

Açık Uçlu Maddeleri Otomatik Puanlamak Ne Kadar Güvenilirdir: Türk Dilinde Bir Uygulama *

İbrahim UYSAL **

Nuri DOĞAN ***

Öz

Özellikle geniş ölçekli testlerde açık uçlu maddelerin kullanılması açık uçlu maddeleri puanlama konusunda zorluk ortaya çıkarmıştır. Ancak açık uçlu maddelerin otomatik puanlanmasına dayalı yaklaşımla bu zorluğun üstesinden gelinebilmektedir. Araştırmanın amacı açık uçlu maddeleri otomatik puanlayarak elde edilen verilerin güvenilirliğini incelemektir. Alt amaçlardan birisi otomatik puanlamada makine öğrenmesine dayalı farklı algoritmaları (destek vektör makineleri, lojistik regresyon, çok terimli sade bayes, kısa-uzun süreli bellek ve iki yönlü kısa-uzun süreli bellek) karşılaştırmaktır. Diğer alt amaç ise otomatik puanlamaya özgü sistemin test edilmesinde (%33, %20 ve %10) kullanılan veri oranının farklılaştırılmasıyla otomatik puanlamanın güvenilirliğinin değişimini incelemektir. Otomatik puanlamaların güvenilirliği incelenirken gerçek puanlayıcılardan elde edilen verilerin güvenilirliği ile karşılaştırma yapılmıştır. Türk dilindeki açık uçlu maddelerin ilk otomatik puanlama denemesini gösteren bu çalışmada Millî Eğitim Bakanlığı tarafından uygulanan Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) programı Türkçe test verileri kullanılmıştır. Sistemin test edilmesinde çapraz geçerlikten yararlanılmıştır. Güvenirliği gösterecek uyum katsayılarına yönelik olarak uyum yüzdesi, otomatik puanlama araştırmalarında sıklıkla kullanılan karesel ağırlıklı Kappa ve verinin kategorilere dağılımındaki dengesizlik sorunundan etkilenmeyen Gwet'in AC1 katsayısı kullanılmıştır. Araştırma sonuçları otomatik puanlama algoritmalarından faydalanılabileceğini göstermiştir. Otomatik puanlamada kullanılabilecek en iyi algoritmanın iki yönlü kısa-uzun süreli bellek olduğu bulunmuştur. Kısa-uzun süreli bellek ve çok terimli sade bayes algoritmaları; destek vektör makineleri, lojistik regresyon ve iki yönlü kısa-uzun süreli bellek algoritmalarından daha düşük performans sergilemiştir. Otomatik puanlamada %33 test veri oranında uyum katsayılarının %10 ve %20 test veri oranlarına göre biraz daha düşük olduğu ancak istenilen aralıkta olduğu belirlenmiştir.

Anahtar Kelimeler: Açık uçlu madde, makine öğrenme algoritmaları, otomatik puanlama, puanlayıcılar arası güvenilirlik, uyum katsayıları.

GİRİŞ

Bireyler yaşamları boyunca çok sayıda testle karşılaşır. Testler, bireylerin bilgi, beceri ve yetenekleri arasındaki farkları gösterir. Böylece bireyler hakkında kararlar alabilir (Geisinger ve Usher-Tate, 2016). Son yıllarda testlerde birden fazla madde formatı kullanımı daha çok rağbet görmektedir. Karma test olarak adlandırılan bu yaklaşımda aşına olunan çoktan seçmeli maddelerin yanında yanıtı sınırlandırılmış ya da sınırlandırılmamış açık uçlu maddeler kullanılmaktadır. Çoktan seçmeli maddelerde bireyler bir problemle ilgili bir doğru, birden fazla yanlış cevapla karşılaşmaktadır. Yanıtı sınırlandırılmış açık uçlu maddelerde bireyler sorulara birkaç kelime, cümle ya da paragrafla cevap verirken yanıtı sınırlandırılmamış maddelerde bireyler sorulara istedikleri uzunlukta cevap vermektedir (Downing, 2009). Madde türlerinin bir arada kullanımı her bir formatın kısıtlı yönlerinin ortadan kaldırılmasını sağlamaktadır (Messick, 1993). Örneğin testlerde sadece çoktan seçmeli maddelerin kullanımı öğretime etki etmekte, bireyler çoktan seçmeli testlere yönelik çalışma yapmaktadır. Bu durum özgün, eleştirel ve üst düzey düşünme becerilerini kısıtlayabilmektedir. Ancak açık uçlu maddelerin kullanımı bu kısıtlılığı ortadan kaldırabilmektedir.

* Bu makale Nuri DOĞAN danışmanlığında İbrahim Uysal tarafından 2019 yılında tamamlanan “Açık Uçlu Maddelerde Otomatik Puanlamanın Güvenirliği ve Test Eşitleme Hatalarına Etkisi” isimli doktora tezinin bir bölümünden türetilmiştir.

** Dr. Arş. Gör., Bolu Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi, Bolu-Türkiye, e-posta: ibrahimuysal06@gmail.com, ORCID ID: 0000-0002-6767-0362

*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

Bu makaleye atıfta bulunmak için:

Uysal, İ., & Doğan, N. (2021). Açık uçlu maddeleri otomatik puanlamak ne kadar güvenilir? Türk dilinde bir uygulama. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 12(1), 28-54. doi: 10.21031/epod.817396

Geliş Tarihi: 28.10.2020

Kabul Tarihi: 14.02.2021

Açık uçlu maddelerin uygulanması zordur ve puanlaması uzun zaman ve çaba gerektirmektedir (Gierl, Latifi, Lai, Boulais ve Champlain, 2014). Puanlanacak birey ve açık uçlu maddelerinin sayısı arttıkça da daha fazla puanlayıcıya ihtiyaç duyulmaktadır. Bunun yanı sıra çok sayıda puanlayıcının puanlama konusunda eğitilmesi gerekmektedir. Puanlayıcı sayısı arttıkça puanlamadaki öznellik güvenilirliği de düşürmektedir (Ebel ve Frisbie, 1991; Hagge, 2010). Geniş kapsamlı test uygulamaları düşünüldüğünde açık uçlu maddelerin puanlanmasının sınava dair maliyeti de oldukça arttıracak göz önünde tutulmalıdır (Cohen, Ben-Simon ve Hovav, 2003).

Otomatik puanlama, alanyazında ve test uygulayıcıları tarafından son yıllarda oldukça popülerlik kazanan bir yaklaşımdır. Otomatik puanlamada bilgisayar destekli analizlerle yazılı bir metin otomatik olarak değerlendirilmektedir (Shermis, 2010). Otomatik madde puanlama fikri puanlama zorluğunu azaltmak üzere yaklaşık 50 yıl önce bir ortaokul öğretmeni olan Page (1966) tarafından ortaya atılmıştır (Ramineni ve Williamson, 2013). Page (1966) Project Essay Grade (PEG) programının geliştiricisidir. Geliştirilen bu ilk programda kompozisyonlar puanlanırken kelime uzunluğu, kompozisyon uzunluğu, virgül ve edat sayısı, çoğunlukla kullanılmayan kelimeler üzerinden puanları tahmin etme yoluna gidilmiştir (Wang ve Brown, 2007).

Otomatik puanlama sistemleri kısa cevaplı maddelerden kompozisyonlara kadar farklı uzunluktaki cevaplar üzerinde çalışabilmektedir (Gierl vd., 2014). Yani otomatik puanlama yanıtı sınırlandırılmış ve sınırlandırılmamış açık uçlu maddeleri puanlayabilmektedir. Şu an okullarda yer alan yazma becerisi görevlerinin %90'ının otomatik madde puanlama sistemleri ile değerlendirilebileceği belirtilmektedir (Shermis ve Burnstein, 2003). Sınıf içi uygulamaların yanı sıra otomatik puanlama sistemleri ile geniş ölçekli testlerde de puanlama yapılabilmektedir. Uluslararası GMAT (Graduate Management Admission Test), TOEFL (Test of English as a Foreign Language) ve GRE (Graduate Record Examination) gibi geniş ölçekli testlerde bu yaklaşımdan faydalanılmaktadır. Otomatik puanlama sistemlerinin en önemli avantajı bireylere hemen dönüt verebilmesidir (Gierl vd., 2014). Otomatik puanlama işleminde puanlama özellikleri bilgisayara el ile tanımlanabileceği gibi (örneğin Page'in ilk çalışmaları) puanlama davranışları bilgisayara gerçek puanlayıcıların yaptığı puanlamalardan otomatik olarak haritalandırılabilir. Otomatik puanlamada kullanılan ve puanlama özelliklerini öğrenen denetlenen (supervised) makine öğrenme algoritmalarında genellikle dört adımdan oluşan bir süreçten yararlanılmaktadır (Powers, 2015). Bu adımlar; 1) bilgisayarı eğitmek üzere nitelikli olduğu bilinen bir puanlamanın metine dayalı bir kitaplık ile tanımlanması, 2) eğitim verilerindeki yazılardan çeşitli özelliklerin çıkarılması, 3) yazının tüm nitelikleriyle ilgili bir model geliştirilmesi, 4) değerlendirilmemiş yazılara kurulan modeli kullanarak puan ataması yapılması ya da kategorilere ayrılmasıdır. Denetlenen makine öğrenmesi sürecinde kullanılabilecek farklı algoritmalar bulunmaktadır. Bu çalışmada klasik makine öğrenmesini temel alan üç algoritma (lojistik regresyon [logistic regression - LR], çok terimli sade bayes [multinomial naive bayes - MNB], destek vektör makineleri [support vector machine - SVM]) ve yapay sinir ağlarını temel alan iki derin öğrenme algoritması (kısa-uzun süreli bellek [long-short term memory - LSTM], iki yönlü kısa-uzun süreli bellek [bidirectional long-short term memory - BLSTM]) kullanılmıştır. Berg ve Gopinathan (2017), Gierl vd. (2014), Jang, Kang, Noh, Kim, Sung ve Seong (2014) ve Lilja (2018) kaynaklarından bu algoritmalarla ilgili detaylı bilgi edinilebilir.

Açık uçlu maddelerde otomatik puanlama sistemlerinin kullanılması kaynakların verimli kullanılmasını sağlayacak, puanlama süresini azaltacak ve iş gücü kaybını önleyecektir (Attali ve Burstein, 2006; Chen, Xu ve He, 2014). Bu sistemin kullanımı çok sayıda puanlayıcı kullanma zorunluluğunu ortadan kaldıracaktır. Böylece açık uçlu sorular içeren geniş ölçekli testler için büyük kolaylık sağlanacaktır. Bu nedenle araştırma önem taşımaktadır. Bazı durumlarda karşılaşılan puanlama yanlışlığının otomatik puanlama ile önüne geçilebilecektir. Farklı eğitim alan puanlayıcılardan kaynaklanan, güvenilirlikle ilgili sorunlar giderilebilecek ve genellenebilirlik konusunun da üstesinden gelinebilecektir (Adesiji, Agbonifo, Adesuyi ve Olabode, 2016). Ancak otomatik puanlama sistemlerinin kullanılabilmesi elde edilen puanların mümkün olduğunca gerçek puanlayıcılara benzemesine ve güvenirliliğinin düşmemesine bağlıdır. Gerçek puanlayıcılar, otomatik puanlama sistemleri için bir önemli bir kriter konumundadır (Cohen, Levi ve Ben-Simon, 2018). Güvenirliliği düşük, gerçek puanlayıcılarla uyum sağlamayan otomatik puanlama sonuçları bireyler

hakkında yanlış kararlar verilmesine neden olabilir. Bu açıdan yaklaşıldığında araştırma gerçek puanlayıcılar ve otomatik puanlama arasında karşılaştırma yaparak sistemin kullanımını değerlendirdiğinden önemlidir. Otomatik puanlama koşulları değiştiğinde (örn. sistemin eğitilmesinde ve test edilmesinde kullanılan veri sayısı) otomatik puanlama ve gerçek puanlayıcılar arasındaki uyumda değişikliklerin olması muhtemeldir. Bu doğrultuda otomatik puanlamanın yapılabileceği puanların yeterince güvenilir olacağı veri miktarının belirlenmesi gerekmektedir. Bu durum araştırmanın önemini arttırmaktadır. Araştırmanın amacı açık uçlu maddeleri otomatik puanlayarak elde edilen verilerin güvenilirliğini incelemektir. Alt amaçlardan birisi otomatik puanlamada makine öğrenmesine dayalı farklı algoritmaları (destek vektör makineleri, lojistik regresyon, çok terimli sade bayes, kısa-uzun süreli bellek ve iki yönlü kısa-uzun süreli bellek) karşılaştırmaktır. Diğer alt amaç ise otomatik puanlamaya özgü sistemin test edilmesinde (%33, %20 ve %10) kullanılan veri oranının farklılaştırılmasıyla otomatik puanlamanın güvenilirliğinin değişimini incelemektir. Sonuçların uygun olduğu koşulların belirlenmesi otomatik puanlama çalışmaları gerçekleştirilmesinin önünü açacaktır.

