### **TOBB ETÜ Yapay Zeka Mühendisliği Bölümü**

**YAP 470/BİL 570 Dönem Projesi Ara Rapor**

**2022-23 Yaz Dönemi**

**Proje Üyesi: Beyza Akdeniz Date: 20 /07 /2023**

| **1. Proje ismi** |
| --- |
| Bilim Kurgu Kitapları Tema Tahmini |
| **2. Proje tanımı [[1]](#footnote-0) (kısa özet, problemin açıklanması, beklenen sonuçlar)** |
| Bu projede kitapların tanıtıcı yazıları, kitap ismi ve kitap türleri modele verilerek supervised olarak Random Forest algoritmasıyla eğitilecektir. Bu şekilde overfit riskini azaltmak planlanmaktadır.Bu model, metinlerin içerdiği kelimelerin sıklığını dikkate alarak bir özellik vektörü oluşturur (BOW), aynı zamanda dil bilgisinin de önemli olduğu durumlar için Syntactic N-gram'lar kullanarak daha fazla bilgi elde eder. Bu yaklaşım, metinlerin daha kapsamlı incelenmesi yoluyla metinlerin bilim kurgu alt türünü tespit etme görevinde daha başarılı olmasını sağlaması beklenmektedir. |
| **3. Proje metodolojisi, daha önce benzer projelerde kullanılan metotlar hakkında bilgiler, literatür taraması** |
| Daha önce yapılan çalışmalarda, romanların türlerini belirlemek için farklı özellikler (sözcük sıklığı, karakter sayısı, kelime uzunluğu vb.) kullanılmıştır. Makine öğrenimi teknikleri (SVM, Naive Bayes, Random Forest, K-Nearest Neighbor and Logistic Regression) kullanılarak tür sınıflandırması yapılmıştır.  Kitap tanıtıcı yazılarının kullanılarak kitap türü sınıflandırmasının yapıldığı bazı çalışmalarda ise Recurrent Neural Network (RNN) ve Word2Vec gibi yöntemler kullanılarak kitapların tema tahmini yapılmıştır.  Bu projede ise Random Forest modeli, Bag-of-Words (BOW) özellik vektörleriyle genişletilmiş Syntactic N-gram'lar ile beraber kullanılmaktadır.  Yararlanılan makaleler:  <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3805945>  <https://link.springer.com/article/10.1007/s41870-022-00937-6>  <https://www.researchgate.net/publication/371884351_Book_Genre_Classification_System_Using_Machine_Learning_Approach_A_Survey>  <http://uu.diva-portal.org/smash/get/diva2:1599825/FULLTEXT01.pdf> |
| **4. Projede kullanılan veriler, dış bağlantılar hakkında bilgiler, verilerin tamamını veya birkaç örneği gösterebilirsiniz.** |
| Kullanılan veri seti bağlantısı:  <https://www.kaggle.com/datasets/tanguypledel/science-fiction-books-subgenres>  “sf\_time\_travel.csv”den alınmış ve kolonları seçilmiş verilerden bir örnek: |
| **5. Projenin ayrıntılı planı, akış şeması, veri işleyişi vs., projede karşılaşılan sorunlar, projenin şu andaki durumu (varsa elde edilen sonuçlar veya ekran çıktıları), bundan sonra yapılacaklar** |
| **Projenin Ayrıntılı Planı ve Bundan Sonra Yapılacaklar:**  Literatür taraması ve metodolojinin belirlenmesi:   * Paragraftan tema tahmini konusunda yapılan araştırmalar incelenmeli. * Makine öğrenmesi ve doğal dil işleme teknikleri üzerinde durulmalı. * Proje için en uygun metodoloji belirlenmeli.   Verilerin analiz edilmesi ve rapora işlenmesi:   * Kaggle'dan indirilen veri setleri okunmalı ve içeriği analiz edilmeli. * Veri ön işleme adımları uygulanmalı (temizleme, düzenleme, öznitelik mühendisliği). * Verilerin özellikleri ve dağılımları incelenmeli. * İstatistiksel analizler yapılmalı. * Analiz sonuçları raporlanmalı.   Sorunların raporlanması:   * Karşılaşılan sorunlar ve engeller belirlenmeli. * Sorunların nedenleri ve çözüm önerileri raporlanmalı.   Projenin güncel durumunun raporlanması ve sonraki planın oluşturulması:   * Projenin mevcut aşaması detaylı bir şekilde raporlanmalı. * Gerçekleştirilen adımlar, sonuçlar ve elde edilen bulgular açıklanmalı. * Sonraki aşamaların planı yapılmalı.   Projeyi sonuçlandırma:   * Veri setinin uygun bir şekilde bölünmesi (eğitim, doğrulama, test setleri). * Makine öğrenimi veya doğal dil işleme algoritmalarının uygulanması. * Modelin eğitimi ve hiperparametre ayarlaması. * Modelin performansının değerlendirilmesi.   Paper hazırlama:   * Projenin detaylı bir şekilde açıklanacağı bir makale hazırlanmalı. * Makalede problemin tanımı, kullanılan yöntemler, veri analizi, sonuçlar ve tartışmalar yer almalı. * Makale, IEEE makale formatına uygun bir şekilde düzenlenmeli. * Makale gönderim sürecine uygun şekilde son düzenlemeler yapılmalı.    Projede Karşılaşılan Sorunlar:  * Metin üzerinde çalışılan bir proje olduğundan datasetteki verilerin sayısallaştırılması için ekstra yöntemler gerekmiştir. * Kullanıcıların koyduğu tagler subjektiftir ve verinin kalitesini belirler. Bu da modelin doğruluğunu düşüren bir faktördür. * Bazı bilim kurgu alt türleri birbirine benzemektedir. Bu durum, doğru sınıflandırma yapmayı zorlaştırmıştır ve modelin bu benzer türleri birbirinden ayırt etmekte zorlanmasına neden olmuştur.   **Projenin Şu Anki Durumu:**  Veri seti üzerinde bir EDA yapılmış ve hangi özniteliklerin öne çıktığı ve hangilerinin kullanılmaması gerektiği yapılan analizlere göre belirlenmiştir.  Proje için Support Vector Machine, Random Forest ve Lineer Regression modelleri denenmiş ve en iyi performans gösteren modelin Random Forest olduğu gözlenmiştir. Yapılan literatür taramasıyla da beraber metinlerin sayısal verilere dökülebilmesi için kullanılan yöntemler incelenmiş ve iyi performans verebilmesi için Bag-of-Words ve Syntactic N-gram tekniklerine başvurulmuştur. (Bunun için kullanılan Bag-of-Words (BOW) modeli, metinlerdeki kelimelerin frekansını sayısal bir özellik uzayına dönüştürme yöntemidir. Metindeki her bir kelimenin sayısını sayar ve bu sayıları bir vektör olarak temsil eder. Ayrıca kullanılan Syntactic N-gramlar ise metindeki kelime gruplarının veya dil bilgisi birimlerinin (örneğin, cümleler veya paragraflar) frekansını dikkate alır.)  Projenin bundan sonraki kısmında EDA ile daha fazla görselleştirme yapılacak, modelin daha optimal çalışması ve süregelen sorunların çözülmesi için çalışılacak ve sonuçlar raporlanacaktır. |

1. Eğer gerekiyorsa ayrı bir sayfa kullanın [↑](#footnote-ref-0)