Bilim Kurgu Kitapları Tema Tahmini

| Beyza Akdeniz |  |
| --- | --- |
| *Bilgisayar Mühendisliği Departmanı* |  |
| *Tobb Economy and Technology University* |  |
| *Ankara, Türkiye* |  |
| [beyzaakdeniz@etu.edu.tr](mailto:beyzaakdeniz@etu.edu.tr) |  |

***Abstract* - Bilim kurgu alt türlerinin çok-etiketli sınıflandırması projesi gerçekleştirilmiştir. Veri seti, çeşitli bilim kurgu alt türlerini içeren kitapların açıklama metinlerini ve ilgili alt tür etiketlerini içermektedir. Amaç, bu açıklama metinlerini analiz ederek her bir kitabın alt türlerini doğru bir şekilde tahminlemektir. Projede, metin işleme teknikleri (veri temizleme, önişleme), kelime vektörleştirme ve farklı makine öğrenimi modelleri (Random Forest, Lineer Regresyon, Destek Vektör Makineleri) kullanılmıştır. Veri seti temizlenip işlendikten sonra, metin verileri vektörleştirilmiş ve modeller üzerinde eğitim gerçekleştirilmiştir. Her bir alt tür için, doğruluk, F1 skoru, Hamming kaybı ve ROC eğrisi altında alan değeri (AUC ROC) gibi metrikler hesaplanmıştır. Bu metrikler, modellerin performansını değerlendirmek ve karşılaştırmak için kullanılmıştır. Sonuçlar Lineer Regresyon modelinin en iyi performansı sergilediğini göstermektedir. Bu model, diğer iki modele göre daha yüksek recall ve F1 skoru değerleri elde etmiştir. Bununla birlikte, ROC eğrisi altında alan değeri (AUC ROC) en düşük olan model ise Destek Vektör Makineleri olmuştur. Bu sonuçlar, metin madenciliği ve çok-etiketli sınıflandırma alanlarında daha fazla araştırma yapılmasının önemini vurgulamaktadır.**

**Bu proje, bilim kurgu alt türlerini analiz etmek ve tahminlemek için metin madenciliği ve makine öğrenimi tekniklerini birleştirmiş ve farklı modellerin performansını karşılaştırmıştır.**

***Keywords- Multi-label Classification, Data Mining, Text Analysis, Science Fiction***

I. Giriş

Bilim kurgu, edebiyat ve pop kültürün ayrılmaz bir parçasıdır ve geniş bir yelpazede alt türlere sahiptir. Ancak, bu alt türlerin karmaşıklığı ve çeşitliliği, doğru bir şekilde sınıflandırılmasını zorlaştırabilir. Bir roman veya öykü, birden fazla alt türe sahip olabilir ve bu da onu birden fazla etiketle tanımlanabilen bir veri noktası haline getirir.

Bu proje, metin madenciliği ve makine öğrenimi tekniklerini kullanarak bilim kurgu eserlerini farklı alt türlere sınıflandırmayı amaçlamaktadır. Çok-etiketli sınıflandırma yaklaşımıyla, her bir eser birden fazla alt türe ait olabilir ve her bir alt türün varlığını tahmin etmek için farklı metin özellikleri kullanılabilir.

Projede kullanılan ana yöntem, metin önişleme, kelime vektörleştirme ve farklı makine öğrenimi modellerinin uygulanmasıdır. Metin madenciliği teknikleriyle, eserlerin açıklamaları temizlenir, özgün kelimeler ve kökler çıkarılır, gereksiz bilgiler filtrelenir ve son olarak kelime vektörlerine dönüştürülür. Ardından, çok-etiketli sınıflandırma modelleri kullanılarak eserlerin her bir alt türü tahmin edilir.

Bu çalışmanın amacı, bilim kurgu eserlerinin alt türlere ayrılmış ve etiketlenmiş verilerle nasıl sınıflandırılabileceğini göstermektir. Farklı makine öğrenimi modellerinin performansı karşılaştırılarak, alt tür tahmini konusundaki etkinlik değerlendirilir. Bu çalışma, metin madenciliği ve çok-etiketli sınıflandırma alanındaki uygulamalara örnek teşkil ederek, daha geniş bir metin verisi kümesinde benzer yöntemlerin nasıl kullanılabileceğini aydınlatmayı amaçlamaktadır.

