T.C.
AKDENİZ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
EĞİTİM BİLİMLERİ ANA BİLİM DALI

YÜKSEK LİSANS TEZİ

OTOMATİK PUANLAMA SİSTEMLERİ: SİSTEMATİK BİR GÖZDEN GEÇİRME

ÇAĞRIŞIM HELHEL BAYRAKTAROĞLU

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME
TEZLİ YÜKSEK LİSANS PROGRAMI

Antalya, 2025

T.C.
AKDENİZ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
EĞİTİM BİLİMLERİ ANA BİLİM DALI
EĞİTİMDE ÖLÇME VE DEĞERLENDİRME
TEZLİ YÜKSEK LİSANS PROGRAMI

OTOMATİK PUANLAMA SİSTEMLERİ: SİSTEMATİK BİR GÖZDEN GEÇİRME

YÜKSEK LİSANS TEZİ

Çağrışım Helhel Bayraktaroğlu

Danışman

Doç. Dr. Bilal Barış Alkan

Antalya, 2025

DOĞRULUK BEYANI

Yüksek lisans tezi olarak sunduğum bu çalışmayı, bilimsel ahlak ve geleneklere aykırı düşecek bir yol ve yardıma başvurmaksızın yazdığımı, yararlandığım eserlerin kaynakçalardan gösterilenlerden oluştuğunu ve bu eserleri her kullanışımda alıntı yaparak yararlandığımı belirtir; bunu onurumla doğrularım. Enstitü tarafından belli bir zamana bağlı olmaksızın, tezimle ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara katlanacağımı bildiririm.

27/01/2025

Çağrışım Helhel Bayraktaroğlu

T.C.

AKDENİZ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Çağrışım HELHEL BAYRAKTAROĞLU'nun bu çalışması 27/01/2025 tarihinde jürimiz tarafından Eğitim Bilimleri Ana Bilim Dalı Eğitimde Ölçme ve Değerlendirme Tezli Yüksek Lisans Programında **Yüksek Lisans Tezi** olarak **oy birliği/oy çokluğu** ile kabul edilmiştir

İMZA

Başkan :

Üye :

Üye (Danışman**) :**

YÜKSEK LİSANS TEZİNİN ADI: Otomatik Puanlama Sistemleri: Sistematik Bir Gözden Geçirme

ONAY: Bu tez, Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüş ve Enstitü Yönetim Kurulunun …….. tarihli ve …..……. sayılı kararıyla kabul edilmiştir.

TEŞEKKÜR

ÖZET

OTOMATİK PUANLAMA SİSTEMLERİ: SİSTEMATİK BİR GÖZDEN GEÇİRME

HELHEL BAYRAKTAROĞLU, Çağrışım

Yüksek Lisans, Eğitim Bilimleri Ana Bilim Dalı,

Tez Danışmanı: Doç. Dr. Bilal Barış Alkan

Ocak 2025, 61 sayfa

Eğitimde ölçme ve değerlendirme süreçlerindeki hız, tutarlılık ve objektiflik gibi gereksinimler; otomatik metin puanlama sistemlerinin önemini her geçen gün artırmaktadır. Bu bağlamda tez çalışmasında, otomatik metin puanlama sistemlerinin mevcut durumunun ve eğitimdeki katkılarının, metin madenciliği teknikleri ile entegre bir sistematik literatür taraması yöntemiyle analiz edilmesi amaçlanmaktadır.

Çalışmada, ACM, ACL, IEEE Explore, Springer ve Science Direct gibi akademik veri tabanlarından, belirli anahtar kelimelerle elde edilen 397 makale, önceden belirlenen dahil etme ve hariç tutma kriterleri doğrultusunda 182 makaleye indirgenmiştir. Bu makaleler; kullanılan veri setleri, otomatik puanlama türleri, değerlendirme metrikleri, içerik ve makine öğrenme teknikleri gibi kategoriler altında detaylı bir şekilde incelenmiş ve kapsamlı bir veri tablosunda sınıflandırılmıştır. Ayrıca, Latent Dirichlet Allocation temelli metin madenciliği teknikleri kullanılarak, makalelerdeki temel konular ve eğilimler analiz edilmiş ve kelime bulutları ile görselleştirilmiştir.

Yapılan sistematik literatür taraması neticesinde; otomatik puanlama sistemleriyle ilgili temel eğilimler ve odak noktaları ayrıntılı bir şekilde ortaya konulmuştur. Çalışmaların yayın yıllarına ilişkin analizlerde, özellikle 2010 yılından itibaren bu alanda yapılan çalışmaların belirgin bir şekilde artış gösterdiği tespit edilmiştir. Veri setleri açısından yapılan değerlendirmelerde, TOEFL, ASAP ve Kaggle gibi geniş ölçekli ve çeşitli kaynakların araştırmalarda yoğun olarak kullanıldığı belirlenmiştir. Otomatik puanlama türleri incelmiş ve "semantic," "deep learning," ile "rubric " temelli yaklaşımların öne çıktığı görülmüştür. Ayrıca, makine öğrenmesi tekniklerinde, "neural networks," "BERT," ve "LSTM" gibi modern yöntemlerin sıklıkla tercih edildiği saptanmıştır. Doğrulama metrikleri bakımından ise "Quadratic Weighted Kappa (QWK)" ve "correlation" gibi güvenilirlik ölçütlerinin yaygın olarak kullanıldığı ortaya konmuştur.

Son olarak, Türkiye'de bu alanda yapılan akademik çalışmaların sınırlı sayıda olması göz önüne alındığında, bu çalışma hem literatüre hem de eğitimde ölçme ve değerlendirme alanında teknoloji temelli bu uygulamaların kullanımı konusunda; araştırmacılara ve eğitim üzerine karar vericilere rehber niteliği taşımaktadır.

Anahtar Sözcükler: *Otomatik puanlama, Metin puanlama, Doğal dil işleme, Yapay zeka, Derin öğrenme*

ABSTRACT

AUTOMATIC SCORING SYSTEMS: A SYSTEMATIC REVIEW

HELHEL BAYRAKTAROĞLU, Çağrışım

Master of Arts, Department of Educational Sciences

Supervisor: Assoc. Prof. Dr. Bilal Barış Alkan

January 2025, 61 pages

Requirements such as speed, consistency and objectivity in measurement and assessment processes in education are increasing the importance of automated text assessment systems from day to day. Against this background, this study aims to analyse the current status of automated text scoring systems and their contribution to education by means of a systematic literature review using text mining methods.

In the study, 397 articles retrieved with specific keywords from academic databases such as ACM, ACL, IEEE Explore, Springer and Science Direct were reduced to 182 articles according to the previously defined inclusion and exclusion criteria. These articles were analysed in detail according to categories such as datasets used, automatic scoring types, evaluation metrics, content and machine learning techniques and classified in a comprehensive data table. In addition, text mining techniques based on latent Dirichlet Mapping were used to analyse the most important topics and trends in the articles and visualise them with word clouds.

As a result of the systematic literature research, the most important trends and focal points in automatic evaluation systems were identified in detail. When analysing the publication years of the studies, it was found that the number of studies conducted in this area has increased significantly, especially since 2010. When analysing the data sets, it was found that large and diverse sources such as TOEFL, ASAP and Kaggle were used intensively in the studies. Different types of automated assessment were analysed and it was found that "semantic"," "deep learning" and "rubric-based" approaches were predominant. It was also found that modern methods such as "neural networks"," "BERT" and "LSTM" were often favoured among the machine learning methods. In terms of verification metrics, it was found that reliability measures such as "Quadratic Weighted Kappa (QWK)" and "Correlation" were widely used.

Considering the limited number of academic studies conducted in this field in Turkey, this study serves as a guide for researchers and decision makers in the field of education for the

use of these technology-based applications in the field of measurement and assessment in education as well as for the literature.

Keywords: *Automatic scoring, Text scoring, Natural language processing, Artificial intelligence, Deep learning*

İÇİNDEKİLER

BÖLÜM II

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

BÖLÜM III

YÖNTEM

BÖLÜM IV

BULGULAR

BÖLÜM V

SONUÇ, TARTIŞMA VE ÖNERİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

KISALTMALAR LİSTESİ

OMP: Otomatik Metin Puanlama

MÖ: Makine Öğrenmesi

DDİ: Doğal Dil İşleme

YZ: Yapay Zeka

BÖLÜM I
GİRİŞ

1.1. Problem Durumu

Son yıllarda, online eğitim sistemleri eğitim ve öğretimin ayrılmaz bir parçası haline gelmiş ve özellikle bazı alanlarda hızla yerini sağlamlaştırmıştır. Kurum çalışanlarına yönelik eğitim programları, üniversitelerdeki uzaktan eğitim dersleri ve kişisel gelişim kursları gibi birçok farklı bağlamda online eğitim yaygın bir şekilde kullanılmaktadır. Teknolojinin hızlı gelişimiyle birlikte, online eğitim süreçlerinin değerlendirilmesi ve ölçme-değerlendirme uygulamalarının otomatik hale getirilmesine yönelik çalışmalar önemli bir ivme kazanmıştır. Özellikle TOEFL, IELTS, GRE, e-YDS ve e-ALES gibi geniş katılımlı sınavların yanı sıra, bu tür sınavların online ortamlara taşınması, otomatik değerlendirme sistemlerine olan ihtiyacı artırmıştır. Otomatik puanlama sistemlerinin kullanım alanları yalnızca bilgisayar ortamında gerçekleştirilen sınavlarla sınırlı kalmamaktadır. Günümüzde el yazısını algılayabilen teknolojilerin gelişmesi, klasik yazılı sınavların da dijital ortamda değerlendirilebilmesine olanak sağlamaktadır (Dong vd., 2017). Çoktan seçmeli sınavların otomatik olarak okunup değerlendirilmesi uzun süredir uygulanabilir bir süreç olmakla birlikte, yazılı yanıtların değerlendirilmesinde hâlâ büyük ölçüde alanında uzman insanlara ihtiyaç duyulmaktadır. Özellikle büyük ölçekli sınavlarda, birden fazla değerlendirici aynı cevapları puanlamakta ve bu süreç, çeşitli hatalara açık bir yapıya sahiptir (Attali & Burstein, 2006). Bu bağlamda, değerlendirme süreçlerinde oluşabilecek tutarsızlıkları en aza indirmek için önceden belirlenmiş ve netleştirilmiş kriterlerin varlığı büyük önem taşımaktadır (E. B. Page, 1966). Ancak, insan değerlendirmesine dayalı süreçlerde, puanlama sonrası yapılan itirazlar, sınavların tekrar değerlendirilmesi ve kimi durumlarda puanların değiştirilmesi gerekebilmektedir. Bu tür süreçlerin otomatik hale getirilmesi, sınav sonuçlarının daha tutarlı ve güvenilir olmasını sağlayabilir (Burstein & Chodorow, 1999). Ayrıca, değerlendirme süreçlerinde yer alan geniş uzman kadrosuna ödenen maliyetlerin azaltılması ve sınav sonuçlarının daha kısa sürede açıklanması gibi avantajlar da göz önünde bulundurulduğunda, otomatik puanlama sistemlerinin eğitim dünyasına önemli katkılar sunduğu görülmektedir (Zhang & Litman, 2018). Otomatik puanlama sistemleri, yalnızca hız açısından değil, aynı zamanda ölçme-değerlendirme süreçlerinde insan hatasını en aza indirerek değerlendirme

kalitesini artırma potansiyeline de sahiptir. Bu sistemler, online eğitimin yaygınlaşmasıyla birlikte, eğitimde değerlendirme süreçlerinin etkinliğini ve ekonomikliğini artırma gereksinimiyle doğrudan ilişkilidir. Otomatik puanlama sistemleri, insan müdahalesine duyulan ihtiyacı azaltarak, daha hızlı, daha güvenilir ve daha az maliyetli ölçme-değerlendirme uygulamaları sunmaktadır (Kitchenham & Charters, 2007). Eğitimde dijitalleşmenin bir parçası olarak, bu teknolojiler yalnızca akademik çevrelerde değil, aynı zamanda iş dünyasında ve kişisel gelişim alanlarında da giderek önem kazanmaktadır.

Otomatik metin puanlama sistemleri, kısa cevapların ve kompozisyonların değerlendirilmesinde kullanılan yenilikçi teknolojiler arasında yer almakta ve eğitim süreçlerinde giderek daha yaygın bir şekilde kullanılmaktadır. Kısa cevapların ve kompozisyonların otomatik değerlendirilmesi sürecinde karşılaşılan en büyük zorluklardan biri, aynı soruya verilebilecek pek çok farklı doğru cevabın bulunmasıdır. Bu durum, basit programlama sistemleri ile üstesinden gelinmesi oldukça zor bir problem teşkil etmektedir. Otomatik puanlama sistemleri, bu zorluğu aşmak için genellikle doğal dil işleme (Natural Language Processing, NLP) ve makine öğrenimi (Machine Learning, ML) tekniklerini bileştiren karmaşık algoritmalar ve sistemler kullanır. Bu sistemlerin geliştirilmesine yönelik çalışmaların tarihi, 1960'lara kadar uzanmakta olup, bu alandaki ilk örneklerden biri olan Project Essay Grade (PEG), Ellis B. Page tarafından 1966 yılında geliştirilmiştir (Larkey, 1998). PEG'in ilk versiyonu, kelime uzunluğu, nadir kelime kullanımı ve noktalama işaretlerinin sıklığı gibi metrikleri analiz ederek, insan değerlendiricilerin değerlendirme biçimini taklit etmeyi amaçlamıştır (E. B. Page, 1966). Daha sonra 1973 yılında geliştirilen ikinci versiyonu ise regresyon modelleri kullanılarak, yazının "özellik kümelerini" (proxes) insan değerlendirmeleriyle eşleştirmiştir (Shermis ve Burstein, 2003).

1999 yılında Foltz, Laham ve Landauer tarafından geliştirilen Intelligent Essay Assessor (IEA), semantik benzerlik ölçümleri yapmak için Öğrenme Anlamsal Analizi (Latent Semantic Analysis, LSA) yöntemini kullanarak öğrenci kompozisyonlarını değerlendirme sürecine yeni bir boyut kazandırmıştır (Landauer vd., 1998). Aynı yıl, Educational Testing Service (ETS) tarafından geliştirilmeye başlanan E-rater, NLP tekniklerini kullanarak özellikle TOEFL gibi standartlaştırılmış sınavlarda öğrenci yazılarını insan değerlendiricilere paralel bir şekilde puanlama kapasitesine sahip olmuştur (Attali ve Burstein, 2006). 2000 yılında geliştirilen BETSY (Bayesian Essay Test Scoring sYstem) ise Bayes teoremi temelli istatistiksel

modelleme yöntemlerini dil özellikleriyle birleştirerek bu alandaki öncü sistemlerden biri olmuştur (Rudner ve Liang, 2002).

2010'lu yıllardan sonra, derin öğrenme tabanlı yaklaşımlar otomatik puanlama alanında köklü bir değişim yaratmıştır. Örneğin, Dong ve çalışma arkadaşları tarafından önerilen bir modelde, tekrarlayan sinir ağları (Recurrent Neural Networks, RNN) ve evrişimli sinir ağları (Convolutional Neural Networks, CNN) bir arada kullanılarak dikkat mekanizması yardımıyla önemli metin parçalarına odaklanılmıştır (Dong vd., 2017). Bunun yanı sıra, Zhang ve Litman'ın çalışması olan "Co-attention Based Neural Network for Source-dependent Essay Scoring", kaynak metin ile öğrenci yazıları arasındaki bağlamsal ilişkileri analiz ederek, içerik uyumu ve anlamını değerlendirme sürecine entegre etmiştir (Zhang ve Litman, 2018). Geleneksel otomatik puanlama sistemlerinin yalnızca yüzeysel metriklere odaklanmasının aksine, bu yenilikçi yaklaşımlar, yazılı içeriklerin bağlam içerisindeki anlamını ve kaynak metinle olan semantik ilişkisini ortaya çıkarma yeteneği sunmuştur.

Günümüzde otomatik puanlama sistemleri, öğrencilerin yazma becerilerinin değerlendirilmesi sürecinde yaygın bir şekilde kullanılmaktadır. Örneğin, öğrencilerin metin kanıtlarını nasıl kullandığını ve argümanlarını ne kadar iyi organize ettiğini ölçen Response-to-Text Assessment (RTA) testi, bu sistemlerin bir uygulaması olarak öne çıkmaktadır (Rahimi vd., 2017). Ayrıca, TOEFL ve GMAT gibi standartlaştırılmış testlerde OMP sistemleri hem insan değerlendirmesini desteklemek hem de bağımsız bir ölçüm aracı olarak hizmet etmektedir (Attali ve Burstein, 2006).

Otomatik puanlama sistemleri, yalnızca ölçme-değerlendirme süreçlerini daha etkin ve ekonomik hale getirmekle kalmamakta, aynı zamanda insan hatalarını azaltarak sınav sonuçlarının güvenilirliğini artırmakta ve daha hızlı bir değerlendirme süreci sunmaktadır. Bu nedenle, eğitim ve değerlendirme alanında gelecekteki uygulamaları şekillendirecek önemli bir teknoloji olarak dikkat çekmektedir.

Literatür incelendiğinde, otomatik puanlama sistemleri üzerine yapılan çalışmaların uluslararası düzeyde yoğunlaştığı, ancak Türkiye'de bu alandaki akademik üretimin sınırlı olduğu gözlemlenmiştir. Özellikle, doğal dil işleme ve makine öğrenimi tabanlı yaklaşımlar uluslararası literatürde ayrıntılı şekilde ele alınırken (Deane, 2013), Türkiye bağlamında bu sistemlerin teorik ve uygulamalı yönlerini inceleyen çalışmaların eksik olduğu dikkat çekmektedir. Bu durum, Türkiye'deki eğitim süreçlerinde otomatik puanlama sistemlerinin

daha etkin kullanımını sağlamak için kapsamlı ve bağlamsal araştırmalar yapılması gerekliliğini ortaya koymaktadır.

Bu doğrultuda çalışmanın temel amacı, metin madenciliği ile entegre bir sistematik literatür taraması yöntemi kullanarak otomatik puanlama sistemleri üzerine yapılan akademik çalışmaları kapsamlı bir şekilde incelemek ve bu alandaki mevcut durumu detaylı bir çerçevede ortaya koymaktır. Özellikle, otomatik metin puanlama sistemlerinde kullanılan veri setleri, otomatik puanlama türleri, içerik analizi, makine öğrenimi teknikleri ve doğrulama metrikleri açısından değerlendirilmesi hedeflenmiştir. Çalışma, bu alandaki yöntem ve uygulamaları değerlendirmenin yanı sıra, otomatik puanlama sistemlerinin gelişimine katkı sağlayacak öneriler sunmayı amaçlamaktadır. Araştırmanın önemi, eğitimde ölçme ve değerlendirme süreçlerinde otomatik puanlama sistemlerinin sunduğu potansiyeli vurgulamasıyla öne çıkmaktadır. Bu sistemlerin geliştirilmesi, ölçme ve değerlendirme süreçlerini daha hızlı, tutarlı ve ekonomik bir hale getirme potansiyeline sahiptir. Ayrıca, bu sistemlerin insan değerlendirmesiyle uyumlu ve güvenilir sonuçlar üretme kapasitesi, eğitim alanında daha adil ve objektif değerlendirmeler yapılmasının önünü açmaktadır. Bu bağlamda, tez çalışması, teknolojik yeniliklerin eğitim süreçlerine uygulanabilirliğini göstermesi ve mevcut literatürdeki eksiklikleri gidermesi açısından önemli bir katkı sunmaktadır.

Çalışmada sistematik literatür taraması yöntemi benimsenmiş, ACM, ACL, IEEE Explore, Springer ve Science Direct gibi akademik veri tabanlarından toplam 182 makale detaylı bir şekilde analiz edilmiştir. Analiz sürecinde, R programı kullanılarak belge-terim matrisleri oluşturulmuş, Latent Dirichlet Allocation (LDA) yöntemi ile konular modellenmiş ve kelime bulutları yardımıyla bulgular görselleştirilmiştir. Bu yöntemle elde edilen sonuçlar, otomatik puanlama sistemlerinin mevcut eğilimlerini, karşılaşılan zorlukları ve gelişim potansiyelini kapsamlı bir şekilde ortaya koymuştur.

1.2. Araştırmanın Amacı ve Problemleri

Bu araştırmanın genel amacı; otomatik puanlama sistemlerinin mevcut durumunu metin madenciliği yöntemleriyle entegre bir sistematik literatür taraması ile analiz etmek ve bu sistemlerin eğitimdeki değerlendirme süreçlerine katkısını ortaya koymaktır.

