Doğal Dil İşleme ile Haber Başlığı Üretme

Ş. Gülen Keçeli ve Pervin Mine Gökşen
Bilgisayar Mühendisliği Bölümü
TOBB Ekonomi ve Teknoloji Üniversitesi
{s.keceli, pgoksen}@etu.edu.tr

Abstract

Bu çalışmanın amacı, Doğal Dil İşleme (NLP) teknolojisini kullanarak haber başlıkları oluşturmak için Haber Başlık Oluşturucu NLP projesini tanıtmak ve değerlendirmektir. Bu makalede, haber metinlerindeki her cümlenin önem düzeyi çıkarılan öznitelikler'e makine öğrenmesi yöntemleri uygulanarak belirlenmiştir. Bu cümle, metnin temel anlamını yansıtan bir başlık oluşturmak için kullanılmıştır.

1 Giriş

Haber başlıkları, okuyucuların haber makalelerine ilgi göstermesine yardımcı olur. Bugünün hızlı tempolu dünyasında, başlıklar okuyucuların dikkatini çekme ve bir haber hikayesinin esasını iletmek için kritik bir rol oynar. Basılı gazetelerden online haber websitesine kadar, başlıklar okuyucunun ve içerik arasındaki ilk temas noktasıdır ve bilgi kapısının görevini yaparlar. Ancak, etkileyici başlıklar üretmek her zaman kolay değildir ve genellikle önemli denilebilecek bir miktarda zaman ve çaba gerektirir. Haber endüstrisinde, haber başlıklarını oluşturmak için genellikle profesyonel yazarlar ve editörlerden yardım alınmaktadır. Bununla birlikte, bu süreçte yapılan hatalar ve yanlış anlamalar okuyuculara yanlış bilgi sunabilir ve haber makalelerinin etkililiğini azaltabilir. Bu noktada Doğal Dil İşleme (NLP) Teknikleri devreye girmektedir.

Daha önceki çalışmalarda, derin öğrenme tabanlı yöntemler, metinleri özetlemek ve anlamlı başlıklar üretmek için etkili bir şekilde kullanılmıştır. Ancak bu yaklaşımların bazı dezavantajları bulunmaktadır. Derin öğrenme modelleri genellikle büyük miktarda veri ve hesaplama gücü gerektirir. Ayrıca başlık üretme kategorisinde birçok çalışma olmasına rağmen(Xiaotao Gu, 2020)(Ruqing Zhang, 2020) Türkçe başlık üretme problemi için yapılan çalışmalar sınırlıdır.

Biz bu çalışmada, haber metinlerinden başlık oluşturmak için metindeki önemli cümleye odaklanan bir yaklaşım benimsedik. Öncelikle, metindeki cümlelerin önem düzeylerini belirlemek için bir yöntem geliştirildi. Öznitelik olarak cümlenin içerdiği Named Entity Recognition sayısı,TF-IDF değeri yüksek olan kelime sayısı, isim tamlaması sayısı, cümle uzunluğu ve metindeki konumu verilmiştir. Ardından cümlenin önemi bu özniteliklerle eğitilerek cümlenin başlık ile benzerliği öğretilmiştir. Bu önemli cümleden yola çıkarak başlık oluşturmak için dependancy parser kullanıldı. Bu yaklaşımın avantajı, haber metninin içeriğini kapsayan ve de bilgi içeren başlıklar üretmemize yardımcı olmuştur.Yapılan deneylerde, geliştirilen yöntemin başarı oranı ve oluşturulan başlıkların kalitesi gibi başarım ölçütleri kullanılarak değerlendirilmiştir. Elde edilen sonuçlar, geliştirilen yöntemin istenildiği kadar etkili çalışmasa da başlık üretirken elle tutulur sonuçlare gözlemlenmiş, ve Baseline metodu ile yakın başarımlar alınmıştır.

2 İlgili Çalışmalar

Literatür taramamızda metin başlığı oluşturma ve başlık oluşturmaya benzer bir yöntem olan metin özetleme ile ilgili araştırmalardan bahsedeceğiz.

