Model Explainability in Turkish Sentiment Analysis

Ayşe Irmak Erçevik ve Halil İbrahim Akgün
Computer Engineering
TOBB ETÜ
ayseirmakercevik@etu.edu.tr
halilibrahimakgun@etu.edu.tr

Özetçe

Yapay Zeka (AI) alanında yapılan ciddi gelişmelerle birlikte başarımı daha yüksek olan karmaşık modeller geliştirilmiştir. Geliştirilen bu modellerden birisi de Transformer yapılı BERT modelidir. BERT tabanlı duygu analizi modellerinin, sözlük tabanlı (LB) ve Makine öğrenimi (ML) yaklaşımları kullanılarak oluşturulan duygu analizi modellerine kıyasla daha başarılı olduğu gözlemlenmiştir. Bu çalışmada Türkçe veriseti üzerinde çalışan BERT tabanlı duygu analizi modellerinin açıklanabilirliği için oluşturulmuş yerel yaklaşımlı model-agnostik bir açıklanabilirlik metodu geliştirilmiştir. Cümlelerdeki sıfatlar, zarflar ve yüklemler belirlenerek olumsuzlarıyla değiştirilmiş ve bu sözcük türlerinin modelin son tahmini üstündeki etkisi gözlemlenmiştir.

Anahtar Kelimeler : Açıklanabilir Yapay Zeka (XAI), Duygu Analizi, Doğal Dil İşleme, LIME, SHAP, Model-Agnostik Açıklama, BERT.

1 GİRİŞ

Son yıllarda Yapay Zeka (AI) alanında ciddi gelişmeler yaşanmıştır. Her yıl daha yeni ve daha karmaşık modellerin geliştirilmesiyle, daha doğru ve kesin sonuçlar elde edilmiştir. Ancak bu durumla beraber, modellerin karar verme sürecinde yürüttükleri karmaşık matematiksel adımları açıklamak zorlaşmıştır. Model karar süreçlerinin doğruluğu, güvenilirliği, savunulabilirliği ve kesinliği sorgulanmaya başlanmış ve son yıllarda da üstünde çokça çalışmalar yapılan Açıklanabilir Yapay Zeka (XAI) kavramı ortaya çıkmıştır.

Duygu Analizi, bir yazı parçasının olumlu veya olumsuz olup olmadığını belirleme sürecidir. Bir duygu analiz modeli, metin analizi için doğal dil işleme (NLP) ve makine öğrenme tekniklerini birleştirerek bir cümle veya tümce içindeki varlıklara, konulara, temalara ve kategorilere ağırlıklı duygu puanları atar. Ancak bir metindeki duygunun olumlu mu yoksa olumsuz mu olduğunu anlamak bir model için oldukça zordur. İnsanlar bile karmaşık ve belirsiz olan duyguları anlamakta zorlanırken bir modelin bir metindeki ironik ifadeleri ve karmaşık anlatım biçimlerini bir insan gibi algılayarak metni olumlu ve olumsuz olarak sınıflandırması zordur.

Tokcaer (2021)'ın çalışmasında 2020 yılına kadar literatürde yer alan sözlük tabanlı (LB) ve Makine öğrenimi (ML) yaklaşımları kullanılarak oluşturulan duygu analizi modelleri yer almaktadır. Bu çalışmada, Makine öğrenimi yaklaşımları kullanılarak oluşturulan duygu analizi modellerinin sözlük tabanlı modelerden daha iyi performans gösterdiği ancak BERT tabanlı modellerin, çoğu ML modellerinden de daha başarılı olduğu (savaş yıldırım, 2021) ve (Utku Umur Açıkalın, 2020) çalışmaları üzerinden anlatılır.

Duygu analizi modelleri her ne kadar duyguyu belirlese de, bu kararı nasıl aldığı hakkında bilgi sağlamazlar. Bu nedenle bu modeller "kara kutu modeller" olarak adlandırılır. Bu modellerin amaçlandığı gibi çalıştığından emin olunması ve kararların mümkün olduğunca şeffaf olması gerekir. Ancak son yıllardan itibaren modellerin yorumlanabilirlik düzeyi, derin öğrenme yaklaşımları ve dil gömme özellikleri(language embedding features) gibi kara kutu tekniklerinin kullanılmasıyla azalmıştır. (Sayedi, 2021)

İngilizce metinler üzerinde çalışan duygu analizi modellerinin açıklanabilirliğinde daha çok LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro, 2020) ve SHAP (SHapley Additive exPlanation) adında (Sikand, 2020) açık kaynaklı Model-Agnostik Açıklama sistemlerinden yararlanılır (Marco Tulio Ribeiro, 2016). Her iki yöntemde de modellerin karar vermesinde faydalı

olduğu tespit edilen metin özelliklerinin önemi ve etkisi görselleştirilir.

Yapılan literatür taramasında Türkçe veri seti üstünde çalışan BERT tabanlı duygu analizi modellerinin açıklanabilirliği için oluşturulmuş yerel yaklaşımlı model-agnostik bir model oluşturulmadığı gözlemlenmiştir. Bunun yanında Hila Chefer (2021)'ın BERT tabanlı modeller için LRP adında Model-Spesifik yaklaşımlı açıklanabilirlik modeli geliştirdiği ancak modelde eksiklikler olduğu gözlemlenmiştir. Bu nedenle araştırma sorumuz şudur:
BERT tabanlı Türkçe duygu analizi modellerine uygulanabilecek bir açıklanabilirlik sisteminin oluşturtulması mümkün müdür ?

2 BACKGROUND

2.1 Yorumlanabilirlik Yaklaşımları:

Algoritma Şeffaflığı: Algoritma şeffaflığı olan bir modelin, veriyi veya öğrenilen modeli dikkate almadan algoritmayı bilmesi beklenir. Modelde uygulanan algoritma modeli nasıl oluşturur sorusuna cevap verir ? (Diogo V Carvalho and Cardoso, 2019)

Modüler Düzeyde Küresel Model Yorumlanabilirliği: Bu seviyede yanıtlanmak istenilen soru, modelin parçalarının modelin tahminlerini nasıl etkilediğidir. Bu tür bir yaklaşım, model özelliklerinin çok boyutlu olduğu veya kara kutu özelliklere sahip modeller için uygun değildir. (Honegger, 2018)

Tek Bir Tahmin için Yerel Yorumlanabilirlik: Bu seviyede, "Model neden bir test verisi için belirli bir tahmin yaptı?" sorusuna cevap vermeye çalışılır. Bir açıklamanın tek bir örneği ayrıntılı olarak analiz edilir. Yerel olarak bakıldığında, bu seviyede açıklama doğrusal olarak birkaç özelliğe bağlı olduğundan, modelin görevinin karmaşıklığını azaltmak mümkündür. (Diogo V Carvalho and Cardoso, 2019)

Bir Tahmin Grubu için Yerel Yorumlanabilirlik: Bir tahmin grubunu açıklamak için iki farklı yaklaşım mevcuttur. İlkinde global yöntemler uygulanarak bütün global olarak ele alınırken, diğer durumda tahminler tek tek ele alınmaktadır. Ardından sonuçlar toplanmaktadır. (Honegger, 2018)

2.2 Açıklamaların Özellikleri:

Sadakat: Açıklamanın makine öğrenimi modelinin tahminine ne kadar yakın olduğunun değerlendirilmesidir. Açıklama ile tanımlanan özellikler, modelin gerçekte seçimini temel aldığı özellikler midir?

Anlaşılabilirlik: Bu bileşen açıklama yöntemi için çok önemlidir. İnsanla ilgili olduğu için öznel bir konudur. Bir kişinin açıklamayı ne ölçüde anlayabildiğini araştırır.

Tutarlılık: Aynı metodoloji ile gerçekleştirilen iki açıklamanın, aynı görev tarafından yönlendirilen ve benzer sonuçlar veren iki model üzerinde yapıldığında ne ölçüde benzer sonuçlar elde ettiğini değerlendirilmesidir.

Kararlılık: Kararlılık sabit bir model için iki benzer örneğin değerlendirilmesine odaklanır. Bu durumda tespit edilen özellikler arasındaki benzerlik değerlendirilir.

Doğruluk: Test kümesinde olmayan verilerin ne kadar iyi tahmin edildiği merak edilir. Doğruluğu yüksek olan makine modelleri için açıklamanın doğruluğunun yüksek olması beklenir. Eğer model düşük doğruluğa sahipse açıklamanın doğruluğunun da düşük olması beklenir.

Kesinlik: Genellikle model tahmininden ne kadar emin olduğunu da verir. Bu kavram modelin kesinliğiyle ilgilidir.

Önem Derecesi: Açıklama yönteminin makine öğrenimi modelinin daha fazla ağırlığa sahip özelliklerini ne ölçüde dikkate aldığının değerlendirilmesidir.

Yenilik: Bu kriter verimlilik ve memnuniyet kavramlarına çok yakındır. Bu kavram yapılan açıklamanın yenilik derecesini ifade eder.

Temsil edilebilirlik: Açıklama yönteminin modeli ne ölçüde değerlendirdiğini ölçer. "Kara kutu modeli bütünüyle mi açıklamış yoksa sadece bir örneğini mi açıklamış?" sorusunun değerlendirilmesidir.

3 İLGİLİ ÇALIŞMALAR

3.1 Duygu Analizi

Duygu analizi, duygu yönelimini belirlemek ve duygusal gücü ölçmek için kullanılmaktadır (Khan, 2019). Duygu analizi birçok araştırmacı tarafından, bir sınıflandırma problemi olarak da tarif edilmektedir. Bu sınıflandırma ikili (pozitif, negatif) olabileceği gibi üçlü (pozitif, negatif, nötr) şeklinde de olabilmektedir (Gezici and Yanıkoğlu, 2018) (Sun et al., 2017).

3.1.1 Duygu Analizinin Uygulama Alanları

Genel hatlarıyla duygu analizinin uygulama alanları günümüzde ve gelecekte şu şekilde sıralanabilir:

- Ürün ve hizmete yönelik değerlendirmeler doğrultusunda en doğru reaksiyonun belirlenerek müşteri ilişkilerinin etkin yönetilebilmesi (Bougie et al., 2003) (Douglas Cirqueira, 2020).
- Hedef varlıklara (politikacılar, filmler, ürünler, ülkeler vb.) yönelik bütün dijital platformlardaki duygu algılarının zaman eksenindeki değişiminin izlenebilmesi (Pang and Lee, 2008) (Mohammad et al., 2013).
- Kullanıcıların duygu durumlarına göre uygun diyalog sistemlerinin geliştirilebilmesi (Ravaja et al., 2006).
- Çevrimiçi eğitim platformlarında, eğitim alan bireyin duygusuna göre sistemin eğitim içeriğini güncelleyebilen akıllı sistemlerin geliştirilebilmesi (Litman and Forbes-Riley, 2004).
- Duygu tonlaması da yapabilen gerçeğe daha yakın metin okuma sistemlerinin geliştirilebilmesi (Francisco and Gervás, 2006) (Bellegarda, 2010).