1.2.1. Araştırmanın Alt Problemleri

Araştırmanın amacı doğrultusunda cevap aranacak olan araştırma problemleri şunlardır:

i. Otomatik puanlama sistemleri üzerine yapılan araştırmalar için mevcut veri setleri nelerdir?
ii. Otomatik puanlama sistemlerine ilişkin çalışmalarda kullanılan otomatik puanlama türleri nelerdir?
iii. Otomatik puanlama sistemlerine ilişkin çalışmalarda ve ilgili sistemlerin geliştirilmesinde kullanılan veri setleri nelerdir?
iv. Otomatik puanlama sistemlerine ilişkin çalışmalarda hangi konu ve içerikler değerlendirilmiştir?
v. Otomatik puanlama sistemlerine ilişkin çalışmalarda kullanılan makine öğrenme teknikleri nelerdir?
vi. Otomatik puanlama sistemlerine ilişkin çalışmalarda sıklıkla kullanılan değerlendirme metrikleri nelerdir?

1.3. Araştırmanın Önemi

Otomatik metin puanlama sistemleri, güncel teknolojik yöntem ve gelişmelerin ışığında geleneksel puanlama yöntemlerindeki öznellik, tutarsızlık, zaman alıcılık gibi sınırlamaları gidererek ölçme ve değerlendirme alanına alternatif bir bakış açısı getirmiştir. Özellikle son dönemlerde, adından sıklıkla söz ettiren gelişmiş yapay zeka (YZ) teknolojilerinden ve doğal dil işleme (DDİ) tekniklerinden yararlanan OMP sistemleri; hızlı, tutarlı ve ölçeklenebilir notlandırma çözümleri sunmaktadır. Ancak kullanımı ve uygulama alanlarının genişletilmesi doğrultusunda, eğitimde ölçme ve değerlendirme bağlamında çok ciddi katkıları bulunacağı gerçeğine rağmen; ilgili konuda, ülkemizde oldukça sınırlı sayıda çalışmanın yer aldığı ve özellikle bu alanda oldukça sınırlı sayıda tez ve çalışmanın bulunduğu görülmektedir. Bu doğrultuda tezin; OMP sistemlerinin eğitim bağlamında hangi konu ve alanlarda, hangi amaçlarla kullanılabileceği ve ne tür katkılar sağlayabileceği yönünde kapsamlı şekilde bilgi vermesi ve bu bilginin ilgili alandaki çalışmaları etkili bir sistematik literatür taraması yöntemiyle sunması hedeflendirmiştir. Bu bağlamda tezin en önemli katkısı; OMP sistemlerinden faydalanma veya ilgili sistemleri geliştirme noktasında etkili ve güvenilir bir yol

haritasına ihtiyaç duyan araştırmacılara, eğitimcilere ve eğitimle ilgili karar vericilere sistematik bir kaynak sunmaktır.

1.4. Araştırmanın Sayıltıları

Otomatik metin puanlama sistemlerinin, doğal dil işleme ve yapay zeka gibi veri madenciliği temelli tekniklerden yararlanarak; insan değerlendiriciler kadar tutarlı, etkili ve doğru değerlendirme yapabileceği kabul edilmektedir.

1.5. Araştırmanın Sınırlılıkları

Otomatik metin puanlama sistemlerine ilişkin çalışmalara ulaşma ve sistematik literatür taramasına dahil etme noktasında, erişim sağlanamayan akademik çalışmalar bulunmaktadır. Ancak bu çalışmaların sayı ve niteliği, araştırma sonuçlarını kayda değer ölçüde etkilememektedir.

1.6. Tanımlar

Otomatik Puanlama Sistemi: Otomatik puanlama sistemi, dilbilimsel, yapısal ve içerik tabanlı özellikleri analiz etmek için doğal dil işleme ve makine öğrenimi tekniklerini kullanarak yazılı metinleri değerlendiren ve puanlayan bilgisayar tabanlı bir uygulamadır.

Doğal Dil İşleme: Doğal dil işleme; makinelerin insan dilini işlemesini ve analiz etmesini sağlamaya odaklanan; metin analizi, duygu tespiti ve dil üretimi gibi görevleri kolaylaştıran veri madenciliğinin bir alt alanıdır.

Rubrik Tabanlı Değerlendirme: Rubrik tabanlı değerlendirme, yazılı metinlerin nesnel ve tutarlı puanlanmasını sağlamak için tutarlılık, organizasyon, dil bilgisi ve kelime bilgisi gibi önceden tanımlanmış ölçütleri kullanan yapılandırılmış bir değerlendirme çerçevesini ifade eder.

Dönüştürücü Modeller: BERT ve GPT gibi dönüştürücü modeller, metin dizilerini işlemek için dikkat mekanizmalarını kullanan, gelişmiş dil anlayışı için bağlamsal ve anlamsal ilişkileri yakalayan derin öğrenme mimarileridir.

Değerlendirme Metriği: Bir sistemin veya modelin performansını ölçmek ve sonuçlarını objektif olarak değerlendirmek için kullanılan bir ölçüm yöntemidir.

BÖLÜM II
KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

2.1. Otomatik Puanlama Sistemlerinin Tanımı ve Kapsamı

2.1.1. Otomatik Metin Puanlama Sistemlerinin Tanımı

Öğrencilerin yazdığı kısa cevapların ve kompozisyonların değerlendirilmesi sürecinin otomatikleştirilmesi üzerine pek çok çalışma yürütülmüştür. Fakat çoktan seçmeli sorulara verilen cevapların aksine açık uçlu sorulara verilecek birbirinden farklı birden fazla doğru cevap olabilir. Bu durum kısa cevap ve kompozisyonların otomatik olarak değerlendirilip puanlandırılmasının önündeki en büyük zorluklardan biridir. Basit programlama sistemleri ile bu zorluğun üstesinden gelinmesi oldukça zordur. Otomatik puanlama sistemleri içerik, organizasyon, kelime dağarcığı ve dilbilgisi gibi çeşitli kriterlere göre kompozisyonları değerlendirmek ve/veya puanlamak için Doğal Dil İşleme ve Makine öğrenimi tekniklerini kullanan karmaşık sistemlerdir (Westera vd., 2018). Otomatik puanlama sistemleri öğrencilere yazı ve kompozisyonları ile ilgili geri bildirim vererek kendilerini geliştirmelerini sağlamak amacıyla kullanılabileceği gibi büyük ölçekli sınavların değerlendirme sürecinde daha hızlı ve tutarlı puanlama sağlamak amacıyla da kullanılabilir. Aynı zamanda otomatik puanlama sistemlerinin kullanımı insan değerlendiricilere olan bağımlılığı da azaltacağı için bu bağımlılıktan kaynaklanan ekonomik yükü de hafifletecektir.

2.1.2. Eğitim ve Değerlendirme Süreçlerindeki Rolü

Otomatik Metin Puanlama Sistemleri yazma becerilerinin ölçülmesinde kullanılabilecek oldukça güçlü bir araçtır. Bu becerilerin başında dilbilgisi kullanımı gelmektedir. OMP sistemleri, öğrencilerin yazma becerilerini sadece dilbilgisi açısından değil, aynı zamanda içerik organizasyonu, fikir geliştirme ve bağlam uygunluğu gibi daha karmaşık kriterlere göre değerlendirebilir. Değerlendirme süreçlerinde “BERT ” gibi ileri düzey dil modellerinin kullanılmasıyla öğrenci cevapları anlam açısından daha derinlemesine analiz edilebilmiştir (Mardini vd., 2024). Bu tür teknolojiler, yazma becerilerinin standartlaştırılmış ve şeffaf bir şekilde değerlendirilmesine olanak tanır, böylece eğitimde ölçme ve değerlendirme süreçlerinin güvenilirliği artar.

Otomatik Puanlama Sistemleri öğrencilerin yazdıkları kompozisyonların değerlendirilmesi sürecinde hali hazırda kullanılmaktadır. Örneğin, TOEFL ve GMAT gibi standartlaştırılmış testlerde OMP sistemleri, yazılı bölümleri değerlendirmede hem insan değerlendirmesini desteklemede hem de bağımsız bir ölçüm yöntemi olarak kullanılmaktadır (Attali ve Burstein, 2006).

2.1.3. İlk Nesil Sistemler

Otomatik puanlama sistemleri, öğrenci yazılarının objektif, hızlı ve tutarlı bir şekilde değerlendirilmesini sağlamak amacıyla geliştirilmiş teknolojilerdir. Bu sistemlerin temelleri 1960'larda atılmış olup, erken dönem örneklerinde yazıların yüzeysel özelliklerine dayalı analizler ön planda olmuştur.

Project Essay Grade (PEG), 1966 yılında Ellis B. Page tarafından geliştirilen ilk otomatik puanlama sistemi olarak literatürde önemli bir yer tutmaktadır. PEG, kelime uzunluğu, nadir kelime kullanımı ve noktalama işaretlerinin sıklığı gibi metrikleri analiz ederek insan değerlendiricilerin puanlama biçimlerini taklit etmeyi amaçlamıştır (E. B. Page, 1966). PEG'in ikinci versiyonu, yazılardaki "özellik kümelerini" insan değerlendiricilerin puanlamalarıyla eşleştirmek için regresyon analizine dayalı bir sistem sunmuştur (Shermis vd., 2010). 1990'lı yıllarda, daha gelişmiş algoritmaların kullanıldığı sistemler ortaya çıkmıştır. Intelligent Essay Assessor (IEA), Landauer vd., (1998) tarafından geliştirilen ve LSA (Latent Semantic Analysis) algoritmasını kullanan bir sistemdir. Bu sistem, öğrenci yazılarındaki anlamlı içerik ile referans metinler arasındaki anlamsal benzerlikleri analiz ederek yazıların içerik doğruluğunu değerlendirmiştir (Ramesh ve Sanampudi, 2022). Aynı yıl, Educational Testing Service (ETS) tarafından geliştirilen E-rater, doğal dil işleme (NLP) tekniklerinden yararlanarak yazıların dil bilgisi, organizasyon ve kelime seçimi gibi yönlerini incelemiştir (Attali ve Burstein, 2006). Bu sistem, özellikle TOEFL gibi standart sınavlarda yazıların hızlı ve tutarlı bir şekilde değerlendirilmesini sağlamıştır.

İlk nesil otomatik puanlama sistemleri, yüzeysel metriklere odaklanarak yazıları analiz etmede önemli adımlar atmıştır. Ancak bu sistemlerin bağlamsal anlam ve içerik derinliği konularında sınırlı kaldığı bilinmektedir. Yine de PEG, IEA ve E-rater gibi öncü yaklaşımlar, daha karmaşık ve bağlama duyarlı sistemlerin geliştirilmesine zemin hazırlamıştır.

2.1.4. Modern Yaklaşımlar

Geleneksel otomatik puanlama sistemlerinden farklı olarak, modern sistemler doğal dil işleme (NLP) ve makine öğrenimi tekniklerini kullanarak daha yenilikçi ve etkili çözümler sunmaktadır. Bu sistemler derin öğrenme temelli olup, özellikle LSTM (Uzun Kısa Süreli Bellek), Transformer tabanlı modeller ve hibrit yaklaşımlar öne çıkmaktadır.

LSTM (Uzun Kısa Süreli Bellek) modelleri, yazıların dilbilgisel doğruluğunu, düzenini ve yapısını değerlendirirken; Transformer tabanlı modeller, bağlamsal anlamayı güçlendiren yenilikçi özelliklere sahiptir. Örneğin, BERT (Bidirectional Encoder Representations from Transformers) gibi modeller, metinlerin bağlamını etkili bir şekilde analiz ederek daha ayrıntılı değerlendirmeler yapmaktadır (Suresh vd., 2023).

Son yıllarda, LSTM ve Transformer mimarilerinin birleştirilmesiyle oluşturulan hibrit modeller, kompozisyon ve kısa cevap analizlerinde önemli avantajlar sağlamıştır. Bu modeller, kelime ve metin düzeyinde yapılan değerlendirmelerde yüksek doğruluk oranlarına ulaşarak hata payını azaltmaktadır (Chassab vd., 2024). Örneğin, LSTM modelleri dilbilgisel doğruluğu ön planda tutarken, Transformer modelleri yazının bağlamsal anlamını ve organizasyonunu detaylı bir şekilde inceleyebilmektedir.

Modern otomatik puanlama sistemleri, yalnızca yazıları değerlendirmekle kalmaz, aynı zamanda öğrencilere geri bildirim sunarak yazılı ifade becerilerinin gelişimine de katkı sağlar. Böylece bu sistemler, eğitim süreçlerini daha verimli, hızlı ve etkili hale getirebilir.

2.1.5. Büyük Dil Modellerinin (Large Language Models - LLM) Etkisi

Doğal Dil İşleme (NLP) alanındaki gelişmeler, büyük dil modellerinin (LLM) otomatik puanlama sistemlerinde kullanımını önemli ölçüde artırmıştır. Bu modeller geniş veri kümeleri üzerinde eğitildiği için daha esnek bir yaklaşım sunarak matematik ve dil testlerinde doğru yanıtların yanı sıra çözüm sürecinin de değerlendirilebilmesini mümkün kılar (Morris vd., 2024). Büyük dil modeli dendiğinde akla ilk gelen modellerden biri GPT-4'dür. GPT-4'ün zincirleme düşünme (chain-of-thought) teknikleri kullanılarak daha karmaşık soruları çözme kabiliyeti geliştirilmiş ve bu durum, yazılı yanıtların daha derinlemesine analiz edilmesini mümkün kılmıştır (Martínez, 2024). LLM'ler, büyük veri kümeleri üzerinde önceden eğitildikleri için hem eğitim süresini kısaltmakta hem de daha az veriyle yüksek performans sağlayabilmektedir (Dong vd., 2017). LLM'lerin otomatik puanlama sistemlerinin geçerlilik

boyutuna olan katkısı da dikkat çekmektedir. Bu modeller, kriter ve yapısal geçerliliği destekleyen güçlü kanıtlar sunabilmektedir. Böylece değerlendirme süreçlerinin nesnelliğini ve güvenilirliğini artırmaktadır. Chassab vd. (2024), LSTM tabanlı bir dil modelini optimize etmek için Fox algoritmasını kullanarak geliştirdikleri FLSTM-ALM modelinin, otomatik yazı değerlendirme süreçlerinde yüksek doğruluk oranlarına ulaştığını göstermiştir.

2.2. Otomatik Puanlama Sistemlerinin Teknik Temelleri

2.2.1. Doğal Dil İşleme (Natural Language Processing - NLP) Teknikleri

Kelime gömme, kelimeleri matematiksel bir biçimde temsil etmek için kullanılan bir doğal dil işleme (NLP) tekniğidir. Kelime gömme, kelimeler arasındaki anlamsal ve bağlamsal ilişkileri modellemek için eğitilir. Google'ın Word2Vec modeli ve BERT gibi bağlama duyarlı gömme teknikleri, cümle ve kelime seviyesinde daha zengin bilgi sağlayarak özellikle cümle benzerliklerini hesaplama ve yazma kalitesini değerlendirme süreçlerinde kullanılmıştır (Suresh vd., 2023). Mikolov vd. (2013) tarafından geliştirilen Word2Vec, kelimelerin sabit boyutlu vektör temsillerini oluşturur ancak bağlam duyarlılığı sağlamaz. Örneğin 'hücre' kelimesinin her bağlamdaki temsili aynıdır. BERT, bağlam duyarlı bir kelime gömme yöntemidir ve kelimenin çevresindeki diğer kelimelere bağlı olarak farklı vektörler üretir (Ejima ve Takeuchi, 2022). Örneğin 'hücre' kelimesinin çevresindeki kelimelere göre biyoloji ile alakalı olup olmadığını belirleyebilir. RoBERTa, BERT gibi Transformer tabanlı bir model olup BERT'ten farklı olarak büyük miktarda veri üzerinde önceden eğitilir ve dil modellerini daha iyi anlamak için optimize edilmiştir (Beseiso vd., 2021). Beseiso vd. (2021), RoBERTa'nın bağlam içerisindeki tutarlılığı artırdığı ve otomatik puanlama sistemlerinde daha yüksek doğruluk sağladığını belirtmiştir.

Pennington vd. (2014) tarafında geliştirilen GloVe, kelimelerin anlamsal düzenliliklerini modellemek için büyük veri setlerinden eşdizim (co-occurrence) matrislerini kullanır. GloVe, Word2Vec gibi kelimeler arasında anlamsal benzerlikleri yakalar.

2.2.2. Değerlendirme Kriterleri

Rubrik tabanlı otomatik puanlama sistemlerinde kompozisyon/yazı değerlendirme süreci çeşitli kriterlere göre yapılır; dilbilgisi, içerik organizasyon, akıcılık gibi. Yazım

doğruluğunun kontrolü için kelimeler gruplandırılabilir. Örneğin, NAPLAN rubriği kelimeleri basit, yaygın, zor ve çok zor olarak sınıflandırarak yazım hatalarının ağırlığını değerlendirir (Fazal vd., 2013). İçerik ve organizasyon kontrolü ise daha zorlu bir süreçtir. Çoğu model, içerik uygunluğunu yüzeysel kelime eşleşmelerine dayandırmaktadır (Zupanc ve Bosnić, 2017). Örneğin, Attali ve Burstein (2006), makale değerlendirme modellerinde içerik uygunluğunu ölçmek için "E-rater" sistemini geliştirmiştir. Ancak bu model de yeni sistemlerle karşılaştırıldığında içerik uygunluğunu değerlendirmede yüzeysel kalmıştır. Bu yeni sistemler, metindeki anlam, organizasyon ve fikirlerin geliştirilmesini değerlendirebilmek için doğal dil işleme (NLP) tekniklerini kullanır. Yeni çalışmalarda, BERT ve GPT gibi bağlama duyarlı dil modellerinin kullanımıyla beraber derin anlam analizi yapılarak içerik uygunluğu daha doğru şekilde yapılabilir (Devlin vd., 2019). Örneğin, IntelliMetric gibi sistemler, içerik geliştirme ve organizasyon özelliklerini ayrı ayrı değerlendirir (Fazal vd., 2013).

Rubrik tabanlı değerlendirme sistemleri katmalı bir değerlendirme sistemine sahiptir. Sistemler, öğrencilere anında geri bildirim sağlayarak yazma sürecini destekler ve anında geri bildirim vererek öğretmenlerin yüksek seviyeli geri bildirimlere odaklanmasına olanak tanır (Wilson vd., 2016). Bazı modeller, özellikle GPT-4 gibi gelişmiş dil modelleri, rubrik tabanlı değerlendirme süreçlerini anlamak için ara mantık adımları oluşturarak kompozisyonlardaki kriterlerin daha iyi değerlendirilmesini sağlar (Lee vd., 2024).

2.2.3. Kullanılan Algoritmalar

Otomatik puanlama sistemleri, yazıların değerlendirilmesinde çeşitli makine öğrenimi ve derin öğrenme tekniklerinden yararlanmaktadır. Bu teknikler, metinlerin dil bilgisi doğruluğu, içerik düzeni ve bağlamsal anlam açısından analiz edilmesine olanak tanır. Regresyon ve destek vektör makineleri (Support Vector Machine - SVM) gibi geleneksel yöntemlerle başlayan otomatik puanlama sistemlerinin gelişimi, günümüzde derin öğrenme modellerinin gücüyle devam etmektedir.

Erken dönem otomatik puanlama sistemlerinde regresyon modelleri yaygın bir şekilde kullanılmıştır. Örneğin, Project Essay Grade (PEG), yazılardaki yüzeysel özellikleri analiz ederek insan değerlendiricilerin puanlama süreçlerini taklit etmek için regresyon modellerinden yararlanmıştır (E. B. Page, 1966). Bu sistem, kelime uzunluğu, nadir kelime kullanımı ve noktalama işaretlerinin sıklığı gibi değişkenlere dayanarak değerlendirmeler yapmıştır.

SVM, metin sınıflandırma ve yazı değerlendirme görevlerinde sıklıkla tercih edilen geleneksel yöntemlerden biridir. Bu algoritma, yazıların dil bilgisi, kelime sıklığı ve organizasyon özelliklerine dayanarak sınıflandırma yapar. Araştırmalara göre, SVM'in özellikle küçük veri kümelerinde hızlı eğitim süreci ve etkili sonuçlar verdiği belirtilmiştir (Suriyasat vd., 2023).

BETSY, Bayes teoremini temel alan bir yöntem olarak öne çıkmıştır. Bu sistem, dil özelliklerini istatistiksel modelleme ile birleştirerek yazıların anlam ve yapısını analiz etmiştir (Rudner ve Liang, 2002).

Uzun kısa süreli bellek ağları (Long Short-Term Memory - LSTM), yazıların dil bilgisi doğruluğu, içerik düzeni ve organizasyonu gibi faktörlere odaklanarak etkili sonuçlar sağlarken Bi-LSTM (Çift Yönlü LSTM) ise, metinlerin bağlamını ileri ve geri yönde analiz ederek daha derin bir anlam çıkarımı sunar (Suresh vd., 2023).

