Erciyes Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 41, Sayı 2, 2025

Erciyes University
Journal of Institue Of Science and Technology
Volume 41, Issue 2, 2025

Film Açıklamalarında Anlamsal Benzerlik Kullanımının Tür Sınıflandırma Performansına Etkisi

Zeynep KOYUN[1], **Fatih Ahmet ŞENEL***[2]

*[1] Isparta Uygulamalı Bilimler Üniversitesi Keçiborlu Meslek Yüksekokulu, ISPARTA
[2] Süleyman Demirel Üniversitesi Mühendislik ve Doğa Fakültesi Bilgisayar Mühendisliği, ISPARTA

(Alınış / Received: 20.06.2025, Kabul / Accepted: 04.08.2025, Online Yayınlanma / Published Online: 30.08.2025)

Anahtar Kelimeler
Doğal Dil İşleme,
FastText,
Kelime Gömmesi,
MEtin Sınıflandırma,
TF-IDF

Öz: Bu çalışmada, İngilizce film açıklamalarının altı farklı türde (aksiyon, komedi, romantik, bilim kurgu, drama, animasyon) otomatik olarak sınıflandırılması hedeflenmiştir. Dengeli dağılıma sahip veri kümesi, doğal dil işleme (NLP) teknikleri ile işlenmiştir. İlk aşamada, FastText'in önceden eğitilmiş kelime gömme vektörleri kullanılarak açıklamalardaki anlamsal olarak düşük benzerlik gösteren kelimeler filtrelenmiş, böylece her tür için bağlama özgü içerik korunarak metinsel verinin kalitesi artırılmıştır. Bu ön işleme adımının ardından, metinlerin sayısal temsili TF-IDF (Term Frequency–Inverse Document Frequency) yöntemiyle elde edilmiştir. TF-IDF, sık tekrar eden ancak ayırt edici gücü düşük kelimelerin etkisini azaltarak sınıflandırma algoritmalarına daha anlamlı öznitelikler sunmaktadır. Elde edilen vektörler, Softmax Regresyon, Multinomial Naive Bayes, Destek Vektör Makineleri, Rastgele Orman ve Çok Katmanlı Algılayıcı modelleri ile sınıflandırılmış; sırasıyla %89,07, %89,75, %89,75, %85,40 ve %90,02 doğruluk oranlarına ulaşılmıştır. FastText tabanlı anlamsal kelime filtreleme ile TF-IDF vektörlerinin birleştirilmesi, çok sınıflı metin sınıflandırma problemleri için etkili bir çözüm sunmuştur. Ayrıca, farklı makine öğrenmesi algoritmalarının karşılaştırılması yoluyla model seçimine yönelik ampirik katkı sağlanmış olup , anlamca sadeleştirilmiş açıklamaların model başarısını olumlu etkilediği gösterilmiştir. Bu yönüyle çalışma, metin ön işleme sürecine yenilikçi bir yaklaşım getirmektedir.

Impact of Semantic Similarity on Genre Classification Performance in Movie Descriptions

Keywords
FastText,
Natural Language Processing,
Text Classification,
TF-IDF,
Word Embedding

Abstract: The objective of this study is to develop a system that automatically categorizes English movie descriptions into six distinct genres: Action, Comedy, Romance, Science Fiction, Drama, and Animation. The dataset, characterized by a balanced distribution across these genres, was processed using natural language processing (NLP) techniques. In the initial phase, pre-trained FastText word embedding vectors were employed to filter out words exhibiting low semantic similarity within the descriptions. This step aims to preserve context-specific content relevant to each genre, thereby enhancing the quality of the textual data. Subsequently, the numerical representation of the texts was obtained using the Term Frequency–Inverse Document Frequency (TF-IDF) method, which is known to improve the performance of classification algorithms by reducing the influence of frequently occurring but less informative words. The resulting vectors were classified using Softmax Regression, Multinomial Naive Bayes, Support Vector Machines, Random Forest, and Multilayer Perceptron models, achieving accuracy rates of 89.07%, 89.75%, 89.75%, 89.75%, 85.40%, and 90.02%, respectively. The integration of FastText-based semantic filtering with TF-IDF vectorization proved to be an effective approach for multi-class text classification tasks. Moreover, an empirical comparison of different machine learning algorithms provides valuable insights for model selection. The findings suggest that semantically simplified

descriptions contribute positively to model performance. In this regard, the study makes a significant contribution to text preprocessing in genre classification tasks.

*İlgili Yazar, email: fatihsenel@sdu.edu.tr

1. Giriş

Günümüzde dijital platformlarda ve kurumsal veri tabanlarında hızla artan metin tabanlı veri miktarı, etkin bilgi yönetimi ve analiz süreçlerinin önünü açmaktadır. Bu kapsamda, metin sınıflandırma algoritmalarının geliştirilmesi ve uygulanması, büyük ölçekli veri setlerinde anlamlı bilgi çıkarımı için zorunlu hale gelmiştir. Metin sınıflandırma, verilen metinlerin önceden tanımlanmış sınıflara otomatik olarak atanması işlemidir ve doğal dil işleme alanında temel bir problem olarak kabul edilmektedir.

Metinlerin manuel olarak kategorize edilmesi hem zaman alıcı hem de insan hatasına açık bir süreçtir. Bu nedenle, özellikle haber ajansları, e-ticaret platformları, sosyal medya analizleri, müşteri geri bildirimleri ve otomatik içerik yönetimi sistemlerinde, otomatik metin sınıflandırma tekniklerine olan ihtiyaç kritik bir hal almıştır. Bu teknikler, bilgi erişimini hızlandırmak, veri organizasyonunu kolaylaştırmak ve üst düzey karar destek sistemlerine altyapı oluşturmak için kullanılmaktadır. Teknik olarak, metin sınıflandırma probleminin başarısı, metinlerin etkili ve doğru sayısal temsiline bağlı olmaktadır. Bu amaçla kelime gömme (word embedding) yöntemleri, özellikle FastText gibi önceden eğitilmiş modeller, metinlerin anlamsal özelliklerini yakalamada yaygın olarak tercih edilmektedir [1]. Aynı zamanda, klasik metin temsil yöntemlerinden olan TF-IDF algoritması, sık kullanılan ancak ayırt edici özelliği düşük kelimelerin etkisini minimize ederek sınıflandırma modellerine daha anlamlı öznitelikler sunmaktadır [2].

Metin sınıflandırma alanında, özellikle İngilizce veri setleri üzerinde birçok farklı yöntem geniş çapta incelenmiştir. Geleneksel makine öğrenmesi algoritmaları ve derin öğrenme algoritmaları gibi yöntemler, özellikle sınırlı veri ve düşük hesaplama kaynaklarında etkili sonuçlar vermektedir. Bu yöntemler büyük ölçüde İngilizce veri setleri üzerinde geliştirilip test edilmekle birlikte, Türkçe gibi morfolojik açıdan daha karmaşık ve sözdizimsel farklılıklara sahip diller için uyarlanabilirliği ve performansları konusunda da çalışmalar yapılmaktadır [3-14]. Bu nedenle, Türkçe diline özgü önceden eğitilmiş modeller (örneğin BERTurk) ve Türkçe doğal dil işleme alanındaki çalışmalar, İngilizce ağırlıklı literatüre tamamlayıcı nitelikte olup, farklı dil yapılarının metin sınıflandırma üzerindeki etkisini ortaya koymak açısından önem taşımaktadır.

Gao ve arkadaşları [15], metin sınıflandırma performansını artırmak amacıyla çok seviyeli dikkat mekanizması ile karşıt öğrenmeyi birleştiren yeni bir yöntem önermişlerdir. Transformer mimarisi üzerine inşa edilen bu modelde, farklı soyutlama düzeylerinde dikkat yapıları kullanılarak metinlerdeki anlamlı ögelerin daha etkili biçimde temsil edilmesi sağlanmıştır. Ayrıca, karşıt öğrenme ile benzer örneklerin birbirine yakın, farklı örneklerin ise uzak konumlandırılması hedeflenmiş ve böylece modelin ayrıştırma kabiliyeti güçlendirilmiştir. AG News, 20 Newsgroups ve Yahoo! Answers veri setleriyle yapılan deneylerde, model özellikle AG News üzerinde %95,2 doğruluk ve %94,7 $F1$ skoru ile BERT ve RoBERTa gibi güçlü taban modellere kıyasla daha başarılı sonuçlar üretmiştir. Bu bulgular, çok seviyeli dikkat ve karşıt öğrenmenin birlikte kullanımının bağlamsal temsil gücünü artırarak metin sınıflandırma görevlerinde anlamlı performans iyileştirmeleri sağladığını ortaya koymaktadır.

Zhang vd. [16], klinik metin sınıflandırma görevlerinde karşılaşılan görevler arası bilgi dengesizliği sorununa çözüm olarak MTTL-ClinicalBERT adını verdikleri simetrik çok görevli öğrenme tabanlı bir model geliştirmişlerdir. ClinicalBERT üzerine inşa edilen bu model, görevler arası dengeli bilgi aktarımını hedefleyen simetrik bir öğrenme mekanizması ile optimize edilmiştir. Model, her bir görevi hem kaynak hem hedef olarak ele alması sayesinde, görevler arası bilgi paylaşımı çift yönlü şekilde gerçekleştirilmiştir. i2b2 2010, Medical NER, ShARe/CLEF ve n2c2 gibi klinik veri setleri üzerinde yapılan deneylerde, önerilen yöntemin özellikle görev dengesizliğinin yüksek olduğu senaryolarda ClinicalBERT, BioBERT ve BioClinicalBERT gibi güçlü modellerden daha iyi performans gösterdiği görülmüştür. Örneğin, i2b2 veri seti üzerinde elde edilen %88,6 $F1$ skoru, önerilen yaklaşımın etkili ve genelleştirilebilir bir çözüm sunduğunu göstermektedir. Bu çalışma, çok görevli öğrenme yapısında simetrik bilgi aktarımının klinik NLP uygulamalarında başarımı artırıcı etkisini ortaya koymuştur.

Jamshidi vd. [17], metin sınıflandırma görevlerinde doğruluk ve açıklanabilirliği birlikte artırmak amacıyla BERT, MTM-LSTM ve Karar Ağacı algoritmalarını birleştiren çok bileşenli bir yöntem önermişlerdir. Önerilen yaklaşımda, BERT ile elde edilen bağlamsal temsiller MTM-LSTM yapısında çok görevli öğrenmeye tabi tutulmuş, ardından bu temsiller Karar Ağacı algoritması ile sınıflandırılmıştır. Bu yapı sayesinde hem derin öğrenmenin genelleme gücü hem de karar ağaçlarının yorumlanabilirliği bir araya getirilmiştir. BBC News, 20 Newsgroups ve Reuters-21578 veri setleriyle yapılan deneylerde model, özellikle 20 Newsgroups üzerinde %96,4 doğruluk ve %95,8 $F1$ skoru ile klasik yöntemlerden üstün performans sergilemiştir. Bu çalışma, çok görevli öğrenme ile klasik

sınıflandırıcıların entegrasyonunun, metin sınıflandırma problemlerinde hem doğruluk hem de açıklanabilirlik açısından anlamlı katkılar sunduğunu göstermektedir.