3.1.2 Duygu Analizi Yaklaşımları

Duygu analizinde kullanılan yaklaşımlar genel olarak iki ana gruba ayrılır (Vohra and Teraiya, 2013): makine öğrenmesi temelli yaklaşımlar ve sözlük temelli yaklaşımlar. Dilin kendine özgü kuralları, içeriğin türü, dildeki mevcut kaynak ve araçların durumu dikkate alındığında, bu iki yaklaşımın güçlü ve zayıf oldukları noktalar vardır. Farklı problemler ve çözüm arayışları sonucunda, iki yaklaşımı barındıran hibrit yöntemler de kullanılmaktadır (Tokcaer, 2021) . Genel olarak, sözlük temelli yaklaşımlar ölçeklenebilirliği ile ön plana çıkarken (Sağlam, 2019), makine öğrenmesi temelli yaklaşımlar ise alana özgü çalışmalarda (Medhat et al., 2014) yoğun olarak tercih edilmektedir. Literatürde üzerinde mutlak uzlaşı sağlanan bir yaklaşım mevcut değildir (Thelwall and Paltoglou., 2011).

3.2 Post-Hoc Açıklanabilirlik Metodları

Post-hoc açıklanabilirlik, tasarımları ile kolayca yorumlanamayan modeller için tasarlanmıştır. Halihazırda geliştirilmiş bir modelin belirli bir girdi için nasıl tahminler ürettiği hakkında anlaşılır bilgi aktarmayı amaçlar. Modelinin çalışma prensibi, model eğitimi sonrasında farklı bir yöntem ile elde edilen bulgular üzerinden açıklanır. "Bize modelin başka neler anlatabileceğini söyler" (Lipton, 2017). Bu yaklaşım modelden bağımsız ve modele özgü yöntemlere ayrılmıştır.

3.2.1 Modelden Bağımsız(Agnostik) Metodlar

Modelden bağımsız metodlardan bazı bilgileri çıkarmak için tüm modellere uygulanacak şekilde tasarlanmıştır. Açıklamayı oluşturmak için yalnızca modelin girdi ve çıktısının bilinmesi yeterlidir. Ağırlıklar veya yapısal bilgiler gibi modelin iç işleyişine erişemeyen yöntemlerdir (Molnar, 2019). Yerel yorumlanabilir modelden bağımsız açıklama tekniği LIME (Marco Tulio Ribeiro, 2016), bu yaklaşıma en iyi bilinen katkılardan biridir. Bu teknik, modeli yorumlanabilir bir doğrusal modelle yerel olarak benzeterek anlamaya çalışmaktadır (Hanjie Chen, 2019). SHAP (SHapley Additive ExPlanation) kara kutu modelleri için bir açıklama sistemidir (Lundberg and Lee, 2017). LIME'a benzer şekilde, modelden bağımsızdır ve farklı alanlara uygulanabilir. Modelden bağımsız metodlardan öne çıkan bir tanesi, girdiye farklı rastgele maskeler uygulayarak ve bunların hedef sınıfın çıktı olasılığını nasıl etkilediğini kontrol ederek girdi önemini deneysel olarak tahmin eden RISE'dır (Vitali Petsiuk and Saenko, 2018).Bizim geliştirdiğimiz metod post-Hoc modelden bağımsız kategorisi altında yer almaktadır.

3.2.2 Modele Özgü(Spesifik) Metodlar

Makina öğrenimi modelinin girdi ve çıktısına ek olarak modelin kendi iç bileşenlerinin analiziyle açıklama sağlanır. Bu metodlar gradyan tabanlı ve alaka tabanlı olmak üzere ikiye ayrılır.

Gradyan Tabanlı Metodlar Gradyan tabanlı yöntemlerde, her katmanın girdisine göre giriş aktivasyon değerleri ile çarpılarak geri yayılım yoluyla gradyan değeri hesaplanır. Bu yaklaşım baz alınarak ilk olarak Gradyan Input yöntemi geliştirilmiştir (Avanti Shrikumar and Kundaje, 2016).Grad-CAM çalışmasının arkasındaki fikir, tahmin için önemli olan bir girdinin değiştirilmesi çıktıyı etkilerken alakasız bir girdinin çıktıyı etkileyemeyecek olmasıdır (Ramprasaath R. Selvaraju, 2017).Sınıfa özgü olan ve tutarlı sonuçlar sağlayan bu yöntem, zayıf denetimli anlamsal bölümleme (Kunpeng Li, 2018) gibi uygulamalar tarafından kullanılır. Bununla birlikte, yöntemin hesaplaması yalnızca en derin katmanların gradyanlarına dayanmaktadır (Danilevsky et al., 2020).

Alaka(Relevance) Tabanlı Metodlar Alaka tabanlı metodlar arasında tahmin fonksiyonunun değerini girdi değişkenleri üzerinde yeniden dağıtmaya çalışan ayrıştırma (decomposition) yaklaşımları bulunmaktadır (Grégoire Montavon, 2016). Layer-wise relevance propagation (LRP), girdi özelliklerinin sinir ağının çıktısına katkısını değerlendirmek için kullanılan bir tekniktir (S. Bach, 2015). Bu teknikte ters yönde yayılma(back propagation) uygulanır. Her bir katmanın aktivasyon puanı aşamalı olarak bir önceki katmana dağıtılır. Geri yayılma işlemi koruma ilkesine(conservation principle) dayanır (W. Samek and Müller, 2021). Belirli bir katmandaki nörona akan tüm alaka(relevance), o nörondan çıkan alakanın toplamına eşittir. Üst katmanların alaka puanları verildiğinde alt katmanların alakalarını hesaplamak için kural bazlı yaklaşımlar, normalizasyon ve pooling işlemleri gerçekleştirilir (Leila Arras, 2017). Bu işlem giriş katmanına ulaşana kadar devam eder. Bu yöntemin yalın halinde alaka puanları, çıktıdan, son attention haritasına kadar atanmaktadır, girdi katmanına kadar atanmamaktadır. (Hila Chefer, 2021) çalışmasında yalın LRP modelini geliştirerek ”tam” LRP modeli yapılmış ve bu şekilde nöronların alaka puanları girdi değişkenlerine kadar aktarılabilinmiştir. LRP metodunda alaka puanlarının, çıktıdan girdi katmanına, geriye doğru yayılmasıyla hesaplanmasından ve yayılma kurallarının seçiminin sezgisel olmasından, güçlü bir teorik gerekçeden yoksun olmasından dolayı girdiler nicel olarak yüksek puan alabilmektedir. (Sandor Berglund, 2021) (Jung et al., 2021).

3.3 Açıklanabilirlik Metodlarını Değerlendirme Yöntemleri

3.3.1 Otomatik Değerlendirme

Bu yöntemde, insan katılımı olmadan bir açıklamanın amacına ulaşmada ne kadar iyi olduğunu değerlendirmek için vekil görevler tanımlanır (Sokol and Flach, 2020). Bu yaklaşım daha önce XAI literatüründe açıklama yöntemlerini değerlendirmek için benimsenmiş ve kabul edilmiştir (Weerts et al., 2019) (Honegger, 2018). Açıklanabilirliği değerlendirmek için sözcükler önem sırasına göre silinerek yerel doğruluk ölçülür. Belirlenen kelimelerin silinmesi ile yanlış tahminlerde doğruluk artar, doğru tahminlerde azalır (Leila Arras, 2016). Tahmin başka bir sınıfa (geçiş noktası) geçmeden önce silinmesi gereken kelime sayısı ölçülür ve belgedeki kelime sayısı ile normalleştirilir (Nguyen, 2018).Bu değere ortalama geçiş noktası(average switching point) denir. Örneğin, 0.10 değeri, tahmin değişmeden önce kelimelerin %10'unun silinmesi gerektiğini gösterir. Bu değer ne kadar düşükse açıklanabilirlik metodu o kadar iyidir. Açıklanabilirlik metodunun en önemli kelimeleri tespit edebildiğini belirtir.

3.3.2 İnsan Temelli Değerlendirme

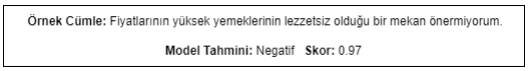
Bir açıklamanın kalitesine ilişkin daha genel kavramları test etmek istendiğinde, insan temelli değerlendirme en uygundur (Lakkaraju et al., 2016)(Lertvittayakumjorn and Toni, 2019). Bir açıklamayı değerlendirmenin bir yolu, insanlardan açıklama ve girdiye dayalı olarak bir modelin çıktısını tahmin etmelerini istemektir. Doshi-Velez and Kim (2017) bu değerlendirmeyi ileri tahmin (forward prediction) olarak adlandırıyor. Bu yöntemde insanlara yerel açıklama yaklaşımları tarafından belirlenen en önemli kelimelerin vurgulandığı metinler gösterilir. İnsanlardandan güvenlerini beşli Likert ölçeğinde belirtmeleri istenir (”Cevabımdan eminim”: kesinlikle katılmıyorum ... kesinlikle katılıyorum). Daha sonrasında insanların güven düzeylerinin ortalamaları alınır.

Bu çalışmada dönüştürücü modelleri kullanıyoruz. Bu modeller kelimelerin birbiriyle ilişkisine bakıyor. Metodu değerlendirirken kelime silme işlemi modelin çalışma mantığına ters düşecekektir. Bu nedenle çalışmamızda otomatik değerlendirme kullanılmayacak, insan temelli değerlendirme kullanılacaktır.

4 PROBLEM

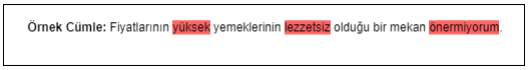
Duygu analizi modelleri bir dökümanın pozitif, negatif veya nötr bir duyguya sahip olduğunu tahmin eder. Bu modeller tahminini neden yaptığını açıklamazlar. Bu nedenle kullanıcılar modeli yorumlamakta zorlanırlar. Modele güvenmeleri için modelin o kararı nasıl aldıklarını bilmek isterler.

Figure 1: Örnek Cümle:



Figür 1'de verilen örnek cümlede model cümlenin negatif duyguya sahip olduğunu 0.97 skorla tahmin etmiştir. Bu kararı almasında hangi kelimelerin ne kadar etkide bulunduğunu açıklamamıştır.

