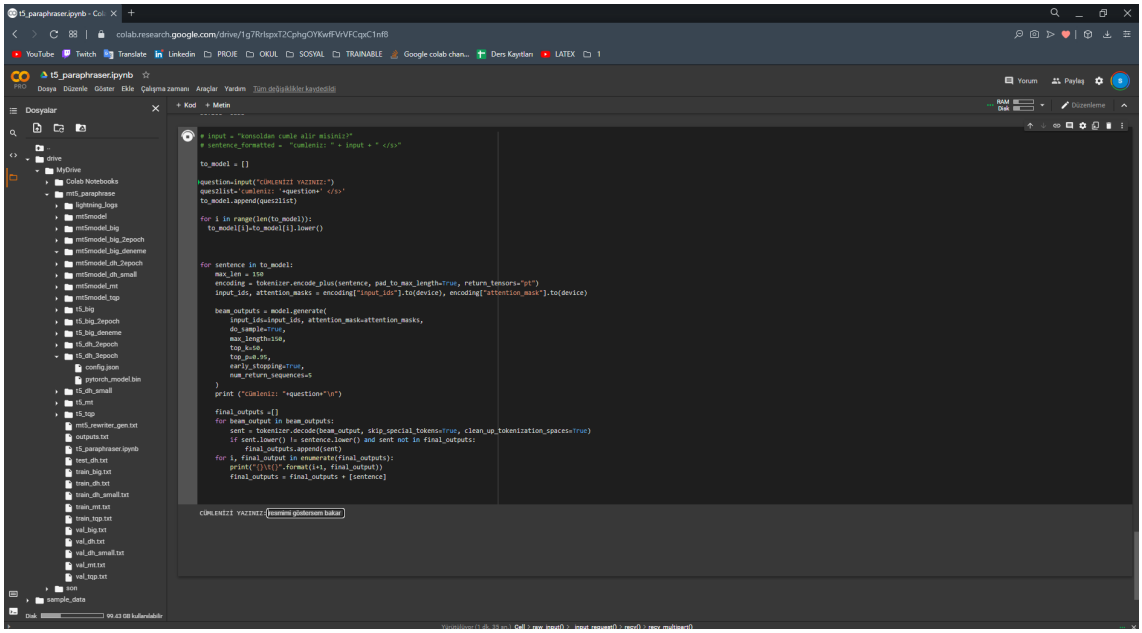
Şekil 6.4 Modelin Eğitim Aşaması - 1



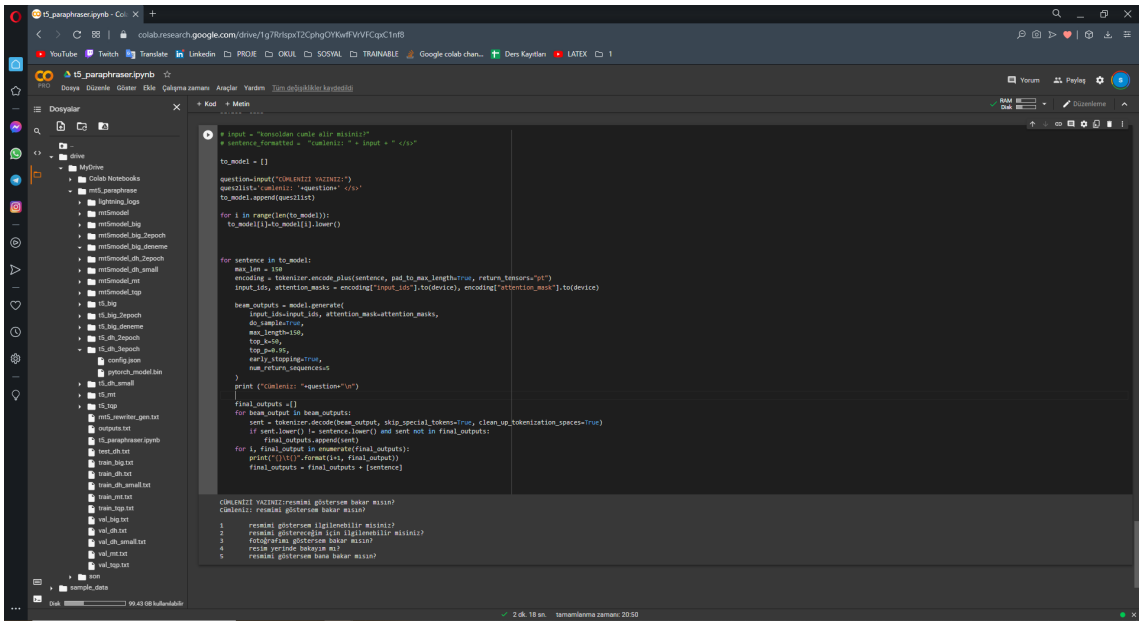
Şekil 6.5 Modelin Eğitim Aşaması - 2



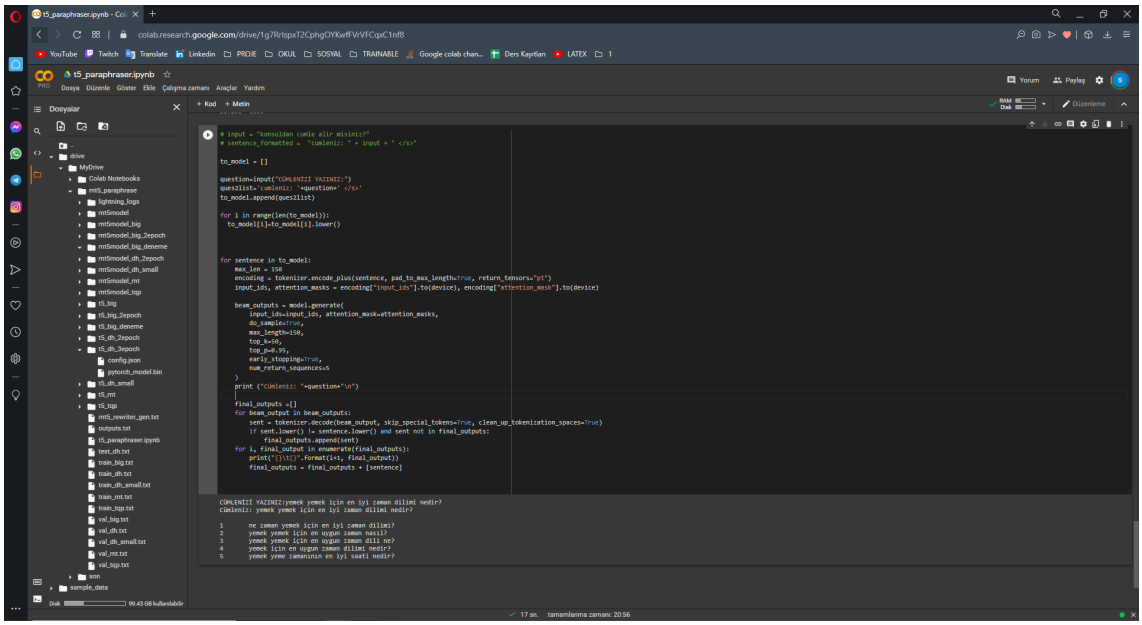
Şekil 6.6 Eğitilen Modelin Yüklenmesi



Şekil 6.7 Konsoldan Cümle Girişi



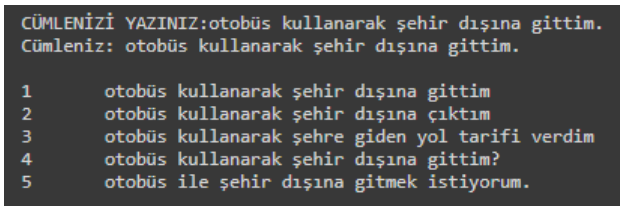