2.1 Otomatik Metin Başlık Oluşturma

2.1.1 Derin Öğrenme Yöntemleri

Gu makalesinde (Xiaotao Gu, 2020), haber başlığı oluşturmak için makine öğrenmesi ve yapay sinir ağı kullanılmıştır. BERT pre-training kullanarak veri seti oluşturur. NHNET MODEL'ini kullanarak haber başlığı oluşturulmuştur ve kendi kendine oylama tabanlı sistem kullanarak

ağırlıklar belirlenmiştir. Kedia, Mantha, Guo ve Achan ise(Mansi Ranjit Mane, 2020) BERT yanı sıra seq2seq + Attention, Ptr-Net ve Transformer modelleriyle haber başlığı oluşturulmuştur. (Ruqing Zhang, 2018)Seq2seq modeline modeline benzer olarak dikkat mekanizması ve çift yönlü dikkat mekanizması kullanan DASeq2Seq modeli kullanmışlardır. Başka bir yaklaşım (Ruqing Zhang, 2020) SLGen modelidir. Fakat NHNet, özetleme yapmak için bir hiyerarşik kodlayıcı-decoder yaklaşımı kullanırken, SLGen graf tabanlı sinir ağı kullanır. Aynı şekilde Encoder-Decoder kullanan PENS (Xiang Ao, 2021) başlık tahmini üzerine kişiselleştirme de eklemişler. Click-bait tık tuzağı olarak adlandırılan okuyucuların dikkatlerini çekerek içeriğe gitmelerini sağlama oranına vurgu yapar. Li , Wu and Miao (Zhengpeng Li, 2022) çözüme PENS ile benzer şekilde yaklaşırken,problemi genel olarak ele alarak İngilizce haberlere haber başlığı üretmeyi işlemişlerdir. Kedia, Mantha, Guo ve Achan ise(Mansi Ranjit Mane, 2020) çözüm modellerine hiyerarşik graf tabanlı sinir ağı kullanan TD-NHG modeli kullanmışlardır.

2.1.2 İstatiksel Yaklaşımlı Yöntemler

Sethi, Agrawal, Madaan, Singh ve Kumar (Nandini Sethi and Kumar, 2016) içeriklere ilk önce dil analizi için POS Tagging , söylem analizi, token sıklığı teknikleri uygulamışlardır ve birden fazla başlık önerisinde bulunmuştur. Öneriler için tokenların sıklıklarına bakılarak, sıfat isim öbeğinden içeren başlıklar ve bizim de çalışmamızda deneyini yaptığımız atasözü içeren cümlelerde atasözü olan başlıklar olmak üzere 3 çıktı veren bir modeldir. Shao ve Wang(Shao and Wang, 2017), başlık oluşturmak için DTATG yöntemini kullanır. Merkezi cümleler belirlenir ve dependancy tree oluşturulur. Başlık adayları WCO,iki veya daha fazla kelimenin aynı bağlamda birlikte geçme sıklığı, ve RAKE,önem skorlarına göre anahtar kelimeleri sıralama ve en önemli olanları seçme, algoritmalarıyla bulunur. Xu, Yang ve Lau(Songhua Xu, 2010) çalışmalarında, bir belgenin türünün etkisini dikkate alarak, keyword çıkarma ve başlık oluşturma için Wikipedia'dan türetilen yeni kelime özellikleri kullanılmaktadır. Bu özellikler arkaplan bilgisi, link, kategori ve infobox bilgilerini içerir. Kelimenin anahtar kelime olduğunu bulmak için SVM kullanır. Zajic ve Bonnie (Schwartz et al., 2002) de bizim çözüm yöntemimizle benzer olarak metin içindeki kelimelerden oluşan bir haber başlığı önerisinde bulunma olarak soruna yaklaşmışlardır. Çözüm olarak daha önce yazım kontrolü, POS Tagging , dil tanımlama ve özetleme gibi alanlarda kullanılmış Noisy Channel Model yöntemi metinden kelime seçme için kullanılmıştır. Üretme kısmında ise Hidden markov model kullanılmıştır.