Figure 2: Örnek Çıktı Gösterimi:



Figür 2'de modelin bu kararı almasında etkili olan kelimeler sıcaklık haritasıyla görselleştirilmiştir. Model, tahminini açıklarken girdi cümledeki hangi kelimelerin bu kararı almakta etkili olduğunun çıktısını verebilirse, kullanıcıların modele güveni artar. Kullanıcı bu sayede birçok farklı modeli karşılaştırarak hangi modelin kendisi için iyi olduğuna karar verebilir. Bazı durumlarda model yanlış nedenlerle doğru kararlar verebilir. Yani kullanıcılar için sadece modelin tahmini ve skoru önemli değildir. Aynı zamanda modelin o kararı alırken mantıklı nedenlere dayandırması önemlidir.

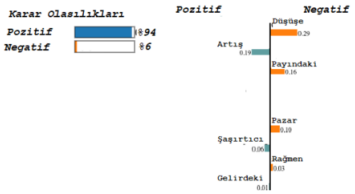
Literatürde ise Türkçe veri seti üzerinde çalışan BERT tabanlı duygu analizi modellerinin açıklanabilirliği için oluşturulmuş yerel yaklaşımlı, model-agnostik bir model oluşturulmamıştır. Bunun yanında Hila Chefer (2021)'ın BERT tabanlı modellerin açıklanabilirliği için oluşturduğu LRP metodunda da ciddi eksiklikler vardır.

5 ÇÖZÜME GENEL BAKIŞ

Bu problem için iki çözüm önerisi oluşturduk. İlk olarak BERT tabanlı duygu analizi modelinin, modele girdi olarak verilen metni sınıflandırma sürecinde, metindeki hangi kelimelerden ne derece etkilendiğini analiz eden ve görselleştiren model-agnostik bir sistem tasarlamak. İkinci olarak Hila Chefer (2021)'ın BERT tabanlı modellerin açıklanabilirliği için oluşturduğu LRP metodundaki eksiklikleri kaptıcak bir yöntem geliştirmek.

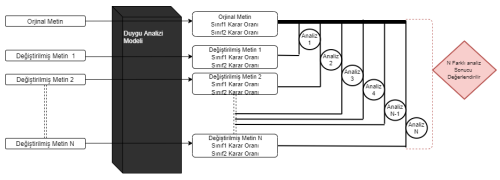
5.1 Çözüm 1 - Model-Agnostik Yapıda Açıklanabilirlik Modeli Tasarlamak:

Tasarlanabilecek açıklanabilirlik modelinin;
Girdileri:
Türkçe Metin Girdisi ve
Daha önceden oluşturulmuş BERT tabanlı duygu analizi modelidir.
Beklenen Çıktısı:
Modele girdi olarak verilen metnin içerdiği kelimelerin, modelin sınıflandırmasına etkisinin analiz edildiği ve analizin görselleştirildiği ekran çıktısıdır.
Örnek:
Metin Girdisi: "Pazar payındaki düşüşe rağmen gelirdeki şaşırtıcı artış"
Model: Finansal metnin duyarlılığını analiz etmek için önceden eğitilmiş bir NLP modeli
Beklenen Çıktı:



Bu sistemde, BERT modelinin karar verme algoritmasına herhangi bir müdahalede bulunmadan, modele verilen girdi metninde değişiklik yaparak, model sonucunun bu değişikliklerden nasıl etkilendiği analiz etmeyi ve görselleştirmeyi amaçladık.

Figure 3: Genel Yaklaşım:



BERT modeli Masked Language Modeling (MLM) ve Next Sentence Prediction (NSP) adı verilen iki teknikle eğitilmektedir. MLM tekniğinde, cümle içerisindeki kelimeler arasındaki ilişki üzerinde durulurken, ikinci teknik olan NSP'de ise cümleler arasındaki ilişki kurulmaktadır (Uçar, 2020). Bundan dolayı, modele verilen metin içerisinde değişiklik yapılırken, cümle yapısının ve kelimeler arasındaki sözdizimsel ilişkinin bozulmaması dikkate edilmesi gerektiğini düşündük.

5.1.1 Girdi Metninin Değiştirilmesinde Düşünülen Yaklaşımlar:

BERT modelinde her bir kelimenin karar sonucuna kattığı ağırlık sağındaki ve solundaki kelimelerin ağırlıklarından etkilenir (Jacob Devlin, 2018). Bundan dolayı kelimeler üzerinde yapılması planlanan değişiklik için Figür 4'de verilen genel yaklaşım oluşturduk:

Figure 4: Genel Yaklaşım Sözde Kod:

```
InputModification(text):
 list input_original=[k1,k2,k3,k4,....kN]
// N kelimeden oluşan metin girdisi
list predictions=[]
// Her bir kelime değişimi sonucu modelin döndüğü karar olasılıklarının tutulduğu liste
for (int i=0; i<N; i++)
   {
     input_new = change(i, input_original)
     //orjinal metindeki i. eleman değişir, geri kalan kelimeler aynı kalır
     prediction = model(input_new)
    //yeni metin modele girdi olarak verilerek yeni karar olasılıkları alınır
     predictions.add(prediction)
   }
prediction_original = model(input_original)
//modele girdi olarak orijinal metin verilerek karar olasılıkları alınır
impact_eachWord = analize(prediction_original,predictions)
// her bir kelimenin değişimi sonucu elde edilen karar olasılıkları ile
// orijinal karar olasılıklarının farkına bakılır ve her bir kelimenin etki /sonucu alınır
```

Bu yaklaşım ile her iterasyonda girdi içerisindeki tek bir kelime değiştirilir ve geri kalan kelimeler aynı kalır. Değiştirilmesi planlanan kelimenin komşuları orijinal metindeki gibi kaldığından, kelimenin komşularından gelen ağırlık onlar üzerinde yapılan/yapılacak olan değişimin etkisini içermez sadece kelimenin anlamsal değişiminden kaynaklı etkiyi içerir.

Kelime Değiştirme Yaklaşımları:
İlk olarak hedeflediğimiz yaklaşım her iterasyonda etkisi gözlemlenecek olan kelimenin etkisini model üzerinde sıfırlamaktı.

Yaklaşım 1- Kelime Silinmesi: İlk olarak etkisi gözlemlenecek kelimenin metinden silinmesini düşündük ancak kelimenin silinmesiyle cümlenin bir yapı elamanın (sıfat, zarf, zamir,yüklem...) silinmiş olacağı bundan dolayı da komşu kelimeler ve cümleler arasındaki ilişkiye müdahale edileceğini düşündük. Bu yaklaşımla, girdi olarak verilen orijinal metin sonucunda elde edilen karar olasılıkları ile metinde yapılan değişiklik sonucunda elde edilen karar olasılıkları arasındaki fark cümle yapısının eksilmesinin yarattığı değişimi de içerecekti.

Yaklaşım 2- Kelimenin Nötr Kelime ile Yer Değiştirmesi İkinci yaklaşım olarak, her bir kelimenin model için nötr değeri olan, etkisinin 0 olacağı başka bir kelime ile yer değiştirmesini düşündük. Bu yaklaşımının Naive Bayes gibi sözlük tabanlı (kelimenin model üzerindeki etkisinin sözlükte bulunma sıklığına bağlı olduğu) basit modellerde kullanılabileceğini öngördük. Ancak bu yaklaşımın, BERT gibi kelimenin model üzerindeki etkisinin, diğer kelimelere ve cümle yapılarına bağlı olduğu modellerde başarısız olacağını düşündük. Bu tarz kompleks modellerde, kelimenin modele etkisi, corpusta beraber yer aldığı cümlelere ve birlikte kullanıldığı komşu kelimelerin ağırlıklarına bağlı olduğu için her bir kelime için aynı yapıda nötr kelime bulmanın ölçeklenebilir olmayacağı tanısına vardık.

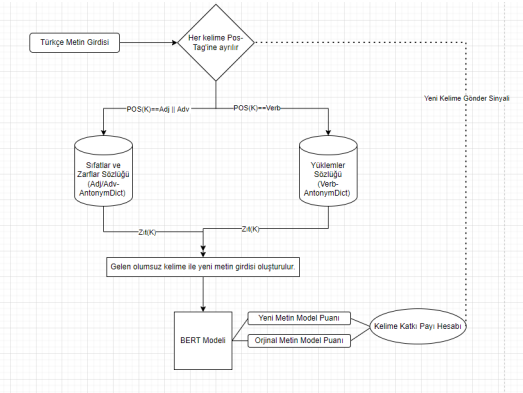
Yaklaşım 3- Kelimenin Zıttı Kelime ile Yer Değiştirmesi Bu yaklaşımda, kelimelerin zıt anlamlarının kelimenin metinle ilişkisini 180 derece değiştireceğini varsaydık. Bu yaklaşımda, orijinal metin sonucunda elde edilen karar olasılıkları ile metinde yapılan değişiklik sonucunda elde edilen karar olasılıkları arasındaki fark sadece kelimenin semantik değişiminin (komşulardan gelen ağırlıkların değişiminin) göstergesi olacaktı. Bunun yanında cümlenin yapısı da korunacaktı. Bu yaklaşımdaki problem ise her bir cümle yapısının zıttının olmamasıydı. Sıfatların ve yüklemlerin zıt anlamlıları ya da olumsuz anlamlıları olsa bile isimlerin zıt anlamlıları yoktu. Ancak Türkçe bir cümlenin duygusunu ifade eden yapıların sıfatlar ve yüklemler olduğu düşünüldüğünde, (Kerimoğğlu, 2016) ideal bir modelin duygu analiz sürecinde isimlerden ve diğer yapılardan daha az etkilenmesini bekledik. Bundan dolayı bu yaklaşımda kelimelerin etkisini analiz ederken sıfat ve yüklem kelime türlerinin sonuca etkisini ayrı ayrı hesaplamaya, diğer yapıdaki kelimelerin etkisini ise kümülatif olarak hesaplamaya karar verdik.(Cümlenin karar puanı –(sıfatların ve yüklemlerin karar puanı))

5.1.2 Yaklaşım 3- Detaylı Çözüm Önerisi:

Bu aşamada çözüm önerisi için gerekli yapılar:
Girdiler:
Türkçe Metin
Sistem için gerekli diğer yapılar:
a) BERT Türkçe Duygu Analizi Modeli (Demir, 2021)
b) Türkçe sıfatların, zarfların ve karşılık gelen zıt anlamlılarının olduğu bizim oluşturacağımız sözlük (Adj/Adv-AntonymDict),
c) Türkçe geçmiş, şimdiki ve gelecek zamanlı bazı

yüklemlerin ve bu yüklemlerin olumsuzluklarının yer aldığı bizim oluşturacağımız sözlük (Verb-AntonymDict).

