TÜRKÇE İÇİN MAKİNE ÖĞRENME TABANLI DOĞAL DİL İŞLEME MODELİ GELİŞTİRİLMESİ

Melek Gezer

İnönü Üniversitesi,Yazılım Mühendisliği, Malatya, Türkiye, melekgezer99@gmail.com

Özet— Doğal dil, insanları diğer canlılardan ayıran ve iletişim kurmalarını sağlayan en temel özelliklerden biridir. Dil, duygu ve düşüncelerin ifade edilmesinde kullanılan, aynı zamanda kültürlerin nesiller boyunca aktarılmasını sağlayan bir araçtır. Günlük hayatta karşılaşılan yazılar ve sesler birer doğal dil örneğidir. Doğal dil dinamik bir yapıya sahiptir; kelimeler zamanla değişir, bazıları kaybolurken yenileri türetilir. Bu nedenle, doğal dil işleme (DDİ) süreçleri insanlar için bile karmaşıkken, bilgisayar ortamında modellenmesi önemli bir zorluk teşkil etmektedir.

Dil bilim, insanların dili nasıl kullandığını incelerken, doğal dil işleme çalışmaları dil bilimciler ile bilgisayar bilimcilerin ortak çabasını gerektirir. Bu çalışmalar, insan-bilgisayar etkileşiminde kritik bir rol oynar. Geleneksel makine öğrenmesi yöntemlerinden biri olan **lojistik regresyon**, doğal dil işlemede metin sınıflandırma, duygu analizi ve sözcük etiketleme gibi görevlerde yaygın olarak kullanılmaktadır. Lojistik regresyon, özellikle daha küçük veri kümelerinde veya daha az karmaşık dil modellerinde etkili sonuçlar verebilen, yorumlanması kolay bir istatistiksel yöntemdir.

Bu çalışmada, Türkçe sözcük etiketleme için **lojistik regresyon tabanlı bir model** önerilmiştir. Model, doğal dil araştırmacılarına kendi analizlerini gerçekleştirebilecekleri ve uygulayabilecekleri bir platform sunmayı hedeflemektedir. Geliştirme aşamasında uzman geri bildirimleri alınarak modelin hata oranı optimize edilmiş ve dilin yapısal özelliklerine uygun bir etiketleme performansı sağlanmıştır. Lojistik regresyonun sadeliği ve performansı, bu çalışmada doğal dil işleme görevleri için alternatif bir yaklaşım olarak değerlendirilmiştir.

Anahtar Kelimeler— Türkçe metin sınıflandırma, tema analizi, lojistik regresyon, TF-IDF, chunklama, doğal dil işleme.

DEVELOPMENT OF MACHİNE LEARNING BASED NATURAL LANGUAGE PROCESSING MODEL FOR TURKISH

Abstract— Natural language is one of the most fundamental features that distinguishes humans from other living beings and enables communication. Language serves as a tool for expressing emotions and thoughts while facilitating the transmission of cultures across generations. Written texts and spoken words encountered in daily life are examples of natural language. Natural language has a dynamic structure—words evolve over time, some disappear, and new ones emerge. As a result, natural language processing (NLP) is inherently complex, even for humans, and poses significant challenges when modeled computationally.

Linguistics studies how humans use language, while NLP research requires collaboration between linguists and computer scientists. These studies play a crucial role in human-computer interaction. **Logistic regression**, a traditional machine learning method, is widely used in NLP tasks such as text classification, sentiment analysis, and word tagging. It is particularly effective for smaller datasets or less complex language models, offering interpretability and computational efficiency.

In this study, a **logistic regression-based model** is proposed for Turkish word tagging. The model aims to provide researchers with a platform to perform and apply their own linguistic analyses. During development, expert feedback was incorporated to optimize error rates and ensure accurate

tagging performance aligned with the structural features of the language. The simplicity and performance of logistic regression present it as a viable alternative approach for NLP tasks.

Keywords— Turkish text classification, theme analysis, logistic regression, TF-IDF, chunking, natural language processing.

