T.C. ÖLÇME, SEÇME VE YERLEŞTİRME MERKEZİ BAŞKANLIĞI

TÜRKÇE YETERLİLİK SINAVININ KONUŞMA VE YAZMA BECERİLERİNİN YAPAY ZEKÂ İLE DEĞERLENDİRİLMESİ PROJESİ

Proje Amacı ve Kapsamı

T.C. Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı (ÖSYM); ulusal ölçekte yıllık ortalama 10 milyonun üzerinde adaya sınav hizmeti sunan ve yaptığı sınav sayısı ile hizmet verdiği aday kitlesinin büyüklüğü bakımından dünyanın önde gelen kurumlarından biridir. ÖSYM, uyguladığı sınavlarda geçerli, güvenilir ve adil ölçme, seçme ve yerleştirme yapan kurum olma misyonu ile yoluna devam etmektedir.

Bu doküman, ÖSYM'nin dört temel dil becerisini (dinleme, konuşma, okuma ve yazma) ölçen yabancı dil olarak Türkçe yeterlik sınavının konuşma ve yazma becerilerinin yapay zekâ tabanlı bir otomatik değerlendirme sistemi ile puanlanma uygulamasının geliştirilmesi için projelendirme süreçleri hakkında bilgiler içermektedir.

Türkçe Yeterlik Sınavının Konuşma ve Yazma Becerilerinin Yapay Zekâ İle Değerlendirilmesi Projesi; hızlı, objektif ve doğruluk oranı yüksek bir otomatik değerlendirme sistemi oluşturmayı amaçlamaktadır.

Mevcut Durum

Türkçe Dört Becerili Dil sınavı için konuşma ve yazma becerileri çeşitli eğitimleri tamamlamış ve standardizasyon çalışmalarına katılmış uzman değerlendiriciler tarafından yapılmaktadır. Aday performansları iki farklı değerlendirici tarafından sınav kapsamında geliştirilmiş olan *Dereceli Puanlama Anahtarları* ile puanlanır. İki değerlendirici arasında sınav kapsamında belirlenmiş olan puan farkının aşılması durumunda aday performansı "Uzman Değerlendirici" olarak sınıflandırılmış üçüncü değerlendiriciye gönderilir ve yeniden değerlendirilir. Mevcut durumda bir adayın tüm konuşma ve yazma cevaplarının değerlendirme süresi ortalama 3 saattir.

Gizlilik Sözleşmesi (NDA) imzalanmasının ardından, pilot sınavlardan elde edilen konuşma ve yazma testlerine ilişkin veriler paylaşılacaktır. Bu veriler hem ön eğitim hem de modelin son halinin oluşturulması için kullanılacaktır. Eğer pilot veri yeterli bulunmazsa, ek veri toplama süreci başlatılabilecektir. Bu süreçte, yeni deneme sınavları düzenlenerek daha fazla veri elde edilmesi planlanmaktadır.

Proje kapsamında 4 deneme sınavı yapılmış ve yaklaşık 1000 adayın konuşma ve yazma testlerinde elde edilen verileri toplanmıştır. Bu sınavlar, Yapay Zekâ modelinin eğitilmesi ve performansının artırılması için kullanılacaktır. Konuşma ve yazma testlerinden elde edilen veriler Yapay Zekâ modelinin önce Diller için Avrupa Ortak Başvuru Metni (D-AOBM/CEFR) seviyelerine göre sınıflandırma problemini çözme performansı precision, recall, accuracy, F1 measure, bakımından değerlendirilecek ve ROC eğrisi çizdirilerek bu metrikler bakımından performans başarımı hem geçerleme hem de test işlemlerinde en az %90 olarak beklenmektedir. Ardından modelin rubrik bazlı puan tahmini açısından değerlendirilecek ve uzman değerlendirmeleri ile karşılaştırılarak geçerlik doğrulama testleri yapılacaktır. Yapay Zekâ modeli, uzman puanlamaları ile yüksek korelasyon (+0.80) göstermelidir. Modelin değerlendirme hatası (MAPE – Ortalama Mutlak Yüzde Hatası) %20'den düşük olmalıdır ve R-squared (R^2) değeri bakımından %80 olmalıdır. Geliştirilecek Yapay Zekâ modeli, Multi Faceted Rasch Model ile kestirimler

yapılacak ve in-fit, out-fit indexlerine bakılacaktır. Geliştirilecek Yapay zekâ modelinde güvenilirlik için bu değerlerin 0,5 ile 1,5 arasında olması beklenecektir.

