ArXiv Sitesindeki Makaleleri Kullanarak Başlık Oluşturma

Arman Kuyucu[1], Recep Aydın[2], Buğra Burak Önal[3], Samet Yavuz[4], Recep Kaya[5]

[1,2,3,4,5] *Bilgisayar Mühendisliği, Kocaeli Üniversitesi*
Kabaoğlu Mahallesi, Baki Komsuoğlu Bulvarı No:515, Umuttepe, 41001 İzmit/Kocaeli

[1]190201099@kocaeli.edu.tr
[2]200202093@kocaeli.edu.tr
[3]190201034@kocaeli.edu.tr
[4]190201066@kocaeli.edu.tr
[5]190201027@kocaeli.edu.tr

***Abstract*— Doğal Dil İşleme(Natural Language Processing) insanların birbirleri aralarında konuşmak için kullandıkları dili, bilgisayar-insan etkileşiminde kullanmak için yapılan uygulamaları inceleyen bir yapay zeka alt alanıdır. Başlık oluşturma (Title Generator), metinlerin içeriğini analiz ederek otomatik olarak uygun başlıklar oluşturan bir sistemdir. ArXiv içerisindeki makalelerin içeriğini ve başlıklarını barındıran etiketli bir veri seti kullanılmıştır. Çalışmada çok bilinen bir doğal dil işleme modeli olan BART yapay zeka dil modeli ve bu modelin eğittiğimiz hali kullanılarak başlık oluşturma işlemi yapılmıştır. Deneysel Sonuçlara bakılacak olursa en yüksek %36.39 RougeLSum skoru ve %9.81 BLEU skoru elde edilmiştir.**

***Keywords*— Başlık Oluşturma, Doğal Dil İşleme, Yapay Zeka, Dil Modeli, ROUGE, BLEU.**

I. Giriş

Metindeki anahtar kelimeleri, cümleleri veya önemli bilgileri tanımlamak için dilbilgisi, kelime dağarcığı, anlam çıkarımı ve bağlam anlayışı gibi NLP tekniklerini kullanıp, başlık oluşturma işlemi gerçekleştirilir.

Başlık oluşturucunun amacı, metnin anlamını özetleyen ve okuyucunun dikkatini çeken etkili bir başlık oluşturmaktır. Bu başlıklar genellikle haber makaleleri, blog yazıları, sosyal medya paylaşımları ve web sayfaları gibi içerikler için kullanılır. Başlık oluşturucular, içeriğin özünü yansıtmak için anahtar noktaları seçerken aynı zamanda okuyucunun ilgisini çekmeyi hedefler.

Başlık oluşturucular, metinlerin hızla büyüyen hacmiyle başa çıkmak ve içerik oluşturuculara zaman kazandırmak için önemli bir araç haline gelmiştir. Ayrıca haber ajansları, yayıncılar ve içerik pazarlamacılar gibi profesyoneller tarafından da kullanılır.

Kullanılan veri seti[1] 'input_text' ve 'target_text' adındaki sütunlardan ve 19.973 ingilizce örnekten oluşmaktadır. Veri setinin içeriği şekil 1'de verilmiştir.

	input_text	target_text
0	We present in this article a detailed quanti...	A unified analysis of the reactor neutrino pro...
1	The aim of the KArlsruhe TRItium Neutrino ex...	The KATRIN sensitivity to the neutrino mass an...
2	In this letter, we propose three different s...	Penguin-mediated B_(d,s)->VV decays and the Bs...
3	We find non-supersymmetric metastable vacua ...	Non Supersymmetric Metastable Vacua in N=2 SYM...
4	We show how to test whether a graph with n v...	Recognizing Partial Cubes in Quadratic Time
...	...	...
19968	Schroedinger (Nature, v.169, p.538 (1952)) d...	Real-Valued Charged Fields and Interpretation ...
19969	Given a collection of states (rho_1, ..., rh...	How many copies are needed for state discrimin...
19970	Hu, Paz and Zhang [B.L. Hu, J.P. Paz and Y....	Alternative Derivation of the Hu-Paz-Zhang Mas...
19971	We demonstrate guiding of cold neutral atoms...	Guiding Neutral Atoms with a Wire
19972	We show that {\it any} entanglement measure ...	Limits for entanglement measures

Şekil 1. Veri Setinin İçeriği

Projede çok bilinen bir dil modeli olan BART dil modeli ve bu modelin eğittiğimiz hali kullanılmıştır.

