*Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi*
*Sayı 8, S. 309-317, 4, 2024*

***Araştırma Makalesi***

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

*International Journal of Advanced Natural Sciences and Engineering Researches*
*Volume 8, pp. 309-317, 4, 2024*

***Research Article***

# Varlık İsmi Tanıma Probleminin Sosyal Ağlarda Reklam Sistemine Uyarlanması

Harun Berkin ÇETİN [1*], Feyza Zeynep SALAM [1], Kadriye MARANGOZ [1] ve Burak GÖZÜTOK [1]

[1]*Veri Bilimi Takımı Donanım Haber Elektronik Yayıncılık İstanbul, Türkiye*

[*]*(berkincetin@donanimhaber.com.tr)*

*(Geliş Tarihi: 15 Mayıs 2024, Kabul Tarihi: 25 Mayıs 2024)*

(3rd International Conference on Engineering, Natural and Social Sciences ICENSOS 2024, May 16-17, 2024)

**ATIF/REFERENCE:** Çetin, H. B., Salam, F. Z., Marangoz, K. & Gözütok, B. (2024). Varlık İsmi Tanıma Probleminin Sosyal Ağlarda Reklam Sistemine Uyarlanması. *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(4), 309-317.

***Özet –*** Sosyal ağlarda reklam sisteminin yaygınlaşmasıyla içeriklerden anlam çıkarılma ihtiyacı artmıştır. Bu çalışma kapsamında da forum temelli sosyal ağ platformları için kullanıcıların yazdıkları metinlerden yararlanılarak akıllı bir reklam sisteminin ortaya çıkarılmasına destek olunmaktadır. Geliştirilen sistem, Varlık İsmi Tanıma (Named Entity Recognition) teknikleri kullanılarak, metinlerdeki ürün adlarını ve bunların ait olduğu kategorileri tanımlamak için tasarlanmıştır. Metinlerin vektörleştirilmesi için de BERT mimarisinden yararlanılmaktadır. Bu sayede, forum temelli sosyal ağ platformlarındaki kullanıcılar tarafından paylaşılan içeriklerde yer alan ürünlerin ve kategorilerin otomatik olarak tespit edilmesi mümkün hale gelmiştir. Çalışma sonuçlarının, pazarlama ve reklam stratejileri için faydalı olacağına inanılmaktadır.

*Anahtar Kelimeler – Varlık İsmi Tanıma, İsim Varlık Tanıma, Reklam Sistemi, Sosyal Ağ, BERT.*

## I. GİRİŞ

Son yıllarda sosyal ağ platformları, teknolojinin de gelişmesiyle milyonlarca kullanıcısı ile dünyanın her yerinde popüler hale gelmiştir. Bu platformların özellikle forum temelli olanları, kullanıcıların insanlarla iletişim kurmalarını, bilgi paylaşımında bulunmalarını ve farklı, özgün ve kişisel içerikleri keşfetmelerini sağlamaktadır. Bunun yanı sıra, işletmeler ve reklam verenler de bu platformları pazarlama stratejilerinde kullanmaktadırlar. Ancak reklamların doğru hedef kitleye ulaştırılmasında sıkıntı yaşanması ile reklam bütçelerinin de etkin kullanımı engellenebilmektedir. eMarketer tarafından yapılan bir araştırmada da, reklamların yaklaşık %30'unun hedef kitleye ulaşamadığı belirtilmiştir [1]. Bu nedenle, "Varlık İsmi Tanıma Probleminin Sosyal Ağlarda Reklam Sistemine Uyarlanması" çalışması, doğru kitleye doğru ürünleri öneren bir reklam sistemi geliştirmeyi hedeflemektedir.