TEKNİK İSTERLER

1. Sistemde kullanılacak veri kümesi, İDARE (ÖSYM) bünyesinde MSSQL veri tabanında depolanan Türkçe dilinde konuşma ve yazma örneklerini içerecektir. Model oluşturulması, doğrulama (validation) ve test için gereken veriler yerel sunucular (on-premises) üzerinde saklanacaktır.
   - Konuşma ve Yazma Dereceli Puanlama anahtarında yer alan 5 farklı ölçüt için farklı puanlara sahip aday performansları mevcuttur. Puanlama anahtarları 5 ölçütten ve her bir ölçüte bağlı 5 puandan oluşmaktadır. Her bir ölçüt için beklenen yeterlik tanımlamaları ve belirlenen puana karşıt gelen örnek performanslar bulunmaktadır.
   - Konuşma verileri ses kayıtları "mp3, wav" gibi yaygın kullanılan formattadır.
   - Yazma testi aday cevapları "txt, csv" gibi yaygın formattadır.
   - Her bir veri örneği, sınav kapsamında geliştirilen Dereceli Puanlama Anahtarında yer alan ölçütlere göre en az iki farklı değerlendirici tarafından puanlanmıştır.
   - Konuşma ve yazma görevlerinin her biri için farklı deneme sınavlarındaki aday cevaplarının bir bölümü *Diller için Avrupa Ortak Başvuru Metni (D-AOBM)* dil düzeylerine göre kalibre edilmiş ayrıca dereceli puanlama anahtarı ile puanlanmıştır. Kalibre edilmiş aday performansları da modelin geliştirilmesi için gizlilik sözleşmesinin imzalanmasından sonra paylaşılacaktır.
2. Projede kullanılacak verilerin gizliliği ve güvenliği, KVKK (Kişisel Verilerin Korunması Kanunu) ve uluslararası veri güvenliği standartlarına uygun şekilde sağlanacaktır.
3. Pilot veriler, halihazırda var olan verilerdir. Türkçe dört becerili dil sınavı için gönüllülük esası ile adayların sınava başvurması sayesinde toplanmaktadır. Pilot veri paylaşımı, Gizlilik Anlaşması (NDA) imzalanmasının ardından gerçekleştirilecektir. NDA imzalanması ile verilerin yalnızca proje kapsamında kullanılması ve üçüncü taraflarla paylaşılmaması garanti altına alınmaktadır. Paylaşılacak veriler, anonimleştirilmiş ve kişisel bilgilerden arındırılmış şekilde teslim edilecektir. Pilot veri, yapay zekâ modelinin ön eğitiminde kullanılacaktır.
4. Sistemde iki farklı değerlendirme modülü yer almalıdır. Bunlar performansa ilişkin hem D-AOBM düzeyini hem de DPA puanını sağlamalıdır. Dolayısıyla geliştirilen sistem iki farklı türde değerlendirme yapmalıdır:
   a. Sınav görevlerine uygun olarak geliştirilen Dereceli Puanlama Anahtarı (DPA) ile uyumlu puanlama (0-25 arası puan değeri)
   b. Puanlama güvenirliğinin kontrolü için D-AOBM uyumluluğuna sahip sınıflandırma (A1-C2 seviyeleri)
5. Geliştirilecek çözüm altyapısında Türkçe dilini anlama ve Türkçe dilinde içerik üretme yeteneği en yüksek olan açık kaynak büyük dil modelleri (LLM) kullanılacaktır. Geliştirilecek çözümde, mevcut verilerden faydalanarak ince ayar (fine-tuning) yapılacak ve konuşma ve yazma değerlendirmelerinde doğrulama ve test aşamalarında 4-a maddesinde belirtilen değerlendirme için %80 ve üzeri doğruluk oranı sunacaktır. Konuşma ve yazma testlerinden elde edilen veriler yapay zekâ modelinin CEFR seviyelerine göre sınıflandırma problemini çözme performansı precision, recall, accuracy, F1 measure bakımından değerlendirilecek ve bu metrikler bakımından performans başarımı hem geçerleme hem de test işlemlerinde 4-b maddesindeki değerlendirme için en az %90 olarak beklenmektedir. Ardından modelin rubrik bazlı puan tahmini açısından değerlendirilecek ve uzman değerlendirmeleri ile karşılaştırılarak geçerlik doğrulama testleri yapılacaktır. Yapay zekâ modeli, uzman puanlamaları ile yüksek korelasyon (+0.80) göstermelidir. Modelin değerlendirme hatası (MAPE – Ortalama Mutlak Yüzde Hatası) %20'den düşük olmalıdır ve R-squared (R^2) değeri bakımından %80 olmalıdır. Geliştirilecek yapay zekâ modeli, Multi Faceted Rasch Model ile kestirimler yapılacak ve in-fit, out-fit indexlerine bakılacaktır. Geliştirilecek yapay

zekâ modelinde güvenilirlik için bu değerlerin 0,5 ile 1,5 arasında olması beklenecektir. Modelin güvenirliği bu ölçüm değerleri kullanılarak test edilecektir.