2.2 Otomatik Metin Özetleme

2.2.1 Derin Öğrenme Yöntemleri

(Kaikhah, 2004) derin öğrenme için başlık, paragraf konumu, cümle konumu, paragrafın ilk cümlesi, cümle uzunluğu, tematik kelime sayısı, başlıktaki kelime sayısı öznitelikleri çıkarılmıştır. Neural Networka cümlenin özette olup olmayacağı öğretilmiştir.

2.2.2 İstatiksel Yaklaşımlı Yöntemler

Kutlu , Cığır ve Çiçekli (Kutlu et al., 2010) bizim çözüm önerimize benzer olarak metin özetleme için metni cümlelere ayırıp, kendi elde ettikleri bir skora göre sıralayarak özet çıkartmışlardır. Kulnarni ve Apte (Kulkarni and Apte, 2013) İngilizce metin özetleme için preproccessing, fuzzification, rule base , defuzzification , cümle seçimi ve montaj ve özet oluşturma aşamalarından geçiyor. Fuzzification verilerdeki belirsizliği, defuzzification ise çıktılardaki belirsizliği azaltmak için matematiksel bir modele çevirir. İki makalede de öznitelik bazında başlık ile metin benzerliği, cümlenin metindeki konumu, cümlelerin benzerlikleri ve konuyu, metni kapsayan kelimelerin geçme sıklığını alırken (Kutlu et al., 2010) yapay öğrenme ile öznitelik ağırlıklarını bulabilmek için eğitmişlerdir. Çiçekli, Ozsoy ve Alpaslan (**?**) Türkçe metinlerin özetlenmesinde cümle benzerliğine odaklanmış olup bunun için kelimelerin ve cümlelerin anlam yapısını çıkarmak için Latent Semantic Analysis, cümleler arası ilişkilerini çıkarmak için de SVD algebra metodu kullanılmışlar. Meru Brunn,Yllias Chali ve Christopher J. Pinchak (Brunn et al., 2002); metin özetleme için isim filtreleme, keyword çıkarılması, WordNet kullanımı ile zincir oluşturma ve cümle puanlaması ile özet oluşturmuşlardır. Contoy ve OLeary(Conroy and O'leary, 2001) ise metin özetleme probleminde 3 özniteliği baz alarak yine HMM model ile metinde olma olasılığı hesaplamışlardır. (Kutlu et al., 2010)(Kulka-

rni and Apte, 2013) makalelerine benzer olarak cümlenin sırası, cümledeki term sayısı, ve kelimelerin özette olma olasılıkları özniteliktir. (Yavuz Selim Kartal, 2020) makine öğrenmesi kullanarak seçilen cümlelerden özetleme yapan bir model geliştirilmişlerdir. Öznitelik olarak bizim de kullandığımız kelime sıklığı, NER, konum kullanmış olup ek olarak konuşma ifadeleri başlık benzerliği kullanmışlardır. Hovy ve Lin(Eduard Hovy, 1999), SUMMARIST adlı metin özetleme sistemini tanımlamıştır. Bu sistem, konu belirleme, yorumlama ve üretme işlemlerini içermektedir. Özetleme işlemi sırasında, belirleyici terimler kullanılarak cümleler sıralanır ve anlamlı bir özet oluşturulur. (Ferreira et al., 2014) cümle puanlama için kelime bazlı, cümle bazlı ve diyagram bazlı puanlanmıştır. Kelime bazlı puanlama için kelime sıklığı, tfidf, kelimelerin birlikte geçme sıklığı, lexical benzerlik ,büyük harf ,özel isim sayısı bakılmıştır. Cümle bazlı skor hesabı için işaret ifadeleri, cümlenin konumu, başlık ile benzerliği, diğer cümlelerle aynı kelimeleri içermesi,uzunluğu ve sayısal veri içermesine bakılmıştır. Diyagram bazlı puanlarken Text Rank, Bushy Path of the Node ve aggregrate similarity değerleri ele alınmıştır. Bu makaleye benzer olarak Bushy Path of Node kullanan (Yeh et al., 2005) 2 yöntem önermişlerdir. İlk yöntemde cümlelerin sırası, pozitif/negatif kelimeler içermesi, başlık ile benzerliği ve diğer cümlelere benzerliği öznitelik olarak alınmış ve genetic algoritma ile eğitilip cümle skorları hesaplanmıştır. İkinci yöntemde ise bir cümlenin anlam gösterimini çıkarmak için Latent semantic analysis, cümlenin metinlerle anlam ilişkisini çıkarmak için de Text Relation Map kullanmışlardır. Sonrasında ise cümlelerin bushinesslarına göre, yani cümleden çıkan anlam linklerinin sayısı yerine linklerin ağırlıklarının toplandığı bir önem hesabı ile cümle seçilmiştir.

