Çok-Etiketli Film Türü Sınıflandırması İçin Türkçe Konu Modellemesi Veri Kümesi
A Turkish Topic Modeling Dataset For Multi-label Classification of Movie Genre

Elgun Jabrayilzade
İzmir Yüksek Teknoloji Enstitüsü
Bilgisayar Mühendisliği
elgunjabrayilzade@std.iyte.edu.tr

Algın Poyraz Arslan
İzmir Yüksek Teknoloji Enstitüsü
Bilgisayar Mühendisliği
alginarslan@std.iyte.edu.tr

Hasan Para
İzmir Yüksek Teknoloji Enstitüsü
Bilgisayar Mühendisliği
hasanpara@std.iyte.edu.tr

Ozan Polatbilek
İzmir Yüksek Teknoloji Enstitüsü
Bilgisayar Mühendisliği
ozanpolatbilek@iyte.edu.tr

Erhan Sezerer
İzmir Yüksek Teknoloji Enstitüsü
Bilgisayar Mühendisliği
erhansezerer@iyte.edu.tr

Selma Tekir
İzmir Yüksek Teknoloji Enstitüsü
Bilgisayar Mühendisliği
selmatekir@iyte.edu.tr

***Özetçe* —İstatistiksel konu modellemesi, gözetimsiz bir şekilde dokümanlara konu ataması yapmayı amaçlar. Gizli Dirichlet Ayırımı (GDA) konu modellemesinde standart modeldir. Uzun dokümanlardan oluşan derlemlerde yüksek başarım gösterirken kısa metinlerde başarılı sonuçlar vermez. Kısa metinler üzerinde konu modellemesi sosyal medyanın rolü nedeniyle yükseliştedir. Dolayısıyla hem uzun hem de kısa metinler üzerinde konu tespiti yapan yaklaşımlar aranmaktadır. Bununla birlikte, aynı kesin referans kategorilere sahip uzun ve kısa metinlerin birlikte yer aldığı veri kümeleri eksikliği görülmektedir. Önerilen çalışmada, filmlerin hem kısa tanımlarını hem de uzun altyazılarını içeren bir Türkçe film veri kümesi[1] sunulmaktadır. Ayrıca sunulan veri kümesi için GDA doküman-konu veya Doc2Vec gösterimlerini girdi olarak alan bir Tam Bağlı Sinir Ağı (TBSA) kullanılarak çok-etiketli film türü sınıflandırması sonuçları verilmektedir.**

Anahtar Kelimeler—kısa metin sınıflandırma, uzun metin sınıflandırma, metin sınıflandırma veri kümesi, GDA, Doc2Vec, tam bağlı sinir ağı, film, altyazı, özet.

***Abstract*—Statistical topic modeling aims to assign topics to documents in an unsupervised way. Latent Dirichlet Allocation (LDA) is the standard model for topic modeling. It shows good performance on document collections, documents being relatively long texts but it has poor performance on short texts. Topic modeling on short texts is on the rise due to the potential of social media. Thus, approaches that are able to find topics on short texts as well as long texts are sought. However, there is a lack of datasets that include both long and short texts which have the same ground-truth categories. In this work, we release a Turkish movie dataset which contain both short film descriptions and long subscripts where film genre can be considered as topic. Furthermore, we provide multi-label movie genre classification results using a Feed Forward Neural Network (FFNN) taking LDA document-topic or Doc2Vec dense representations.**

Keywords—short text classification, long text classification, text classification dataset, LDA, Doc2Vec, feed-forward neural networks, movie, subtitle, plot.

TABLO I: Veri Kümesi İstatistikleri.