6. Türkçe için iyi yanıtları veren modeli belirlemek için karşılaştırmalı performans testleri (benchmark) yapılacak ve raporlanacaktır.
7. Farklı modellerin cevaplarını karşılaştırmak için test arayüzleri tasarlanacaktır.
8. Geliştirilecek çözüm ile anlam analizi, yazım kontrolü, bağlam değerlendirmesi, göreve uygunluk, içerik ve dil bilgisi kontrolü yapılacaktır. Tüm rubrikler regresyon değerlendirme metriklerine göre değerlendirilecektir. Yapay Zekâ modeli, uzman puanlamaları ile yüksek korelasyon (+0.80) göstermelidir. Modelin değerlendirme hatası (MAPE – Ortalama Mutlak Yüzde Hatası) %20'den düşük olmalıdır ve R-squared (R^2) değeri bakımından %80 olmalıdır. Konuşma testi veya konuşma tabanlı değerlendirme sistemlerinde, anlam analizi, yazım kontrolü, bağlam değerlendirmesi göreve uygunluk, içerik, akıcılık, sesletim ve dil bilgisi kontrolü gibi başlıklar, modelin verdiği yanıtların doğru, anlamlı ve tutarlı olmasını sağlamak için kritik öneme sahiptir.
   a. Anlam analizi, dil modelinin verilen soruya uygun, tutarlı ve anlamlı yanıtlar üretmesini sağlar.
   b. Semantik Benzerlik, modelin verdiği yanıt ile doğru cevap arasında anlam benzerliği olup olmadığını ölçer. Yanıtların içerik açısından doğru ve anlamlı olması gerekmektedir. Anlam analizi ve semantik benzerlik ile yanıtların anlamlı ve tutarlı olmasını sağlamak ve metnin veya yanıtın sorunun bağlamına uygun olması değerlendirilir.
   c. Yazım kontrolü, dildeki olası yazım hatalarını, imla yanlışlıklarını, harf hatalarını, eksik kelimeleri ya da yanlış kullanılan karakterleri tespit eder. Bu kontrol, konuşma testi sistemlerinin doğru ve profesyonel cevaplar sunması için önemlidir. Yazım kontrolünde, doğrudan yazım hataları, büyük/küçük harf hataları, noktalama hataları tespit edilmektedir. Dereceli Puanlama Anahtarları ile hatalı yazımın tespit edilmesi, cümlelerin anlaşılır ve profesyonel bir şekilde yazılması ve okuyucunun veya dinleyicinin anlam kaybını ölçmek hedeflenmektedir.
   d. Bağlam değerlendirmesi, bir konuşma testi veya yazılı metnin bağlamını doğru şekilde anlamak ve doğru şekilde yanıt vermek için kritik öneme sahiptir. Bağlam, dilin doğru anlaşılmasını sağlar; çünkü bir kelimenin anlamı, genellikle çevresindeki kelimelerle ilişkilidir. Bağlam değerlendirmesi, modelin gerçek bir konuşma veya test ortamındaki soruları doğru şekilde anlamasını sağlar. Bağlam değerlendirmesi, ile konuşmaların ya da yazılı metinlerin bağlamını doğru bir şekilde analiz etmek, önceki sorular veya yanıtlarla tutarlı bir şekilde yeni soruları ele almak, dilin farklı bağlamlarda doğru kullanılmasını sağlamak amaçlanmaktadır. Bağlam değerlendirmesi ile adayların konuşmaları ya da yazılı metinlerin bağlamını doğru bir şekilde analiz etmek, önceki sorular veya yanıtlarla tutarlı bir şekilde yeni soruları ele almak, dilin farklı bağlamlarda doğru kullanılmasını ölçmek amaçlanmaktadır.
   e. Dil bilgisi kontrolü ile cümle yapısı, zaman uyumu, biçim, birimlerin doğru kullanımı, bağdaşıklık ögelerinin uygun kullanımı analiz edilmektedir. Dil bilgisi kontrolü ile dil bilgisi kurallarına uygun, doğru yapıda cümleler oluşturmak, yanlış dil bilgisel yapıları, sözdizimsel hataları ve zaman uyumsuzluklarını tespit etmek ve kullanıcının verdiği yanıtları dil bilgisi açısından doğru hale getirmek amaçlanmaktadır.
   f. Konuşma performanslarının değerlendirilmesinde akıcılık ve sesletim de değerlendirilecektir. Akıcılık kontrolü ile konuşmadaki duraklamalar, sık tekrarlanan sözcükler ve ifadeler belirlenecektir. Sesletim kontrolünde adayların telaffuzlarındaki hatalar belirlenecektir. Her bir başlık, modelin sözlü veya yazılı dildeki kaliteyi değerlendirmeye yönelik farklı yönleri ele alır. Konuşma ve yazma testi için geliştirilen bir çözümde, anlam analizi, yazım kontrolü, bağlam değerlendirmesi ve dil bilgisi kontrolü, modelin verdiği yanıtların doğruluğu, anlamlılığı ve dilsel kalitesini arttırmak için kritik öneme sahiptir. Bu dört başlık, doğru ve anlamlı yanıtların üretilmesinde birbirini tamamlayan süreçlerdir ve her biri modelin performansını geliştirmeye yönelik önemli adımlardır. Modelin bu alanlarda yetkin olması, kullanıcı deneyimini iyileştirir ve