II. Literatur araştırması

Bilim kurgu türündeki eserlerin sınıflandırılması ve alt türlere ayrılması, metin madenciliği ve makine öğrenimi alanlarında sıklıkla araştırılan bir konudur. Alt tür tahminlemesi, özellikle çok-etiketli sınıflandırma (multi-label classification) ve doğal dil işleme (NLP) tekniklerinin birleşimini gerektirir. Aşağıda, bu alanda yapılan bazı önemli çalışmalar ve yöntemler özetlenmiştir:

*A. Alt Türlere Göre Özellik Çıkarımı:*

Bazı çalışmalar, alt tür tahmini için özel olarak tasarlanmış özelliklerin çıkarılmasını içerir. Bu özellikler, metinlerin kelime kullanımı, cümle yapısı, semantik anlamı vb. gibi yönlerini yakalamayı amaçlar. Bu tür yaklaşımlar, geleneksel özellik mühendisliği prensiplerini kullanarak alt tür tahminini gerçekleştirmeye çalışır.

*B. Kelime Gömme (Word Embedding):*

Kelime gömme teknikleri, metinleri semantik olarak anlamlı vektörlerle temsil etmek için kullanılır. Bu yaklaşım, kelime benzerliklerini ve ilişkilerini yakalayarak metinlerin anlamını daha iyi yakalamayı amaçlar. Özellikle Word2Vec, GloVe ve FastText gibi popüler kelime gömme modelleri, bu tür uygulamalarda yaygın olarak kullanılır. [3]

*C. Sinir Ağları ve Derin Öğrenme:*

*Derin öğrenme, büyük veri setleri üzerinde otomatik özellik çıkarımı ve karmaşık örüntüleri tanıma yeteneği nedeniyle alt tür tahminlemesi için sıkça kullanılır. Özellikle evrişimli sinir ağları (CNN) ve uzun-kısa vadeli bellek (LSTM) gibi sinir ağı mimarileri, metin verilerini etkili bir şekilde işlemek ve alt tür tahminini gerçekleştirmek için kullanılır. [1]*

*D. Çok-Etiketli Sınıflandırma (Multi-Label Classification):*

*Bilim kurgu eserleri gibi çoklu alt türlere sahip metinlerin sınıflandırılması için çok-etiketli sınıflandırma yöntemleri yaygın olarak kullanılır. Bu yaklaşımda, her bir metin birden fazla etiketle ilişkilendirilebilir. Çok-etiketli sınıflandırma yöntemleri, etiket etkileşimlerini ve korelasyonlarını dikkate alarak tahminler yapabilir.[2]*

*E. Model Karşılaştırması ve Değerlendirme Metrikleri:*

*Literatürde, farklı alt tür tahminlemesi yöntemlerinin performansını karşılaştıran çalışmalar mevcuttur. Bu çalışmalar genellikle doğruluk (accuracy), hassasiyet (precision), duyarlılık (recall), F1 skoru ve ROC eğrisi altındaki alan (AUC-ROC) gibi metriklerle modelleri değerlendirir.*

*Bu proje ise bu alanın öne çıkan uygulamalarından birini temsil ederek, metin madenciliği ve çok-etiketli sınıflandırma tekniklerinin nasıl birleştirilebileceğini göstermeyi amaçlamaktadır.*

III. Veri seti

Kullanılan veri seti bağlantısı:

https://www.kaggle.com/datasets/tanguypledel/science-fiction-books-subgenres

1. *Veri Seti Tanımı:*

*Kullanılan veri seti, bilim kurgu romanlarının metin verilerini ve bu romanların alt tür etiketlerini içermektedir. Her bir veri noktası, bir kitap hakkındaki metin verileri (kitap tanıtıcı yazıları, açıklamalar) ve alt tür etiketlerinden oluşmaktadır.*

1. *Öznitelik Çıkarımı:*

*Metin verileri özellik çıkarımı için Bag-of-Words (BOW) yöntemi kullanılmıştır. BOW, metinlerde geçen kelimelerin sıklığını temsil eden bir özellik vektörüdür.*