3 Önerilen Yöntem

Çalışma kapsamında hedefimiz, makine öğrenimi tekniklerini kullanarak Türkçe haber metinlerinden uygun cümleleri seçerek bir başlık oluşturmaktır. Başlık önerisi için haber metinleri cümlelere ve her cümle kelimelere ayrılıp önişlemlere tabi tutuldu(BÖLÜM 3-3.1). Önişlemden geçmiş olan cümleler için model eğitiminde kullanacağımız öznitelikler oluşturuldu(BÖLÜM 3-3.2). Özniteliklerle birlikte makine öğrenmesinde kullanacağımız etiketleri elimizde bulunan verisetinden başlık ve cümlenin benzerliğini kullanarak oluştruduk(BÖLÜM 3-3.3). Etiketlenen cümleler ile özniteliklerimizi Rassal Orman (Random Forest) ve Lineer Regression (LR) makine öğrenmesi modeli ile eğittik. Bu sayede haber metninde bulunan cümleler için en önemli cümle öğrenilmiş oldu. En önemli cümle ile dependancy parsing yöntemi önerilen başlık oluşturulmuştur. Bu aşamada dependancy parsing ile başlık elde edilmiştir. Dependancy parsing ile elde edilen örnek sonuç şu şekildedir:

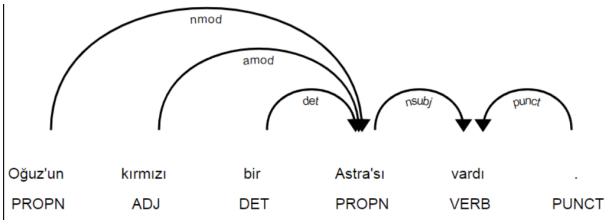

Figure 1: Dependancy Parsing

Başlık oluşturulurken deneysel gösterilemeyen yöntemler uygulanıp en iyi sonucu elde ettiğimiz şu adımları uyguladık. İlk olarak root(fiil) başlığa ekleniyor. Daha sonra roota bağlı olan özneyi ve eğer fiil birleşik fiil ise yani compound olarak etikeketlenmiş kelimeyi ve fiilden hemen önce bağlaç var ise onu da başlığa ekliyorum. Başlığa eklenen kelimelere bağlı isim veya sıfat tamlamalarını ve birleşik kelimeleri de ekledikten sonra bunlara bağlı yan cümle ve numeric sayı bilgisini de başlığıma ekleyip nihai başlığımı elde etmiş bulunuyorum.

3.1 Ön İşlem

Elimizdeki veri setini sadece metin metni ve başlık bilgisinden oluşması üzerine temizledik. Daha sonra metni cümlelerine ve her cümleyi kelimlerine ayırıp 2 farklı şekilde kaydettik. Bu işlemleri yaparken hangi habere ait olduğu bilgisini kaybetmemek için haber numarası atadık. Noktalama işaretlerini hatalara sebep olduğundan dolayı çıkarttık. hem veri setini hem de eldeki veri tokenlarına ayrıldıktan sonra zembereğin zeyrek aracı ile (Akın and Akın, 2007) lemmatization işlemine tabi tuttuk. Son olarak tüm kelimeler VNLP(vnl) kullanılarak stop wordslardan arındırıldı.

