# Türkçe Doğal Dil İşleme ile Parafrazlama Sistemi: Transformer Tabanlı Modellerin Uygulama ve Değerlendirilmesi

Danışman: Dr. Öğr. Üyesi Esin Ayşe ZAİMOĞLU

Nefise İMAMNİYAZ
Bilişim Sistemleri Mühendisliği Bölümü
Sakarya Üniversitesi
naifeisa.abudurexiti@oge.sakarya.edu.tr

## Giriş

Parafrazlama, bir cümlenin veya kelimenin anlamı bozulmadan farklı biçimde ifade edilebilmesidir. Doğal dil işleme (NLP) alanının önemli bir araştırma konusudur ve bir metnin anlamını koruyarak farklı biçimlerde yeniden ifade edilmesi sürecini kapsar. Parafrazlama, dilin biçimsel çeşitliliğini desteklemenin yanı sıra metin özetleme, dil öğrenme, otomatik içerik üretimi ve veri artırma gibi pek çok uygulamanın temelini oluşturmaktadır. Bu sistem, dilin bağlamsal ve sözdizimsel özelliklerini dikkate alarak, metinlerin yeniden yapılandırılmasına olanak tanır. Ancak, Türkçe gibi eklemeli bir dilde, bu süreç oldukça zorlu hale gelir. Türkçede kelime türetme ve anlam ilişkilerinin çeşitliliği, NLP sistemlerinin doğruluğunu artırmayı zorlaştırmaktadır.

Bu çalışmada, Türkçe dilinde anlamı koruyarak parafrazlama yapabilen bir sistem geliştirilmiştir. T5 Transformer modelinin kullanıldığı bu sistemde, 10.000 satırlık Türkçe veri kümesi ile model eğitilmiş ve modelin başarısı BLEU ve ROUGE gibi metriklerle değerlendirilmiştir. Bu çalışmanın amacı, Türkçe cümleleri anlamını kaybetmeden farklı biçimlerde ifade edebilen bir sistem geliştirmek ve bu sistemin çeşitli NLP uygulamalarında kullanılabilirliğini göstermektir.

## Literatür

Yapılan literatür taraması sonucunda, Türkiye'de bu alana yönelik yalnızca tek çalışmaya ulaşılmıştır. Söz konusu çalışma Hilal Tekgöz'e ait "Türkçe dilinde eşanlatım oluşturma derlemi ve doğal dil işleme modellerinin karşılaştırılması" isimli Yüksek Lisans Tezidir. İlgili Tez çalışmasında, MSCOCO ve QQR gibi hazır veri setleri Türkçeye uyarlanarak parafraz üretimi gerçekleştirilmiş, çeşitli Transformer tabanlı modeller (T5, BART, Seq2Seq) karşılaştırılmıştır. Bu durum, konunun Türkiye'de henüz sınırlı düzeyde ele alındığını göstermektedir.

Dünya literatüründe ise Parafrazlama, Rahul Bhagat ve Eduard Hovy'nin 2013 yılında yayınladığı 'What Is a Paraphrase?" adlı çalışmada şu şekilde tanımlanmıştır: 'Parafrazlar, aynı anlamı farklı ifadelerle aktaran cümleler veya kelime öbekleridir [2]. Transformer tabanlı (dönüştürücü) modellerin bu alana entegre edilmesi ise 2017 yılında Google tarafından yayınlanan "Attention Is All You Need" çalışmasından sonra başlamış ve bu tarihten itibaren doğal dil işleme alanında büyük bir devrim yaratmıştır [3]. Transformer mimarisi; başlangıçta makine çevirisi, sonrasında parafraz üretimi, görsel erişim ve nefret söylemi tespiti gibi çok sayıda doğal dil işleme (NLP) görevinde yoğun olarak kullanılmıştır. Parafrazlama üzerine yapılan güncel çalışmalar, yalnızca sözdizimsel değil, anlamsal eşdeğerlik odaklı sistemlerin geliştirilmesine odaklanmakta; BERT, GPT, T5 gibi modellerin yüksek başarı sağladığı görülmektedir. Ayrıca, bu çalışmaların birçoğu yüksek kaynaklı diller üzerine yapılmıştır ve Türkçe gibi düşük kaynaklı diller için hâlâ önemli bir açık bulunmaktadır.

