Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 7(2) (2024) 159-173
DOI: 10.38016/jista.1509968

Fenomen-Hedef Kitle Eşleştirmesinin Otomatikleştirilmesi: Sosyal Medya Gönderilerinin Sınıflandırılması ile Reklama Yönelik Hedef Kitle Analizi

Mehmet Varan[1], Aslı Yatkınoğlu[2*], Amine Gonca Toprak[3], Fatih Soygazi[4], Bora Mocan[5]

[1, 2, 3, 5] AdresGezgini A.Ş., İzmir, Türkiye

[4] Bilgisayar Mühendisliği Bölümü, Adnan Menderes Üniversitesi, Aydın, Türkiye

mehmetvaran@adresgezgini.com, aslicankut@adresgezgini.com, goncatoprak@outlook.com, fatih.soygazi@adu.edu.tr, boramocan@adresgezgini.com

Öz

İnternet kullanımının son yıllarda yaygınlaşması, bireylerin ve toplumların iletişimden alışveriş alışkanlıklarına kadar neredeyse her alanda davranışlarının evrilerek büyük değişikliklerin ortaya çıkmasına sebep olmuştur. Böylece geleneksel iletişim yöntemleri de dönüşüme uğramıştır. Bu gelişmeler sonucunda, günümüzde en yaygın iletişim aracı olarak kabul edilen sosyal medya kavramı doğmuştur. Yeni bir iletişim şekli olan sosyal medya, kurum ve kuruluşların hedef kitleleri ile yer ve zaman kısıtı olmaksızın doğrudan iletişim kurabilmelerini mümkün kılarak reklam verenler için ürünlerini tanıtabilecekleri oldukça etkili bir kanal haline gelmiştir. Sosyal medyada ürün pazarlamak "fenomen" olarak adlandırılan kişiler sayesinde gerçekleşmektedir ve her fenomenin hitap ettiği bir hedef kitle bulunmaktadır. Bu bağlamda, fenomenlerin hitap ettiği hedef kitle ile reklamı yapılacak ürünün hedef kitlesinin doğru bir şekilde eşleşmesi, sosyal medya üzerinden yapılan ürün pazarlamasında kritik bir rol oynamaktadır. Bu çalışmada en doğru fenomen-ürün hedef kitle eşleşmesini gerçekleştirebilmek adına, Instagram fenomenlerinin paylaşmış olduğu gönderileri analiz ederek fenomenin hedef kitlesini kategorize eden bir metin sınıflandırma modeli geliştirilmiştir. Bu amaç doğrultusunda veri gizliliğini ihlal etmemek adına Instagram profili herkese açık olan 1.005 farklı fenomenin üçüncü taraf bir yazılım ile gönderileri elde edilerek bu gönderilerdeki açıklamalar BERTopic mimarisi ile kümelenmiştir. Oluşturulan kümelerin temsilleri ve içeriği incelenerek temsil ettiği kategoriye göre etiketlenmiştir. Etiketlenen veriler ile BERTurk sınıflandırma modeli geliştirilmiştir. Sınıflandırma model performans değerlendirilmesi sonucunda ölçülerek 0,92 doğruluk ve 0,91 F1 skor değeri elde edilmiştir. Elde edilen sonuçlar doğrultusunda yüksek sınıflandırma doğruluğu ile fenomen gönderilerini otomatik olarak kategorize edebilen bir sistem geliştirilmiş ve fenomen-ürün hedef kitle eşleştirilmesinde başarıyla kullanılmıştır.

Anahtar Kelimeler: Metin sınıflandırma, Kümeleme Analizi, BERTopic, BERTurk, Instagram

Automating Influencer-Target Audience Matching: Target Audience Analysis for Advertising through Classification of Social Media Posts

Abstract

The widespread adoption of the internet has led to significant transformations in individual and societal behaviors, influencing everything from communication to shopping habits. As a result, traditional communication methods have evolved, giving rise to social media as a dominant medium today. Social media enables organizations to engage directly with target audiences without geographical or temporal constraints, making it an effective platform for advertisers. Social media marketing is often facilitated by "influencers," individuals who have built their own audience. Accurate matching between the influencer's target audience and the advertised product's audience is essential for effective social media marketing. This study aims to develop a text classification model that categorizes the target audiences of Instagram influencers by analyzing their posts, in order to achieve the most accurate influencer-product target audience matching. To avoid violating data privacy, posts from 1.005 distinct influencers with publicly accessible Instagram profiles were collected using a third-part software, and the descriptions in these posts were clustered using the BERTopic architecture for topic modeling. A BERTurk classification model was developed using the labeled data. The representations and content of the resulting clusters were analyzed and labeled according to the categories they represented. These labeled records were then used for classification purposes. The performance of the classification model was evaluated, achieving an

* Sorumlu yazar
E-posta adresi: aslicankut@adresgezgini.com

Gönderme : 4 Temmuz 2024
Revizyon : 12 Ağustos 2024
Kabul : 7 Eylül 2024

accuracy of 0,92 and an F1 score of 0,91. The results demonstrate the development of a system that can automatically categorize influencer posts with high classification accuracy and has been successfully applied for influencer-product target audience matching.
Keywords: Text Classification, Clustering Analysis, BERTopic, BERTurk, Instagram

1. Giriş (Introduction)

Gün geçtikçe teknolojinin gelişmesi ve bireylerin günlük yaşamlarında internetin yaygınlaşmasıyla birlikte arkadaşlıklar, gündem takibi, boş zaman değerlendirme, alışveriş gibi günlük aktiviteler internet ortamına da taşınmıştır (Yıldırım ve Yıldırım, 2022). Özellikle mobil cihaz (akıllı telefon, tablet vb.) kullanımının artmasıyla internet yer ve zaman fark etmeksizin ulaşılabilir bir hale gelerek bireylerin internette daha fazla vakit geçirmesine sebep olmuştur. Teknolojinin ve internetin günlük yaşamda bu denli büyük bir role sahip olması, insanlık tarihi boyunca sürekli değişerek evrilen iletişim yöntem ve kanallarını da etkileyerek sosyal medyayı doğurmuştur (Şahinkayası ve Şahinkayası, 2017).

Sosyal medya bireylerin birbirleriyle video, mesaj veya fotoğraf içerikleriyle paylaşımda bulunmalarına ve iletişim kurmalarına olanak sağlayan çeşitli çevrimiçi platformlar olarak tanımlanabilir (Carr ve Hayes, 2015). Günümüzde sosyal medya bireylerin daha fazla iletişime geçmesine olanak sağlayarak vazgeçilmez bir iletişim aracı haline gelmiştir. Sosyal medya platformlarının yaygınlaşmasıyla birlikte farklı yaş gruplarından oluşan sosyal medya kullanıcıları, çeşitli ilgi alanları doğrultusunda hedef kitleleri oluşturmaktadır.

Sosyal medyanın en çok etkilediği alanlardan bir tanesi ürün pazarlama faaliyetleridir (Terkan, 2014). Geleneksel ürün pazarlama yöntemlerine kıyasla sosyal medya platformları, işletme büyüklüğü ve sektör fark etmeksizin tüm işletmelere her farklı yaş grubu ve sosyo-ekonomik gruptan tüketicilerle yani hedef kitlelerle iletişim kurabilme olanağı sağlamaktadır (Arslan, 2017). Instagram, Facebook, Twitter, Youtube, Tiktok, Linkedin gibi farklı sosyal medya ağları üzerinden oluşturulan çeşitli içeriklerle işletmeler, sosyal medyanın sağladığı çift taraflı ve etkileşimli iletişim sayesinde hedef kitlelere daha etkili ve daha az maliyetli bir şekilde doğrudan ulaşabilmektedir.

Ürün pazarlamanın temel araçlarından biri olan reklam, bir ürün veya hizmetin medya kanalları aracılığıyla kitlelere tanıtılması olarak tanımlanmaktadır (Bagwell, 2007). Farklı kanallar üzerinde belirli bir ücret karşılığında yapılan reklamların temel amacı, tüketicilerin ilgisini reklamı yapılan ürün veya hizmete yönlendirerek ilgili ürün veya hizmet satışının gerçekleştirilmesidir. Gazete, televizyon, radyo, dergi gibi farklı iletişim kanalları aracılığıyla yayınlanan reklamların günümüzde internetin etkisi ile sosyal medyada yaygınlaşmasıyla, sosyal medya araçları önemli bir ürün pazarlama aracı haline evrilmiştir (Özdemir vd., 2014).

Sosyal medya üzerinden ürün tanıtımları dijital reklamlar ve sosyal medya ağlarında fazla sayıda takipçisi olan hesaplar olarak adlandırılan fenomenler üzerinden gerçekleşmektedir.

Sosyal medya fenomenleri, oluşturdukları içerikler ve yaptıkları paylaşımlarla geniş kitlelere ulaşarak bireylerin düşünce, tutum ve davranışlarını etkileyebilmektedir. Sosyal medya fenomenlerinin aynı zamanda tüketicilerin satın alma kararları üzerinde de önemli bir etkisinin olması, ürün pazarlama literatürüne yeni bir kavram kazandırmıştır. Literatürde sosyal medya platformlarında fenomenler aracılığıyla gerçekleştirilen ürün pazarlama faaliyetleri şeklinde tanımlanan bu yeni kavram, fenomen pazarlaması olarak adlandırılmaktadır (Leung vd., 2022).

Fenomen pazarlaması, fenomenin herhangi bir sosyal medya platformu aracılığıyla bir ürün veya hizmete dair sunduğu, tüketicinin satın alma motivasyonunu etkileyen pazarlama aktiviteleri üzerinden gerçekleşmektedir. Tüketiciler bir ürün veya hizmet satın alırken gerçek tüketici deneyimlerine çok önem verdiğinden fenomen pazarlaması, işletmelerin hedef kitlelere ulaşabilmek için tercih ettiği en yaygın ürün pazarlama yöntemlerinden biri haline gelmiştir (Çopuroğlu, 2022). Bir diğer deyişle, fenomenler, sosyal medya platformları üzerinde işletmelerin reklam yüzü olarak içerik oluşturduğu alanda (ör. seyahat, yemek, moda vb.) hedef kitlesinin satın alma motivasyonunu olumlu bir şekilde etkilemek için içerik üreterek iş birliği yaptığı işletmenin ürün veya hizmetinin tüketicilere ulaşmasını sağlamaktadır. İşletmelerin, tüketicilerin satın alma niyetini olumlu yönde etkilemek istedikleri takdirde fenomenler ile iş birliği yaparak kazanç sağlayabilecekleri görülmüştür. Yapılan araştırmalarda fenomenlerin reklamlarına yönelik pozitif bir tutum olduğu, satışları arttırmaya yönelik olumlu etkileri olduğu ortaya çıkmıştır (Karataş ve Eti, 2022).