Bu çalışmada, sosyal medya platformlarında bulunan ürünlerin doğru kişilere önerilmesi için bir makine öğrenimi yaklaşımı olan "Varlık İsmi Tanıma" (Named Entity Recognition, NER) yöntemini kullanmaktadır. Bu yöntem, belirli bir metinde bulunan önemli kelimeleri belirleyerek, bunların anlamını ve ilişkilerini analiz eder. Bu problemde önemli kelimeler ürünler ve kategori olmak üzere 2 ana başlıkta incelenmektedir. Bu sayede de, kullanıcıların ilgileri tespit edilip buna göre uygun ürünler önerilmektedir.

Çalışmada, özellikle BERT (Bidirectional Encoder Representations from Transformers) [2] modeli kullanılmaktadır. BERT modeli, önceki modellere kıyasla daha uzun metinlerde daha yüksek doğruluk oranlarına sahip olduğundan, bu çalışmada kullanılmak için seçilmiştir.

Bu çalışmada, reklamların hedef kitleye ulaşmasını kolaylaştırırken, aynı zamanda kullanıcıların daha ilgi çekici ve uygun ürünler keşfetmelerine yardımcı olmaktadır. Bu nedenle, kullanıcıların sosyal medya platformlarından daha fazla fayda sağlamalarını sağlayacak önemli bir çalışma olarak görülmektedir.

İlgili çalışmada [3], Çin'de adlandırılmış varlık tanıma için iyi tanımlanmış bir ince çözünürlüklü veri kümesi olan CLUE organizasyonunun (CLUENER2020) NER veri kümesi tanıtılmıştır. CLUENER2020, 10 kategoriden (kişi ismi, organizasyon, pozisyon, şirket, adres, oyun, hükümet, sahne, kitap ve film) oluşmaktadır. Bunlar yaygın etiketlerin yanı sıra daha çeşitli kategoriler içerdiği için daha zorlayıcı ve gerçek dünya uygulamalarını daha iyi yansıtabilmektedir. Çalışmada BILSTM+CRF, BERT, RoBERTa ve bu yöntemlerin kıyaslanması için insan performansı kullanılmıştır. Çalışmanın sonuçları, RoBERTa yönteminin on başlık için en yüksek doğruluk oranını sağladığı gözlemlenmiştir. En yüksek doğruluk oranı kişi ismi için %89.09 olarak elde edilmiştir. Diğer başlıklar için doğruluk oranları sırasıyla organizasyon %82.34, pozisyon %79.62, şirket %83.02, adres %62.63, oyun %86.80, hükümet %88.17, sahne %70.49, kitap %74.60 ve film %87.46 olarak açıklanmıştır.

İlgili çalışmada [4], ürün başlıklarındaki marka isimlerini tanımak için bir yöntem önerilmektedir. Çalışmada, bir Bi-LSTM-CRF modeli kullanılmıştır. Model, her kelimeyi bir dizi özellikle temsil edilmekte ve ardından Bi-LSTM tabakaları ile özelliklerin bağlantılarını öğrenmektedir. Daha sonra, CRF tabakası, kelime düzeyinde etiketlerin belirlenmesinde kullanılmaktadır. Çalışma, iki farklı veri kümesi üzerinde test edilmiş ve farklı marka isimlerini tanımada %82 ila %84 arasında değişen doğruluk oranları elde edilmiştir. Ayrıca bu çalışmada marka tanıma performansının artırılması adına farklı ön işleme adımları ve kelime temsilleri de denenmiştir. Özellikle, kelime düzeyinde karakter seviyesinde özellikler ve kelimeler arasındaki çift yönlü bağlantılar kullanılarak elde edilen kelime temsillerinin, modelin doğruluğunu artırdığından bahsedilmiştir. Sonuç olarak; bu çalışma, ürün başlıklarındaki marka isimlerini tanımak için bir Bi-LSTM-CRF modelinin etkili bir şekilde kullanılabileceğini ve doğruluğun, özellikle farklı ön işleme adımları ve kelime temsilleri ile birlikte optimize edildiğinde artırılabileceğini göstermektedir.