II. Yöntem

A. BART Dil Modeli

Etkisiz kelimeleri elde etmek için NLTK kütüphanesi kullanılmıştır. BART (Bidirectional and AutoRegressive Transformers), dil anlama ve doğal dil işleme alanında kullanılan bir yapay zeka modelidir. BART, OpenAI tarafından geliştirilmiştir ve GPT-3 modelinin bir uzantısı olarak düşünülebilir.

BART, otomatik kodlama, makine çevirisi, özetleme, metin düzeltme ve dil oluşturma gibi çeşitli görevleri gerçekleştirebilir. Hem doyurucu öğrenme (unsupervised learning) hem de gözetimli öğrenme (supervised learning) için kullanılabilir.

Model, bir transformatör (transformer) mimarisi kullanır. Transformatörler, özellikle doğal dil işleme görevleri için son derece etkili olan bir derin öğrenme modelidir. BART, kodlama ve kod çözme (decoding) aşamalarında özerk olarak çalışabilen bir yapıdadır. İlk olarak, girdi metni "kodlama" aşamasında bir dilin temsili olarak anlaşılır hale getirilir. Ardından, metin "çözme" aşamasında başka bir dile çevrilebilir, özetlenebilir veya yeniden düzenlenebilir.

BART'ın "bidirectional" (iki yönlü) özelliği, hem giriş hem de çıkış metnini dikkate alarak anlamayı ve üretmeyi öğrenmesi anlamına gelir. Bu, BART'ın metinleri okurken veya yazarken bağlamı anlamasına yardımcı olur. Bu, çeviri veya özetleme gibi görevlerde daha iyi sonuçlar elde edilmesini sağlar.

Ayrıca, BART modeli "autoregressive" (özyinelemeli) bir yapıya sahiptir. Bu, modelin çıktıyı üretirken önceki çıktıları kullanmasını sağlar. Örneğin, bir metin parçası üretirken, BART, daha önce üretilen kelime ve cümleleri dikkate alır. Bu, daha akıcı ve tutarlı çıktılar elde edilmesini sağlar.

BART, büyük bir önceden eğitim aşamasından geçer. Büyük miktarda metin verisiyle beslenir ve dilin yapısını ve ilişkilerini öğrenir. Bu önceden eğitim, BART'ın genel dil anlama yeteneklerini geliştirmesine yardımcı olur. Ardından, gözetimli öğrenme aşamasında belirli bir görev için etiketli veriler kullanılarak modelin özelleştirilmesi sağlanır.

B. Veri Önişleme

Veri bilimi projelerinde sahip olunan veriler düzensiz halde bulunabilir, bozuk kayıtlar içerebilir ya da uygulayacağınız analize uygun halde bulunmayabilirler. Dahası daha iyi sonuçlar elde edebilmek üzere çeşitli filtrelerden geçirmek, anomalilerden işin başında kurtulmak istenilebilir. Bu ve benzeri birçok duruma müdahaleyi içeren yöntemler bütününe veri ön işleme denir. Veri ön işleme, bir veri kümesini temizleme, düzenleme ve hazırlama işlemlerini içerir. Bu adımlar, veri kümesinin kalitesini artırır, modelin daha iyi sonuçlar üretmesini sağlar.

Öncelikle, 'abstracts' ve 'titles' adlı iki liste verilir. 'abstracts' listesi, her biri bir araştırma makalesinin özetini içeren bir dizi string içerirken, 'titles' listesi, her biri bir makalenin başlığını içeren bir dizi string içerir.