Transformer tabanlı modeller, bağlama duyarlı metin analizinde önemli bir devrim yaratmıştır. Özellikle BERT, yazıların semantik anlamını, organizasyonunu ve bağlamını detaylı bir şekilde analiz ederek yüksek doğruluk oranlarına ulaşmıştır. Zhang ve Litman (2018), kompozisyonlar ile kaynak metinler arasındaki bağlamsal ilişkileri analiz etmek için eş-dikkat (co-attention) mekanizmasını kullanan bir model geliştirmiştir. Bu yöntem, metinlerin yalnızca yapısal özelliklerini değil, içerik uyumunu da göz önünde bulundurmaktadır.

Derin öğrenme tabanlı hibrit modeller, LSTM ve Transformer mimarilerini birleştirerek yazıları daha kapsamlı bir şekilde değerlendirmeyi mümkün kılmıştır. Dong vd. (2017) geliştirdiği model, RNN (Recurrent Neural Network) ve CNN (Convolutional Neural Network) yöntemlerini dikkat mekanizması ile birleştirerek önemli metin parçalarına odaklanmıştır. Bu tür hibrit yaklaşımlar, özellikle uzun ve karmaşık metinlerin analizinde etkili sonuçlar sunmaktadır.

2.3. Eğitimde Otomatik Puanlama Sistemlerinin Kullanımı

2.3.1. Yazma Eğitiminin Desteklenmesi

Otomatik puanlama sistemleri, eğitimde kişiselleştirilmiş geri bildirim sağlama konusunda önemli bir katkı sunabilir. Bu sistemler, yazılı çalışmaların hızlı ve ayrıntılı bir

şekilde değerlendirilmesine olanak tanıyarak öğrencilerin öğrenme sürecine doğrudan destek sağlar. Özellikle BERT gibi Transformer tabanlı modeller ve LSTM gibi yöntemler, yalnızca dilbilgisel hataları belirlemekle sınırlı kalmayıp içerik organizasyonu ve bağlamsal analiz yaparak daha kapsamlı geri bildirimler sunar (Suresh vd., 2023; Zhang ve Litman, 2018). Bunun yanı sıra, bu sistemler her öğrencinin öğrenme düzeyine ve hızına uyum sağlayarak bireysel ihtiyaçlara yönelik geri bildirimler (Wilson vd., 2016) oluşturur ve öğretmenlerin iş yükünü azaltır (Attali ve Burstein, 2006). PEG Writing gibi sistemler, öğretmenlerin manuel değerlendirme yükünü azaltarak, yazım, organizasyon ve dilbilgisi gibi alt boyutlarda hızlı ve tutarlı geri bildirim sağlamayı amaçlar (Wilson vd., 2016). Böylelikle otomatik puanlama sistemleri, öğrencilerin yazı yazma becerilerini geliştirmelerine sürekli destek sağlayarak kişiselleştirilmiş öğrenme deneyimlerini daha etkili hale getirme potansiyeline sahiptir.

2.3.2. Standart Sınavlarda Kullanım

Standart sınavlarda kullanılan otomatik metin puanlama sistemleri (OMP), yazılı değerlendirme süreçlerini hızlandırma ve objektifliği artırma konusunda önemli bir rol oynamaktadır. TOEFL, GRE ve IELTS gibi sınavlarda yaygın olarak kullanılan E-rater, dil bilgisi, içerik düzeni ve organizasyon gibi metrikleri değerlendirerek insan değerlendiricilerle yüksek bir uyum sağlamıştır. Literatürde, E-rater sisteminin Quadratic Weighted Kappa (QWK) skorlarının 0.87 ile 0.94 arasında değiştiği belirtilmiş ve bu doğruluk oranları sistemin güvenilirliğini kanıtlamıştır (Reddy Chavva vd., 2024). Son dönemde, transformer tabanlı modellerin geliştirilmesiyle birlikte OMP sistemleri daha sofistike hale gelmiştir. Özellikle RoBERTa tabanlı modeller, bağlam analizi ve metin derinliği üzerinde daha başarılı bir değerlendirme yaparak QWK skorunda 0.815 seviyesine ulaşmıştır (Reddy Chavva vd., 2024). Bu tür sistemler, yalnızca sınav süreçlerini hızlandırmakla kalmayıp, aynı zamanda öğrencilere eşit ve adil bir değerlendirme sunarak eğitimde önemli bir katkı sağlamaktadır.

2.3.3. Çok Dilli ve Kültürlerarası Uygulamalar

Otomatik puanlama sistemlerinin başarısı, kullanılan modelin yapısı, veri miktarı ve dilin kendine has özelliklerine bağlı olarak değişiklik gösterebilir. İngilizce için geniş çaplı veri setlerinin bulunabilirliği, bu dildeki modellerin eğitilmesini kolaylaştırırken, diğer dillerde bu tür veri setleri genellikle daha sınırlıdır.

Transformer tabanlı modeller, İngilizce gibi dillerde oldukça etkili sonuçlar üretirken, eklemeli yapıya sahip Japonca ve Türkçe gibi dillerde, kelime tabanlı yaklaşımlar daha iyi performans gösterebilmektedir. Örneğin, Japonca yazılarda BERT tabanlı modeller başarılı olsa da kelime tabanlı yöntemler (örneğin, Bag-of-Words) doğruluk oranlarında üstünlük sağlayabilmektedir (Ejima ve Takeuchi, 2022).

Ayrıca, farklı dil ve kültürlere sahip bireylerin yazılı çalışmalarını değerlendirirken, bu faktörlerin değerlendirme sürecine etkisinin göz önünde bulundurulması önemlidir. İngilizce öğrenen farklı kültürel geçmişlere sahip öğrencilerin yazdığı kompozisyonların değerlendirme sonuçları, dil ve kültürel farklılıklardan etkilenebilir (Morris vd., 2024). Bu nedenle, modellerin bu çeşitliliğe uyum sağlayabilecek şekilde geliştirilmesi gerekmektedir.

Bunun yanı sıra, karmaşık veya belirsiz soruların değerlendirilmesinde bağlamı daha iyi anlamaya yönelik iyileştirmelere ihtiyaç duyulmaktadır (Martínez, 2024). Bu tür geliştirmeler, sistemlerin daha adil ve doğru değerlendirmeler yapmasını sağlayacaktır.

2.4. Avantajlar ve Sınırlamalar

2.4.1. Teknik Avantajlar

Otomatik metin puanlama (OMP) sistemleri, insan değerlendiricilere kıyasla daha tutarlı, tarafsız ve hızlı bir şekilde değerlendirme yapabilme avantajına sahiptir (Myers ve Wilson, 2023). Bu sistemler, algoritmalara dayalı olarak çalıştıkları ve belirli özelliklere odaklandıkları için, insan değerlendiricilerde karşılaşılan yorgunluk ya da içerikten etkilenme gibi faktörlerden etkilenmezler. Ayrıca, geniş katılımlı sınavlarda görev alan değerlendiricilerin eğitim düzeyi, ruh hali ve bireysel önyargıları gibi değişkenler, sonuçların tutarlılığını olumsuz etkileyebilir. Otomasyon sayesinde büyük ölçekli sınavların değerlendirilmesi hızlanırken, öğretmenlerin iş yükü de önemli ölçüde azalır.

Özellikle kısa cevaplı soruların değerlendirilmesinde otomatik puanlama sistemleri, insanlara göre çok daha hızlı sonuçlar sunar (Morris vd., 2024). Bu hız, sınav sonuçlarının daha kısa sürede elde edilmesini sağlarken, aynı zamanda insan değerlendiricilerin aksine daha tutarlı sonuçlar verebilir (Myers & Wilson, 2023).

Ayrıca, otomatik puanlama sistemleri sayesinde çok sayıda kompozisyon veya kısa yanıt hızlıca değerlendirilebildiği için hem zaman hem de maliyet açısından tasarruf sağlanır.

Bu hızlı süreç, öğrencilerin daha çabuk geri bildirim almasını sağlayarak öğrenme süreçlerini destekler. Aynı zamanda, bireysel öğrenci ihtiyaçlarını analiz etme ve özelleştirilmiş öğrenme çözümleri sunma imkânı da sunar.

2.4.2. Sınırlamalar ve Sorunlar

Otomatik puanlama sistemleri, eğitimde önemli bir rol oynasa da bazı sınırlılıkları ve zorlukları bulunmaktadır. Hallucination (yanlış bilgi üretimi), özellikle transformer tabanlı modellerde, yazının içeriğiyle tutarsız veya hatalı geri bildirimlerin ortaya çıkmasına neden olabilmektedir (Reddy Chavva vd., 2024). Bunun yanı sıra, etik sorunlar arasında veri gizliliği, yapay zeka önyargıları ve dijital eşitsizlik gibi konular öne çıkmaktadır. Örneğin, yapay zekanın eğitildiği veri setlerindeki önyargılar, farklı kültürel ve dilsel bağlamlarda haksız değerlendirmelere yol açabilir. Eğer kullanılan modellerin eğitimi özellikle bazı yanıt sınıflarının az temsil edildiği veri setleri kullanılarak yapıldıysa, modeller yanlış sınıflandırmalar yapabilir (Morris vd., 2024). Japonca ve İngilizce yazıların değerlendirilmesinde kullanılan modellerin, kültürel bağlama duyarsız olabileceği ve bu durumun adil değerlendirmeyi zorlaştırdığı söylenebilir (Obata vd., 2023). Ayrıca, teknolojik altyapı eksiklikleri nedeniyle dijital eşitsizlik, bu sistemlerin küresel çapta kullanımını kısıtlayabilir. Bu nedenle, OMP sistemlerinin daha adil, güvenilir ve kapsayıcı hale getirilmesi için çalışmaların devam etmesi gerekmektedir.

2.4.3. İnsan Değerlendiriciler ile Karşılaştırma

İnsan değerlendiricilerin değerlendirmeleri bazı etkenlere göre değişiklik gösterebilir. Örneğin değerlendiricin birkaç gün veya uzun süreler boyunca puanlama ve değerlendirme yaptığı durumlarda yorgunluk, ruh hali değişiklikleri, yazı temasına bakış açısının değişmesi gibi etkenler verilen notu etkileyebilir. Başka bir örnekte geniş çaplı bir sınavda birden fazla değerlendiricinin olduğu durumlarda, değerlendiricilerin geçmiş deneyimleri ve bireysel görüşleri aynı kağıdın farklı değerlendirilmesiyle sonuçlanabilir. Öte yandan OMP sistemleri bu tür etkenlerden etkilenmeyecek bu da sonuçların daha tutarlı olmasını sağlayacaktır. OMP sistemlerinin insan değerlendiricilerle oldukça uyumlu puanlamalar yaptığı pek çok araştırmada gösterilmiştir. Örneğin, TOEFL ve benzer sınavların (GRE, IELTS gibi) otomatik değerlendirme sistemleriyle entegre edilmesi, yazma görevlerinin daha tutarlı bir şekilde değerlendirilmesine olanak tanımıştır (Beseiso vd., 2021). Özellikle kısa cevaplı soruların değerlendirilmesinde otomatik puanlama sistemleri, insanlara göre çok daha hızlı sonuçlar

sunar (Morris vd., 2024). İnsan değerlendiricilerin aynı değerlendirmeyi yapması için çok daha uzun bir zamana ihtiyacı vardır. Ancak, OMP sistemlerinin insan değerlendiricilerinin arkasında kaldığı durumlar da mevcuttur. Bağlama dayalı yaratıcı yazılar ve kültürel ifadeler, OMP sistemleri için önemli derecede bir zorluk olmaya devam etmektedir. Örneğin, Ejima ve Takeuchi (2022), OMP sistemlerinin Japonca yazım hatalarını etkili bir şekilde tespit ettiğini, ancak bağlamsal anlamlandırma konusunda insan değerlendiricilere kıyasla henüz yeterince başarılı olmadığını belirtmiştir. Bu nedenle, bağlama duyarlı ve daha gelişmiş NLP tekniklerine ihtiyaç vardır.

2.5. Güncel Çalışmalar

2.5.1. Geleneksel Yaklaşımlar

Otomatik puanlama sistemlerinin değişimi ve gelişimi, her dönemde yazı değerlendirme süreçlerini daha etkili ve tutarlı hale getirme çabalarıyla şekillenmiştir. Project Essay Grade (PEG), IntelliMetric, ve E-rater bu alanda önemli kilometre taşları olarak kabul edilmektedir.

PEG, Ellis B. Page tarafından 1966 yılında geliştirilen ve yazı değerlendirme süreçlerini otomatikleştirme konusunda bir çığır açan ilk sistemlerden biridir. PEG, yazıların kelime uzunluğu, nadir kelime kullanımı ve noktalama sıklığı gibi yüzeysel özelliklerini analiz ederek insan değerlendiricilerin puanlama davranışlarını taklit etmiştir (E. B. Page, 1966). Sistem, regresyon modellerini kullanarak yazılardaki bu yüzeysel özellikler ile insan değerlendirmeleri arasındaki ilişkiyi modellemiştir. Bu yöntem, o dönem için yenilikçi olsa da yazının bağlamsal anlamını dikkate almaması nedeniyle eleştirilmiştir.

IntelliMetric, yazı değerlendirme alanında daha ileri bir adım olarak kabul edilmektedir. Bu sistem, bağlamsal analiz yeteneklerini kullanarak insan değerlendiricilerle uyumlu sonuçlar sunmayı hedeflemiştir. IntelliMetric, yüzeysel metriklerin ötesine geçerek dil bilgisi, organizasyon ve içerik gibi çeşitli unsurları analiz etmiştir (Rudner ve Liang, 2002). Literatürde, IntelliMetric'in hem akademik hem de kurumsal yazı değerlendirmelerinde başarılı olduğu belirtilmektedir.

E-rater, 1999 yılında Educational Testing Service (ETS) tarafından geliştirilmiş ve doğal dil işleme (NLP) tekniklerini kullanarak yazıların dil bilgisi, organizasyon ve kelime seçimi gibi yönlerini değerlendirmiştir. E-rater, TOEFL gibi standart sınavlarda geniş bir

uygulama alanı bulmuş ve insan değerlendiricilerle tutarlı sonuçlar elde etmiştir (Attali ve Burstein, 2006). Bu sistem, yazıların hızlı ve tutarlı bir şekilde değerlendirilmesini sağlaması nedeniyle büyük beğeni toplamıştır.

Akademik değerlendirmelerde, PEG, IntelliMetric ve E-rater gibi sistemler, yazıların sadece yüzeysel özelliklerini değil, aynı zamanda içerik doğruluğunu ve organizasyonunu da dikkate alarak otomatik puanlama alanında önemli bir katkı sağlamıştır. Bu sistemler, daha sofistike ve bağlama duyarlı modellerin geliştirilmesine öncülük etmiştir.

2.5.2. Yeni Nesil Sistemler

Yeni nesil otomatik puanlama sistemleri, son yıllarda yapılan literatür çalışmalarında geniş bir yer bulmuştur. Bu sistemler, geleneksel modellerin sınırlamalarını aşmak ve yazılı metinleri daha derinlemesine analiz edebilmek amacıyla gelişmiş yapay zeka tekniklerini kullanmaktadır.

LSTM ve Transformer tabanlı modeller, bu alandaki araştırmaların temelini oluşturur. Transformer tabanlı modeller, özellikle bağlama duyarlı analizlerde önemli bir avantaj sağlarken, LSTM modelleri dilbilgisel yapıların ve yazı organizasyonunun değerlendirilmesinde öne çıkmaktadır. Bu iki mimarinin bir araya getirilmesiyle ortaya çıkan hibrit modeller hem yazının dilsel özelliklerini hem de bağlamını dikkate alarak daha kapsamlı değerlendirmeler yapabilmektedir (Suresh vd., 2023).

Hibrit modellerin başarısı, bağlama duyarlı gömme teknikleriyle desteklenerek daha ileri bir seviyeye taşınmıştır. Bu yaklaşımlar, karmaşık ve çok katmanlı yazıları daha iyi anlamlandırarak yalnızca yazının yüzeysel özelliklerini değil, aynı zamanda derin anlamını ve yazının hedefini de analiz etmektedir. Literatürde, bu tür sistemlerin otomatik değerlendirmelerde hata oranlarını düşürdüğü ve doğruluk oranlarını artırdığı belirtilmektedir (Chassab vd., 2024).

Gelecekteki çalışmalar, bu sistemlerin yaratıcılık ve eleştirel düşünme gibi daha soyut özellikleri değerlendirebilecek şekilde geliştirilmesine odaklanmaktadır. Örneğin, Martínez (2024), bağlama duyarlı sistemlerin eğitimde daha kapsamlı bir dönüşüme katkı sağlayabileceğini öne sürmektedir. Bu bağlamda, yeni nesil otomatik puanlama sistemleri yalnızca yazı değerlendirme değil, aynı zamanda öğrenme süreçlerini destekleme açısından da büyük bir potansiyele sahiptir.

Makine öğrenimi ve derin öğrenme yaklaşımlarını birleştiren hibrit modeller, hataları azaltma ve genelleme yeteneğini artırma potansiyeline sahiptir. Örneğin, XGBoost gibi yöntemlerin derin öğrenme tabanlı çıkarım teknikleriyle entegre edilmesi, performansı optimize edebilir (Suriyasat vd., 2023).

2.5.3. Farklı Diller ve Kültürlerdeki Araştırmalar

İngilizce, doğal dil işleme (NLP) tekniklerinin en sık çalışıldığı ve en çok uygulama alanı bulunan dillerden biri olduğundan, Otomatik Puanlama Sistemlerinin (OMP) İngilizce metinlerdeki performansı genellikle diğer dillere kıyasla daha yüksektir. OMP sistemleri, İngilizce yazılardaki dilbilgisi ve yazım hatalarını tespit etmede büyük başarı sağlamaktadır. Örneğin, E-rater sistemi, TOEFL ve GRE gibi uluslararası standart sınavlarda uzun yıllardır kullanılmakta ve insan değerlendiricilerle yüksek bir tutarlılık göstermektedir (Shermis vd., 2010)

Bunun yanı sıra, PEG ve IntelliMetric gibi diğer OMP sistemleri de, yazılı metinlerin değerlendirilmesinde insan değerlendiricilerle güçlü bir uyum sağladıklarını bu sistemlerin içerik, organizasyon ve dilbilgisi gibi ölçütlerde başarılı bir performans sergiledikleri ortaya konulmuştur (Shermis vd., 2010).

Ancak, bağlama dayalı yaratıcı yazılar ve kültürel ifadeler, OMP sistemleri için önemli bir meydan okuma olmaya devam etmektedir. Bu tür metinler, yalnızca yüzeysel dil özelliklerinin değil, aynı zamanda derin anlam ilişkilerinin de değerlendirilmesini gerektirdiği için, mevcut sistemlerin semantik ve bağlam analizi konusundaki sınırlılıklarını ortaya koymaktadır. Bu nedenle, bağlama duyarlı ve daha gelişmiş NLP tekniklerine olan ihtiyaç devam etmektedir.

Otomatik Puanlama Sistemlerinin (OMP) Japonca yazılar üzerindeki performansı, dilin karmaşık yapısal özellikleri, kullanılan eğitim verilerinin kapsamı ve sistemlerin teknolojik altyapısına bağlı olarak farklılık göstermektedir. Japonca'nın kelime sınırlarının açıkça belirgin olmaması, OMP sistemlerinin metinleri doğru bir şekilde analiz etmesini güçleştiren önemli bir faktördür. Ancak, son yıllarda doğal dil işleme (NLP) teknolojilerindeki ilerlemeler, bu zorluğun üstesinden gelinmesine olanak sağlamıştır (Ejima ve Takeuchi, 2022)

Japonca'nın bağlama duyarlılığı yüksek bir dil olması, OMP sistemlerinin yalnızca dil bilgisi veya yüzeysel dil özelliklerini değil, aynı zamanda anlam ilişkilerini de dikkate almasını

gerektirir. Okgetheng ve Takeuchi (2024), Japonca yazma becerilerinin değerlendirilmesinde çeşitli modellerin insan değerlendiricilerle olan tutarlılığını Quadratic Weighted Kappa (QWK) ölçütüyle değerlendirmiştir. En iyi performansı gösteren Calm2-7b modeli, insan değerlendiricilerle olan tutarlılık açısından 0.5982 QWK değeri elde etmiştir (Okgetheng ve Takeuchi, 2024). Bu sonuçlara göre yaklaşık %60'lık bir tutarlılık gösterdiği söylenebilir.

Diğer bir çalışmada, OMP sistemlerinin Japonca yazım hatalarını etkili bir şekilde tespit ettiğini, ancak bağlamsal anlamlandırma konusunda insan değerlendiricilere kıyasla henüz yeterince başarılı olmadığını belirtmiştir (Obata vd., 2023).