İlgili çalışmada [5], Bengalce akıllı telefon markaları ve modelleri üzerine insanların yazdığı olumlu ve olumsuz yorumları içeren metinler veri seti olarak kullanılmıştır. Varlık isim tanıma eğitimi için Spacy'nin NER modeli ve Amazon Comprehend servisine ait NER modeli kıyaslanmıştır. Çalışmanın sonuçları, Spacy Named Entity için %87.99 doğruluk oranı ve Amazon Comprehend NER için %95.51 doğruluk oranı olarak açıklanmıştır.

İlgili çalışmada [6], NER probleminin e-ticaret alanındaki önemi vurgulanmaktadır. TripleLearn adında bir model eğitim çerçevesi kullanarak üç ayrı eğitim veri seti üzerinden öğrenen bir NER modelinin performansı gösterilmektedir. En iyi model, holdout test verisindeki F1 skoru %69.5'ten %93.3'e yükseltmektedir. Modelin yayınlandığı süre boyunca, kullanıcı etkileşimlerini, arama dönüşümlerini ve geliri arttırdığı belirtilmiştir. TripleLearn çerçevesinin model ve problemden bağımsız olduğu ve endüstriyel uygulamalara genelleştirilebileceği de vurgulanmıştır.

Tüm bu çalışmalar, NER alanında e-ticaret ürünleri için çeşitli yöntemlerin önerildiği ve performanslarının ölçüldüğü çalışmalardır. Özellikle marka adı ve ürün türü gibi kategoriler için yüksek başarı oranları elde etmişlerdir. Bu sonuçlar, NER teknolojilerinin e-ticaret sektöründe oldukça etkili olduğunu göstermektedir. Bu çözümlerin daha da geliştirilmesiyle, müşterilerin aradıkları ürünleri daha hızlı ve kolay bir şekilde bulabilecekleri bir e-ticaret deneyimi sunulabilir. Ayrıca kullanıcı etkileşimleri, arama dönüşümleri ve gelir arttırılabilir.

## II. MATERYAL VE YÖNTEM

### A. *Veri Kümesi*

Bu çalışmada, Forum platformundaki konu başlıklarının altında olan mesajların etiketlenmesiyle 20697 örnekten oluşan bir veri seti oluşturulmuştur. Oluşturma sürecinde Doccano isimli bir veri etiketleme platformu kullanılmıştır. Doccano, verilerin etiketlenmesine olanak tanıyan bir araçtır. Doccano platformundan alınan ekran görüntüsü Şekil 1'de verilmiştir.


Şekil 1. Doccano Veri Etiketleme Platformu

Etiketlenmiş veriler, JSONL formatında Doccano platformundan alınmış ve CoNNL formatına dönüştürülmüştür. Bu dönüşüm sırasında, her isimli varlığın ilk sözcüğüne "B-etiket_ismi" etiketi, sonraki her sözcüğe "I-etiket_ismi" etiketi ve geri kalan isimsiz sözcüklere ise "O" etiketi verilerek gerçekleştirilmiştir. Bu işleme "BIO tagging" adı verilmektedir. Bu formatın örneği Tablo 1' ve Şekil 2'de verilmiştir.