Şekil 6.8 Cümlelere Yapılan Tahminler - 1



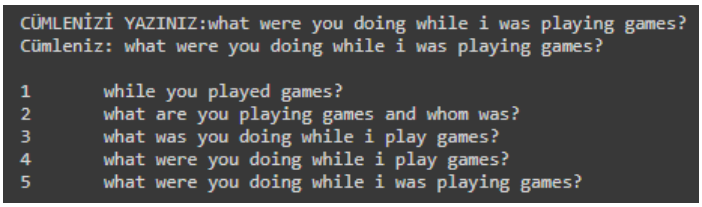
Şekil 6.9 Cümlelere Yapılan Tahminler - 2

7
Deneysel Sonuçlar

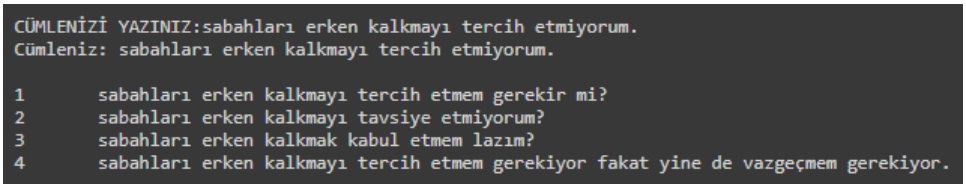
Sisteme verilen uç noktalara ait girdiler ve alınan sonuçların ekran çıktıları aşağıdaki şekillerde gösterilmiştir.