Lepagnol vd. [18], büyük dil modellerinin yaygın kullanımı karşısında, küçük dil modellerinin (SLM) de sıfır-atış (zero-shot) metin sınıflandırma görevlerinde rekabetçi performans gösterebileceğini deneysel olarak ortaya koymuşlardır. Çalışmada, BERT-tabanlı ve DistilBERT gibi hafifletilmiş modeller, herhangi bir yeniden eğitim yapılmaksızın yalnızca doğal dil yönlendirmeleriyle test edilmiş; uygun görev tanımlamaları sayesinde anlamlı sınıflandırma sonuçları elde edilmiştir. AG News, DBPedia, Yahoo! Answers ve SST-2 gibi veri setleri üzerinde gerçekleştirilen analizlerde, bazı görevlerde %80'in üzerinde doğruluk sağlanmış; örneğin DBPedia veri setinde DistilBERT modeliyle %88 doğruluk elde edilmiştir. Bu bulgular, düşük kaynaklı ortamlarda küçük modellerin pratik ve ekonomik bir alternatif olabileceğini göstermekte; NLP alanında daha erişilebilir çözümler için önemli bir katkı sunmaktadır.

Literatürde İngilizce veriseti ile yapılan çalışmaların sayısı artırılabilmektedir [19-23]. Genel olarak literatür incelendiğinde yapılan çalışmalarda IMDb, AG News, DBPedia, Yahoo! Gibi platformlardan verilerin toplandığı görülmektedir. Ayrıca yapılan çalışmalarda anlamsal bağ kurabildiği için BERT tabanlı modellerin kullanıldığı da görülmektedir. Bununla birlikte BERT tabanlı modellerin her zaman iyi sonuçlar vermediği ve iyileştirme adımlarına ihtiyaç duyulduğu da anlaşılmaktadır. Bu çalışmada, iki farklı yöntem kullanarak İngilizce film açıklamalarını altı farklı türe (aksiyon, komedi, romantik, bilim kurgu, drama ve animasyon) TF-IDF temelli sayısal temsil yöntemlerinin entegrasyonunu kullanarak otomatik şekilde sınıflandırmak ve farklı makine öğrenmesi algoritmalarının performanslarını karşılaştırmaktır. Yöntem 1 ve Yöntem 2 olarak iki farklı metot amaçlanmıştır. Yöntem 1'de temel ön işlem aşamaları olan durak kelimelerin kaldırılması ve tüm harflerin küçük harfe çevrilmesi işlemi gerçekleştirilirken, Yöntem 2'de ise FastText katmanı ön işlem aşamasından sonra uygulanmış, film açıklamaları tür adı ile bağlamsal benzerliği dikkate alınarak yeniden oluşturulmuştur. Yöntem 2 ile önerilen yöntemin daha başarılı sonuçlar verdiği gözlemlenmiştir. Ayrıca yapılan karşılaştırmalara anlamsal bağlamlara dikkat ederek kelime gömme işlemini yapan BERT modelinin sonuçları da eklenmiştir. Elde edilen sonuçlara göre anlamsal bağlama dikkat etmeyen FastText yöntemi, BERT modelinden daha başarılı sonuçlar vermiştir.

Bu çalışmanın ikinci bölümünde kullanılan materyal metotlar detaylı şekilde açıklanmıştır. Üçüncü bölümde elde edilen analiz sonuçlarına yer verilmiştir. Son bölümde ise çalışmanın sonuçları tartışılmıştır.

2. Materyal ve Metot

Bu çalışmada film açıklamalarının türlerine göre otomatik olarak sınıflandırılması amacıyla NLP ve makine öğrenmesi tabanlı bir yaklaşım önerilmiştir. Önerilen yaklaşımda iki farklı ön işleme adımı uygulanmıştır. Yöntem 1 ve Yöntem 2 olarak isimlendirdiğimiz ön işleme adımlarında Yöntem 1, FastText modeli içermemektedir. Yöntem 2 modelimiz, Yöntem 1 modeline ek olarak FastText katmanı içermektedir. Yöntem 1 ve Yöntem 2 modellerimizin detaylı analizleri yapılarak karşılaştırmalar yapılmış ve FastText modelinin uygulanmasının sınıflandırma başarısına etkisi incelenmiştir. Bu bağlamda, yöntemsel süreç dört ana aşamada ele alınmıştır: veriseti hazırlığı, metin ön işleme ve özellik çıkarımı, sınıflandırma modellerinin eğitimi ve performans değerlendirmesi aşamalarıdır.

2.1. Veriseti ve Ön İşleme

Bu çalışmada, IMDb platformundan elde edilen film açıklamaları ve tür bilgileri geliştirilen bir script aracılığı ile toplamda altı türe ait film verileri otomatik olarak toplanmıştır. Toplanan veri üzerinde gerçekleştirilen ön incelemelerde, bazı filmlerin birden fazla türe ait olduğu tespit edilmiştir. Bu durumu ortadan kaldırmak amacıyla, verisetimizde sadece tek bir türe ait filmler kalacak şekilde eleme yapılmıştır. Her bir tür için 1225 açıklama içerecek şekilde dengeli bir dağılım sağlanmış ve toplamda 7350 açıklamadan oluşan homojen bir veri kümesi elde edilmiştir. Topladığımız verisetine ait dağılım istatistik verileri Tablo 1'de verilmiştir.

Tablo 1. Verisetine Ait Dağılım İstatistikleri

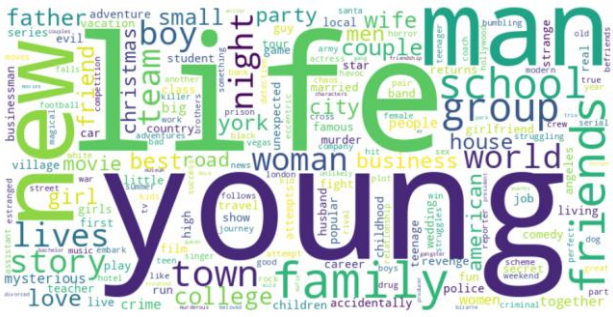
Tür	Veri Sayısı	Ortalama Kelime Sayısı	Minimum Kelime Sayısı	Maksimum Kelime Sayısı	Kelime Sayısının Medyan Değeri	Kelime Sayısı Dağılımının Standart Sapması
Aksiyon	1225	27,04	6	163	26	9,62
Animasyon	1225	35,65	5	497	30	29,55
Komedi	1225	26,47	8	130	25	9,33
Drama	1225	26,21	7	64	25	8,27
Romantik	1225	26,47	5	181	25	11,31
Bilim-Kurgu	1225	27,94	5	134	26	12,94

Tablo 1 incelendiğinde genel olarak, açıklamalar ortalama 28,29 kelime uzunluğundadır. Bu uzunluklar minimum 5 ve maksimum 497 kelime arasında değişmektedir. Açıklama uzunluklarının standart sapması ortalama 15,71 olup, bu da veri içinde açıklama uzunluklarının orta derecede çeşitlilik içerdiğini göstermektedir. Ayrıca, açıklamaların %50'si (medyan) 26 kelime veya daha kısadır. Bu durum, veri kümesindeki açıklamaların genellikle kısa tutulduğunu göstermektedir. Tür bazında incelendiğinde, her tür eşit sayıda film içermektedir. En yüksek ortalama kelime sayısına sahip tür animasyon türüdür; bu türde açıklamalar ortalama 35,65 kelime olup, açıklama uzunluklarının standart sapması da diğer türlere kıyasla oldukça yüksektir. Bu da animasyon türündeki açıklamaların hem daha uzun hem de daha değişken olduğunu göstermektedir. Diğer yandan, drama, komedi ve romantik türleri ortalama 27 kelime ile daha kısa açıklamalara sahiptir. Bu türlerdeki açıklamalar daha tutarlı olup, standart sapmaları da düşüktür. Aksiyon ve bilim-kurgu türleri ise ortalama kelime sayısı bakımından orta seviyede yer almaktadır. Sonuç olarak, IMDb kaynaklı bu veri setinde açıklama uzunlukları türlere göre farklılık göstermekle birlikte genel eğilim kısa açıklamalar yönündedir.

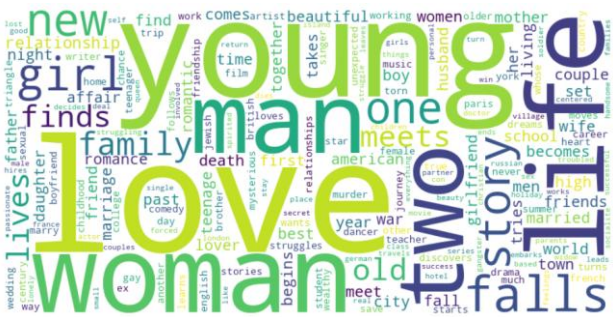
Verisetini Tablo 1'de sunulduğu şekilde yapılandırmamızın temel amacı, geliştirilecek makine öğrenmesi ve derin öğrenme modellerinin her türe yönelik dengeli ve adil bir öğrenme süreci gerçekleştirmesini sağlamaktır. İlk olarak açıklamalarda bulunan durak kelimeler temizlenmiştir. Daha sonra her bir film açıklamasındaki kelimeler FastText modeli kullanılarak vektör temsillerine dönüştürülmüştür. Daha sonra, her kelimenin vektörü ile ait olduğu tür adının vektörü arasındaki kosinüs benzerliği hesaplanmış ve eşik değeri 0,1 ve üzeri olan kelimeler kullanılarak yeni açıklamalar oluşturulmuştur. Böylece veriseti kullanıma hazır hale gelmiştir.