Alanyazındaki araştırmalarda otomatik puanlama işlemlerinin Türk dili dışındaki dillerde gerçekleştirildiği görülmektedir. Bunlar arasından makine öğrenmesinde farklı algoritmaları kullanan araştırmalara Gierl vd. (2014), Adesiji vd. (2016), Taghipour ve Tou Ng (2016) örnek olarak verilebilir. Gierl vd. (2014) otomatik puanlamada denetlenen makine öğrenmesine dayalı SVM algoritmasından, Adesiji vd. (2016) otomatik puanlamada özelliklerin elle tanımlandığı üç modülden oluşan bir yapıdan ve Taghipour ve Tou Ng (2016) denetlenen makine öğrenmesine dayalı üç yinelenen sinir ağı algoritmasından (temel yineleyen birimler, aralıklı yineleyen birimler ve LSTM birimleri) yararlanmıştır. Dillerin yapılarının birbirinden farklılık göstermesi otomatik puanlama üzerinde etkili olabilecek bir faktördür. Bu nedenle Türk dilinde otomatik puanlamanın araştırılması gerekmektedir. Türkçe'nin içerisinde yer aldığı Altay dil ailesi ses uyumu, bitişiklik, son ek, cümle sıralaması, tamlayanın tamlanandan önce gelmesi, sıfat tamlamalarında tamlayan ile tamlanan arasında hal, cinsiyet ve sayı bakımından fark bulunmaması, çokluk bildiren sayılardan sonra gelen isimlerin çokluk eki almaması, kelimelerde cinsiyetin belli olmaması gibi özellikler barındırmaktadır. Bu özelliklerin diğer dil ailelerinden farklılaşması Altay dil ailesindeki otomatik puanlama araştırmalarını gözden geçirmeyi gerektirmektedir. Jang vd. (2014) Kore dilinde, Ishioka ve Kameda (2006) Japon dili üzerinde araştırma gerçekleştirmiştir. Bahsedilen iki araştırmada özelliklerin elle tanımlandığı algoritmalar kullanılmıştır. Araştırma Türk dili üzerinde ilk otomatik puanlama denemesi özelliğiyle özgünlük taşımaktadır.

YÖNTEM

Araştırmada gerçek puanlayıcılar ve otomatik puanlama algoritmalarına ait puanların güvenilirliği karşılaştırıldığından ilişkisel bir araştırma gerçekleştirilmiştir. Creswell (2012) ilişkisel araştırmalarla bir değişkendeki farklılığın diğer değişkeni nasıl etkilediğini görmenin mümkün olduğunu belirtmektedir.

Araştırmada Kullanılan Yazılımın Geliştirilmesi

Araştırmada, araştırmacının da içerisinde bulunduğu bir ekip tarafından geliştirilen otomatik puanlama yazılımı kullanılmıştır. Yazılım geliştirilirken Milli Eğitim Bakanlığı (MEB) tarafından uygulanan “Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme Projesi”nde Türkçe testine ait yanıtı sınırlandırılmış açık uçlu maddeler kullanılmıştır. Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme Projesi Türkçe testi bu araştırmada kullanılan testlerden (ABİDE) bağımsızdır. Bu test beşinci sınıf öğrencilerine yönelik olup 5 açık uçlu madde içermektedir. Yazılım hazırlanırken 0-1 ve 0-1-2 şeklinde puanlanan 5 yanıtı sınırlandırılmış açık uçlu maddeden yararlanılmıştır. Bu testte tüm öğrenci cevapları iki puanlayıcı tarafından derecelendirilmiş ve gerektiğinde üst puanlayıcıya ulaşarak nihai bir puan elde edilmiştir. Puanlama işlemlerinde dereceli puanlama anahtarlarından yararlanılmıştır.

Yazılımın hazırlanmasında kullanılan maddelerden ikisine ait sonuçlar örnek olarak burada sunulmuştur. Sonucu sunulan iki kategorili maddeye (madde 16) ve dereceli puanlama anahtarına ek A'da, üç kategorili maddeye (madde 20) ve dereceli puanlama anahtarına ek B'de yer verilmektedir. Türkçe testinde yer alan 16. madde için 303 ve 20. madde için 637 öğrenciye ait veri kullanılmıştır. Madde 20 üç kategorili puanlandığı için daha fazla veri üzerinde deneme yapılmıştır. Linux işletim sistemi üzerinde Python programı kullanılarak otomatik puanlama sistemi oluşturulmuş ve denemeler yapılmıştır. Otomatik puanlamada SVM, LR, MNB, LSTM ve BLSTM olmak üzere beş algoritma kullanılmıştır. Yazılımda Keras ve scikit-learn isimli iki kütüphaneden yararlanılmıştır. Verinin %90'ı sistemi eğitmek, %10'u ise sistemi test etmek amacıyla kullanılmıştır. Çapraz geçerlik ile rastgele örnekleme yöntemine başvurulmuştur. 10 kat çapraz geçerleme ile test verileri ve eğitim verileri 10 kez birbirinden farklı olacak şekilde değiştirilerek veri sayısı kadar otomatik puanlama yapılmıştır. Böylece 303 veri üzerinde yapılan denemede 303 puanlanma sonucuna, 637 veri üzerinde yapılan denemede 637 puanlama sonucuna ulaşılmıştır. Otomatik puanlama ile gerçek puanlayıcıların nihai puanları arasındaki uyum incelenerek yazılımın kullanılabilirliği araştırılmıştır. Tablo 1'de ikili (0-1) puanlanan madde 16 ve üçlü (0-1-2) puanlanan madde 20'ye ait sonuçlara yer verilmektedir.

Tablo 1. Yazılım Oluşturulurken Elde Edilen Uyum Yüzdeleri

	Veri Sayısı	Kategori Sayısı	SVM (%)	LR (%)	MNB (%)	LSTM (%)	BLSTM (%)
Madde 16	303	2	98.0	98.3	96.1	99.0	99.0
Madde 20	637	3	85.5	82.4	75.1	87.3	88.7

Not: Uyum yüzdelерinin %80'in üzerinde olması kabul edilebilir bir uyumu göstermektedir (Hartmann, 1977).

Tablo 1 incelendiğinde madde 16 için elde edilen uyum yüzdelерinin oldukça yüksek olduğu görülmektedir. 16. madde için en yüksek uyum yüzdesini gösteren algoritmalar LSTM ve BLSTM'dir. Madde 20 için elde edilen uyum yüzdelерinin yeterli düzeyde olduğu belirlenmiştir. Madde 20'de en iyi uyumu gösteren algoritmanın BLSTM olduğu görülmektedir. Elde edilen sonuçlar oluşturulan sistemin yapılandırılmış cevap maddelerini puanlamada yeterli olacağını göstermiştir. Böylece araştırma kapsamındaki ABİDE veri setleri için otomatik puanlama işlemine geçilmiştir.

Araştırmanın Veri Kaynağı

Araştırmanın veri kaynağını Türkiye'de 2016 yılında Millî Eğitim Bakanlığı (MEB) tarafından uygulanan Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) 8. sınıflar araştırması oluşturmaktadır. Öğrencilerin üst düzey düşünme becerilerinin incelenmesinin amaçlandığı testlerde çoktan seçmeli ve yanıt sınırlandırılmış açık uçlu maddeler birlikte yer almaktadır. Araştırma A₁ ve B₁ kitapçığı Türkçe testlerinde yer alan yanıt sınırlandırılmış açık uçlu maddeler üzerinde yürütülmüştür. A₁ testinde dokuz ve B₁ testinde 10 madde açık uçludur. A₁ ve B₁ testlerindeki beş açık uçlu madde ise ortaktır. Açık uçlu maddeler 0-1 ve 0-1-2 şeklinde puanlanmaktadır. Açık uçlu maddelerin puanlama işlemi iki gerçek puanlayıcı tarafından yapılmıştır. Puanlar arasında uyum olmadığında ise cevap bir üst puanlayıcıya gönderilmiştir. Böylece nihai puanlar elde edilmiştir. Puanlama yapılırken dereceli puanlama anahtarlarından yararlanmıştır. A₁ ve B₁ kitapçığında yer alan açık uçlu maddelerin Cramer's V katsayılarının sırasıyla .83 ile .98 ve .87 ile .99 aralığında değiştiği açıklanmıştır. Katsayıların .80'in üzerinde olmasının puanlayıcıların tutarlılığının yüksek olduğunu gösterdiği belirtilmektedir (MEB, 2017a; MEB, 2017b). ABİDE testinden örnek maddelere ve dereceli puanlama anahtarlarına ek C ve ek D'de yer verilmiştir.

Verinin Bilgisayar Ortamına Aktarımı

Öncelikle MEB'den yukarıda özellikleri açıklanan veriler talep edilmiştir. Bu talebe istinaden veriler içerisinden seçkisiz şekilde belirlenen 1000 veri araştırmacılarla paylaşılmıştır. Veri içerisinde iki farklı puanlayıcı grubuna ve nihai puanlara ait puan matrisleri ve jpeg formatında öğrenci cevapları bulunmaktadır. Öğrenci cevap kâğıtları bilgisayar ortamına elle (manuel) girilmiştir. Bu durumun nedeni öğrenci yazılarının okunmasının zor olması ve bitişik el yazısı kullanımı nedeniyle optik karakter tanıma sistemlerinden (OCR) yeterince destek alınamamasıdır. Ayrıca OCR programlarından kaynaklanacak hataları bertaraf etmektir. Elle girilen verilerin öğrenci cevaplarıyla tamamen eşleşmesi için veriler lisans öğrencilerinden oluşan bir çalışma grubu tarafından kontrol edilerek hatalar giderilmiştir. Öğrenci cevapları doğrudan aktarılmış olup herhangi bir düzeltmeye tabi tutulmamıştır.

Verilerin Analizi

Araştırma verileri analiz edilmeden önce MEB'den alınan 1000 öğrenciye ait veri incelenmiştir. Açık uçlu maddelerden alınan puanların kategorilere dengeli dağılımı temel alınarak veri girişi yapılmıştır. Bu işlem veride açık uçlu maddelere ilişkin yaygınlık (kategorilere dağılımda dengesizlik [prevalence]) probleminin mümkün olduğunca önüne geçmek için gerçekleştirilmiştir. A₁ kitapçığı için 9, B₁ kitapçığı için 10 açık uçlu maddenin tamamı dikkate alınarak A₁ kitapçığından 697 ve B₁ kitapçığından 701 veri girişi yapılmıştır. Ardından testteki açık uçlu maddelerin yarısına veya yarısından fazlasına cevap veren öğrenciler seçilmiştir. Bu işlem sonrasında açık uçlu maddelerin her birisi için kayıp veri oranı hesaplanmıştır. Kayıp veri oranı %5'in altında kalacak şekilde veri temizlenmiştir. Bu işlem otomatik puanlamada uyum katsayılarının normalden yüksek çıkmasının önlenmesi amacıyla gerçekleştirilmiştir. Veriler temizlenirken kategorilere dağılım dikkate alınmıştır. Bazı kategorilerde az sayıda veri bulunduğu için mümkün olduğunca bu kategorilerde puan alan bireylerin araştırma kapsamından çıkarılmamasına dikkat edilmiştir. Yukarıda belirtilen kriterler dikkate alınarak A₁ kitapçığından 84 ve B₁ kitapçığından ise 96 kişiye ait veri temizlenmiştir. Ardından gerçek puanlayıcı grubu 1 ve gerçek puanlayıcı grubu 2'nin öğrencilere verdiği puanlar incelenmiştir. Burada karşılaşılan kayıp puanlar nedeniyle bir grup öğrenci de araştırma kapsamı dışına çıkarılmıştır. Bu yönde sırasıyla A₁ ve B₁ kitapçığından toplam 6 kişi çıkarılmıştır. Son olarak çoktan seçmeli maddelerdeki kayıp veri sayısı değerlendirilmiş testteki toplam madde sayısının yarısından ve çoktan seçmeli maddelerin yarısından fazlasına cevap vermeyen öğrenciler araştırma kapsamından çıkarılmış ve kayıp veri oranının %5'in altında kalması sağlanmıştır. Bu yönde A₁ kitapçığından veri çıkarılmamış olup B₁ kitapçığından 15 kişi araştırma kapsamı dışında bırakılmıştır. Son durumda A₁ kitapçığından 90 kişi, B₁ kitapçığından ise 117 kişi çıkarılmıştır. Böylece veri hazırlık süreci tamamlanarak A₁ kitapçığından 607, B₁ kitapçığından 584 veri ile otomatik puanlama işlemine geçilmiştir.

ABİDE açık uçlu verilerinin otomatik puanlanması

Otomatik puanlama aşamasında nihai puanlardan bir kısmı kullanılarak otomatik puanlama sistemi eğitilmiştir. Bu şekilde otomatik puanlama sisteminin gerçek puanlayıcılardan nasıl puanlama yapıldığını öğrenmesi sağlanmış ve sisteme puanlama özellikleri haritalandırılmıştır. Ardından sistemin eğitilmesinde kullanılmayan veriler otomatik olarak puanlanmıştır. Yazılımda elle herhangi bir özellik tanımlaması yapılmamıştır. Sistemin eğitimi/test edilmesinde kullanılan veri oranı araştırmada etkisi incelenen bir faktördür. Test için kullanılan veri oranları %10, %20 ve %33 olarak belirlenmiştir. Dolayısıyla sistemin eğitilmesinde kullanılan veri oranı sırasıyla %90, %80 ve %67'dir. Bu değerler A₁ kitapçığı için 607 verinin sırasıyla 61, 121 ve 200'ünün sistemi test etmek amacıyla kullanıldığını; sırasıyla 546, 486 ve 407'sinin ise sistemi eğitmek amacıyla kullanıldığını göstermektedir. Benzer bir hesaplama B₁ kitapçığı için yapılabilir. Sonuçlar hesaplanırken %10 test veri oranı için 10 kat, %20 test veri oranı için 5 kat ve %33 test veri oranı için 3 kat çapraz geçerlik kullanılmıştır. Bu şekilde eğitim ve test verileri farklılaştırılarak A₁ kitapçığı için 607 verinin tümü, B₁ kitapçığı için 584 verinin tümü test verisi haline getirilmiştir. Araştırma sonuçları başka

çalışmalarla karşılaştırılırken veri oranından ziyade veri sayıları kullanılmalıdır. Oran ile sonuç belirtilmesinin nedeni çapraz geçerlik uygulaması ve anlaşılabilirliği arttırmaktır.