Figure 5: Sistem Akış Diyagramı:



Çözüm Aşamaları:

1) Sistemimize girdi olarak verilen Türkçe metin doğrudan BERT tabanlı duygu analizi modeline girdi olarak verilir. Modelin, Türkçe metin için tahmin sonucu, sınıflandırma puanı kaydedilir.

2) Model tahmininde hangi kelimelerin katkıda bulunduğunu anlamak için metindeki tüm kritik kelimeleri (Sıfatları, zarfları ve eylemleri) bulmamız gerekiyor. Bu nedenle ilk olarak cümleyi part-of-speech (POS) tag'lerine ayırıyoruz. Böylece her bir kelimenin cümle içindeki dilbilimsel türünü buluyoruz. Daha sonrasında metinde belirlenen her bir kritik kelime için aşağıdaki adımlar izlenir:

- Kelimenin sözcük türü POS-Tag'i aracılığıyla sıfat mı yoksa zarf mı diye bakılır. Eğer sıfat yada zarf ise Adj/Adv-AntonymDict sözlüğü üzerinden zıt anlamlısı çekilir. Eğer kelimenin zıt anlamlısı sözlükte yok ise WordNet üzerinden kelimenin zıt anlamlısı çekilir. Metinde kelimenin yerine sözlükten çekilen zıt anlamlısı koyulur. Yeni oluşturulan metin BERT modeline girdi olarak verilir.Modelin yeni metin için tahmini kaydedilir. Figür 6' da verilen Kelime Katkı Denklemi kullanılarak kelimenin model tahminine katkısı bulunur. Sözcük türü, Sıfat yada zarf değilse bu adım atlanır.

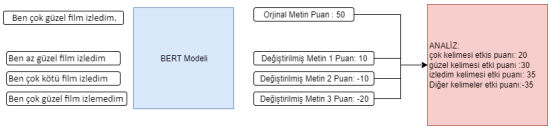
- Kelimenin sözcük türü POS-Tag'i aracılığıyla yüklem mi diye bakılır. Eğer yüklem ve olumlu ise olumsuzu, olumsuz ise olumlusu Verb-AntonymDict sözlüğü üzerinden çekilir. Eğer sözlükte yüklemin olumlu/olumsuz karşılığı yok ise "Yüklem Sözlükte Bulunamadı" hatası verilir ve sonraki sıfat/zarf için aynı işlemler tekrar yapılır. Metinde kelimenin yerine sözlükten çekilen olulu/olumsuz yüklem koyulur. Yeni oluşturulan metin BERT modeline girdi olarak verilir. Modelin yeni metin için tahmini kaydedilir. Figür 6' da verilen Kelime Katkı Denklemi kullanılarak kelimenin model tahminine katkısı bulunur. Sözcük türü Verb değilse bu adım atlanır.

Figure 6: Kelime Katkı Denklemi:

$$\text{O Kelimenin Katkısı} = \frac{\text{İlk Cümlenin Tahmin Sonucu - İkinci Cümlenin Tahmin Sonucu}}{2}$$

3) Metindeki tüm sıfat,zarf ve eylemlerin BERT modelinin sonucuna katkısı hesaplandıktan sonra, Modelin kararını en çok etkileyen kelimeden en az etkileyen kelimeye doğru bir sıralama yapılır ve kelimeler, sonuca etkisine göre görselleştirilir.

Figure 7: Örnek Süreç:



5.2 Çözüm 2 - LRP Metodundaki Eksikleri Gidermek:

Layer-wise relevance propagation (LRP), girdi özelliklerinin sinir ağının çıktısına katkısını değerlendirmek için kullanılan bir tekniktir (S. Bach, 2015). Bu tekniğe göre sinir ağlarının son katmanına ulaşana kadar aktivasyon değerleri hesaplanır. Son katmanın aktivasyon değeri tahmin değerini oluşturur. Daha sonrasında her bir girdinin alakasını hesaplamak için ters yönde yayılma(back propagation) uygulanır. Her bir katmanın aktivasyon puanı aşamalı olarak bir önceki katmana dağıtılır. Geri yayılma işlemi koruma ilkesine(conservation principle) dayanır (W. Samek and Müller, 2021). Belirli bir katmandaki nörona akan tüm alaka(relevance), o nörondan çıkan alakanın toplamına eşittir. Üst katmanların alaka puanları verildiğinde alt katmanların alakalarını hesaplamak için kural bazlı yaklaşımlar, normalizasyon ve pooling işlemleri gerçekleştirilir (Leila Arras, 2017). Bu işlem giriş katmanına ulaşana kadar

devam eder. (Hila Chefer, 2021) çalışmasında LRP metodunun eksik yönlerinden bahsedilmiş ve bu eksiklikler giderilmeye çalışılmıştır.Bunun yanında yapılan geliştirmelerin kaynak kodları GitHub üzerinden paylaşılmıştır (chefer, 2021). Paylaşılan kaynak kodlardan yararlanarak, İngilizce dilinde yapılan bu çalışma Türkçe veriler üzerinde ve BerTurk ile duygu analizi için geliştirilen daha önce eğitilmiş model (savaş yıldırım, 2021) ile çalıştırıldı .

Figure 8: LRP Test Çıktısı 1:

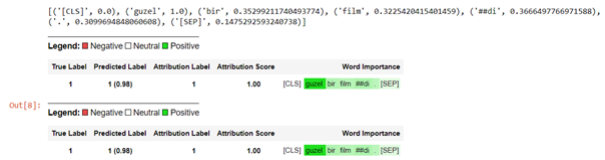
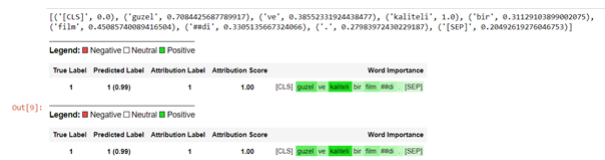


Figure 9: LRP Test Çıktısı 2:



Figür 7 ve Figür 8'de yaptığımız testlerden iki tanesinin sonucu görülmektedir. Geliştirilmiş LRP metodu Türkçe için duygu analizini doğru olarak açıklayabilmektedir. Her bir token'ın sonuca etkisini sayısal olarak bize vermektedir. Ayrıca sıcaklık haritası ile de sonuçları göstermektedir. Fakat bu yöntemde görmüş olduğumuz eksiklik noktalama işaretleri ve stop words'lere yüksek alaka puanı atamasıdır. Birçok dökümanda geçen "ve, veya, bir" gibi kelimelerin duygu analizine etkide bulunmaması gerekir. Şekil 4'te geçen "bir" kelimesine 0.35, Şekil 5'te geçen "ve" kelimesine 0.38, "bir" kelimesine verilen 0.31 alaka puanlarının daha düşük olması gerekirdi. Bu nedenle çözüm önerisi aşağıda verilen aşamalardan oluşmaktadır.

1) Önceden hazırlanmış stop word listesi hazırlanacak.

2) LRP metodunda ters yönde yayılma yapılırken stop words nöronlarına puan verilmeyecek veya sabit küçük bir puan verilecektir.

Bu çözüm yöntemi ile kural bazlı olan LRP metoduna yeni bir kural ekleyip onu geliştirerek kullanıcıya daha açıklanabilir bir yöntem geliştirmeyi amaçlanmaktadır.

Bu proje kapsamında BERT, BERTurk gibi transformer yapılarından oluşan modellerin karmaşık attention mekanizmaları üzerinde değişiklik yaparak açıklanabilirlik modeli geliştirmenin zor olacağı düşünüldü. Bu nedenle projede modellerin algoritmasında değişiklik yapmadan, modele verilen girdinin değiştirilerek, çıktıdaki değişim etkisinin analiz edilmesi yaklaşımı benimsendi.

6 DENEYSEL DEĞERLENDİRME VE SONUÇLAR

Bu bölümde, kullanıcı simülasyon deneyleri üzerinden tasarlanan açıklanabilirlik modelinin sonuçları değerlendirildi. Simülasyon deneylerinin tasarımı için özellikle aşağıdaki araştırma soruları ele alındı: (1)Açıklamalar modeli doğru temsil ediyor mu? (2) Model Ölçeklenebilir mi? (3) Açıklamalar kullanıcıların model tahminine olan güvensizliği ortadan kaldırıyor mu? (4) Oluşturduğumuz sistem model-agnostik mi?

6.1 İmplementasyon Detayları

Sistemimizde BERT tabanlı 2 farklı duygu analizi modeli kullandık. Bunlardan ilki savaş yıldırım (2021)'ın Türkçe metinlerin duygu analizi için eğittiği modeldir. Diğeri ise gürkan şahin (2022)'in Türkçe ürün yorumları duygu sınıflandırması için eğittiği modeldir.Kullandığımız iki modeldeki parametreler şu şekildedir:

- max seq length = 128
- per gpu train batch size = 32
- learning rate = 0.00002
- num train epochs = 3.0

Sistemimize girdi olarak verilen Türkçe metinlerin sıfat ve zarf türündeki sözcüklerinin tespitinde Onur Yılmaz'ın Türkçe sözcükler üzerinde çalışan **POS etiketleme** aracını (Yılmaz, 2016) kullandık. Araç, test veri setimizdeki yorumların sıfat(adj.) ve zarf(adv.) türündeki sözcüklerin tespitinde %96,2 doğruluk oranı gösterirken, eylem ve eylemsi tespitlerinde %8.9 doğruluk oranı gösterdiğini fark ettik. Araç, emir kipi dışında kullanılan bütün eylemleri "NP" olarak etiketlemekteydi. Bunun üzerine yorumlardaki eylem ve eylemsilerin tespiti için **Zemberek-python** kütüphanesinin **TurkishMorphology** modülünü kullandık. (harun Loodos, 2021). Bu modül ile test yorumlarındaki eylem ve eylemsileri %92,7 doğruluk oranıyla tespit edebildik. Sözcük türü

tespit edilen her bir sözcüğün zıt anlamlısına Bakay et al. (2021) tarafından oluşturulan **Türkçe WordNet** üzerinden ulaştık (Software, 2022). Bu wordnet üzerinden test yorumlarındaki sıfatların ve zarfların %76'sının zıt anlamlı karşılığını (antonym ilişkisi ile bağlı olduğu kelimeleri) tespit edebildiğimizi gözlemledik. Bunun üzerine **Türkçe WordNet**'le birlikte yorumlardaki sıfatların ve zarfların %82.6'sını tespit eden ancak günlük 100 request hakkımız olan **Lexicala Web API** (K-DICTIONARIES, 2020) kullanmaya karar verdik. Türkçe WordNet ve Lexicala Web API sonuçlarına kendimizde zıt anlamlı kelime çiftleri ekleyerek **5.1.2 Yaklaşım 3- Detaylı Cözüm Önerisi** başlığı altında yer alan **Adj/Adv-AntonymDict** sözlüğünü oluşturduk. Bununla beraber test orumlarımızda bulunan eylem ve eylemsilerin çekimlerini göz önüne alarak olumlu/olumsuz eylem çiftlerini içeren **Verb-AntonymDict** sözlüğünü oluşturduk.