daha tutarlı, profesyonel ve anlamlı sonuçlar elde edilmesini sağlar. Modelin eğitilmesi ve beklenen kontroller gizlilik sözleşmesinin imzalanmasından sonra üstlenici ile detaylı olarak paylaşılacaktır.

9. Geliştirilecek olan model en az bir gerçek sınavın değerlendirmesinde kullanılacaktır. Multi Faceted Rasch Model ile kestirimler yapılacak ve in-fit, out-fit indexlerine bakılacaktır. Bu değerlerin 0,5 ile 1,5 arasında değerler olması beklenecektir.
10. Değerlendirme sonuçları, Microsoft SQL Server ilişkisel veri tabanına aktarılabilecek şekilde tasarlanacaktır.
11. Geliştirilecek çözüm, kurum içi kapalı devre (on-prem) çalışacak şekilde tasarlanacaktır.
12. Geliştirilecek çözüm, farklı dil seviyelerinde testlere tabi tutulacak ve test sonuçları raporlanacaktır. Geliştirilen Yapay Zekâ modeli, Multi Faceted Rasch Model ile kestirimler yapılacak ve in-fit, out-fit indexlerine bakılacaktır. Geliştirilecek Yapay zekâ modelinde güvenilirlik için bu değerlerin 0,5 ile 1,5 arasında olması beklenecektir. Benzer biçimde Pearson korelasyonu, Yapay Zekâ ve uzman değerlendirici arasındaki korelasyon değerinin doğrulama ve test aşamalarında +0.80 ve üzeri olması beklenmektedir. Yapay zekâ değerlendirme performansının doğrulama ve test aşamalarında R-kare (R^2) değerinin 0.80 ve üzeri olması beklenmektedir. Bu metriklerin raporlanması beklenmektedir.
13. Geliştirilecek çözüm, OWASP Top 10 standartlarına uygun olarak güvenlik testlerinden geçirilecek ve bu standartlara uygun bir şekilde hiçbir güvenlik açığı içermeyecektir.
14. Proje kapsamında geliştirilecek sistem log altyapısı içerecektir. Bu log sistemi İDARE (ÖSYM)'nin log politikalarına uygun olarak tasarlanmalıdır. Loglama ve diğer proje gereksinimi için ilişkisel veri tabanı olarak MSSQL kullanılacaktır. Ancak farklı bir veri tabanı gereksinimi olursa MSSQL veri tabanına aktarım sorgularının hazırlanması sağlanacaktır.
15. Geliştirilecek çözümün çalışma performansı ve kaynak kullanımı, gerçek zamanlı olarak izlenebilir bir arayüz üzerinden takip edilecektir. Sistemdeki olası sorunlar ve performans problemleri bu arayüzde görüntülenebilir olacaktır. Sistem yeni ölçme ve değerlendirme gereksinimlerine uygun olarak rubrik güncellemelerine uyumlu olması beklenmektedir.
16. Geliştirilecek çözüm, açık kaynak teknolojiler kullanılarak geliştirilecek ve herhangi bir lisans bağımlılığı taşımayacaktır. Sistemlerin herhangi bir aşamasında lisans maliyetleri oluşması durumunda bu maliyet YÜKLENİCİ (Proje Yürütücüsü Kuruluş) tarafından sağlanacaktır. Geliştirilecek çözümün çalışacağı donanım ve altyapı özellikleri YÜKLENİCİ tarafından belirlenecek ve sözleşmenin imzalanma tarihinden sonra 3 ay içinde İDARE'ye iletilecektir.
17. Proje takvimine uygun olarak gerekli donanım ve altyapı İDARE tarafından sağlanacaktır.
18. Projede kullanılabilecek olan donanım gereksinimi için Intel(R) Xeon(R) Silver 4210 CPU işlemci, Nvidia RTX6000 24 GB VRAM ekran kartı, 256 GB RAM bellekli bir sunucu tahsis edilecektir. Proje için bu donanım yeterli olmadığı durumda İDARE tarafından Intel(R) Xeon(R) Gold 6430 2.10 GHz 64 GB RAM Nvidia RTX A6000 48 GB VRAM bir cihaz temini sağlanabilecektir.
19. Geliştirilecek sistem, dış uygulamalarla entegrasyon sağlayabilecek şekilde API servisleri sunmalıdır. API servislerinin test edilmesi için ayrı bir test arayüzü (test endpoint) sağlanmalıdır. Bu test arayüzü, kullanıcıların API'yi güvenli bir şekilde test etmelerini ve entegrasyon süreçlerini doğrulamalarını kolaylaştıracaktır. Test arayüzleri, API'de yapılacak olası hata durumlarını simüle etmek ve geliştirme sürecinde daha fazla doğrulama yapılmasını sağlamak için kullanılacaktır. API üzerinden yapılacak işlemler için kimlik doğrulama ve yetkilendirme amaçlı token ile güvenlik sağlanmalıdır.