1. GİRİŞ

Doğal dil, insanların kendilerini ifade etmeleri ve iletişim kurabilmeleri için kullanılan bir araçtır. Chomsky, dilin çocukluk yıllarında duyulandan, doğal bir dile dönüşümünün, insanın genetik yapısıyla ilişkili olduğunu ifade etmektedir (Chomsky, 1986). İnsanların dili nasıl edindiği, ürettiği ve anladığı dil biliminin araştırma alanıdır. Dil bilimciler dilsel ifadeleri işlemek için kural tabanlı yaklaşımlar öne sürmüşlerdir. Ancak dil kullanımının üreticiliği kurallara her zaman uymamaktadır. Bu doğrultuda dil ifadelerine dilbilgisi kuralları uygulamak yerine istatistiksel yaklaşımlar uygulanarak dil kullanımında ortak kalıplar elde edilmeye çalışılmıştır. Dilin istatistiksel modelleri, dil bilimi ve bilgisayar biliminin alt bilim dalı olan doğal dil işleme (DDİ) çalışmalarında başarı ile uygulanmıştır (Schütze & Manning, 1999). DDİ'nin amacı, doğal dilleri otomatik olarak oluşturma ve anlamadaki problemleri incelemektir (Young, Hazarika, Poria, & Cambria, 2018). DDİ insanlar tarafından üretilen sesleri ve metinleri işleyerek insan bilgisayar etkileşiminin sağlanmasına yardımcı olmaktadır.

DDİ insan dilinin otomatik analizi ve gösterimi için teorik olarak motive edilmiş hesaplama teknikleridir (Cambria & White, 2014). Her yaşta insanın sosyal medyaya ulaşabildiği bir ortamda, üretilen veri miktarı, her geçen gün artarak devam etmektedir. İnsanlar tarafından doğal olarak oluşturulan veriler, doğrudan işlenecek durumda değildir. Bu yüzden insan makine iletişimini sağlamak için verileri anlamlandırma ve verimli kullanabilme çabası birçok alanın birlikte çalışmasını gerektirmiştir. İlk zamanlarda yapılan çoğu DDİ çalışmaları, tek tek sözcüklere odaklanmışken 19. yüzyılın sonlarına doğru sözcüklerin birbirleriyle olan ilişkisine ve bütün üzerinde anlam bilim çalışmalarına yönelmiştir (Cambria & White, 2014).

DDİ çalışmaları ilk olarak metni anlamak için ses veya metinden özellik çıkarımı yapan bir ön işlemeden geçmektedir. Ardından şekilbilim (morphological), sözdizim (syntactic), anlambilim (semantic) ve söylev (discourse) işleme çalışmaları gerçekleştirilebilmektedir. Bu çalışma alanları sözcük kökleri, sözcük bağlamları ve anlambilim açısından bazı zorluklara sahiptir. Zorlukları aşmak için geliştirilen dilbilgisine dayalı kural tabanlı DDİ çalışmaları (Brill 1992), (J. Gimenez and L. Marquez 2004), el yapımı özelliklere dayanmaktadır. El yapımı özellikler zaman almakta ve yetersiz kalmaktadır (Young, Hazarika, Poria, & Cambria, 2018). Eksikliklerin giderilmesi için geliştirilen yapay zekâ yöntemlerinin önemli bir uygulaması olan derin öğrenme, yapay sinir ağı yapısı, güçlü donanımı ve büyük veri girdisi ile daha iyi sonuçlar elde edilmesini sağlamıştır (Song & Lee, 2013).