Fenomen pazarlamasının başarısı için en önemli faktörlerden biri, iş birliği yapacak olan fenomen ve işletmenin hedef kitlesinin örtüşmesidir (Öztek vd., 2021). Örneğin, mobilya üreten bir işletmenin hedef kitlesi kozmetik alanında ürün pazarlamaya daha uygun olan bir fenomen ile iş birliği yapması, ürün pazarlamada istenen getiriyi sağlamayacaktır. İşletme veya marka ile fenomen arasında gerçekleştirilecek iş birliğinde, fenomenin sosyal medya platformundaki paylaşımlarının içerik analizi yoluyla incelenmesi sonucunda, fenomenin hitap ettiği hedef kitle tespit edilebilir. Literatürde doğal dil işleme, yapay zeka ile sınıflandırma gibi güncel çalışma alanları sayesinde sosyal medyadaki fenomenlerin gönderileri gözetilerek otomatik bir şekilde gerçekleştirilebilmektedir (Kim vd., 2020).

Bu çalışmada Instagram fenomenlerinin sosyal medya hesap gönderilerindeki açıklamaların (caption) doğal dil işleme yöntemleri ile analiz edilerek hitap ettiği kitlenin tespiti amaçlanmıştır. Bu kapsamda öncelikle Apify (Apify, 2022) web veri çıkarma (web scraping) platformu ile Instagram'daki halka açık olan hesap gönderilerinin elde edilmesiyle bir veri seti oluşturulmuştur. Açıklama içermeyen gönderilerin ayrılması gibi veri ön işleme adımları ile veri seti, dil modellerinin eğitimine hazır hale getirilmiştir. Elde edilen veri setindeki açıklamalar, topik modelleme algoritması kullanılarak kümelenmiştir. BERTopic modelinin ürettiği kümeler, veri etiketleme ekibi tarafından titizlikle incelenmiş ve her bir kümenin içeriğine uygun etiketler önceden belirlenmiş olan 18 kategoriye göre etiketlenmiştir. Bu süreçte, kümelerin temsil ettiği temalar ve içerikler dikkatle değerlendirilerek, etiketleme işlemi her bir kümenin anlamını en iyi şekilde yansıtacak biçimde gerçekleştirilmiştir. BERT (Bidirectional Encoder Representations from Transformers) modeli hazırlanmış veri seti ile eğitilmiştir Eğitilen model yardımıyla fenomenin gönderileri sınıflandırılarak hitap ettiği kitle yüksek doğruluk oranı ile tespit edilebilmektedir.

Bu çalışma, sosyal medya kullanımının hızla arttığı bu dönemde, Türkçe dili için sosyal medya analitiği ve metin sınıflandırma alanında önemli bir boşluğu doldurmaktadır. Çalışmanın katkıları iki ana başlıkta değerlendirilebilir: Türkçe metinlerin BERTopic mimarisiyle sınıflandırılması, Türkçe dilinde daha fazla araştırma ve uygulamayı teşvik ederek akademik bir katkı sunmaktadır. Ayrıca, elde edilen sınıflandırma başarısıyla fenomen-ürün hedef kitle eşleşmesini mümkün kılarak sosyal medya pazarlamasına ticari bir katkı sağlamaktadır.

2. Literatür Taraması (Related Work)

Bu bölümde öncelikle literatürde yer alan farklı modeller ile başarı elde edilmiş metin sınıflandırma çalışmalarına yer verilmiştir. Daha sonra BERT modeli ile yapılan metin sınıflandırma çalışmaları ile devam edilmiş olup son bölümde ise sosyal medya uygulamalarına yapılan metin sınıflandırma çalışmalarının detaylarına yer verilmiştir.

Literatürde metin sınıflandırma için birçok yöntem ve uygulama alanı bulunmaktadır. Örneğin Türkçe haber metinlerinin sınıflandırılması için Destek Vektör Makinesi, Rastgele Orman ve Naive Bayes sınıflandırma algoritmalarını karşılaştıran çalışmada, 4.900 satırlık haber metinlerinden oluşan veri seti 7 kategoriye ayrılmıştır. İşlemler sonucunda %91 doğruluk oranı ile Naive Bayes algoritması diğer algoritmalara göre en başarılı performansı göstermiştir (Uslu ve Özmen-Akyol, 2021).

Bir internet sitesinin e-ticaret sitesi olup olmadığına karar veren bir uygulama için ön işleme aşamaları gerçekleştirilerek etiketlenen 273 adet site verisi K-En Yakın Komşu ve Naive Bayes algoritmaları ile eğitilmiş, diğer algoritmalara göre Naive Bayes algoritmasının daha iyi sonuç verdiği görülmüştür (Kaşıkçı ve Gökçen, 2019).

Günümüzde metin sınıflandırmada klasik makine öğrenmesi yöntemleri yerine büyük veri setleri ve karmaşık görevler için daha uygun olan derin öğrenme yöntemleri daha fazla kullanılmaktadır. Türkçe haber metinlerinin sınıflandırılması için Konvolüsyonel Sinir Ağları ve Word2Vec metodu kullanılarak yapılan metin sınıflandırma çalışmasında, klasik makine öğrenmesi sınıflandırma algoritmalarından daha iyi bir performans (%93,3 doğruluk) elde edildiği belirtilmiştir. (Acı ve Çırak, 2019).

10.517 e-postadan oluşan veri seti ile, alınan e-postaları önemine göre sınıflandırmak için Word2Vec algoritması kullanılmıştır. 200 e-postadan oluşan test verisi ile sistem başarısı test edilmiş ve %91 oranında doğruluk ile başarı elde edildiği ifade edilmiştir (Sel ve Hanbay, 2019).

Türkçe dilinde yazılan bilimsel metinlerin sınıflandırılması ile ilgili gerçekleştirilen çalışmada önceden eğitilmiş Türkçe bir BERT modeli üzerinde ince ayar yapılmış ve model %96 doğruluk oranı göstermiştir (Özkan ve Kar, 2022).

BERT modeli ve geleneksel makine öğrenmesi modelleri kullanılarak dört farklı deney ile metin sınıflandırma yapılmıştır. Çalışmada sosyal medyada paylaşılan iletiler, film dizi eleştirileri, haber içerikleri gibi veri setleri üzerinde karşılaştırmalı analizler gerçekleştirilerek sonuçlar sunulmuş BERT modelinin diğer modellere göre başarı gösterdiği görülmüştür (González-Carvajal ve Garrido-Merchán, 2020).

Web sitesi URL'lerinden çıkarılan metinler üzerinde önceden eğitilmiş BERT modeli ile sınıflandırılan çalışmada %98 doğruluk ve %67 F1 skoru elde edildiği belirtilmiştir. (Çepni vd., 2023).

Sosyal medya fenomenlerini ve gönderilerini; Naive Bayes, K-En Yakın Komşu, Destek Vektör, Rastgele Orman ve BERT modelleri ile sınıflandırarak karşılaştırılan çalışmada, BERT modelinin diğer modellere göre fenomenleri %98, gönderilerini ise %96 doğruluk ile daha başarılı bir şekilde sınıflandırdığı belirtilmiştir (Kim vd., 2020).

Instagram yorumlarını otomatik olarak sınıflandıran sistem için Destek Vektör Makineleri (Support Vector Machine kısaca SVM) ve Evrişimli Sinir Ağı (Convolutional Neural Network kısaca CNN) algoritmaları karşılaştırılmıştır. %84,23 doğruluk oranı ile CNN algoritmasının daha iyi sonuç verdiği belirtilmiştir (Prabowo ve Purwarianti, 2017).

Yapılan literatür taramasında, metin sınıflandırma alanında çeşitli algoritmalar ve derin öğrenme modellerinin başarıları vurgulanmıştır. Bu çalışmalar incelendiğinde klasik makine öğrenmesi yöntemlerinin ve GaussianNB (Gaussian Naive Bayes), K-En Yakın Komşu (K-Nearest Neighbors kısaca KNN), Destek Vektör Sınıflandırması (Support Vector Classifier kısaca SVC) ve Rastgele Orman (Random Forest) gibi

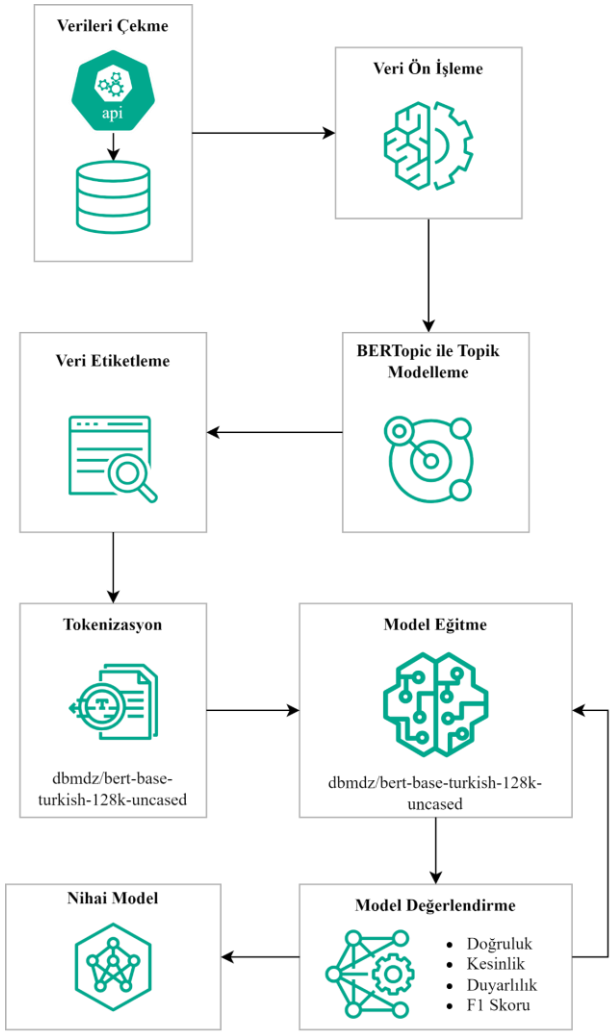
yaygın olarak kullanılan sınıflandırma modelleri kullanılarak geliştirilen modellerde en fazla %40.70 başarı oranı elde edildiği gözlemlenmiştir ve bu modellerin çalışmamız için yeterli doğruluk oranlarını veremeyeceği sonucu çıkarılmıştır. Özellikle BERT modeli, Türkçe metin sınıflandırılmasında yüksek doğruluk oranları elde ederek literatürde öne çıkmıştır. Literatürde BERT modeli ile yapılmış sınıflandırma çalışmalarının doğruluk oranları yapılan çalışmanın başarıya ulaşmasına referans olmuştur. Çalışmamız, BERT modelinin sosyal medya analizleri ve metin sınıflandırma alanındaki uygulamalarına yeni bir perspektif kazandırarak literatüre önemli bir katkı sağlamaktadır.