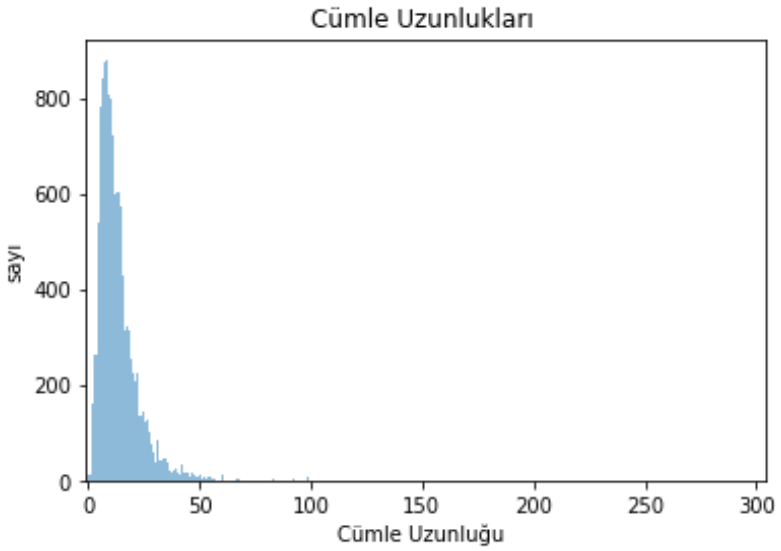
Tablo 1. BIO Tagging

| **Tokens** | **IO** | **BIO** |
|---|---|---|
| Dün | O | O |
| öğlen | O | O |
| , | O | O |
| Ayşe | I_PER | B_PER |
| A | I_PER | I_PER |
| yapay | I_PER | I_PER |
| zeka | I_PER | I_PER |

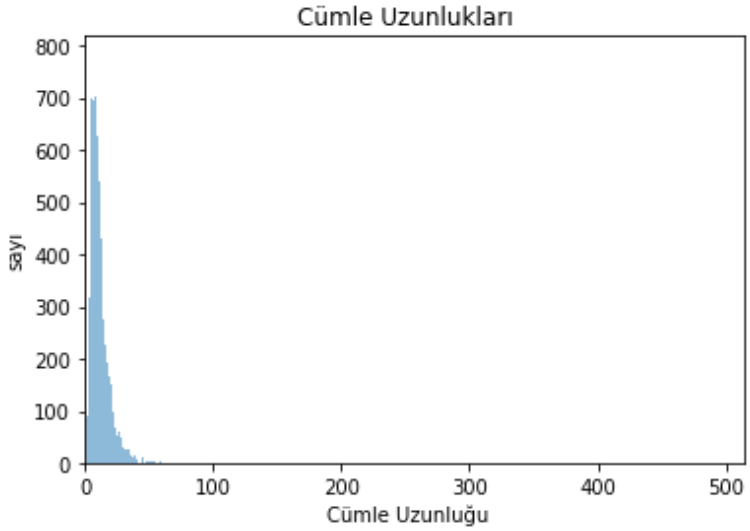
| | -DOCSTART- | -X- | -X- | O |
|---|---|---|---|---|
| 0 | NaN | NaN | NaN | NaN |
| 1 | Müzik | NN | NN | O |
| 2 | Şenliği | NN | NN | O |
| 3 | ' | NN | NN | O |
| 4 | ne | NN | NN | O |
| 5 | hazırlanın | NN | NN | O |
| 6 | POZİTİF | NN | NN | B-ORG |
| 7 | ve | NN | NN | I-ORG |
| 8 | Açık | NN | NN | I-ORG |
| 9 | Radyo | NN | NN | I-ORG |

Şekil 2. CoNNL Formatı

Veri setindeki her forum mesajı, Zemberek Türkçe dil işleme kütüphanesi [7] kullanılarak cümlelere ayrılmıştır. Ayrıca, veri setinde aşırı uzun veya kısa forum mesajları bulunduğu için bu mesajlar veri setinden çıkarılmıştır. Veri setinden bu mesajlar çıkarılmadan önceki ve sonraki cümle uzunluk dağılımı Şekil 3 ve 4'te yer almaktadır. Bu önişleme adımları sonrasında, veri seti modelin kullanabileceği bir formata dönüştürülmüş ve model eğitimi için kullanılmıştır.