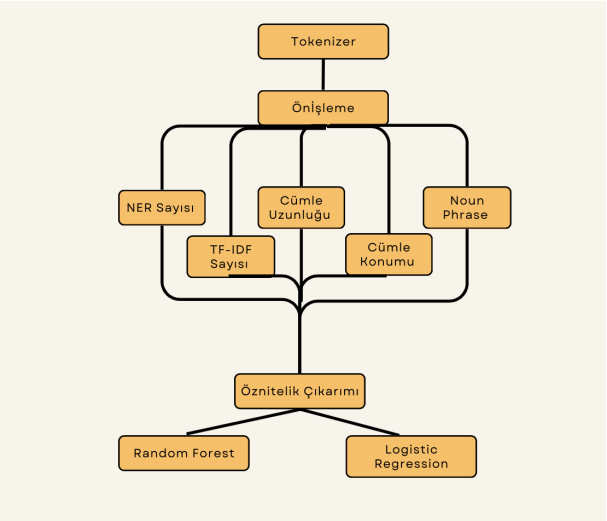

Figure 2: Diagram

3.2 Öznitelik Çıkarımı

Model eğitiminde kullanılacak özniteliklerin şu şekildedir:

- NER (Named Entity Recognition) metin içindeki özel olarak belirtilmiş kurum, kuruluş, yer, özel isim, tarih para birimleri gibi pek çok adlandırılmış isimleri bulup sınıflandırır. NER tespiti için Zemberek-NLP'nin(Akın and Akın, 2007) sağladığı veri seti ve sağladıkları basit perceptron model kullanılarak eğitilerek kullanılmıştır. Cümlede bulunan NER sayısı arttıkça cümlenin metin içindeki önemide artmaktadır. Metnin her cümlesinde bulunan NER sayısı öznitelik olarak verilmek üzere sayılmıştır.

- Noun Phrase, metin içersindeki isim tamlamamlarıdır. İlgili Çalışmalar başlığı altında da bahsettiğimiz (Nandini Sethi and Kumar, 2016) gibi isim tamlamaları başlık üretmede kullanılan özniteliklerden biridir.Bu sebeple her cümlede geçen isim tamlaması sayısı birer öznitelik olarak belirlenmiştir. Noun phrase detection için dependancy parsing yöntemi kullanılmıştır.

- Bir diğer feature olarak metinin bütünüyle TF-IDF skorları hesaplanmıştır. TF-IDF ise bir kelimenin belirli bir haber metinde ne kadar önemli olduğunu belirlemek için kullanılan bir ölçüttür. TF-IDF, bir kelimenin TF (term frequency) değerini ve IDF (inverse document frequency) değerini çarpıp elde edilir. Bu sayede, nadir kullanılan ancak belirli bir metinde sıkça geçen kelimeler, daha yüksek bir TF-IDF değerine sahip olabilir. Bu sayede kelimenin metinde ne kadar önemli olduğu belirlenmiştir. Haber metninde her cümledeki keliemeler için bir TF-IDF skoru hesaplanmıştır. Belli bir thresholdun üzerinde TF-IDF skoruna sahip olan kelime sayısı öznitelik olarak verilmiştir.

- Uzun cümleler metin hakkında genellikle daha fazla bilgi içerirken, kısa cümlelerin daha spesifik olabildiği gözlemlenmiştir. Cümle uzunluğu, metindeki bilgi düzeyini yansıtmaktadır. Bu sebeple cümle uzunluğu bir öznitelik olarak eklenmiştir.

- Bir cümlenin konumu, onun metindeki diğer cümlelerle ilişkisini gösterebilir. Metnin ilk cümlesi genellikle genel bir giriş sağlar ve metnin son cümlesi genellikle sonuç veya özeti verir. Bu durumda cümlenin konumu, cümlenin başında veya sonunda olsun, cümlenin genel anlamı hakkında ipuçları verdiği anlamına gelmektedir. Bu sebeple modelimize vereceğimiz son öznitelik olarak cümlenin metindeki konumunun bilgisi seçilmiştir.