Literatürde Türkçe dili için metin sınıflandırma çalışmalarının sayısı oldukça sınırlıdır. BERTurk ile yapılan bu çalışmada, yalnızca gönderi açıklama metinleri kullanılarak %92 doğruluk değeri elde edilmiştir. Bu sonuç, benzer çalışmalarda (Kim vd., 2020) rapor edilen %60 doğruluk değerini önemli ölçüde aşmaktadır. Dolayısıyla, bu çalışma hem BERTurk'ün etkinliğini vurgulamakta hem de gönderi açıklama metinlerinin derin öğrenme modelleri için güçlü bir veri kaynağı olabileceğini ortaya koymaktadır.

Türkçe dilinde metin sınıflandırma çalışmaları ve fenomen-ürün hedef kitle eşleştirmesi konusunda literatürde makine öğrenmesi yöntemleri ile yapılmış çalışmaların yeterli seviyede olmaması bu alanda yapılacak çalışmalara duyulan ihtiyacı göstermektedir. Çalışmamız, Türkçe dilinde yapılacak metin sınıflandırma çalışmaları için referans oluşturmaktadır. Sosyal medya pazarlamasında doğru fenomen-hedef kitle eşleşmesini sağlayarak, markaların daha etkili kampanyalar oluşturmasına olanak tanımaktadır. Literatüre yenilikçi bir yaklaşım sunmakta olan bu çalışma, ilgili amaçla yapılacak gelecekteki çalışmalara önemli bir referans oluşturmaktadır.

3. Materyal ve Metod (Material and Method)

Bu bölümde çalışma kapsamında kullanılan yöntemler açıklanmıştır. Çalışmada kullanılan metodoloji Şekil 1'de verilmiştir.



Şekil 1. Metodoloji

3.1 Veri Toplama

Çalışma kapsamında kullanılacak olan veri setini oluşturmak için, kişisel verilerin ihlalini önlemek adına Instagram hesapları halka açık olan fenomenlerin gönderileri elde edilmiştir. Instagram hesapları halka açık olan fenomenlerin gönderileri, Apify adlı üçüncü taraf hizmeti olan veri kazıma (web scraping) platformu kullanılmıştır. Apify, kullanıcıların e-ticaret siteleri, haber siteleri, sosyal medya platformları gibi çeşitli web sitelerinden veri çekmelerine olanak tanıyarak bu verilerin çeşitli amaçlar doğrultusunda kullanılabilmesini mümkün kılar (Apify, 2022). Bu çalışma kapsamında Apify, Instagram üzerinden 1005 farklı fenomenin gönderilerinin çekilmesinde kullanılmıştır. Böylelikle çalışma kapsamında elde edilen veri seti, güvenilirlik ve etik kurallara uygunluk çerçevesinde oluşturulmuştur.

Bu yöntemle toplamda 1.005 farklı fenomenin 647.951 adet gönderi verisi elde edilmiştir. Veri ön işleme adımında sonra analiz için toplamda 610.600 gönderi verisi, çalışmada kullanılan nihai veri setini oluşturmuştur.

3.2 Veri Ön İşleme

Elde edilen veri setine, model eğitimi ve analizler için veri ön işleme adımı uygulanmıştır. Bu adımdaki amaç, model performansını maksimize etmek amacıyla daha kaliteli bir veri seti elde etmektir. Bu bağlamda, fenomen gönderilerinin açıklama metinleri işlenmiştir. Uygulanan adımlar Şekil 2'de verilmiştir.

Şekil 2. Veri ön işleme adımları

Öncelikle, açıklama metin uzunluğu 30 karakterden kısa olan veriler, veri setinden çıkarılmıştır. Türkçe dilinde ortalama kelime uzunluğu göz önüne alındığında, 30 karakter genellikle yaklaşık 4 ila 5 (Dalkılıç vd., 2003) kelimeye denk gelmektedir. Bu nedenle, 30 karakterden kısa metinlerin veri setinden çıkarılması, yeterli bilgi ve bağlam sağlayan daha anlamlı açıklamaların analiz edilmesine olanak tanımaktadır. Bu seçim, metinlerin yeterli bilgi içermesini sağlamak ve analizde daha anlamlı sonuçlar elde etmek amacıyla yapılmıştır. Daha sonra, açıklama metinlerinde herhangi bir büyük harf olmaması adına metinlerin tamamı küçük harflere dönüştürülmüştür çünkü kullanılacak model büyük-küçük harf hassasiyeti taşımamaktadır. Bu dönüşüm, metinlerdeki büyük ve küçük harf farklılıklarını ortadan kaldırarak, modelin tüm metinleri aynı şekilde değerlendirmesini ve karşılaştırmasını sağlar. Bu nedenle, büyük harflerin küçük harflere dönüştürülmesi, modelin doğruluğunu ve işlem sürecini iyileştirmek adına uygulanmıştır. Noktalama işaretleri/emojiler kaldırılmıştır. Bu adım, metinlerin tutarlılığını artırmak ve dil modelinin sadece anlamlı kelimelere odaklanmasını sağlamak için önemlidir. Noktalama işaretleri ve emojiler, modelin analizini karmaşıklaştırabileceğinden, bunların temizlenmesi gereklidir. Ek olarak gereksiz boşluklar ve özel karakterler de temizlenerek tüm açıklama metinleri aynı formata getirilmiştir. Sonrasında, gönderi etiketi (hashtag) içermeyen gönderiler veri setinden çıkarılmıştır. Çünkü sosyal medya üzerinden ürün veya hizmet tanıtımı yapan fenomenler, ürün veya hizmetin reklam olduğunu belirtmek zorundadır ve aynı zamanda reklamı yapılan gönderinin daha fazla kişiye ulaşması amacıyla gönderi etiketi kullanma eğilimindedirler. Bu nedenle, gönderi etiketi içermeyen verilerin veri setinden çıkarılması, reklam hedef kitlesini daha doğru bir şekilde belirlemede önemli bir rol oynamaktadır. Bu adım, veri setinin kalitesini artırmaya ve sonuçların doğruluğunu artırmaya yönelik bir önlem olarak uygulanmıştır.

3.3 Veri Etiketleme

Veri ön işleme adımından sonra elde edilen nihai veri seti, fenomen gönderilerinin açıklamalarını baz alarak verileri etiketlemek amacıyla BERTopic modeli ile kümelenmiştir.

BERTopic mimarisi, doğal dil işleme alanında yaygın olarak kullanılan metin konularının temsillerini (topic) tespit etmek ve bu konuları kümeler halinde gruplamak için kullanılan BERT dil modeli mimarisine dayalı bir modeldir (Grootendorst, 2022).

3.4 Mimari ve Model

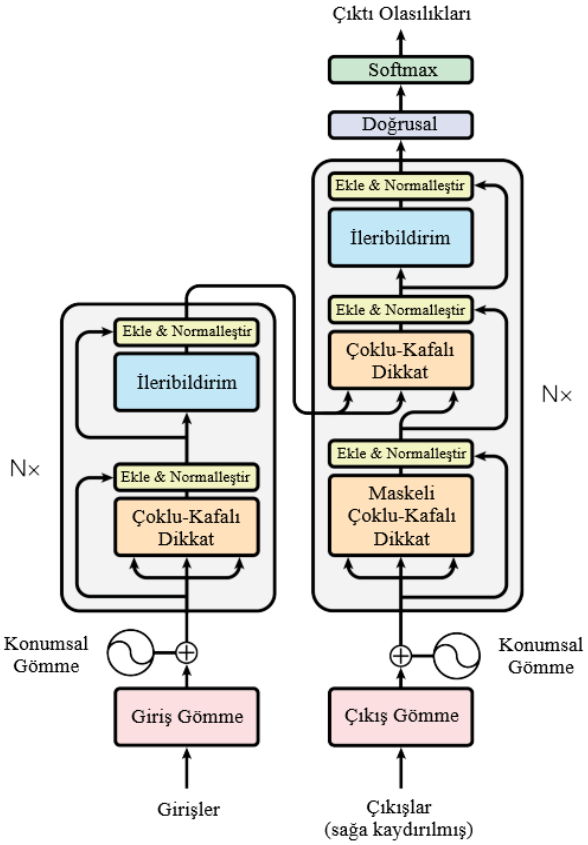
Çalışma kapsamında iki farklı amaçla iki farklı dönüştürücü (transformer) mimarisi kullanılmıştır;

1. Binlerce satır veriden oluşan veri setini önceden belirlenmiş kategoriler doğrultusunda etiketlemek için BERTopic mimarisi kullanılmıştır.
2. Veri seti etiketleri elde edildikten sonra sınıflandırma modelinin fenomenlerin hitap ettiği hedef kitle doğrultusunda sınıflandırabilmek için BERT mimarisi kullanılmıştır.