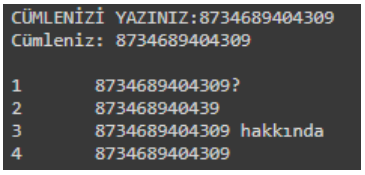
Şekil 7.1 Uç Durum: Soru Cümlesi Olmayan Cümle



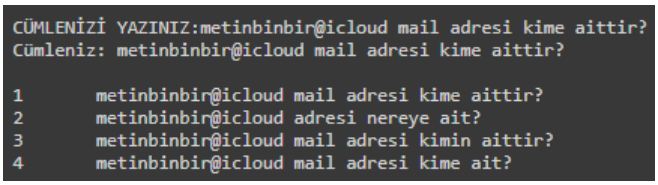
Şekil 7.2 Uç Durum: İngilizce Soru Cümlesi



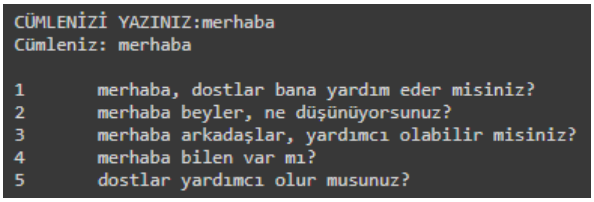
Şekil 7.3 Uç Durum: Soru Cümlesi Olmayan Olumsuz Cümle



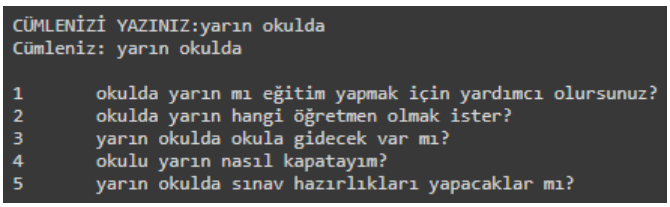
Şekil 7.4 Uç Durum: Yalnızca Sayı Girişi



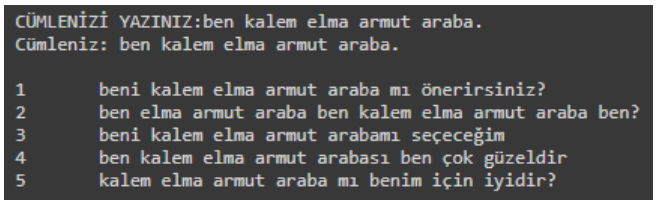
Şekil 7.5 Uç Durum: Özel Karakter İçeren Girdi Durumu



Şekil 7.6 Uç Durum: Tek Kelimelik Girdi Durumu



Şekil 7.7 Uç Durum: Eksik Cümle



Şekil 7.8 Uç Durum: Yüklem İçermeyen Cümle

Veri seti tasarımı kısmındaki Tablo 5.14'teki DH-S, DH, MT, TQP ve BIG kodlu 5 adet veri setiyle MT5 modeline ince ayar yapılmış modellere ait girdiler ve ürettikleri aynı anlamlı cümlelerin çıktıları Tablo 7.1, Tablo 7.2, Tablo 7.3 ve Tablo 7.4'te verilmiştir.

Referans	**yaşadığın şehirde en güzel manzaraya sahip yer neresidir?**
DH-S	Şehirdeki en güzel manzaraya sahip mekan neresi?
DH	hayatını sürdürdüğüm şehirde en güzel manzara alan bölge nedir?
TQP	Senin yaşadığın şehirde en iyi manzaraya sahip yer nedir?
MT	Yaşadığınız şehirde en güzel manzara alabileceğin yer neresi?
BIG	yaşadığın şehirde en güzel manzaraların neler olduğunu biliyor musun?
Referans	**hangi saatlerde kitap okumak daha verimlidir?**
DH-S	hangi saatlerde kitap okunabilir?
DH	hangi saatlerde kitap okumak daha yararlı?
TQP	Ne zaman kitap okumak daha güvenli?
MT	Özel kitap okumak için hangi saatler uygun olacak?
BIG	kitap okumak için en uygun saat nedir?
Referans	**son model arabalar çevreye daha mı az zarar veriyor?**
DH-S	son model arabaların çevreye zararları daha mı az oluyor?
DH	son arabaların çevreye etkisi az mi?
TQP	Son model araba çevreye daha az zarar verir mi?
MT	Son model arabalar çevreye zarar veriyor mu?
BIG	yeni arabalar çevreye daha az zarar veriyor mu?
Referans	**insanlar boş zamanlarını neler yaparak daha verimli hale getirebilir?**
DH-S	insanlar boş zamanlarını iyileştirebilecek şeyler?
DH	insanlar boş zamanlarını nasıl iyileştirir?
TQP	İnsanlar boş zamanını nasıl geçirir?
MT	boş zamanımızı en verimli hale getirmenin en iyi yolları nelerdir?
BIG	insanlar boş zamanlarını nasıl daha verimli yapabilir?
Referans	**ülkemizdeki konut fiyatlarında ileriki zamanlarda nasıl bir artış bekleniyor?**
DH-S	ülkemizdeki konut fiyatlarında ilerleyen haftalarda ne kadar artış olur?
DH	ülkemizdeki konut fiyatları gelecek tarihte nasıl artacak?
TQP	ülkemizdeki konut fiyatlarına kısa bir zaman için biraz artış bekliyor musun?
MT	Fransa'da konut fiyatlarının ileriki zamanlarda nasıl artacağını düşünüyorsunuz?
BIG	ülkemizdeki konut fiyatları gelecekte nasıl artacaktır?