Belirlenen bu düşük eşik değeri, yalnızca türüyle güçlü semantik bağa sahip kelimeleri değil, aynı zamanda bağlamsal açıdan anlamlı ancak zayıf ilişkiler kuran kelimeleri de korumamıza olanak tanımaktadır. Bu sayede, açıklamalarda yer alan semantik çeşitlilik ve bağlamsal zenginlik muhafaza edilmiştir. Nitekim Singh vd. [24] de belirttiği üzere, kelime gömme tabanlı sistemlerde sabit ve yüksek eşikler yerine, bağlama duyarlı daha düşük eşiklerin tercih edilmesi daha verimli sonuçlar doğurmaktadır. Singh vd. [24]'nin çalışmasında olduğu gibi bu çalışmada da 0,1 eşik değeri seçilmiştir. Ayrıca deneysel sonuçlar başlığı altında farklı eşik değerleri ile analizler gerçekleştirilerek 0,1 eşik değerinin en iyi seçim olduğu teyit edilmiştir. Bu bağlamda, örneğin "bıçak" kelimesi doğrudan "aksiyon" kelimesiyle eş anlamlı olmasa da bağlamsal ilişkisi nedeniyle açıklamada tutulmuştur. Eşik değeri olarak 0,1'in seçilmesi, türüyle anlamlı ilişkisi bulunan kelimelerin belirlenmesini sağlarken, aynı zamanda metnin semantik bütünlüğünün korunmasına da katkıda bulunmuştur. Uygulanan bu ön işleme süreci sonucunda, her film açıklamasında yalnızca ait olduğu türüyle anlamlı bağ kuran kelimeler korunmuş ve böylece her bir film için, türe özgü anlam taşıyan kelimelerden oluşan sadeleştirilmiş açıklamalar elde edilmiştir. Son olarak, elde edilen bu açıklamalar ve karşılık gelen tür etiketleri bir liste halinde kaydedilmiş, ardından pandas kütüphanesi kullanılarak bir DataFrame formatına dönüştürülmüş ve makine öğrenmesi ile derin öğrenme modellerinde kullanılmak üzere hazır hale getirilmiştir. Oluşturulan temizlenmiş veri kümesinde, her türe özgü önemli kelimeleri görselleştirmek amacıyla TF-IDF vektörleri kullanılarak kelimelerin metin içerisindeki önem skorları hesaplanmıştır. Bu skorlar, kelimelerin ilgili türüyle olan bağını yansıtmakta ve görselleştirme sürecine katkı sağlamaktadır. Elde edilen kelimeler ve önem dereceleri, kelime bulutlarının oluşturulmasında kullanılmış; her türe ait tematik içeriği yansıtan ayrı görseller üretilmiştir. Böylece, veri kümesindeki her bir türün özgün dilsel yapısı ve tematik eğilimleri görsel olarak ortaya konmuştur. İlgili kelime bulutlarına Şekil 1, 2, 3, 4, 5 ve 6'da yer verilmiştir.

Şekil 1. Aksiyon Türü için TF-IDF Öne Çıkan Kelimeler



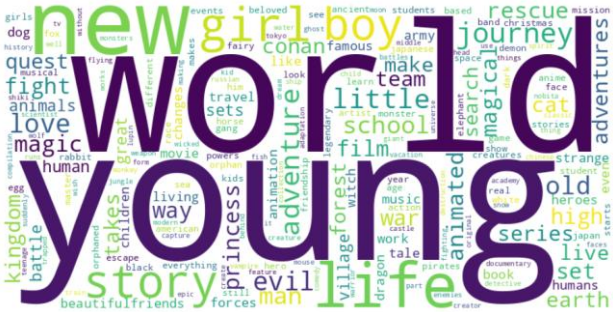
Şekil 2. Komedi Türü için TF-IDF Öne Çıkan Kelimeler



Şekil 3. Romantik Türü için TF-IDF Öne Çıkan Kelimeler

Şekil 4. Bilim-Kurgu Türü için TF-IDF Öne Çıkan Kelimeler

Şekil 5. Drama Türü için TF-IDF Öne Çıkan Kelimeler



Şekil 6. Animasyon Türü için TF-IDF Öne Çıkan Kelimeler

Tablo 2. Film Türlerine Göre Temsili Açıklamalar ve Sayısal Etiketler

Film Açıklamaları	Sayısal Etiketler
Blacksmith Will Turner teams up with eccentric pirate \"Captain\" Jack Sparrow to save Elizabeth Swann, the governor's daughter and his love, from Jack's former pirate allies, who are now undead.	0
Unscrupulous boxing promoters, violent bookmakers, a Russian gangster, incompetent amateur robbers and supposedly Jewish jewelers fight to track down a priceless stolen diamond.	1
A former San Francisco police detective juggles wrestling with his personal demons and becoming obsessed with the hauntingly beautiful woman he has been hired to trail, who may be deeply disturbed.	2
After investigating a mysterious transmission of unknown origin, the crew of a commercial spacecraft encounters a deadly lifeform.	3
A banker convicted of uxoricide forms a friendship over a quarter century with a hardened convict, while maintaining his innocence and trying to remain hopeful through simple compassion.	4
During her family's move to the suburbs, a sullen 10-year-old girl wanders into a world ruled by gods, witches and spirits, and where humans are changed into beasts.	5

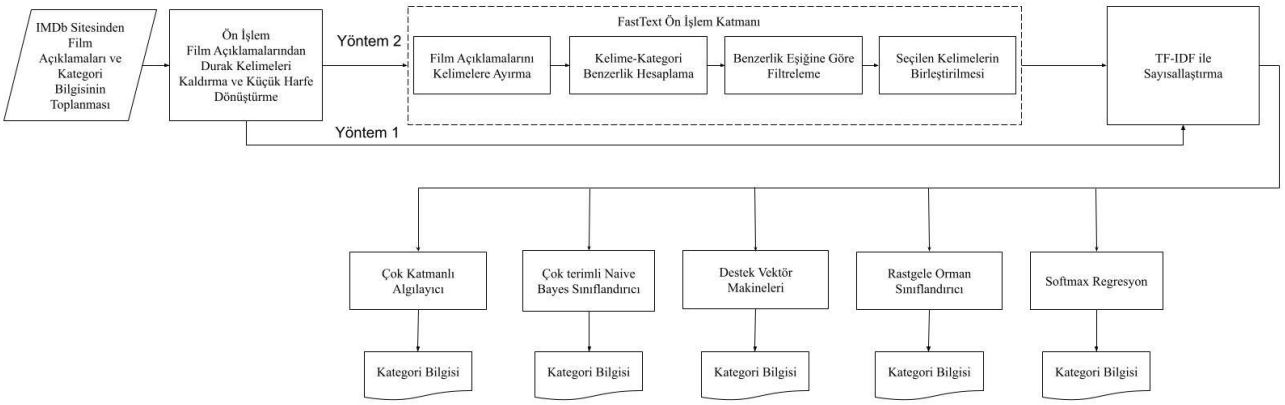
Verisetindeki film türleri sayısal etiketlerle eşleştirilmiş olup, sırasıyla aksiyon (0), komedi (1), romantik (2), bilim-kurgu (3), drama (4) ve animasyon (5) olarak kodlanmıştır. Verisetine ait temsili açıklamalar ve sayısal etiketler Tablo 2'de gösterilmiştir.

2.2. Model Mimarisi ve Algoritmalar

IMDb platformundan elde edilen film açıklamaları ve bu açıklamalara ait tür bilgilerine dayalı olarak yürütülen film türü sınıflandırma süreci, birden fazla aşamadan oluşan sistematik bir veri işleme ve modelleme yaklaşımıyla gerçekleştirilmiştir. Bu sürecin ilk adımında, film açıklamaları metinsel olarak temizlenmiş; durak kelimeler metinlerden çıkarılmış ve tüm sözcükler küçük harfe dönüştürülerek dilsel standartlaşma sağlanmıştır. Bu temel ön işleme adımlarının ardından, veriler iki farklı işleme hattına (Yöntem 1, Yöntem 2) yönlendirilmiştir.

İlk hat, yalnızca temel işlemleri içeren geleneksel bir metin hazırlama sürecidir. İkinci hat ise, FastText modeli kullanılarak daha gelişmiş bir anlamsal filtreleme yaklaşımı sunmaktadır. Bu aşamada, açıklamalar sözcüklerine ayrıldıktan sonra her bir kelime ile hedef tür arasındaki semantik benzerlikler FastText vektörleri yardımıyla hesaplanmış, düşük benzerlik değeri taşıyan kelimeler metinden filtrelenmiş ve yalnızca yüksek anlam ilişkisine sahip sözcükler birleştirilerek açıklamalar yeniden yapılandırılmıştır. Bu yöntem, her bir türüyle bağlamsal olarak daha ilişkili içeriklerin öne çıkarılmasını amaçlamaktadır.

Her iki ön işleme yolunun çıktıları, TF-IDF yöntemiyle sayısal vektörlere dönüştürülmüş ve bu temsiller çeşitli makine öğrenmesi algoritmalarıyla eğitilmiştir. Kullanılan modeller arasında Softmax Regresyonu (SR), Rastgele Orman Sınıflandırıcı (RF), SVM, Çok Terimli Naive Bayes Sınıflandırıcı (MNB) ve MLP bulunmaktadır. Eğitim süreci sonucunda, her film açıklaması için uygun film türü tahmin edilerek sınıflandırma işlemi gerçekleştirilmiştir. Tüm bu işlem adımlarını ve bileşenlerini görsel olarak bir araya getiren akış diyagramı Şekil 7'de sunulmuştur.



Şekil 7. Akış Diyagramı

Şekil 7'deki akış diyagramında verilen FastText ön işlem aşaması sonucunda elde edillen film açıklamaları ile orijinal film açıklamaları Tablo 3'te verilmiştir. Tablo 3 ile her tür için seçilen örnek cümlelerin, FastText benzeri yöntemlerle ön işlemden geçirilerek, sadece türle ilişkili anahtar kelimelerin bırakıldığı sadeleştirilmiş versiyonları yer almaktadır. Böylece, metin sınıflandırma modelleri için tür ayırt edici özelliklerin belirlenmesi amaçlanmıştır.

Tablo 3. Film Türlerine Ait Örnek Açıklamalar ve Ön İşlem Sonuçları

Sayısal Etiketler	Orijinal Film Açıklamaları	FastText Ön İşlem Adımı Sonrası Film Açıklamaları
0	when jason bourne is framed for a cıa operation gone awry, he is forced to resume his former life as a trained assassin to survive.	jason bourne framed cıa operation trained assassin survive.
1	three buddies wake up from a bachelor party in las vegas with no memory of the previous night and the bachelor missing. they must make their way around the city in order to find their friend in time for his wedding.	three buddies wake up bachelor party las vegas with no memory bachelor missing find friend wedding.
2	mansoor, a reserved and reticent pithoo, helps pilgrims make an arduous journey upwards to the temple town. his world turns around when he meets the beautiful and rebellious mukku who draws him into a whirlwind of intense love.	reticent pilgrims journey temple town. world turns meets beautiful rebellious mukku draws whirlwind intense love.
3	while scavenging the deep ends of a derelict space station, a group of young space colonists come face to face with the most terrifying life form in the universe.	derelict space station, young space colonists terrifying life universe.
4	two detectives, a rookie and a veteran, hunt a serial killer who uses the seven deadly sins as his motives.	two rookies hunt serial killer deadly motives.
5	little jack is a young fox living happily with his family in the woods, but everything changes when his father is captured by a circus troupe in order to be part of their show. the rest of the animals from the wood will share the same destiny. with the help from a nature-loving disabled boy in a wheelchair and a young acrobat girl, little jack will try to rescue the animals.	little young fox living everything changes captured circus troupe show. rest animals wood disabled boy wheelchair young acrobat girl little rescue animals.