Türkçe, eklemeli bir dil olduğu için kelimelerin yapısı ve anlamları eklerle değişebilir. Bu durum, OMP sistemleri için Türkçe yazıları doğru bir şekilde analiz etmeyi zorlaştıran önemli bir faktördür. Ancak, son yıllarda doğal dil işleme (NLP) teknolojilerindeki gelişmeler, Türkçe yazılar üzerinde OMP sistemlerinin performansını artırmıştır. Önceden eğitilmiş modeller kullanıldığında daha başarılı puanlamaların yapıldığı belirlenmiştir ve daha büyük veri setleriyle doğruluk oranlarının daha da yükselebileceği düşünülmektedir (Aksoy, N 2021) Bununla birlikte, Türkçe'nin de Japonca gibi bağlama dayalı bir dil olması, sistemlerin performansını sınırlayan bir faktör olmaya devam etmektedir.

BÖLÜM III
YÖNTEM

3.1. Araştırmanın Modeli

Araştırma kapsamında, belirli veri tabanlarında uygun arama dizinleri kullanılarak otomatik metin puanlama sistemleri üzerine yapılmış mevcut makale ve çalışmalar belirlenmiş, belirli kriterlere göre seçilmiş ve çeşitli kategoriler açısından analiz edilmiştir.

Bu doğrultuda araştırma modeli, sistematik literatür taraması yöntemine dayanmaktadır. Bu model ile genel olarak; hedef alandaki çalışmalar kapsamlı ve özetleyici nitelikte incelenerek alanın genel durumunun ortaya konulması amaçlanmaktadır (Kitchenham ve Charters, 2007).

Tablo 3.1. *Veri tabanlarında kullanılan arama dizinleri*

Veri Tabanı	Kullanılan arama dizini
ACL	site:aclanthology.org (“Automated essay grading” OR “Automated essay scoring” OR “short answer scoring systems” OR “essay scoring systems” OR “auto matic essay evaluation”) AND ("validity" OR "validation study") -review
ACM	("automated essay scoring" OR "automatic essay evaluation" OR "computerized essay scoring" OR "machine scoring") AND ("validity" OR "validation study" OR "construct validity" OR "criterion validity")
IEEE Explore	"automated essay scoring" OR "essay scoring systems" OR "automatic essay evaluation" OR “short answer scoring systems” AND "validity"
Springer	"automated essay scoring" OR "essay scoring systems" OR "automatic essay evaluation" OR “short answer scoring systems” AND "validity" NOT “medical” NOT “review”
Science Direct	"automated essay scoring" OR "essay scoring systems" OR "automatic essay evaluation" AND "validity"

3.2. Dahil Etme ve Hariç Tutma Kriterleri

Araştırmada ACM, ACL, IEEE Xplore, Springer ve ScienceDirect veri tabanları üzerinden sistematik bir literatür taraması gerçekleştirilmiştir. Tarama sürecinin daha odaklı ve güvenilir olmasını sağlamak amacıyla dahil etme ve hariç tutma kriterleri belirlenmiş, bu kriterlerin bir kısmı doğrudan arama dizinlerine entegre edilmiştir.

Araştırmaya 2000-2024 yılları arasında yayımlanmış makaleler dahil edilmiştir. Bu tarih aralığı, otomatik puanlama sistemlerinin gelişim sürecini ve güncel değişiklikleri

kapsayacak şekilde seçilmiştir. İngilizce yazılmış veya İngilizce çevirisi bulunan çalışmalar değerlendirmeye alınmıştır. Araştırmada yalnızca yazılı metin değerlendirme sistemlerine odaklanılması sebebiyle konuşma analizi veya farklı veri türleriyle çalışan makaleler dışarıda bırakılmıştır. Ayrıca, dört sayfadan kısa olup özgün veri veya bulgu içermeyen çalışmalar araştırmaya dahil edilmemiştir. Derleme ve anket makaleleri, literatüre katkılarının doğrudan birincil bulgular içermemesi nedeniyle çalışmaya alınmamıştır. Eğitim alanında OMP sistemleri kullanımına yönelik ampirik veya teorik araştırmalar sunan, güvenirlik ve geçerlilik niteliklerini karşılayan makaleler araştırmaya dahil edilmiş, bu koşulları sağlamayan makaleler çalışma dışında bırakılmıştır.

3.3. Araştırmanın Materyali

Araştırmanın materyalini; ACM, ACL, IEEE Explore, Springer ve Science Direct adlı akademik veri tabanlarında, çeşitli arama dizinleri kullanılarak elde edilen "makale" türündeki akademik çalışmalar oluşturmaktadır. Arama dizinleri kullanılarak yapılan taramada 1791 makale belirlenmiştir. Bu makalelerin 636'sı görüntülenmiş, başlıklarına ve kullanılan anahtar kelimelere göre değerlendirilerek 397 makale indirilmiştir. Belirlenen dahil etme ve hariç tutma kriterlerine göre tekrar değerlendirilen makalelerden 182'si çalışmaya dahil edilmiş ve çalışmanın materyali olarak belirlenmiştir.

Tablo 3.2. *Araştırma kapsamında incelenen makale sayıları ve kaynakları*

Akademik Kaynak	Nihai Makale Sayısı
ACL	23
ACM	7
IEEE Explore	56
Springer	74
Science Direct	22

3.4. Veri Analizi Aracı

Araştırma kapsamında incelenen akademik belgeler; metin madenciliği alanında sıklıkla kullanılan açık kaynak kodlu R (versiyon:4.4.2.) yazılımında, metin madenciliği uygulamaları entegre bir sistematik literatür taraması çalışması ile detaylı olarak incelenmiştir.

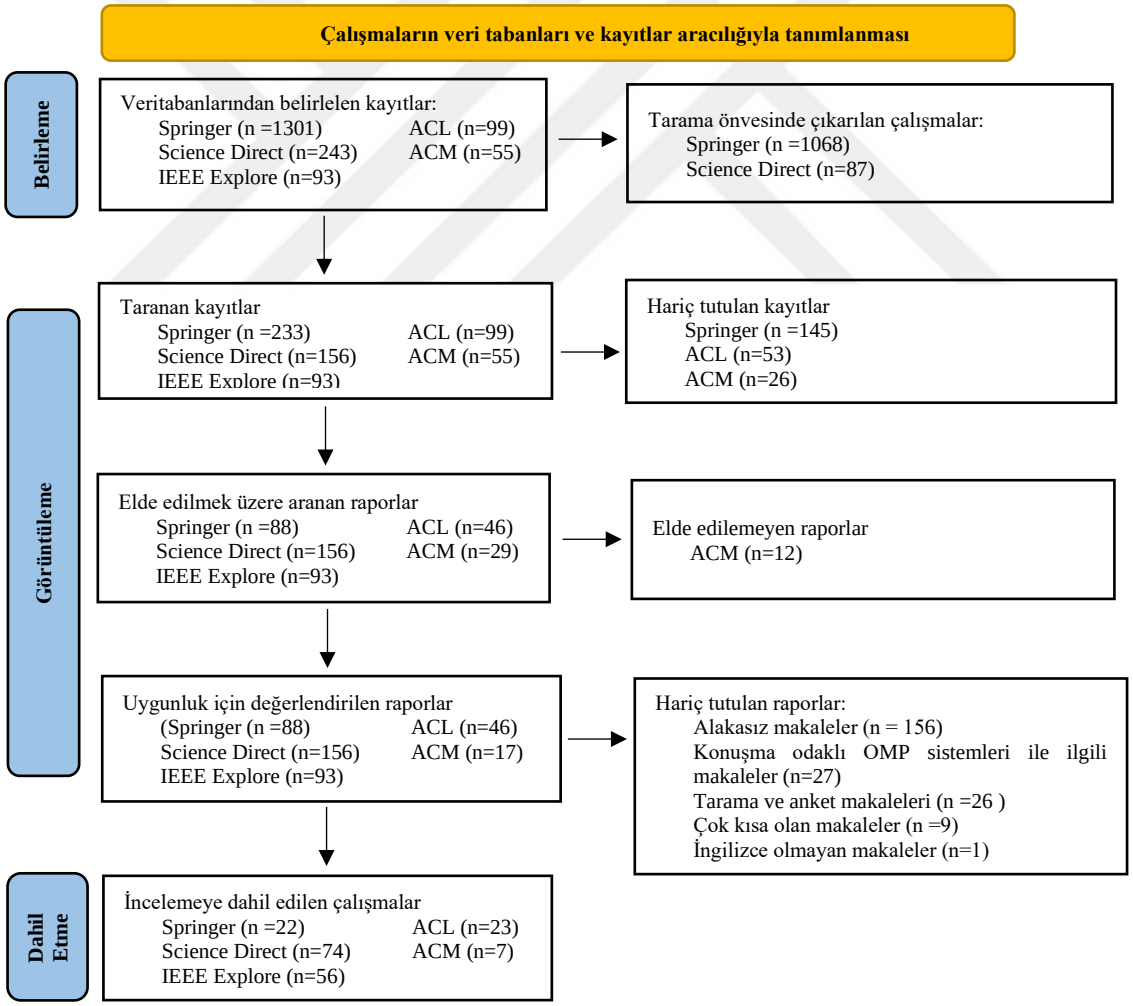
3.5. Veri Analizi Süreci

Çalışmanın bu bölümünde, metin madenciliği yöntem ve teknikleri kullanılarak yapılan sistematik literatür taramasına ilişkin veri analizi sürecine detaylı olarak yer verilmiştir.

3.5.1. Akademik çalışmaların elde edilmesi

ACM, ACL, IEEE Explore, Springer ve Science Direct adlı akademik veri tabanlarında, çeşitli arama dizinleri ile aratılarak 1791 makele belirlenmiştir. Bu makalelerin 636'sı görüntülenerek 397makale indirilmiştir. 397 akademik makale sayısı; araştırmacı tarafından yapılan eleme/hariç tutma işleminin ardından 182'e indirgenmiştir. Gerçekleştirilen sistematik literatür taraması süreci Tablo 3.3 verilen PRISMA Diagramı ile gösterilmiştir.

Tablo 3.3 *PRISMA Diagramı* (M. J. Page vd., 2021)



3.5.2. Ana veri tablosunun oluşturulması

Araştırmaya dahil edilen 182 akademik çalışma; aynı zamanda araştırmanın alt problemlerini oluşturan ve ilgili makalelere ilişkin inceleme kriteri/değişkeni olan çeşitli kategoriler altında yapılan içerik analizi sınıflandırılmış ve çalışmanın ana materyalini oluşturacak olan veri tablosu oluşturulmuştur. İlgili veri tablosuna ilişkin temsili gösterim, Tablo 3.4 ile sunulmuştur. İlgili veri tablosu, diğer bir deyişle, sistematik literatür taramasına ilişkin makale incelemeleri **Ek-1** ile ayrıca sunulmuştur.

Tablo 3.4. *Veri tablosu ve değişkenlere göre çalışmaların incelenmesine ilişkin örnek*

title	**pub_date**	**scoring_type**	**dataset_used**	**content**	**valid_metric**	**ml_technique**
Automated Arabic Essay Evaluation	2020	Arabic essay scoring using multiple criteria	Essays written by university students and journalists	Evaluates essays on spelling, structure, coherence, style, and punctuation without requiring model essays.	Correlation with human scores (0.87)	Support Vector Regression (SVR)
Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts	2019	Text similarity for machine-generated summaries	Human-authored essays and summaries	Introduces sentence mover's similarity metric combining word and sentence embeddings for better alignment.	Correlation with human judgments	Sentence embedding-based similarity metrics
⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.5.3. Korpusların oluşturulması

Veri tablosundaki her bir sütun, akademik çalışma metinlerinin ayrı açılardan/ayrı konular bakımından incelemek demektir. Bu bağlamda her bir sütun verisi ayrı ayrı kullanılarak, 6 farklı "Korpus" adı verilen, satırlarda belgelerin sütunlarda belge içeriklerinin yer aldığı metin veri seti oluşturulmuştur. Dolayısıyla bundan sonraki tüm veri analizi adımları her bir korpus için ayrı ayrı uygulanmıştır. Korpus örneği Tablo 3.5'te ayrıca verilmiştir.

Tablo 3.5. *Korpus örneği*

ACL	**Automated Arabic** Essay Evaluation
ACL	Language Proficiency Scoring
ACL	Toward Evaluation of Writing Style Finding Overly Repetitive Word Use in Student Essays
ACL	Onthe Effectiveness of Using Syntactic and Shallow Semantic Tree Kernels for Automatic Assessment of Essays

3.5.4. Korpus verilerinin analize hazırlanması

Veri tablosundaki sütunların ayrı şekilde çeşitli kodlamalar yardımıyla R yazılımına aktarımının ardından; "Korpus" adı verilen, satırlarda belgelerin sütunlarda belgelere ilişkin sütun içeriklerinin yer aldığı metin veri setleri üzerinde, metin madenciliğinin en temel adımı olan "metin ön işleme" adımları yürütülmüştür. Bu adımlar sırasıyla şunlardır:

- Harf Dönüşümü: Metindeki tüm harfler küçük harfe dönüştürülür.
- Durak Kelimeleri Kaldırma: Dildeki yaygın ve anlam taşımayan durak kelimeler (and, the, so, I vb.) metinden çıkarılır.
- Noktalama İşaretlerini Kaldırma: Katkısız karakterler temizlenir.
- Sayıları Kaldırma: Metindeki sayısal ifadeler temizlenir.
- Fazla Boşlukları Temizleme: Art arda gelen boşluklar tek boşluğa indirilir. Satır sonları ve satır başları kaldırılır.

3.5.4.1. Minimum kelime frekansının belirlenmesi

Metni daha anlamlı hale getirmek amacıyla nadir kullanılan kelimelerin çıkarılması işlemidir. Bu doğrultuda frekansı 5'ten az kelimeler korpustan çıkarılmıştır.

3.5.4.2. Belge-Terim matrisinin oluşturulması

"DocumentTermMatrix (DTM)" fonksiyonu kullanılarak, veri seti, bir belge-kelime matrisine dönüştürülür. Bu matriste, satırlar belgeleri, sütunlar ise bu belgelerde geçen kelimeleri temsil etmektedir. Bu doğrultuda kelimelerin belgelerdeki dağılımının kolaylıkla analiz edilmesi sağlanır.

3.5.4.3. Konu modellemenin uygulanması (LDA)

LDA fonksiyonu yardımıyla uygulanan Latent Dirichlet Allocation (LDA) yöntemiyle, Gibbs Sampling algoritması kullanılarak her belgenin konular üzerindeki dağılımı modellenmiştir. Diğer bir deyişle; ilgili değişken açısından her bir belgede en sık işlenen konular tespit edilmiştir.

Araştırmalarda LDA, hem içerik analizi hem de veri görselleştirme/modelleme süreçlerinde kullanılmaktadır. Sistematik literatür taraması çalışmaları sırasında belgelerin gizli

temalarını çıkararak, literatürdeki eğilimleri ve öncelikli araştırma konularını ortaya koymak için yaygın şekilde uygulanmaktadır (Blei et al., 2003). LDA'nın temel avantajı, metinlerdeki karmaşık kelime dağılımını anlaşılabilir temalara dönüştürerek analiz süreçlerini kolaylaştırmasıdır.

3.5.4.4. Sonuçların Görselleştirilmesi

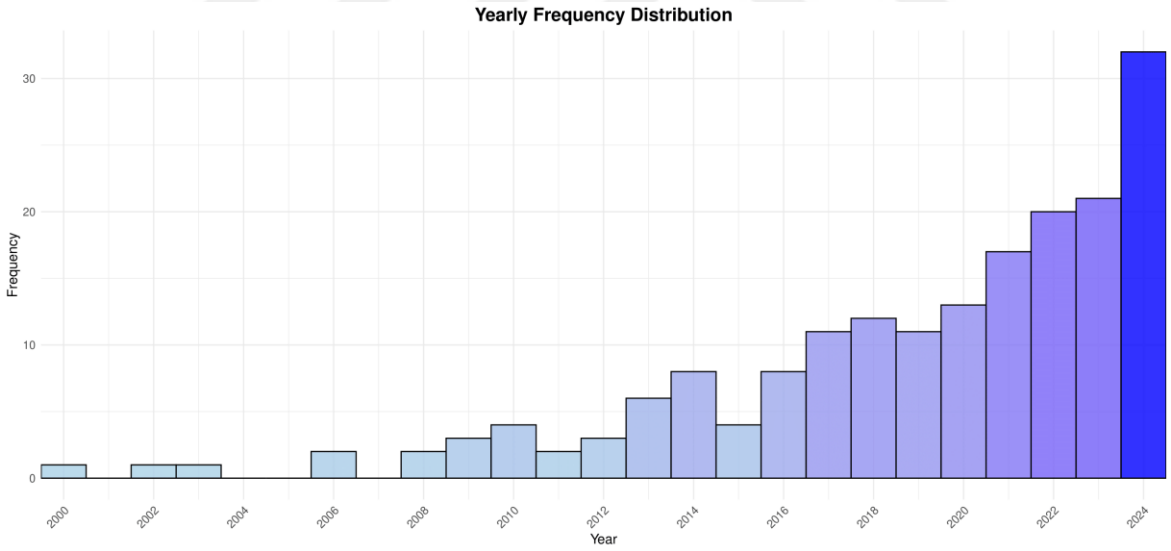
Wordcloud fonksiyonu kullanılarak, konu modelleme uygulaması ile elde edilen sonuçların görsel olarak daha kolay yorumlanabilmesi için kelime bulutları oluşturulmuştur. Bu kelime bulutları, her bir konuya ilişkin en sık kullanılan kelimeleri daha belirgin ve büyük puntolarla sunmaktadır.

BÖLÜM IV
BULGULAR

Tezin bu bölümünde öncelikle, araştırmanın amaçları doğrultusunda; otomatik metin puanlama sistemlerine ilişkin akademik makaleler, metin madenciliği teknikleriyle entegre sistematik literatür taraması yöntemiyle analiz edilmiştir. Çalışmalara ilişkin bilgiler doğrultusunda oluşturulan veri seti üzerinde; yayın tarihi, puanlama türü, veri setleri, içerik, makine öğrenmesi teknikleri ve doğrulama metrikleri gibi kategorilerde metin madenciliği yöntemleri uygulanmış; kelime bulutlar ile bulgular görselleştirilmiştir. Amaç, bu alandaki ana eğilimleri ve öne çıkan özellikleri belirlemektir.

4.1. Yayın tarihi değişkenine ilişkin bulgular

Sistematik literatür taramasına ilişkin veri tablosu üzerinde, "yayın yılı" değişkeni açısından yapılan analiz sonuçlarına ilişkin bulgular Şekil 4.1'deki histogram grafiği ile sunulmuştur.

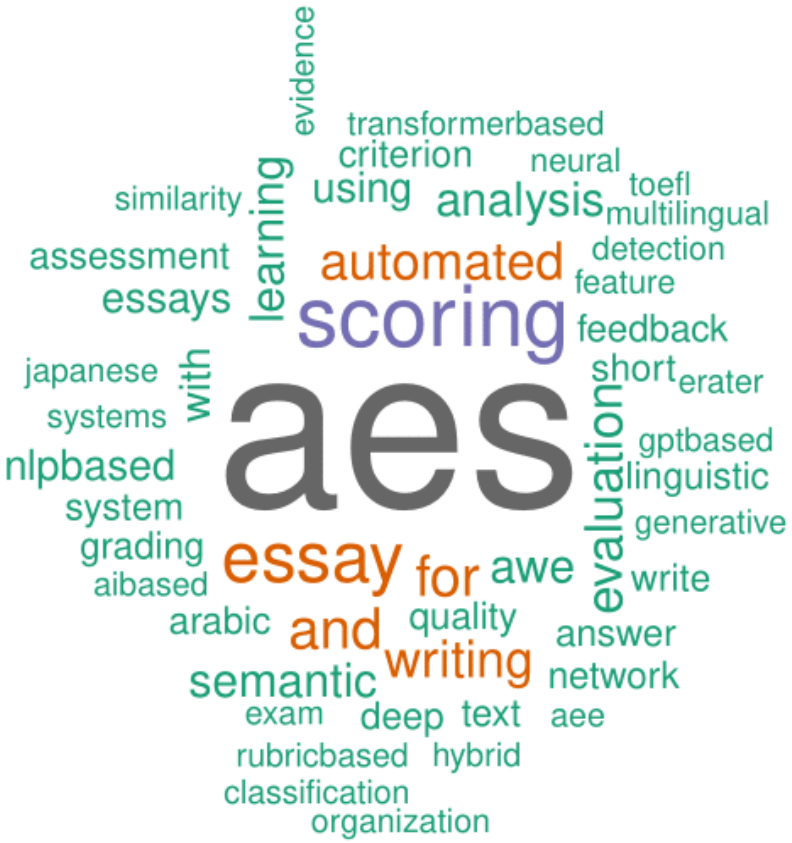


Şekil 4.1 *Yayın tarihlerine ilişkin histogram grafiği*

Yayın tarihlerine ilişkin histogram grafiği incelendiğinde; Otomatik puanlama sistemleri üzerine yapılan çalışmaların 2000 yılından itibaren başladığı ve 2010 yılından sonra belirgin bir artış gösterdiği gözlemlenmiştir. 2020 sonrası dönemde, çevrimiçi eğitimin ve teknolojik çözümlerin yaygınlaşmasıyla birlikte ilgili çalışmaların yoğunlaştığı görülmektedir.