1965 yılında derin öğrenmenin temeli kabul edilen çok katmanlı bir perceptron türü algoritma (Ivakhnenko & Lapa, 1965) önerilmiştir. Fakat o yıllardaki basit bir ağın eğitiminin uzun sürmesi ve yüksek hesaplama maliyetlerinden dolayı, destek vektör makinaları gibi el ile hazırlanmış özelliklere sahip modeller (Cortes & Vapnik, 1995)) kabul görmüştür (Şeker, Dirib, & Balık, 2017). Yakın zamanda grafik işleme birimi (GPU) ve diğer donanımsal gelişmeler sayesinde hesaplama maliyetleri düştüğünden, çok sayıda gizli katmandan oluşan yapay sinir ağları tekrar kullanılmaya başlanmıştır (Schmidhuber, 2015). Bu doğrultuda DDİ çalışmalarında, makine

çevirisi, bilgi alma, metin özetleme, soru cevaplama, bilgi çıkarma, konu modelleme ve sözcük etiketleme gibi görevlere, derin öğrenme uygulanmasına odaklanılmaktadır (Young, et al., 2018) (Şeker, Dirib, & Balık, 2017).

Basit bir derin öğrenme çerçevesinin, adlandırılmış varlık tanıma, anlamsal rol etiketleme ve sözcük etiketleme gibi birçok DDİ görevinde, en modern yaklaşımlardan daha iyi performans gösterdiği ortaya konulmuştur (Collobert, ve diğerleri, 2011). DDİ alanında istatistiksel yöntemler, kural tabanlı yöntemlerden daha başarılı olmaktadır. Bu alanda sözcük etiketlemesi, sözcük türlerinin sözcüklere atanması ile gerçekleştirilmektedir (sözcük/isim, sıfat, fiil, vb.). Etiketler bilgisayarların cümlede ifade edileni anlamasında kolaylık sağlamaktadır. Ancak sözcükler farklı bağlamlarda kullanıldığı zaman farklı anlamlar ifade edebilmektedir. Örneğin 'yüz' sözcüğü kullanıldığı bağlama göre isim veya fiil etiketini alabilmektedir. Verilen örneği incelediğimizde, her sözcüğü etiketlemenin söz konusu olmadığı görülmektedir. Verilen örneği incelediğimizde, her sözcüğü etiketlemenin söz konusu olmadığı görülmektedir. Sözcük etiketlemede belirsizliği gidermek için etiketlenecek sözcüğün öncesinde ve sonrasında kullanılan sözcüklere bakılarak doğru etiket sınıfı belirlenebilmektedir. Türkçe doğal dil işleme için geliştirilen Zemberek kütüphanesi, bu tür bağlamsal analizlerde önemli bir rol oynamaktadır. Zemberek, Türkçe'nin morfolojik yapısını işleyebilen, sözcük köklerini çözümleyen ve bağlama duyarlı etiketleme yapabilen açık kaynaklı bir kütüphanedir (Akyürek & Güngör, 2010).
Zemberek, özellikle Türkçe gibi eklemeli ve bağlamsal anlam değişikliklerinin sık görüldüğü diller için optimize edilmiş bir araçtır. Sözcük etiketleme, kök analizi ve cümle parçalama gibi temel DDİ görevlerinde yüksek doğruluk sunar. Dil modelleme (Sundermeyer, Schlüter, & Ney, 2012), makine çevirisi ve metin anlama gibi alanlarda Türkçe için özelleştirilmiş çözümler sunan Zemberek, akademik ve endüstriyel uygulamalarda yaygın olarak kullanılmaktadır.

Bu çalışma 4 bölümden oluşmaktadır. 2. Bölümde literatür taraması anlatılmıştır. 3. bölümde önerilen modelin geliştirilme süreçlerinden bahsedilmiştir. 4. bölümde ise sonuçlara değinilmiştir.