Şekil 3. Temizleme Öncesi Veri Seti Cümle Uzunluk Dağılımı

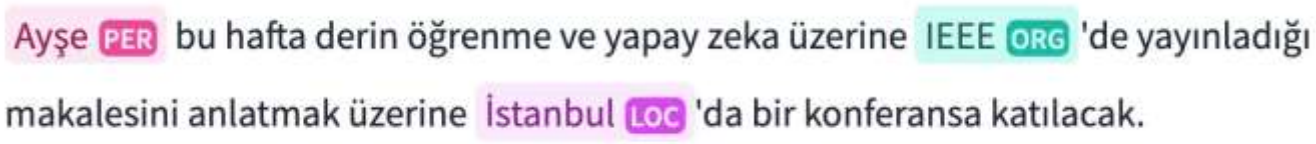



Şekil 4. Temizleme Sonrası Veri Seti Cümle Uzunluk Dağılımı

### B. *Deney Yöntemi*

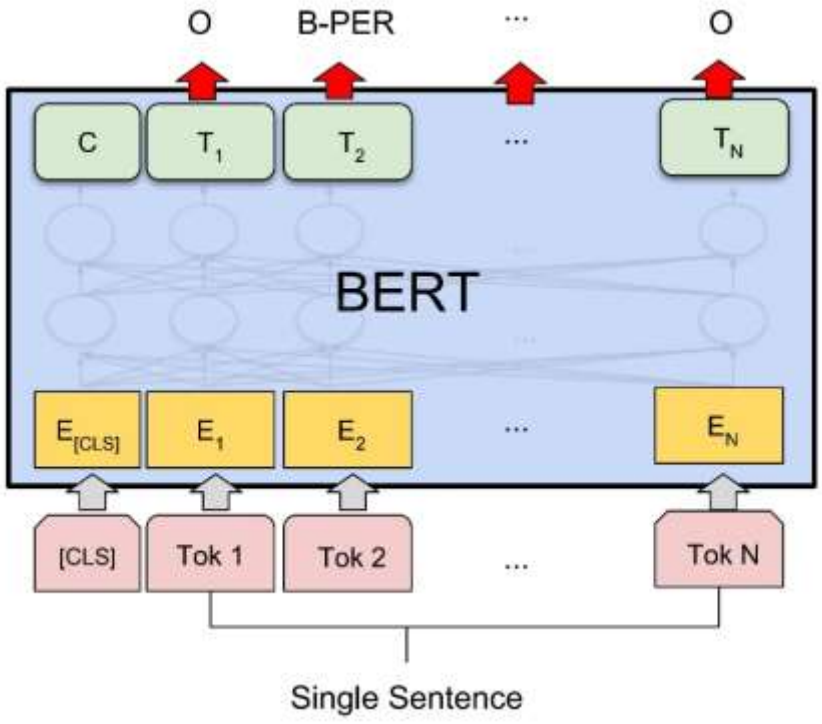
Çalışmanın ana amacı, akıllı bir reklam sistemi geliştirmektir. Bu reklam sistemi ile kullanıcılara daha akıllı reklamlar sunmak hedeflenmiştir. Bu reklam sistemi, kullanıcılara daha akıllı reklamlar sunmak için tasarlanmaktadır. Bu nedenle, çalışmada kullanılan yöntemler de bu amaç doğrultusunda seçilmiştir.

NER problemi, çalışmanın merkezinde yer alan bir sorundur. Bu yöntem genellikle kişi, yer, organizasyon isimleri gibi bilgileri tespit etmek için kullanılmaktadır ve bu çalışmanın temelini oluşturmaktadır. Şekil 5'teki örnekte insan ismi (PER), organizasyon ismi (ORG) ve yer ismi (LOC) tanıma örneği sunulmuştur. Bu çalışma kapsamında problem, bir metin içerisinde geçen varlıkların (ürün adları, markalar, kategoriler vb.) tanınması ve sınıflandırılması olarak tanımlanmaktadır.