4.2. Otomatik puanlama türüne ilişkin bulgular

Sistematik literatür taramasına ilişkin veri tablosu üzerinde, “otomatik puanlama türü” değişkeni açısından yapılan analiz sonuçlarına ilişkin bulgular, Şekil 4.2’deki kelime bulutu ile sunulmuştur.

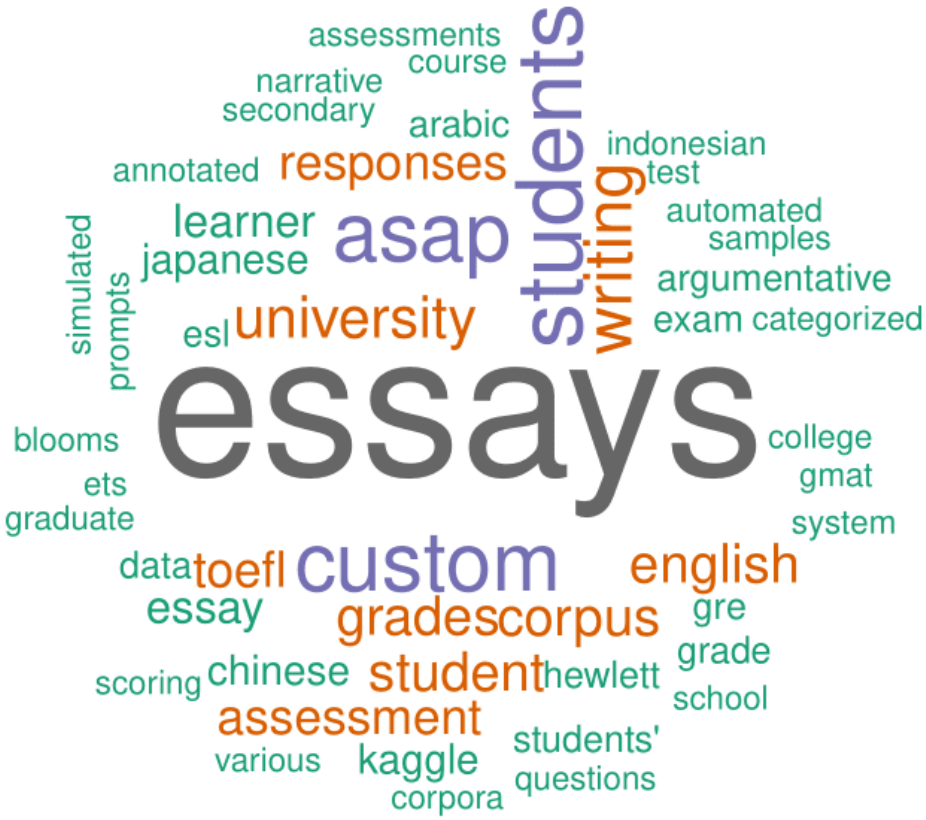


Şekil 4.2 *Çalışmalarda en sık kullanılan otomatik puanlama türlerine ilişkin kelime bulutu*

Çalışmalarda kullanılan otomatik puanlama türüne ilişkin olarak oluşturulan kelime bulutu incelendiğinde; "OMP," "semantic," "deep," "rubric-based," ve "feedback" terimlerinin sıklıkla geçtiği belirlenmiştir. "Arabic," "Japanese," ve "multilingual" gibi terimlerin yer alması, çok dilli otomatik puanlama sistemlerinin araştırmalarda yer bulduğunu göstermektedir. "Neural," "evaluation," ve "short answer scoring" terimleri, kısa yanıtlar ve derin öğrenme odaklı sistemlerin önemini vurgulamaktadır.

4.3. Kullanılan veri setine ilişkin bulgular

Sistematik literatür taramasına ilişkin veri tablosu üzerinde, “kullanılan veri seti” değişkeni açısından yapılan analiz sonuçlarına ilişkin bulgular, Şekil 4.3’teki kelime bulutu ile sunulmuştur.

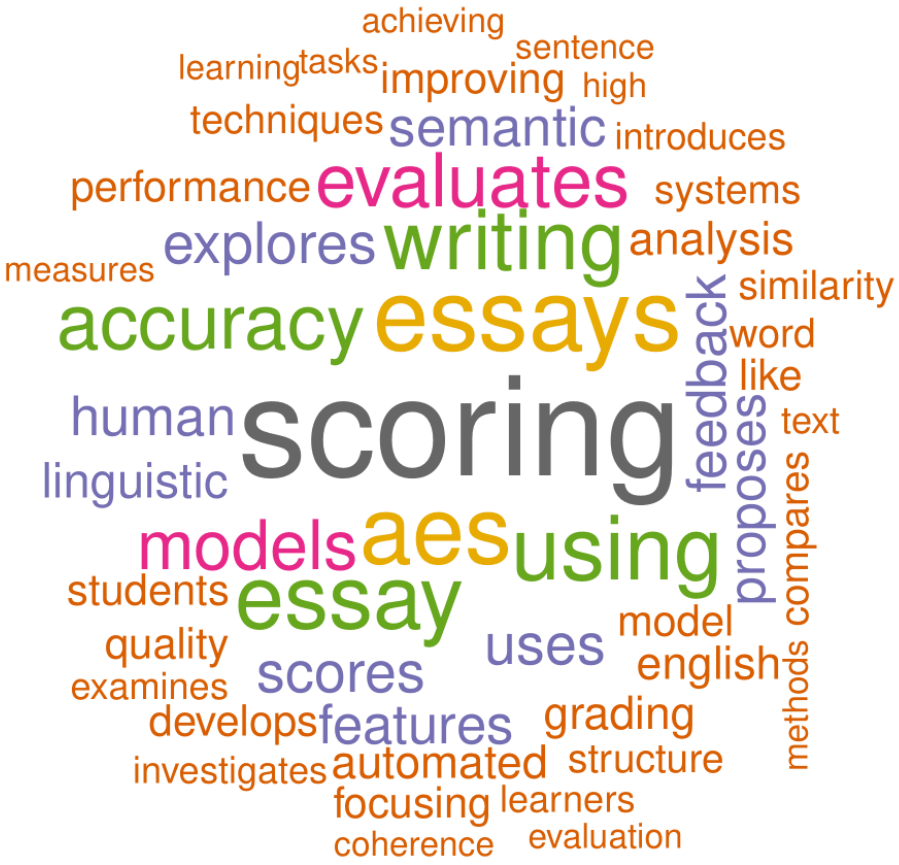


Şekil 4.3 *Çalışmalarda en sık kullanılan veri setlerine ilişkin kelime bulutu*

Çalışmalarda kullanılan veri setlerine ilişkin oluşturulan kelime bulutu incelendiğinde; "TOEFL," "Kaggle," "ASAP," ve "Hewlett" terimlerinin öne çıktığı görülmüştür. "Student," "learner," ve "responses" terimleri, öğrenci yanıtlarından oluşan veri setlerinin araştırmalarda yoğun olarak kullanıldığını işaret etmektedir. "Narrative," "argumentative," ve "categorized" terimleri, farklı yazma türlerinin analiz edildiğini göstermektedir.

4.4. Otomatik puanlama yapılan içeriğe ilişkin bulgular

Sistematik literatür taramasına ilişkin veri tablosu üzerinde, “içerik” değişkeni açısından yapılan analiz sonuçlarına ilişkin bulgular, Şekil 4.4’teki kelime bulutu ile sunulmuştur.



Şekil 4.4 *Çalışmalarda otomatik puanlama yapılan içerik türlerine ilişkin kelime bulutu*

Çalışmalarda üzerinde otomatik puanlama yapılan içeriklere ilişkin oluşturulan kelime bulutu incelendiğinde; "Writing," "students," "coherence," ve "semantic analysis" gibi terimlerin öne çıktığı görülmüştür. "Structure," "evaluation," ve "feedback" gibi terimler, yazılı içeriğin organizasyonu ve geri bildirim süreçlerinin vurgulandığını göstermektedir.

4.5. Otomatik puanlamada kullanılan makine öğrenmesi tekniğine ilişkin bulgular

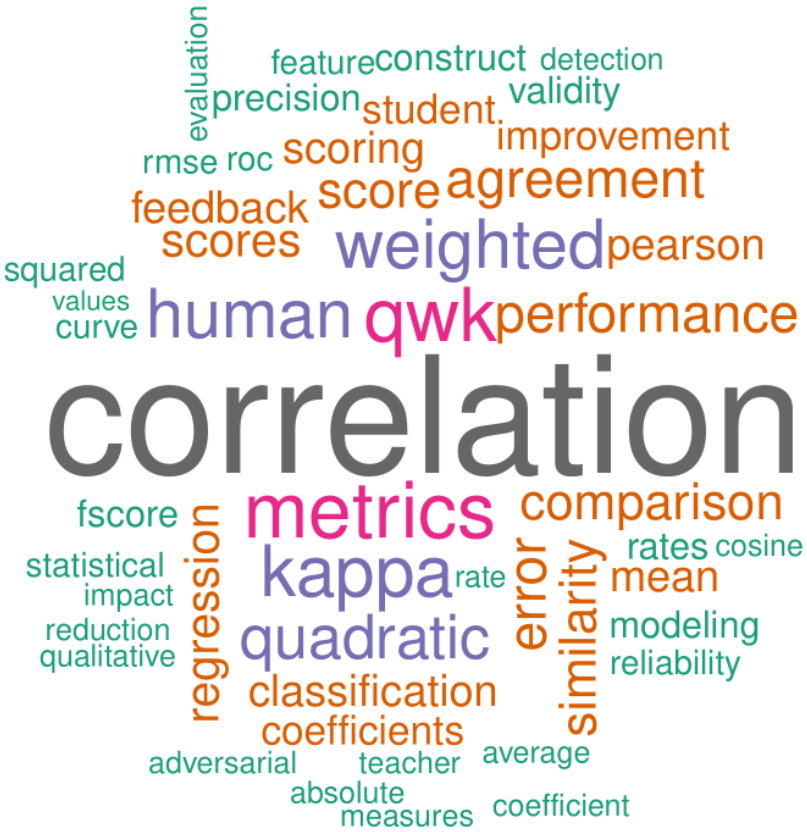
Sistematik literatür taramasına ilişkin veri tablosu üzerinde, “otomatik puanlamada kullanılan makine öğrenmesi tekniği” değişkeni açısından yapılan analiz sonuçlarına ilişkin bulgular, Şekil 4.5’teki kelime bulutu ile sunulmuştur.

Şekil 4.5 *Çalışmalarda otomatik puanlamada kullanılan makine öğrenmesi tekniğine ilişkin kelime bulutu*

Çalışmalarda otomatik puanlamada kullanılan makine öğrenmesi tekniklerine ilişkin oluşturulan kelime bulutu incelendiğinde; "Neural networks," "BERT," "LSTM," "SVM," ve "regression" gibi terimler, makine öğrenmesi yöntemleri arasında sıklıkla öne çıkmıştır. "Attention," "embedding," ve "feature engineering" gibi tekniklerin yer alması, derin öğrenme yöntemlerinin yaygın olarak kullanıldığını göstermektedir. "Adversarial," "clustering," ve "similarity" gibi terimlerin bulunması, yenilikçi yöntemlerin araştırmalarda uygulandığını göstermektedir.

4.6. Çalışmalarda kullanılan doğrulama metriklerine ilişkin bulgular

Sistematik literatür taramasına ilişkin veri tablosu üzerinde, “kullanılan doğrulama metriği” değişkeni açısından yapılan analiz sonuçlarına ilişkin bulgular, Şekil 4.6’daki kelime bulutu ile sunulmuştur.



Şekil 4.6 *Çalışmalarda kullanılan doğrulama metriklerine ilişkin kelime bulutu*

Çalışmalarda kullanılan doğrulama metriklerine ilişkin oluşturulan kelime bulutu incelendiğinde; "Quadratic Weighted Kappa (QWK)," "correlation," "precision," ve "validity" terimleri doğrulama metrikleri arasında sıkça yer almıştır. "Reliability," "accuracy," ve "RMSE" gibi terimlerin bulunması, model güvenilirliğini ve doğruluğunu değerlendirme süreçlerinin ön planda olduğunu göstermektedir. "Cosine similarity" ve "agreement" gibi metrikler, insan değerlendirmeleri ile sistem çıktılarının uyumunu değerlendirme odaklıdır.

BÖLÜM V

BÖLÜM V

SONUÇ, TARTIŞMA VE ÖNERİLER

5.1. Sonuç ve Tartışma

Otomatik metin puanlama alanı temelinde metin madenciliği teknikleriyle entegre bir sistematik literatür taraması yürütülen bu tez çalışmasında; genel olarak amaç, öğrenci cevaplarını daha tutarlı ve geçerli bir şekilde puanlamayı hedefleyen otomatik puanlandırma sistemleri üzerine yapılan çalışmaları, kullanılan veri setleri, otomatik puanlama türü, içerik, değerlendirme metrikleri ve kullanılan makine öğrenmesi teknikleri açısından yenilikçi uygulamalarla analiz etmek ve bu sistemlerin ölçme değerlendirme süreçlerindeki etkinliği noktasında alanyazına bir kaynak sağlamaktır. Bu doğrultuda tezde, otomatik metin puanlama sistemleri üzerine yapılan çalışmaları analiz etmek için sistematik literatür taraması ve metin madenciliği yöntemleri uygulanmıştır. ACM, ACL, IEEE Explore, Springer ve Science Direct gibi akademik veri tabanlarından elde edilen 182 makale, belirli kriterler doğrultusunda seçilmiş ve yayın tarihi, veri seti, puanlama türü, içerik, doğrulama metrikleri ve makine öğrenmesi teknikleri gibi kategoriler altında detaylı bir şekilde incelenmiştir. Metin madenciliği sürecinde R programı kullanılarak, makalelere ilişkin bir veri tablosu üzerinden, her bir değişken doğrultusunda, belge-terim matrisleri oluşturulmuş, Latent Dirichlet Allocation (LDA) yöntemiyle konular modellenmiş ve kelime bulutlarıyla görselleştirme yapılmıştır. Bu yöntemle otomatik puanlama sistemlerinin mevcut durumu ve genel eğilimleri kapsamlı bir şekilde ortaya konmuştur.

Otomatik metin puanlama çalışmalarının yayın tarihlerine ilişkin kelime bulutu sonuçları incelendiğinde; OMP çalışmalarının 2000'li yılların başında sınırlı sayıda iken, 2010 sonrası hızla artış gösterdiği görülmektedir. Bu artışın en önemli nedenleri arasında yapay zeka, derin öğrenme ve doğal dil işleme yöntemlerinin hızlı gelişimi yer almaktadır. Özellikle çevrimiçi eğitimin yaygınlaşması ve pandemi döneminde dijital eğitim araçlarının zorunluluk haline gelmesi, OMP araştırmalarını daha da önemli bir konuma taşımıştır. PRISMA 2020 (M. J. Page vd., 2021) gibi modern yönergeler, bu alandaki sistematik literatür taramalarını ve araştırmaların yapılandırılmasını kolaylaştırmıştır. Derin öğrenme tabanlı modellerin, özellikle dikkat mekanizmaları ve RNN tabanlı yaklaşımların kullanılması, metin analizi süreçlerinde

önemli bir dönüm noktası olmuştur (Dong vd., 2017; Zhang ve Litman, 2018). Ayrıca, Zhang ve arkadaşlarının çalışmaları, 2010 sonrası dönemde geliştirilen neural network tabanlı modellerin OMP sistemlerinde geniş kapsamlı kullanılabilirliğini ortaya koymaktadır (Zhang et al., 2018). Burstein & Chodorow, (1999) ise daha erken dönemlerde TOEFL veri setlerini kullanarak bu sistemlerin standart sınavlarda uygulanabilirliğini göstermiştir. Mizumoto ve Eguchi (2023) tarafından yapılan çalışmalar, çevrim içi eğitimde kullanılan OMP sistemlerinin pandemi sonrası dönem için önemini vurgulamaktadır.

Çalışmalarda en sık kullanılan otomatik puanlama türlerine ilişkin kelime bulutu sonuçları incelendiğinde; OMP sistemlerinde yaygın olarak rubric tabanlı ve semantik analiz odaklı yaklaşımlar kullanıldığı gözlemlenmiştir. Özellikle derin öğrenme ve semantik uyum odaklı modeller, sadece yüzeysel metin özelliklerini değil, içerik bağlamını da analiz ederek değerlendirme süreçlerini iyileştirmektedir (Zhang ve Litman, 2018). Rubric tabanlı yaklaşımlar, öğrenci yazılarının daha tutarlı bir şekilde değerlendirilebilmesini sağlarken, semantik analiz odaklı yöntemler anlam bütünlüğünü değerlendirmeye odaklanmıştır (Shermis, 2014). Bunun yanı sıra, dikkat mekanizmalarına sahip derin öğrenme tabanlı modeller, OMP sistemlerinin metin içeriği ve organizasyonunu daha etkin bir şekilde değerlendirmesine imkan tanımıştır (Dong vd., 2017). Ayrıca, kısa cevapların puanlanmasında kullanılan Bayes tabanlı modeller gibi geleneksel yöntemler, bu tür sistemlerin esnekliğini artırmıştır (Rudner ve Liang, 2002). Zhang ve arkadaşları (2018), neural network tabanlı eş-dikkat mekanizmalarının öğrenci yazılarında içerik uyumunu nasıl değerlendirdiğini açıklamaktadır.

Çalışmalarda en sık kullanılan veri setlerine ilişkin kelime bulutu sonuçları incelendiğinde; OMP sistemlerinde kullanılan veri setleri arasında TOEFL, ASAP ve Kaggle gibi standart veri kaynaklarının öne çıktığı görülmektedir. Bu veri setleri, genellikle öğrenci yanıtlarını içermekte ve OMP sistemlerinin gerçek dünya uygulamalarıyla uyumunu test etmek için kullanılmaktadır (Foltz et al., 1999). Ayrıca, bu veri setlerinin farklı yazı türlerini (örneğin, "argumentative" ve "narrative") içermesi, sistemlerin çeşitliliğini artırmaktadır (Attali ve Burstein, 2006). Zhang ve Litman (2018), bu veri setlerinin semantik analiz süreçlerindeki kritik rolünü ele almıştır.

Çalışmalarda otomatik puanlama yapılan içeriklere ilişkin kelime bulutu sonuçları incelendiğinde; OMP sistemlerinin metin analizi süreçlerinde "writing," "coherence," ve "evaluation" gibi kriterlerin öne çıktığı görülmektedir. Bu, sistemlerin yalnızca dilbilgisel doğruluk değil, aynı zamanda organizasyon ve bağlamsal uyum gibi karmaşık özellikleri

değerlendirdiğini göstermektedir. Zhang ve Litman (2018), semantik uyum ve içerik analizine odaklanan neural network tabanlı yaklaşımların bu süreçteki önemini vurgulamaktadır. Huang ve Zhao (2023) tarafından yapılan çalışmalar, OMP sistemlerinin bağlam ve anlam bütünlüğü analizindeki başarılarını detaylandırmaktadır.

Çalışmalarda otomatik puanlamada en sık kullanılan makine öğrenmesi tekniklerine ilişkin kelime bulutu sonuçları incelendiğinde; OMP sistemlerinde kullanılan makine öğrenmesi teknikleri arasında öne çıkanlar, derin öğrenme (Deep Learning), dikkat mekanizmaları (Attention Mechanisms), RNN (Recurrent Neural Networks), CNN (Convolutional Neural Networks), ve BERT (Bidirectional Encoder Representations from Transformers) tabanlı modellerdir. Bu yöntemler, metinlerin bağlamsal anlamını, dilbilgisel doğruluğunu ve organizasyonunu analiz ederek insan değerlendiricilerle yüksek uyum sağlayan puanlama sonuçları üretmektedir (Zhang ve Litman, 2018). Özellikle RNN ve CNN'in birlikte kullanıldığı modeller, metinlerdeki sekans verilerini ve yerel özellikleri daha etkili bir şekilde yakalayabilmektedir (Dong vd., 2017). BERT tabanlı yaklaşımlar ise dil bağlamını ve semantik ilişkileri daha derinlemesine analiz ederek OMP performansını artırmaktadır (Huang ve Zhao, 2023). Modern yöntemlerin yanında, daha geleneksel olan Bayes tabanlı modeller ve regresyon analizleri de kısa metin değerlendirmelerinde hala kullanılmaktadır (Rudner ve Liang, 2002).