6.2 Veri Kümesi

Sistemimizin başarımının ölçülmesinde film ve kitap yorumlarını içeren bir test veri seti hazırladık. Veri setimiz 10 pozitif, 10 negatif olmak üzere 20 yorumdan oluşmaktadır. Film yorumları beyazperde.com sitesinden kişilerin çeşitli filmler için yaptıkları yorumlardan aldık. Kitap yorumları ise, hepsiburada.com sitesinden çeşitli kişilerin satın aldıkları kitaplar için yazdıkları değerlendirmelerden oluşturduk. Test yorumlarındaki sıfatlarda, zarflarda ya da eylemlerde bulunan yazım hatalarını sisteme vermeden önce düzelttik. Bunun dışında yorumlar üzerinde bir değişiklik yapmadık.

6.3 Referans Yöntemler

Transform'lar için açıklanabilirlik yöntemleri, karmaşık attention mekanizmaları nedeniyle zordur. Bu nedenle bunları kara kutu yöntemleriyle karşılaştırmayı uygun bulduk. Oluşturulması planan açıklanabilirlik metodu için LRP baseline olarak seçilmiştir. Bölüm 3.2.2'nin altında bulunan alaka tabanlı metodlarda LRP hakkında detaylı bilgi verilmiştir. Hila Chefer (2021) çalışmasında LRP yönteminin gelişmiş halini GitHub üzerinden paylaşmıştır (chefer, 2021). Paylaşılan kaynak kodlardan yararlanarak, İngilizce dilinde yapılan bu çalışma Türkçe veriler üzerinde ve BerTurk ile duygu analizi için geliştirilen daha önce eğitilmiş modeller (savaş yıldırım, 2021) (gürkan şahin, 2022) ile çalıştırdık.

6.4 Değerlendirme Ölçütleri

Sistemimizin başarımını değerlendirmek için iki farklı yöntem kullandık. (1) Test veri setindeki yorumları kendi sistemimize ve baseline modelimiz olan LRP modeline girdi olarak verdik. İki yaklaşımın her bir test girdisi için oluşturdukları sonuçları belli ölçütler doğrultusunda karşılaştırdık. Bu yöntemle yaklaşımımızın baseline modelimizle ne kadar paralel çalıştığını gözlemleyerek modelin doğruluğu konusunda bir bilgi sahibi olmaya çalıştık. (2) Test veri setimizde bulunan 5 yorum için kendi sistemimizin sonuçları ve baseline modelimiz olan LRP'nin sonuçlarını içeren bir anket hazırladık. Ankette, cevaplayıcılardan her iki yaklaşım için yaklaşımın sonucuna ne kadar güvendiklerini 1 ila 5 arasında puanlamalarını istedik. 1 yaklaşıma hiç güvenmediklerini, 5 ise çok güvendiklerini temsil etmekteydi. Bu yöntemle insan bakış açısıyla modelimizin doğruluğunu ve güvenilirliğini ölmeyi amaçladık. Örnek bir anket sorusu Figür 16'da belirtilmiştir.

Yöntem-1 Baseline model ile Sonuç Kıyaslaması İki model arasındaki sonuçlar kıyaslanırken kullanılan ölçütler ve tanımları aşağıda verilmiştir.

Sözcük Katkı Puanı (SKP): Sistemimizin/LRP modelinin orjinal metne verdiği tahmin sonucu (İlk Cümlenin tahmin sonucu) ve sözcüğün olumsuzunun kullanılmasıyla oluşturulan yeni metin için Sistemimizin/LRP modelinin verdiği tahmin sonucu(İkinci cümlenin tahmin sonucu) kullanılarak Figür 6'da verilen Kelime Katkı Denklemi gerçekleştirlir. Elde edilen sonuç Sözcük Katkı Puanı'dır.

Sözcüklerin Toplam Sonuca Etki Oranı (SSEO): Sözcük Katkı Puanı'nın Sistemimizin/LRP modelinin orjinal metne verdiği tahmin sonucuna oranı.

Sıralama Sapma Derecesi (SSP): Her iki modelde eylemlerin/eylemsilerin, sıfatların ve zarfları **SSEO** değerlerine göre büyükten küçüğe sıralanarak indexlenir.Her bir sözcük için $\mid indeksLRP - indeksMETOD \mid$ işlemi yapılarak elde edilen sonuç.

Bu yöntemde, yukarıda verilen ölçütler kullanılarak sistemimize ve LRP modeline girdi olarak verilen yorumun seçili sözcüklerinin **SSP** değerlerinin Ortalaması ve Standart Sapmasını hesaplayark iki yaklaşımın sonuçlarını gözlemledik. Bunun yanında her iki modelin

sözcüklerinin SSEO değerlerine göre yapılan sıralamalarında Minimum Edit Distance (MED) hesabı yaparak, Figür 10'daki gibi durumlarda daha başarılı kıyaslama yapabileceğimizi düşündük.

Figure 10: Örnek Sıralama:

LRP Modeli Sıralama	Sistem Sıralama	Sözcüklerin SSP değeri
Çok	Gitmeseydim	1
Güzel	Çok	1
Başarısız	Güzel	1
Komik	Başarısız	1
İzledim	Komik	1
Gitmeseydim	İzledim	5

Standart Sapma: 1.63

Ortalama: 1.6

MED: 1 (replace çok-gitmeseydim)

Ölçütleri Appendix A Tablo 11 üzerinden incelersek; Tablodaki puan sütununun sözcüklerin SKP değerlerini, oran sütununun ise SSEO değerlerini gösterdiği gözlemlenir. Buna göre, bu örnek yorum için sözcüklerin SSEO değerine göre LRP modelinde sıralanışı:(0)Başarılı, (1)Açan, (2)İnanılmaz, (3)Eğelenceli, (4)Zeki iken sistemimize göre sıralanışları:(0)Başarılı, (1)Açan, (2)Eğelenceli, (3)İnanılmaz, (4)Zeki'dir. Bu örnekte sadece "Eğelenceli" ve "İnanılmaz" sözcüklerinin sıralamasında bir farklılık olduğu gözlemlenir. Buna bağlı olarak. "Eğelenceli" ve "İnanılmaz" sözcüklerinin SSP değeri 1, diğer sözcüklerin SSP değeri 0'dır. Bunun yanında Ortalama SSP değeri 0.4 ve Standart Sapma 0.547'dir SSP değerlerine bakılarak sistem sonucu ile LRP modelinin neredeyse paralel çalıştığı gözlemlenir.

Yöntem-2 Anket Üstünden Sistem Güvenilirliğini Ölçmek Bölüm 3.3.2'de detaylı olarak anlatılan insan temelli değerlendirme yapılmıştır. Değerlendirme ölçütümüz insanların metodlara olan güvenleridir. Ankette yer alan cümlelerin açıklamalarına kullanıcıların güvenleri sorulur. Her bir yorum için verilen güven skorları toplanır ve ankete katılan kişi sayısına bölünür. Böylece metodumuz ve referans model için her bir yoruma ortalama güven skoru atanmış olur. Daha sonrasında metodumuz ve referans modelin skorları kendi arasında toplanır ve anketteki cümle sayısına bölünür. Böylece metodumuz ve referans modelin ayrı ayrı ortalama güvenini hesaplamış oluyoruz.

6.5 Deney Sonuçları

Bu bölümde deney sonuçlarımızı belirlediğimiz araştırma sonuçlarının cevapları üzerinden inceledik. Bölüm 6.4 Yöntem-1'de belirtilen değerlendirme ölçütleri doğrultusunda örnek verilen test yorumlarının detaylı sonucu Tablo 2, 3, 4, 5, 6, 7, 8, 9, 10, 11'de gösterilmiştir. Her bir tabloda sözcüklerin katkı puanı (SKP), ve Sözcüklerin katkı puanının, modelin karar puanına oranı (SSEO) verilmiştir. Bunun yanında Yöntem-2' de belirtilen değerlendirme ölçütleri doğrultusunda oluşturulan 23 kişinin katıldığı anket sonuçları da Tablo 1'de gösterilmiştir.

Table 1: Anket Sonuçları (Kullanıcıların 1-5 arasında puan verebildiği ankette ortalama güvenler gösterilmektedir.)

	metodumuz ortalama güven	LRP ortalama güven
1.Cümle(Tablo 6)	3.60	3.13
2.Cümle(Tablo 2)	3.21	3.47
3.Cümle(Tablo 7)	2.52	2.95
4.Cümle(Tablo 4)	3.26	3.13
5.Cümle(Tablo 8)	3.17	3.69
Ortalama	3.15	3.27

Açıklamalar modeli doğru temsil ediyor mu? Sistemimizin oluşturduğu açıklamaların modeli doğru temsil edip edmediğini 2 başlık altında değerlendirdik.

1) İlk olarak, sistemde savaş yıldırım (2021)'ın BERT tabanlı duygu analizi modelinin kullanıldığı durumda elde edilen sonuçların Bölüm 6.4 Yöntem-1 ölçütleri üzerinden değerlendirdik. Figür 11' de yer alan tabloda, sistemimizde savaş yıldırım (2021)'ın BERT tabanlı duygu analizi modeli kullanılırken, test veri setimizde yer alan 4 negatif, 4 pozitif olmak üzere 8 örnek yorumdan elde edilen sonuçların Yöntem-1 Ölçütleri üzerinden değerlendirilmesi yer almaktadır. Tabloya bakılarak sistemimizin pozitif cümlelerdeki sonuçlarının negatif cümlelere kıyasla baseline modelimize daha yakın olduğu gözlemlenmiştir. Bunun yanında, tüm veri setimizin sisteme girdi olarak verilmesiyle elde edilen sonuçların Yöntem-1 Ölçütleri üzerinden aldıkları değerler Figür 12 ve 13 üzerinden bakılabilir. Grafiklerde test veri setimizdeki pozitif/negatif her bir yorum için hesaplanan SSP Standart Sapması, SSP Ortalaması ve MED ölçütlerinin değeri verilmiştir. Grafiklerden de görüldüğü üzere pozitif yorumlarda metriklerin değerleri negatif yorumlara kıyasla daha küçüktür. Bundan yola çıkarak sistemimizin pozitif yorumlarda negatif yorumlara kıyasla baseline modelimizin

sonuçlarından daha az saptığı gözlemlenmiştir.