1. *Önişleme Aşamaları:*
2. *Metin Temizleme: Metin verilerindeki gereksiz semboller, sayılar ve özel karakterler temizlenmiş ve sadece harf ve boşluk karakterleri bırakılmıştır.*
3. *Tokenizasyon: Temizlenen metin verileri cümlelere ve kelimelere bölünmüştür.*
4. *Stop Words Kaldırma: İngilizce stop kelimeleri (anlam taşımayan yaygın kelimeler) kaldırılarak metinler daha anlamlı hale getirilmiştir.*
5. *Kelime Kökleri Çıkarma (Lemmatizasyon): Kelimelerin köklerine dönüştürülerek kelime varyasyonlarından arındırılmıştır.*
6. *Bag-of-Words (BOW) Özellikleri: Metinlerden elde edilen temizlenmiş ve lemmatize edilmiş kelimeler, BOW yöntemi ile özellik vektörlerine dönüştürülmüştür.*

*D. Öznitelik Açıklamaları:*

1. *Hatalı Veriler ve Eksik Veriler:*

*Kullanılan veri seti üzerinde öncelikle eksik veri analizi yapılmıştır. Eksik veriler gerekli önişleme adımlarıyla temizlenmiştir.*

1. Structured / Unstructured:

Kullanılan veri setindeki metin verileri "unstructured" formattadır, çünkü doğal dil metinlerinden oluşmaktadır ve belirli bir yapı içermez.

1. Sayısal Değerler:

Kullanılan özellikler metin verileri üzerinden çıkarıldığı için sayısal değerler içermemektedir.

1. Ordinal ve Nominal Veriler:

Projede nominal veriler kullanılmıştır.

1. Sayısal Olmayan Veriler:

Kullanılan özellikler metin verilerinden oluşturulmuştur:

IV. Kullanılan modeller

Bu projede, bilim kurgu romanlarının alt türlerine sınıflandırılması için farklı makine öğrenimi modelleri kullanılmıştır. Aşağıda kullanılan modellerin tanımları, metodolojileri ve neden bu modellerin seçildiği açıklanmıştır:

1. Eğitim / Test Ayırımı:

Veri seti, ve test olmak üzere iki parçaya ayrılmıştır. Veri setinin %80'i eğitim için kullanılmış ve %20'si test için ayrılmıştır.

1. Kullanılan Modeller ve Metodolojiler:
2. Random Forest:

Random Forest, ağaç tabanlı bir ensemble öğrenme algoritmasıdır. Her ağaç, verilerin rastgele örneklemlerini ve özniteliklerini kullanarak eğitilir. En sonunda tüm ağaçların tahminleri bir araya getirilir. [4]

Projede alt tür tahminlerinde kullanılan bu model, çoklu etiket (multi-label) sınıflandırma problemini ele almak için uygundur.

C. Linear Regression:

Linear Regression, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi doğrusal bir denklemle modeller. Logistic Regression gibi çoklu sınıfların tahmin edilmesinde de kullanılabilir.

Projede alt tür tahminlerinde kullanılan bu model, çoklu etiketli sınıflandırma için uygun bir alternatif olarak seçilmiştir.

D. Support Vector Machine (SVM):

Support Vector Machine, verileri sınıflara bölmek için hiper-düzlemler kullanır. Lineer ve non-lineer sınıflandırma problemlerini çözmek için kullanılır.

Projede alt tür tahminlerinde kullanılan bu model, çoklu etiketli sınıflandırma için özgün bir yaklaşım olarak seçilmiştir.

Bu projede bu modellerin tercihinin başlıca nedenleri:

Random Forest: Özellikle çoklu etiket (multi-label) sınıflandırma problemlerinde iyi sonuçlar verdiği için tercih edilmiştir.

Linear Regression: Çoklu sınıfların tahmininde kullanılabilirliği ve basitliği nedeniyle tercih edilmiştir.

Support Vector Machine (SVM): Hem lineer hem de non-lineer sınıflandırma problemleri için etkili bir seçenek olduğu için tercih edilmiştir.

Modellerin Seçimi ve Literatür Bilgisi:

Bu modeller literatürde yaygın olarak kullanılan ve çoklu etiketli sınıflandırma problemlerinde iyi performans gösteren modellerdir. Özellikle metin verileri üzerinde sınıflandırma yaparken bu modellerin tercih edildiği literatürde sıkça belirtilir.