Tablo 7.1 Farklı Veri Setlerine Göre Üretilen Aynı Anlamlı Cümle Çıktıları - 1

Referans	**okul bahçesinde top oynayan çocukları gördünüz mü?**
DH-S	Okul bahçesinde top oynayan çocuklar var mı?
DH	okul bahçesinde top oynayan çocuklardan haberdar mısınız?
TQP	okula gidene katlanan çocukları gördün mü?
MT	Okul bahçesinde top oynananlar neler?
BIG	okul bahçesinde top oynayan çocuklardan haberiniz var mı?
Referans	**salgın hastalıkların etkisi ne kadar sürecek?**
DH-S	salgın hastalıklarının etkisi kaçtır?
DH	salgın hastalıkları kaç yılda etki eder?
TQP	sağlık hastalıklarının sebebi ne kadar sürecek?
MT	salgın hastalıkları nasıl sürecek?
BIG	salgın hastalıklar ne kadar zaman alır?
Referans	**sıcak havalarda parkta yürüyen insanlar görüyor musunuz?**
DH-S	parkta sıcak havalarda yürüyen var mı?
DH	sıcak havalarda parkta duran insanlara rastlanır mı?
TQP	Sıcak havalarda parkta yürüyüşe çıkan insanlar görüyor musun?
MT	Yüksekliklere yakın parkta yürüyen insanlar görüyor musunuz?
BIG	sıcak havalarda parkta yürüyebilenler bulunuyor mu?
Referans	**deniz kenarında uçan kuşlar çok mutlu gözükmüyorlar mı?**
DH-S	deniz kenarında uçan kuşlar çok mutlu görünmezler mi?
DH	deniz kenarında uçan kuşlar fazla mutlu değil mi?
TQP	Mavi deniz kenarında uçan kuşlar çok mutlu görünmüyor mu?
MT	Denizlerde uçan kuşlar mutlu mu?
BIG	deniz kenarında uçan kuşlar gerçekten mutlu değil mi?
Referans	**ülke genelinde bisiklet kullanımının artması konusunda ne düşünüyorsun?**
DH-S	ülke genelinde bisiklet kullanımı artış yönünde düşünceniz nedir?
DH	bisiklet kullanımının yükselmesi hakkında fikirleriniz nedir?
TQP	Dünya genelinde bisiklet kullanımının artırılması konusunda fikrin nedir?
MT	Bisiklet kullanımının artmasının altındaki görüşleriniz nelerdir?
BIG	ülke genelinde bisiklet kullanımının artması hakkındaki görüşleriniz nedir?

Tablo 7.2 Farklı Veri Setlerine Göre Üretilen Aynı Anlamlı Cümle Çıktıları - 2

Referans	**Askerliğimin tecillenmesi için nereye başvurmalıyım?**
DH-S	askerliğimi nereye tecilleyebilirim?
DH	askerliğimi tecillemek için kurum önerisi?
TQP	Nereden bir askerliğini tecillemek için gerekli olduğunu bana söyleyebilir misin?
MT	Biri kendi askerlik işlerini tecillemek için nereye başvurabilir?
BIG	Askerlik durumu nasıl tecillenebilir?
Referans	**Bayrampaşaya nereden gidebilirim?**
DH-S	bayrampaşaya yol tarifi?
DH	bayrampaşada hangi yerden gidebilirim?
TQP	Bayrampaşaya giden yolu bana söyleyecek misin?
MT	Bayrampaşaya gitmenin en iyi yolu nedir?
BIG	Bayrampaşaya nasıl gideceğimi bana söyleyebilir misiniz?
Referans	**Telefonumun ekranı kırıldı tamire nereye götürmeliyim?**
DH-S	Telefonumun ekranı kırıldı nereye götüreyim?
DH	sıcak havalarda parkta duran insanlara rastlanır mı?
TQP	Android telefonumun ekranı kırıldı tamire nereye gitmeli?
MT	Telefonumun ekranı kırıldı nereye göndermeliyim?
BIG	Telefonumun ekranı kırıldı, tamire nereye götürmem gerekir?
Referans	**Motorlu taşıtların çevreye verdiği zarar ile ilgili fikriniz nedir?**
DH-S	motorlu taşıtların çevreye verdiği zarar hakkında ne düşünüyorsunuz?
DH	motorlu taşıtların çevreye verdiği zarar hakkında ne düşünüyorsunuz?
TQP	motolu taşıtların çevreye verdiği zarar hakkında ne düşünüyorsun?
MT	Motorlu taşıtlar çevreye verdiği zarar ile ilgili düşüncelerin neler?
BIG	Motorlu taşıtların çevrelere verdiği zarar hakkında ne düşünüyorsunuz?
Referans	**paramızın değerinin düşmesi hakkında ne düşünüyorsunuz?**
DH-S	paramız değerinin düşmesiyle ilgili düşünceniz nedir?
DH	paramızın değerinin çok olarak düşmesi ile ilgili ne düşünüyorsunuz?
TQP	Paramızın değeri nedir?
MT	Paranın değeri düşüyor mu?
BIG	paranın değerinin düşürülmesine ne düşünüyorsunuz?