Figure 11: Örnek Sonuç Değerlendirme Tablosu:

	Pozitif 1 (Tablo1)	Pozitif2 (Tablo4)	Pozitif3 (Tablo 5)	Pozitif4 (Tablo8)
Standart sapma	0	1.75	1.16	0.5
Ortalama	0	2.25	0.8	0.4
med	0	7	2	1
İşlenen kelime sayısı	2	8	9	5
Eylem		5	2	1
Sıfat	2	3	6	4
Bağlaç	0	0	1	0
	Negatif 1 (Tablo 3)	**Negatif2 (Tablo7)**	**Negatif3 (Tablo6)**	**Negatif4 (Tablo2)**
Standart sapma	1.6	1.9	2.23	1.5
Ortalama	1.33	2.33	2.4	1.3
med	4	7	12	2
İşlenen kelime sayısı	9	10	12	6
Eylem	2	3	6	3
Sıfat	2	7	5	1
Bağlaç	0	0	1	1

Figure 12: Pozitif Yorumlar Sonuç Değerlendirme Grafiği:

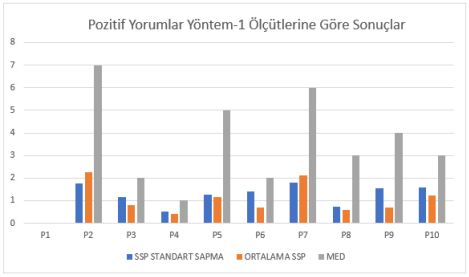
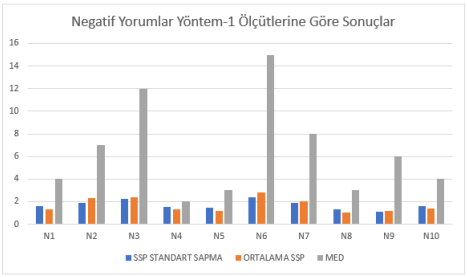


Figure 13: Negatif Yorumlar Sonuç Değerlendirme Grafiği:



2) İkinci olarak, Test veri setimizde bulunan 5 yorum için kendi sistemimizin sonuçlarını ve baseline modelimiz olan LRP'nin sonuçlarını içeren anketten elde edilen cevapları Bölüm 6.4 Yöntem-2 ölçütleri üzerinden değerlendirdik. Tablo 1'de katılımcıların, seçilen 5 yorum cümlesine sistemimiz ve LRP tarafından yapılan açıklamalara duyduğu ortalama güven yer almaktadır. Bu tabloya bakılaraksistemimizin örnek pozitif yorumlarda (1, 4 ve 5.Cümleler) elde ettiği sonuçların LRP'ye göre daha güvenilir olduğu, negatif yorumlarda (2 ve 3. Cümleler) ise LRP modelinin sistemimize göre daha güvenilir sonuçlar verdiği gözlemlenebilir.

Bu bilgiler doğrultusunda sistemimizin savaş yıldırım (2021)'ın BERT tabanlı duygu analizi modeli kullanıldığı durumlarda, pozitif yorumlarda negatif yorumlara kıyasla daha doğru çalıştığını söyleyebiliriz.

Metod Ölçeklenebilir mi? Bu soru ile metodumuzun artan yükle başa çıkma yeteneğini değerlendirmek istedik. Bu amaçla birden fazla cümleden oluşan uzun test metinlerini her iki metod ile denedik. Tablo 4, 5 görüldüğü üzere uzun cümlelerde sıfat varsa tasarladığımız metod her iki modelde de "güzel" kelimesine 0.87 ve 0.98 skorları vererek etkili kelimeleri bulabilmiştir. Bu kelimelere yüksek puan verdiğinden LRP'den daha fazla başarım sağlamıştır. Uzun cümlelerde eğer cümlede sıfat yoksa veya az varsa tablo 7, 9, 10'da olduğu gibi bizim metodumuz tahminde etkili olan kelimeleri bulamamaktadır. Bu durumda LRP metodu daha fazla başarım göstermiştir. Bunun yanında Figür 15'e bakılarak tüm test yorumları için, Sözcüklerin SSP değerine bağlı metriklerin, metindeki toplam eylem/eylemsi, sıfat ve zarf sözcük türlerinin sayısıyla ilişkili olmadığı gözlemlenebilir. Figür 14'e bakılarak da Sözcüklerin SSP değerine bağlı metriklerin metindeki sıfat oranına bağlı olduğu gözlemlenir. Bu bilgiler ışığında sistemimizin artan metin uzunluğuyla, metindeki sıfat oranının yüksek olması durumunda başa çıkabildiği söylenebilir.

Biz bu çalışmada sistemimizi ve baseline modelimizi değerlendirirken oluşturduğumuz, test yorumlarındaki sıfatların, zarfların ve yüklemlerin olumsuzlarını Adj/Adv-AntonymDict ve Verb-AntonymDict Sözlüklerine eklemiş bulunduk. Siztemimizde, Adj/Adv-AntonymDict sözlüğünde zıtı bulunmayan sıfat ve zarfların zıt anlamlıları Bakay et al. (2021) tarafından oluşturulan **Türkçe WordNet** üzerinden (Software, 2022) aranır. Bu WordNet üzerinden bir sonuç alınamadığında da **Lexicala Web API** (K-DICTIONARIES, 2020) üzerinden aranır. Bu yöntemle sistemimizi sıfat ve zarfların sözlükte bulnumaması durumuna karşı olabildiğince ölçeklemeye çalıştık. Ancak sistemimiz yüklemrin olumlu/olumsuz anlamlılarının çekilmesinde tamamen oluşturduğumuz Verb-AntonymDict Sözlüğüne bağlıdır. Bu konu için yapılabilecek öneriler **8 Gelecekteki İşler** başlığı

altında detaylıca incelenmiştir.

Figure 14: Metindeki Sıfat Oranına Bağlı Yöntem-1'deki Ölçütlerin Değişimi

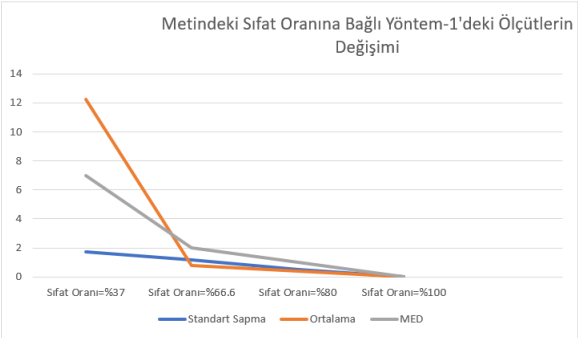
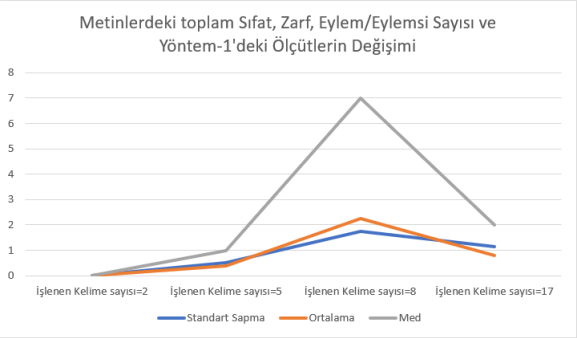


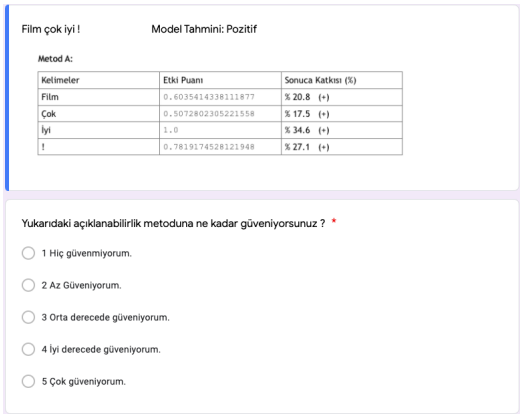
Figure 15: Metinlerdeki toplam Sıfat, Zarf, Eylem/Eylemsi Sayısı ve Yöntem-1'deki Ölçütlerin Değişimi



Açıklamalar kullanıcıların model tahminine olan güvensizliği ortadan kaldırıyor mu? Açıklamalar insanlara sunulmak içindir. Bu nedenle, insanları kullanarak açıklamaları değerlendirmek istiyoruz. Bölüm 3.3.2'de ayrıntılı olarak açıkladığımız insan temelli değerlendirmenin bir yöntemi olan ileri tahmini kullanıyoruz. Bulduğumuz sayısal kelime katkılarına insanların ne kadar güvendiğini analiz ediyoruz. Bu değerlendirmeyi google form üzerinde anket yaparak gerçekleştiriyoruz (sur, 2022). Bu ankette açıklama metodu tarafından belirlenen kelimelerin öneminin belirtildiği sayısal değerler insanlara gösterildi. Ayrıca modelin duygu tahmini paylaşıldı. İnsanlara güvenlerini beşli Likert ölçeğiyle belirtmeleri istendi. Örnek bir anket sorusu Figür 16'da belirtilmiştir.

Ayrıca modellerin yanlış karar verebileceği insanlarla paylaşılmıştır. Anketi gerçekleştiren insanların kaliteli olmasından emin olmak amacıyla

Figure 16: Örnek Anket Sorusu:



Bilgisayar Mühendisliği yüksek lisans, doktora öğrencileri veya mezunlarından seçilmiştir. Ankette farklı uzunlukta cümleler, farklı sayıda dilbilimsel türlerin bulunduğu veya bulunmadığı cümleler kullanılmıştır. Tablo 1'de anket sonuçları ortalama güven olarak verilmiştir. 5 cümlenin 2 tanesinde metodumuza LRP'ye göre insanlar tarafından daha fazla güvenilmektedir. Metodumuzun 5 cümle için ortalama güveni 3.15 iken, LRP'nin ortalama güveni 3.27 çıkmıştır. Metodumuz ortalamada LRP'ye yakın bir başarımı vardır. Bazı cümlelerde ise başarımı LRP'ye göre daha iyidir.