2. LİTERATÜR TARAMASI

Doğal dil işleme (DDİ) alanında, anlamsal benzerlik, metin çıkarımı ve makine çevirisi gibi görevlerde derin öğrenme tabanlı modellerin kullanımı giderek yaygınlaşmaktadır. Bu bağlamda, özellikle dikkat (attention) mekanizmaları ve tekrarlayan sinir ağları (RNN) üzerine yapılan çalışmalar ön plana çıkmaktadır.
Rocktäschel vd. (2016), iki cümle arasında anlamsal ilişkiyi belirlemek amacıyla sinir ağı tabanlı bir dikkat modeli önermiştir. Bu model, her iki cümleyi okuyarak içeriklerini anlamlandırmakta ve aralarındaki çıkarımsal ilişkileri tespit etmektedir. Benzer şekilde, Luong vd. (2015), dikkat mekanizmasının nöral makine çevirisinde nasıl etkili bir şekilde kullanılabileceğini göstermiştir. Özellikle kaynak cümlenin belirli kısımlarına odaklanarak yapılan çevirilerde anlam bütünlüğünün korunduğu ve nadir kelimelerin köklerine inerek doğru biçimde işlendiği vurgulanmıştır.
Mueller ve Thyagarajan (2016), cümle benzerliğini ölçmek amacıyla Siamese mimarili tekrarlayan sinir ağları önermiştir. Bu model, önceden çıkarılan özelliklere dayalı yöntemleri geride bırakarak, cümle çiftleri arasındaki anlamsal yakınlığı daha yüksek doğrulukla tespit etmektedir. Wang ve Jiang (2016) ise doğal dil çıkarımı görevinde LSTM tabanlı modeller kullanarak kelime kelime eşleştirme yöntemini önermiştir. Bu yaklaşım, önemli kelimelere ağırlık vererek ve anlam açısından belirleyici olan uyumsuzlukları hatırlayarak daha isabetli tahminler yapabilmektedir.
Angeli ve Manning (2014), yapay zekâ uygulamalarında özellikle sağduyu (common sense) çıkarımı için doğal mantık tabanlı bir sistem geliştirmiştir. Geniş ölçekli bir bilgi tabanından sağduyu çıkarımı yapan bu sistem, %91 doğruluk oranına ulaşarak bu alandaki önemli bir boşluğu doldurmuştur.
Bowman vd. (2015), doğal dil çıkarımı (natural language inference - NLI) problemlerinin çözümü için geniş kapsamlı ve etiketlenmiş bir veri kümesi olan SNLI'yi (Stanford Natural Language Inference) sunmuştur. Bu veri kümesi, cümle çiftleriyle zenginleştirilmiş olup, anlamsal ilişki analizleri için makine öğrenimi modellerine önemli bir kaynak sağlamaktadır.
Nadir kelimelerin doğru şekilde çevrilmesi, makine çevirisi sistemlerinin başlıca problemlerinden biridir. Bu doğrultuda Luong vd. (2015), nadir kelime problemini çözmek amacıyla etkili bir yöntem önermiş ve geleneksel sistemlere kıyasla daha başarılı sonuçlar elde etmiştir. Sutskever vd. (2014) ise sıralı veri üzerinde çalışan Sequence-to-Sequence öğrenme yaklaşısını tanıtarak, giriş verilerini vektör temsiline dönüştürerek modelin bu verileri daha etkin öğrenmesini sağlamıştır.
Kelime temsilleri açısından önemli bir katkı sunan Pennington vd. (2014), GloVe (Global Vectors) modelini tanıtmıştır. Bu model, kelimeleri vektörlere dönüştürerek, anlamsal ilişkileri daha iyi yansıtan bir temsil sunmaktadır. Bu tür vektör temsilleri, kelimelerin bilgisayar tarafından işlenmesini kolaylaştırmakta ve çok sayıda doğal dil işleme görevinde kullanılmaktadır.
Zhao vd. (2014) tarafından geliştirilen Enchu modeli, hem anlamsal benzerlik hem de metin çıkarımı görevleri için kullanılabilecek heterojen