Tablo 7.3 Farklı Veri Setlerine Göre Üretilen Aynı Anlamlı Cümle Çıktıları - 3

Referans	**Bu yaz başlayan hangi yaz dizisini izleyeceksiniz?**
DH-S	yazın başlayan hangi yaz dizisini izleyeceksiniz?
DH	bu yaz başlayan yaz dizileri?
TQP	Bu yaz ne tür dizileri izliyorsun?
MT	Bir yaz başlamak için izlemesi gereken en iyi yaz dizileri hangileridir?
BIG	Bu yaz çıkacak en iyi TV şovlarından bazıları nelerdir?
Referans	**Bir kızdan çok hoşlanıyorum nasıl çıkma teklifi etmemi önerirsiniz?**
DH-S	çok hoşlanıyorum şimdi çıkma teklifi yapımı önerir misiniz?
DH	bir kıza çok hoşlanıyorum, ne yapmalıyım?
TQP	Nasıl bir kızdan hoşlanacağımı bana söyleyebilir misiniz?
MT	Bir kızın çok hoşlandığım için nasıl çıkma teklifi etmeliyim?
BIG	bir kadından hoşlanıyorum, nasıl çıkma teklifi etmemi tavsiye edersiniz?
Referans	**Aranızda daha önce hiç balık yememiş var mı?**
DH-S	aranızda hiç balık yememiş bulunuyor mu?
DH	aranızda daha önce hiç balık tüketmemiş var mı?
TQP	Şu anda hiç balık yememiş var mı?
MT	Bugüne kadar hiç balık yememiş var mı?
BIG	aranızda daha önce hiç balık tüketmemiş kimse var mı?
Referans	**Ölmeden kesin deneyimlemem lazım dediğiniz şeyler var mı?**
DH-S	ölmeden önce ne deneyimlemek gerekir?
DH	ölmeden önce kesin neler görmemi önerir misiniz?
TQP	Ölmeden gerçekten deneyimlemem gereken şeyler var mı?
MT	Ölmeden kesin deneyimlemem gereken şeyler nelerdir?
BIG	Ölmeden önce deneyimlememi gerektiren şeylerden bazıları nelerdir?
Referans	**Dünya üzerinde insanlar tarafından konuşulan dillerin sayısı nedir?**
DH-S	toplumdan konuşulan dillerin sayısı nedir?
DH	insanların dünya üzerinde konuşma dillerinin ne kadar olduğunu düşünüyorsunuz?
TQP	Dünya üzerinde insanlar için konuşulan dilleri kaç tane?
MT	Dünyada konuşulan dillerin sayısı nedir?
BIG	dünyada insanlar tarafından konuşulan dillerin sayısı ne kadar?

Tablo 7.4 Farklı Veri Setlerine Göre Üretilen Aynı Anlamlı Cümle Çıktıları - 4

Farklı epoch değerleriyle (2 ve 3) ve DH kodlu veri setiyle eğitilmiş iki farklı modelin aynı cümleye dair ürettiği çıktılar aşağıdaki şekillerde gösterilmiştir.

```
CÜMLENİZİ YAZINIZ:günde ne kadar su içmeliyim?

Cümleniz: günde ne kadar su içmeliyim?
1       hangi günde suyu içmeliyim?
2       Bir günde ne kadar su içmeli miyim?
3       günde hangi su içmeliyim?
4       günde günde ne kadar su içebiliriz?
5       günde ne kadar su içebiliriz?
```

Şekil 7.9 1.Cümle - Epoch Sayısı=2

```
CÜMLENİZİ YAZINIZ:günde ne kadar su içmeliyim?

Cümleniz: günde ne kadar su içmeliyim?
1       günlük ne kadar su içmelim?
2       günde kaç litre su içmeliyim?
3       günde kaç su içeyim?
4       günde ne kadar su içmem gerekiyor?
5       gün içerisinde ne kadar su tüketmeliyim?
```

Şekil 7.10 1.Cümle - Epoch Sayısı=3

```
CÜMLENİZİ YAZINIZ:kendinizi huzursuz hissettiğiniz anlar oluyor mu?

Cümleniz: kendinizi huzursuz hissettiğiniz anlar oluyor mu?
1       bu kadar hiçbir şey hissettiğiniz anlar oluyor mu?
2       özünüzdeki kişisi bir insana dikkat eder misiniz?
3       kendinizi birin huzursuz hissettiğiniz şeyler?
4       kendinizi uzursuz hissedeceğiniz birini düşünüyor musunuz?
5       kendinizi mutlu hissedecek biri var mı?
```