Otomatik puanlama sonuçlarını değerlendirmek için gerçek puanlayıcıların nihai puanları ile uyum hesaplanmıştır. Karşılaştırma imkânı sunması açısından gerçek puanlayıcı grubu 1 ve gerçek puanlayıcı grubu 2'nin nihai puanlarla olan uyumu da incelenmiştir. Her bir madde ayrı ayrı incelenmiştir.

Uyum katsayıları

Puanlayıcılar arası uyum incelenirken uyum yüzdesi (percentage of agreement), karesel ağırlıklı Kappa (quadratic weighted Kappa [QWK]) ve Gwet'in AC1 (Gwet's AC1) katsayılarından yararlanılmıştır. Aşağıda detaylı bilgilere yer verilmektedir.

Uyum Yüzdesi: Uyum yüzdesi basit ve hızlı bir şekilde hesaplanabilen anlaşılması ve yorumlanması kolay bir katsayıdır. Bu nedenle araştırmada yer verilmiştir. Bu yöntemde katılımcıların birinci ve ikinci puanlayıcıdan aldıkları puan dizileri karşılaştırılmakta, puanlayıcıların üzerinde tam olarak anlaşılabilirlik derecelendirme sayısının tüm derecelendirmelerin sayısına oranı hesaplanmakta ve sonuç yüzde cinsinden ifade edilmektedir. Elde edilen sonuçlar %0 ile %100 aralığında değişmektedir. Bu katsayı şans eseri oluşabilecek anlaşmaları hesaba katmadığından eleştirilmektedir. Çünkü bu durum uyumun olduğundan fazla bulunmasına yol açabilmektedir. Ayrıca puanlayıcılar arası anlaşmazlığı ele almamaktadır. Bu yöntem tüm ölçek düzeyleri (sınıflama, sıralama, eşit aralıklı ve oran) ve puan kategorisi sayısı iki ya da daha fazla olduğunda kullanılabilir (Araujo ve Born, 1985; Goodwin, 2001; Graham, Milanowski ve Miller, 2012; Meyer, 1999). Araştırmacılar kesin bir kural olmamakla beraber uyum yüzdesinin %80'in üzerinde olması gerektiği konusunda görüş birliğine sahiptir (Hartmann, 1977).

Karesel Ağırlıklı Kappa: Kappa katsayısı en sık kullanılan uyum katsayılarından birisidir. Kappa katsayısı puanlayıcılar arasındaki şans eseri anlaşma olasılığını dikkate alan bir uyum katsayısıdır. Fakat puanlayıcıların anlaşmama olasılığını dikkate almamaktadır. Bu nedenle Kappa katsayısı ağırlıklandırılmıştır. Kappa katsayısı ağırlıklandırılırken uyumsuzluğun derecesine göre ağırlıklar kullanılmaktadır. En sık kullanılan iki ağırlıklandırma tekniği doğrusal (linear) ve kareseldir (quadratic). Doğrusal ağırlıklandırmada ağırlıklar puanların standart sapması ile orantılı iken karesel ağırlıklandırmada ağırlıklar puanların standart sapmasının karesi (varyans) ile orantılıdır. Yorumlanması kolay olduğundan uygulamada karesel ağırlıklı Kappa (QWK) kullanımı oldukça fazladır. QWK otomatik puanlama araştırmalarında sıklıkla kullanılmaktadır. Bu nedenle araştırmada yer verilmiştir. İki ya da daha fazla puan kategorisi bulunduğu kullanılabilen bu katsayı puanlardan birisi diğerinden veya diğerlerinden çok daha fazla sayıda ise yanıltıcı bir şekilde düşük olabilmektedir. Bu durum alanyazında yaygınlık (prevalence) sorunu olarak tanımlanmaktadır ve Kappa katsayısıyla ilgili en çok raporlanan sorundur. Yaygınlığın yanı sıra yanlılık (bias) Kappa değeri üzerinde etkilidir. Yanlılık sorunu puanlayıcıların bir duruma ilişkin değerlendirmelerine ait frekanslar arasında farklılık olduğunda ortaya çıkmaktadır (Byrt, Bishop ve Carlin, 1993; Eugenio ve Glass, 2004). Karesel ağırlıklı Kappa otomatik puanlama sistemi puanları ile üzerinde karara varılmış gerçek puanlayıcı puanları arasındaki uyumu değerlendirmekte de kullanılabilir ve 0 ile 1 aralığında değişen değerler alır. 0 katsayısı puanlayıcılar arasında uyum olmadığını gösterirken, 1 katsayısı puanlayıcılar arasındaki tam uyumu göstermektedir. Değerlendiriciler arasında şans eseri ortaya çıkacak değerden az uyuma rastlandığında bu değer 0'ın altında düşebilmektedir (Altman, 1991; Brenner ve Kliebsch, 1996; Graham, Milanowski ve Miller, 2012; Preston ve Goodman, 2012; Sim ve Wright, 2005; Vanbelle, 2016). Landis ve Koch (1977) Kappa katsayısının yorumlanması için bir ölçüt belirtmiştir. Altman (1991) ise bu ölçütün uyarlamasını yapmıştır. Buna göre .20 altındaki değerler zayıf, .21 ile .40 aralığındaki değerler kayda değer, .41 ile .60 aralığındaki değerler orta, .61 ile .80 aralığındaki değerler iyi ve .81 ile 1.00 aralığındaki değerler çok iyi uyumu göstermektedir. Williamson, Xi ve Breyer (2012) gerçek puanlayıcılar ve otomatik puanlama sistemleri arasındaki uyumun .70'in üzerinde olmasını önermektedir. Karesel ağırlıklı Kappa değeri hesaplanırken Wang, Wei, Zhou ve

Huang (2018) ile Preston ve Goodman (2012) tarafından kullanılan denklemlerden yararlanılmıştır. Bu kaynaklardan detaylı bilgiye ulaşılabilir.

Gwet'in AC1 Katsayısı: Gwet'in AC1 katsayısı (Gwet, 2008) Cohen'in Kappa katsayısında karşılaşılan paradokslar doğrultusunda ortaya çıkmıştır. Verilerin kategorilere dağılımındaki çarpıklık (yaygınlık), puanlayıcılardan kaynaklanan yanlılık, puanlayıcıların duyarlılık (sensitivity) ve özgüllüğünün (specificity) farklılaşması Kappa değerinin puanlayıcılar arasındaki uyumu tespit etme yeteneğini düşürmektedir (Eugenio ve Glass, 2004; Gwet, 2008). AC1 katsayısı Kappa katsayısından her bir kategori için marjinal olasılık ortalamaları ile şans uyumunun beklenen oranı üzerinde yaptığı düzeltme ile ayrılmaktadır. Böylece Kappa değerine göre paradokslardan daha az etkilenmekte, kategoriler arasındaki çarpıklık yani kategoriler arasındaki değişkenliğe karşı daha kararlı olmaktadır (Hoek ve Scholman, 2017).

AC1 katsayısının kategorilerde dengesizlik ve simetri eksikliği bulunduğu puanlayıcılar arası uyumu tespit etme yeteneği fazladır (Shankar ve Bangdiwala, 2014). Gwet'in AC1 katsayısı kategorik veride ve herhangi bir sayıda puanlayıcı bulunması durumunda kullanılabilir (Wongpakaran, Wongpakaran, Wedding ve Gwet, 2013). AC1 katsayısı; uyum yüzdesinden daha düşük, Kappa katsayısından ise daha yüksek değerler almaktadır (Lacy, Watson, Riffe ve Lovejoy, 2015). Gwet'in AC1 katsayısı Landis ve Koch (1977) tarafından Kappa katsayısı için sunulan kriterler aracılığıyla yorumlanabilmektedir (Senay, Delisle, Raynauld, Morin ve Fernandes, 2015; Siriwardhana, Walters, Rait, Bazo-Alvarez ve Weerasinghe, 2018). Hoek ve Scholman (2017) araştırmacılara, araştırmalarında Kappa değeri ile birlikte AC1 değerinin kullanımını önermektedir. Bunun yanı sıra Haley (2007) AC1 katsayısının otomatik puanlama sistemlerini değerlendirmede iyi bir çözüm olduğunu belirtmektedir. Bu nedenlerle araştırmada bu katsayıya yer verilmiştir. Gwet'in AC1 katsayısını hesaplamak için kullanılan denklem Gwet (2016) kaynağından incelenebilir.

Uyum katsayıları yorumlanırken puanların yaygınlığı ve puanlayıcıların yanlılığı önem taşımaktadır. Bu nedenle yaygınlık ve yanlılık indeksleri hesaplanmaktadır. Byrt, Bishop ve Carlin (1993) Kappa katsayısının yanıltıcı olmaması için yaygınlık ve yanlılık indekslerinin de tartışılması gerektiğini belirtmektedir. Yaygınlık indeksi -1 ile 1 aralığında değişse de mutlak değeri kullanıldığından elde edilen katsayıların 1'e yaklaşmasının Kappa değerini düşüreceği belirtilebilir. Yanlılık indeksinin mutlak değeri ise 0 ile 1 aralığında değişmekte olup yanlılık katsayılarının yükselmesinin Kappa değerini yükselteceği belirtilebilir (Byrt, Bishop ve Carlin, 1993). A₁ ve B₁ kitapçıklarında yer alan tüm yapılandırılmış cevap maddelerinin yaygınlık ve yanlılık katsayıları incelenmiştir. A₁ kitapçığında yer alan madde 2, madde 7, madde 14 ve madde 19'un, B₁ kitapçığında yer alan madde 3 ve madde 5'in yaygınlık katsayısının yüksek olduğu ve bu nedenle bu maddelerde QWK değerinin olduğundan düşük olabileceği öngörülmektedir. A₁ kitapçığında madde 10 ve madde 11, B₁ kitapçığında madde 8, madde 9 ve madde 18'in en düşük yaygınlık katsayısına sahip maddeler olduğu ve bu nedenle QWK değerinin gerçekte var olan uyuma daha yakın olacağı öngörülmektedir. A₁ ve B₁ kitapçığında yer alan maddelerin tamamının yanlılık değerleri çok düşüktür ve bu nedenle QWK değerinin olduğundan yüksek bulunma ihtimali oldukça azdır.

Uyum yüzdesi, QWK ve AC1 katsayıları hesaplanırken sırasıyla R (R Core Team, 2018) programında bulunan "irr" (Gamer, Lemon, Fellows ve Singh, 2010), "rel" (LoMartire, 2017) ve "Metrics" (Hamner ve Frasco, 2018) paketlerinden yararlanılmıştır. Uyum katsayıları için tüm maddelerin ortalaması alınarak algoritmaların performansı karşılaştırılmıştır. Ayrıca sistemi test etmede kullanılan veri oranlarının ortalaması alınarak algoritmaların performansı gözden geçirilmiştir.

BULGULAR

A₁ kitapçığında yer alan açık uçlu maddelere ilişkin uyum katsayıları öncelikle gerçek puanlayıcı grubu 1 ve 2 ile gerçek puanlayıcıların nihai puanları arasında hesaplanmıştır. Ardından beş farklı otomatik puanlama algoritması ile nihai puanlar arasındaki uyum otomatik puanlama sisteminin test edilmesinde kullanılan veri oranları değiştirilerek incelenmiştir. Sonuçlar tablo 2'de gösterilmektedir. A₁ kitapçığında yer alan bir maddenin (madde 2) yorumuna örnek olarak yer verilmiştir. Örnek madde yaygınlık probleminin bulunduğu bir duruma ilişkindir. A₁ kitapçığında bulunan diğer maddelere

ilişkin sonuçlar tablo 2 üzerinden değerlendirilebilir. Tablo 2’de her bir uyum katsayısı türüne göre en yüksek uyum değerine sahip üç katsayı koyu, en düşük uyum değerine sahip üç katsayı ise italik olarak gösterilmiştir.

Tablo 2’de madde 2’ye ait değerler incelendiğinde ilk gerçek puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesinin .980, AC1 indeksinin .976 ve QWK değerinin ise .880 olduğu görülmektedir. İkinci gerçek puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesi .979, AC1 indeksi .975 ve QWK değeri ise .862 olarak bulunmuştur.

OP ile gerçek puanlayıcıların nihai puanları arasındaki uyum %10 test veri oranı için incelendiğinde en yüksek uyum yüzdesinin BLSTM algoritması ile .941 olarak elde edildiği ve bu değeri .921 ile MNB algoritmasının izlediği görülmektedir. En düşük uyum yüzdesi ise .913 ile LSTM algoritmasında elde edilmiştir. Uyum yüzdeleri genel olarak incelendiğinde değerlerin birbirine yakın ve kabul edilebilir düzeyde ($>.80$) olduğu sonucuna ulaşılmaktadır. AC1 indeksi incelendiğinde en yüksek uyumun elde edildiği algoritma .931 ile BLSTM olup bu algoritmayı .910 ile LR algoritması izlemektedir. En düşük AC1 değeri ise .904 ile SVM ve LSTM algoritmalarına aittir. AC1 değerlerinin tüm algoritmalar için birbirine yakın olduğu ve çok iyi uyum gösterdiği ($>.80$) gözlemlenmiştir. En yüksek QWK değeri BLSTM algoritması ile .569 olarak bulunmuş olup bu değeri .448 ile MNB algoritması izlemektedir. En düşük QWK değeri .061 ile LSTM algoritmasına aittir ve bu değeri .223 ile LR algoritması izlemektedir. QWK değerlerinin algoritmalar arasında oldukça değişkenlik gösterdiği, ranjının .508 olduğu ve AC1 indeksi ile uyum yüzdesinden ayrıştığı sonucuna ulaşılmıştır. QWK değeri genel olarak değerlendirildiğinde BLSTM ve MNB algoritmalarının orta düzeyde ($<.60 \wedge >.40$), LR ve SVM algoritmalarının ($<.40 \wedge >.20$) kayda değer, LSTM algoritmasının ise zayıf ($<.20$) uyum gösterdiği belirtilebilir.