Tür	Film Sayısı
Animasyon	979
Macera	2532
Komedi	6444
Aile	1426
Fantezi	1807
Romantik	4544
Suç	3377
Dram	11957
Gerilim	5273
Western	642
Aksiyon	3738
Korku	2639
Tarih	1005
Biyografi	1112
Gizem	1986
Bilim-Kurgu	1905
Savaş	1133
Müzikal	1194
Spor	497
Belgesel	831
Film-Noir	192
Haber	27

Sınıf Sayısı	Film Sayısı
1	3612
2	6196
3	5888
4	3108
5	1206
6	372
7	93
8	26
9	1
10	1
Toplam	20503

	Ortalama Uzunluk
Özet	76.20
Altyazı	4671.36

I. GIRIŞ

İstatistiksel konu modellemesi, dokümanların kategorilendirmesinde kullanılan temel gözetimsiz yaklaşımdır. Dokümanların ait olduğu kategorilerin belirlenmesi, dokümanları sınıflandırmak üzere üstveriler üretilmesini ve bu üstveriler kullanılarak hedefe yönelik veri dağıtımını sağlar. Örneğin ekonomi alanındaki haberler ekonomi ile ilgilenen kullanıcılara dağıtılır. Sosyal medyanın temel iletişim aracı haline geldiği günümüzde kısa metinlerin analiz edilerek ön plana çıkan konuların belirlenmesi gündemin takip edilmesini, bu metinler üzerinden otomatik

[1] https://cloud.iyte.edu.tr/index.php/s/dUb03a6GiuiDynS

978-1-7281-7206-4/20/$31.00 ˜©˜ 2020 IEEE

olay tespiti yapılması acil durumlara anında müdahale edilmesini kolaylaştırır.

Geleneksel konu modellemesi yaklaşımı olan Gizli Dirichlet Ayırımı (GDA), dokümanlardan konulara, konulardan kelimelere geçişin yapıldığı üretimsel bir modeldir [1]. Modelin varsayımı, her dokümanın belli konuları işlemek üzere yazıldığı ve her konunun da kendisine ait bir terminolojisi olduğudur. Dokümanlar, konular ve kelimeler koşullu olasılıklar üzerinden birbirine bağlanmıştır. Model varsayılan halinde dokümanlar üzerinde yüksek başarım göstermektedir. Modelin kısa metinler üzerinde kabul edilebilir başarım gösterebilmesi için belli uyarlamaların yapılması gerekmektedir. Kısa metinler üzerinde istatistiksel konu modellemesi yükseliştedir ve bu konuda yeni çözümler önerilmektedir [2], [3]. Bu alanda özellikle kısa metinler içeren etiketli veri kümelerine gereksinim vardır.

Önerilen çalışmada, istatistiksel konu modellemesinde kullanılmak üzere Türkçe bir film veri kümesi sunulmaktadır. Kategorileri film türleri olan bu veri kümesi, hem kısa film tanımlarını hem de filmlere ait Türkçe altyazıları içermektedir. Bu sayede mevcut konu modellemesi algoritmalarının aynı kesin referans kategoriler için kısa ve uzun metinler üzerindeki başarımı aynı anda sınanabilecek ve karşılaştırmalı bir analiz mümkün hale gelecektir.

II. LİTERATÜR TARAMASI

Film türü sınıflandırması makine öğrenmesi alanında üzerinde sık çalışılan bir konudur. Görsel ve metin girdiler başta olmak üzere, birçok türde girdi kullanılarak sınıflandırma yapmaya çalışan modeller ve veri kümeleri mevcuttur.

İmge ve video gibi görsel bilgileri girdi olarak kullanan metotlar arasından [4] ve [5], film afişlerindeki imgelerden yaralanırken, [6], [7], [8] ve [9] film fragmanlarını kullanarak tür bilgisini tahminlemeye çalışmaktadır.

Öte yandan metin bazlı bilgileri kullanan birçok yöntem de bulunmaktadır. Bunların arasından [10], işitme engelliler için oluşturulan tarif bilgilerini (ing. captions) kullanırken, [11], [12] ve [13] film özeti bilgisini kullanarak sınıflandırma yapmaktadır. [14] ise filmlerin altyazı bilgilerini çeşitli metotlarla sınayarak film türü tespiti yapmaktadır.