3.4.1 Dönüştürücü Mimarisi (Transformer Architecture)

Dönüştürücüler, doğal dil işleme ve diğer sıralı veri işleme görevlerinde kullanılan bir sinir ağı mimarisidir. Özellikle, uzun mesafe bağımlılıkları ele almak ve büyük veri kümeleri üzerinde paralel işlem yapmak için etkilidir. Bu mimari, dikkat mekanizmasını (attention mechanism) içeren bir yapıya sahiptir ve daha önceki dil modellerinden önemli ölçüde farklılık gösterir (Vaswani vd., 2017). Dönüştürücü mimarisi, birçok tekrarlayan katman içerir ve her bir katman, birbiriyle bağlantılıdır. Her katman, dikkat mekanizmasını kullanarak girdi verilerini işler. Dikkat mekanizması, her bir girdi öğesinin, diğer tüm öğelerle olan ilişkisini hesaplar ve bu ilişkilere göre ağırlıklar

atar. Bu sayede, her bir öğenin önemi belirlenir ve dikkate alınır. Dönüştürücü mimarisi, genellikle bir kodlayıcı (encoder) ve bir çözücü (decoder) olarak iki ana bileşenden oluşur (Aitken vd., 2021). Kodlayıcı, girdi verilerini temsil eden vektörler oluştururken, çözücü, bu vektörleri hedef çıktılara dönüştürür. Her bir bileşen, birçok tekrarlayan katmana sahiptir ve her katman, birden fazla dikkat mekanizması içerir. Dikkat mekanizması, girdi vektörlerinin birbiriyle olan ilişkilerini hesaplar. Her bir girdi vektörü, diğer tüm vektörlere olan benzerliklerine göre ağırlıklar alır. Bu ağırlıklar, her bir vektörün diğerleri üzerindeki etkisini belirler. Özellikle, uzun mesafe bağımlılıkları ele almak için etkilidir ve önceki dil modellerinden daha iyi sonuçlar verir. Dönüştürücü mimarisi, dil modelleri ve diğer sıralı veri işleme görevlerinde genellikle kullanılır. Büyük metin veri kümeleri üzerinden eğitilmiş olan modeller, genellikle az miktarda etiketlenmiş veri ile yüksek doğruluk sağlar.

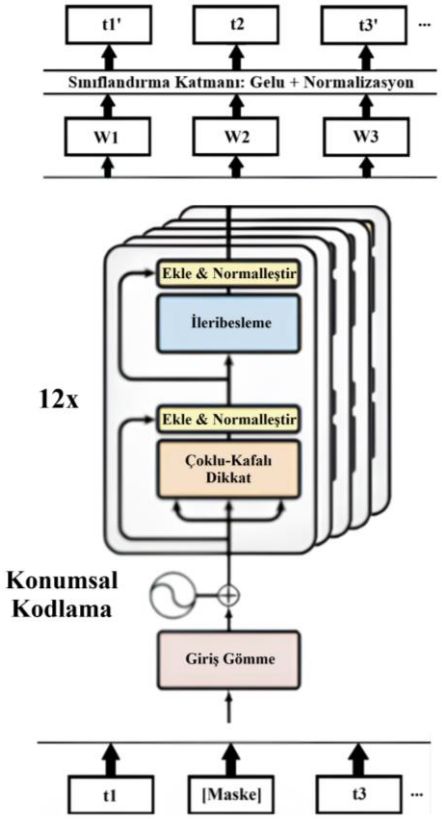


Şekil 3. Dönüştürücü - model mimarisi (Vaswani vd., 2017)

BERT, dönüştürücü mimarisi ile geliştirilmiştir ve sadece kodlayıcı kısmını kullanarak metni çift yönlü (hem ileri hem de geri yönde) analiz eden bir dil dönüştürücü modelidir. Dönüştürücü model mimarisi Şekil 3'te verilmiştir.

3.4.2 BERT Mimarisi

BERT, Google tarafından geliştirilen ve doğal dil işleme alanında devrim niteliğinde bir ilerleme olarak kabul edilen bir modeldir. Bu model, büyük miktarda metin verisi üzerinde ön eğitilmiş bir dil modelidir ve sıralı veri işleme için son derece etkilidir. BERT, hem sol hem sağ bağlamı dikkate alan bir biçimde kelimeleri bir araya getirir. Bu, metin içerisindeki her kelimenin anlamını, hem önceki hem de sonraki kelimelerin bağlamından elde eder. Bu şekilde, metnin daha geniş bir bağlamını anlayabilir ve daha derin bir semantik anlam çıkarabilir.



Şekil 4. On iki kodlayıcı bloğa sahip, dönüştürücü tabanlı BERT temel mimarisi (Khalid vd., 2021)

Şekil 4'te on iki kodlayıcı bloğa sahip, dönüştürücü tabanlı BERT temel mimarisi verilmiştir. BERT, sadece kodlayıcı kısmını içeren bir dil modelidir. Bu, dönüştürücü mimarisinin sadece kodlayıcı bileşenini içerdiği anlamına gelir. Kodlayıcı, girdi verilerini temsil eden vektörler oluşturur, ancak bu vektörlerin nasıl kullanılacağı veya çözümleneceği konusunda herhangi bir bilgi bulunmaz. Bu özellik, BERT'in önceden eğitilmiş bir dil modeli olarak kullanılmasını sağlar. BERT, büyük metin veri kümeleri üzerinde eğitimli olduğu için, genel dil yapısını ve anlamını öğrenir. Ancak, belirli bir görev için kullanılmak üzere eğitilmesi gerekebilir. Örnek olarak, sınıflandırma görevleri için, BERT modelinin kodlayıcı kısmı, sınıflandırma modeliyle birleştirilerek yeniden eğitilir

ve öğrenilmiş temsiller kullanılarak sınıflandırma yapılır. Bu nedenle, BERT modeli, kodlayıcı bileşeninin özelliklerinden yararlanarak çeşitli doğal dil işleme görevleri için kullanılabilir. Kodlayıcı, metin verilerini temsil eden vektörler oluştururken, bu vektörlerin çözümlenmesi veya belirli bir görev için kullanılması, modelin yeniden eğitilmesini gerektirir (Devlin vd., 2018).

3.4.3 BERTopic Mimarisi

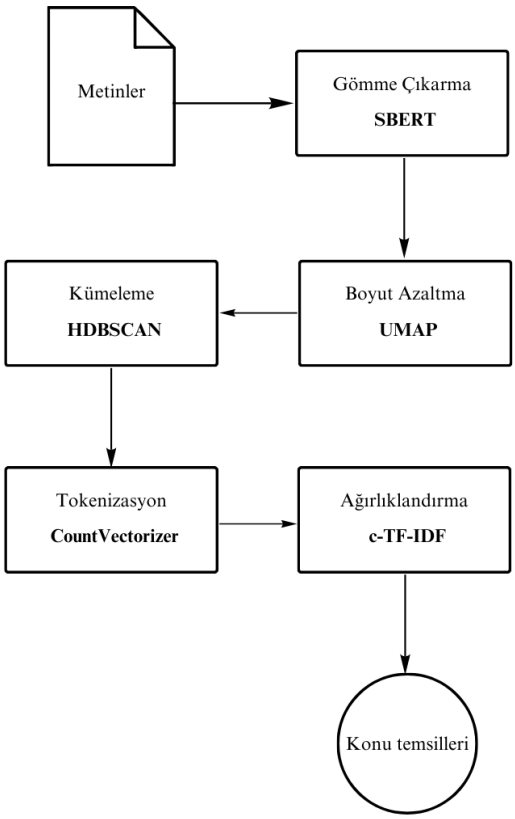
BERTopic modeli, BERT ve Sınıf tabanlı kelime frekansı–ters doküman frekansı (Class-based term frequency- inverse document frequency kısaca c-TF-IDF) tekniklerini kullanarak metin konularını tespit edip kümeleyen bir konu modelleme tekniğidir (Grootendorst, 2022). Temel olarak metinlerden cümle gömmelerini elde ettikten sonra benzer cümleleri kümeleyerek kümeleri oluşturur. Şekil 5'te BERTopic modeli ile küme oluşturma adımları verilmiştir. Kümeler; gömme çıkarma, boyut azaltma, kümeleme, tokenizasyon ve ağırlıklandırma olmak üzere beş temel adımdan geçerek oluşturulmaktadır.

BERTopic, metinlerden cümle gömmelerini elde etmek için Cümlü Dönüştürücü (Sentence-BERT kısaca SBERT) tekniğini kullanmaktadır. SBERT mimarisi, BERT mimarisinin geliştirilmiş versiyonudur. BERT mimarisi cümleleri kelime kelime işlerken SBERT cümleleri tek seferde işlediği için metin semantiklerini daha kapsamlı bir şekilde anlayabilir (Reimers ve Gurevych, 2019).

Kümeleme adımında yüksek boyutlu gömme vektörlerinin kümelenmesi daha zor ve komplike olduğundan, kümeleme adımından önce Tekdüze Manifold Yaklaşımı ve Yansıtması (Uniform Manifold Approximation and Projection kısaca UMAP) kullanılarak cümle gömmelerinin boyutları azaltılır.

Cümle gömmelerinin boyutu azaltıldıktan sonra yoğunluk temelli bir kümeleme algoritması olan Gürültülü Uygulamalar için Hiyerarşik Yoğunluk Tabanlı Uzamsal Kümeleme (Hierarchical Density-Based Spatial Clustering of Applications with Noise kısaca HDBSCAN) algoritması ile benzer metin gömmeleri kümelenir.

BERTopic mimarisi, tokenizasyon adımında ise doğal dil işleme alanında yaygın olarak kullanılan kelimelerin metinde geçme sıklığı ve metin içerisindeki önemini dikkate alarak tokenizasyon işlemi için Sayı Vektörleştirici (CountVectorizer) metin işleme tekniğini kullanır.



Şekil 5. BERTopic modeli ile kümeleme adımları

Sonraki adımda ise elde edilen tokenler, c-TF-IDF tekniği ile kümeleri oluşturulur. Kelime frekansı–ters doküman frekansı (Term frequency- inverse document frequency kısaca TF-IDF) tekniği, metin belgelerini kelimenin alaka düzeyine göre vektörleştirirken, c-TF-IDF aynı işlemi tek bir kategorideki tüm belgeleri tek bir belge olarak ele alır ve yapar. Böylelikle ilgili küme özelinde kümeyi en iyi şekilde temsil eden kelimeler elde edilmiş olur (Liu vd., 2018).