%20 test veri oranında en yüksek uyum yüzdesini .942 ile BLSTM algoritması göstermiş olup en düşük uyum yüzdesini ise .913 ile MNB algoritması göstermiştir. Tüm algoritmalarda uyum yüzdelerinin birbirine oldukça yakın ve kabul edilebilir düzeyde ($>.80$) olduğu görülmektedir. AC1 indeksi açısından uyum değerlendirildiğinde en yüksek uyuma .933 ile BLSTM algoritmasında rastlanmış olup en düşük uyuma ise .899 ile MNB algoritmasında rastlanmıştır. AC1 indeksi değerlerinin genel olarak yakın olduğu ve tamamının çok iyi uyum gösterdiği ($>.80$) belirtilebilir. QWK değerleri incelendiğinde en yüksek uyuma sahip algoritmanın .593 ile BLSTM olduğu en düşük uyuma sahip algoritmanın ise .147 ile LSTM olduğu belirtilebilir. En düşük uyuma sahip ikinci algoritma ise .212 ile SVM’dir. Görüldüğü üzere %20 test veri oranında, %10 test veri oranına benzer şekilde QWK değerleri düşük ve algoritmalar arasında farklılık olacak şekilde elde edilmiştir. %20 test veri oranında QWK değerlerinin ranjı .446’dır. QWK değerleri genel olarak incelendiğinde BLSTM algoritmasının orta düzeyde uyuma ($<.60 \wedge >.40$), MNB, LR ve SVM algoritmalarının kayda değer uyuma ($<.40 \wedge >.20$), LSTM algoritmasının ise zayıf uyuma ($<.20$) işaret ettiği görülmektedir.

%33 test veri oranı için uyum yüzdesi en yüksek algoritma .934 ile BLSTM’dir. En düşük uyum yüzdesine sahip algoritma ise .909 ile SVM’dir. Genel olarak uyum yüzdeleri yüksek, birbirine yakın ve kabul edilebilir ($>.80$) bulunmuştur. AC1 indeksleri genel olarak yüksek bulunmakla birlikte en yüksek uyum .924 ile BLSTM, en düşük uyum .899 ile SVM algoritmalarına aittir. Tüm algoritmalar için elde edilen değerler birbirine yakın olup çok iyi uyum ($>.80$) göstermektedir. QWK değerleri değerlendirildiğinde en yüksek uyum .522 ile BLSTM algoritmasında en düşük iki uyum ise .128 ile SVM, .000 ile LSTM algoritmalarında elde edilmiştir. %33 test veri oranında QWK değerleri düşüktür, algoritmalar arasında fazla değişkenlik göstermektedir ve ranjı .522’dir. Elde edilen değerler ele alındığında BLSTM algoritmasının orta düzeyde uyuma ($<.60 \wedge >.40$), MNB ve LR algoritmalarının kayda değer uyuma ($<.40 \wedge >.20$), LSTM ve SVM algoritmalarının ise zayıf uyuma ($<.20$) sahip olduğu görülmektedir.

Tablo 2. A₁ Kitapçığındaki Açık Uçlu Maddelere Yönelik Gerçek Puanlayıcı Grupları ve Otomatik Puanlama Algoritmaları ile Nihai Puanlar Arasındaki Uyum Katsayıları

Madde Kodu	Gerçek Puanlayıcı Grubu ile Nihai Puanlar Arası Uyum				Test verisi seçim yöntemi	Otomatik Puanlama Algoritmaları ile Nihai (Gerçek Puanlayıcıların Üzerinde Anlaştıkları) Puanlar Arası Uyum														
				SVM		LR			MNB			LSTM			BLSTM					
	UY	AC1	QWK	UY		AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK				
Madde 2	P ₁ -P _N	.980	.976	.880	ÇG %10	.914	.904	.226	.919	.910	.223	.921	.908	.448	.913	.904	.061	.941	.931	.569
	P ₂ -P _N	.979	.975	.862	ÇG %20	.916	.906	.212	.923	.914	.273	.913	.899	.347	.916	.907	.147	.942	.933	.593
					ÇG %33	.909	.899	.128	.921	.912	.208	.918	.906	.337	.911	.903	.000	.934	.924	.522
Madde 7*	P ₁ -P _N	.979	.970	.974	ÇG %10	.845	.782	.862	.822	.752	.836	.735	.642	.720	.720	.629	.683	.881	.833	.884
	P ₂ -P _N	.970	.958	.971	ÇG %20	.855	.796	.859	.815	.743	.832	.731	.639	.720	.735	.647	.744	.881	.833	.892
					ÇG %33	.827	.756	.825	.822	.752	.832	.722	.625	.705	.728	.638	.726	.875	.823	.877
Madde 8*	P ₁ -P _N	.997	.995	.997	ÇG %10	.928	.894	.910	.936	.906	.915	.896	.849	.859	.779	.687	.701	.957	.937	.937
	P ₂ -P _N	.987	.981	.985	ÇG %20	.936	.906	.917	.931	.899	.911	.901	.856	.868	.776	.683	.684	.946	.921	.899
					ÇG %33	.931	.899	.909	.931	.899	.896	.875	.819	.839	.771	.676	.672	.942	.916	.912
Madde 10*	P ₁ -P _N	.944	.891	.885	ÇG %10	.837	.682	.665	.845	.699	.681	.827	.667	.641	.840	.688	.672	.863	.733	.720
	P ₂ -P _N	.947	.897	.892	ÇG %20	.840	.689	.672	.842	.693	.675	.835	.681	.660	.829	.662	.652	.842	.695	.673
					ÇG %33	.817	.642	.626	.819	.649	.626	.830	.673	.648	.824	.657	.637	.835	.680	.660
Madde 11*	P ₁ -P _N	.985	.972	.968	ÇG %10	.870	.755	.723	.875	.769	.726	.843	.720	.648	.924	.860	.835	.956	.917	.904
	P ₂ -P _N	.985	.972	.968	ÇG %20	.873	.761	.730	.881	.779	.744	.835	.708	.626	.934	.879	.855	.962	.929	.918
					ÇG %33	.871	.757	.727	.865	.748	.708	.825	.693	.600	.870	.759	.717	.946	.898	.883

* A₁ ve B₁ kitapçıklarında yer alan ortak maddeleri göstermektedir.

Not 1: P₁: İlk puanlayıcı grubu, P₂: İkinci puanlayıcı grubu, P_N: Nihai puanlar anlamına gelmektedir.

Not 2: UY uyum yüzdesini, AC1 Gwet'in AC1 katsayısını, QWK karesel ağırlıklı Kappa değerini göstermektedir.

Not 3: ÇG: Çapraz geçerlik anlamına gelmekte, %10, %20 ve %33 ise test veri oranlarını göstermektedir.

Tablo 2 (devamı). A₁ Kitapçığındaki Açık Uçlu Maddelere Yönelik Gerçek Puanlayıcı Grupları ve Otomatik Puanlama Algoritmaları ile Nihai Puanlar Arasındaki Uyum Katsayıları

Madde Kodu	Gerçek Puanlayıcı Grubu ile Nihai Puanlar Arası Uyum				Test verisi seçim yöntemi	Otomatik Puanlama Algoritmaları ile Nihai (Gerçek Puanlayıcıların Üzerinde Anlaştıkları) Puanlar Arası Uyum														
				SVM		LR			MNB			LSTM			BLSTM					
	UY	AC1	QWK	UY		AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK				
Madde 14	P ₁ -P _N	.975	.959	.937	ÇG %10	.901	.839	.744	.911	.857	.764	.890	.828	.695	.792	.709	.318	.929	.884	.818
	P ₂ -P _N	.969	.948	.921	ÇG %20	.895	.829	.724	.904	.847	.747	.881	.817	.667	.873	.807	.635	.928	.880	.816
					ÇG %33	.893	.825	.725	.906	.849	.752	.876	.811	.646	.792	.710	.315	.916	.864	.781
Madde 15*	P ₁ -P _N	.972	.960	.971	ÇG %10	.708	.585	.683	.720	.603	.686	.687	.563	.613	.560	.428	.224	.766	.666	.714
	P ₂ -P _N	.960	.943	.943	ÇG %20	.717	.595	.678	.712	.593	.664	.672	.544	.589	.539	.415	.137	.740	.628	.707
					ÇG %33	.677	.539	.656	.690	.562	.625	.680	.557	.564	.516	.397	.000	.741	.628	.711
Madde 18	P ₁ -P _N	.997	.995	.997	ÇG %10	.956	.937	.952	.924	.893	.914	.867	.811	.790	.718	.616	.517	.970	.958	.961
	P ₂ -P _N	.998	.998	.994	ÇG %20	.941	.916	.937	.921	.888	.904	.868	.813	.796	.761	.672	.599	.965	.951	.952
					ÇG %33	.924	.893	.912	.923	.891	.906	.863	.807	.756	.671	.544	.515	.960	.944	.947
Madde 19	P ₁ -P _N	.997	.996	.997	ÇG %10	.919	.892	.900	.936	.915	.918	.815	.752	.807	.802	.739	.749	.939	.918	.922
	P ₂ -P _N	.995	.993	.996	ÇG %20	.914	.886	.897	.931	.908	.909	.822	.762	.820	.797	.736	.720	.937	.916	.936
					ÇG %33	.918	.890	.904	.921	.895	.899	.820	.760	.800	.778	.719	.624	.919	.891	.918

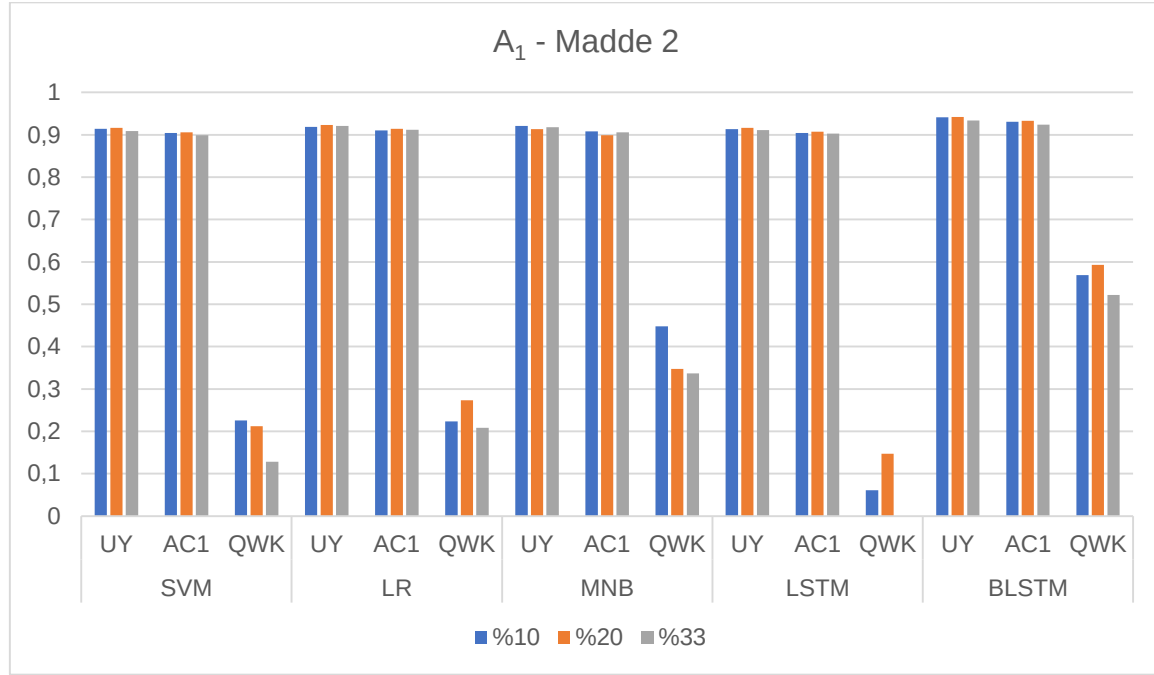
* A₁ ve B₁ kitapçıklarında yer alan ortak maddeleri göstermektedir.

Not 1: P₁: İlk puanlayıcı grubu, P₂: İkinci puanlayıcı grubu, P_N: Nihai puanlar anlamına gelmektedir.

Not 2: UY uyum yüzdesini, AC1 Gwet'in AC1 katsayısını, QWK karesel ağırlıklı Kappa değerini göstermektedir.

Not 3: ÇG: Çapraz geçerlik anlamına gelmekte, %10, %20 ve %33 ise test veri oranlarını göstermektedir.

Şekil 1 otomatik puanlama algoritmaları ve test veri oranlarına göre A₁ kitapçığında yer alan madde 2 için elde edilen uyum değerlerini göstermektedir.