3.3 Veri Etiketlemesi

Veri etiketlenmesi için daha önceden eğitilmiş bir BERT modeli kullanılmıştır. Transformer modelinden karşılaştırılacak iki cümlelerin gömülü halleri alır, daha sonra cosine similarity ile cümleler arasında ne kadar anlamsal benzerlik olduğunu belirtmek için skor hesaplanması yapılmaktadır.(sen)

4 Deneyler

4.1 Deney Düzeneği

Veri Seti: Projede üzerinde analiz ve test yapabilmek üzere, zaman kısıtı sebebiyle, MLSUM olarak adlandırılan birden fazla dilde haber metinlerine, özetlerine ve de başlıklarına yer verildiği bir veri kümesi tercih edilmiştir. Non-commercial araştırma amacıyla kullanılmak üzere TDD (Turkish Data Depository) aracılığıyla kullanıma sunulan ve de 5 farklı dilde haber metinleri içeren büyük bir veri kümesi olan MLSUM'dan (ver) biz Türkçe haber metinlerinin ayrıştırıldığı haber metinlerini, metinlerin başlıklarını, özetlerin, ve de metinlerin

çekildiği site URL'lerini içeren hazır dataseti kullanacağız. Detayları **Veri İstatisiği** tablosunda görülmektedir.

Dataset Statistics

trsum	Cleaned
Avg. article length	258.4
Avg. summary length	18.3
Splits	
Training	246490
Validation	10852
Test	11897
Total	269239

Figure 3: Veri İstatistiği

Kurulumlar: Dependancy Parsing, Stop wordlerin atılması ve Stemming yapmak için (vnl) kurulumu yapılıp kullanılmıştır. NER için (Akın and Akın, 2007) ile kurulumu gerçeklenmiştir. Zembereği çalıştırmak için java'yı pythonda jpype kullanılmıştır. Makine öğretim modellerinde (sci) kütüphanesi kullanımıştır.

Parametre Ayarlama: Deneylerimizde makine ögrenmesi algoritmaları olarak Logistic Regression (LR) ve Random Forest (RF) modellier Scikit-Learn kütüphanesi(sci) kullanılarak modellenmiştir. Bu modeller farklı parametrelerle çalıştırılmış ve bunun sonucunda RF'de **n_estimators** değeri olarak 600 ağaç kullanılmıştır. LR için parametreler varsayılandır.

Baseline: Haber başlığı ya da başlık üretmenin otomatik metin özetleme problemi ile benzerliği olduğunu ve de çözüm yöntemlerinin benzerliğini gözlemlediğimizi daha önce belirtmiştik (Tan et al., 2017) (Putra and Khodra, 2017). Bizim problemimize ürettiğimiz çözümle elde edilen sonuçların başarımını karşılaştırmak için bir metin özetleme problemine getirilen çözüm seçilmiştir. Variations of the Similarity Function of TextRank for Automated Summarization (Barrios et al., 2016) makalesi seçilmiştir. Baseline olarak seçtiğimiz makalede (Barrios et al., 2016), Barrios, L´opez, Argerich, Wachenchauzer TextRank'ın benzerlik fonksiyonunu geliştirerek metin özetlemeyi geliştirmeyi planlamışlardır. Orijinal TextRanking algoritmasının üzerine tanımlanan modifikasyonlar: 1. longest commen substring: cümleler arası en uzun eşleşen substring'in belirlenmesi ve de uzunluğun raporlanması 2. cosine distance: Bu formül textlerin vektör formatlarının karşılaştırılması için kullanılıyor. Metni vektör haline getirebilmek için orijinal TF-IDF methodu kullanılmıştır. 3. BM25: BM25, bir belgenin bir sorguya olan uygunluğunu ölçmek için kullanılan bir skorlama fonksiyonudur. BM25, sorgunun özelliklerine göre belgeleri puanlar ve her belgenin sorguya olan uygunluğunu sıralar. Projede, metin özetleme için kullanılan özellikler İngilizce dilinin dil anlamsal yapısından bağımsız olduğundan dolayı Türkçe'ye uyarlanması için herhangi bir sorun teşkil etmemektedir. Bu sebeple projenin Türkçe'ye uyarlanması için bir işlem yapılmamıştır ve çıkan sonuçlar değerlendirildiğinde bir problemle karşılaşılmamıştır.