2.2.1. Softmax Regresyonu

Softmax Regresyonu (SR), çok sınıflı (multi-class) sınıflandırma problemlerinde kullanılan bir doğrusal modeldir [25]. Temel amacı, bir veri örneğinin k adet sınıftan hangisine ait olduğunu olasılıksal olarak tahmin etmektir. Denklem 1'de SR'nin olasılık tahmini gösterilmiştir.

$$P(y = k|x) = \frac{exp(w_k^T x + b_k)}{\sum_{j=1}^{k} exp\left(w_j^T x + b_j\right)} \quad (1)$$

Denklem 1 ile x örnek verisinin k sınıfına ait olma olasılığı hesaplanırken;

- $x \in \mathbb{R}^d - d$ boyutlu bir özellik vektörünü,
- k sınıf sayısını,
- $w_k \in \mathbb{R}^d$, her sınıf için tanımlanan ağırlık vektörünü,
- b_k ise her sınıf için tanımlanan sabiti ifade etmektedir.

2.2.2. Rastgele Orman Sınıflandırıcı

Rastgele Orman Sınıflandırıcı (RF), çok sayıda karar ağacının bir araya gelmesiyle oluşturulan bir topluluk (ensemble) öğrenme yöntemidir. Toplulukta yer alan her bir karar ağacı, eğitim verisetinden bootstrap yöntemiyle rastgele seçilen örneklerle eğitilmektedir. Düğüm bölünmeleri sırasında ise rastgele alt özellik setleri kullanılmaktadır. Bu çift yönlü rastgelelik mekanizması, ağaçlar arasındaki korelasyonu azaltarak modelin genellenebilirliğini artırmakta ve aşırı öğrenme riskini minimize etmektedir. Sınıflandırma sürecinde, her ağaç ayrı ayrı tahmin yapmakta ve RF, tüm ağaçların verdiği sınıf tahminleri arasından oy çokluğu esasına göre nihai sınıf kararını vermektedir. Bu yöntem hem yüksek doğruluk sağlamakta hem de farklı veri tipleriyle uyumlu çalışabilmeye imkân sunmaktadır [26]. Denklem 2'de RF yöntemine ait eşitlik verilmiştir.

$$Sınıf = \arg max_{k \in \{1,\ldots,K\}} \sum_{m=1}^{M} 1\{h_m(x) = k\} \quad (2)$$

Burada, M karar ağacı sayısını, x sınıflandırılacak örnek vektörünü, $h_m(x)$, x örneğinin m. sınıfa ait olma olasılığını hesaplayan fonksiyondur. $1\{.\}$, x örneği ilgili sınıfa aitse 1 değilse 0 değeri döndüren yapıyı ifade etmektedir. $\arg max$e her bir örneğin en çok hangi sınıfa dahil edildi ise sınıflandırma sonucunu seçmek için kullanılmaktadır.

2.2.3. Destek Vektör Makineleri

Destek Vektör Makineleri (SVM), denetimli öğrenme kapsamında sınıflandırma ve regresyon problemlerinde kullanılan güçlü ve popüler bir algoritmadır. SVM'nin temel amacı, farklı sınıflara ait veri noktalarını en geniş marj ile ayıran doğrusal veya doğrusal olmayan bir sınır (hiperdüzlem) bulmaktır. Doğrusal ayrılabilir verilerde, SVM en geniş marjı sağlayarak sınıflar arasındaki ayırıcı sınırı maksimize eden bir hiper düzlem bulmaktadır. Bu hiper düzlem, destek vektörleri olarak adlandırılan, karar sınırına en yakın eğitim örnekleri tarafından belirlenmektedir. Maksimum marj, modelin genel performansını ve genelleme yeteneğini arttırmaktadır. Doğrusal olarak ayrılmayan durumlarda ise SVM, veriyi çekirdek (kernel) fonksiyonları aracılığıyla yüksek boyutlu bir uzaya dönüştürür ve böylece karmaşık sınıflandırma sınırları oluşturabilir özellik kazanmış olmaktadır. SVM, ayrıca sınıf dengesizliği ve gürültüye karşı dayanıklılık sağlamak için ceza parametresi (C) ve yumuşak marj kavramları kullanılmaktadır. Ceza parametresi ile bazı eğitim hatalarına izin verilerek aşırı öğrenme önlenmektedir [27]. SVM yöntemi Denklem 3 ile verilen optimizasyon problemini, Denklem 4'te verilen kısıtlar altında çözerek sınıflandırma işlemini gerçekleştirmektedir.

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{N} \xi_i \quad (3)$$

$$y_i(w^T \phi(x_i) + b) \geq 1\text{-}\ \xi_i\ , \quad \xi_i \geq 0, i = 1, \ldots, \mathrm{N} \quad (4)$$

Burada, $\frac{1}{2}\|w\|^2$ hiper düzlemin normunun karesini ifade etmektedir. $\mathrm{C}\sum_{i=1}^{N} \xi_i$ ise sınıflandırma hatalarını cezalandırmak için kullanılır. N verisetindeki örnek sayısını, y, i. örneğin sınıf etiketini temsil etmektedir. $\phi(.)$ veriyi daha yüksek boyutlu uzaya taşıyan kernel fonksiyonunu, ξ_i slack değişkenleri ise marj belirlemek için kullanılmaktadır. b terimi ise sabit terimdir.

2.2.4. Çok Terimli Naive Bayes Sınıflandırıcı

Çok Terimli Naive Bayes (MNB), özellikle metin sınıflandırma, doküman sınıflandırma ve spam tespiti gibi NLP problemlerinde yaygın olarak kullanılan, olasılıksal bir makine öğrenme algoritmasıdır. Naive Bayes modelleri, Bayes Teoremi'ne dayanmakta ve her özelliğin (kelimenin) sınıf etiketinden bağımsız olarak diğer kelimelerle olan ilişkisinin göz ardı edilmesi varsayımı ile çalışmaktadır. MNB'de ise, özellikle özelliklerin (örneğin kelime frekanslarının) sayısal ve sıklık temelli (discrete count-based) olduğu durumlar hedeflenmektedir. Bu Model, her sınıf için belirli bir belge içinde her kelimenin görünme olasılıklarını öğrenmektedir. Eğitim aşamasında, her sınıfa

ait belgelerdeki kelimelerin frekansları kullanılarak olasılıklar hesaplanıp sınıflandırma sırasında, test dokümanındaki kelimelere karşılık gelen bu olasılıklar kullanılarak her sınıf için bir olasılık puanı hesaplanır ve en yüksek olasılığa sahip sınıfa atanmaktadır [28]. Bayes teoremi Denklem 5 ile verilmiştir.

$$P(C_k|x) = \frac{P(C_k).\ P(x|C_k)}{P(x)} \quad (5)$$

Denklem 5'te: k sınıf sayısını, $(C_k|x)$, x örneğinin C_k sınıfına ait olma olasılığını, $P(C_k)$, C_k'nın ön olasılığını, $P(x|C_k)$ ise x girdisinin C_k'ya ait olma olasılığını ifade etmektedir.

Çok terimli modelde, özellik vektörü x=$(x_1, x_2, \dots, x_n)$ belge içindeki kelime frekanslarını temsil eder. Bu durumda, koşullu olasılık Denklem 6 ve 7 ile hesaplanmaktadır:

$$P(x|C_k) = \prod_{i=1}^{n} P(w_i|C_k)^{x_i} \quad (6)$$

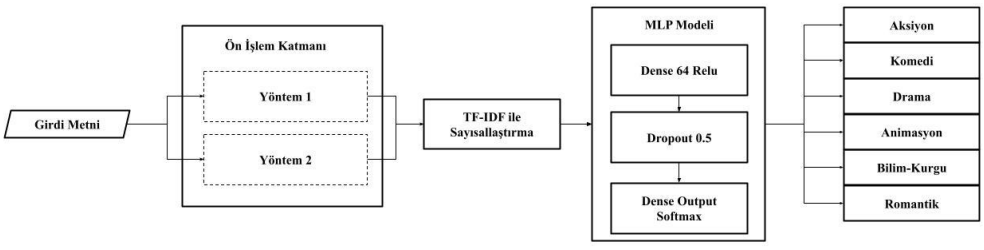
$$P(w_i|C_k) = \frac{N_{ik}+1}{\sum_j N_{jk}+V} \quad (7)$$

Denklem 6 ile $P(w_i|C_k)$ kelime w_i'nin sınıf C_k'da görülme olasılığıdır ve x_i, o kelimenin belgede kaç kez geçtiği bilgisidir. N_{ik}, sınıf C_k'ya ait belgelerde w_i kelimesinin toplam frekansı, V ise toplam kelime sayısıdır.

2.2.5. Çok Katmanlı Algılayıcı

Çok Katmanlı Algılayıcı (MLP) yapay sinir ağlarının en temel ve güçlü yapılarından biridir ve hem sınıflandırma hem de regresyon problemlerinde yaygın olarak kullanılmaktadır. Nitekim bazı çalışmalarda, evrişimsel sinir ağlarının (CNN) kelime sırasına duyarlılığının her zaman avantaj sağlamadığı, özellikle kelime torbası (bag-of-words) temelli TF-IDF gibi yöntemlerle Dense tabanlı modellerin daha stabil ve yorumlanabilir sonuçlar verdiği vurgulanmaktadır [29]. Özellikle sıralı bağımlılıklardan ziyade içerik ağırlıklı temsillere (TF-IDF gibi) dayanan veri kümelerinde, CNN yerine Dense mimarilerin daha uygun olduğu çeşitli çalışmalarla desteklenmiştir [30].