Bu iki liste kullanılarak, pandas kütüphanesindeki DataFrame sınıfı kullanılarak bir veri çerçevesi olan 'papers' oluşturulur. Bu veri çerçevesinin iki sütunu vardır: 'input_text' ve 'target_text'. 'input_text' sütunu, 'abstracts' listesindeki her bir özet string'ini içerir ve 'target_text' sütunu, 'titles' listesindeki her bir başlık string'ini içerir.

Daha sonra, 'nan_count' adlı bir değişken oluşturulur ve her bir sütundaki eksik değerlerin sayısını içerir. Bu değişken, özellikle veri setindeki eksik verilerin sayısını belirlemek için kullanılır.

Sonrasında, 'papers' veri setindeki tüm satırların eksik değer içerip içermediğini kontrol etmek için 'isna().sum()' fonksiyonu kullanılarak 'nan_count' değişkenine atama yapılır. Eksik değerlerin olduğu tespit edilirse, bu satırlar 'dropna()' fonksiyonu kullanılarak veri setinden çıkarılır.

Son olarak, 'titles' ve 'abstracts' listeleri bellekten silinir ve 'papers' veri seti ile devam edilir.

Tokenizasyon işlemi ise, metin verilerindeki kelimeleri belirli bir sözlük yapısına dönüştürmek için kullanılan bir işlemdir. Bu işlem genellikle makine öğrenimi modellerinde kullanılmadan önce gerçekleştirilir. Tokenizer, bu işlemi gerçekleştiren bir araçtır ve metinleri verilen bir sözlük yapısına göre ayırır.

C. K-Fold Çapraz Doğrulama

K-Fold Çapraz Doğrulama, sınıflandırma modellerinin değerlendirilmesi ve modelin eğitilmesi için veri setini parçalara ayırma yöntemlerinden biridir. Veriyi belirlenen bir K sayısına göre eşit parçalara böler, her bir parçanın hem eğitim hem de test için kullanılmasını sağlar, böylelikle dağılım ve parçalanmadan kaynaklanan sapma ve hataları asgariye indirir. Ancak modeli k kadar eğitmek ve test etmek gibi ilave bir veri işleme yük ve zamanı ister. Bu durum eğitim ve testi kısa süren küçük ve orta hacimli veriler için sorun olmasa da büyük hacimli veri setlerinde hesaplama ve zaman yönünden maliyetli olabilir [2].

Bu projede, kullanılan dil modelinin eğitilmesinin çok uzun sürmesi ve çapraz doğrulamayı yapacak yeteri kadar zamanın olmamasından dolayı çapraz doğrulama yapılamamıştır.

D. Hiper Parametre Ayarlama

Makine öğrenmesinde hiper parametre ayarlaması çok önemlidir. Daha efektif ve başarılı sonucu alabilmek için hiper parametre ayarlaması yapılmıştır.

Ayarlama için özel bir fonksiyon kullanılmamıştır. Bunun başlıca sebebi modelin eğitiminin uzun sürmesidir. Hiper parametre ayarlaması yapmak için modelin parametrelerinden en önemli olan 3 tanesi belirlenmiştir ve bunların model üzerindeki etkilerinin incelenmesi hedeflenmektedir.

BART modelinin en önemli parametreleri "max_seq_length", "train_batch_size" ve "num_train_epochs" olarak belirlenmiştir.

E. Değerlendirme Ölçütleri

Projede yapılan işlemlerin sonucunu değerlendirmek için sıkça kullanılan ROUGE ve BLEU değerlendirme ölçütleri kullanılmıştır.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metriği, doğal dil işleme alanında otomatik özetleme sistemlerinin performansını değerlendirmek için kullanılan bir metriktir. ROUGE metriği, otomatik olarak oluşturulan bir özetin referans (gerçek) özetle karşılaştırılmasını sağlar.

ROUGE metriği, özetleme performansını değerlendirmek için farklı seviyelerde kullanılabilir. Örneğin, Rouge-1, metnin tek kelimelik bir birleşimini karşılaştırırken, Rouge-2 metnin

iki kelimelik bir birleşimini karşılaştırır. RougeL ise en uzun ortak alt dize (longest common subsequence) kullanarak kelime seviyesinde karşılaştırma yapar. RougeLSum metriği, Rouge-1, Rouge-2 ve RougeL metriklerini içerir. Bu metriklerin kullanımıyla, özetin doğruluğu, kapsamlılığı ve bütünlüğü değerlendirilebilir.