Çalışmalarda en sık kullanılan doğrulama metriklerine ilişkin kelime bulutu sonuçları incelendiğinde; OMP sistemlerinde doğrulama metrikleri, insan değerlendiricilerle sistem puanlamaları arasındaki uyumu ve güvenilirliği ölçmek için kritik bir role sahiptir. En yaygın kullanılan metrikler arasında Quadratic Weighted Kappa (QWK), Pearson Korelasyonu ve Mean Squared Error (MSE) bulunmaktadır. Özellikle QWK, insan değerlendiricilerin puanlamalarıyla OMP sistemlerinin uyumunu değerlendirmede en güvenilir metriklerden biri olarak öne çıkmaktadır (Huang ve Zhao, 2023). Pearson korelasyonu, metinlerin dilbilgisel doğruluğu ve içerik uyumunun karşılaştırılmasında yaygın olarak kullanılmaktadır (Attali ve Burstein, 2006). Ayrıca, bazı çalışmalar ROC eğrisi ve F1 skoru gibi makine öğrenmesi odaklı metriklerden yararlanarak sistemlerin doğruluğunu ve genel performansını analiz etmektedir (Dong vd., 2017). Bu metrikler, OMP sistemlerinin sadece yüzeysel analizlerle değil, bağlamsal ve semantik açıdan da doğru puanlama yapmasını sağlamaktadır.

Sonuçlar, genel olarak; OMP sistemlerinin 2010 yılından itibaren, yapay zeka, derin öğrenme ve doğal dil işleme alanlarındaki teknolojik yeniliklerle paralel olarak gelişim gösterdiğini ortaya koymaktadır. Özellikle TOEFL, ASAP ve Kaggle gibi geniş ölçekli veri

setlerinin yaygın olarak kullanıldığı; rubric tabanlı, semantik analiz odaklı ve neural network tabanlı yöntemlerin öne çıktığı belirlenmiştir. Ayrıca, doğrulama metrikleri açısından Quadratic Weighted Kappa ve Pearson Korelasyonu gibi güvenilir ölçütlerin sıklıkla tercih edildiği görülmüştür. Bu bulgular, OMP sistemlerinin yalnızca dilbilgisel doğruluk değil, içerik bağlamı ve organizasyonu gibi karmaşık özellikleri değerlendirme kapasitesine sahip olduğunu ortaya koymaktadır.

Tez kapsamında yapılan analizler, OMP sistemlerinin insan hatalarını azaltma, değerlendirme süreçlerini hızlandırma ve daha tutarlı sonuçlar sunma gibi önemli avantajlar sağladığını göstermektedir. Ancak, Türkiye'de bu alandaki çalışmaların sınırlı olması, bu sistemlerin eğitim teknolojilerine entegrasyonu ve adaptasyonunda önemli bir boşluğa işaret etmektedir. Bu tez, bu boşluğu doldurma yönünde önemli bir adım atmış, hem alanyazına hem de uygulamaya yönelik somut katkılar sağlamıştır.

Kısacası bu çalışma, otomatik metin puanlama sistemlerinin eğitimdeki potansiyelini değerlendiren kapsamlı bir çerçeve sunmuş ve bu alandaki gelecekteki araştırmalara ışık tutacak değerli bir rehber oluşturmuştur. Bulgular, eğitimde teknolojinin etkin kullanımı için OMP sistemlerinin geliştirilmesi ve yaygınlaştırılması gerektiğini vurgulamakta, bu alandaki akademik ve pratik çalışmaların önemini bir kez daha ortaya koymaktadır.

5.2. Öneriler

Otomatik puanlama sistemleri, eğitimde önemli bir rol oynasa da bazı sınırlılıkları ve zorlukları bulunmaktadır. Özellikle transformer tabanlı otomatik puanlama sistemleri, hatalı değerlendirmelerin yapılmasına neden olabilmektedir. Bunun yanını sıra, veri gizliliği, yapay zeka önyargıları ve dijital eşitsizlik gibi etik sorunlar henüz çözüme ulaştırılmamıştır. Ayrıca, teknolojik altyapı eksiklikleri nedeniyle dijital eşitsizlik, bu sistemlerin küresel çapta kullanımını kısıtlayabilir.

Otomatik Puanlama Sistemleri, yazılı metinlerin değerlendirilmesinde sağladığı hız, tutarlılık ve ölçeklenebilirlik gibi avantajlarla ölçme değerlendirme alanında önemli bir potansiyele ve yere sahiptir. Özellikle İngilizce gibi doğal dil işleme uygulamalarında yoğun olarak çalışılan dillerde, OMP sistemlerinin insan değerlendiricilerle yüksek düzeyde tutarlılık gösterdiği tespit edilmiştir. Bununla birlikte, Türkçe ve Japonca gibi dil yapısı açısından daha

karmaşık ve bağlama dayalı dillerde, OMP sistemlerinin performansı dilin kendine özgü zorlukları nedeniyle sınırlı kalabilmektedir. OMP sistemlerinin performansı, kullanılan teknoloji, eğitildikleri veri setlerinin kapsamı ve dilin yapısal özellikleri gibi faktörlere bağlıdır. İngilizce için geliştirilen sistemler, TOEFL ve GRE gibi standart sınavlarda başarıyla kullanılırken, Türkçe ve Japonca gibi eklemeli ya da bağlama dayalı dillerde, daha ileri Doğal Dil İşleme teknikleri gereksinimi dikkat çekmektedir. Ayrıca, İngilizce dışındaki dillerde kullanılabilecek veri setlerinin daha sınırlı sayıda olduğu da görülmektedir.

Otomatik puanlama sistemlerinin daha adil, güvenilir ve kapsayıcı hale getirilmesi için çalışmaların devam etmesi gerekmektedir. Otomatik puanlama sistemlerinin performansını artırmak için daha çeşitli, kapsayıcı ve dilin farklı bağlamlarını içeren eğitim veri setlerinin oluşturulması gerektiği tavsiye edilmektedir.

KAYNAKÇA

Abdi Tabari, M., & Johnson, M. D. (2023). Exploring new insights into the role of cohesive devices in written academic genres. *Assessing Writing, 57*, 100749. https://doi.org/10.1016/j.asw.2023.100749

Ahmed, H. M. M., & Sorour, S. E. (2024). Classification-driven intelligent system for automated evaluation of higher education exam paper quality. *Education and Information Technologies, 29*(15), 19835–19861. https://doi.org/10.1007/s10639-024-12555-9

Ajetunmobi, S. A., & Daramola, O. (2017). Ontology-based information extraction for subject-focused automatic essay evaluation. *2017 International Conference on Computing Networking and Informatics (ICCNI)*, 1–6. https://doi.org/10.1109/ICCNI.2017.8123781

Ajit Tambe, A., & Kulkarni, M. (2022). Automated essay scoring system with grammar score analysis. *2022 Smart Technologies, Communication and Robotics (STCR)*, 1–7. https://doi.org/10.1109/STCR55312.2022.10009053

Aksoy, N. (2021). Türkçe dilinde yapılmış açık uçlu sınavların doğal dil işleme ile otomatik olarak değerlendirilmesi (Order No. 31286168) [Master's thesis, Balıkesir Üniversitesi]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/türkçe-dilinde-yapılmış-açık-uçlu-sınavların/docview/3110348647/se-2

Allen, L. K., Likens, A. D., & McNamara, D. S. (2018). A multi-dimensional analysis of writing flexibility in an automated writing evaluation system. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 380–388. https://doi.org/10.1145/3170358.3170404

Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., & McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information: The essay, the writer, and keystrokes. *Proceedings of the Sixth*

International Conference on Learning Analytics and Knowledge, 114–123. https://doi.org/10.1145/2883851.2883939

Alqahtani, A., & Alsaif, A. (2020). Automated Arabic essay evaluation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)* (pp. 181–190). NLP Association of India (NLPAI). https://aclanthology.org/2020.icon-main.24/

Alsanie, W., Alkanhal, M. I., Alhamadi, M., & Alqabbany, A. O. (2022). Automatic scoring of Arabic essays over three linguistic levels. *Progress in Artificial Intelligence, 11*(1), 1–13. https://doi.org/10.1007/s13748-021-00257-z

Andersen, M. R., Kabel, K., Bremholm, J., Bundsgaard, J., & Hansen, L. K. (2023). Automatic proficiency scoring for early-stage writing. *Computers and Education: Artificial Intelligence, 5*, 100168. https://doi.org/10.1016/j.caeai.2023.100168

Arifin, A. R., Purnamasari, P. D., & Putri Ratna, A. A. (2021). Automatic essay scoring for Indonesian short answers using Siamese Manhattan long short-term memory. *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 1–6. https://doi.org/10.1109/ICECCE52056.2021.9514223

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3), Article 3. https://ejournals.bc.edu/index.php/jtla/article/view/1650

Azmi, A. M., Al-Jouie, M. F., & Hussain, M. (2019). AAEE – Automated evaluation of students' essays in Arabic language. *Information Processing & Management, 56*(5), 1736–1752. https://doi.org/10.1016/j.ipm.2019.05.008

Balyan, R., McCarthy, K. S., & McNamara, D. S. (2020). Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. *International Journal of Artificial Intelligence in Education, 30*(3), 337–370. https://doi.org/10.1007/s40593-020-00201-7

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International*

Journal of Educational Technology in Higher Education, 21(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration. *Assessing Writing, 22*, 48–59. https://doi.org/10.1016/j.asw.2014.06.001

Bernius, J. P., Krusche, S., & Bruegge, B. (2022). Machine learning-based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence, 3*, 100081. https://doi.org/10.1016/j.caeai.2022.100081

Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education, 33*(3), 727–746. https://doi.org/10.1007/s12528-021-09283-1

Bhatt, R., Patel, M., Srivastava, G., & Mago, V. (2020). A graph-based approach to automate essay evaluation. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 4379–4385. https://doi.org/10.1109/SMC42975.2020.9282902

Blake, R., & Gutierrez, O. (2011). A semantic analysis approach for assessing professionalism using free-form text entered online. *Computers in Human Behavior, 27*(6), 2249–2262. https://doi.org/10.1016/j.chb.2011.07.004

Bojamma, A. M., Bhattacharya, M., & Liu, T. C. F. (2024). A lexical model for automated essay evaluation system using ensemble learning techniques. *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 1–4. https://doi.org/10.1109/I-SMAC61858.2024.10714597

Bonthu, S., Sree, S. R., & Krishna Prasad, M. H. M. (2024). Framework for automation of short answer grading based on domain-specific pre-training. *Engineering Applications of Artificial Intelligence, 137*, 109163. https://doi.org/10.1016/j.engappai.2024.109163

Bridgeman, B., & Ramineni, C. (2017). Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings. *Assessing Writing, 34*, 62–71. https://doi.org/10.1016/j.asw.2017.10.001

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Burstein, J., & Wolska, M. (2003, April). Toward evaluation of writing style: Overly repetitious word use. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education, 55*(2), 489-499. https://doi.org/10.1016/j.compedu.2010.02.012

Catulay, J. J. J. E., Magsael, M. E., Ancheta, D. O., & Costales, J. A. (2021). Neural-network architecture approach: An automated essay scoring using Bayesian linear ridge regression algorithm. *2021 8th International Conference on Soft Computing and Machine Intelligence (ISCMI)*, 196-200. https://doi.org/10.1109/ISCMI53840.2021.9654801

Chali, Y., & Hasan, S. A. (2013). On the effectiveness of using syntactic and shallow semantic tree kernels for automatic assessment of essays. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 767–773. Asian Federation of Natural Language Processing. https://aclanthology.org/I13-1092.pdf

Chang, T.-H., & Lee, C.-H. (2009). Automatic Chinese essay scoring using connections between concepts in paragraphs. *2009 International Conference on Asian Language Processing*, 265-268. https://doi.org/10.1109/IALP.2009.63

Chang, T.-H., Tsai, P.-Y., Lee, C.-H., & Tam, H.-P. (2008). Automated essay scoring using set of literary sememes. *2008 International Conference on Natural Language Processing and Knowledge Engineering*, 1-5. https://doi.org/10.1109/NLPKE.2008.4906764

Chassab, R. H., Zakaria, L. Q., & Tiun, S. (2024). An optimized LSTM-based augmented language model (FLSTM-ALM) using Fox algorithm for automatic essay scoring prediction. *IEEE Access, 12*, 48713-48724. https://doi.org/10.1109/ACCESS.2024.3381619

Chen, H., He, B., Luo, T., & Li, B. (2012). A ranked-based learning approach to automated essay scoring. *2012 Second International Conference on Cloud and Green Computing*, 448-455. https://doi.org/10.1109/CGC.2012.41

Chen, M., & Li, X. (2018). Relevance-based automated essay scoring via hierarchical recurrent model. *2018 International Conference on Asian Language Processing (IALP)*, 378-383. https://doi.org/10.1109/IALP.2018.8629256

Chen, Z., & Zhou, Y. (2019). Research on automatic essay scoring of composition based on CNN and OR. *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 13–18. https://doi.org/10.1109/ICAIBD.2019.8837007

Cheng, Y., Lyons, K., Chen, G., Gašević, D., & Swiecki, Z. (2024). Evidence-centered assessment for writing with generative AI. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 178–188. https://doi.org/10.1145/3636555.3636866

Clark, E., Celikyilmaz, A., & Smith, N. A. (2019). Sentence mover's similarity: Automatic evaluation for multi-sentence texts. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2748–2760. https://doi.org/10.18653/v1/P19-1264

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*(1), 100–108. https://doi.org/10.1016/j.asw.2012.11.001

Conijn, R., Cook, C., van Zaanen, M., & Van Waes, L. (2022). Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education, 32*(4), 835–866. https://doi.org/10.1007/s40593-021-00268-w

Contreras, J. O., Hilles, S., & Abubakar, Z. B. (2018). Automated essay scoring with ontology based on text mining and NLTK tools. *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, 1–6. https://doi.org/10.1109/ICSCEE.2018.8538399

Contreras, J. O., Hilles, S., & Bakar, Z. A. (2021). Essay question generator based on Bloom's taxonomy for assessing automated essay scoring system. *2021 2nd International*

Conference on Smart Computing and Electronic Enterprise (ICSCEE), 55–62. https://doi.org/10.1109/ICSCEE50312.2021.9498166

Das, L. B., Raghu, C. V., Jagadanand, G., George, R. A. R., Yashasawi, P., Kumaran, N. A. A., & Patnaik, V. K. (2022). FACToGRADE: Automated essay scoring system. *2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 42–48. https://doi.org/10.1109/IAICT55358.2022.9887447

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

del Río, I. (2019). Linguistic features and proficiency classification in L2 Spanish and L2 Portuguese. In D. Alfter, E. Volodina, L. Borin, I. Pilan, & H. Lange (Eds.), *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning* (pp. 31–40). LiU Electronic Press. https://aclanthology.org/W19-6304/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding (*arXiv:1810.04805*). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In R. Levy & L. Specia (Eds.), *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 153–162). Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-1017

Eid, S. M., & Wanas, N. M. (2017). Automated essay scoring linguistic feature: Comparative study. 2017 International Conference on Advanced Control Circuits Systems (ACCS) & 2017 International Conference on New Paradigms in Electronics & Information Technology (PEIT), 212–217. https://doi.org/10.1109/ACCS-PEIT.2017.8303043

Ejima, C., & Takeuchi, K. (2022). Statistical learning models for Japanese essay scoring toward one-shot learning. *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 313–318. https://doi.org/10.1109/IIAIAAI55812.2022.00070

ElNaka, A., Nael, O., Afifi, H., & Sharaf, N. (2021). AraScore: Investigating response-based Arabic short answer scoring. *Procedia Computer Science, 189*, 282-291. https://doi.org/10.1016/j.procs.2021.05.091

Fazal, A., Hussain, F. K., & Dillon, T. S. (2013). An innovative approach for automatically grading spelling in essays using rubric-based scoring. *Journal of Computer and System Sciences, 79(7)*, 1040-1056. https://doi.org/10.1016/j.jcss.2013.01.021

Fiacco, J., Jiang, S., Adamson, D., & Rosé, C. (2022). Toward automatic discourse parsing of student writing motivated by neural interpretation. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 204-215. https://doi.org/10.18653/v1/2022.bea-1.25

Filho, A. H., Concatto, F., Nau, J., Prado, H. A. D., Imhof, D. O., & Ferneda, E. (2019). Imbalanced learning techniques for improving the performance of statistical models in automated essay scoring. *Procedia Computer Science, 159*, 764-773. https://doi.org/10.1016/j.procs.2019.09.235

Filho, A. H., Do Prado, H. A., Ferneda, E., & Nau, J. (2018). An approach to evaluate adherence to the theme and the argumentative structure of essays. *Procedia Computer Science, 126*, 788-797. https://doi.org/10.1016/j.procs.2018.08.013

Filighera, A., Ochs, S., Steuer, T., & Tregel, T. (2024). Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs. *International Journal of Artificial Intelligence in Education, 34(2)*, 616-646. https://doi.org/10.1007/s40593-023-00361-2

Gaheen, M. M., ElEraky, R. M., & Ewees, A. A. (2021). Automated students Arabic essay scoring using trained neural network by e-jaya optimization to support personalized system of instruction. *Education and Information Technologies, 26(1)*, 1165-1181. https://doi.org/10.1007/s10639-020-10300-6

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System, 80*, 176-187. https://doi.org/10.1016/j.system.2018.12.001

Goura, V. M. K. P., Moulesh, M., Madhusudanarao, N., & Gao, X.-Z. (2022). An efficient and enhancement of recent approaches to build an automated essay scoring system. *Procedia Computer Science, 215*, 442-451. https://doi.org/10.1016/j.procs.2022.12.046

Gunawansyah, Rahayu, R., Nurwathi, Sugiarto, B., & Gunawan. (2020). Automated essay scoring using natural language processing and text mining method. *2020 14th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 1-4. https://doi.org/10.1109/TSSA51342.2020.9310845

Hellman, S., Andrade, A., & Habermehl, K. (2023). Scalable and explainable automated scoring for open-ended constructed response math word problems. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 137-147. https://doi.org/10.18653/v1/2023.bea-1.12

Hirao, R., Arai, M., Shimanaka, H., Katsumata, S., & Komachi, M. (2020). Automated essay scoring system for nonnative Japanese learners. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1250–1257. European Language Resources Association. https://aclanthology.org/2020.lrec-1.157.pdf

Horbach, A., Scholten-Akoun, D., Ding, Y., & Zesch, T. (2017). Fine-grained essay scoring of a complex writing task for native speakers. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 357-366. https://doi.org/10.18653/v1/W17-5040

Huang, J., Zhao, X., Che, C., Lin, Q., & Liu, B. (2023). Enhancing essay scoring with adversarial weights perturbation and metric-specific attention pooling. *2023 International Conference on Information Network and Computer Communications (INCC)*, 8-12. https://doi.org/10.1109/INCC58754.2023.00009

Huang, Y., & Wilson, J. (2021). Using automated feedback to develop writing proficiency. *Computers and Composition, 62*, 102675. https://doi.org/10.1016/j.compcom.2021.102675

Humphry, S., & Heldsinger, S. (2019). Raters' perceptions of assessment criteria relevance. *Assessing Writing, 41*, 1-13. https://doi.org/10.1016/j.asw.2019.04.002

Husni, Rachman, F. H., Suzanti, I. O., & Sari, M. K. (2022). Word ambiguity identification using POS tagging in automatic essay scoring. *2022 IEEE 8th Information Technology International Seminar (ITIS)*, 140-144. https://doi.org/10.1109/ITIS57155.2022.10009034

Idhom, M., Buditjahjanto, I. G. P. A., Munoto, & Samani, M. (2022). Performance evaluation of automated essay scoring online system for competency assessment of community academy. *2022 Fifth International Conference on Vocational Education and Electrical Engineering (ICVEE)*, 206–210. https://doi.org/10.1109/ICVEE57061.2022.9930387

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 the 44th Annual Meeting of the ACL*, 233–240. https://doi.org/10.3115/1220175.1220205

James, C. L. (2006). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing, 11*(3), 167–178. https://doi.org/10.1016/j.asw.2007.01.002

James, C. L. (2008). Electronic scoring of essays: Does topic matter? *Assessing Writing, 13*(2), 80–92. https://doi.org/10.1016/j.asw.2008.05.001

Jansen, T., Vögelin, C., Machts, N., Keller, S., Köller, O., & Möller, J. (2021). Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language. *Teaching and Teacher Education, 97*, 103216. https://doi.org/10.1016/j.tate.2020.103216

Jin, C., & He, B. (2015). Utilizing latent semantic word representations for automated essay scoring. 2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing, 2015 IEEE 12th International Conference on Autonomic and Trusted Computing, and 2015 IEEE 15th International Conference on Scalable Computing and

Communications and Its Associated Workshops (UIC-ATC-ScalCom), 1101–1108. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.202

Johnsi, R., & Kumar, G. B. (2024). Digital essay assessment using machine learning algorithms. *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, 1–13. https://doi.org/10.1109/ICONSTEM60960.2024.10568843

Kabra, A., Bhatia, M., Kumar, Y., Li, J. J., & Shah, R. R. (2021). Evaluation toolkit for robustness testing of automatic essay scoring systems (No. arXiv:2007.06796). *arXiv*. https://doi.org/10.48550/arXiv.2007.06796

Ke, Z., Inamdar, H., Lin, H., & Ng, V. (2019). Give me more feedback II: Annotating thesis strength and related attributes in student essays. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3994–4004. https://doi.org/10.18653/v1/P19-1390

Kharkwal, G., & Muresan, S. (2014). Surprisal as a predictor of essay quality. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 54–60. https://doi.org/10.3115/v1/W14-1807

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Technical Report EBSE 2007-001). Keele University & Durham University Joint Report.