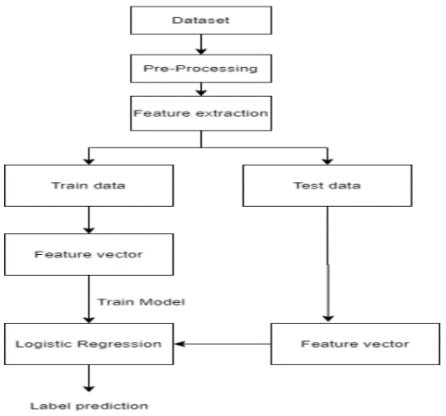
ölçütleri bir araya getiren bir topluluk (ensemble) yaklaşımı sunmuştur. Yedi farklı öznitelik üzerinden yapılan hesaplamalar, iki farklı görev için ortak benzerlik ölçütleri ile başarılı sonuçlar elde etmiştir.
Yukarıda özetlenen çalışmalar, doğal dil işlemede kullanılan derin öğrenme modellerinin hem anlamsal analiz hem de metin çıkarımı konularında nasıl yüksek performans sağladığını göstermektedir. Bu literatür, projenin amaçlarına yönelik en uygun model ve yaklaşımın seçilmesi noktasında önemli bir bilgi birikimi sunmaktadır. Özellikle cümle benzerliği, çıkarım gibi alt görevlerde dikkat mekanizmalarının ve kelime temsillerinin başarısı dikkat çekicidir.

3. ÖNERİLEN MODEL

Yandaki Şekil 1, önerilen modeli tanımlamaktadır. Modeli eğitmek için Lojistik Regresyon kullanılacaktır.
Bu çalışmada kullanılan veri seti Opensubtitles.org üzerinden alınmıştır.
Alınmış olup veri seti, ön işleme tabi tutulmuştur; bu işlemler noktalama işaretlerinin kaldırılması, tokenizasyon (parçalama), stopword'lerin (gereksiz kelimelerin) çıkarılması, köklerine ayırma (stemming) ve verinin vektörleştirilmesi (tf-idf) adımlarını içermektedir.

Fonksiyonların detayları aşağıda listelenmiştir:

Tokenizasyon – Her kelimeyi benzersiz bir tam sayı ile eşleyen büyük bir sözlük oluşturan bir süreçtir. Bu sözlük, cümleleri karakter dizileri dizisinden tamsayı dizilerine dönüştürmek için kullanılır.
Köklerine Ayırma (Stemming) – Uzatılmış kelime ifadelerinin kök halini elde etme sürecidir. Bu algoritma, örneğin “Likes, Liked, Likely, Liking” gibi kelimeleri “Like” köküne indirger.
Vektörleştirme – Bir algoritmanın tek bir değer yerine aynı anda birden fazla değer üzerinde çalışacak şekilde dönüştürülmesidir.
Tf-idf – “Term Frequency - Inverse Document Frequency” ifadesinin kısaltmasıdır ve metni anlamlı bir sayısal temsile dönüştürmek için yaygın olarak kullanılan bir algoritmadır. Bu teknik, çeşitli doğal dil işleme (NLP) uygulamalarında özellik çıkarımı için sıkça kullanılır.



Şekil 1. Model İş Akışı

3.1. VERİ SETİNİN OLUŞTURULMASI VE ÖN İŞLEME

3.1.1. Veri Toplama (Data Collection)

- Kaynak**:** Opensubtitles.org üzerinden 16 tema kategorisinden (aksiyon, dram, bilim kurgu vb.) 3’er film olmak üzere toplam 48 SRT dosyası indirilmiştir.
- Etiketleme (Labeling)**:** Her cümle, manuel olarak 16 sınıftan birine atanmıştır.

3.1.2. Metin Ön İşleme (Text Preprocessing)

- Tokenizasyon**:** Metinler kelimelere ayrılmıştır.
- Stop-word Eliminasyonu: Türkçe için hazırlanan stop-word listesi (bağlaçlar, edatlar vb.) çıkarılmıştır.
- Kök Bulma (Stemming)**:** Zemberek NLP kütüphanesi ile kelimeler kök formlarına indirgenmiştir (örneğin, "koşuyor" → "koş").
- Normalizasyon**:**
  Küçük harfe dönüştürme,
  Noktalama işaretlerinin kaldırılması,
  Sayılar ve özel karakterlerin temizlenmesi.