Bu alanda önerilmiş film özeti bilgisi sunan İngilizce birkaç veri kümesi mevcuttur. Doshi ve Zadrozny [12], 60000 film özetinden oluşan bir özet veri kümesi sunmaktadır, fakat bu veri kümesinde herhangi bir tür çakışması olmaksızın sadece dört film türü mevcuttur. Bamman vd. [15] ise 42000 film özeti barındıran çok-etiketli bir İngilizce film özeti veri kümesi sunmaktadır. Son olarak ise, Hoang [13], 250000 film özetinden oluşan bir veri kümesi sunmaktadır. Bunlara ek olarak, İngilizce için, altyazı bilgisi ile sınıflandırma imkanı sunan bir veri kümesi de mevcuttur [14]. Fakat bu veri kümesinde yalnızca üç türe ait toplam 14000 film altyazısı bulunmaktadır.

Türkçe'de de bu konuda yapılan bir çalışma mevcuttur. Ertuğrul ve Karagöz [11], kendi topladıkları film özeti bilgileri ile film türü tespitini hedeflemişlerdir, fakat bizim çalışmamızdan farklı olarak bu çalışmada sadece 4 türe ait tek-etiketli 6000 film özeti bulunmaktadır.

Bilgimiz dahilinde, filmlere ait hem özet hem de altyazı bilgisi sunan başka veri kümesi mevcut değildir, ayrıca Türkçe'de çok-etiketli ve bu denli geniş çaplı bir veri kümesi henüz bulunmamaktadır.

III. VERI KÜMESI

Çalışma kapsamında sunulan film veri kümesinin oluşturulmasındaki ilk adım, derlem içerisinde yer alacak filmlerin listesinin belirlenmesidir. Bu amaçla Kaggle tarafından yayınlanan "The Movies Dataset" [16] esas alınmıştır. Bu film listesinden, çok az veriye sahip "yetişkin" sınıfına ait filmler ve bir konu belirtilmediğinden "kısa filmler" çıkarılmıştır.

Film listesinin belirlenmesinin ardından IMDB portalından[2] filmlere ait temel üstveriler çekilmiştir. Bu üstveriler arasında tür, puan, yönetmen ve senarist bilgileri yer almaktadır.

Filmlerin özetleri ve Türkçe altyazıları ise planetdp[3] sitesinden paletlenmiştir. Bu site üzerinde filmleri aramak ve arama sonucunda ilgili dosyaları indirmek üzere Selenyum kütüphanesini[4] kullanan bir Python betiği yazılmıştır. Filmlerin altyazılarının indirilmesi işlemi sırasında, bazı filmlerde birden fazla Türkçe altyazı seçeneğinin bulunduğu ve yine bazen Türkçe altyazı yerine başka dilde sonuç alındığı gözlemlenmiştir. Bu hususların çözümünde, birden fazla altyazı durumunda en popüler olanı tercih edilmiş, farklı dildeki sonuçları ayıklamak içinse fasttext [17] tabanlı dil algılama modeli kullanılmıştır. Ayrıca, eksik altyazısı bulunan (20000 karakterden az) filmler de kaldırılmıştır.

Sonuç olarak, 22 türden 20503 filmi içeren bir veri kümesi oluşturulmuştur. Filmlerden 2966'sının altyazısı, 4750'sinin özeti, 301'inin senaristi, 24'ünün yönetmeni ve 1'inin puanı yoktur. Bir film birden fazla tür ile eşleşebilmektedir. Veri kümesine ait temel istatistikler Tablo I'de verilmektedir.

IV. YÖNTEM

Çalışmada, veri kümesinin zorluk derecesini göstermek ve referans noktası oluşturmak adına çeşitli temel yöntemler hem kısa metinler (film özetleri) hem de uzun metinler (film altyazıları) üzerinde denenmiştir. Öncelikle GDA [1] ve Doc2Vec [18] modelleri ile altyazı ve kısa metinlerin doküman vektörleri elde edilmiş ve ardından Tam Bağlı Sinir Ağı kullanılarak film türü sınıflandırması yapılmıştır. Doküman vektörü oluşturmada GDA istatistiksel bir yöntem iken Doc2Vec sinir ağı tabanlı bir modeldir. Dolayısıyla her iki gösterim yöntemi de denenerek farklı yaklaşımların veri kümesi üzerindeki başarımları gösterilmiştir.