Klebanov, B. B., & Higgins, D. (2012). Measuring the use of factual information in test-taker essays. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 63–72). Association for Computational Linguistics. https://aclanthology.org/W12-2007.pdf

Klobucar, A., Elliot, N., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing*, *18*(1), 62–84. https://doi.org/10.1016/j.asw.2012.10.001

Koza, J. R., Keane, M. A., Yu, J., & et al. (2000). Automatic creation of human-competitive programs and controllers by means of genetic programming. *Genetic Programming and Evolvable Machines, 1*(2), 121–164. https://doi.org/10.1023/A:1010076532029

Kwako, A., & Ormerod, C. (2024). Can language models guess your identity? Analyzing demographic biases in AI essay scoring. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 78–86). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.7.pdf

Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing, 45*, 100467. https://doi.org/10.1016/j.asw.2020.100467

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. *2016 4th International Conference on Cyber and IT Service Management*, 1–6. https://doi.org/10.1109/CITSM.2016.7577578

Laudenbach, M., Brown, D. W., Guo, Z., Ishizaki, S., Reinhart, A., & Weinberg, G. (2024). Visualizing formative feedback in statistics writing: An exploratory study of student motivation using DocuScope Write and Audit. *Assessing Writing, 60*, 100830. https://doi.org/10.1016/j.asw.2024.100830

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, 100213. https://doi.org/10.1016/j.caeai.2024.100213

Li, A. (2023). The construction and application of an automatic scoring system for English writing. *Procedia Computer Science*, *228*, 872-881. https://doi.org/10.1016/j.procs.2023.11.115

Li, C., Lin, L., Mao, W., Xiong, L., & Lin, Y. (2022). An automated essay scoring model based on stacking method. *2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI)*, 248-252. https://doi.org/10.1109/SEAI55746.2022.9832246

Li, X., Chen, M., & Nie, J.-Y. (2020). SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Systems, 210*, 106491. https://doi.org/10.1016/j.knosys.2020.106491

Li, X., Wen, Q., & Pan, K. (2017). Unsupervised off-topic essay detection based on target and reference prompts. *2017 13th International Conference on Computational Intelligence and Security (CIS)*, 465-468. https://doi.org/10.1109/CIS.2017.00108

Li, Y., & Yan, Y. (2010). Notice of retraction: Automated essay scoring system for CET4. *2010 Second International Workshop on Education Technology and Computer Science*, 94-97. https://doi.org/10.1109/ETCS.2010.80

Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. *System, 99*, 102505. https://doi.org/10.1016/j.system.2021.102505

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*, 66-78. https://doi.org/10.1016/j.system.2014.02.007

Liang, G., On, B.-W., Jeong, D., Heidari, A. A., Kim, H.-C., Choi, G. S., Shi, Y., Chen, Q., & Chen, H. (2021). A text GAN framework for creative essay recommendation. *Knowledge-Based Systems, 232*, 107501. https://doi.org/10.1016/j.knosys.2021.107501

Lim, F. V., & Phua, J. (2019). Teaching writing with language feedback technology. *Computers and Composition, 54*, 102518. https://doi.org/10.1016/j.compcom.2019.102518

Liu, B., Schlegel, V., Thompson, P., Batista-Navarro, R. T., & Ananiadou, S. (2023). Global information-aware argument mining based on a top-down multi-turn QA model. *Information Processing & Management, 60*(5), 103445. https://doi.org/10.1016/j.ipm.2023.103445

Liu, M., Li, Y., Xu, W., & Liu, L. (2017). Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies, 10*(4), 502-513. https://doi.org/10.1109/TLT.2016.2612659

Lotfy, N., Shehab, A., Elhoseny, M., & Abu-Elfetouh, A. (2023). An enhanced automatic Arabic essay scoring system based on machine learning algorithms. *Computers, Materials & Continua, 77*(1), 1227-1249. https://doi.org/10.32604/cmc.2023.039185

Madnani, N., Cahill, A., & Riordan, B. (2016). Automatically scoring tests of proficiency in music instruction. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 217-222. https://doi.org/10.18653/v1/W16-0524

Mahmoud, S., Nabil, E., & Torki, M. (2024). Automatic scoring of Arabic essays: A parameter-efficient approach for grammatical assessment. *IEEE Access*, *12*, 142555-142568. https://doi.org/10.1109/ACCESS.2024.3470728

Mardini, G. I. D., Quintero, C. G. M., Viloria, C. A. N., Percybrooks, W. S., Robles, H. S. N., & Villalba, R. K. (2024). A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer questions. *Education and Information Technologies, 29*(4), 4565-4590. https://doi.org/10.1007/s10639-023-11890-7

Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09396-9

Matta, M., Keller-Margulis, M. A., & Mercer, S. H. (2022). Cost analysis and cost-effectiveness of hand-scored and automated approaches to writing screening. *Journal of School Psychology, 92*, 80-95. https://doi.org/10.1016/j.jsp.2022.03.003

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35-59. https://doi.org/10.1016/j.asw.2014.09.002

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781

Mim, F. S., Inoue, N., Reisert, P., Ouchi, H., & Inui, K. (2021). Corruption is not all bad: Incorporating discourse structure into pre-training via corruption for essay scoring. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 2202-2215. https://doi.org/10.1109/TASLP.2021.3088223

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics, 3*(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Morris, W., Holmes, L., Choi, J. S., & Crossley, S. (2024). Automated scoring of constructed response items in math assessment using large language models. *International Journal of Artificial Intelligence in Education.* https://doi.org/10.1007/s40593-024-00418-w

Myers, M. C., & Wilson, J. (2023). Evaluating the construct validity of an automated writing evaluation system with a randomization algorithm. *International Journal of Artificial Intelligence in Education*, *33*(3), 609-634. https://doi.org/10.1007/s40593-022-00301-6

Nadeem, F., Nguyen, H., Liu, Y., & Ostendorf, M. (2019). Automated essay scoring with discourse-aware neural models. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 484-493. https://doi.org/10.18653/v1/W19-4450

Nagata, R., Kakegawa, J., & Yabuta, Y. (2009). A topic-independent method for automatically scoring essay content rivaling topic-dependent methods. *2009 Ninth IEEE International Conference on Advanced Learning Technologies*, 88-92. https://doi.org/10.1109/ICALT.2009.90

Nagata, R., Sato, T., & Takamura, H. (2018). Exploring the influence of spelling errors on lexical variation measures. *Proceedings of the 27th International Conference on*

Computational Linguistics (pp. 2391–2398). Association for Computational Linguistics. https://aclanthology.org/C18-1202.pdf

Nair, P. S., & Naga Malleswari, T. (2024). Automated essay scoring for IELTS using classification and regression models. *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, 478-483. https://doi.org/10.1109/AIC61668.2024.10731045

Nakamoto, S., Okamoto, Y., Nakakouchi, T., & Shimada, K. (2024). Towards human-level evaluation: Assessing the potential of GPT-4 in automated evaluation and feedback generation on Japanese essays. *156-161*. https://doi.org/10.1109/IIAI-AAI63651.2024.00039

Nyomi, X., & Moccozet, L. (2022). Anatomy of a large-scale real-time peer evaluation system. *2022 20th International Conference on Information Technology Based Higher Education and Training (ITHET)*, 1-9. https://doi.org/10.1109/ITHET56107.2022.10032005

Obata, A., Tagawa, T., & Ono, Y. (2023). Assessment of ChatGPT's validity in scoring essays by foreign language learners of Japanese and English. *2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, 105-110. https://doi.org/10.1109/IIAI-AAI-Winter61682.2023.00028

Okgetheng, B., & Takeuchi, K. (2024). Modeling score estimation for Japanese essays with generative pre-trained transformers. *In Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)* (pp. 64–73). Association for Computational Linguistics.

Page, E. B. (1966). The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan*, *47*(5), 238–243. http://www.jstor.org/stable/20371545

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Haffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated

guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology, 54,* 255-270. https://doi.org/10.1016/j.cedpsych.2018.07.002

Parker, L., Carter, C., Karakas, A., Loper, A. J., & Sokkar, A. (2024). Graduate instructors navigating the AI frontier: The role of ChatGPT in higher education. *Computers and Education Open, 6,* 100166. https://doi.org/10.1016/j.caeo.2024.100166

Peng, X., Ke, D., Chen, Z., & Xu, B. (2010). Automated Chinese essay scoring using vector space models. *2010 4th International Universal Communication Symposium,* 149-153. https://doi.org/10.1109/IUCS.2010.5666229

Peng, Y., Sun, J., Quan, J., Wang, Y., Lv, C., & Zhang, H. (2023). Predicting Chinese EFL learners' human-rated writing quality in argumentative writing through multidimensional computational indices of lexical complexity. *Assessing Writing, 56,* 100722. https://doi.org/10.1016/j.asw.2023.100722

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21*, 104–111. https://doi.org/10.1016/j.asw.2014.05.001

Perera, G. R., Perera, D. N., & Weerasinghe, A. R. (2015). A dynamic semantic space modelling approach for short essay grading. *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 43–49. https://doi.org/10.1109/ICTER.2015.7377665

Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education, 28*(1), 56–78. https://doi.org/10.1007/s40593-016-0122-z

Persing, I., Davis, A., & Ng, V. (2010). Modeling organization in student essays. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 229–239). Association for Computational Linguistics. https://aclanthology.org/D10-1023.pdf

Pino Castillo, P. A., Soto, C., Asún, R. A., & Gutiérrez, F. (2023). Profiling support in literacy development: Use of natural language processing to identify learning needs in higher education. *Assessing Writing, 58*, 100787. https://doi.org/10.1016/j.asw.2023.100787

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping *e-rater*: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*(2), 103–134. https://doi.org/10.1016/S0747-5632(01)00052-8

Qiu, X., Liao, S., Xie, J., & Nie, J.-Y. (2022). Tapping the potential of coherence and syntactic features in neural models for automatic essay scoring. *2022 International Conference on Asian Language Processing (IALP)*, 407–412. https://doi.org/10.1109/IALP57159.2022.9961308

Rababah, H., & Al-Taani, A. T. (2017). An automated scoring approach for Arabic short answers essay questions. *2017 8th International Conference on Information Technology (ICIT)*, 697–702. https://doi.org/10.1109/ICITECH.2017.8079930

Rahimi, Z., & Litman, D. (2016). Automatically extracting topical components for a response-to-text writing assessment. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 277–282. https://doi.org/10.18653/v1/W16-0532

Rahimi, Z., Litman, D., Correnti, R., Wang, E., & Matsumura, L. C. (2017). Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education, 27*(4), 694–728. https://doi.org/10.1007/s40593-017-0143-2

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Ramineni, C. (2013). Validating automated essay scoring for online writing placement. *Assessing Writing, 18*(1), 40–61. https://doi.org/10.1016/j.asw.2012.10.005

Reddy Chavva, R. K., Reddy Muthyam, S., Seelam, M. S., & Nalliboina, N. (2024). A transformer-based approach for enhancing automated essay scoring. *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, 1–6. https://doi.org/10.1109/ACET61898.2024.10730000

Reimers, N., Beyer, P., & Gurevych, I. (2016). Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 87–96). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1009.pdf

Riyanto, S., Lestari, F. A., Putri Riyanto, C., Ferdiansyah, A., Irfiani, E., Akbar Maulana Siagian, A. H., Sriyadi, Siahaan, F. B., & Fitria Apriani, N. (2024). Employing SBERT for essay assessment. *2024 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, 255–260. https://doi.org/10.1109/IC3INA64086.2024.10732430

Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior, 70*, 207–221. https://doi.org/10.1016/j.chb.2016.12.076

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment, 1*(2), Article 2. https://ejournals.bc.edu/index.php/jtla/article/view/1668

Ruipérez-Valiente, J. A., Staubitz, T., Jenner, M., Halawa, S., Zhang, J., Despujol, I., Maldonado-Mahauad, J., Montoro, G., Peffer, M., Rohloff, T., Lane, J., Turro, C., Li, X., Pérez-Sanagustín, M., & Reich, J. (2022). Large scale analytics of global and regional MOOC providers: Differences in learners' demographics, preferences, and

perceptions. *Computers & Education, 180,* 104426. https://doi.org/10.1016/j.compedu.2021.104426

Ruseti, S., Paraschiv, I., Dascalu, M., & McNamara, D. S. (2024). Automated pipeline for multi-lingual automated essay scoring with ReaderBench. *International Journal of Artificial Intelligence in Education, 34*(4), 1460-1481. https://doi.org/10.1007/s40593-024-00402-4

Schneider, J., Richner, R., & Riser, M. (2023). Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education, 33*(1), 88-118. https://doi.org/10.1007/s40593-022-00289-z

Sendra, M., Sutrisno, R., Harianata, J., Suhartono, D., & Asmani, A. B. (2016). Enhanced latent semantic analysis by considering mistyped words in automated essay scoring. *2016 International Conference on Informatics and Computing (ICIC)*, 304-308. https://doi.org/10.1109/IAC.2016.7905734

Sethi, A., & Singh, K. (2022). Natural language processing based automated essay scoring with parameter-efficient transformer approach. *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 749-756. https://doi.org/10.1109/ICCMC53470.2022.9753760

Sharma, A., & Jayagopi, D. B. (2018). Automated grading of handwritten essays. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 279-284. https://doi.org/10.1109/ICFHR-2018.2018.00056

Sharma, A., & Jayagopi, D. B. (2024). Modeling essay grading with pre-trained BERT features. *Applied Intelligence, 54*(6), 4979-4993. https://doi.org/10.1007/s10489-024-05410-4

Shermis, M. D. (2014a). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53-76. https://doi.org/10.1016/j.asw.2013.04.001

Shermis, M. D. (2014b). The challenges of emulating human behavior in writing assessment. *Assessing Writing, 22*, 91-99. https://doi.org/10.1016/j.asw.2014.07.002

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective.* Lawrence Erlbaum Associates Publishers.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In *International Encyclopedia of Education* (pp. 20-26). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00233-5

Song, W., Wang, D., Fu, R., Liu, L., Liu, T., & Hu, G. (2017). Discourse mode identification in essays. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 112-122. https://doi.org/10.18653/v1/P17-1011

Stephen, T. C., Gierl, M. C., & King, S. (2021). Automated essay scoring (AES) of constructed responses in nursing examinations: An evaluation. *Nurse Education in Practice, 54*, 103085. https://doi.org/10.1016/j.nepr.2021.103085

Suresh, V., Agasthiya, R., Ajay, J., Gold, A. A., & Chandru, D. (2023). AI-based automated essay grading system using NLP. *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 547-552. https://doi.org/10.1109/ICICCS56967.2023.10142822

Suriyasat, S., Chanyachatchawan, S., & Tuaycharoen, N. (2023). A comparison of machine learning and neural network algorithms for an automated Thai essay scoring. *2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 55-60. https://doi.org/10.1109/JCSSE58229.2023.10201964

Syed Abuthahir, S., Swarnamughi, K., Bhasin, N. K., Lawrance, J. C., Rengarajan, M., & Alphonse, F. R. (2024). Building a deep neural network for automated essay scoring in English language teaching. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1-6. https://doi.org/10.1109/ICCCNT61001.2024.10725936

Tashu, T. M. (2020). Off-topic essay detection using C-BGRU Siamese. *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, 221–225. https://doi.org/10.1109/ICSC.2020.00046

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence, 7*, 100255. https://doi.org/10.1016/j.caeai.2024.100255

Thongyoo, T., Saelee, S., & Krootjohn, S. (2016). Automated Thai online assignment scoring. *2016 Fifth ICT International Student Project Conference (ICT-ISPC)*, 33–36. https://doi.org/10.1109/ICT-ISPC.2016.7519229

Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics, 4*(1), 100089. https://doi.org/10.1016/j.acorp.2024.100089

Ulum, Ö. G. (2020). A critical deconstruction of computer-based test application in Turkish State University. *Education and Information Technologies, 25*(6), 4883–4896. https://doi.org/10.1007/s10639-020-10199-z

Uto, M., Aomi, I., Tsutsumi, E., & Ueno, M. (2023). Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Transactions on Learning Technologies, 16*(6), 983–1000. https://doi.org/10.1109/TLT.2023.3253215

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education, 28*(1), 79–105. https://doi.org/10.1007/s40593-017-0142-3

Vista, A., Care, E., & Griffin, P. (2015). A new approach towards marking large-scale complex assessments: Developing a distributed marking system that uses an automatically scaffolding and rubric-targeted interface for guided peer-review. *Assessing Writing, 24*, 1–15. https://doi.org/10.1016/j.asw.2014.11.001

Vo, Y., Rickels, H., Welch, C., & Dunbar, S. (2023). Human scoring versus automated scoring for English learners in a statewide evidence-based writing assessment. *Assessing Writing, 56*, 100719. https://doi.org/10.1016/j.asw.2023.100719

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition, 28*(2), 97–111. https://doi.org/10.1016/j.compcom.2011.04.004

Walia, T. S., Josan, G. S., & Singh, A. (2019). An efficient automated answer scoring system for Punjabi language. *Egyptian Informatics Journal, 20*(2), 89–96. https://doi.org/10.1016/j.eij.2018.11.001

Wang, D., Song, Y., Li, J., Han, J., & Zhang, H. (2018). A hybrid approach to automatic corpus generation for Chinese spelling check. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2517–2527. https://doi.org/10.18653/v1/D18-1273

Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., Magooda, A., & Quintana, R. (2020). eRevis(ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing, 44*, 100449. https://doi.org/10.1016/j.asw.2020.100449

Wang, Q. (2024). A multifaceted architecture to automate essay scoring for assessing English article writing: Integrating semantic, thematic, and linguistic representations. *Computers and Electrical Engineering, 118*, 109308. https://doi.org/10.1016/j.compeleceng.2024.109308

Wangkriangkri, P., Viboonlarp, C., Rutherford, A. T., & Chuangsuwanich, E. (2020). A comparative study of pretrained language models for automated essay scoring with adversarial inputs. *2020 IEEE Region 10 Conference (TENCON)*, 875–880. https://doi.org/10.1109/TENCON50793.2020.9293930

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*(1), 85–99. https://doi.org/10.1016/j.asw.2012.10.006

Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., & Trausan-Matu, S. (2018). Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online

training. *Computers and Education, 123*, 212–224. https://doi.org/10.1016/j.compedu.2018.05.010

Wild, F., Hoisl, B., & Burek, G. (2009). Positioning for conceptual development using latent semantic analysis. *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics - GEMS '09*, 41–48. https://doi.org/10.3115/1705415.1705421

Wilkens, R., Seibert, D., Wang, X., & François, T. (2022). MWE for essay scoring English as a foreign language. In Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference (pp. 62–69). European Language Resources Association. https://aclanthology.org/2022.readi-1.9.pdf

Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology, 68*, 19–37. https://doi.org/10.1016/j.jsp.2017.12.005

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers and Education, 168*, 104208. https://doi.org/10.1016/j.compedu.2021.104208

Wilson, J., & Huang, Y. (2024). Validity of automated essay scores for elementary-age English language learners: Evidence of bias? *Assessing Writing, 60*, 100815. https://doi.org/10.1016/j.asw.2024.100815

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing, 27*, 11–23. https://doi.org/10.1016/j.asw.2015.06.003

Wilson, J., & Rodrigues, J. (2020). Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology, 82*, 123–140. https://doi.org/10.1016/j.jsp.2020.08.008

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing, 34*, 16–36. https://doi.org/10.1016/j.asw.2017.08.002

Wu, Y. (2023). English composition assisted writing training based on big data text mining. *2023 International Conference on Industrial IoT, Big Data and Supply Chain (IIoTBDSC)*, 130–134. https://doi.org/10.1109/IIoTBDSC60298.2023.00031

Xue, S., Zhang, J., Zhou, J., & Ren, F. (2022). Robust automated essay scoring by using attentive capsule. *2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, 595–599. https://doi.org/10.1109/CCIS57298.2022.10016365

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics, 3*(3), 100133. https://doi.org/10.1016/j.rmal.2024.100133

Yang, S. (2023). Automated English essay scoring based on machine learning algorithms. *2023 2nd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)*, 263–266. https://doi.org/10.1109/ICDACAI59742.2023.00056

Yoon, S.-Y., Cahill, A., Loukina, A., Zechner, K., Riordan, B., & Madnani, N. (2018). Atypical inputs in educational applications. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, 60–67. https://doi.org/10.18653/v1/N18-3008

Zainal, N., & Abu Hassan, M. H. (2022). Automated essay scoring (AES) using English essay question. *2022 IEEE 20th Student Conference on Research and Development (SCOReD)*, 1–6. https://doi.org/10.1109/SCOReD57082.2022.9973989

Zhang, H., & Litman, D. (2018). Co-attention based neural network for source-dependent essay scoring. In J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building*

Educational Applications (pp. 399–409). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0549

Zhao, C., & Huang, J. (2020). The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers. *Studies in Educational Evaluation, 67,* 100911. https://doi.org/10.1016/j.stueduc.2020.100911

Zhao, R., Zhuang, Y., Zou, D., Xie, Q., & Yu, P. L. H. (2023). AI-assisted automated scoring of picture-cued writing tasks for language assessment. *Education and Information Technologies, 28*(6), 7031–7063. https://doi.org/10.1007/s10639-022-11473-y

Zhou, L. (2022). Construction of English writing hybrid teaching model based on machine learning automatic composition scoring system. *Procedia Computer Science, 208*, 384–390. https://doi.org/10.1016/j.procs.2022.10.054

Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems, 120,* 118–132. https://doi.org/10.1016/j.knosys.2017.01.006

.