Şekil 1. Otomatik Puanlama Algoritmaları ve Test Veri Oranlarına göre A₁ Kitapçığı Madde 2'ye İlişkin Uyum Değerlerini Gösteren Grafik

Şekil 1 incelendiğinde; madde 2 için tüm test veri oranları ve OP algoritmalarında QWK katsayısı, AC1 değerleri ve uyum yüzdesinden oldukça düşüktür. QWK katsayılarının tamamında düşük değerlerle karşılaşılmasının ve bazı koşullarda katsayının .000'a kadar yaklaşmasının sebebi yaygınlık problemi. Bu nedenle QWK değerleri değerlendirilmemiştir. Bu durum araştırmada öngörülen durumlardan birisidir. Tüm test veri oranları ve otomatik puanlama algoritmaları dikkate alınarak karşılaştırma yapıldığında uyum değerlerinin %20 test veri oranında biraz daha yüksek, %33 test veri oranında ise biraz daha düşük olduğu gözlemlenmektedir. Ancak aralarındaki farklar oldukça azdır. Uyum yüzdesi tüm koşullarda kabul edilebilir sınır olan .80'in üzerindedir. AC1 indeksi ise tüm koşullarda (>.80) çok iyi uyuma işaret etmektedir. AC1 değerleri Kappa katsayısı ile aynı yönde değerlendirilmektedir. Buna göre tüm AC1 katsayıları otomatik puanlama ile gerçek puanlayıcılar arasında bulunması gereken uyum değerinden (>.70, Williamson vd., 2012) yüksektir. Tablo 2'de madde 2 için tüm koşullar dikkate alındığında en yüksek uyum yüzdesi (.942) ve en yüksek AC1 değeri (.933) BLSTM algoritmasında ve %20 test veri oranında elde edilmiştir. Elde edilen bu değerler gerçek puanlayıcı grupları ile nihai puanlar arasındaki uyum yüzdesi ve AC1 değerine yakındır. Madde 2'de karşılaşılan yaygınlık sorunu nedeniyle gerçek puanlayıcılar ile nihai puanlar arasında hesaplanan QWK değerleri de düşüktür. Bu durum makine öğrenmesine daha olumsuz şekilde yansımıştır.

B₁ kitapçığında yer alan açık uçlu maddelere ilişkin uyum katsayıları A₁ kitapçığıyla aynı şekilde hesaplanmıştır. Sonuçlar tablo 3'de gösterilmektedir. B₁ kitapçığında yer alan bir maddenin (madde 5) yorumuna örnek olarak yer verilmiştir. B₁ kitapçığında bulunan diğer maddelere ilişkin sonuçlar tablo 3 üzerinden değerlendirilebilir. Tablo 3'de her bir uyum katsayısı türüne göre en yüksek uyum değerine sahip üç katsayı koyu, en düşük uyum değerine sahip üç katsayı ise italik olarak gösterilmiştir.

Tablo 3. B₁ Kitapçığındaki Açık Uçlu Maddelere Yönelik Gerçek Puanlayıcı Grupları ve Otomatik Puanlama Algoritmaları ile Nihai Puanlar Arasındaki Uyum Katsayıları

Madde Kodu	Gerçek Puanlayıcı Grubu ile Nihai Puanlar Arası Uyum	UY	AC1	QWK	Test verisi seçim yöntemi	Otomatik Puanlama Algoritmaları ile Gerçek Puanlayıcıların Üzerinde Anlaştıkları Puanlar Arası Uyum														
						SVM			LR			MNB			LSTM			BLSTM		
						UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK
Madde 3	P ₁ -P _N	.966	.952	.877	ÇG %10	.911	.879	.665	.913	.882	.667	.906	.871	.653	.913	.880	.678	.923	.894	.719
	P ₂ -P _N	.973	.962	.900	ÇG %20	.914	.883	.683	.911	.879	.665	.904	.869	.642	.921	.891	.716	.913	.879	.686
					ÇG %30	.916	.885	.688	.906	.872	.644	.901	.865	.623	.911	.878	.671	.911	.878	.671
Madde 5*	P ₁ -P _N	.971	.960	.972	ÇG %10	.866	.818	.884	.836	.778	.864	.779	.710	.740	.836	.778	.861	.918	.888	.925
	P ₂ -P _N	.979	.972	.979	ÇG %20	.863	.814	.882	.837	.781	.855	.781	.712	.743	.825	.766	.846	.902	.866	.913
					ÇG %30	.870	.823	.878	.844	.790	.866	.786	.720	.744	.784	.718	.783	.892	.853	.904
Madde 6*	P ₁ -P _N	.991	.988	.981	ÇG %10	.942	.915	.909	.954	.933	.924	.884	.833	.861	.740	.628	.654	.959	.940	.939
	P ₂ -P _N	.993	.990	.995	ÇG %20	.945	.920	.919	.947	.923	.915	.873	.819	.848	.752	.649	.645	.949	.925	.923
					ÇG %30	.937	.908	.916	.947	.923	.906	.846	.781	.832	.719	.593	.682	.952	.930	.926
Madde 8*	P ₁ -P _N	.950	.902	.899	ÇG %10	.827	.659	.649	.818	.645	.629	.820	.649	.632	.834	.673	.663	.854	.713	.704
	P ₂ -P _N	.957	.916	.913	ÇG %20	.812	.629	.618	.800	.608	.591	.832	.673	.656	.805	.618	.601	.858	.719	.713
					ÇG %30	.820	.646	.634	.793	.593	.578	.827	.662	.646	.793	.590	.582	.842	.691	.679
Madde 9*	P ₁ -P _N	.985	.971	.967	ÇG %10	.846	.711	.670	.836	.696	.642	.796	.637	.538	.877	.772	.732	.885	.788	.751
	P ₂ -P _N	.993	.987	.985	ÇG %20	.844	.706	.668	.844	.714	.658	.796	.641	.533	.873	.767	.722	.882	.779	.746
					ÇG %30	.849	.716	.679	.837	.698	.647	.796	.643	.531	.868	.760	.707	.872	.766	.716

* A₁ ve B₁ kitapçıklarında yer alan ortak maddeleri göstermektedir.

Not 1: P₁: İlk puanlayıcı grubu puanları, P₂: İkinci puanlayıcı grubu puanları, P_N: Nihai puanlar anlamına gelmektedir.

Not 2: UY uyum yüzdesini, AC1 Gwet'in AC1 katsayısını, QWK karesel ağırlıklı Kappa değerini göstermektedir.

Not 3: ÇG: Çapraz geçerlik anlamına gelmekte, %10, %20 ve %33 ise test veri oranlarını göstermektedir.

Not 4: Bu tabloda yer alan madde 5, madde 6, madde 8 ve madde 9 sırasıyla A₁ kitapçığındaki madde 7, madde 8, madde 10 ve madde 11'e karşılık gelmektedir.

Tablo 3 (devamı). B₁ Kitapçığındaki Açık Uçlu Maddelere Yönelik Gerçek Puanlayıcı Grupları ve Otomatik Puanlama Algoritmaları ile Nihai Puanlar Arasındaki Uyum Katsayıları

Madde Kodu	Gerçek Puanlayıcı Grubu ile Nihai Puanlar Arası Uyum				Test verisi seçim yöntemi	Otomatik Puanlama Algoritmaları ile Gerçek Puanlayıcıların Üzerinde Anlaştıkları Puanlar Arası Uyum														
				SVM		LR			MNB			LSTM			BLSTM					
	UY	AC1	QWK	UY		AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK				
Madde 11	P ₁ -P _N	.986	.981	.987	ÇG %10	.918	.886	.912	.911	.876	.902	.861	.807	.867	.882	.838	.887	.940	.916	.925
	P ₂ -P _N	.990	.986	.989	ÇG %20	.902	.865	.893	.913	.878	.900	.863	.810	.863	.878	.833	.880	.943	.920	.929
					ÇG %30	.904	.867	.901	.914	.881	.899	.861	.808	.860	.885	.843	.894	.930	.901	.927
Madde 12	P ₁ -P _N	.949	.923	.932	ÇG %10	.736	.606	.667	.757	.637	.719	.707	.566	.606	.654	.490	.663	.793	.690	.749
	P ₂ -P _N	.938	.908	.937	ÇG %20	.759	.640	.718	.764	.647	.740	.682	.528	.559	.649	.481	.674	.784	.677	.741
					ÇG %30	.755	.634	.718	.755	.635	.719	.683	.531	.573	.634	.467	.654	.774	.662	.738
Madde 17*	P ₁ -P _N	.974	.963	.966	ÇG %10	.707	.580	.653	.693	.565	.631	.635	.492	.522	.541	.393	.171	.743	.634	.705
	P ₂ -P _N	.978	.968	.974	ÇG %20	.729	.612	.675	.678	.543	.609	.610	.456	.488	.545	.391	.302	.716	.595	.671
					ÇG %30	.680	.543	.617	.700	.575	.637	.616	.471	.478	.575	.430	.339	.697	.567	.644
Madde 18	P ₁ -P _N	1.000	1.000	1.000	ÇG %10	.712	.425	.429	.748	.497	.497	.740	.480	.485	.784	.568	.571	.786	.572	.572
	P ₂ -P _N	.995	.990	.990	ÇG %20	.711	.421	.425	.741	.483	.483	.726	.453	.458	.759	.517	.520	.767	.535	.534
					ÇG %30	.719	.439	.442	.731	.463	.462	.731	.463	.466	.755	.510	.512	.769	.538	.538
Madde 20	P ₁ -P _N	.969	.945	.929	ÇG %10	.818	.687	.569	.817	.685	.563	.760	.562	.471	.834	.703	.623	.839	.717	.627
	P ₂ -P _N	.969	.946	.929	ÇG %20	.815	.681	.562	.830	.708	.597	.750	.544	.447	.820	.683	.585	.837	.710	.629
					ÇG %30	.789	.640	.495	.810	.674	.545	.740	.527	.421	.793	.630	.529	.820	.691	.572

* A₁ ve B₁ kitapçıklarında yer alan ortak maddeleri göstermektedir.

Not 1: P₁: İlk puanlayıcı grubu puanları, P₂: İkinci puanlayıcı grubu puanları, P_N: Nihai puanlar anlamına gelmektedir.

Not 2: UY uyum yüzdesini, AC1 Gwet'in AC1 katsayısını, QWK karesel ağırlıklı Kappa değerini göstermektedir.

Not 3: ÇG: Çapraz geçerlik anlamına gelmekte, %10, %20 ve %33 ise test veri oranlarını göstermektedir.

Not 4: Bu tabloda yer alan madde 17, A₁ kitapçığındaki madde 15'e karşılık gelmektedir.

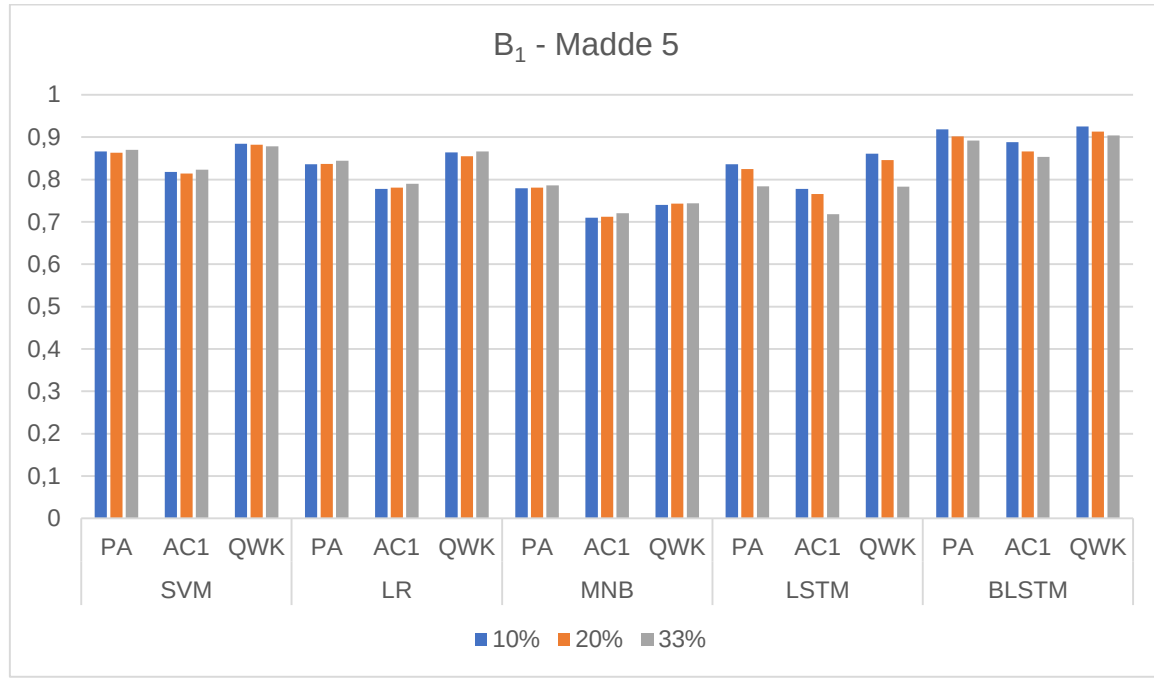
Tablo 3'te madde 5'e ait değerler incelendiğinde ilk gerçek puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesinin .971, AC1 indeksinin .960, QWK değerinin ise .972 olduğu görülmektedir. İkinci gerçek puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesi .979, AC1 indeksi .972 ve QWK değeri ise .979 olarak bulunmuştur.