3.5 Model Eğitme

Bu bölümde, model eğitimi aşamasında kullanılan model tanıtılmıştır.

3.5.1 BERTurk ile Sınıflandırma

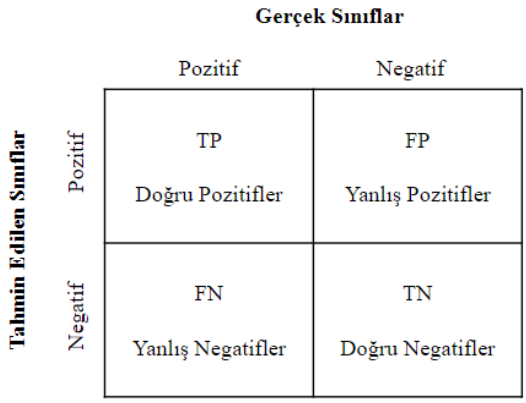
Sıralı veri işleme için BERTurk modelinin kodlayıcı (encoder) kısmını kullanır. Bu kodlayıcı, girdi metin dizisini bir dizi temsil vektörüne dönüştürür. Bu vektörler, girdi metnin anlamını ve bağlamını yansıtır. Ardından, bu temsil vektörleri, bir sınıflandırma katmanına beslenir. Sınıflandırma katmanı, bu temsil vektörlerini alır ve belirli sınıflara ait olasılık değerlerini tahmin eder. Bu, tipik olarak bir "softmax" aktivasyon fonksiyonu ile gerçekleştirilir.

3.6 Model Performans Değerlendirmesi

Sınıflandırma model performansı karışıklık matrisi (confusion matrix), doğruluk (accuracy), kesinlik

(precision), duyarlılık (recall) ve f1 skoru (f1-score) performans metrikleri ile değerlendirilmiştir.

Karışıklık matrisi, sınıflandırma modelinin performansını değerlendirmek için kullanılan, model tahminlerini özetleyen bir performans değerlendirme metriğidir. Sınıflandırma modeli performansı değerlendirilirken modelin yaptığı hataları ve zayıflıkları hakkında daha kapsamlı çıkarımlar yapılmasına yardımcı olur. Karışıklık matrisi Şekil 6'daki gibi ifade edilir.



Şekil 6. Karışıklık matrisi

Karışıklık matrisinde doğru pozitifler doğru şekilde sınıflandırılan pozitif örnekleri, yanlış pozitifler yanlış şekilde sınıflandırılan negatif örnekleri, yanlış negatifler yanlış şekilde sınıflandırılan pozitif örnekleri, doğru negatifler ise doğru şekilde sınıflandırılan negatif örnekleri belirtmektedir.

Bir sınıflandırma probleminde doğruluk metriği, modelin ne kadar örneği doğru tahmin ettiğini ifade eder ve aşağıdaki gibi hesaplanır;

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Kesinlik metriği, bir sınıflandırma modelinin pozitif olarak tahminlediği örneklerin kaç adetinin gerçekten pozitif olduğunu belirtir ve aşağıdaki gibi hesaplanır;

$$Kesinlik = \frac{TP}{TP + FP} \quad (2)$$

Duyarlılık metriği ise sınıflandırma modelinin pozitif olarak tahmin etmesi gereken sınıfların ne kadarını pozitif olarak tahmin ettiğini belirtir ve aşağıdaki gibi hesaplanır;

$$Duyarlılık = \frac{TP}{TP + FN} \quad (3)$$

F1 skoru ise kesinlik ve duyarlılık metriklerinin harmonik ortalaması alınarak hesaplanır ve uç durumların göz ardı edilmemesini, daha doğru analizler yapılabilmesini sağlar. F1 skoru aşağıdaki gibi hesaplanır;

$$F1\ Skoru = 2\ x\ \frac{Kesinlik\ x\ Duyarlılık}{Kesinlik + Duyarlılık} \quad (4)$$

4. Değerlendirme (Evaluation)

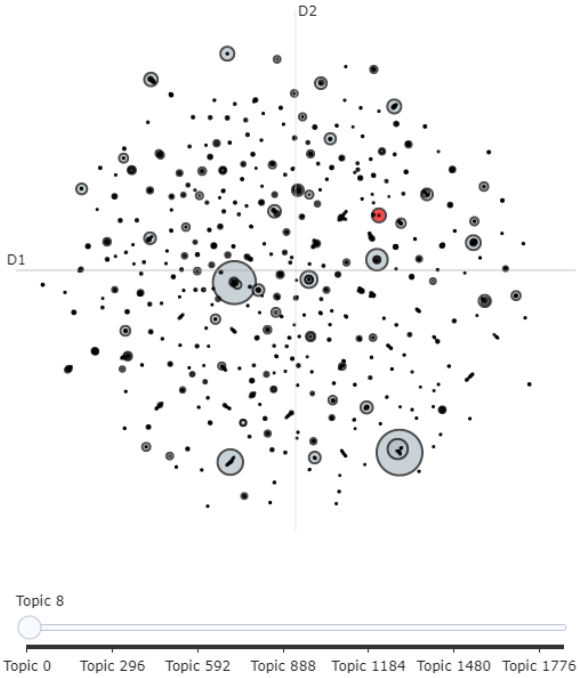
Bu çalışma, fenomenleri Instagram'daki gönderilerinde bulunan açıklamalar doğrultusunda hitap ettiği kategoriyi tanımlayabilmek adına doğal dil işleme teknikleriyle sınıflandırmayı amaçlamaktadır. Bu amaç doğrultusunda öncelikle çalışma kapsamında kullanılan veri seti, Apify ile Instagram hesabı halka açık olan fenomenlerin gönderi bilgilerinin çekilmesi ile oluşturulmuş ve veriyi doğal dil işleme model eğitimine hazırlamak adına veri ön işleme gerçekleştirilmiştir.

Elde edilen veri seti ile BERTopic modelini kullanarak gönderi açıklamalarının kümeleri (topic) oluşturulmuştur. Toplamda 1862 adet farklı küme oluşmuştur. BERTopic modeli tarafından belirli bir küme ile ilişkilendirilemeyen verileri temsil eden "-1" numaralı küme görmezden gelinmiştir. Kümeye ait anahtar kelimeler ve örnek verilerin detaylı incelenmesi ile kümeler önceden belirlenmiş olan 18 kategoriye göre etiketlenmiştir. Tablo 1'de, örnek olması için model tarafından oluşturulan 21 kümenin etiketlenmiş hali, bu kümelere ait veri setindeki örnek sayısı ve kümeye ait anahtar kelimeler verilmiştir.

Tablo 1. Model tarafından oluşturulan ilk 21 kümenin etiketlenmiş hali

Etiket	Kümeler	Örnek Sayısı	Kümelere Ait Anahtar Kelimeler
-	0	21524	seni, sen, ki, hep, ne, şey, anne, diye, hayat, sana
Yiyecek ve İçecek	1	19078	kaşığı, bardağı, yemek, yumurta, su, ekleyip, gr, adet, paket, şeker
Seyahat ve Ulaşım	2	6826	seyahat, travel, gezi, şehir, tatil, antik, deniz, yer, burası, gezgin
Yiyecek ve İçecek	3	4795	kahve, coffee, coffeetime, coffeerem, coffeelover, coffeeoftheday, kahvekeyfi, kahvesi,coffeeculture, coffeeshots
Kitaplar ve Edebiyat	4	4161	kitap, kitabı, kitap önerisi, roman, edebiyat, kitapkurdu, kitaptavsiyesi, kitaplar, kitapları, neokuyorum
Otomobiller ve Araçlar	5	3065	otomobil, otomobiltutkusu, suv, porsche, bmw, arabatutkusu, otoparkcom, ford, arabasevdası, elektrikli
Güzellik ve Egzersiz	6	2229	diyet, kilo, diyetisyen, beslenme, onlinediyet, diyetyemekleri, zayıflama, roket, diyeti, diyetisyenpinardemirkaya
Güzellik ve Egzersiz	7	2051	makeup, makyaj, lipstick, makeuptutorial, ruj, maccosmeticsturkiye, maybelline, matte, makyajı, eyeliner
Güzellik ve Egzersiz	8	2024	saç, saçlarımı, hair, hairstyle, saçlar, saçlarım, saçları, şampuan, bakım, saçbakımı
Güzellik ve Egzersiz	9	1956	cilt, cildi, bakım, serum, cildin, nemlendirici, kremi, ciltbakımı, asit, cilde
Çevrimiçi Topluluklar	10	1917	reelsinstagram, iphone, reels, reelsvideo, instagram, samsung, pro, reelsindia, reelesviral, apple
Çevrimiçi Topluluklar	11	1593	youtube, video, videonun, videoyu, kanalımda, yayında, youtubeda, kanalda, abone, videosu
Güzellik ve Egzersiz	12	1555	gununegzersizi, bacak, egzersiz, kalça, core, kaslarını, egzersizler, hareket, çalıştıran, omuz
İşler ve Eğitim	13	1483	tongucakademi, ogrenci, lgs, öğrenci, okul, ders, teog, öğretmenler, öğretmenlerimizin, seyev
Seyahat ve Ulaşım	14	1402	kamp, kampalani, kampvedogahayati, kampvedogadakiler, camping, kamptagram, kampmudavimleri, kampturkiye, karavan, yolacikyolacik
Haberler	15	1390	galatasaray, futbol, football, nba, fenerbahçe, maç, süperlig, messi, gol, transfer
Güzellik ve Egzersiz	16	1386	modanisa, renk, indirim, far, kodu, rengi, paleti, renkler, palet, makyaj
Ev ve Bahçe	17	1383	dekorasyon, bekliyorumsayfamızı, interiordesign, dekorasyonönerileri, yorumlarınızı, türkkahvesikeyfi, sunumvetarif, interior, balkon
Alışveriş	18	1368	hediye, çekiliş, cekilis, arkadaşınızı, etiketlemek, kişiye, çekillişvar, yapmanız, gerekenler, ceikilisvar
Güzellik ve Egzersiz	19	1357	moda, fashion, outfitoftheday, outfit, elbise, dress, ootd, kombin, style, ekinde
İnsan ve Toplum	20	1242	gelisimadam, girisimcilik, startup, fazlası, internettenparakazanmak, sosyalmedyayonetimi, makemoney, etiketlesosyalmedyauzmani, girisim, gelisim

Model tarafından oluşturulan tüm kümelere ait mesafe haritası Şekil 7 üzerinde verilmiştir. Örnek olarak kırmızı renkle belirtilen kümeye ait anahtar kelimeler "saç, saçlarımı, hair, hairstyle, saçlar, saçlarım, saçları, şampuan, bakım, saçbakımı" olarak listelenmiştir. Yanında bulunan diğer kümeye ait anahtar kelimeler ise "cilt, cildi, bakım, serum, cildin, nemlendirici, kremi, ciltbakımı, asit, cilde" olarak belirlenmiştir. Bu iki kümenin birbirine yakın olmasının nedeni, her iki kümenin de kişisel bakım ürünleri ile ilgili olmasıdır. Saç ve cilt bakımı, kişisel bakım kategorisi altında ortak bir ilgi alanı oluşturur, bu da kümelerin yakınlığını açıklar.