Şekil 7.11 2.Cümle - Epoch Sayısı=2

```
CÜMLENİZİ YAZINIZ:kendinizi huzursuz hissettiğiniz zamanlar oluyor mu?

Cümleniz: kendinizi huzursuz hissettiğiniz zamanlar oluyor mu?
1       kendinizi hiç huzursuz hissedtiğiniz zamanlar olur mu?
2       kendinizi huzursuz hissettiğiniz zamanlar yaşanıyor mu?
3       kendinizi huzursuz hisseden var mı?
4       kendinizi uyumsuz hissettiğiniz zamanlar olur mu?
5       kendinizi hayal ettiğiniz zamanlar oluyor mu?
```

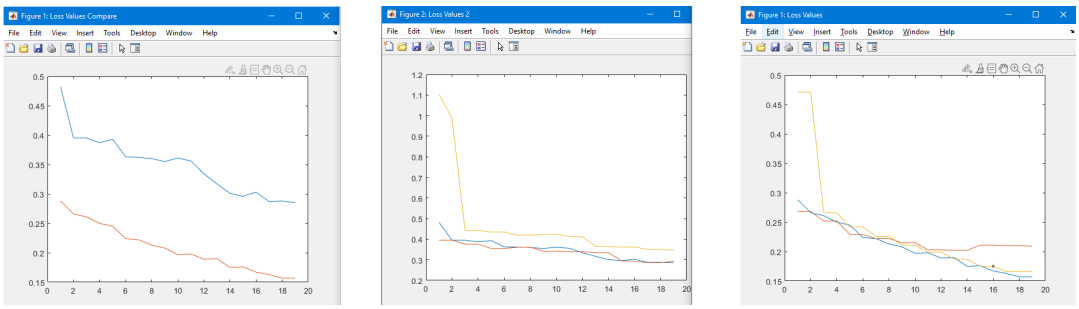
Şekil 7.12 2.Cümle - Epoch Sayısı=3

8
Performans Analizi

8.1 T5 İngilizce Dil Modelleme

Tarafımızca oluşturulmuş İngilice T5 dil modeli üzerinde ince ayar işlemi için dil modelinin kayıp (loss) değerleri veri setinin büyüklüğüne ve tekrarlara göre Şekil 8.1'deki grafiklerde ifade edilmiştir.

Sarı renkli çizgi: avg_train_loss
Mavi renkli çizgi: loss
Kırmızı renkli çizgi: val_loss
x axis: epoch
y axis: loss değerleri



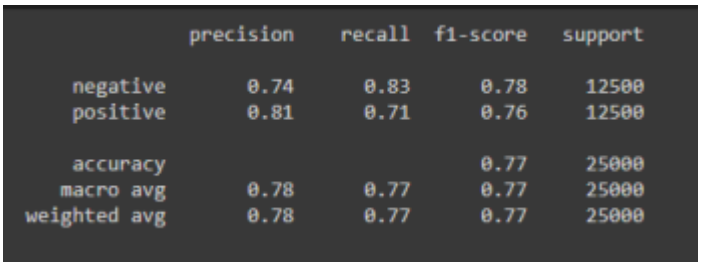
Şekil 8.1 T5 İngilizce Dil Modellemesinde Loss Değerleri
(Karşılaştırma-Küçük Veri Seti-Büyük Veri Seti)

8.2 T5 ile İngilizce Duygu Analizi

Büyük veri setiyle oluşturulmuş İngilizce dilindeki T5 dil modeli üzerinde uygulanan ince ayar işleminde, cümle üzerinden duygu analizi işlemine ait Hesaplanan-Gerçeklik matrisinde (Tablo 8.1) gösterilmiştir. Duygu analizi işlemine ait F1 skoru da Şekil 8.2'de gösterilmiştir.

		Gerçek	
		Pozitif	**Negatif**
Hesaplanan	**Pozitif**	9824	3078
	Negatif	2676	9422

Tablo 8.1 Hesaplanan-Gerçeklik Matrisi



	precision	recall	f1-score	support
negative	0.74	0.83	0.78	12500
positive	0.81	0.71	0.76	12500
accuracy			0.77	25000
macro avg	0.78	0.77	0.77	25000
weighted avg	0.78	0.77	0.77	25000

Şekil 8.2 İnce Ayar İşleminin Başarısına Ait Metrik Değerler

8.3 İngilizce Aynı Anlamlı Cümle Üretimi Değerlendirmesi

T5, BoW ve Seq2Seq ile İngilizce diline ait aynı anlamlı cümle üretiminin başarısı insan değerlendirmesiyle ölçülmüştür ve en başarılı sonuçların önceden eğitilmiş T5 dil modeliyle yapıldığı saptanmıştır.

8.4 MT5 ile Türkçe Aynı Anlamlı Cümle Üretimi

8.4.1 BLEU Skoru

BLEU skoru, girilen cümle ve oluşturulan cümledeki kelimelerin n-gramlarını kontrol ederek ek alıp almadığına bakmaksızın bire bir aynı kelimeleri içermesi durumunda arttığı için aynı anlamlı cümle üretiminin başarısını doğrudan ölçen bir metrik olmayacağı kanaatine varılmıştır. MT5 dil modeline uygulanan ince ayar işlemi sonucu en yüksek BLEU skorunu DH-S veri setiyle eğitilen model almıştır (Tablo 8.2).