Sistemin BERT Tabanlı Farklı Bir duygu Analizi Modeli İçin Oluşturduğu Açıklamalar: Sistemimiz BERT tabanlı duygu analizi modelini açıklamak için model-agnostik bir yöntem kullanır. Modelin algoritması üzerinden herhangi bir değişiklik yapmadan, modele verilen girdileri değiştirerek, modele verilen metin girdisindeki sıfatların, zarfların ve eylemlerin modelin kararına etkisini hesaplar. Sistemin, bu yaklaşımı benzer modeller üzerinde gerçekleştirdiğinde, benzer açıklamalar oluşturmasını bekleriz.

Bu durumu değerlendirmek için ilk önce belirlediğimiz 1 pozitif ve 1 negatif yorumu, sistemimiz savaş yıldırım (2021)'ın duygu analizi modelini kullanırken ve sistemimiz gürkan şahin (2022)'in Türkçe sınıflandırması için eğittiği modeli kullanırken girdi olarak verdik. Appendix A'da yer alan Tablo 2, 3, 4 ve 5'da her iki modelin, bu iki yorumun tahmin puanını hesaplarken hangi sıfatlardan,zarflarden ve eylemlerden ne derece etkilendiği gözlemlenebilir.

Sistemimizin her iki durumda negatif yorum için oluşturduğu çıktıyı gözlemlemek için Tablo

4((savaş yıldırım, 2021) modeli kullanıldığındaki sonuçları içeririr) ve Tablo 3'e ((gürkan şahin, 2022) modeli kullanıldığındaki sonuçları içerir) bakılabilir. Tablo 4'de sıfatların,zarfların ve eylemlerin etki oranına göre sıralanışı:(0)Sıkıcı, (1)Değmez, (2)Çok, (3)Olur, (4)İzlenmeye. Tablo 3'de sıfatların,zarfların ve eylemlerinetki oranına göre sıralanışı:(0)Sıkıcı, (1)Olur, (2)İzlenmeye, (3)Çok, (4)Değmez. Sonuçları Bölüm 6.4 Yöntem-1 Ölçütleri üzerinden incelersek. Sözcüklerdeki toplam SSP 7, Ortalama SSP 1.75, Standart Sapma 0.89 ve MED 4'dür.

Sistemimizin her iki durumda pozitif yorum için oluşturduğu çıktıyı gözlemlemek için Tablo 2((savaş yıldırım, 2021) modeli kullanıldığındaki sonuçları içeririr) ve Tablo 5'e ((gürkan şahin, 2022) modeli kullanıldığındaki sonuçları içerir) bakılabilir. Tablo 2'de sıfatların,zarfların ve eylemlerin etki oranına göre sıralanışı:(0)Güzel, (1)Ediyorum, (2)Okuması, (3)Okuyun, (4)Çok, (5)Anlatan, (6)Gereken, (7)Varın. Tablo 5'de sıfatların,zarfların ve eylemlerinetki oranına göre sıralanışı:(0)Ediyorum, (1)Güzel, (2)Okuyun, (3)Okuması, (4)Anlatan, (5)Çok, (6)Varın, (7)Gereken. Sonuçları Bölüm 6.4 Yöntem-1 Ölçütleri üzerinden incelersek. Sözcüklerdeki toplam SSP 8, Ortalama SSP 1, Standart Sapma 0 ve MED 5'dir.

Yukarıdaki sonuçlar incelendiğinde, her iki girdi için, sistemde (savaş yıldırım, 2021) modeli kullanıldığındaki açıklamalar ile sistemde (gürkan şahin, 2022) modeli kullanıldığındaki açıklamalar arasındaki sapmanın küçük olduğu gözlemlenir.

7 SONUÇ

Bu çalışmada Türkçe veriseti üzerinde çalışan BERT tabanlı duygu analizi modellerinin açıklanabilirliği için oluşturulmuş yerel yaklaşımlı model-agnostik bir açıklanabilirlik metodu geliştirilmiştir. Hila Chefer (2021)'ın LRP modeli de çalışmanın baseline modeli kabul edilmiştir. Oluşturulan modelin başarımı insan temelli değerlendirme ile ölçülmüştür. Bunun yanında metodun hangi durumlarda baseline modele paralel çalıştığı, hangi durumlarda baseline modele kıyasla farklı sonuçlar elde ettiği değerlendirilmiştir. Yapılan değerlendirmeler sonucunda metodun BERT tabanlı farklı duygu analizi modellerinde benzer sonuçlar sergilediği ve sıfat/zarf oranının yüksek olduğu metinlerin açıklamalarını oluştururken daha iyi performans gösterdiği tespit edilmiştir.

8 GELECEK İŞLER

Önerdiğimiz metodun bazı eksiklikleri bulunmaktadır. Wordnet sıfatların zıt anlamlısını en sık geçen manaya(most frequent sense) bakarak veriyor. Cümlemizdeki sıfatların birden fazla manası olabilir. Bu mana Wordnet'teki ilk manaya denk gelmeyebilir. O kelime eş sesli kelime olabilir, deyim içinde kullanılabilir veya mecaz anlamda kullanılabilir. Bu durumda açıklanabilirlik metodumuz doğru çalışmaz. Gelecek çalışma olarak denetimli öğrenme ile kelime mana ayrımı yapılabilir. Böylece Wordnet'ten zıt anlam çekilirken doğru mana elde edilir. Böyle bir çalışma metodumuzun başarımını artıracaktır.

İkinci gördüğümüz eksiklik metodumuzun fill'lerin olumsuzunu bulurken sözlük tabanlı bir yaklaşım gerçekleştirmesidir. Sözlüğümüzde bütün kelimeler olmadığı için uygulanabilirlik açısından bu durum sorun teşkil edebilir. Bu nedenle gelecek çalışma olarak cümledeki fiil'in olumsuzunu bulan bir algoritma geliştirilebilir. Böyle bir yapı metodumuzun daha geniş alanda kullanılabilmesini sağlayacaktır.

Metodumuzun son eksik yanı kullanıcı dostu özelliğinin azlığıdır. Açıklamalar kullanıcılara sunulacağı için bu durum büyük sorun teşkil edebilir. Literatürdeki diğer çalışmalarda sıcaklık haritasıyla(heatmap) kelimelerin katkıları duygu polaritesine göre yeşil ve kırmızı renklerle gösterilmektedir. Bizim metodumuza bu özelliği eklememiz kullanıcının anlamasını ve memnuniyetini artıracaktır.

Kaynaklar

2022. Değerlendirme anketi.

Anna Shcherbina Avanti Shrikumar, Peyton Greenside and Anshul Kundaje. 2016. " not just a black box: Learning important features through propagating activation differences".

Özge Bakay, Özlem Ergelen, Elif Sarmış, Selin Yıldırım, Bilge Nas Arıcan, Atilla Kocabalcıoğlu, Merve Özçelik, Ezgi Sanıyar, Oğuzhan Kuyrukçu, Begüm Avar, and Olcay Taner Yıldız. 2021. Turkish WordNet KeNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 166–174, University of South Africa (UNISA). Global Wordnet Association.

Jerome Bellegarda. 2010. Emotion analysis using latent affective folding and embedding. In *Proceedings of*

the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, pages 1–9.

Roger Bougie, Rik Pieters, and Marcel Zeelenberg. 2003. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. *Journal of the academy of marketing science*, 31(4):377–393.

Hila chefer. 2021. Transformer-explainability.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.

Izzet Emre Demir. 2021. turkish-sentiment-analysis-with-bert.

Eduardo M Pereira Diogo V Carvalho and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8):832.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Gültekin Cakir Antonio Jacob Fabio Lobato Marija Bezbradica Markus Helfert Douglas Cirqueira, Fernando Almeida. 2020. Explainable sentiment analysis application for social media crisis management in retail. In *Conference: 4th International Conference on Computer-Human Interaction Research and Applications - Volume 1: WUDESHI-DR*.

Virginia Francisco and Pablo Gervás. 2006. Automated mark up of affective information in english texts. In *International Conference on Text, Speech and Dialogue*, pages 375–382. Springer.

Gizem Gezici and Berrin Yanıkoğlu. 2018. Sentiment analysis in turkish. In *Turkish natural language processing*, pages 255–271. Springer.

Alexander Binder Wojciech Samek Klaus-Robert Müller Grégoire Montavon, Sebastian Lapuschkin. 2016. " explaining nonlinear classification decisions with deep taylor decomposition".

Yangfeng Ji Hanjie Chen. 2019. Improving the explainability of neural sentiment classifiers via data augmentation. NeurIPS 2019 Workshop on Robust AI in Financial Services.

Lior Wolf Hila Chefer, Shir Gur. 2021. " transformer interpretability beyond attention visualization".

Milo Honegger. 2018. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. arXiv:1808.05054.

Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. 2018. "bert: Pre-training of deep bidirectional transformers for language understanding". arXiv:1810.04805.

Yeon-Jee Jung, Seung-Ho Han, and Ho-Jin Choi. 2021. Explaining cnn and rnn using selective layer-wise relevance propagation. *IEEE Access*, 9:18670–18681.

K-DICTIONARIES. 2020. Lexicala api.

Caner Kerimogğlu. 2016. "türkçe dil bilgisi öğretimindeki cümle öğeleriyle ilgili değerlendirmeler".

Lee Y. Khan, J. 2019. Lessa: A unified framework based on lexicons and semisupervised learning approaches for textual sentiment classification. *Applied Sciences*, 9(24), 5562.

Kuan-Chuan Peng Jan Ernst-Yun Fu Kunpeng Li, Ziyan Wu. 2018. "tell me where to look: Guided attention inference network".

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684.

Klaus-Robert Müller Wojciech Samek Leila Arras, Grégoire Montavon. 2016. "what is relevant in a text document?". arXiv 1612.07.

Klaus-Robert Müller Wojciech Samek Leila Arras, Grégoire Montavon. 2017. "explaining recurrent neural network predictions in sentiment analysis".

Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. *arXiv preprint arXiv:1908.11355*.

Zachary C. Lipton. 2017. The mythos of model interpretability. arXiv:1606.03490v3.

Diane Litman and Kate Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 351–358.

harun Loodos. 2021. Zemberek-py.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. arXiv:1705.07874v2.

Carlos Guestrin Marco Tulio Ribeiro, Sameer Singh. 2016. "why should i trust you?": Explaining the predictions of any classifier. arXiv:1602.04938v3.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Saif M Mohammad et al. 2013. Tracking sentiment in mail: How genders differ on emotional axes. *arXiv preprint arXiv:1309.6347*.

Christoph Molnar. 2019. *Interpretable Machine Learning*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

B Pang and L Lee. 2008. Opinion mining and sentiment analysis. foundations and trend. *Information Retrieval*, 2(1-2).