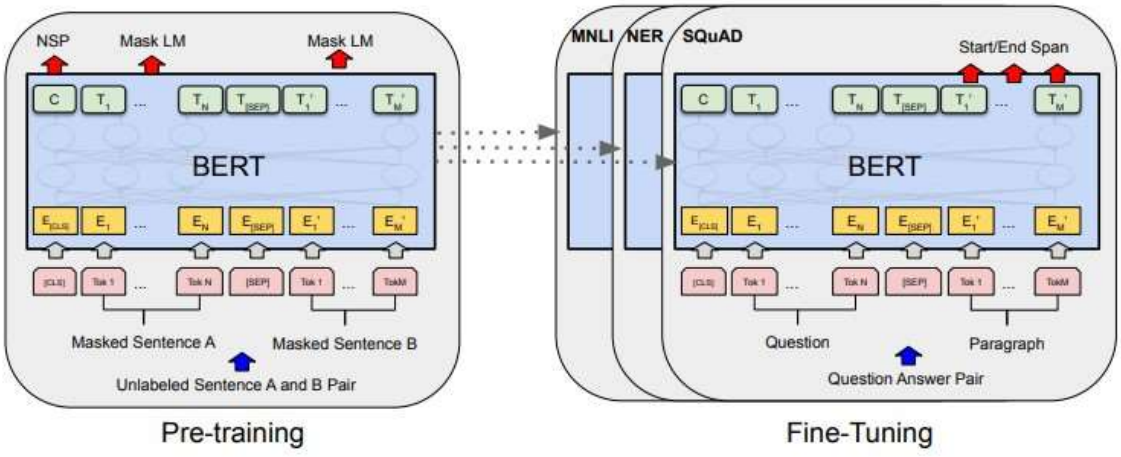

Şekil 5. Varlık İsim Tanıma

Bu çalışmada, akıllı reklam sistemi için forum mesajları üzerinde NER problemine odaklanan BERT mimarisini kullanan bir model geliştirilmiştir. BERT, Google tarafından geliştirilmiş olan ve doğal dil işleme görevleri için yaygın olarak kullanılan önceden eğitilmiş bir dil modelidir. BERT, dönüştürücü tabanlı bir sinir ağı mimarisine sahip olup büyük miktarda metin verisinde eğitilerek bağlamsal metin vektörleri oluşturmak için kullanılmaktadır. Önceden eğitilmiş dil modelleri sayesinde, daha küçük veri setleriyle bile iyi sonuçlar elde etmek mümkündür. Şekil 6'de bu mimari görselleştirilmiştir.




Şekil 6. BERT modeli ile NER [2]

NER problemine çözüm sunmak için, önceden eğitilmiş BERT modeline belirteç sınıflandırma yapabilen bir katman eklenmiş ve modelin ince ayarlaması gerçekleştirilmiştir. Bu sayede, model metindeki her bir belirteç için tahminde bulunabilmektedir. Bu çalışma kapsamında Türkçe BERT modeli olan "dbmdz/bert-base-turkish-cased" [8] üzerinde ince ayarlama yapılmıştır. Şekil 7'de BERT üzerinde ince ayarlama gösterilmiştir.




Şekil 7. BERT modeli ince ayar [2]

Çalışmanın amacına uygun olarak, veri seti olarak forum mesajları seçilmiş ve Türkçe BERT modeli, forum mesajlarında geçen "ürün adı", "marka" ve "kategori" varlıklarını tanımlayıp sınıflandırabilen bir NER sistemi geliştirmek için kullanılmıştır. Modelin daha iyi performans göstermesi için her cümle, alt sözcüklere bölünmüş; karşılık gelen etiketler belirlenmiş ve cümle uzunluğundan artan kısımlar "PAD"

etiketli belirteçlerle doldurulmuştur. Türkçe BERT, alana özgü bir veri kümesinde ince ayarlanarak performansı artırılmıştır. Model eğitimi için şu parametre değerleri kullanılmıştır:

- Epok (Epoch) : 30
- Öğrenme katsayısı (Learning rate): 1e-05
- Toplu boyut (Batch size) : 32

Eğitim ve test setleri, %90 ve %10'luk dağılımlarla ayarlanmıştır. Bunun yanında, veri setinin ilk 5000'lik kısmıyla ayrı bir eğitim yapılmış ve artan veri sayısının model performansına etkisi ölçülmüştür.