Veri Seti	BLEU-1	BLEU-2	BLEU-3
DH-S	0.926	0.787	0.586
DH	0.768	0.606	0.420
TQP	0.415	0.255	0.147
MT	0.471	0.312	0.197
BIG	0.752	0.585	0.420

Tablo 8.2 Farklı Veri Setleri Kullanılarak Eğitilmiş MT5 Modellerin BLEU Skorları

8.4.2 İnsan Değerlendirmesi

5 farklı veri setiyle eğitilmiş 5 adet MT5 modeli ile 20 cümlenin her biri için 5'er tane aynı anlamlı çıktı üretilmiştir ve ana dili Türkçe olan 5 farklı insan tarafından, Excel dosyasında; her satırda en iyi ve en kötü üretilmiş cümleye ait çıktılar işaretlenmiştir. Sonuç olarak en başarılı modelin tarafımızca ekleme yapılmış olan manuel hazırlanmış DH kodlu veri setiyle eğitilmiş model olduğu sonucuna varılmıştır.

Veri Seti ID	Başarılı	Başarısız
DH-S	%19.8	%14.2
DH	%26.6	%14.6
TQP	%16	%32.2
MT	%13.4	%27.2
BIG	%24.2	%11.8

Tablo 8.3 Farklı Veri Setleri Kullanılarak Eğitilmiş MT5 Modellerin İnsan Değerlendirmesi

8.4.3 Sonuçların Değerlendirilmesi

8.4.3.1 Uç Noktaların Değerlendirilmesi

Şekil 7.1, Şekil 7.3, Şekil 7.5 görsellerinde üretilen cümlelerin referans edilmiş cümle ile aynı veya yakın anlam içerdiği görülmüştür. Bunun anlamı model eğitilirken yalnızca ikili soru cümleleri kullanılsa dahi soru olmayan cümlelerde de başarı elde edebilecek olması, ayrıca eş anlamlı cümlesi üretilecek referans cümlenin özel karakter bulundurması ile ilgili bir problemin olmadığıdır.

Şekil 7.4, Şekil 7.6, Şekil 7.7 görsellerinde eş anlamlı cümle üretilmesi beklenen girdilerin, 1 ya da 2 kelime olduğu görülmektedir. Modelin eğitildiği veri setinde bu tarz kısa cümleler genellikle öneri almak veya yardım istemek için kullanılmıştır. Veri

setindeki kısa cümlelerin bu anlamı taşımasının etkisi çıktılarda görülmektedir.

Şekil 7.2 görselinde Türkçe için eğitilmiş veri setinin İngilizce içeriği olan bir cümleyi işleyemediği görülmektedir.

Şekil 7.8 görselinde insanlar için de bir anlam taşımayan cümle referans olarak verilmiştir. Neticesinde modelin cümlenin içerdiği kelimelerden öneriler beklediği görülmektedir.

8.4.3.2 Epoch Sayılarına Göre Değerlendirme

Epoch bir modelin tek bir veri setini birden fazla kullanarak kendini eğitmesini ayarlamaktadır. Bu sebeple belirli bir eşik değerine kadar epoch sayısını arttırmak modelin öğrenimini ve başarısını arttıracaktır. Modelin epoch için eşik değerinin 2 olmadığı ve 3 yapıldığında daha başarılı sonuçlar ürettiği Şekil 7.9 ile Şekil 7.10 arasında ve Şekil 7.11 ile Şekil 7.12 arasında görülmektedir.

9
Sonuç

Tamamlanan proje kapsamında aşamalı olarak İngilizce dili için aynı anlamlı cümle üretimi, İngilizce dil modelleme, modellenen dilin test edilmesi, Türkçe için aynı anlamlı cümle üretimi ve bu görev için eğitilmiş modelin başarısının ölçülmesi tamamlanmıştır.

Projenin ilk aşamasında kullanılacak mimari olarak T5 seçilmiş, Türkçe versiyonu olarak kullanabilmek adına MT5 dil modelinden yararlanılmıştır. Modellenen İngilizce dilin başarısını görmek için duygu analizi görevi çalıştırılmış ve başarısı F1 skor yöntemi ile ölçülmüştür. Nihayetinde, Türkçe için aynı anlamlı cümleler üretilmiş ve başarısını ölçmek için iki farklı yöntem denenmiştir. BLEU skor yönteminin projede beklenenlere ters düştüğü tespit edilip değerlendirmede ölçüt olarak kabul edilmemiştir. Diğer yöntem olan insan değerlendirmesinde ise beklendiği üzere insan tarafından oluşturulmuş veri seti ile eğitilmiş model en başarılı seçilmiştir.