A. Gizli Dirichlet Ayırımı

Gizli Dirichlet Ayırımı (GDA), dokümanların konularını belirlemek üzere tasarlanmış gözetimsiz bir modeldir. Model; dokümanlar (D), konular (T) ve kelimeler (W) bileşenlerinden oluşan bir nedensellik modelidir. Bileşenler arasındaki ilişki $D \rightarrow T \rightarrow W$ patikası ile verilebilir. Modelin varsayımı,

[2]https://www.imdb.com/

[3]https://www.planetdp.org/

[4]https://robotframework.org/SeleniumLibrary/SeleniumLibrary.html

her dokümanın belli konuları işlemek üzere yazıldığı ve her konunun da kendisine ait bir terminolojisi olduğudur. Dolayısıyla bileşenler D ve T ile T ve W koşullu olasılıklarla birbirine bağlıdır. Dokümanlardaki kelime dağılımı $p(W|D)$, model patikası temel alınarak $p(T|D)$ ve $p(W|T)$ koşullu olasılıklarının çarpımı olarak formüle edilmiştir. Bayes teoremi kullanılarak ters olasılıklar hesaplanabilir, bir başka ifade ile kelimelerden konular çıkarsanabilir.

GDA kullanılarak gerek dokümanlar gerekse de kelimeler için boyutları konular olan vektörler elde edilebilir ve bu vektörlerin semantik gösterimler olarak işe yararlığı bazı doğal dil işleme görevlerinde ortaya konmuştur [19]. Literatürde GDA'dan elde edilen doküman-konu olasılık dağılış vektörleri dokümanların konu sınıflandırmasında kullanılmıştır. Sarıoğlu vd. [20] çalışmasında, doküman vektörlerinin boyutu arttıkça SVM ile sınıflandırma başarımının arttığı gözlemlenmiştir. Pérez vd. [21] tarafından yapılan çalışmada GDA modeli asıl sınıf sayısından daha fazla sınıf sayısı ile çalıştırılarak elde edilen doküman vektörleri öznitelik vektörü olarak kullanılmış ve bu vektörler sınıflandırıcılara beslenerek yüksek başarım elde edilmiştir. Bu yaklaşım temel alınarak, GDA modeli, verilen metinleri 200 sınıfa ayıracak şekilde çalıştırılmış ve elde edilen 200 boyutlu olasılık dağılış vektörleri doküman gösterimleri olarak ele alınmıştır.

B. Doc2Vec

Doc2Vec [18], doküman gösterimleri öğrenmek üzere tasarlanmış bir yapay sinir ağı dil modelidir. DM ve DBOW olmak üzere iki varyasyonu vardır. DM modelinde girdi olarak aynı pencere içerisinde yer alan kelimeler (son kelime hariç) ile birlikte doküman vektörü alınır ve pencere içerisindeki son kelime tahmin edilir. DBOW modelinde ise doküman vektörü girdi olarak alınıp doküman içerisindeki bir pencere seçilerek o penceredeki bir kelime tahminlenmeye çalışılır. Çalışmada Wikipedia verisi ile eğitilmiş bir Doc2Vec modeli kullanılarak hem film özetleri hem de altyazılar için Doc2Vec gösterimleri hesaplanmıştır.