Şekil 7. Kümeler arası mesafe haritası

Başlangıçta veri çekme aşamasında fenomenlerin tüm gönderileri çekilmiştir. Bu gönderiler arasında, ürün veya hizmet tanıtımı içermeyen bazı kişisel gönderiler de bulunmuştur. Örneğin, kişinin kişisel hayatına dair bilgiler ve tatil paylaşımları gibi gönderiler kişisel içeriklere örnek teşkil etmektedir. Veri etiketleme aşamasında bu gönderiler veri setinden çıkarılmıştır. Bu sayede sadece reklam içeren gönderilerden oluşan bir veri seti elde edilmiştir. Kümelere ait anahtar kelimeler incelendiğinde, örnek sayısının azalmasıyla birlikte bu kelimeler arasındaki benzerliğin de azaldığı gözlemlenmiştir. Bu sebeple, kümelere ait örnek sayıları için iki farklı eşik değeri (threshold) belirlenmiştir: Bu değerler 190 ve 140 olarak belirlenmiştir. Belirlenen bu sınır değerleri, her bir kümenin temsil ettiği anahtar kelimelerin içeriklerine dayanmaktadır. Örneğin, 185 örneğe sahip bir kümenin anahtar kelimeleri ("zafer, vatan, şehitlerimizi, ağustos, tayyare, minnetle, oy, zaferbayramı, kutlu, anıyoruz") belirli bir temayı net bir şekilde temsil etmesine rağmen, reklam ile ilgili belirlediğimiz 18 kategoriye tam olarak uymamaktadır. Bu nedenle bu tür kümeler yoğun veri setine dahil edilmemiştir. Benzer şekilde, 178 örneğe sahip bir küme, ("amigurumitarifleri, kisiyeozelhediye, dogumgunuhediyesi, amigurumiteknikleri, amigurimitarifleri, freetarif, crocheting, dogumgunu, elemegi, loveislove") anahtar kelimelerini içermekte ve bu kelimeler kısmen "yiyecek ve içecek" kategorisine denk gelmektedir. Ancak bu küme, "yiyecek ve içecek" kategorisi dışındaki kategoriler ile de örtüşebileceği için yoğun veri setine dahil edilmemiş, ancak orta yoğun veri setine dahil edilmiştir. Orta yoğun veri seti için 140 sınırının belirlenmesindeki temel neden ise bu noktada kümelerin 2 kategoriyi de içermeye başlamasıdır. 140'ın altında, örneğin 136 örneğe sahip bir kümenin anahtar kelimeleri ("flowers, çiçekler, tongucakademi, boardmasters, repostapp, tene, tonguçlamaya, teog, flower") "işler ve eğitim", "ev ve bahçe" ve "hobiler ve boş zaman uğraşları" gibi 2'den fazla kategoriye ait veriler içermekte olup, veri setinin dengesini bozabileceğinden bu küme herhangi bir veri setine dahil edilmemiştir. Kümelere ait örnekler incelendiğinde 140-190 aralığında örneğe sahip olan kümelerin 1 veya maksimum 2 kategoriyi temsil ettiği, 140'dan az örnek bulunan kümelerin ise en az 2 veya daha fazla kategoriyi temsil ettiği görülmüştür.

Sonuç olarak 1 kategoriyi temsil eden kümeleri içerdiği için 190 ve üzeri örneğe sahip kümeler etiketlenerek yoğun veri setine dahil edilmiş, 1 veya maksimum 2 kategoriye ait verileri içerdiği için 140 üzeri örneğe sahip kümeler ise etiketlenerek orta yoğun veri setine dahil edilmiştir. Bu kriterler, daha homojen ve belirli kategorilere net bir şekilde uyan veri setleri oluşturmak amacıyla belirlenmiştir. Oluşturulan veri setlerine ait veri miktarları Tablo 2'de gösterilmiştir.

Tablo 2. Veri setlerine ait veri miktarları

Veri Seti	Eğitim	Doğrulama	Test
Orta Yoğun Veri Seti	124,726	14,506	14,507
Yoğun Veri Seti	85,666	10,111	10,113

Veri setlerine, Türkçe metinler ile ön-eğitimli olan BERTurk modelleri arasından "dbmdz/bert-base-turkish-128k-uncased" modeli ile tokenizasyon uygulanmıştır (tokenizing). Kullanılan BERTurk modeli diğer modellere kıyasla daha büyük ve daha performanslıdır. Türkçe tıbbi metin sınıflandırmasında BERTurk modelinin üstün performansı, yapılan bir çalışmada 0.93 F-skoru ile kanıtlanmıştır; bu, çok dilli BERT modelinin 0.82 F-skoruna kıyasla önemli ölçüde daha yüksektir (Celikten vd., 2021). Tokenizasyon işlemi ile metin verileri sayısal vektör temsillerine dönüştürülerek modelin anlayabileceği formata getirilmiştir.

Tokenizasyon işlemi uygulanmış olan veri setleri, modelin eğitimi ve performansının değerlendirilmesi için %80 eğitim, %10 doğrulama ve %10 test verisi olarak bölünmüştür. Bu bölme işlemi, modelin genel

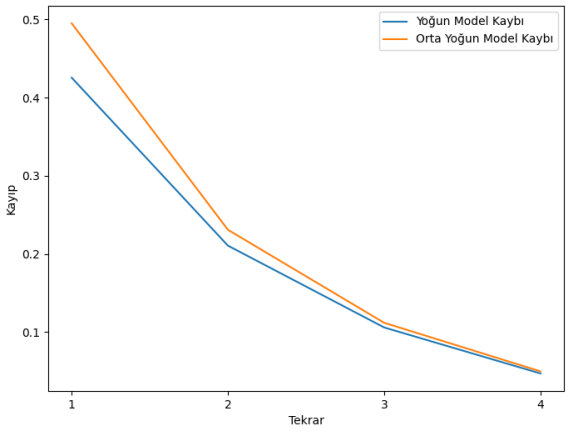
bir performans ölçütü elde etmesine olanak tanırken, aynı zamanda modelin aşırı uyum (overfitting) gibi sorunlara karşı direncini test etmek için de gereklidir.

Eğitim aşamasında, tokenizasyon işleminden geçmiş veriler BERTurk modeli ile eğitilmiştir. Tablo 3'te verilmiş olan parametreler ile eğitim gerçekleştirilmiştir. Bu parametreler, modelin eğitim verisi üzerinde doğru bir şekilde öğrenmesini ve genelleme yapmasını sağlamak için parametre optimizasyonu sonucuna göre seçilmiştir.

Tablo 3. Model eğitim parametreleri

Şifreleyici (Encoder) Model	Tekrar	Yığın Boyutu	Optimize Edici	Öğrenme Oranı
dbmdz/bert-base-turkish-128k-uncased	4	32	AdamW	5e-5

Tablodaki parametreler kullanılarak 2 farklı veri seti ile 2 farkı model eğitilmiştir. Bu modellerden yoğun veri seti ile eğitilen model "Yoğun Model", orta yoğun veri seti ile eğitilen model ise "Orta Yoğun Model" olarak adlandırılmıştır.



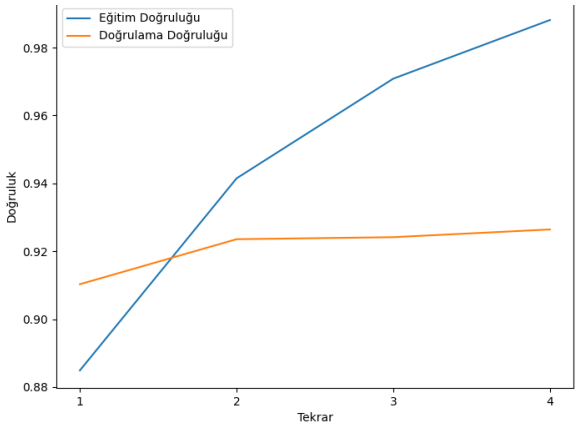
Şekil 8. Yoğun ve orta yoğun model eğitim kaybı

Şekil 8'deki grafikte, "Yoğun Model Kaybı" ve "Orta Yoğun Model Kaybı" olarak iki farklı modelin eğitim kayıplarının dört tekrar boyunca değişimi gösterilmiştir. İlk tekrarda orta yoğun modelin kaybı (yaklaşık 0,5), yoğun modelinkinden (yaklaşık 0,4) daha yüksektir, bu da başlangıçta orta yoğun modelin daha kötü performans gösterdiğini işaret eder. Tekrar sayısı arttıkça her iki modelin de kayıp değerleri azalarak öğrenme sağlanmıştır. Tüm tekrarlar boyunca yoğun modelin kaybı, orta yoğun modelin kaybından daha düşük seyretmiştir, bu da yoğun modelin daha iyi performans sergilediğini gösterir. Son tekrarda ise kayıp değerleri neredeyse eşitlenerek (yaklaşık 0.1) her iki modelin de yeterli eğitim sonrası benzer performans seviyesine ulaştığı görülmüştür.