3.1.3. Chunklama (Metin Parçalama)

- Problem**:** Tek cümlelerle eğitimde düşük doğruluk (%30) elde edilmesi.
- Çözüm**:** Bağlamsal bilgiyi korumak için 20 cümlelik sabit uzunlukta ve **10 cümle** örtüşmeli (stride) parçalar oluşturulmuştur.
- Etiket Belirleme**:** Her chunk’ın etiketi, içerdiği cümlelerin mod (en sık geçen) etiketi ile atanmıştır.
- Sonuç**:** Doğruluk %85’e yükselmiştir.

3.2. MODEL MİMARİSİ

Önerilen model, Scikit-learn Pipeline yapısı ile uygulanmıştır:

3.2.1. Özellik Çıkarımı (Feature Extraction)

TF-IDF Vektörleştirici (Term Frequency-Inverse Document Frequency):

- Amaç: Metinleri sayısal vektörlere dönüştürerek anlamsal önem ağırlıkları atamak.
- Parametreler ve Açıklamaları:
  ngram_range=(1, 3): Unigram (tek kelime), bigram (iki kelime) ve trigram (üç kelime) kombinasyonlarını içerir. Dilbilimsel örüntüleri yakalamak için kritiktir.
  max_features=4000: En yüksek TF-IDF skoruna sahip 4000 terim seçilir. Hesaplama verimliliği için sınırlandırma.
  min_df=5: En az 5 belgede geçen terimler filtrelenir. Seyrek (sparse) terimlerin elenmesi.
  max_df=0.6: Terimlerin en fazla %60 belgede geçmesi sınırı. Stop-word benzeri sık terimlerin çıkarılması.
  sublinear_tf=True: Terim frekanslarının logaritmik ölçeklenmesi.

3.2.2. Sınıflandırıcı (Classifier)

Lojistik Regresyon (Logistic Regression):

- Amaç: Linear bir modelle çok sınıflı sınıflandırma (one-vs-rest stratejisi).
- Parametreler ve Açıklamaları:
  penalty='elasticnet': L1 (Lasso) + L2 (Ridge) regularizasyonu. Aşırı öğrenmeyi (overfitting) önlemek için katsayıları cezalandırır.
  l1_ratio=0.2: Regularizasyonun %20 L1, %80 L2 olarak uygulanması. L1, özellik seçimi yapar; L2, katsayıları küçültür.
  C=0.2: Regularizasyon gücü. Düşük C (0.2) = Daha güçlü regularizasyon.
  class_weight='balanced': Dengesiz sınıflar için otomatik ağırlıklandırma.

3.3. HİPERPARAMETRE OPTİMİZASYONU

GridSearchCV ile 5 katlı çapraz doğrulama (StratifiedKFold) yapılmıştır:

3.3.1. Optimize Edilen Parametreler

Parametre	Değer Aralığı	Seçilen Optimal Değer	Açıklama
tfidf__max_features	[3500, 4000, 4500]	4000	Özellik uzayının boyutunu düşürerek he
clf__C	[0.1, 0.2, 0.3]	0.2	Düşük C, daha güçlü regularizasyon ile
clf__l1_ratio	[0.1, 0.2, 0.3]	0.2	L2 ağırlıklı regularizasyon, model kararlı

Şekil 2. Parametre Değişim Grafiği

3.3.2. Parametre Değişimlerinin Sonuçları

Deney	max_features	C	l1_ratio	Eğitim Doğruluğu	Test Doğruluğu
1	5000	0.1	0.5	**100%**	89.85%
2	3000	0.05	0.5	71.16%	65.94%
3	**4000**	**0.2**	**0.2**	**95.84%**	**92.21%**

Şekil 3. Sonuç Grafiği

Deney 1: Aşırı öğrenme (overfitting) nedeniyle **test** doğruluğu düşük. Çözüm: C değeri düşürülerek regularizasyon artırıldı.
Deney 2: Aşırı regularizasyon nedeniyle model underfit yapmış.
Çözüm: C ve l1_ratio dengelendi.
Deney 3 (Optimal): Dengeli regularizasyon ile overfitting minimize edilirken, yüksek genelleme başarısı sağlandı.