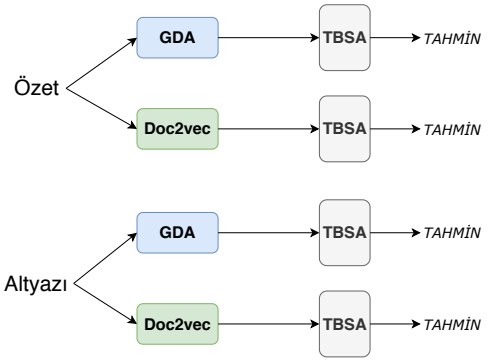
C. Tam Bağlı Sinir Ağı

GDA ve Doc2Vec modelleri ile elde edilen doküman vektörlerini işleyip çok-etiketli konu sınıflandırması yapmak üzere Tam Bağlı Sinir Ağı (ing. Fully-Connected Neural Layer) kullanılmıştır. Tam Bağlı Sinir Ağı (TBSA), çok-etiketli sınıflandırma problemlerindeki başarımı dikkate alınarak tercih edilmiştir.

GDA ve Doc2Vec modellerinden elde edilen öznitelik vektörleri TBSA modeline girdi olarak verilip çok-etiketli sınıflandırma yapılmıştır. Her iki model için de öznitelik vektörlerinin boyutu 200 olduğundan TBSA modelinin de girdi katmanı büyüklüğü 200'dür. Gizli katman büyüklüğü bir üst değişken olarak ele alınmış olup 100 ve 150 değerleri ile ayrı ayrı sınanmış ve aktivasyon fonksiyonu olarak Relu kullanmılmıştır. Son katmanın büyüklüğü ise sınıf sayısı olan 22'dir ve sigmoid fonksiyonu ile sınıf olasılıkları elde edilmiştir. Film özetleri ve altyazılar için uygulanan sınıflandırma iş akışı Şekil 1'de gösterilmektedir.

TABLO II: GDA+TBSA ile Sınıflandırma Örneği.

Film	Tahmin	Gerçek Değer
Yüzüklerin Efendisi	macera, fantezi, drama	macera, fantezi, drama
Yıldız Savaşları	macera, fantezi, aksiyon, animasyon	macera, fantezi, aksiyon, bilim-kurgu
Görünmez adam	korku, bilim-kurgu	gerilim, aksiyon



Şekil 1: Çok-etiketli Sınıflandırma İş Akışı.

D. Rassal Taban

Rassal Taban başarım seviyesi sayesinde, veri kümesi üzerinde çalıştırılacak herhangi bir modelin başarımının en az ne kadar olması gerektiği belirlenebilir. Rassal taban belirlenirken her bir etiketin diğerlerinden bağımsız olduğu ve görülme olasılığının %50 olduğu varsayımı ile analiz yapılmıştır. Bu analiz sonucunda rastgele çalışan bir modelin doğruluk oranı %50 ve F1 skoru 0.1880'dir.

V. Sonuçlar

Veri etiketlerinin sıkça birden fazla olması, dağılışı koruyan bir eğitim ve test kümesi ayrımı yapmayı zorlaştırmıştır. Hem bu sebepten hem de yöntemlerin kuvvetini ölçmek üzere K-Katlamalı Çapraz Doğrulama tekniği uygulanmıştır. K olarak 5 seçilmiş olup, oluşturulan 5 kattaki veriler her bir model için korunmuştur.

Çok-etiketli film sınıflandırması başarımının değerlendirmesinde kesinlik, duyarlılık ve F1 metriğinin yanında kesin eşleşmenin (ing. exact match) kullanımının uygun olacağı değerlendirilmiştir. Zira tam bağlı sinir ağından elde edilen tahmin vektörü ile kesin referans vektörün kesin eşleşmesi bir filmin ait olduğu tüm türlerin doğru tahmin edilmesi anlamına gelirken kısmi eşleşmeler filmin türlerinin bir alt kümesinin doğru tahmin edilmesi demektir. Kesinlik, duyarlılık ve F1 metrikleri ise çoklu etiketlerden her birinin doğru tahmin edilmesine odaklanır. Örneğin Tablo II'de de görüleceği üzere "Yüzüklerin Efendisi" filminin kesin referans türleri macera, fantezi ve dramdır. Sınıflandırıcı bu türlerin hepsini doğru tahminlemişse kesin eşleşme olarak değerlendirilir ve aynı şekilde kesinlik, duyarlılık ve F1 değerleri tamdır. "Yıldız Savaşları" filminin kesin referans tür bilgisi macera, fantezi, aksiyon ve bilim kurgu olarak verilmiştir. Sınıflandırıcının filmin türünü animasyon, macera, fantezi ve aksiyon olarak tahminlemesi durumunda kesin eşleşme gerçekleşmemiştir çünkü bilim kurgu türü yerine animasyon etiketlemesi yapılmıştır. Filmin animasyon olarak kategorilendirmesi kesinlik değerini azaltırken, bilim kurgu