Çalışmada kullanılan MLP, üç katmandan oluşmaktadır bunlar; girdi katmanı, 64 nöronlu bir gizli katman ve çıkış katmanıdır. Girdi katmanı verisetindeki verilerin uygulandığı katmandır. Gizli katmanlar, öğrenme sürecinde girdi verisinin dönüştürülmesini ve karmaşık ilişkilerin modellenmesini sağlamaktadır. Çıkış katmanı ise girdi verisine karşılık sınıf etiketini üretmektedir. Her katmandaki nöronlar, bir önceki katmandaki tüm nöronlara bağlıdır ve bu bağlantılar üzerinde güncellenen ağırlıklar yer almaktadır. Her nöron, girişlerinin ağırlıklı toplamını alıp, üzerine bir sapma (bias) değeri ekleyerek sonucu bir aktivasyon fonksiyonuna uygulamaktadır. Aktivasyon fonksiyonları, nöron çıktılarını belirli bir aralığa normalize etme görevi görür. Gizli katmanlarda sıklıkla ReLU (Rectified Linear Unit), çıkış katmanında ise çok sınıflı sınıflandırma problemleri için softmax fonksiyonu tercih edilmektedir. Bu çalışmada da bu aktivasyon fonksiyonları tercih edilmiştir. Softmax fonksiyonu her sınıf için bir olasılık üretmekte ve bu olasılıkların toplamı 1 olmaktadır. Modelin eğitimi, geri yayılım algoritması ve gradyan inişi yöntemiyle gerçekleştirilmektedir. Bu süreçte model, tahminleri ile gerçek değerler arasındaki farkı ölçen bir kayıp fonksiyonunu minimize etmeye çalışmaktadır. Bu çalışmada, çoklu sınıflandırma problemi için uygun olan categorical-crossentropy kayıp fonksiyonu kullanılmıştır. Optimizasyon sürecinde learning_rate değeri 0,0001 olarak belirlenmiştir. Eğitim süresince kullanılan maksimum epok sayısı 100 olarak tanımlanmış ve EarlyStopping özelliği aktif edilerek eğitim doğrulama kaybında üç ardışık epok boyunca iyileşme gözlenmezse eğitim sonlandırılmaktadır. Buna ek olarak dropout katmanı eklenerek modelin hem genelleme yeteneği artırılmış hem de aşırı öğrenmesinin önüne geçilmiştir. Eğitimde kullanılan batch boyutu 128 olarak seçilmiş, bu sayede eğitim süreci bellek açısından dengelenmiş ve öğrenme davranışı istikrarlı hale getirilmiştir. MLP, özellikle doğrusal olmayan ve karmaşık ilişkileri öğrenebilmesi sayesinde, büyük veri setlerinde etkili sonuçlar verebilen güçlü bir modeldir [31]. Çalışmamızda kullanılan çok katmanlı algılayıcı modeline ait akış diyagramı Şekil 8'de verilmiştir.



Şekil 8. MLP Akış Diyagramı

2.3. FastText Yöntemi

Bu çalışmada kullanılan FastText yöntemi, Facebook AI Research (FAIR) tarafından geliştirilen ve önceden eğitilmiş bir kelime gömme modelidir [32]. Kullandığımız model İngilizce diline özgü kelime vektörleri üretmek amacıyla eğitilmiştir. Bu model, Common Crawl veri kümesi üzerinde yaklaşık 600 milyar kelime içeren çok büyük bir metin koleksiyonu kullanılarak eğitilmiştir. Eğitim verisi çok alanlı ve ham metinlerden oluşmakta olup, veri

ön işleme adımları gerçekleştirilmiş ve yüksek çeşitlilikte içerikler barındırmaktadır. FastText modeli, her kelimeyi 300 boyutlu bir vektör ile temsil eder. Bu vektörler, kelimeler arası semantik benzerlikleri sayısal düzeyde ifade edebilmekte olup, örneğin "king"-"man"+"woman" matematiksel işlemi sonucunun "queen" vektörüne yakınsaması gibi anlamsal çıkarımlar yapabilmektedir. Modelin temel avantajlarından biri, karakter n-gramlarına dayalı alt-kelime temsilleri kullanmasıdır. Bu sayede model, eğitim sırasında doğrudan karşılaşmadığı yeni ya da nadir kelimeler için bile benzer alt parçalara (örneğin "com", "put", "tion") dayanarak vektör tahmini yapabilmektedir. Bu özelliği ile doğal dil işleme uygulamalarında karşılaşılan sözlük dışı kelimeler sorununu önemli ölçüde azaltmaktadır. FastText modeli bağlamdan bağımsızdır; bu da her kelimenin yalnızca tek bir sabit vektöre sahip olduğu anlamına gelmektedir. Dolayısıyla çok anlamlı kelimeler (örneğin "bank") tüm bağlamlarda aynı vektörle temsil edilir. Bu durumun sınırlayıcı etkisine rağmen model, hafif yapısı ve CPU üzerinde dahi yüksek işlem verimliliği ile öne çıkmaktadır. Kullanımı oldukça pratik olan FastText modeli, büyük veri kümesi üzerinde eğitilmiş, hızlı, esnek ve geniş kapsama sahip bir kelime gömme yaklaşımı olarak bu çalışmada kullanılmıştır.

2.4. Değerlendirme Metrikleri

Sınıflandırma modellerinin performansını kapsamlı bir şekilde değerlendirmek amacıyla, öncelikle karmaşıklık matrisi kullanılmıştır. Karmaşıklık matrisi, modelin tahmin ettiği sınıflar ile gerçek sınıflar arasındaki ilişkiyi sayısal olarak gösteren $n \times n$ boyutunda bir matristir; burada n sınıf sayısını temsil etmektedir. Matrisin ana diagonalindeki elemanlar doğru sınıflandırılan örneklerin sayısını (TP_i, doğru pozitifler) gösterirken, diğer hücreler yanlış sınıflandırmaları (yanlış pozitifler, yanlış negatifler, doğru negatifler) ifade etmektedir. Karmaşıklık matrisine dayanılarak, sınıf bazında aşağıdaki değerlendirme metrikleri hesaplanmıştır:

- **Kesinlik**, bir sınıf için modelin pozitif olarak tahmin ettiği örnekler içindeki doğru pozitiflerin oranını ifade etmekte ve Denklem 8 ile tanımlanmaktadır.

$$Kesinlik_i = \frac{TP_i}{TP_i + FP_i} \quad (8)$$

Burada TP_i doğru pozitif, FP_i ise yanlış pozitif sayısını belirtmektedir.

- **Duyarlılık veya Hassasiyet**, gerçek pozitif örneklerin model tarafından doğru tespit edilme oranıdır ve Denklem 9 ile formüle edilmiştir.

$$Duyarlılık_i = \frac{TP_i}{TP_i + FN_i} \quad (9)$$

Burada FN_i yanlış negatif sayısını göstermektedir.

- ***F*1-Score**, kesinlik ve duyarlılık değerlerinin harmonik ortalaması olup, her iki metriğin dengeli şekilde göz önüne alınmasını sağlar ve formülü Denklem 10 ile verilmiştir.

$$F1_i = 2x \frac{Kesinlik_i \, x \, Duyarlılık_i}{Kesinlik_i + Duyarlılık_i} \quad (10)$$

$F1$ metriği özellikle dengesiz sınıf dağılımı bulunan veri setlerinde modelin performansını daha sağlıklı değerlendirmede kullanılmaktadır.

Bunların yanı sıra, Alıcı İşletim Karakteristiği Eğrisi (Receiver Operating Characteristic - ROC) ve bu eğrinin altında kalan alan (Area Under Curve - AUC) metrikleri de değerlendirmeye dahil edilmiştir. ROC eğrisi, sınıflandırma modelinin farklı eşik değerlerindeki doğru pozitif oranı (True Positive Rate - TPR) ile yanlış pozitif oranı (False Positive Rate - FPR) arasındaki ilişkiyi grafiksel olarak göstermekte ve Denklem 11 ile verilmiştir.

$$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{FP+TN} \quad (11)$$

ROC eğrisi altında kalan alan (AUC), modelin genel ayırt edici gücünü tek bir değer olarak özetler ve 0.5 (rasgele tahmin) ile 1 (mükemmel ayırt etme) arasında değişir. Yüksek AUC değeri, modelin sınıfları ayırt etme başarısının yüksek olduğunu gösterir.

Sonuç olarak, karmaşıklık matrisi tabanlı kesinlik, duyarlılık, $F1$-skoru ve destek metrikleri ile $ROC - AUC$ analizi birlikte kullanılarak model performansı çok boyutlu ve güvenilir bir şekilde değerlendirilmiştir. Bu sayede, modelin hem genel doğruluğu hem de dengesiz sınıf dağılımına karşı dayanıklılığı ölçülmüş ve yorumlanmıştır.

3. Deneysel Sonuçlar

Bu çalışmada IMDb sitesinden alınan film açıklamaları kullanılarak, film türlerinin altı farklı türe sınıflandırılması gerçekleştirilmiştir. Toplamda 7350 film açıklamasından oluşan verisetimiz %80-%20 eğitim-test verisi olacak şekilde rastgele ayrılmıştır.

Beş farklı sınıflandırıcı kullanılarak yapılan sınıflandırma başarısının iyileştirilmesi için FastText yönteminden oluşan bir ön işleme katmanı kullanılmıştır. Yöntem 1 ve Yöntem 2 olarak isimlendirdiğimiz modellerimizden Yöntem 1, temel metin ön işlem adımları ile yapılan sınıflandırmayı, Yöntem 2 ise FastText katmanının eklendiği ön işlem adımları ile yapılan sınıflandırmayı ifade etmektedir. Bu bölümde ilk olarak FastText modelinin neden seçildiğinin anlaşılması için diğer kelime gömme modelleri ile bir karşılaştırma yapılmıştır. Daha sonra film açıklamalarında yer alan kelime ile tür adı vektörleri arasındaki kosinüs benzerliği için belirlenen 0,1 eşiğinin doğruluğu incelenmiştir. Son olarak tüm yöntemlerin sınıflandırma sonuçları detaylı olarak analiz edilmiştir.

3.1. Farklı Kelime Gömme Modellerinin Karşılaştırılması

Bu bölümde, FastText modeline ek olarak Word2Vec, GloVe ve BERT modelleri kullanılarak elde edilen kelime vektörleri ile sınıflandırma işlemleri gerçekleştirilmiştir. Word2Vec ve GloVe modelleri FastText modeli gibi bağlamdan bağımsız oldukları için karşılaştırma işlemine dahil edilmiştir. Buna ek olarak film açıklamalarındaki anlamsal ve bağlamsal bilgilerinde dikkate alınarak sınıflandırma başarısına etkisini inceleyebilmek için BERT modeli kullanılmıştır. Böylece hem daha kapsamlı hem de daha objektif bir karşılaştırma yapılabilmiştir. Tablo 4'de elde edilen sınıflandırma sonuçlarına ait doğruluk değerleri verilmiştir.

Tablo 4. Farklı Kelime Gömme Modellerinin Sınıflandırma Başarı Doğrulukları

Model	Modellerin Doğruluk Değerleri				
	SR	**RF**	**SVM**	**MNB**	**MLP**
Word2Vec	0,8565	0,8152	0,8674	0,8668	0,8674
BERT	0,5397	0,4657	0,5207	0,5635	0,5696
GloVe	0,7489	0,6088	0,7697	0,7481	0,7735
FastText	0,8907	0,8540	0,8975	0,8975	0,9002

Tablo 4'deki karşılaştırmalar yapılırken tüm modellerde kosinüs benzerliği eşik değeri 0,1 alınarak sınıflandırmalar gerçekleştirilmiştir. Tablodaki verilerden FastText modelinin en iyi sonuçlar verdiği görülmektedir. Bu çalışmada tercih edilmesinin doğru bir karar olduğu bu analizler ile ortaya koyulmuştur. Her ne kadar bağlamdan bağımsız olarak çalışan kelime gömme yöntemleri yakın sonuçlar vermiş olsa da FastText modeli en iyi sonuçları elde etmiştir. Ayrıca BERT modeli anlamsal ve bağlamsal ilişkileri dikkate almasına rağmen diğer üç modelin gerisinde kalmıştır. Film açıklamalarına göre yapılan sınıflandırma işlemlerinde kelimelerin anlamsal ve bağlamsal ilişkilerine bakmaya gerek olmadığı yapılan analizlerden anlaşılmaktadır. Şekil 9'da gösterilen türlere göre kelime dağılım kümeleri de bunu doğrulamaktadır.