Otomatik puanlama ile nihai puanlar arasındaki uyum %10 test veri oranı için incelendiğinde en yüksek uyum yüzdesi BLSTM algoritması ile .918 olarak elde edilmiştir. Bu uyum yüzdesi değerini .866 ile SVM algoritması takip etmektedir. En düşük uyum yüzdesi ise .779 ile MNB algoritmasında elde edilmiştir. Uyum yüzdeleri genel olarak incelendiğinde SVM, LR, LSTM ve BLSTM algoritmaları için kabul edilebilir değerlere ($>.80$) ulaşıldığı görülmektedir. AC1 indeksi incelendiğinde en yüksek uyumun elde edildiği algoritma .888 ile BLSTM'dir. En düşük AC1 değeri ise .710 ile MNB algoritmasına ait olup bu algoritmayı .778 ile LR ve LSTM algoritmaları izlemektedir. AC1 değerlerinin BLSTM ve SVM algoritmaları için çok iyi ($>.80$), LR, LSTM ve MNB algoritmaları için iyi uyuma ($>.60 \wedge <.80$) işaret ettiği bulunmuştur. En yüksek QWK değeri BLSTM algoritması ile .925 olarak bulunmuş olup bu değeri .884 ile SVM algoritması izlemektedir. En düşük QWK değeri .740 ile MNB algoritmasına aittir. QWK değerlerinin AC1 indekslerinden büyük olduğu görülmüştür. QWK değeri SVM, LR, LSTM ve BLSTM algoritmaları için çok iyi ($>.80$), MNB algoritması için iyi uyum ($>.60 \wedge <.80$) bulunduğunu göstermektedir.

%20 test veri oranında en yüksek uyum yüzdesini .902 ile BLSTM algoritması göstermiş olup en düşük uyum yüzdesini ise .781 ile MNB algoritması göstermiştir. Uyum yüzdesine göre BLSTM, LR, LSTM ve SVM algoritmaları kabul edilebilir uyum ($>.80$) gösterirken MNB algoritması kabul edilebilir uyum göstermemiştir. AC1 indeksi açısından en yüksek uyuma .866 ile BLSTM algoritmasında rastlanmış olup en düşük uyuma ise .712 ile MNB algoritmasında rastlanmıştır. AC1 indeksi değerlerinin BLSTM ve SVM algoritmaları için çok iyi ($>.80$), LR, LSTM ve MNB algoritmaları için ise iyi uyum ($<.80 \wedge >.60$) gösterdiği belirtilebilir. QWK değerleri incelendiğinde en yüksek uyuma sahip algoritmanın .913 ile BLSTM, en düşük uyuma sahip algoritmanın ise .743 ile MNB olduğu belirtilebilir. En düşük QWK değerine sahip ikinci algoritma ise .846 ile LSTM'dir. Görüldüğü üzere QWK açısından MNB algoritması için iyi ($<.80 \wedge >.60$), BLSTM, LR, LSTM ve SVM algoritmaları için çok iyi uyuma ($>.80$) ulaşılmıştır. QWK değerlerinin %20 test veri oranında AC1 indekslerinden büyük olduğu görülmektedir.

%33 test veri oranı için uyum yüzdesi en yüksek algoritma .892 ile BLSTM'dir. En düşük uyum yüzdesine sahip algoritma ise .784 ile LSTM'dir. LSTM ve MNB algoritmaları dışında tüm algoritmalarda uyum yüzdesi kabul edilebilir ($>.80$) bulunmuştur. AC1 indekslerine göre en yüksek uyum .853 ile BLSTM algoritmasına aittir. En düşük uyum ise .718 ile LSTM algoritmasına aittir ve bu algoritmayı .720 ile MNB algoritması izlemektedir. AC1 indeksleri açısından BLSTM ve SVM algoritmaları için çok iyi uyuma ($>.80$), LR, LSTM ve MNB algoritmaları için ise iyi uyuma ($<.80 \wedge >.60$) ulaşıldığı görülmektedir. QWK katsayısına göre en yüksek uyum .904 ile BLSTM algoritmasında, en düşük iki uyum ise .744 ile MNB, .783 ile LSTM algoritmalarında elde edilmiştir. QWK değerleri BLSTM, LR ve SVM algoritmaları için çok iyi ($>.80$); LSTM ve MNB algoritmaları için iyi uyuma ($<.80 \wedge >.60$) işaret etmektedir. QWK değerlerinin %33 test veri oranında da AC1 indekslerinden büyük olduğu görülmektedir.

Şekil 2 otomatik puanlama algoritmaları ve test veri oranlarına göre B₁ kitapçığında yer alan madde 5 için elde edilen uyum değerlerini göstermektedir.



Şekil 2. Otomatik Puanlama Algoritmaları ve Test Veri Oranlarına göre B₁ Kitapçığı Madde 5'e İlişkin Uyum Değerlerini Gösteren Grafik

Şekil 2 incelendiğinde tüm koşullarda MNB algoritmasına ait uyum katsayıları diğer algoritmalarla ait uyum katsayılarından daha düşük, BLSTM algoritmasına ait uyum katsayıları ise diğer algoritmalarla ait uyum katsayılarından daha yüksektir. QWK değeri BLSTM, LR ve SVM algoritmaları için tüm test veri oranlarında ve LSTM algoritması için %10 ve %20 test veri oranında çok iyi ($>.80$), MNB algoritmasında tüm test veri oranlarında ve LSTM algoritması %33 test veri oranında iyi ($<.80 \wedge >.60$) uyum göstermiştir. Tüm koşullarda AC1 değerleri BLSTM ve SVM algoritmaları için çok iyi ($>.80$) uyum; LR, MNB ve LSTM algoritmaları için iyi uyum ($<.80 \wedge >.60$) bulunduğunu göstermektedir. Madde 5 için tüm AC1 katsayıları QWK katsayılarından daha düşüktür. Uyum yüzdesi BLSTM, LR ve SVM algoritmaları için tüm test veri oranlarında; LSTM algoritmasında ise %10 ve %20 test veri oranlarında kabul edilebilir ($>.80$) değerler göstermiştir. Williamson, Xi ve Breyer (2012) tarafından gerçek puanlayıcılar ile otomatik puanlama arasındaki Kappa uyum katsayısının en az .70 olması kriterine göre tüm algoritmalar ve test veri oranlarında QWK değerleri kabul edilebilir bulunmuştur. Aynı kriter AC1 katsayısı için kullanıldığında tüm algoritmalar ve test veri oranlarında kabul edilebilir değerlere ulaşılmıştır. Madde 5 için tüm koşullarda en yüksek uyum yüzdesi (.918), AC1 değeri (.888) ve QWK katsayısı (.925) BLSTM algoritmasında ve %10 test veri oranında elde edilmiştir. Bu değerler gerçek puanlayıcı grupları ile nihai puanlar arasında bulunan uyum yüzdesi, AC1 ve QWK değerlerine yakındır.

Otomatik puanlama algoritmaları arasında genel bir karşılaştırma yapabilmek amacıyla algoritmaların her bir maddedeki performanslarının ortalaması alınmıştır. Tablo 4, otomatik puanlama algoritmalarının farklı test veri oranlarındaki performanslarını ve bu performansların ortalamalarını göstermektedir. Tablo 4'de her bir test veri oranında ve ortalama performansta tüm uyum katsayılarında en yüksek uyumu gösteren katsayılar koyu, en düşük uyumu gösteren katsayılar ise italik olarak gösterilmiştir.

Tablo 4. Otomatik Puanlama Algoritmalarının Ortalama Performansları

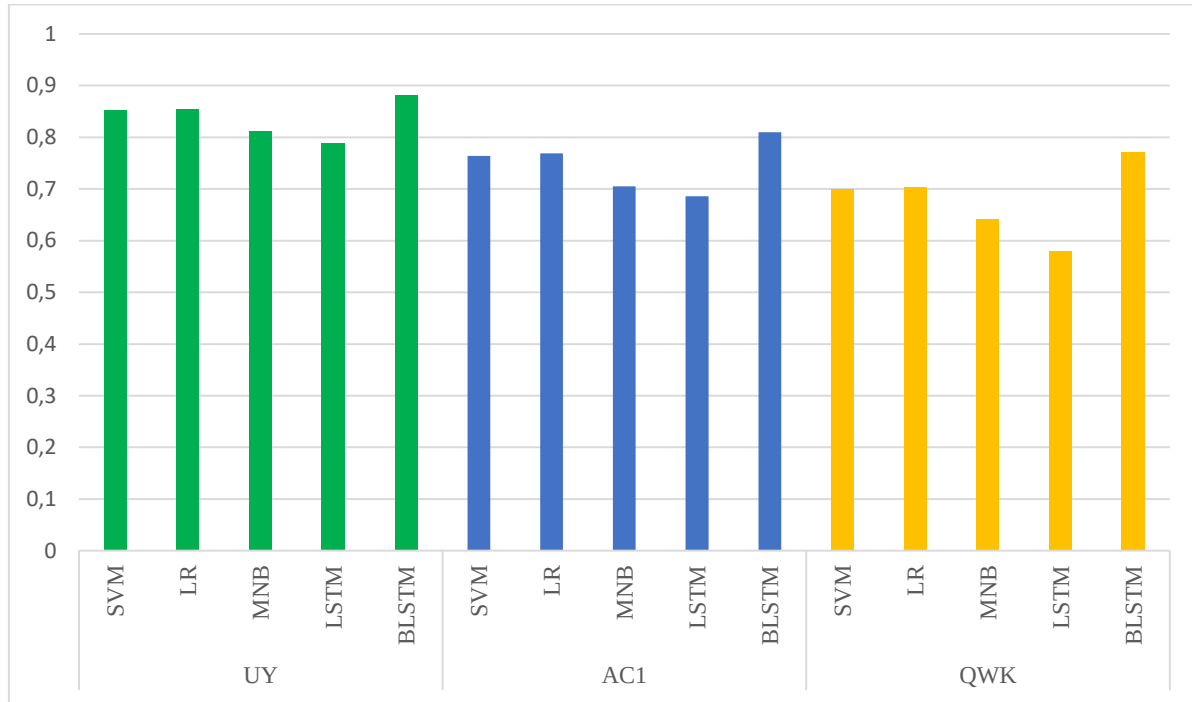
Uyum Katsayıları	Otomatik Puanlama Algoritması	%10	%20	%33	Ortalama
UY	SVM	.855	.855	.848	.853
	LR	.857	.856	.851	.855
	MNB	.816	.810	.807	.811
	LSTM	.794	.799	.775	.789
	BLSTM	.889	.883	.874	.882
AC1	SVM	.768	.767	.756	.764
	LR	.773	.771	.762	.769
	MNB	.712	.704	.700	.705
	LSTM	.694	.698	.665	.686
	BLSTM	.822	.810	.798	.810
QWK	SVM	.705	.704	.689	.699
	LR	.710	.710	.692	.704
	MNB	.658	.640	.627	.642
	LSTM	.583	.612	.545	.580
	BLSTM	.782	.775	.755	.771

Tablo 4’de her bir test veri oranına ilişkin uyum yüzdeleri incelendiğinde değerlerin birbirine yakın olduğu görülmekle birlikte %33 test veri oranında değerlerde biraz azalma olmuştur. LSTM algoritması dışında tüm algoritmalar uyum yüzdesi açısından kabul edilebilir değerler göstermiştir. LSTM algoritması ise kabul edilebilir uyuma yakın değerler göstermiştir.

AC1 değerleri incelendiğinde %33 test veri oranında biraz azalma olduğu, bunun yanı sıra SVM, LR, MNB ve LSTM algoritmalarının ortalama performanslarının iyi uyuma işaret ettiği görülmektedir. BLSTM algoritması ise %10 ve %20 test veri oranlarında çok iyi, %33 test veri oranında iyi uyum göstermiştir.

QWK değerleri incelendiğinde AC1 ve uyum yüzdesine benzer şekilde %33 test veri oranında azalma olduğu, bunun yanında tüm test veri oranlarında yakın değerlerin elde edildiği görülmektedir. QWK değeri açısından SVM, LR, MNB ve BLSTM algoritmaları iyi uyum bulunduğunu göstermektedir. LSTM algoritması ise %20 test veri oranında iyi uyum, %10 ve %33 test veri oranlarında orta düzeyde uyum göstermiştir.

Tüm test veri oranlarının ortalamaları her bir OP algoritması ve uyum katsayısı açısından incelendiğinde en yüksek uyum yüzdesi, AC1 ve QWK değerine sahip algoritmanın BLSTM olduğu görülmektedir. BLSTM algoritması kabul edilebilir uyum yüzdesine sahip olmakla birlikte, AC1 katsayısına göre çok iyi, QWK katsayısına göre iyi uyum göstermektedir. SVM, LR ve MNB algoritmaları kabul edilebilir uyum yüzdesi, AC1 katsayısına ve QWK katsayısına göre ise iyi uyum göstermektedir. LSTM algoritması kabul edilebilir uyum yüzdesine sahip olmayıp, AC1 indeksi açısından iyi uyum, QWK katsayısı açısından ise orta uyum göstermektedir. Hem madde ortalamalarının değerlendirilmesi hem de madde kapsamında yapılan değerlendirmeler sonucunda en iyi üç otomatik puanlama koşulu BLSTM algoritması %10 test veri oranı, BLSTM algoritması %20 test veri oranı ve BLSTM algoritması %33 test veri oranı olarak belirlenmiştir. Şekil 3’te algoritmaların test veri oranlarına göre alınan ortalamaları gösterilmektedir.