TABLO III: Sınanan Yöntemlerin Veri Kümesindeki Sonuçları.

Yöntem	P (Kesinlik)	R (Duyarlılık)	F1	Doğruluk	Kesin Eşleşme (Exact Match)
	Altyazı				
Rassal Taban	0.500	0.121	0.188	0.500	$2.8\text{x}10^{-7}$
GDA + TBSA	**0.6** ($\pm$0.012)	**0.737** ($\pm$0.014)	**0.623** ($\pm$0.007)	**0.912** ($\pm$0.003)	**0.147** ($\pm$0.012)
Doc2Vec + TBSA	0.473 ($\pm$0.019)	0.516 ($\pm$0.033)	0.458 ($\pm$0.018)	0.890 ($\pm$0.003)	0.089 ($\pm$0.012)
	Film Özeti				
Rassal Taban	**0.500**	0.121	0.188	0.500	$2.8\text{x}10^{-7}$
GDA + TBSA	0.482 ($\pm$0.025)	**0.538** ($\pm$0.013)	**0.468** ($\pm$0.009)	**0.888** ($\pm$0.001)	0.083 ($\pm$0.018)
Doc2Vec + TBSA	0.459 ($\pm$0.040)	0.496 ($\pm$0.070)	0.439 ($\pm$0.022)	**0.887** ($\pm$0.003)	**0.092** ($\pm$0.005)

türünde olduğu bilgisinin kaçırılması duyarlılığa zarar verir. Dolayısıyla değerlendirmede bu metriklerin kesin eşleşme metriği ile birlikte kullanılması başarımın açıklanmasında tam eşleşmelerin payının belirlenmesine olanak tanır.

Tablo III'te de görülebileceği üzere, GDA altyazılarda daha iyi kesin eşleşme değerleri verirken, film özetlerinde Doc2Vec daha başarılı sonuçlar vermiştir. Kesin eşleşme değerleri ile ilgili bir başka dikkate değer olgu ise, Doc2Vec ile elde edilen sonuçlarda iki kategori arasında ciddi bir fark gözlenmezken, GDA'nın sonuçlarının çok daha değişken olmasıdır. Buradan, beklendiği üzere, GDA'nın metin uzunluğundan daha fazla etkilendiği çıkartılabilir. Bilindiği üzere, GDA genellikle uzun metinlerde çok iyi sonuç verirken, metin kısaldıkça bu sonuçların kötüleştiği gözlemlenmiştir. Öte yandan, Doc2Vec metodu metin uzunluğundan fazla etkilenmeyip kısa metinlerde de başarılı sonuçlar üretmiştir.

Öte yandan F1 değerlerinde her iki kategoride de GDA'nın daha başarılı olduğu gözlemlenmiştir. Kesin eşleşme değerlerine benzer olarak, GDA yine uzun metinlerde daha büyük bir başarı göstermiş ve daha önemli bir farkla Doc2Vec'i geride bırakmıştır.

VI. Sonuç

Bu çalışmada çok-etiketli konu modellemesi problemi kısa ve uzun metinler şeklinde ele alınmış olup, farklı türlerde filmlerin altyazı ve özetleri derlenerek bir veri kümesi önerilmiştir. Önerilen veri kümesi üzerinde hem sinir ağı hem istatistiksel temel yöntemler kullanılarak öznitelikler öğrenilip Tam Bağlı Sinir Ağı ile sınıflandırma yapılmıştır. GDA tabanlı model uzun metinlerde önemli ölçüde başarılı olurken, Doc2Vec tabanlı model metin uzunluğundan etkilenmeyip kısa metinlerde ufak bir farkla daha başarılı olabilmiştir.