3.4. PERFORMANS DEĞERLENDİRMESİ

3.4.1. Metrikler

- Accuracy (Doğruluk): 92.21%
- Weighted F1-Score: 92%
- Sınıf Bazında Performans:
  En Yüksek F1: Animasyon (0.99), Bilim Kurgu (0.97).
  En Düşük F1: Romantik (0.85), Komedi (0.87).

4.2. Karışım Matrisi Analizi

- En Çok Karışan Sınıflar:
  Romantik → Müzik (%8.89)
  Komedi → Müzik (%4.72)
- Sebep: Bu temaların diyalog yapılarının benzerliği (örneğin, duygusal ifadeler).

Şekil 5. Film Tema Analizi Ekranı

4. SONUÇ

Bu çalışmada, Türkçe film alt yazı metinlerinden otomatik tema sınıflandırması için bir makine öğrenmesi modeli geliştirilmiş ve başarılı bir şekilde uygulanmıştır. Önerilen model, lojistik regresyon tabanlı bir sınıflandırıcı olup, TF-IDF vektörleştirme ve metin chunklama teknikleri ile desteklenmiştir. Yapılan deneysel çalışmalar sonucunda, model %92.21 test doğruluğu ve %92 ağırlıklı F1-skoru elde etmiştir.

4.1. Başlıca Katkılar

Özgün Veri Seti: Türkçe film alt yazılarından oluşan, 16 tema kategorisinde elle etiketlenmiş bir veri seti oluşturulmuştur.
Metin Ön İşleme ve Chunklama: Tek tek cümleler yerine örtüşmeli metin parçaları (chunk) kullanılarak bağlamsal bilgi korunmuş ve doğruluk %30'dan %85'e çıkarılmıştır.
Hiperparametre Optimizasyonu: GridSearchCV ile en uygun parametreler belirlenmiş, ElasticNet regularizasyonu sayesinde overfitting kontrol altına alınmıştır.
Kullanıcı Dostu Arayüz: Geliştirilen model, kullanıcıların SRT dosyalarını yükleyerek film teması analizi yapabileceği bir web arayüzüne entegre edilmiştir.

4.2. Modelin Sınırlılıkları

Veri Seti Boyutu: Her tema için yalnızca 3 film kullanılmış olup, daha fazla veri ile modelin genelleme performansı artırılabilir.
Derin Öğrenme Eksikliği: Mevcut model, linear bir yaklaşım kullanmaktadır. LSTM, BERT gibi derin öğrenme modelleri dil yapılarını daha iyi öğrenebilir.
Karışan Sınıflar: Özellikle romantik-komedi-müzik temaları arasında yüksek karışım oranları gözlemlenmiştir. Bu sınıflar için ek özellik mühendisliği (örn. duygu analizi) gerekebilir.

4.3. Gelecek Çalışmalar İçin Öneriler

Veri Setinin Genişletilmesi: Her tema kategorisi için film sayısı artırılarak modelin kararlılığı iyileştirilebilir.

Derin Öğrenme Modelleri: Transformers (BERT, DistilBERT) veya CNN-LSTM hibrit modelleri ile dilsel örüntülerin daha karmaşık temsilleri öğrenilebilir.
Çok Dilli Modeller: İngilizce-Türkçe çift dilli veri setleri ile karşılaştırmalı tema analizi yapılabilir.
Etki Analizi: Film temaları ile izleyici beğenisi arasındaki korelasyonlar incelenilir.

Son Söz

Bu çalışma, Türkçe doğal dil işleme (NLP) alanında metin tabanlı tema sınıflandırması için yenilikçi bir yaklaşım sunmuştur. Geliştirilen model, yüksek doğruluk ve pratik uygulanabilirlik ile öne çıkmakta olup, sinema endüstrisi, içerik öneri sistemleri ve dijital kütüphaneler gibi alanlarda kullanılabilir. Gelecek çalışmalarda, derin öğrenme tabanlı modeller ve genişletilmiş veri setleri ile performansın daha da artırılması hedeflenmektedir.