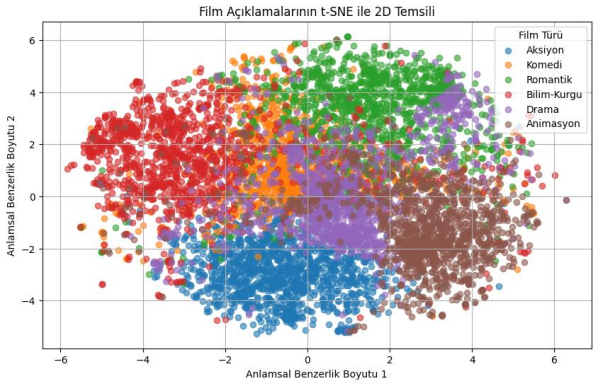
3.2. FastText Modelinde Kosinüs Eşiği Parametresinin Analizi

Film açıklamalarındaki her kelimenin vektörü ile bu filmin ait olduğu tür adı vektörü arasındaki kosinüs benzerliği 0,1 eşik değerinin altında olan kelimelerin açıklamalardan çıkarıldığı önceki bölümlerde ifade edilmişti. Bu bölümde seçilen bu eşik değerinin doğruluğu, farklı eşik değerleri için analizler yapılması ile teyit edilmiştir. Elde edilen sonuçlar Tablo 5'de verilmiştir. Dört farklı eşik değeri için FastText yöntemi kullanılarak elde edilen sonuçlar incelendiğinde, 0,1 değerinin en iyi tercih olduğu anlaşılmaktadır.

Tablo 5. Farklı Eşik Değerleri için Yapılan Analizler

Model	Eşik Değerleri			
	0,05	**0,1**	**0,15**	**0,2**
SR	0,7855	0,8907	0,7985	0,7701
RF	0,6965	0,8540	0,7327	0,7076
SVM	0,7848	0,8975	0,7922	0,7647
MNB	0,7814	0,8975	0,7891	0,7615
MLP	0,7875	0,9002	0,8089	0,7889

Bu bulgular, TF-IDF temsili aracılığıyla her film türüne özgü dilsel ve bağlamsal özelliklerin etkin biçimde ayrıştırılabildiğini göstermektedir. Ayrıca, yüksek boyutlu TF-IDF vektörlerinin t-SNE (t-Distributed Stochastic Neighbor Embedding) yöntemiyle iki boyuta indirgenerek görselleştirilmesi sonucunda, farklı film türlerinin anlamlı kümeler oluşturduğu ve birbirlerinden açık biçimde ayrıştığı gözlemlenmiştir. Türler arası bu ayrışma, dilsel farkların hem istatistiksel hem de görsel düzlemde başarıyla yakalanabildiğine işaret etmektedir. Söz konusu görsel, Şekil 9'da verilmiştir.



Şekil 9. Film Türlerine Göre TF-IDF Temsili Üzerinden 2D t-SNE Kümelenmesi

Şekil 9'dan görüldüğü üzere, FastText katmanının eklenmesi ile düzenlenen açıklamaların, ait oldukları türlerin karakteristik temalarını daha belirgin bir şekilde yansıttığı çıkarımı yapılabilmektedir. Bu durum, her tür için oluşturulan kelime bulutlarıyla da desteklenmektedir. Kelime bulutları incelendiğinde, aksiyon türünde "battle", "weapon", "fight", "soldier" gibi savaş ve çatışma odaklı terimlerin ön plana çıktığı; komedi türünde "funny", "laugh", "joke" gibi mizahi ifadelerin ağırlık kazandığı gözlemlenmiştir. Romantik tür açıklamalarında "love", "relationship", "kiss" gibi duygusal içerikli sözcükler öne çıkarken, bilim kurgu türünde "spaceship", "alien", "technology" gibi uzay ve teknoloji temalı kelimelerin belirginleştiği görülmüştür. Drama türünde "family", "struggle", "loss" gibi duygusal derinlik taşıyan ifadelerin öne çıktığı; animasyon türünde ise "animated", "cartoon", "magic" gibi hayal gücüne ve çocuklara yönelik içerikleri yansıtan terimlerin baskın olduğu belirlenmiştir.

3.3. Kullanılan Yöntemlerin Karşılaştırılması

İlk olarak Yöntem 1 ile sınıflandırma işlemleri gerçekleştirilmiştir. Yapılan sınıflandırma işlemine ait sonuçlar Tablo 6'da verilmiştir. Tablo 6'da verilen veriler tüm türlerin ortalama değerleridir. Elde edilen $F1$ skoru, kesinlik, duyarlılık ve doğruluk değerleri incelendiğinde, genel olarak sınıflandırma performanslarının düşük düzeyde kaldığı gözlemlenmiştir. Özellikle "Drama" ve "Komedi" türlerinde tüm modellerde belirgin performans düşüklüğü dikkat çekmektedir. Örneğin, "Drama" türü için MNB modeli yalnızca 0,3176 duyarlılık değeri üretmiştir. Ayrıca, RF ve SVM gibi güçlü sınıflandırıcılar dahi bu türlerde 0,4'ün altında $F1$ skorları ile sınıflandırma yapmışlardır. Görece daha ayrık dilsel özellikler taşıyan "Bilim Kurgu" ve "Animasyon" türlerinde sonuçlar bir miktar daha yüksek olsa da genel performansın sınırlı kaldığı görülmüştür. Bu durum Tablo 1'de verilen veriseti istatistiklerinden görüldüğü üzere, bu türlerdeki açıklamaların kısa olması ve standart sapmalarının düşük olmasından kaynaklandığı ön görülmektedir. Sonuçların düşük çıktığı türlerde, diğer türlere nispeten türü doğrudan çağrıştıracak güçlü ve ayırt edici anahtar kelimelerin daha az olması sınıflandırma başarısını düşürmüştür.

Genel olarak modellerin performansları incelendiğinde, MNB modelinin en başarılı sonuçları verdiği görülmektedir. Özellikle Bilim Kurgu türünde 0,6835'lik $F1$ Skoru ile MNB, Animasyon ve Romantik gibi diğer türlerde de başarılı bir performans sergilemiştir. MLP'de genel olarak dengeli bir performans göstermiş ve Bilim Kurgu ile Animasyon türlerinde MNB'ye yakın sonuçlar elde etmiştir. Bu iki model, Yöntem 1 için en iyi sınıflandırıcı olmuşlardır. Her ne kadar MNB ve MLP yöntemleri Yöntem 1 için en iyi sınıflandırıcılar olsa da doğruluk sonuçları sırasıyla 0,5635 ve 0,5513 seviyelerinde kalarak çok düşük seviyededir. Bu sonuçlara göre kullanılan sınıflandırıcıların başarılarının artırılması gerekmektedir.

Tablo 6. Yöntem 1 kullanılarak TF-IDF Vektörleri ile Yapılan Sınıflandırma Sonuçlarına Ait Ortalama Metrikler

Model	$F1$ Skoru	Kesinlik	Duyarlılık	Doğruluk
SR	0,5408	0,5572	0,5421	0,5397
RF	0,4675	0,4716	0,4716	0,4657
SVM	0,5209	0,5206	0,5231	0,5207
MNB	0,5585	0,5646	0,5677	0,5635
MLP	0,5513	0,5522	0,5532	0,5513

Bu aşamada verisetine uygulanan ön işlem adımlarına bir ek katman uygulanmıştır. Yöntem 2 olarak isimlendirilen yöntemde FastText ön işlem katmanı kullanılmıştır. Anlamsal filtreleme (FastText tabanlı) yapılmadan gerçekleştirilen ön işlemlerin, sınıflandırma başarısını olumsuz etkilediği ve bağlamsal bilgi

eksikliğinin özellikle semantik olarak örtüşen türlerde önemli performans kayıplarına neden olduğu Yöntem 1'in sonuçlarından anlaşılmaktadır.

Yöntem 2 ile ön işleme sürecine, film açıklamaları ile ait oldukları türler arasındaki anlamsal ilişkilere dayalı bir filtreleme katmanı eklenerek aynı sınıflandırıcılarla tekrar modelleme işlemleri gerçekleştirilmiştir. Yöntem 2'ye ait sınıflandırma sonuçları tüm türlerin ortalama verileri olacak şekilde Tablo 7'de verilmiştir. Tablo 8'de ise en iyi sonuçların elde edildiği MLP yönteminin her tür için elde edilen metrikleri verilmiştir.

Tablo 7. Yöntem 2 kullanılarak TF-IDF Vektörleri ile Yapılan Sınıflandırma Sonuçlarına Ait Ortalama Metrikler

Model	*F*1 Skoru	Kesinlik	Duyarlılık	Doğruluk
SR	0,8919	0,8968	0,8913	0,8907
RF	0,8919	0,8612	0,8553	0,8540
SVM	0,8988	0,9011	0,8981	0,8975
MNB	0,8978	0,8991	0,8995	0,8975
MLP	0,8999	0,9026	0,9013	0,9002

Tablo 8. Yöntem 2 kullanılarak TF-IDF Vektörleri ile Yapılan MLP Sınıflandırıcısına Ait Metrikler

Model	Tür	*F*1 Skoru	Kesinlik	Duyarlılık	Doğruluk
MLP	Aksiyon	0,9602	0,9644	0,9559	0,9002
	Komedi	0,8609	0,8375	0,8855	
	Romantik	0,8773	0,8719	0,8828	
	Bilim Kurgu	0,8980	0,8835	0,9129	
	Drama	0,8916	0,9316	0,8549	
	Animasyon	0,9112	0,9268	0,9157	

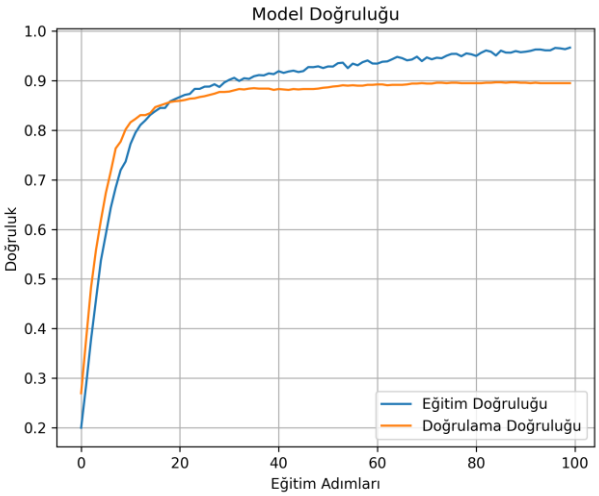
Yapılan sınıflandırma sonuçlarına göre, FastText katmanının eklenmesinin, sınıflandırma başarısı açısından belirleyici bir etkiye sahip olduğu Tablo 7'den görülmektedir. Yöntem 2 ile, her bir film açıklaması, yalnızca ait olduğu film türüyle yüksek TF-IDF değerine sahip kelimeler dikkate alınarak yeniden yapılandırılmıştır. Elde edilen sınıflandırma başarılarında ciddi bir iyileşme söz konusudur.