4.2 Deney Sonuçları

Atasözü içeren cümle, araması bütün haber metinlerinde yapılmıştır. Bir makalede (Nandini Sethi and Kumar, 2016), deyim ve atasözlerinin haber metinlerini kısaca ve zekice özetlediğinden dolayı başlıklarda tercih edildiği ve isim tamlamalarının da yüksek oranda başlıklarda kullanıldığı belirtilmiştir. Bunun için Kaggle sitesinden (Okçular, 2023) türkçe deyim-atasözleri veri setinden sadece atasözleri kalacak şekilde temizleme işlemi yaptık. Daha sonra metinde atasözünün olup olmadığının taramasını yapıldı. Fakat elimizdeki veri setinde hiçbir haber metni atasözü içermemektedir. Yani onucumuzu etkileyen bir öznitelik olmamıştır. Bu sebeple geliştirme aşamasında kullanılamamıştır.
Öznitelikler Kıyaslanması: Bu deneyde, her bir öznitelik sistemden çıkarılıp yeniden bir model eğitilmesi yapılmıştır. Aşağıdaki tabloda RF modeli üzerinde sonuçlar gözlemlenebilmektedir.

	MSE	MAE	R2
Normal	0.028	0.13	0.11
Noun Phrase	0.028	0.13	0.093
NER	0.029	0.13	0.079
TF-IDF	0.03	0.13	0.09
Cümle Uzunluğu	0.03	0.13	0.13
Cümle Konumu	0.03	0.147	-0.07

Model Başarımı: Başlık ile en uyumlu cümle bulunurken kullanılan LR ve RF makine öğrenmesi modellerinin başarımı test edilip elde edilen sonuçlar tabloda gösterilmiştir.

RF	LR
0.3	0.27

Başlık üretme problemine orijinal başlık ile benzerliğini bulma yönünden yaklaşıldığı için

ayrıca, metin içinden en yüksek benzerlik skoruna sahip olan cümleyi bulma oranı ile de başarımı hesaplanmıştır. Çoklu çalıştırmalarda da yukarıda elde edilen sonuçlarla benzer sonuçlar elde edilmiştir.

Baseline ile Kıyaslama: Daha önce belirtilen referans çalışmamız metni özetleyip cümleler halinde özet dönmektedir. Bir bir cümle ile başlığın benzerliğine bakılacak şekilde kendi çalışmamız ile (Barrios et al., 2016) arasında bir karşılaştırma yaptık. Bu karşılaştırma sonucunda baseline yönteminin az bir farkla daha iyi bir sonuç verdiği gözlemlenmiştir.

Baseline	Modelimiz
0.32	0.3

5 Kısıtlamalar

- **Veri setindeki gürültü:** Başlıklarda ve metinlerde yer alan ”devamı arkada”, ”şok!şok!şok!” gibi anlamsız cümleler modelin yanlış öğrenmesine sebep olabilmektedir.
- **Veri miktarı:** Zaman ve kaynak kısıtında dolayı model eğitim yaparken elimizdeki veri setinde bulunan 2500 haber metni kullanılmıştır. Bu verilerin %80'i eğitim aşamasında %20'si test aşamasında kullanılmıştır.
- **Hesaplama Gücü:** İsim tamlamalarını kullanmak için kullandığımız Dependancy-Parsing işleminin hesaplanması uzun sürdüğünden dolayı bize bir zaman aynı zamanda bellek kısıtına neden olmaktadır.
- **Dependancy Parsing:** Dependancy parsing için kullandığımız yöntemde bazı cümleler çok uzun geldiğinden dolayı (vnl) çalışırken hata verip programı sonlandırmaktadır. Bu sebeple cümleyi chunklarına ayırıp o chunklarda isim tamlaması araması yapıldı. Fakat bu yöntem ile metinde bazı isim tamlamalarını bulamamaktadır.

6 Sonuç

Bu çalışmada, haber metinlerindeki her cümlenin önem düzeyi, çıkarılan öznitelikler ve makine öğrenmesi yöntemleri kullanılarak belirlenip metnin ana anlamını yansıtan bir başlık oluşturmak amacıyla kullanılmıştır. Başlık oluşturmada birçok haber metni için görece iyi sonuçlar elde ettiğimizi düşünsek de matematiksel başarımda istenilen yüksek başarımlar elde edilememiştir. Gelecek çalışmalarda daha fazla öznitelik ve makine öğrenmesi modeli ile başarım iyileştirilmesi planlanmaktadır. Baseline ile karşılaştırdığımızda Baseline methodumuz özet verip benzerlik için daha bir uzun bir sonuç dönse de elde etttiğimiz başarımlar Baseline metodumuza benzerdir. Bunun yanı sıra haber metinlerine başlık oluşturmak için derin öğrenme teknikleri proje kapsamına eklenebilir. Gelecek çalışmalarda, daha fazla öznitelik çıkarım yöntemi ve makine öğrenmesi modeli kullanılabilir. Yaptığımız çalışmaların kaynak koduna Referanslar kısmında bulunan (git)'dan ulaşılabilmektedir.