## III. BULGULAR

Çalışma kapsamında kullanıcılara akıllı reklam sunmak için farklı veri sayıları ile iki ayrı NER modeli eğitilmiş ve bu modellerin sınıflandırma sonuçları Tablo 2 ve 3'de gösterilmiştir. Modellerin eğitim veri seti büyüklükleri sırasıyla 5000 ve 20697'dir.

Tablo 2. Model 1 Sınıflandırma Sonuçları

| | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| PAD | 1.00 | 1.00 | 1.00 |
| Kategori | 0.63 | 0.84 | 0.72 |
| Ürün | 0.70 | 0.74 | 0.72 |
| Micro Avg | 0.77 | 0.87 | 0.82 |
| Macro Avg | 0.78 | 0.86 | 0.82 |
| Weighted Avg | 0.79 | 0.87 | 0.82 |

Tablo 3. Model 2 Sınıflandırma Sonuçları

| | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| PAD | 1.00 | 1.00 | 1.00 |
| Kategori | 0.79 | 0.93 | 0.86 |
| Ürün | 0.87 | 0.94 | 0.90 |
| Micro Avg | 0.88 | 0.96 | 0.92 |
| Macro Avg | 0.89 | 0.96 | 0.92 |
| Weighted Avg | 0.89 | 0.96 | 0.92 |

Birinci model ile %82, ikinci model ile %92'lik bir F1 skoru elde edilmiştir. Veri sayısının arttırılması ile model başarımında %10'luk bir artış sağlanmıştır. Performans metrikleri dışında gözle bir değerlendirme yapılması için, model geliştirilme aşamasında ortaya çıkan 2 modelin çıktıları ayrı ayrı incelenmiştir. Tablo 4'de örnek cümleler içerisinde geçen ürünleri tespit etmek amacıyla iki modelden ayrı ayrı alınan çıktılar da gösterilmiştir.

Tablo 4. Örnek Cümlelerin Model Çıktıları

| **Örnek Cümle** | **Model Çıktısı** | **Model** |
|---|---|---|
| Hoover'ın isi pompalı 9 kg kapasiteli kurutma makinesini 3 yıldır kullanıyorum. Faturada belirgin bir fark olmuyor haftada 2 3 kez çalıştırma ile. Tavsiye ederim. | [('Hoover', 'B-Kategori'), ('ınlı makinesini', 'B-Kategori')] | 1 |
| | [("Hoover '", 'B-Kategori'), ('ısı pompalı kurutma makinesini', 'B-Kategori')] | 2 |
| M52 Sahipleri için tavsiyeler. 1- Samsung note 7 faciasının etkisi midir nedir bilinmez yüksek hızlı şarj değerlerine çıkmıyor. Cihazın kutusundan çıkan kablo ve adaptör milyoncudan alınmış gibi. | [('M52', 'B-Urun'), ('Samsung note 7', 'B-Urun'), ('kablo', 'B-Kategori'), ('adaptör', 'B-Kategori'), ("Samsung", "B-Kategori"), ("un", "B- | 1 |

| | | |
|---|---|---|
| Piyasadaki en ucuz 25w şarjı destekleyen şarj aleti dejininki. Samsung'un kendi adaptörü super hızlı şarj diye bildirim çıkarıyor. | Kategori"), ('adaptörü', 'B-Kategori')] | |
| | [('M52', 'B-Urun'), ("Samsung note 7', 'B-Urun'), ('kablo', 'B-Kategori'), ('adaptör', 'B-Kategori'), ('sarj aleti', 'B-Kategori'), ("Samsung", 'B-Kategorisi')] | 2 |

İlk örnek cümlede %92'lik skora sahip 2. modelin daha iyi sonuç verdiği görülmektedir. Yukarıdaki örnekte de görüldüğü üzere, düşük skora sahip model "-ın" ekini ve "pompalı" sözcüğünde sadece "-lı" ekini isimli varlık belirteci olarak sınıflandırmıştır. Fakat yüksek skora sahip modelin "ısı pompalı kurutma makinesini" sözcüğünü doğru şekilde isimli varlık olarak sınıflandırdığı görülmektedir.