    Modelin sürekli olarak eğitilmesini sağlamak için daha sonra gelecek verilerle eğitilmesine açık halde geliştirilmelidir. İDARE, ihtiyaç halinde YÜKLENİCİ'den toplantı, rapor ve kurum lokasyonunda çalışmasını talep edebilecektir. Pilot veriler, İDARE tarafından proje kapsamında Gizlilik Sözleşmesi ile kurum dışında çalışabilmeleri için YÜKLENİCİ'ye sağlanacaktır. YÜKLENİCİ, pilot verilerle yapılacak

model geliştirme süreci sonrasında, yeni verilerle yapılacak kurulumları ve çalışmaları İDARE güvenlik politikaları gereğince fiziksel olarak kurum lokasyonu olan Ankara/Bilkent ÖSYM Başkanlığında gerçekleştirecektir.

Dokümanlar

Proje kapsamında aşağıda belirtilen raporlar hazırlanacak ve İDARE ile proje planında belirtilen tarihlerde paylaşılacaktır. İDARE gerekli görmesi durumunda ek raporlar talep edebilecektir.

- Üst Düzey Sistem Tasarımı Dokümanı

  Geliştirilecek sistemin genel yapısı ve işleyişi tanımlanmalı, sistem mimarisi detaylı şekilde açıklanmalıdır. Ana bileşenlerin işlevleri ve bunlar arasındaki etkileşimler belirlenmeli, veri akışı ve süreçler netleştirilmelidir. Performans, ölçeklenebilirlik ve güvenlik gibi teknik gereksinimler ele alınmalı ve riskler ile varsayımlar ortaya konulmalıdır. Bu doküman, sistemin başarılı şekilde tasarlanıp uygulanabilmesi için yol gösterici bir rehber olmalıdır.

- Analiz ve Tasarım Dokümanı

  Sistemin mevcut durumu analiz edilmeli ve ihtiyaç duyulan özellikler belirlenmelidir. Teknik gereksinimlerden tasarım ilkelerine kadar olan süreç açık bir şekilde sunulmalı, kullanıcı deneyimini iyileştirmek için arayüz tasarımı yapılmalıdır. Ayrıca, veri tabanı yapısı ve sistemin mimari çözümü detaylandırılmalıdır.

- Veri Analiz Raporu

  Bu raporda, analiz edilen veri setleri ve kullanılan yöntemler açıklanmalı, veri setinin temel özellikleri tanımlanmalıdır. Bulgular, grafikler ve tablolarla görselleştirilmeli ve analizden elde edilen sonuçlar yorumlanmalıdır. Stratejik kararlar için öneriler geliştirilmeli ve sonraki adımlar belirlenmelidir. Rapor, veri temelli karar alma süreçlerine rehberlik etmelidir.

- Test Raporları

  Test süreçleri ve sonuçları detaylı şekilde sunulmalıdır. Test amaçları ve kapsamı netleştirilmeli, uygulanan test senaryoları ve kullanılan yöntemler, performans ve hedeflenen çıktılarla uyum açıklanmalıdır.