Yöntem 2, türler arasında farklı sınıflandırma başarıları elde edebilmiştir. *F*1 skorları incelendiğinde, genel olarak tüm sınıflayıcılar yüksek doğruluk seviyelerine ulaşmış olsa da en dikkat çekici sonuçlar MLP ve SVM tarafından elde edilmiştir. Özellikle aksiyon ve bilim kurgu türlerinde MLP modeli, sırasıyla 0,9602 ve 0,8980 gibi yüksek *F*1 skorları ile üstün performans sergilemiştir. Komedi ve romantik türlerde ise görece daha düşük ancak hâlâ iyi düzeyde sonuçlar elde edilmiştir. Bu durum, bu türlerin içerik bakımından daha fazla çeşitlilik ve bağlam karmaşıklığı içermesinden kaynaklanmaktadır.

Kesinlik değerleri dikkate alındığında, SR ve MLP modelleri, özellikle aksiyon ve drama türlerinde yüksek doğruluk sağlamıştır. Örneğin SR, aksiyon türü için 0,9854 gibi oldukça yüksek bir kesinlik değerine ulaşmıştır. Bu durum, modelin özellikle aksiyon türüne ait örnekleri yüksek isabetle tanıyabildiğini göstermektedir. Benzer şekilde, MLP modeli drama (0,9316) ve animasyon (0,9268) türlerinde de güçlü bir doğruluk sergilemiştir.

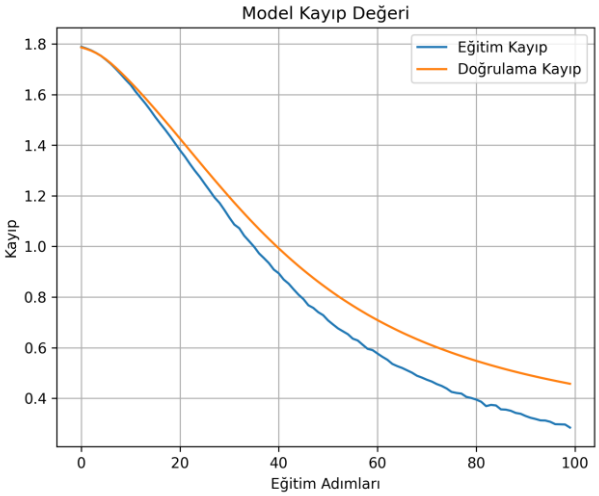
Duyarlılık değerlerine bakıldığında ise MNB sınıflayıcısının aksiyon (0,9824) ve animasyon (0,9558) türlerinde öne çıktığı görülmektedir. Bu sonuç, modelin özellikle bu türlere ait örnekleri kaçırmadan doğru şekilde etiketleyebildiğine işaret etmektedir. Bununla birlikte, RF sınıflayıcısı diğer modellere kıyasla çoğu türde daha düşük duyarlılık ve kesinlik değerleriyle performans açısından geride kalmıştır. Fakat her durumda FastText katmanı olmadan yapılan temel ön işlemlere göre yüksek oranda sonuçlarda iyileştirme mevcuttur.

Genel olarak değerlendirildiğinde, FastText ile zenginleştirilmiş açıklamalardan elde edilen TF-IDF vektörlerinin, film türlerinin sınıflandırılmasında etkili bir özellik çıkarımı sağladığı görülmektedir. Söz konusu vektörler ile eğitilen sınıflayıcılar, özellikle belirgin tematik yapılara sahip türlerde (örneğin aksiyon, animasyon ve bilim kurgu) yüksek başarı göstermiştir. Bu bulgular, türler arası dilsel farklılıkların yalnızca görsel olarak değil, aynı zamanda nicel performans ölçütleriyle de doğrulanabildiğini ortaya koymaktadır. Kullanılan sınıflandırıcılar arasında en iyi performans MLP ile elde edilmiştir. Şekil 10 ve Şekil 11'de sırasıyla MLP'nin eğitim sürecinde elde edilen doğruluk ve kayıp değerleri gösterilmiştir.



Şekil 10. MLP Model Doğruluğu

Şekil 10'da, yaklaşık 20. epok'tan sonra doğrulama doğruluğunun %90 civarında sabitlendiği görülmektedir. Eğitim doğruluğu ise %98 civarında bir doğruluğa ulaşarak sabitlenmiştir.



Şekil 11. MLP Model Kayıp Değeri

Şekil 11'de ise hem eğitim hem de doğrulama kayıplarının sürekli olarak azaldığı ve aralarındaki farkın düşük kaldığı gözlemlenmektedir. Bu sonuçlar, modelin hem eğitim hem de doğrulama verileri üzerinde tutarlı bir şekilde öğrendiğini ve aşırı öğrenmenin olmadığını göstermektedir.

FastText ile ön işlenmiş verilerin TF-IDF temsili kullanılarak eğitilen modellerin çok sınıflı sınıflandırma performansları, ROC AUC skorları üzerinden de değerlendirilip Tablo 9 ile sonuçlar verilmiştir. Bu kapsamda, her bir film türüne ait sınıf bazlı ROC AUC değerlerinin yanı sıra mikro ve makro ortalamalar da hesaplanmıştır. Elde edilen sonuçlar, sınıflayıcı modellerin yüksek ayrıştırma kabiliyetine sahip olduğunu ortaya koymuştur.

Sınıf bazında incelendiğinde, MLP ve MNB modelleri neredeyse tüm türler için 0,98 üzeri AUC skorları ile oldukça yüksek ayırt edicilik göstermiştir. Özellikle MNB modeli, animasyon türü için 0,9955 gibi dikkat çekici bir AUC değeri üretmiş ve genel olarak en yüksek makro ortalama AUC (0,9907) performansını sergilemiştir. Bu sonuç, modelin tüm sınıflar arasında ortalama olarak dengeli bir ayırt edicilik sağladığını göstermektedir.

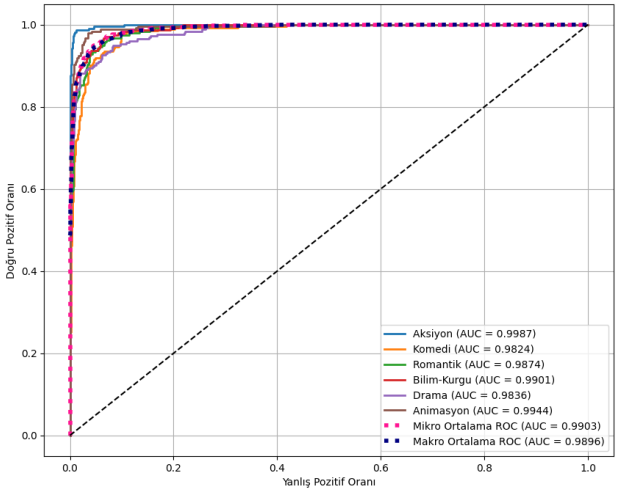
Tablo 9. Model Bazında Çok Sınıflı ROC AUC Sonuçları (Sınıf Bazlı, Makro ve Mikro Ortalama)

Tür	SR	RF	SVM	MNB	MLP
Aksiyon	0,9976	0,9938	0,997	0,9986	0,9985
Komedi	0,9862	0,9792	0,9819	0,9838	0,9834
Romantik	0,9886	0,9805	0,9859	0,9883	0,9881
Bilim-Kurgu	0,9925	0,9819	0,9902	0,9908	0,9899
Drama	0,9823	0,9662	0,9703	0,9866	0,9833
Animasyon	0,9908	0,9816	0,9928	0,9955	0,9949
Mikro Ortalama AUC	0,9892	0,9804	0,987	0,9877	0,9905
Makro Ortalama AUC	0,9898	0,9806	0,9866	0,9907	0,9898

MLP modeli ise özellikle aksiyon (0,9985) ve animasyon (0,9949) türlerinde üst düzey ayrım gücü göstermiştir. Bu yüksek performans, modelin bu türlere ait örnekleri diğerlerinden başarıyla ayırt edebildiğini göstermektedir. MLP'nin mikro ortalama AUC değeri 0,9905 ile tüm örnekler genelinde yüksek doğrulukta bir sınıflandırma yeteneğine sahip olduğunu da teyit etmektedir.

SR, SVM ve RF modelleri de benzer şekilde güçlü AUC performansları sergilemiştir. Özellikle SR ve SVM modelleri, aksiyon, bilim kurgu ve animasyon gibi belirgin temalara sahip türlerde 0,99'a yakın veya üzerinde AUC skorları elde etmiştir. Bu modellerin de mikro ve makro ortalama skorları 0,98'in üzerinde kalarak tutarlı ve güvenilir sınıflandırma yaptıkları sonucuna varılmıştır.

Sonuç olarak, ROC AUC analizleri, FastText destekli TF-IDF temsillerinin türler arasında güçlü bir ayrım yapılmasını sağladığını ve sınıflayıcı modellerin yüksek doğrulukla çalıştığını ortaya koymuştur. Özellikle MNB ve MLP modelleri hem sınıf bazında hem de genel ortalamalarda öne çıkan performanslarıyla dikkat çekmektedir. Bu bulgular, çalışmada geliştirilen metin temelli sınıflandırma yaklaşımının sadece doğruluk ve duyarlılık gibi temel metriklerle değil, aynı zamanda ROC AUC gibi ayrıştırma gücünü ölçen daha üst düzey göstergelerle de başarılı olduğunu kanıtlamaktadır. Bu kapsamlı sonuçlar Tablo 9'da sunulmuş olup, MLP modeline ait ROC eğrisi hem sınıf bazlı hem de makro ve mikro ortalamaları içerecek şekilde hazırlanıp Şekil 12 ile verilmiştir.