Türkçe için aynı anlamlı cümle üretimi çalışması boyunca bu alanda ihtiyaç duyulan veri setine 20.000 cümle katkıda bulunulmuş ve büyütülmüştür. Makine çevirisiyle oluşturulmuş ve manuel olarak oluşturulmuş veri kümeleri karşılaştırılmış, cümle ikililerinin niteliklerinin başarıya etkisi gösterilmiştir.

Çalışma boyunca doğal dil işleme alanına yönelik çok farklı alanlara ait işler ile uğraşılmış ve edinilmesi beklenen tecrübe kazanılmıştır. Belirlenen görev kapsamında ülkemizin teknolojisine, veri kümesine katkıda bulunulmuştur. Tarafımızca katkı sağlanmış veri setinin modele etkisi, katılmayan veri seti ile oluşturulmuş modeller ile kıyaslanarak gösterilmiştir. Proje neticesinde, ana dili Türkçe olan insanlar tarafından değerlendirilen cümlelerde beğenilme oranı oldukça yüksektir. Bunun sonucu olarak büyüklüğü az olan bir veri seti üzerinde aynı anlamlı cümle üretimi görevi çalıştırılarak veri setinin sağlıklı bir şekilde büyütülmesi teorik olarak mümkündür. Proje amacı ile elde edilen veri setine katkılar yapılarak denenmesine rağmen veri seti günlük işlerde kullanılabilecek tüm cümleleri içermez ve uç nokta testlerinde bunun sebebi olarak başarısızlıklar görülmüştür.

Aynı anlamlı cümle üretimi amacı ile kullanılmış T5 mimarisinin diğer modellere göre daha başarılı olduğu, aynı zamanda diğer modellere göre kullanımının daha rahat olduğu belirtilmiştir. Bu hususta ileride bu alanda çalışmak isteyenler için bu modelin ilk tercihlerden olmasının düşünüldüğü belirtilmiştir. Aynı zamanda, üzerine çalışılacak veri seti mümkün oldukça insan emeği ile oluşturulmalı ve genişletilmelidir.

Referanslar

[1] E. Egonmwan and Y. Chali, "Transformer and seq2seq model for paraphrase generation," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*, Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 249–255. DOI: 10.18653/v1/D19-5627. [Online]. Available: https://www.aclweb.org/anthology/D19-5627.

[2] Y. Fu, Y. Feng, and J. P. Cunningham, "Paraphrase generation with latent bag of words," p. 12, 2019.

[3] M. Y. Sarıgöz, *Monatis/tqp: V0.1*, version v0.1, Apr. 2021. DOI: 10.5281/zenodo.4719801. [Online]. Available: https://doi.org/10.5281/zenodo.4719801.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, *Huggingface's transformers: State-of-the-art natural language processing*, 2020. arXiv: 1910.03771 [cs.CL].

[6] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, *mT5: A massively multilingual pre-trained text-to-text transformer*, 2020. arXiv: 2010.11934 [cs.CL].

[7] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An ımperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[8] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: A method for automatic evaluation of machine translation," Oct. 2002. DOI: 10.3115/1073083.1073135.

[9] Q. Lhoest, P. von Platen, T. Wolf, A. V. del Moral, Y. Jernite, S. Patil, M. Drame, J. Chaumond, J. Plu, J. Davison, S. Brandeis, T. L. Scao, V. Sanh, K. C. Xu, L. Tunstall, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, L. Debut, C. Delangue, T. Matussière, S. Bekman, and F. Lagunas, *Huggingface/datasets: 1.7.0*, version 1.7.0, May 2021. DOI: 10.5281/zenodo.4817769. [Online]. Available: https://doi.org/10.5281/zenodo.4817769.

[10] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, *Character-aware neural language models*, 2015. arXiv: 1508.06615 [cs.CL].

[11] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 115–124. DOI: 10.3115/1219840.1219855. [Online]. Available: https://www.aclweb.org/anthology/P05-1015.

[12] L. Sharma, L. Graesser, N. Nangia, and U. Evci, "Natural language understanding with the quora question pairs dataset," p. 10, Jul. 2019.

[13] Y. Scherrer, *TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages*, version 1.0, Zenodo, Mar. 2020. DOI: 10.5281/zenodo.3707949. [Online]. Available: https://doi.org/10.5281/zenodo.3707949.

Özgeçmiş

BİRİNCİ ÜYE

İsim-Soyisim: Metin BİNBİR
Doğum Tarihi ve Yeri: 12.04.1999, Gölcük
E-mail: l1119613@std.yildiz.edu.tr
Telefon: 0530 593 05 92
Staj Tecrübeleri:

İKİNCİ ÜYE

İsim-Soyisim: Sercan AKSOY
Doğum Tarihi ve Yeri: 30.07.1998, İstanbul
E-mail: l1119620@std.yildiz.edu.tr
Telefon: 0534 958 88 74
Staj Tecrübeleri:

Proje Sistem Bilgileri

Sistem ve Yazılım: Windows İşletim Sistemi, Python, Google Colab Pro
Gerekli RAM: 16GB
Gerekli Disk: 200GB