Ek olarak belirtilmelidir ki, bir film için verilen tür bilgileri film için eşit derecede tanımlayıcı olmayabilir. Dolayısıyla çok-etiketli sınıflandırmada her etiketin ağırlığını eşit kabul etmek yanıltıcı olabilir. Örneğin "Yıldız Savaşları" filmini bir kişi temel olarak fantezi türünde değerlendirirken bir başkası bir bilim kurgu filmi olarak görebilir. Bununla birlikte, IMDB'de verilen film tür etiketleri öncelik bilgisi içermediğinden bu ayrımı yapmak mümkün olmamıştır.

Kaynakça

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[2] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '16. New York, NY, USA: ACM, 2016, pp. 165–174.

[3] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 233–242.

[4] W.-T. Chu and H.-J. Guo, "Movie genre classification based on poster images with deep neural networks," in *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, ser. MUSA2 '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 39–45.

[5] K. Kundalia, Y. Patel, and M. Shah, "Multi-label movie genre detection from a movie poster using knowledge transfer learning," *Augmented Human Research*, vol. 5, no. 1, p. 11, Dec 2019.

[6] J. Wehrmann and R. C. Barros, "Convolutions through time for multi-label movie genre classification," in *Proceedings of the Symposium on Applied Computing*, ser. SAC '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 114–119.

[7] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, "Movie genre classification via scene categorization," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 747–750.

[8] G. S. Simões, J. Wehrmann, R. C. Barros, and D. D. Ruiz, "Movie genre classification with convolutional neural networks," in *2016 International Joint Conference on Neural Networks (IJCNN)*, July 2016, pp. 259–266.

[9] K. Sivaraman and G. Somappa, "Moviescope : Movie trailer classification using deep neural networks," 2017.

[10] D. Brezeale and D. J. Cook, "Using closed captions and visual features to classify movies by genre," in *In Poster session of the Seventh International Workshop on Multimedia Data Mining (MDM/KDD2006*, 2006.

[11] A. M. Ertugrul and P. Karagoz, "Movie genre classification from plot summaries using bidirectional lstm," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, Jan 2018, pp. 248–251.

[12] P. Doshi and W. Zadrozny, "Movie genre detection using topological data analysis," in *Statistical Language and Speech Processing*, T. Dutoit, C. Martín-Vide, and G. Pironkov, Eds. Cham: Springer International Publishing, 2018, pp. 117–128.

[13] Q. Hoang, "Predicting movie genres based on plot summaries," *ArXiv*, vol. abs/1801.04813, 2018.

[14] M. Pieters and M. Wiering, "Comparison of machine learning techniques for multi-label genre classification," in *Artificial Intelligence*, B. Verheij and M. Wiering, Eds. Cham: Springer International Publishing, 2018, pp. 131–144.

[15] D. Bamman, B. O'Connor, and N. A. Smith, "Learning latent personas of film characters," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 352–361.

[16] R. Banik. (2017) The movies dataset. [Online]. Available: https://www.kaggle.com/rounakbanik/the-movies-dataset

[17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2016.

[18] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, p. II–1188–II–1196.

[19] T. L. Griffiths, J. B. Tenenbaum, and M. Steyvers, "Topics in semantic representation," *Psychological Review*, vol. 114, p. 2007, 2007.

[20] E. Sarioglu, K. Yadav, and H.-A. Choi, "Topic modeling based classification of clinical reports," in *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 67–73. [Online]. Available: https://www.aclweb.org/anthology/P13-3010

[21] J. Pérez, A. Pérez, A. Casillas, and K. Gojenola, "Cardiology record multi-label classification using latent dirichlet allocation," *Computer Methods and Programs in Biomedicine*, vol. 164, pp. 111 – 119, 2018.