References

Github repository.

scikit-learn.

Sentence similarity with bert.

Veri seti.

Vnlp.

Ahmet Afsin Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source nlp framework for turkic languages. *Structure*, 10(2007):1–5.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606.

Meru Brunn, Yllias Chali, and Christopher Pinchak. 2002. Text summarization using lexical chains.

John M. Conroy and Dianne P. O'leary. 2001. Text summarization via hidden markov models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 406–407, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin Eduard Hovy. 1999. Automated text summarization in summarist.

Rafael Ferreira, Frederico Freitas, Luciano de Souza Cabral, Rafael Dueire Lins, Rinaldo Lima, Gabriel França, Steven J. Simske, and Luciano Favaro. 2014. A context based text summarization system. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 66–70.

K. Kaikhah. 2004. Automatic text summarization with neural networks. In *2004 2nd International IEEE Conference on 'Intelligent Systems'. Proceedings (IEEE Cat. No.04EX791)*, volume 1, pages 40–44 Vol.1.

Anita R. Kulkarni and Sameer Apte. 2013. A domain-specific automatic text summarization using fuzzy logic.

Mücahid Kutlu, Celal Cığır, and Ilyas Cicekli. 2010. Generic Text Summarization for Turkish. *The Computer Journal*, 53(8):1315–1323.

Aditya Mantha Stephen Guo Kannan Achann Mansi Ranjit Mane, Shashank Kedia. 2020. Product title generation for conversational systems using bert.

Vishu Madaan Sanjay Kumar Singh Nandini Sethi, Prateek Agrawal and Anuj Kumar. 2016. Automated title generation in english language using nlp.

Emre Okçular. 2023. Turkish idioms and proverbs.

Jan Wira Gotama Putra and Masayu Leylia Khodra. 2017. Automatic title generation in scientific articles for authorship assistance: A summarization approach. *Journal of ICT Research and Applications*, 11:253.

Yixing Fan Yanyan Lan Jun Xu Huanhuan Cao Xueqi Cheng Ruqing Zhang, Jiafeng Guo. 2018. Question headline generation for news articles.

Yixing Fan Yanyan Lan Xueqi Cheng Ruqing Zhang, Jiafeng Guo. 2020. Structure learning for headline generation.

D. Zajic R. Schwartz, Blanche E. Door, and Richard M. Schwartz. 2002. Automatic headline generation for newspaper stories.

Liqun Shao and Jie Wang. 2017. Dtatg: An automatic title generator based on dependency trees. *Computing Research Repository*, arXiv:1710.00286v1. Version 1.

Francis C.M. Lau Songhua Xu, Shaohui Yang. 2010. Keyword extraction and headline generation using novel word features.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4109–4115.

Ling Luo Ying Qiao Qing He Xing Xie Xiang Ao, Xiting Wang. 2021. Pens: A dataset and generic framework for personalized news headline generation.

Jiawei Han Jialu Liu Hongkun Yu You Wu Cong Yu Daniel Finnie Jiaqi Zhai Nicholas Zukoski Xiaotao Gu, Yuning Mao. 2020. Generating representative headlines for news stories. *CoRR*, abs/2001.09386.

Mucahid Kutlu Yavuz Selim Kartal. 2020. Türkçe haber metinleri için makine Öğrenmesi temelli Özetleme.

Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, and I-Heng Meng. 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing Management*, 41(1):75–95. An Asian Digital Libraries Perspective.

Jiawei Miao Xinmiao Yu Zhengpeng Li, Jiansheng Wu. 2022. News headline generation based on improved decoder from transformer.