Abhishek Das Ramakrishna Vedantam-Devi Parikh Dhruv Batra Ramprasaath R. Selvaraju, Michael Cogswell. 2017. "grad-cam: Visual explanations from deep networks via gradient-based localization".

Niklas Ravaja, Timo Saari, Marko Turpeinen, Jari Laarni, Mikko Salminen, and Matias Kivikangas. 2006. Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and virtual environments*, 15(4):381–392.

Marco Tulio Correia Ribeiro. 2020. Lime.

G. Montavon F. Klauschen K.-R. Müller W. Samek S. Bach, A. Binder. 2015. "on pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". PLOS ONE 10.

Daniel Morales Brotons Sandor Berglund, Francisco Ferrari. 2021. " lrp-based method for transformer interpretability".

Husna Sayedi. 2021. Explainable ai (xai): Nlp edition.

F. Sağlam. 2019. *Otomatik Duygu Sözlüğü Geliştirilmesi ve Haberlerin Duygu Analizi.* Doktora Tezi, Hacettepe, Ankara, Türkiye.

Alex Sikand. 2020. Explainable ai with lime and shap.

Starlang Software. 2022. Turkishwordnet-py.

K. Sokol and P. Flach. 2020. "explainability fact sheets: a framework for systematic assessment of explainable approaches".

Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25.

Kevan Buckley Thelwall, Mike and Georgios Paltoglou. 2011. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2): 406-418.

S. Tokcaer. 2021. Türkçe metinlerde duygu analizi. *Journal of Yasar University*, 16/63, 1514-1534.

Benan Bardak ve Mücahid Kutlu Utku Umur Açıkalın. 2020. "bert modeli ile türkçe duygu analizi".

Kemal Toprak Uçar. 2020. Bert modeli ile türkçe metinlerde sınıflandırma yapmak.

Abir Das Vitali Petsiuk and Kate Saenko. 2018. "rise: Randomized input sampling for explanation of black-box models".

SM Vohra and JB Teraiya. 2013. A comparative study of sentiment analysis techniques. *Journal Jikrce*, 2(2):313–317.

S. Lapuschkin C. J. Anders W. Samek, G. Montavon and K. R. Müller. 2021. "explaining deep neural networks and beyond: A review of methods and applications". 109.

Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. Case-based reasoning for assisting domain experts in processing fraud alerts of black-box machine learning models.

savaş yıldırım. 2021. bert-base-turkish-sentiment-cased.

Onur Yılmaz. 2016. turkish-pos-tagger.

gürkan şahin. 2022. turkish-product-comment-sentiment-classification.

A Appendix

Table 2: Test Cümlesi(model-savaş yıldırım) : Çok sıkıcı ve gerçekten izlenmeye değmez yazık olur zamanınıza.

	metodumuz		LRP	
	puan	oran%	puan	oran%
Çok	-0.00005	0.0056	-0.39	9.5
Sıkıcı	-0.98	98.1	-1.0	24.12
Ve	-0.00003	0.0037	-0.44	10.8
Gerçekten	-0.006	0.316	-0.35	8.5
İzlenmeye	-0.00001	0.001	-0.57	14.5
Değmez	-0.00008	0.0078	-0.48	12.5
Yazık	-0.006	0.316	-0.09	2.4
Olur	-0.00002	0.002	-0.01	0.27
Zamanınıza	-0.006	0.316	-0.55	13.7

Table 3: Test Cümlesi(model-gürkan şahin) : Çok sıkıcı ve gerçekten izlenmeye değmez yazık olur zamanınıza.

	metodumuz		LRP	
	puan	oran%	puan	oran%
Çok	-0.00003	0.003	-0.53	9.94
Sıkıcı	-0.96	96.46	-1.0	18.76
Ve	-0.0002	0.02	-0.57	10.69
Gerçekten	-0.0001	0.017	-0.36	6.75
İzlenmeye	-0.0001	0.01	-0.97	18.19
Değmez	-0.000005	0.0005	-0.77	14.44
Yazık	-0.0001	0.017	-0.38	7.12
Olur	-0.00001	0.014	-0.16	3
Zamanınıza	-0.0001	0.017	-0.59	11.06

Table 4: Test Cümlesi(model-savaş yıldırım) : Hayvan çiftliği bence insanlara kendilerini hayvanların temsiliyle anlatan çok güzel bir kitap. Bence herkesin okuması gereken bir kitap. Kesinlikle tavsiye ediyorum. Okuyun ve gerçeklerin farkına varın.

	metodumuz		LRP	
	puan	oran%	puan	oran%
Anlatan	0.001	0.15	0.02	0.79
Çok	0.001	0.16	0.04	1.59
Güzel	0.98	99.4	0.1	3.37
Okuması	0.003	0.39	0.04	1.42
Gereken	0.001	0.13	0.047	1.56
Ediyorum	0.96	98.2	0.07	2.33
Okuyun	0.02	2.18	1.0	32.9
Varın	0.0006	0.063	0.124	4.08
Diğerleri:				52

Table 5: Test Cümlesi(model-gürkan şahin) : Hayvan çiftliği bence insanlara kendilerini hayvanların temsiliyle anlatan çok güzel bir kitap. Bence herkesin okuması gereken bir kitap. Kesinlikle tavsiye ediyorum. Okuyun ve gerçeklerin farkına varın.

	metodumuz		LRP	
	puan	oran%	puan	oran%
Anlatan	0.001	0.11	0.07	0.62
Çok	0.0007	0.07	0.31	2.77
Güzel	0.87	88.20	0.13	1.16
Okuması	0.002	0.29	0.36	3.22
Gereken	0.000002	0.0002	0.41	3.66
Ediyorum	0.99	99.96	0.48	4.29
Okuyun	0.008	0.87	1.12	10.01
Varın	0.0003	0.03	1.28	11.44
Diğerleri:				62.83

Table 6: Test Cümlesi(model-savaş yıldırım): Film çok iyi !

	metodumuz		LRP	
	puan	oran	puan	oran
film	-0.17	% 10.75	0.60	% 20.8
çok	0.09	% 9,75	0.50	% 17.5
iyi	0.98	% 100.4	1.0	% 34.6
!	0	% 0	0.78	% 27.1

Table 7: Test Cümlesi(model-savaş yıldırım) : Filmi oyuncuları bile kurtaramamış. İzlemek için boşa vakit harcamayın. Raflarınızda çok boş yer varsa alın. Fiyatından başka ilgi çeken yön bulamadım.

	metodumuz		LRP	
	puan	oran%	puan	oran%
Kurtaramamış	-0.00003	0.003	- 0.99	14.6
İzlemek	-0.0000009	0.00089	-0.18	2.56
Harcamayın	-0.00002	0.0019	- 0.65	9.65
Çok	-0.00000155	0.000155	0.018	0.27
Boş	-0.0000086	0.00086	-0.116	1.70
Varsa	-0.0000117	0.00117	-0.04	0.60
Alın	-0.000007	0.00007	-0.06	0.93
Çeken	-0.000006	0.0006	-0.21	3.20
Bulamadım	-0.000013	0.0013	-0.17	2.6
Diğerleri:				65

Table 8: Test Cümlesi(model-savaş yıldırım) : Kitabı okudum kitap çok etkileyici. Bu fiyata da uygun mutlaka okunması gereken bir kitap. Elif Şafağın bütün kitapları ayrı güzel ama buradaki olayları yorumlayışı bence harika.

	metodumuz		LRP	
	puan	oran%	puan	oran%
Okudum	0.002	0.26	0.12	1.94
Çok	0.11	11.6	0.77	12.32
Etkileyici	0.98	101.4	1.0	15.92
Uygun	0.0005	0.06	0.05	0.89
Okunması	0.09	9.47	0.56	8.96
Gereken	0.0006	0.07	0.18	2.89
Güzel	0.003	0.39	0.056	0.9
Ama	0.003	0.4	0.06	1.06
Harika	0.81	0.81	0.25	4.08
Diğerleri:				

Table 9: Test Cümlesi(model-savaş yıldırım) : Bu zamana kadar izlediğim en kötü filmdi sonuna kadar bir umudum vardı bir şeyler olacak diye ama tamamen hüsran. Bence boşuna para harcamışlar bu filme. Yazık diyorum bir şey demiyorum.

	metodumuz		LRP	
	puan	oran%	puan	oran%
İzlediğim	-0.00004	0.004	-0.39	4
En	-0.00003	0.003	-0.34	3.8
Kötü	-0.001	0.159	-1.0	11.45
Sonuna	-0.000001	0.0001	-0.044	0.47
Vardı	-0.00002	0.002	-0.032	0.35
Olacak	-0.00001	0.001	-0.016	0.178
Ama	-0.000009	0.0009	-0.06	0.717
Hüsran	-0.00001	0.001	-0.84	9.1
Boşuna	-0.00004	0.004	-0.61	6.6
Harcamışlar	0.00005	0.005	-0.76	8.35
Diyorum	-0.000007	0.0007	-0.06	0.72
Demiyorum	-0.000006	0.0006	-0.07	0.78

Table 10: Test Cümlesi(model-savaş yıldırım) : Her şeyiyle eksik bir film olmuş. Senaryo tam oturmamış, temposu dengesiz, mesajlar karışık yani bir vampir filmini bu kadar mıncıklamaya gerek yok. Son derece özensiz ve basit bir film.

	metodumuz		LRP	
	puan	oran%	puan	oran%
Eksik	-0.00001	0.001	-0.36	3.94
Olmuş	-0.00001	0.001	-0.36	3.95
Oturmamış	-0.000004	0.0004	-0.26	2.8
Dengesiz	-0.00002	0.002	-0.25	2.81
Karışık	-0.00003	0.003	-0.69	9.2
Mıncıklamaya	-0.001	0.001	-0.47	5.16
Son	-0.000009	0.0009	-0.1	1.0
Yok	-0.00001	0.001	-0.07	0.84
Özensiz	-0.00001	0.001	-0.58	6.2
Basit	-0.00004	0.004	-1.0	10.8

Table 11: Test Cümlesi(model-savaş yıldırım) : Kurgusu, karakterleri, oyunculukları, kamera kullanımı, tam anlamıyla eğlenceli geyik diyalogları ve inanılmaz derecede başarılı zeki kurgusuyla sinemada çığır açan bir film.

	metodumuz		LRP	
	puan	oran%	puan	oran%
Eğlenceli	0.0005	0.05	0.289	2.6
İnanılmaz	0.0002	0.023	0.38	3.57
Başarılı	0.99	100.062	1.8	17.05
Zeki	2.12	0.002	0.27	2.54
Açan	0.79	79.5	0.43	3.9