Kaynakça

Şeker, A., Dirib, B., & Balık, H. H. (2017). Derin Öğrenme Yöntemleri ve Uygulamaları Hakkında Bir İnceleme. Gazi Journal of Engineering Sciences (GJES), 3(3), 47- 64

Schütze, H., & Manning, C. D. (1999). Foundation of Statistical Natural Language Processing.

Chomsky, N. (1986). Knowledge of Language: Its Nature, Origin, and Use. New York.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine, 55-75.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine, 9(2), 48-57.

Song, H., & Lee, S.-Y. (2013). Hierarchical Representation Using NMF. International Conference on Neural Information Processing 2013: Neural Information Processing, (s. 466-473).

Aksan, Y., & Yaldır, Y. (2012). A corpus-based word frequency list of Turkish: Evidence from the Turkish National Corpus. 15 th International Conference on Turkish Linguistics.

Bill, E. (1992). A Simple Rule-Based Part of Speech Tagger. ANLC '92 Proceedings of the third conference on Applied natural language processing. Trento.

Ivakhnenko, A., & Lapa, V. (1965). Cybernetic predicting devices.

Çöltekin, Ç. (2014). A Set of Open Source Tools for Turkish Natural Language Processing. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), (s. 1079–1086).

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, Phil Blunsom. Reasoning about Entailment with Neural Attention. 2016. [Online]. Available: https://arxiv.org/abs/1509.06664

Minh-Thang Luong, Hieu Pham, Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. 2015. [Online]. Available: https://arxiv.org/abs/1508.04025

Jonas Mueller and Aditya Thyagarajan. Siamese Recurrent Architectures for Learning Sentences Similarity. 2016. [Online]. Available: https://dl.acm.org/doi/10.5555/3016100.3016291

Shuohang Wang and Jing Jiang. Learning natural language inference with LSTM. In Proceedings of NAACL, 2016. [Online]. Available: https://www.aclweb.org/anthology/N16- 1170/

Samuel R. Bowman, Gabor Angeli, Christopher D. Manning. A large annotated corpus for learning natural language inference. In EMNLP,2015. [Online]. Available: https://arxiv.org/abs/1508.05326

Minh-Thang Luong, IIlya Sutskever, Quoc V. Le, Oriol Vinyals, Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In ACL, 2015. [Online]. Available: https://arxiv.org/abs/1410.8206

Minh-Thang Luong, IIlya Sutskever, Quoc V. Le, Oriol Vinyals, Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In ACL, 2015. [Online]. Available: https://arxiv.org/abs/1410.8206

IIlya Sutskever, Quoc V. Le and Quoc V Le. Sequence to sequence learning with neural network. In NIPS, 2014. [Online]. Available: https://papers.nips.cc/paper/5346-sequence-tosequence-learning-with-neural-networks.pdf

Jeffrey Pennington, Richard Socher and Christopher D. Manning. 2014. GloVe:Global Vectors for Word Representation. [Online]. Available: https://www.aclweb.org/anthology/D14-1162.pdf

Jiang Zhao, Tian Zhu and Man Lan. Enchu: One stone two birds:Ensemble of heterogenous measures for semantic relatedness and textual entailment. In SemEval, 2014. [Online]. Available: https://www.aclweb.org/anthology/S14-2044.pdf

https://www.opensubtitles.org/tr/search/sublanguageid-all

https://tahtaciburak.medium.com/zemberek-k%C3%BCt%C3%BCphanesi-pythonda-nas%C4%B1l-kullan%C4%B1l%C4%B1r-8993ec1c3f0e

https://github.com/ahmetaa/zemberek-nlp

https://www.kaggle.com/code/leohuntera/text-preprocessing-nlp-steps-to-process-text/