Şekil 12. MLP Modeli ROC Grafiği

Bu çalışmada MLP diğer yöntemlere göre üstünlük sağlamıştır. Bu üstünlük, doğrusal olmayan sınırları öğrenebilme kapasitesi ve yüksek boyutlu, seyrek metin temsillerinden anlamlı özellik çıkarımları yapabilmesiyle ilişkilidir. Çok katmanlı yapısı sayesinde, metin içerisindeki kelimeler arasında bulunan anlamsal ve bağlamsal ilişkileri temsil edebilmektedir. Ayrıca, Dropout katmanı ile aşırı öğrenme riski azaltılabilmekte, softmax çıkış katmanı aracılığıyla da çok sınıflı problemler uygun şekilde çözülebilmektedir. MLP'nin bu yapısal avantajları, kelime bağımlılıklarını göz ardı eden ve basit olasılık modellerine dayanan MNB, yalnızca doğrusal karar sınırlarını öğrenebilen SR, yüksek boyutlu seyrek temsillerde bölünme sorunu yaşayan RF ve büyük veri kümelerinde hesaplama açısından verimsiz hale gelen SVM gibi geleneksel yöntemlere kıyasla daha başarılı sınıflandırma performansı sergilemesini sağlamıştır.

4. Tartışma ve Sonuç

Bu çalışmada, IMDb film açıklamalarının altı farklı türe sınıflandırılması amacıyla farklı kelime gömme teknikleri ve makine öğrenimi sınıflandırıcıları kullanılmıştır. Elde edilen deneysel bulgular, özellikle kelime bağlamları kurulamayacak kadar kısa metinler üzerinde FastText modelinin etkili ve verimli bir temsil sunduğunu ortaya koymuştur. FastText tabanlı ön işleme katmanının, TF-IDF temsilleriyle entegrasyonu sayesinde sınıflandırma başarısında anlamlı bir artış gözlenmiştir. Bu iyileşme, FastText'in alt kelime düzeyinde dilbilgisel yapıları yakaladığı ve türlere özgü anlamsal filtreleme yapması sayesinde elde edilmektedir. Böylece, her bir film türüne ait karakteristik dilsel özellikler belirgin biçimde ayrıştırılabilmiştir.

Buna karşılık, bağlamsal ilişkilere odaklanan ve derin öğrenmeye dayalı olan BERT modeli, kısa ve özet film açıklamalarında beklenen performans seviyesinin altında kalmıştır. Kısa metinlerde bağlamın sınırlı olması, BERT'in güçlü bağlamsal anlayış yeteneğinin kullanımını kısıtlamış ve sonuç olarak bağlamdan bağımsız kelime gömme yöntemlerinin gerisinde kalmasına neden olmuştur. Bu durum, metin sınıflandırma çalışmalarında model seçiminin, veri yapısı ve metin uzunluğu gibi faktörlerle yakından ilişkili olduğunu göstermektedir. Özellikle kısa ve bilgi yoğun metinlerde bağlamdan bağımsız modellerin tercih edilmesinin daha uygun sonuçlar verdiği gözlenmiştir.

Sınıflandırıcılar arasında, MLP modeli, doğrusal ve olasılıksal temelli geleneksel yöntemlere kıyasla üstün bir performans sergilemiştir. MLP'nin çok katmanlı yapısı, yüksek boyutlu ve seyrek metin temsillerindeki karmaşık ilişkileri daha etkili biçimde öğrenmesine olanak sağlamış; bu sayede türler arası ayrım gücü artmıştır. ROC AUC analizleri ise, özellikle MLP ve MNB modellerinin, çok sınıflı sınıflandırmada yüksek ayırt edicilik kapasitesine sahip olduğunu doğrulamıştır.

Tür bazında yapılan değerlendirmelerde, içerik çeşitliliğin az olması ve bağlamsal karmaşıklığın fazla olması modellerin performanslarında kısıtlayıcı bir etken olmuştur. Bu durum drama ve komedi gibi türlerde ortaya çıkmıştır. İçerik çeşitliliğin yeterince yüksek olduğu diğer türlerde ise model performansları daha yüksek elde edilmiştir.

Gelecekte, bu çalışma kapsamında geliştirilen metin sınıflandırma yöntemleri, yalnızca film türü sınıflandırmasında değil, kısa metinlerin yoğun olduğu sosyal medya analizleri, haber sınıflandırması, müşteri geri bildirimlerinin otomatik analizi ve chatbot yanıt sistemleri gibi çeşitli doğal dil işleme uygulamalarında da etkin biçimde kullanılabilecektir. Özellikle sosyal medya içeriklerinde hızlı ve doğru sınıflandırma ihtiyacı dikkate alındığında, FastText destekli ve TF-IDF tabanlı vektör temsili sınıflandırıcıların uygulama alanı oldukça geniş bir yer almaktadır.

Kaynakça

[1] Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. Proceedings of the First Instructional Conference on Machine Learning. https://ciir.cs.umass.edu/pubfiles/ir-88.pdf

[2] Smith, A., Johnson, M., & Wang, Y. (2020). Automated movie genre classification using TF-IDF and machine learning algorithms. Multimedia Tools and Applications, 79(45-46), 33803–33824. https://doi.org/10.1007/s11042-020-09391-7

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

[4] Özdil, U., Arslan, B., Taşar, D. E., Polat, G., & Ozan, Ş. (2021, September). Ad text classification with bidirectional encoder representations. In 2021 6th international conference on computer science and engineering (UBMK) (pp. 169-173). IEEE.

[5] Yilmaz, E. H., & Toraman, C. (2021, June). Intent classification based on deep learning language model in turkish dialog systems. In 2021 29th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[6] Koruyan, K. (2025). Restoran Müşteri Yorumlarının Duygu Analizi: Sıfır-Atış Metin Sınıflandırma Yaklaşımı. Journal of Intelligent Systems: Theory and Applications, 8(1), 47-62.

[7] Veziroğlu, M., & Bucak, İ. (2025). Haber Sınıflandırma Sistemlerinde Naive Bayes ve Makine Öğrenmesi Algoritmaları Arasında Performans Karşılaştırması. Journal of the Institute of Science and Technology, 15(1), 57-70.

[8] Yiğit Açıkgöz, F. A. T. M. A., & Kayakuş, M. E. H. M. E. T. (2025). Süpermarket Zincirlerinin Mobil Uygulamalarının Kurumsal İtibarına Etkisi: Duygu Analizi ve Metin Madenciliği Yöntemleriyle Değerlendirme. SÜLEYMAN DEMIREL ÜNIVERSITESI VIZYONER DERGISI, 16(45).

[9] Akgün, M. K. (2025). Otel firmalarına yapılan yorumların metin madenciliği ve derin öğrenme yöntemleri ile analizi.

[10] Saka, S. O., & Cömert, Z. (2025). Sinir Ağı Dil Modelleri ve Evrensel Cümle Kodlayıcı Kullanarak Havayolu Müşteri Yorumlarının Duygu Analizi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 37(1), 167-182.

[11] Yiğidefe, G., Kaman, S. Ç., & Eken, B. (2025). Classification and Analysis of Employee Feedback with Deep Learning Algorithms. Sakarya University Journal of Computer and Information Sciences, 8(1), 38-46.

[12] Camadan, E., & Şimşek, M. (2025). X Platformunda İstenmeyen Gönderilerin Makine Öğrenmesi, Geniş Dil Modelleri ve Bilgisayarlı Görü ile Tespiti. Bilişim Teknolojileri Dergisi, 18(2), 181-194.

[13] AYDIN, N., ERDEM, O. A., & Tekerek, A. D. E. M. (2025). Comparative Analysis of Traditional Machine Learning and Transformer-based Deep Learning Models for Text Classification. Journal of Polytechnic, 28(2).

[14] Güneş, Y., & Arıkan, M. (2025). X (Twitter) Sentiment Analysis Based on Hybrid Approach: An Application for Online Food Ordering. Bilişim Teknolojileri Dergisi, 18(2), 143-167.

[15] Gao, I., Liu, G., Zhu, B., Zhou, S., Zheng, H., & Liao, X. (2025, February). Multi-level attention and contrastive learning for enhanced text classification with an optimized transformer. In 2025 5th International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 499-503). IEEE.

[16] Zhang, Q., Chen, S., & Liu, W. (2025). Balanced Knowledge Transfer in MTTL-ClinicalBERT: A Symmetrical Multi-Task Learning Framework for Clinical Text Classification. Symmetry, 17(6), 823.

[17] Jamshidi, S., Mohammadi, M., Bagheri, S., Najafabadi, H. E., Rezvanian, A., Gheisari, M., ... & Wu, Z. (2024). Effective text classification using BERT, MTM LSTM, and DT. Data & Knowledge Engineering, 151, 102306.

[18] Lepagnol, P., Gerald, T., Ghannay, S., Servan, C., & Rosset, S. (2024). Small language models are good too: An empirical study of zero-shot classification. arXiv preprint arXiv:2404.11122.

[19] Islam, M. T., Parvin, F., Sazan, S. A., & Amir, T. B. (2024, September). Comparative Analysis of Sentiment Classification on IMDB 50k Movie Reviews: A Study Using CNN, LSTM, CNN-LSTM, and BERT Models. In 2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON) (pp. 512-517). IEEE.

[20] Emanuel, R. H., Docherty, P. D., Lunt, H., & Möller, K. (2024). The effect of activation functions on accuracy, convergence speed, and misclassification confidence in CNN text classification: a comprehensive exploration. The Journal of Supercomputing, 80(1), 292-312.

[21] Cesarini, M., Malandri, L., Pallucchini, F., Seveso, A., & Xing, F. (2024). Explainable ai for text classification: Lessons from a comprehensive evaluation of post hoc methods. Cognitive Computation, 1-19.

[22] Taha, K., Yoo, P. D., Yeun, C., Homouz, D., & Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. Computer Science Review, 54, 100664.

[23] Maragheh, H. K., Gharehchopogh, F. S., Majidzadeh, K., & Sangar, A. B. (2024). A hybrid model based on convolutional neural network and long short-term memory for multi-label text classification. Neural Processing Letters, 56(2), 42

[24] Singh, J., Singh, S., & Sharma, M. (2018). A comprehensive review on text mining: Techniques and applications. International Journal of Information Technology, 10(4), 491–499. https://doi.org/10.1007/s41870-018-0151-2

[25] Sun, Z., Chen, Z., Liu, J., & Yu, Y. (2025). Multi-class feature selection via Sparse Softmax with a discriminative regularization. International Journal of Machine Learning and Cybernetics, 16(1), 159-172.

[26] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[27] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

[28] McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. AAAI-98 Workshop on Learning for Text Categorization, 752(1), 41–48.

[29] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. arXiv preprint arXiv:1607.01759. https://arxiv.org/abs/1607.01759

[30] Wang, S., & Manning, C. D. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 90–94.

[31] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. https://www.deeplearningbook.org/

[32] fasttext.cc, "Word vectors for 157 languages", https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.en.300.bin.gz (Erişim Tarihi: 10.07.2025).