BLEU (Bilingual Evaluation Understudy), bir çeviri sistemine verilen bir metnin ne kadar iyi çevrildiğini tahmin etmeye çalışır. Ölçü, çeviri sisteminin çıktısının, insan yapımı referans çevirileriyle ne kadar benzerlik gösterdiğini hesaplar. Daha yüksek bir BLEU skoru, çevirinin referans çevirilere daha yakın olduğunu gösterir.

BLEU, çeviri sisteminin n-gram eşleme oranlarını, kelime düzeyinde doğruluğu ve diğer dil düzeyinde ölçümleri kullanarak çalışır. Genellikle 1-gram (tekli kelimeler), 2-gram (ikili kelimeler), 3-gram (üçlü kelimeler) ve 4-gram (dörtlü kelimeler) olmak üzere çeşitli n-gramlar kullanılır. Bu çalışmada 4-gram kullanılarak BLEU değerleri hesaplanmıştır.

III. Deneysel Sonuçlar

Metinlerden başlık oluşturma işlemi için BART dil modeli ve bu modelin eğittiğimiz hali test edilmiştir.

Yapılan deneyler sonucunda elde edilen en iyi Rouge-1, Rouge-2, RougeL, RougeLSum, BLEU skorları Tablo 1'de verilmiştir.

Tabloya göre en iyi RougeLSum skoru %36.3909 ve %9.769 BLEU skoru ile "max_seq_length": 512, "train_batch_size": 4, "num_train_epochs": 5 parametrelerine sahip model ile elde edilmiştir. Bu sonuç hiper parametre ayarlaması yapılarak elde edilmiştir.

TABLO I
Modelin Sonuçları

Parametreler	Rouge-1	Rouge-2	RougeL	RougeLSum	BLEU
Varsayılan Parametreler	%24.7378	%10.7278	%21.7826	%22.1096	%4.6328
"max_seq_length": 16, "train_batch_size": 2, "num_train_epochs": 1	%25.5269	%10.7748	%23.0571	%23.3697	%4.4185
"max_seq_length": 64, "train_batch_size": 2, "num_train_epochs": 3	%34.5628	%17.1034	%31.1576	%31.7222	%7.8716
"max_seq_length": 64, "train_batch_size": 4, "num_train_epochs": 4	%34.1303	%16.6208	%30.8134	%31.4566	%7.7807
"max_seq_length": 64, "train_batch_size": 4, "num_train_epochs": 5	%34.0331	%16.8772	%30.6872	%31.2761	%8.0967
"max_seq_length": 256, "train_batch_size": 3, "num_train_epochs": 4	%39.5686	%20.6963	%35.2530	%36.2472	%9.4179
"max_seq_length": 256, "train_batch_size": 16, "num_train_epochs": 4	%39.2945	%20.8772	%35.6195	%36.3879	%9.8086
"max_seq_length": 512, "train_batch_size": 4, "num_train_epochs": 5	%39.6016	%20.7102	%35.5966	%36.3909	%9.769

IV. Sonuçlar

Bu çalışmada, etiketli ArXiv sitesindeki makaleleri içeren bir veri seti üzerinde dil modeli kullanılarak başlık oluşturma işlemi gerçekleştirilmiştir. Metinlerden başlık oluşturma işlemi için BART dil modeli ve bu modelin eğittiğimiz hali kullanılarak yapılmıştır. Deneysel sonuçlara bakılacak olursa hiper parametre ayarlaması yapıldığında tüm algoritmaların sonuçlarında iyileşme görülmüştür.

Kaynaklar

[1] https://www.kaggle.com/datasets/Cornell-University/arxiv/versions/18 (Erişim Zamanı : 7 Mayıs 2023)

[2] https://www.veribilimiokulu.com/bir-bakista-k-fold-cross-validation/ (Erişim Zamanı : 7 Mayıs 2023)