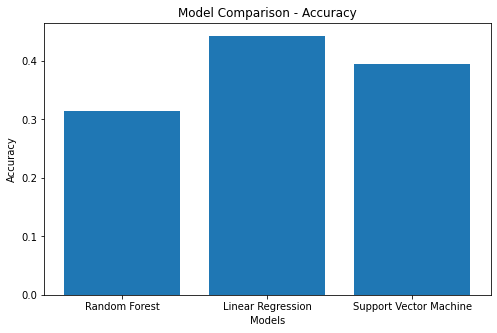
Sınıflandırma Türü:

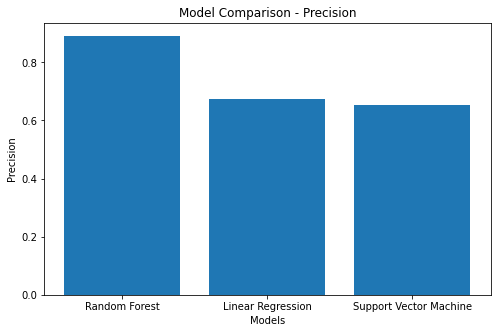
Bu projede çoklu etiket (multi-label) sınıflandırma yapılmaktadır. Her bir roman birden fazla alt türe sahip olabilir, bu nedenle her alt tür etiketi için ayrı bir sınıflandırma yapılır.

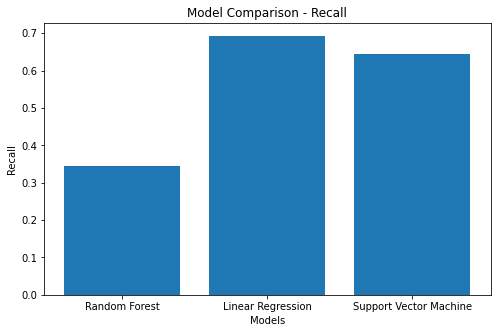
V. Test sonuçları ve sonuçların yorumları, Tartışma

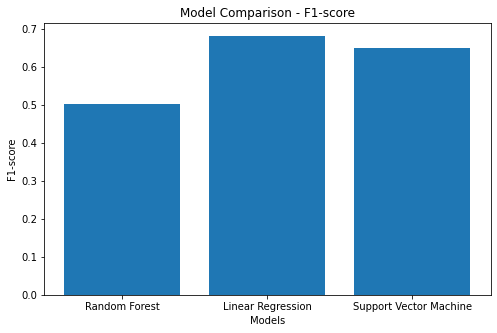
Çok-etiketli sınıflandırmada, her örnek aynı anda birden fazla sınıfa ait olabilir. Bu nedenle, karışıklık matrisi yorumu

daha karmaşık hale gelir. Doğrudan confusion\_matrix işlevini kullanmak yerine, çok-etiketli sınıflandırmayla ilgili çeşitli metrikleri hesaplanmıştır.









Bu sonuçlara dayanarak:

Lineer Regresyon Modeli, en yüksek Accuracy, Precision, Recall ve F1 Score değerlerine sahiptir. Aynı zamanda AUC ROC Score da diğer iki modele göre daha yüksektir. Bu nedenle, modelin genel olarak diğerlerine kıyasla daha iyi performans gösterdiği söylenebilir.

Random Forest Modeli, diğer iki modele kıyasla daha düşük performans sergilemektedir. Düşük Accuracy, Precision, Recall ve F1 Score değerleri ile daha zayıf sonuçlar elde edilmiştir.

Destek Vektör Makinesi (SVM) Modeli, ölçümlerde orta bir yerde bulunmaktadır. Diğer modellere kıyasla daha düşük Accuracy ve Precision değerlerine sahip olsa da, diğer metriklerde orta bir performans sergilemektedir.

Sonuç olarak, Lineer Regresyon Modeli en iyi performansı gösteren model olmuştur. Bu değerlendirmeler metrik değerlerine dayanarak yapılmıştır ve model seçimi konusunda daha fazla istatistiksel analiz yapılabilir.

References

1. K. S. Jones, "Natural language processing: A historical review" in Current Issues in Computational Linguistics: In Honour of Don Walker, Dordrecht, The Netherlands:Springer, pp. 3-16, 1994.
2. Yun Chen, Bo Xiao, Zhiqing Lin, Cheng Dai, Zuochao Li and Liping Yan, "Multi-label Text Classification with Deep Neural Networks", 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp. 409-413, 2018.
3. A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng and C. Potts, "Learning word vectors for sentiment analysis", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-volume 1, pp. 142-150, 2011.
4. Brieman L, Random Forests, Machine Learning, 45, 5-32, (2001)
5. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", Advances in Neural Information Processing Systems, pp. 3111-3119, 2013.
6. Gjorgji Madjarov et al., "An extensive experimental comparison of methodsfor multi-label learning", Pattern recognition, vol. 45, no. 9, pp. 3084-3104, 2012.