Tablo 4. Geliştirilen modeller ve skorları

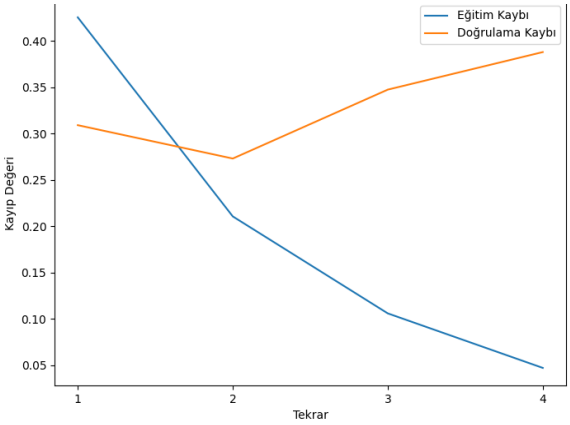
Modeller	F1-Skoru	Kesinlik (Precision)	Duyarlılık(Recall)
Orta Yoğun Model	0,894	0,891	0,897
Yoğun Model	0,912	0,911	0,914

Tablo 4'te geliştirilen iki farklı model arasında yapılan karşılaştırmalar verilmiştir. Yoğun model belirgin bir üstünlük sergilemektedir. Yoğun modelin F1 skoru (0,912) ve kesinlik (0,911) ile duyarlılık (0,914) ölçütleri yüksek ve birbirine yakın değerlerde bulunmaktadır, bu durum modelin hem doğruluğunun hem de hata yapma olasılığının düşük olduğunu göstermektedir. Diğer taraftan, orta yoğun modelin performansı da genel olarak kabul edilebilir düzeydedir (F1 skoru = 0,894, kesinlik = 0,891, duyarlılık = 0,897), ancak yoğun modele kıyasla bir miktar geride kalmıştır. Bu sonuçlar, yoğun modelin özellikle veri dengesizliği gibi zorluklarla daha etkin şekilde başa çıkabildiğini ve sınıflandırma performansının daha istikrarlı olduğunu işaret etmektedir.



Şekil 9. Yoğun model eğitim ve doğrulama doğruluğu

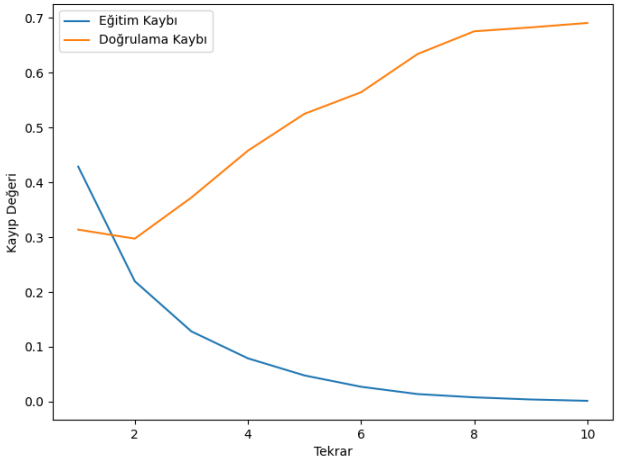
Şekil 9'daki grafikte, yoğun modelin eğitim aşamasındaki doğruluk değerleri görülmektedir. Modelin eğitim doğruluğu her tekrar ile birlikte sürekli artmakta ve dördüncü tekrar sonunda %98'in üzerine çıkmaktadır. Bu, modelin eğitim veri setini oldukça iyi öğrendiğini göstermektedir. Doğrulama doğruluğu ise ilk tekrardan itibaren yavaş bir artış gösterip, ikinci tekrardan sonra yaklaşık %92,5 seviyesinde kalmaktadır. Bunun sebebi ise doğrulama veri setindeki örneklerin eğitim veri setine göre daha çeşitli ve karmaşık olması, modelin doğrulama doğruluğunda daha sınırlı bir artış göstermesine neden olmuştur. Bu durumda, doğrulama doğruluğunun stabil seyretmesi ve aşırı düşüş göstermemesi, modelin genelleme yeteneğinin yeterli olduğunu ve eğitim sürecinde aşırı öğrenme sorunu yaşanmadığını işaret etmiştir.



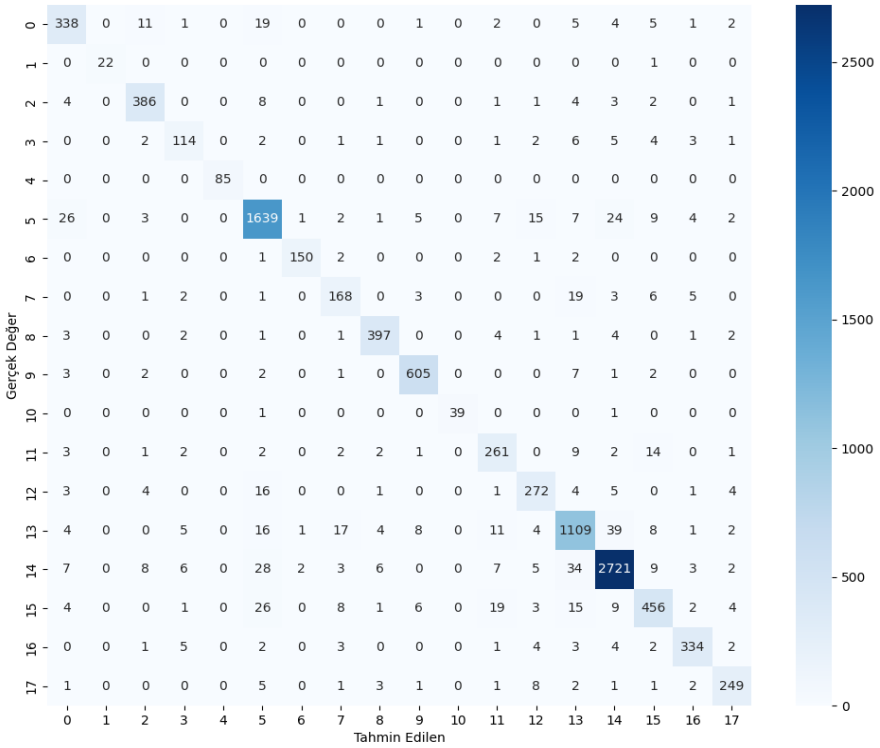
Şekil 10. Yoğun model eğitim ve doğrulama kaybı

Şekil 10'daki grafikte, yoğun modele ait eğitim aşamasındaki kayıp değerleri görülmektedir. Modelin eğitim kaybı her tekrar ile birlikte azalmakta ve dördüncü tekrar sonunda neredeyse sıfıra yaklaşmaktadır. Bu da modelin eğitim veri setini çok iyi öğrendiğini göstermektedir. Doğrulama kaybı ise ilk tekrarda daha düşük başlamakta, ikinci tekrarda bir miktar düşüş göstermekte ancak üçüncü tekrardan itibaren artış eğilimindedir. Doğrulama veri setinde çok çeşitli ve farklı veriler içerdiği düşünüldüğünde, bu artış modelin genelleme yeteneğini zorlayan çeşitliliğe bağlanmaktadır. Bu durum, modelin doğrulama verisindeki hata oranının artmasına neden olmuştur. Ancak, bu çeşitlilik ve doğrulama kaybındaki artışa rağmen, modelin doğrulama doğruluğunun stabil seyretmesi, aşırı öğrenmenin belirgin olmadığını ve modelin geniş veri çeşitliliği karşısında yeterli performans gösterdiğini işaret etmektedir.

Şekil 11'de verilen grafikte 4. tekrardan sonra eğitim kaybı düzenli bir şekilde azalırken doğrulama kaybının belirgin bir şekilde artması, modelin aşırı öğrenme yaşadığını net bir şekilde ortaya koymaktadır. 4. tekrarda doğrulama kaybı, eğitim kaybına kıyasla daha düşükken, takip eden tekrarlar boyunca eğitim kaybı düşmeye devam ederken doğrulama kaybı yükselmiş ve 7. tekrardan itibaren sabitlenmiştir. Bu durum, modelin eğitim verisine aşırı uyum sağladığını ve doğrulama verisinde genelleme yeteneğinin zayıfladığını göstermektedir.



Şekil 11. Yoğun model eğitim ve doğrulama kaybı (10 tekrar



Şekil 12. Karışıklık matrisi

Şekil 12 model performansının karışıklık matrisini göstermektedir. Matrisin satırları gerçek sınıfları, sütunları ise modelin tahmin ettiği sınıfları temsil etmektedir. Diyagonal üzerindeki hücreler doğru tahmin edilen örneklerin sayısını (Doğru Pozitif, TP), diyagonal dışındaki hücreler ise yanlış tahmin edilen örneklerin sayılarını (Yanlış Pozitif, FP ve Yanlış Negatif, FN) gösterir. Örneğin, 0 sınıfının temsil ettiği alışveriş sınıfında doğru tahmin edilen 338 örnek ve 2 sınıfının temsil ettiği ev ve bahçe sınıfında tahmin edilen 11 örnek bulunmaktadır.

5. Tartışma (Discussion)

Bu çalışmada geliştirilen "Yoğun Model" %92,40 doğruluk oranıyla Instagram fenomenlerinin gönderilerini kategorize etmede en yüksek performansı sergilemiştir. Bu başarı, modelin dengeli veri seti üzerinde eğitim alması sayesinde sınıflar arasında adil bir öğrenme gerçekleştirebilmesine ve her bir sınıfı daha doğru tanımlayabilmesine bağlanabilir.

Tablo 5'te verilen sonuçlar, Instagram gönderilerini kullanarak yapılan sınıflandırma çalışmalarının doğruluk oranlarını göstermektedir. Modelin performansı, literatürde yaygın olarak kullanılan sınıflandırma yöntemleri ile karşılaştırıldığında belirgin bir üstünlük göstermektedir. Özellikle GaussianNB, K-En Yakın Komşu, Destek Vektör Sınıflandırması ve Rastgele Orman gibi yöntemlerin doğruluk oranlarının düşük kaldığı göz önüne alındığında, önerilen modelin Instagram gönderilerini sınıflandırmada daha etkili bir yöntem olduğu ortaya çıkmaktadır.