Şekil 3. Otomatik Puanlama Algoritmalarının Ortalama Performanslarını Gösteren Grafik

Şekil 3 incelendiğinde MNB ve LSTM algoritmalarının diğer algoritmalarından biraz daha zayıf performans gösterdiği belirlenmiştir. En düşük performans LSTM algoritmasında, en yüksek performans ise BLSTM algoritmasında gözlemlenmiştir.

SONUÇLAR ve TARTIŞMA

Araştırma otomatik puanlama algoritmalarını sistemin test edilmesinde kullanılan veri oranları üzerinde yapılan değişikliklerle karşılaştırmaktadır. Bu amaçla %10, %20 ve %33 test veri oranlarına göre SVM, LR, MNB, LSTM ve BLSTM algoritmaları birbiriyle karşılaştırılmıştır. Algoritmalar karşılaştırılırken gerçek puanlayıcıların nihai puanlarla olan uyumları dikkate alınmıştır. Böylece gerçek puanlayıcılar ile otomatik puanlama farkı belirlenmiştir. ABİDE verileri dikkate alındığında sonuçlar en iyi otomatik puanlamanın BLSTM algoritmasıyla gerçekleştiğini göstermektedir. LSTM ve MNB algoritmaları, SVM, LR ve BLSTM algoritmalarından daha düşük uyum değerlerine sahiptir. Kumar ve Rama Sree (2014) çeşitli sınıflandırma yöntemleri ile ilgili daha önce yaptıkları deneylerde sade bayes algoritmasının, LR ve SVM algoritmalarından daha düşük uyum yüzdelerine sahip olduğu belirlemiştir. Bu sonuç araştırma bulgularını destekler niteliktedir. Gierl vd. (2014) SVM algoritması ile gerçekleştirdiği otomatik puanlama işleminde QWK değerinin çok iyi uyum gösterdiğini belirtmiştir. Halihazırdaki bu çalışmada ise SVM algoritmasının iyi uyum gösterdiği belirlenmiştir. Taghipour ve Tou Ng (2016) otomatik puanlama işleminde yinelenen sinir ağlarını karşılaştırdıkları çalışmada en yüksek QWK değerine (.746) sahip algoritmanın LSTM olduğunu bulmuştur. Aynı çalışmada en yakın QWK değeri BLSTM algoritmasında (.699) elde edilmiştir. Halihazırdaki bu çalışmada ise BLSTM algoritmasına ait QWK değeri benzer şekilde iyi uyuma işaret etmektedir. Ancak halihazırdaki bu çalışmada LSTM algoritmasının QWK değerine göre orta düzeyde uyum gösterdiği belirlenmiştir. Bu durumun nedeni LSTM algoritmasında cümlelerin tek yönlü, BLSTM algoritmasında ise cümlelerin iki yönlü incelenmesinin Türk dilinde farklılık göstermesi olabilir. Test veri oranlarına göre yapılan karşılaştırmalar %33 test veri oranında uyum katsayılarının biraz düştüğünü gösterse de SVM, LR, MNB ve BLSTM algoritmaları tüm koşullarda iyi ya da çok iyi uyum bulunduğunu göstermektedir.

Otomatik puanlama için kabul edilebilir en düşük uyum katsayısına göre karşılaştırma yapıldığında LR ve BLSTM algoritmalarının istenilen seviyede olduğu, SVM algoritmasının ise istenilen seviyeye çok yakın olduğu belirlenmiştir. Halihazırdaki bu araştırma ile oluşturulan sistemin, uyum yüzdesi dikkate alındığında Adesiji vd. (2016) tarafından hazırlanan ve özelliklerin manuel olarak tanımlandığı algoritmadan daha iyi performans gösterdiği belirtilebilir. Böylece Türk dilinde denetlenen makine öğrenimine dayalı uygun otomatik puanlama algoritmasının seçilmesiyle Türk dilinde açık uçlu maddelerin otomatik olarak puanlanabileceği sonucuna varılmıştır. Türk diline benzer özellik taşıyan dillerde geliştirilen otomatik puanlama sistemleri her ne kadar denetlenen makine öğrenimine dayalı olmasa da benzer şekilde kullanılabilir. Ishioka ve Kameda (2006) Japon dili ve Jang vd. (2014) Kore dili üzerinde oluşturdukları otomatik puanlama sistemiyle gerçek puanlar arasında yüksek düzeyde korelasyon bulunduğunu belirlemiştir.

Türk dilinde kurulan otomatik puanlama sistemi geniş ölçekli testlerde kullanılabilir. Benzer bir dil olan Korece’de oluşturulan otomatik puanlama sisteminin de geniş ölçekli testlerde kullanılabileceği belirtilmektedir (Jang vd., 2014). Araştırma sonucunda elde edilen bulgulara dayalı olarak araştırmacı ve uygulayıcılara aşağıdaki önerilerde bulunmaktadır.

1. Türk dilinde ilk kez denenen ve kullanılabilir olduğu görülen otomatik puanlama, sistemin geliştirilmesi ve pilot uygulamaların yapılmasıyla geniş ölçekli testlerde kullanılarak sınav giderleri azaltılabilir, sonuçlar daha hızlı bir şekilde açıklanabilir.
2. Otomatik puanlama algoritmalarından BLSTM ve LR bu araştırmada kullanılan veriye benzer özellikler taşıyan verilerde tercih edilebilir.
3. Otomatik puanlamada MNB ve LSTM algoritmalarının bu araştırmada kullanılan veriye benzer özellikler taşıyan verilerde kullanılmaması önerilebilir.
4. Bu araştırma en az 400 eğitim verisi ile yapılan otomatik puanlama sonuçlarını yansıtmaktadır. Bundan sonraki araştırmalarda daha az sayıda eğitim verisi ile otomatik puanlama yapılarak bu durumun uyum katsayılarına etkisi değerlendirilebilir. Dahası büyük örneklemelerde (>1000 ya da >3000) fazla sayıda eğitim verisi ile otomatik puanlama işlemi yapıldıktan sonra eğitim verileri kademe kademe azaltılarak bu durumun otomatik puanlamaya etkisi incelenebilir.
5. Sonraki araştırmalarda verideki yazım hatalarının düzeltildiği ve düzeltilmediği durumlarda elde edilen otomatik puanlama sonuçları karşılaştırılabilir.
6. Kâğıt-kalem testleri üzerinde yürütülen sonraki araştırmalarda veri girişi OCR sistemleri aracılığıyla ve elle yapılarak elde edilen sonuçlar karşılaştırılabilir.
7. Araştırma kapsamında iki ve üç kategorili maddeler üzerinde çalışılmıştır. Daha sonraki araştırmalarda kategori sayısının artması durumunda otomatik puanlama sistemlerinin davranışları incelenebilir.

KAYNAKÇA

- Adesiji, K. M., Agbonifo, O. C., Adesuyi, A. T., & Olabode, O. (2016). Development of an automated descriptive text-based scoring system. *British Journal of Mathematics & Computer Science*, 19(4), 1-14. doi: 10.9734/BJMCS/2016/27558
- Altman, D. G. (1991). *Practical statistics for medical research*. Boca Raton: CRC.
- Araujo, J., & Born, D. G. (1985). Calculating percentage agreement correctly but writing its formula incorrectly. *The Behavior Analyst*, 8(2), 207-208. doi: 10.1007/BF03393152
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment*, 4(3). Retrieved from <http://www.jtla.org>.
- Berg, P.-C., & Gopinathan, M. (2017). *A deep learning ensemble approach to gender identification of tweet authors* (Master's thesis, Norwegian University of Science and Technology). Retrieved from <https://brage.bibsys.no/xmlui/handle/11250/2458477>
- Brenner, H., & Kliebsch, U. (1996). Dependence of weighted Kappa coefficients on the number of categories. *Epidemiology*, 7(2), 199-202. <https://doi.org/10.1097/00001648-199603000-00016>
- Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and Kappa. *Journal of Clinical Epidemiology*, 46(5), 423-429. doi: 10.1016/0895-4356(93)90018-V

- Chen, H., Xu, J., & He, B. (2014). Automated essay scoring by capturing relative writing quality. *The Computer Journal*, 57(9), 1318-1330. doi:10.1093/comjnl/bxt117
- Cohen, Y., Ben-Simon, A., & Hovav, M. (October, 2003). *The effect of specific language features on the complexity of systems for automated essay scoring*. Paper presented at the International Association of Educational Administration, Manchester.
- Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education*, 31(3), 241-250. <https://doi.org/10.1080/08957347.2018.1464450>
- Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Boston: Pearson.
- Downing, S. M. (2009). Written tests: Constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 149-184). New York, NY: Routledge.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Eugenio, B. D., & Glass, M. (2004). The Kappa statistic: A second look. *Computational Linguistics*, 30(1), 95-101. <https://doi.org/10.1162/089120104773633402>
- Gamer, M., Lemon, I., Fellows, J., & Singh, P. (2010). *irr: Various coefficients of interrater reliability and agreement* (Version 0.83) [Computer software]. <https://CRAN.R-project.org/package=irr>
- Geisinger, K. F., & Usher-Tate, B. J. (2016). A brief history of educational testing and psychometrics. In C. S. Wells, M. Faulkner-Bond (Eds.), *Educational measurement from foundations to future* (pp. 3-20). New York: The Guilford.
- Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & Champlain, A. D. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, 48, 950-962. doi: 10.1111/medu.12517
- Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, 5(1), 13-34. https://doi.org/10.1207/S15327841MPEE0501_2
- Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Report of the Center for Educator Compensation Reform. Retrieved from <https://files.eric.ed.gov/fulltext/ED532068.pdf>
- Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1), 29-48. doi: 10.1348/000711006X126600
- Gwet, K. L. (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement*, 76(4), 609-637. doi: 10.1177/0013164415596420
- Haley, D. T. (2007). *Using a new inter-rater reliability statistic* (Report No. 2017/16). UK: The Open University.
- Hamner, B., & Frasco, M. (2018). *Metrics: Evaluation metrics for machine learning* (Version 0.1.4) [Computer Software]. <https://CRAN.R-project.org/package=Metrics>
- Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 10(1), 103-116.
- Hoek, J., & Scholman, M. C. J. (2017). *Evaluating discourse annotation: Some recent insights and new approaches*. In H. Bunt (Ed.), *ACL Workshop on Interoperable Semantic Annotation* (pp. 1-13). <https://www.aclweb.org/anthology/W17-7401>
- Ishioka, T., & Kameda, M. (2006). *Automated Japanese essay scoring system based on articles written by experts*. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, 44, 233-240. doi: 10.3115/1220175.1220205
- Jang, E-S., Kang, S-S., Noh, E-H., Kim, M-H., Sung, K-H., & Seong, T-J. (2014). KASS: Korean automatic scoring system for short-answer questions. Proceedings of the 6th International Conference on Computer Supported Education, Barcelona, 2, 226-230. doi: 10.5220/0004864302260230
- Kumar, C. S., & Rama Sree, R. J. (2014). An attempt to improve classification accuracy through implementation of bootstrap aggregation with sequential minimal optimization during automated evaluation of descriptive answers. *Indian Journal of Science and Technology*, 7(9), 1369-1375.
- Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism and Mass Communication Quarterly*, 92(4), 1-21. doi: 10.1177/1077699015607338
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.
- Lilja, M. (2018). *Automatic essay scoring of Swedish essays using neural networks* (Doctoral dissertation, Uppsala University). Retrieved from <http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1213688&dswid=9250>
- LoMartire, R. (2017). *rel: Reliability coefficients* (version 1.3.1) [Computer software]. <https://CRAN.R-project.org/package=rel>

- Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61-73). New Jersey: Lawrence Erlbaum Associates, Inc.
- Meyer, G. J. (1999). Simple procedures to estimate chance agreement and Kappa for the interrater reliability of response segments using the rorschach comprehensive system. *Journal of Personality Assessment*, 72(2), 230-255. doi: 10.1207/S15327752JP720209
- Milli Eğitim Bakanlığı (MEB). (2017a). *Akademik becerilerin izlenmesi ve değerlendirilmesi (ABİDE) 2016 8. sınıflar raporu*. Erişim Adresi: https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_11/30114819_iY-web-v6.pdf
- Milli Eğitim Bakanlığı (MEB). (2017b). *İzleme değerlendirme raporu 2016*. Erişim Adresi: http://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_06/23161120_2016_izleme_degYerlendirme_raporu.pdf
- Page, E. B. (1966). The imminence of grading essays by computers. *Phi Delta Kappan*, 47(5), 238-243. Retrieved from <http://www.jstor.org/stable/20371545>
- Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the “gold standard”. *Applied Measurement in Education*, 28(2), 130-142. doi: 10.1080/08957347.2014.1002920
- Preston, D., & Goodman, D. (2012). *Automated essay scoring and the repair of electronics*. Retrieved from <https://www.semanticscholar.org/>
- R Core Team. (2018). *R: A language and environment for statistical computing* (version 3.5.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.
- Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18(1), 25-39. <https://doi.org/10.1016/j.asw.2012.10.004>
- Senay, A., Delisle, J., Raynauld, J. P., Morin, S. N., & Fernandes, J. C. (2015). Agreement between physicians' and nurses' clinical decisions for the management of the fracture liaison service (4iFLS): The Lucky Bone™ program. *Osteoporosis International*, 27(4), 1569-1576. doi: 10.1007/s00198-015-3413-6
- Shankar, V., & Bangdiwala, S. I. (2014). Observer agreement paradoxes in 2x2 tables: Comparison of agreement measures. *BMC Medical Research Methodology*, 14(100). Advance online publication. <https://doi.org/10.1186/1471-2288-14-100>
- Shermis, M. D. (2010). Automated essay scoring in a high stakes testing environment. In V. J. Shute, B. J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 167-185). New York: Springer.
- Shermis, M. D., & Burnstein, J. (2003). *Automated essay scoring*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257-268. <https://doi.org/10.1093/ptj/85.3.257>
- Siriwardhana, D. D., Walters, K., Rait, G., Bazo-Alvarez, J. C., & Weerasinghe, M. C. (2018). Cross-cultural adaptation and psychometric evaluation of the Sinhala version of Lawton Instrumental Activities of Daily Living Scale. *Plos One*, 13(6), 1-20. <https://doi.org/10.1371/journal.pone.0199820>
- Taghipour, K., & Tou Ng, H. (2016). *A neural approach to automated essay scoring*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 1882-1891. doi: 10.18653/v1/D16-1193
- Vanbelle, S. (2016). A new interpretation of the weighted Kappa coefficients. *Psychometrika*, 81(2), 399-410. <https://doi.org/10.1007/s11336-014-9439-4>
- Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, 6(2). Retrieved from <http://www.jtla.org>
- Wang, Y., Wei, Z., Zhou, Y., & Huang, X. (2018, November). Automatic essay scoring incorporating rating schema via reinforcement learning. In E. Reloff, D. Chiang, H. Julia & T. Jun'ichi (Eds.), *Empirical methods in natural language processing* (pp. 791-797). Brussels, Belgium: Association for Computational Linguistics.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2-13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
- Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13(61), 1-9.