İkinci örnek cümlede de %92 F1 skoruna sahip Model 1'in, %81 F1 skorluk Model 2'ye nazaran daha iyi sonuç verdiği görülmektedir.

İki model çoğunlukla benzer sonuçlar verse de, NER Modeli alt-sözcüklere (belirteç) göre çıkarım yaptığından dolayı, hatalı örneklerde anlamsız sonuçlar çıkabilmektedir. İki örnekte de benzer şekilde %92'lik skora sahip 1. modelin daha iyi sonuç verdiği görülmektedir.

## IV. TARTIŞMA

Elde edilen BERT ve NER modeli, ürün kategorileri ve isimlerinin tanınması için kullanılmıştır. Bu modeller, sosyal ağlardaki reklam sistemine entegre edilerek, kullanıcılara daha doğru ve ilgili reklamlar sunulması amaçlanmıştır.

Deney yöntemleri olarak, öncelikle veri toplama ve temizleme işlemi gerçekleştirilmiştir. Daha sonra, BERT temeli üzerine geliştirilen NER modeli kullanılarak veriler işlenmiş ve kategori sınıflandırması yapılmıştır. Son olarak, geliştirilen model kullanılarak ürün isimleri tespit edilmiştir.

Bu yöntemlerin performansı, farklı ölçütlerle/metriklerle incelenmiştir. Precision, recall, ve F1 score gibi ölçütlere bakılarak modellerin doğruluğu ve başarımı değerlendirilmiştir. Ayrıca, manuel olarak modellerin başarımı gözlemlenmiştir.

## V. SONUÇLAR

Elde edilen sonuçlar, BERT'e ek katman eklenerek geliştirilen ve ince ayar yapılan NER modelinin ürün kategorisi ve isimlerini tanıma konusunda oldukça başarılı olduğunu göstermiştir. Bu sonuçlar, sosyal ağlarda reklam sistemleri için geliştirilen bu yöntemin etkili bir şekilde kullanılabileceğini göstermektedir.

## KAYNAKLAR

[1] eMarketer, "Challenges Facing Advertisers When Targeting Mobile Audiences", 2018.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018.
[3] Liang Xu, Yu tong, Qianqian Dong, Yixuan Liao, Cong Yu, Yin Tian, Weitang Liu, Lu Li, Caiquan Liu, Xuanwei Zhang "CLUENER2020: Fine-grained Named Entity Recognition Dataset and Benchmark for Chinese", 2020.
[4] Priyanka Goyal, Thomas Packer, Faizan Javed, "An End-to-End Solution for Named Entity Recognition in eCommerce Search", 2020.
[5] Md Sabbir Hossain, Nishat Nayla, Annajiat Alim Rassel, "Product Market Demand Analysis Using NLP in Banglish Text with Sentiment Analysis and Named Entity Recognition", 2022.

[6] Xiang Cheng, Mitchell Bowden, Bhushan Ramesh Bhange, Priyanka Goyal, Thomas Packer, Faizan Javed, "An End-to-End Solution for Named Entity Recognition in eCommerce Search", 2020.
[7] Ahmet Afşın Akın, Mehmet Dündar Akın, "Zemberek, an open source NLP framework for Turkic Languages", 2007.
[8] Stefan Schweter, "BERTurk - BERT models for Turkish", Zenodo, 2020.