Buna ek olarak, önceki çalışmalarda geliştirilen "Influencer Profiler" modeli %60,90 doğruluk oranı (Kim vd., 2020) ile sınırlı bir başarı elde ederken, bu çalışmada geliştirilen modelin %92,40 doğruluk oranına ulaşması, fenomen-ürün hedef kitle eşleştirilmesinde önemli bir avantaj sunmaktadır. Bu sonuçlar, önerilen modelin sosyal medya pazarlamasında fenomen seçim sürecini daha etkili bir hale getirdiğini ve pazarlama stratejilerinin başarısını artırabileceğini göstermektedir.

Sonuç olarak, önerilen model, diğer çalışmada geliştirilen ve yaygın sınıflandırma algoritmalarından çok daha yüksek bir doğruluk oranı sunarak, Instagram gönderilerini sınıflandırmada etkili bir yöntem olduğunu kanıtlamıştır.

Tablo 5. Instagram gönderilerindeki açıklama metinleri kullanılarak yapılan sınıflandırma çalışmalarının doğruluk oranları (Kim vd., 2020)

Model	Girdi	Doğruluk
GaussianNB	Metin	%40,70
K-En Yakın Komşu	Metin	%38,85
SVC	Metin	%36,20
Rastgele Orman	Metin	%31,80
Influencer Profiler	Metin	%60,90
Orta Yoğun Model	Metin	%90,57
Yoğun Model	Metin	%92,40

Elde edilen performans metriklerine göre 0,92 doğruluk değeri, modelin elde edilen veri setindeki tüm örneklerin %92'sini doğru sınıflandırıldığını belirtir.

Çalışma kapsamında geliştirilen "Yoğun Model" ve "Orta Yoğun Model" arasında "Yoğun Model" en iyi performans gösteren model olmuştur. Yoğun modele ait diğer performans metrikleri Tablo 6'da verilmiştir.

Tablo 6. Model performans değerlendirme metrikleri

Performans Metriği	Hesaplanan Değeri
Doğruluk	0,92
Kesinlik	0,91
Duyarlılık	0,91
F1 Skoru	0,91

Kesinlik ve duyarlılık değerlerinin 0,91 olması, modelin pozitif sınıflandırmalarda güvenilir olduğunu göstermektedir. F1 skoruna bakıldığında ise modelin hem kesinlik hem duyarlılık açısından güçlü olduğunu göstermektedir. Genel olarak model performans metrikleri 0,91'in üzerinde olduğundan sınıflandırma model performansının yüksek olduğu söylenebilir.

Kesinlik ve duyarlılık metriklerinin birbirine yakın ve yüksek olması, modelin pozitif sınıfları doğru bir şekilde tanımladığını ve çok fazla yanlış pozitif üretmediğini göstermektedir. Bu, modelin etkinliğinin dengeli olduğunu ve herhangi bir ölçümde belirgin bir zayıflığın bulunmadığını göstermektedir. Çalışma sonunda elde edilen sınıflandırma modelinin gerçek dünya uygulamalarında güvenilir bir şekilde kullanılabileceği ve tutarlı sonuçlar vereceği yargısı elde edilmiştir.

6. Sonuçlar (Conclusions)

Gerçekleştirilen çalışma ile, Instagram fenomenlerinin gönderi açıklamalarını analiz ederek otomatik bir şekilde fenomenlerin hitap ettiği hedef kitleyi tespit edebilen dönüştürücü mimarisi tabanlı bir sınıflandırma modeli geliştirilmiştir. Türkçe dilinde oluşturulan veri seti üzerinde eğitilen sınıflandırma model performansı değerlendirildiğinde 0,92 doğruluk ve 0,91 F1 skoru değeri elde edilmiştir. Model performans değerlendirmesi sonucunda, gerçek dünya uygulamalarında başarılı ve güvenilir bir şekilde kullanılabilecek bir sınıflandırma modeli elde edildiği söylenebilir.

Bu çalışma, uygulama alanı değerlendirildiğinde literatüre özgün bir katkı sağlamıştır. Gelecek çalışmalarda, gönderilere ait resimler kullanılarak bir resim sınıflandırma modeli geliştirilecektir. Bu model sayesinde, gönderi resimlerinin ait olduğu kategoriler tespit edilebilecektir. İleriki çalışmalarda, daha büyük bir veri seti elde edilerek model performansı büyük hacimli veri üzerinde değerlendirilecek ve gerekli optimizasyonlar yapılacaktır. Ayrıca, veri setinin büyütülmesi ile birlikte kategori sayısında da artış sağlanacak ve böylelikle daha geniş bir alanda veri

sınıflandırılması mümkün olacaktır. Bu kapsamda, resim ve metin verilerinin bir arada kullanımı, modelin doğruluğunu ve genelleme kabiliyetini artıracak, araştırma alanına önemli katkılar sağlayacaktır.

7. Teşekkür (Acknowledgment)

Bu çalışma Küçük ve Orta Ölçekli İşletmeleri Geliştirme ve Destekleme İdaresi Başkanlığı (KOSGEB) tarafından Ar-Ge, Ür-Ge ve İnovasyon projesi kapsamında desteklenmiştir.

Kaynaklar (References)

Acı, Ç. and Çırak, A., 2019. Türkçe haber metinlerinin konvolüsyonel sinir ağları ve Word2Vec kullanılarak sınıflandırılması. Bilişim Teknolojileri Dergisi, 12(3), pp.219-228.

Aitken, K., Ramasesh, V., Cao, Y. and Maheswaranathan, N., 2021. Understanding how encoder-decoder architectures attend. Advances in Neural Information Processing Systems, 34, pp.22184-22195.

Apify. (2022). Web scraping, data extraction and automation. Apify. Retrieved March 22, 2022, from: https://apify.com/

Arslan, E., 2017, August. The effect of social media on marketing. In International Congress Of Eurasian Social Sciences (ICOESS).

Bagwell, K., 2007. The economic analysis of advertising. Handbook of industrial organization, 3, pp.1701-1844.

Carr, C.T. and Hayes, R.A., 2015. Social media: Defining, developing, and divining. Atlantic journal of communication, 23(1), pp.46-65.

Çelıkten, A. and Bulut, H., 2021, June. Turkish medical text classification using bert. In 2021 29th signal processing and communications applications conference (SIU) (pp. 1-4). IEEE.

Çepni, S., Toprak, A. G., Yatkınoğlu, A., Mercan, Ö. B., & Ozan, Ş. (2023). Performance Evaluation of a Pretrained BERT Model for Automatic Text Classification. Journal of Artificial Intelligence and Data Science, 3(1), 27-35.

Çopuroğlu, F., 2022. Fenomen pazarlamanın satın alma niyeti üzerindeki etkisinde menşei ülkenin aracılık rolü. Gaziantep University Journal of Social Sciences, 21(4), pp.2258-2275.

Dalkılıç, G., & Çebi, Y., 2003. Türkçe külliyat oluşturulması ve Türkçe metinlerde kullanılan kelimelerin uzunluk dağılımlarının belirlenmesi. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 5(1), 1-7.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

González-Carvajal, S. and Garrido-Merchán, E.C., 2020. Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Güler, H., Şahinkayası, Y. and Şahinkayası, H., 2017. İnternet ve mobil teknolojilerin yaygınlaşması: Fırsatlar ve sınırlılıklar. Kilis 7 Aralık Üniversitesi Sosyal Bilimler Dergisi, 7(14), pp.186-207.

Karataş, M. and Eti, H.S., 2022. Dijital pazarlama çağında Instagram fenomenlerinin tüketici satın alma davranışlarına etkisi. AJIT-e: Academic Journal of Information Technology, 13(50), pp.184-219.

Kaşıkçı, T. and Gökçen, H., 2014. Metin madenciliği ile e-ticaret sitelerinin belirlenmesi. Bilişim Teknolojileri Dergisi, 7(1).

Khalid, U., Beg, M.O. and Arshad, M.U., 2021. Rubert: A bilingual roman urdu bert using cross lingual transfer learning. arXiv preprint arXiv:2102.11278.

Kim, S., Jiang, J.Y., Nakada, M., Han, J. and Wang, W., 2020, April. Multimodal post attentive profiling for influencer marketing. In Proceedings of The Web Conference 2020 (pp. 2878-2884).

Leung, F.F., Gu, F.F. and Palmatier, R.W., 2022. Online influencer marketing. Journal of the Academy of Marketing Science, 50(2), pp.226-251.

Liu, C.Z., Sheng, Y.X., Wei, Z.Q. and Yang, Y.Q., 2018, August. Research of text classification based on improved TF-IDF algorithm. In 2018 IEEE international conference of intelligent robotic and control engineering (IRCE) (pp. 218-222). IEEE.

Özdemir, S.S., Özdemir, M., Polat, E. and Aksoy, R., 2014. Sosyal medya kavrami ve sosyal ağ sitelerinde yer alan online reklam uygulamalarinin incelenmesi. Ejovoc (Electronic Journal of Vocational Colleges), 4(4), pp.58-64.

Özkan, M. and Kar, G., 2022. Türkçe Dilinde Yazılan Bilimsel Metinlerin Derin Öğrenme Tekniği Uygulanarak Çoklu Sınıflandırılması. Mühendislik Bilimleri ve Tasarım Dergisi, 10(2), pp.504-519.

Öztek, M., Yerden, N.K., Çolak, E. and Sarı, E., 2021. Fenomen pazarlamasında sosyal medyanın rolü ve moda sektörü üzerine bir içerik analizi. Yaşar Üniversitesi E-Dergisi, 16(62), pp.1053-1077.

Prabowo, F. and Purwarianti, A., 2017, November. Instagram online shop's comment classification using statistical approach. In 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 282-287). IEEE.

Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Sel, S. and Hanbay, D., 2019, April. E-mail classification using natural language processing. In 2019 27th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

Terkan, R., 2014. Sosyal Medya Ve Pazarlama: Tüketicide Kalite Yansimasi. Organizasyon ve Yönetim Bilimleri Dergisi, 6(1), pp.57-71.

Uslu, O. and Özmen-akyol, S., 2021. Türkçe haber metinlerinin makine öğrenmesi yöntemleri kullanılarak sınıflandırılması. Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 2(1), pp.15-20.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017.

Attention is all you need. Advances in neural information processing systems, 30.

Yıldırım, Y. and Yıldırım, H., 2022. Dijital Sınırların Sonsuzluğu: Günlük Hayattan Somut Örnekler. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 10(4), pp.1838-1864.