## Kullanılan Yöntem

Bu çalışmada, Türkçe parafrazlama sistemlerinin geliştirilmesi amacıyla doğal dil işleme (NLP) tekniklerinden yararlanılmıştır. Transformer tabanlı T5 modeli kullanılmış ve model eğitimi için gereken verilerin tamamı manuel olarak toplanmıştır. Geliştirilen model ROUGE ve BLUE metrikleriyle ölçülmüştür.

## Veri Seti

| Asil | Parafraz |
|---|---|
| Kitap okumak, insanın bilgi dağarcığını arttırır. | Bilgiyi arttırmanın yollarından biri kitap okumaktır. |
| Veri analizi, karar verme süreçlerinde önemli bir araçtır. | Veri analizi, karar alma aşamalarında önemli bir yardımcı olmaktadır. |
| Sınava başarılı olmak için çok çalışmam gerekiyor. | Sınavda başarılı olmak için sınava iyi hazırlanmam gerekiyor. |
| Yüksek öğretim, bireylerin kariyer gelişimine katkıda bulunmaktadır. | Üniversite eğitimi, bireylerin mesleki gelişimlerine önemli ölçüde destek vermektedir. |

*Tablo 1. Veri Seti Örneği*

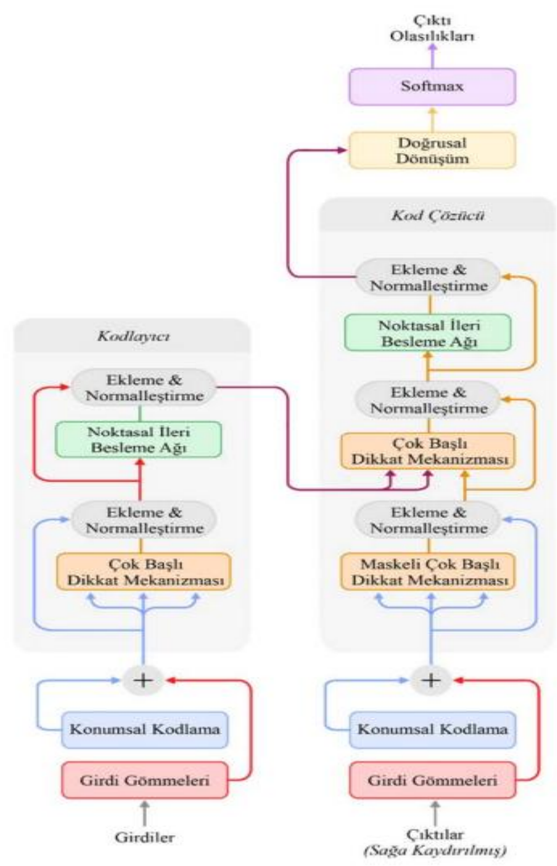
Kullanılan veri setinin tamamı elle oluşturulmuş, kaynak olarak atasözleri, deyimler, günlük konuşma dili, akademik çalışmalardan derlenen metinler ve 18 ayrı dünya klasiğinin alıntıları benimsenmiştir. Oluşturulan veri seti, eğitimden önce veri temizleme, eksik verileri kaldırma, kelimeleri tokenize etme (kelime köklerine ayırma) ve oluşturulan token (kök)leri İD'lerle (benzersiz bir sayısal değerlerle) eşleştirme olmak üzere 4 adımlık kapsamlı bir ön işlemeye tabi tutulmuş ve eğitime hazır hale getirilmiştir.

```
Input Text: paraphrase: Yemekleri zeytinyağı ile pişiriyorum.
Tokens: ['_para', 'phrase', ':', '_Yeme', 'kleri', '_z', 'eytin', 'yağı', '_ile', '_piş', 'ir', 'iyorum', '.']
Token IDs: [435, 59990, 267, 200976, 117763, 397, 151413, 124415, 2222, 37451, 602, 104577, 260]
```