Ek A. Yazılımın Geliştirilmesinde Kullanılan 2 Kategorili Puanlanan Örnek Madde

GÜZEL ATLAR ÜLKESİ: KAPADOKYA



Kapadokya neresidir? Bir şehir, bir ülke yoksa bir bölge midir? Neden her yıl binlerce insan orayı ziyaret eder, yüzlerce kilometre öteden görmeye gelir, dağları geçer, denizleri aşar? Peki, Kapadokya'da ilk önce nereyi ziyaret etmek gerekir? Ne güzel sorular bunlar değil mi! İnsan, öğrenmeye merak etmekle başlar. Sorular sorar, araştırır, bulur, öğrenir. Öğrendikçe de daha bilgili, daha cesur, daha güvenli olur.

Kapadokya, Anadolu ya da Mezopotamya gibi bir bölgenin adı. Nevşehir ilinin sınırları içinde, çok geniş bir alan. 25.000 kilometrekare. Yalnız, oldukça ilginç bir bölge. Bu sebeple binlerce insan her yıl oraya geliyor. Öyle bir bölge ki tarihi "Yontma Taş Devri"ne kadar uzanıyor. Sırasıyla Hititler, Persler, Bizanslılar, Selçuklular ve Osmanlılar yaşamış Kapadokya'da.

Birinci paragraftaki soruların hangisinin cevabı ikinci paragrafta yoktur?

Madde No	16
Bağlam Adı	Güzel Atlar Ülkesi: Kapadokya
Doğru Yanıt (1 Puan) Açıklama	"Kapadokya'da ilk önce nereyi ziyaret etmek gerekir?" sorusuna atıfta bulunan cevaplar doğru cevap olarak kabul edilecektir.
Yanlış Yanıt (0 Puan) Açıklama	Boş cevap ve "Kapadokya'da ilk önce nereyi ziyaret etmek gerekir?" sorusuna atıfta bulunan cevapların haricindeki tüm cevaplar yanlış olarak kabul edilecektir.
Örnek Doğru Yanıtlar	- Peki Kapadokya'da en önce nereyi ziyaret etmek gerekir - Kapadokya'da ilk önce nereyi ziyaret etmek gerekir? sorusunun cevabı yoktu?
Örnek Yanlış Yanıtlar	- Kapadokya'yı ziyarete gelen ilk önce nereye gider? - Kapadokya neresidir? Sorusunun cevabı yok - NEDEN Binlerce insan orayı ziyaret eder? Peki Kapa dokyada ilk nereyi ziyaret etmek gerekir? - Bir şehirmi yoksa bir ülkemidir

Ek B. Yazılımın Geliştirilmesinde Kullanılan 3 Kategorili Puanlanan Örnek Madde

BESLENME

Beslenme çantamda;
Bir dilim ekmek,
Az peynir,
İki bilye, bir topaç
Bir de masal kitabı var.

Gülmeyin arkadaşlar!
Ruhum da doymalı,
Karnımın doyduğu kadar.

Şiire göre, çocuk ruhunu nasıl doyurmaktadır?

Madde No	20
Bağlam Adı	Beslenme
Doğru Yanıt (2 Puan) Açıklama	Çocuğun ruhunu; oyun oynayarak ve kitap okuyarak doyurduğunu ifade eden tüm cevaplar doğru kabul edilir.
Kısmi Doğru Yanıt (1 Puan) Açıklama	Oyun oynar ve kitap okur ifadelerinden sadece birini içeren cevaplar kısmi cevap olarak kabul edilir.
Yanlış Yanıt (0 Puan) Açıklama	Yanlış, ilgisiz ve metinden aynen alınan ifadeler.
Örnek Doğru Yanıtlar	- İki bilyeyi ve bir tane topacı oynayıp, bir masal kitabı okuyarak doyurmaktadır. - 1 bilye bir topaç birde masal kitab okuyup oyunayı Ruhudoyar - Beslenerek, eğlenerek ve okuyarak. - okuyarak ruhunu doyurma isteğiyle - eğlenerek doyuruyo
Örnek Kısmi Doğru Yanıtlar	- Kitap okuyarak, kendini kitabın içine koyarak, ruhunu geliştirip, hissederek. - . iki bilye bir topaç birde masal Kitabı ruhunu doyurmuştur
Örnek Yanlış Yanıtlar	- bir dilim ekmek ,az peynir, iki bilye, bir topaç birde masal kitabı var. - Çocuk ruhunu masal kitabıyla doyurur.

Ek C. ABİDE 2016 Türkçe Testi Örnek Madde Grubu 1

İSTANBUL DEĞİŞİYOR

İstanbul'da beklenmedik bir şekilde nüfusun artması; gecekonduların çoğalmasına, altyapının kurulmasında sorunlar yaşanmasına neden olmaktadır. Kentlerin dokusunda ise önemli değişimler görülmektedir.

İstanbul'un eski semtleri olan Beyoğlu, Sirkeci, Eminönü ve Beyazıt'ta ara sokaklarda taş veya ahşap binalar, birbirini kesen dar sokaklar ve caddeler yer almaktadır. Bakırköy, Caddebostan, Etiler, Nişantaşı, Levent gibi yeni semtlerde çoğu kez doğrusal uzanış gösteren ve birbirini dik kesen cadde ve sokaklar vardır. Ataköy, Bahçeşehir gibi planlı olarak kurulan semtlerde ise daha düzenli caddeler yer almakta, çok katlı binalar yapılmaktadır.

7 - 9. soruları yukarıdaki metne göre yanıtlayınız.

7. Nüfusun olağan dışı artması beraberinde hangi sorunları getirmektedir? Yazınız.

8. Metni göz önünde bulundurduğunuzda fotoğrafta görülen yer İstanbul'un hangi semti olabilir? Gerekçesiyle yazınız.



9. Metinde altı çizili sözcükle anlatılmak istenen aşağıdakilerden hangisidir?

- A) Yapı
- B) Büyüklük
- C) Kapladığı alan
- D) Gelişmişlik düzeyi

“İSTANBUL DEĞİŞİYOR” Bağlamına Ait Puanlama Anahtarı

Soru No:	5
Soru Kodu:	T-2016-0007
Bağlam Adı:	İSTANBUL DEĞİŞİYOR
DOĞRU YANIT- (2 PUAN)Açıklama	Gecekonduların çoğalması VE altyapı problemlerinin artması sorunlarının her ikisine birden vurgu yapan YA DA bu sorunları genelleyen ifadeleri içeren yanıtlar

Ek C (devamı). ABİDE 2016 Türkçe Testi Örnek Madde Grubu 1

Örnek Yanıtlar	Çarpık kentleşme ve imar sorunları
	Gecekonduların artması ve altyapı problemleri
	Gecekonduların artması ve yapılan yolların yeterli olmaması
KISMI DOĞRU- (1 puan) Açıklama	Metinde geçen iki sorundan "gecekonduların çoğalması" YA DA "alt yapı problemlerinin artması" ifadelerinden sadece birini içeren yanıtlar
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar verir
Örnek Yanıtlar	Kentlerin dokusunda önemli değişimler görülmektedir
BOŞ-Açıklama	Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	6
Soru Kodu:	T-2016-0008
Bağlam Adı:	İSTANBUL DEĞİŞİYOR
DOĞRU YANIT- (2 PUAN) Açıklama	"Beyoğlu, Sirkeci, Eminönü, Beyazıt semtlerinden birinin, birkaçının veya hepsinin adını içeren, gerekçe olarak "Ara sokaklarda taş veya ahşap binalar bulunur." YA DA "Birbirini kesen dar sokaklar ve caddeler bulunur." ifadelerinden birini içeren yanıtlar
Örnek Yanıtlar	Beyoğlu çünkü evler ahşap.
	Sirkeci, Eminönü çünkü ara sokaklarda taş veya ahşap binalar bulunur.
KISMI DOĞRU-(1 puan) Açıklama	Sadece semt adını içeren ancak gerekçenin yazılmadığı yanıtlar
Örnek Yanıtlar	Beyoğlu
	Eminönü, Beyazıt
	Beyoğlu, Sirkeci, Eminönü, Beyazıt
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar
BOŞ-Açıklama	Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Ek D. ABİDE 2016 Türkçe Testi Örnek Madde Grubu 2

Soru No:	7
Soru Kodu:	T-2016-0009
Bağlam Adı:	İSTANBUL DEĞİŞİYOR
Doğru Yanıt	A

BASINDA OBEZİTE

10.01.2015

12 Yaş Altı Çocuklarda Mobil Cihazların Kullanımının Yasaklanması İçin Bir Sebep: Obezite

Video oyunları ve televizyon, obezitenin artması ile ilişkilidir. Odasında bu tür cihazları kullanmasına izin verilen çocuklarda obezite görülme sıklığı %30 oranında artmaktadır. Obez olan çocukların %30'unda diyabet ortaya çıkmakta, kalp krizi ve erken felç riski artmakta ve ortalama yaşam süresi kısalmaktadır.

15.12.2014

Çocukluk Döneminde Risk: Obezite

Anne ve babanın obez olması, çocuğun yeme alışkanlığı bakımından anne ve babasını örnek alması, çocukların televizyon ve bilgisayar başında çok zaman geçirmesi, stres, kaygı gibi unsurlar çocukluk döneminde obezitenin oluşmasına neden olmaktadır.

10.11.2014

Çocukları Obez Olan Ailelere Para Cezası Geliyor!

Porto Riko'da hükümet, obeziteyle mücadele amaçlı, çocukları fazla kilolu olan anne ve babalara 800 dolara kadar para cezası verilmesini planlıyor. Gelecek nesillerin daha sağlıklı olması için bu uygulamanın yararlı olacağını düşünenlerin sayısı ülkede oldukça fazla.

10 - 12. soruları yukarıdaki metne göre yanıtlayınız.

10. Gazetelerde obeziteyle ilgili haberlere sıklıkla yer verilmesinin nedeni nedir? Bir ya da iki cümleyle yazınız.

11. Mobil cihazların kullanımı obeziteyi neden artırır? Bir ya da iki cümleyle yazınız.

12. Gazete haberlerine göre aşağıdakilerden hangisi söylenebilir?

- A) Obezite ve diyabet birbirleriyle ilişkilidir.
- B) Televizyon izlemeyen çocuklar obeziteye yakalanmıyor.
- C) Porto Riko'daki para cezası birçok ülkeye örnek olmuştur.
- D) Obezite yalnızca çocukluk döneminde ortaya çıkan bir sorundur.

“BASINDA OBEZİTE” Bağlamına Ait Puanlama Anahtarı

Soru No:	10
Soru Kodu:	T-2016-0010
Bağlam Adı:	BASINDA OBEZİTE

Ek D (devamı). ABİDE 2016 Türkçe Testi Örnek Madde Grubu 2

DOĞRU YANIT- (2 PUAN)Açıklama	Obezite ile ilgili bilinçlendirmeye vurgu yapan yanıtlar
Örnek Yanıtlar	“Obezitenin yaygınlaşmasını önlemek için.” “Halkı bilinçlendirmek için.” “Obezitenin bir hastalık olduğuna dikkat çekmek.” “Halkı uyarmak için.” “Aileleri bilinçlendirmek için.” “Anne ve babaların önlem almasını sağlamak için.” vb.
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar
Örnek Yanıtlar	Para cezasını haber vermek için
BOŞ-Açıklama	- Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	11
Soru Kodu:	T-2016-0011
Bağlam Adı:	BASINDA OBEZİTE
DOĞRU YANIT- (1 PUAN) Açıklama	“Uzun süre hareketsiz kalma, çocukların televizyon ve bilgisayar başında çokça vakit geçirmesi” ifadelerini içeren yanıtlar
Örnek Yanıtlar	“Çocukların bilgisayar ve televizyon başında çok zaman geçirmesi.” “Çocukların bilgisayar başında çok zaman geçirmesinden dolayı hareketsiz kalması.”
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar
BOŞ-Açıklama	Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	12
Soru Kodu:	T-2016-0012
Bağlam Adı:	BASINDA OBEZİTE
Doğru Yanıt	A