*Şekil 1. Önişleme Yapılmış Veri Seti*

## Transformer Mimarisi

Transformer mimarisi, self-attention (öz-dikkat) mekanizmasını kullanarak her bir kelimenin, metindeki diğer kelimelerle olan ilişkisini eş zamanlı olarak öğrenir. Geleneksel Recurrent Neural Networks (RNN) ve Long Short-Term Memory (LSTM) gibi modellerde, her adım bir önceki adımı dikkate alarak hesaplama yapar; oysa Transformer yapısında tüm giriş dizisi eş zamanlı olarak işlenir. Bu mimari günümüzde GPT, BERT, T5, RoBERTa, DeBERTa gibi pek çok güçlü modelin temelini oluşturmuştur. Bu çalışma kapsamında kullanılan T5 modeli de bu mimariye dayalı olarak geliştirilmiştir.




*Şekil 2. Transformer (dönüştürücü) Mimarisi [4]*

## Model Eğitimi

Model eğitimi için kullanılan parametreler de en az veri ön işleme kadar önemlidir. Bu bağlamda bu çalışmada kullanılan hiperparametreler şu şekildedir;

- Epoch sayısı: 3 olarak belirlenmiştir. Bu, modelin veriye ne kadar derinlemesine öğrenme yapabileceğini belirler.
- Batch Size: 8 olarak belirlenmiştir. Her bir eğitim adımında işlenecek veri miktarıdır.
- Learning Rate: 3e-4 seçilmiştir. Eğitim sırasında modelin ağırlıklarınınn ne kadar güncelleneceğini belirler.
- Weight Decay: 0.01 dir. Modelin overfitting (aşırı öğrenme) yapmasının engellemek bu değer belirlenmiştir.
- Eval Strategy: epoch'tur. Yani modelin doğrulama verisi üzerindeki performansını değerlendirmek için epoch sonunda değerlendirme yapılır.

## Test Aşaması

Modelin eğitimi sonrasında yapılan test çalışmasında modelin öğrenme gerçekleştirdiği gözlemlenmektedir.

```
[ ]  # Test
     print(paraphrase("Film gerçekten çok güzeldi."))
     ['Film gerçekten bana iyi geliyordu.']

[ ]  print(paraphrase("bugün güzel bir gün geçirdim."))
     ['Farklı bir gün geçirdim.']

[ ]  print(paraphrase("Uçakla yolculuk yapmayı seviyorum."))
     ['Uçakla yolculuk yapmak hoşuma gidiyor.']

[ ]  print(paraphrase("kitap okumak insanın bilgisini arttırır."))
     ['Kitap okumak insanın bilgisini artırma yolunu sağlar.']
```

*Şekil 3. Modelin Test Edilmesi*

## Sonuç

Çalışmanın sonucunda modelin performansı beklendiği gibi ROUGE skoru 0.46 ve BLUE skoru ise 28.16 olarak sonuç vermiştir. Bu sonuçlar, modelin Türkçe cümleler üzerinde anlam kaybı olmadan parafrazlama yapabilme kapasitesine sahip olduğunu ortaya koymaktadır.

Çalışmanın en önemli katkılarından biri, Türkçe dilindeki parafrazlama görevini yerine getirebilen etkili bir modelin geliştirilmesi ve bu modelin doğal dil işleme alanındaki yeni bir perspektifi ortaya koymasıdır. Ayrıca, modelin Türkçe metinler üzerinde başarı elde etmesi, özellikle eklemeli yapıya sahip Türkçe için başarılı sonuçlar sunduğunu göstermektedir.