EKLER

Ek-1: Otomatik Metin Puanlama Konulu Sistematik Literatür Taramasına İlişkin Veri Tablosu

Author	title	pub_ date	scoring_ type	dataset_ used	content	valid_ metric	ml_techniq ue
(Alqahtani ve Alsaif, 2020)	Automated Arabic Essay Evaluation	2020	Arabic essay scoring using multiple criteria	Essays written by university students and journalists	Evaluates essays on spelling, structure, coherence, style, and punctuation without requiring model essays.	Correlation with human scores (0.87)	Support Vector Regression (SVR)
(Burstein ve Wolska, 2003)	Toward Evaluation of Writing Style: Finding Repetitive Word Use	2003	Writing style evaluation	296 essays (Grade 6 to College)	Detects repetitive word use in essays using human-annotated corpora and machine learning.	F-measure (0.90)	Decision-based machine learning (C5.0 algorithm)
(Chali ve Hasan, 2013)	On the Effectiveness of Using Syntactic and Shallow Semantic Tree Kernels for Automatic Assessment of Essays	2013	Essay grading with structural kernels	Human-scored essays and course materials	Combines syntactic and semantic tree kernels to enhance LSA-based grading models.	Improvement in similarity metrics (Cosine)	Latent Semantic Analysis (LSA), tree kernel methods
(Clark vd., 2019)	Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts	2019	Text similarity for machine-generated summaries	Human-authored essays and summaries	Introduces sentence mover's similarity metric combining word and sentence embeddings for better alignment.	Correlation with human judgments	Sentence embedding-based similarity metrics
(del Río, 2019)	Linguistic Features and Proficiency Classification in L2 Spanish and Prtuguese	2019	Proficiency classificatio n for Spanish and Portuguese	NLI-PT and CEDEL2 corpora	Classifies CEFR levels for L2 Spanish and Portuguese using linguistic features like syntax and morphology.	Pearson correlation with CEFR levels	SVM, Feature-based classificatio n
(Fiacco vd., 2022)	Toward Automatic Discourse Parsing of Student Writing Motivated by Neural Interpretation	2022	Discourse-based essay feedback	RST Discourse Treebank and a new student dataset	Proposes a neural parser for rhetorical structure analysis, focusing on improving coherence feedback.	RST parsing accuracy	Neural network-based RST parsers
(Hellman vd., 2023)	Scalable and Explainable Automated Scoring for Open-Ended Constructed Responce Math Word Problems	2023	Math word problem scoring	34,417 student responses	Develops an explainable scoring model for structured rubrics in math word problems combining text and formulas.	Quadratic Weighted Kappa (QWK)	Feature-aligned scoring models
(Hirao vd., 2020)	Automated Essay Scoring System for Nonnative Japanese Learners	2020	Multi-trait Japanese essay scoring	Japanese learner essay dataset	Uses both feature-based and neural models (BERT, LSTM) to score essays on content, organization, and language.	Root Mean Squared Error (RMSE), QWK	Feature-based models, BERT, LSTM
(Horbach vd., 2017)	Fine-Grained Essay Scoring of a Complex Writing Tasks for Native Speakers	2017	Fine-grained scoring for native German essays	German essays from prospective university students	Scores essays using detailed rubrics for complex writing tasks, addressing high proficiency variability.	Holistic and rubric-specific accuracy	Neural networks, SVM
(Ke vd., 2019)	Give Me More Feedback II: Annotating Thesis Strength and Related Attributes in Stdent Essays	2019	Thesis strength and attribute scoring	ICLE corpus annotated with thesis attributes	Evaluates thesis strength and related dimensions in persuasive essays for instructional feedback.	Agreement with annotated attributes	Feature-based modeling

(Kharkwal ve Muresan, 2014)	Surprisal as a Predictor of Essay Quality	2014	Essay quality prediction using surprisal theory	Learner essays evaluated for processing complexity	Investigates how surprisal, a measure of sentence complexity, predicts essay scores for EFL learners.	Correlation between surprisal and scores	Psycholinguistic surprisal modeling
(Kwako ve Ormerod, 2024)	Can Language Models Guess Your Identity? Analyzing Demographic Biases in AI Essay Scoring	2024	Bias analysis in AES systems	PERSUADE corpus	Analyzes demographic biases in automated scoring using metadata such as race, gender, and socioeconomic status.	QWK, bias analysis	XLNet, transformer-based bias analysis
(Madnani vd., 2016)	Automatically Scoring Tests of Proficiency in Music Instruction	2016	Content scoring for music education essays	Teacher certification test responses	Scores constructed responses for music education certification, focusing on content-based scoring.	Scoring agreement and performance metrics	Simple feature-based models
(Nadeem vd., 2019)	Automated Essay Scoring with Discourse-Aware Neural Models	2019	Discourse-aware AES for essay quality	TOEFL and other essay datasets	Explores neural architectures incorporating discourse structure for improved scoring of longer essays.	Performance improvement over feature-based models	Hierarchical attention models, BERT embeddings
(Nagata vd., 2018)	Exploring the Influence of Spelling Errors on Lexical Variation Measures	2018	Lexical variation metrics with spelling correction	English learner corpora	Examines the impact of spelling errors on measures like TTR and Yule's K for learner assessment.	Stability analysis of lexical metrics	Text preprocessing and correction
(Okgetheng ve Takeuchi, 2024)	Modeling Score Estimation for Japanese Essays with Generative Pre-trained Transformers	2024	Japanese essay scoring with GPT models	Essays on 12 prompts (Japanese GPT dataset)	Examines GPT-based models for scoring Japanese essays with fine-tuning via LoRA, achieving moderate QWK.	Quadratic Weighted Kappa (QWK), accuracy	GPT-based fine-tuning with LoRA
(Persing vd., 2010)	Modeling Organization in Student Essays	2010	Essay organization modeling	1003 essays from ICLE corpus	Proposes computational models for essay organization scoring using alignment and string kernels.	Mean squared error and score correlation	Sequence alignment, alignment kernels
(Rahimi ve Litman, 2016)	Automatically Extracting Topical Components for a Response-to-Text Writing Assessment	2016	Evidence dimension scoring	RTA dataset	Introduces LDA-based automatic topic extraction for response-based essay scoring, reducing expert dependency.	Scoring agreement with manually extracted topics	LDA topic modeling
(Reimers vd., 2016)	Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity	2016	Semantic similarity evaluation for essays	Wikipedia Rewrite Corpus, news corpus	Questions the effectiveness of Pearson correlation in evaluating STS systems and proposes task-specific metrics.	Pearson correlation, alternative evaluation metrics	Task-specific intrinsic evaluation framework
(Song vd., 2017)	Discourse Mode Identification in Essays	2017	Discourse mode analysis	Annotated Chinese student essays	Identifies sentence-level discourse modes like narration, exposition, and argument to improve AES performance.	F1-score for discourse mode classification	Neural sequence labeling
(D. Wang vd., 2018)	A Hybrid Approach to Automatic Corpus Generation for Chinese Spelling Check	2018	Chinese spelling correction corpus creation	OCR-generated Chinese texts	Constructs a hybrid corpus with visually and phonetically induced errors for CSC evaluation.	Spelling error detection accuracy	BiLSTM, OCR, ASR methods
(Wilkens vd., 2022)	MWE for Essay Scoring English as a Foreign Language	2022	Multiword expressions for English essay scoring	Essays annotated with multiword expression data	Investigates the role of MWE concreteness and usage in predicting English proficiency levels.	Accuracy of MWE-based and classical features	Gradient Boosted Trees, feature engineering
(Yoon vd., 2018)	Atypical Inputs in Educational Applications	2018	Atypical input handling in AES systems	ETS assessment datasets for essays and speech	Describes filtering models for atypical or problematic inputs in essay and spoken response scoring systems.	Detection rates for non-scorable responses	NLP-based filters, anomaly detection

(Allen vd., 2016)	Investigating Boredom and Engagement during Writing Using Multiple Sources of Information: The Essay, The Writer, and Keystrokes	2016	Affect detection in writing	132 writing sessions	Uses keystrokes, individual differences, and text indices to predict engagement and boredom in writing sessions.	Classification accuracy (76.5%-77.3%)	NLP-based stealth assessment, feature engineering
(Allen vd., 2018)	A Multi-Dimensional Analysis of Writing Flexibility in an Automated Writing Evaluation System	2018	Writing flexibility analysis	Automated Writing Evaluation System (131 essays)	Studies the role of linguistic flexibility in essays and its relation to feedback and reading comprehension skills.	Correlation and statistical significance testing	Dynamic NLP modeling
(Cheng vd., 2024)	Evidence-Centered Assessment for Writing with Generative AI	2024	Collaborative writing with generative AI	CoAuthor writing tool	Explores human-AI collaboration in writing using epistemic network analysis and evidence-centered design (ECD).	Epistemic network differences	Evidence-centered design and AI interaction modeling
(Ishioka ve Kameda, 2006)	Automated Japanese Essay Scoring System Based on Articles Written by Experts	2006	Japanese essay scoring	Mainichi Daily News articles	Evaluates essays by comparing rhetoric, organization, and vocabulary to expert-written editorials and columns.	Statistical outlier detection, regression analysis	NLP-based models, statistical comparisons
(Kabra vd., 2021)	Evaluation Toolkit for Robustness Testing of Automatic Essay Scoring Systems	2021	Robustness testing for AES systems	ASAP dataset	Proposes adversarial evaluation metrics to test AES models on coherence, grammar, and topic relevance.	Adversarial performance drop, QWK	Black-box adversarial evaluation
(Klebanov ve Higgins, 2012)	Measuring the Use of Factual Information in Test-Taker Essays	2012	Factual content evaluation	Essays with factual information	Studies the correlation between factuality and essay scores; introduces robust and simple measurement techniques.	Medium-strength correlation with essay scores	Fact-based NLP analysis
(Wild vd., 2009)	Positioning for Conceptual Development using Latent Semantic Analysis	2009	Conceptual positioning using semantic analysis	Learner portfolios	Uses LSA to analyze learner progress and align learning materials with individual knowledge development.	Semantic similarity thresholds, cosine similarity values	Latent Semantic Analysis (LSA)
(Lotfy vd., 2023)	An Enhanced Automatic Arabic Essay Scoring System Based on Machine Learning Algorithms	2023	Corpus-based and string-based	270 short answers (sociology course)	Focuses on Arabic essay scoring using similarity measures and adaptive fusion engines to enhance system efficiency.	Pearson's Correlation Coefficient, Willmot's Index	Random Forest, Decision Tree, Adaboost, KNN, Bagging, Lasso
(Perelman, 2014)	When "the state of the art" is counting words	2014	Word count-based AES	Kaggle dataset from the Automated Student Assessment Prize	Critiques AES systems for over-privileging essay length over content quality.		
(Laudenbach vd., 2024)	Visualizing formative feedback in statistics writing: An exploratory Study of Student Motivation Using DocuScope Write ve Audit	2024	Automated Writing Evaluation (AWE)	Data from 30 students in introductory-level statistics course	Explores formative feedback in statistics writing using a visualization tool, improving student motivation and attitudes.	Student motivation construct model survey	
(Wilson ve Huang, 2024)	Validity of automated essay scores for elementary-age English Language Learners: Evidence of Bias?	2024	MI Write AES	2829 students in Grades 3–5, including ELLs and non-ELLs	Examines the predictive validity of automated vs. human scores for elementary English learners in formative assessments.	Predictive validity through regression analyses	Multilevel regression modeling

(James, 2006)	Validating a Computerized Scoring System for Assessing Writing and Placing Students in Composition Courses	2006	Holistic Scoring System	ACCUPLACER™ WritePlacer Plus	Evaluates IntelliMetric™ scores and human grader scores for placing students in composition courses, achieving 77% placement accuracy.	Correlation (0.63), Logistic Regression	IntelliMetric™, Logistic Regression
(Ramineni, 2013)	Validating of Automated Essay Scoring for Online Writing Placement	2013	Criterion Online Writing Evaluation	879 first-year students writing persuasive essays	Validates AES models for online placement in writing programs, comparing AES and human scoring systems.	Inter-rater correlation, model predictive validity	
(Meyer vd., 2024)	Using LLMs to bring evidence-based feedback into classroom: AI-generated feedback increases secondary students' text revision, motivation, and psitive emotions	2024	LLM-based feedback	459 secondary students' argumentative essays	Investigates LLM-based feedback effectiveness in improving revision quality and motivation among secondary students.	Effect size (Cohen's d)	GPT-3.5 Turbo
(Uchida, 2024)	Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations	2024	LLM-based linguistic analysis	Outputs from ChatGPT and COCA datasets	Evaluates LLMs like ChatGPT for linguistic tasks, including frequency lists, collocations, and grammatical patterns.	F1 scores, accuracy rates	GPT-based models
(Y. Huang ve Wilson, 2021)	Using automated feedback to develop writing proficiency	2021	MI Write AES	431 Grades 4–5 students, longitudinal study	Studies the effect of automated feedback on students' independent and aided transfer gains in writing quality over time.	Hierarchical growth curve modeling	
(Wilson, 2018)	Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4	2018	Project Essay Grade (PEG)	230 Grade 3 and Grade 4 students	Explores the use of automated scoring for screening elementary students at risk of failing Common Core ELA standards.	ROC curve analysis, AUC values	Logistic regression
(Z. Li vd., 2014)	The role of automated writing evaluation holistic scores in the ESL classroom	2014	Criterion AWE	67 university ESL students	Investigates correlations between automated scores and human ratings, focusing on holistic feedback utility in ESL classrooms.	Correlation coefficients	Latent Semantic Analysis (LSA), Criterion e-rater
(Kyle, 2020)	The relationship between features of source text use and integrated writing quality	2020	Automated scoring (TOEFL essays)	TOEFL integrated tasks	Examines source text features' relation to writing quality in integrated academic tasks.	Regression analysis, variance explained	
(C. Zhao ve Huang, 2020)	The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers	2020	Generalizability (G-) theory-based	120 CET-4 essays from 60 students	Analyzes scoring reliability and variability in Chinese national EFL exams.	Generalizability coefficients, rater interviews	
(A. Li, 2023)	The Construction and Application of an Automatic Scoring System for English Writing	2023	Statistical and big data models	University-level student essays	Develops an automatic English writing scoring system combining teacher feedback and statistical methods.	Pre- and post-test analysis	Statistical and NLP-based modeling

(Shermis, 2014b)	The challenges of emulating human behavior in writing assessment	2014	Automated essay scoring (Hewlett Trials)	Hewlett dataset	Critiques the construct validity of AES systems, analyzing correlation issues and scoring inconsistencies.	Weighted kappa, correlation coefficients	Regression-based modeling
(Palermo ve Thomson, 2018)	Teacher implementation of Self-Regulated Strategy Development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students	2018	NC Write AWE system	829 middle school students	Evaluates SRSD strategies combined with AWE for improving argumentative writing in middle schools.	Multilevel modeling	Regression-based modeling
(Z. Li, 2021)	Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence	2021	Criterion AWE	University ESL classrooms	Explores teacher perceptions and implementation of AWE in ESL classes and its impact on student revision and grammatical accuracy.	Observational metrics, student behavior analysis	
(Powers vd., 2002)	Stumping e-rater: Challenging the validity of automated essay scoring	2002	E-rater AES system	GRE Writing Assessment prompts	Assesses how e-rater can be manipulated and evaluates potential weaknesses in automated essay scoring.	Human rater comparison, discrepancy analysis	NLP-based modeling
(Lim ve Phua, 2019)	Teaching Writing with Language Feedback Technology	2019	Linguistic Feedback Tool (LiFT)	Singapore secondary school students	Explores LiFT's effectiveness in improving grammatical accuracy and reducing teachers' marking time.	Student progress measures, teacher marking time reduction	NLP-based modeling (Ginger Sof
(Shermis, 2014a)	State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demenstration	2014	Automated Essay Scoring (AES)	Hewlett Foundation ASAP dataset	Compares commercial AES systems and evaluates their performance against human raters.	Quadratic Weighted Kappa	Regression models, ensemble methods
(X. Li vd., 2020)	SEDNN: Shared and enhanced deep neural network model for crass-prompt automated essay scoring	2020	Deep Neural Network AES	Multiple cross-prompt essay datasets	Proposes a cross-prompt AES model leveraging shared and prompt-dependent features.	Pearson correlation, RMSE, Quadratic Weighted Kappa	Shared deep neural networks (SEDNN)
(Mizumoto vd., 2024)	Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment	2024	AI-based language assessment	Cambridge Learner Corpus (CLC FCE)	Explores ChatGPT's ability to assess linguistic accuracy compared to human evaluators and Grammarly.	Correlation coefficients, accuracy rates	ChatGPT as evaluation model
(Azmi vd., 2019)	AAEE – Automated evaluation of students' essays in Arabic language	2019	AEE for Arabic essays	350 essays (Grades 7–12)	Develops an Arabic essay evaluation system using Latent Semantic Analysis and Rhetorical Structure Theory.	Human-machine correlation (0.756)	Latent Semantic Analysis (LSA), Rhetorical Structure Theory

(Butcher ve Jordan, 2010)	A comparison of human and computer marking of short free-text student responses	2010	Short-answer automated marking	Open University student responses	Compares three computerized systems for marking short free-text responses, focusing on marking accuracy.	Accuracy comparison, human-computer agreement	Computational linguistics, NLP-based matching
(McNamara vd., 2015)	A hierarchical classification approach to automated essay scoring	2015	Hierarchical AES	1243 argumentative essays (Grades 9, 11, college freshmen)	Uses hierarchical classification for AES, emphasizing linguistic and semantic features.	Exact/adjacent accuracy (55%/92%)	Coh-Metrix, LIWC, Writing Assessment Tool (WAT)
(Q. Wang, 2024)	Multifaceted architecture for AES for assessing english article writing: Integrating semantic, thematic, linguistic representations	2024	Deep learning hybrid AES	ASAP and CLC-FCE datasets	Integrates deep learning (ConvNet, LSTM) with linguistic and thematic feature extraction for AES.	Quadratic Weighted Kappa, Pearson's correlation	ConvNet, LSTM, Doc2Vec
(Yamashita, 2024)	An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0	2024	ChatGPT-4.0 AES	ICNALE corpus (136 essays)	Evaluates ChatGPT-4.0 using Many-Facet Rasch Measurement to assess rater severity and bias.	Rasch measurement, correlation analysis	ChatGPT-4.0
(Goura vd., 2022)	An efficient and enhancement of recent approaches to build an automated essay scoring system	2022	Regression-based AES	Benchmark datasets	Explores MLR and Xg-Boost for improving AES efficiency and accuracy.	Evaluation metrics comparison quadratic weighted kappa score, pearson correlation coefficient	Multiple Linear Regression (MLR), Xg-Boost
(Filho vd., 2018)	An approach to evaluate adherence to theme and structure in argumentative essays	2018	Theme-based AES	Brazilian ENEM essays	Uses machine learning to assess adherence to themes and argumentative structure of essays.	Classification accuracy, regression analysis	Support Vector Machines (SVM), NLP features
(Walia vd., 2019)	An efficient automated answer scoring system for Punjabi language	2019	NLP-based AES	Punjabi short answer dataset	Explores automated scoring for Punjabi using semantic and grammatical measures with LSTM and RNN architectures.	Accuracy, temporal dependency analysis	Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN)
(Vista vd., 2015)	A new approach towards marking large-scale complex assessments: Developing a distributed marking system that uses an automaticaly scaffolding and rubric-targetted interface for guided peer-review	2015	Distributed peer-marking system	MOOC assessment data	Develops a rubric-guided peer-review system to address reliability issues in large-scale assessments.	Construct validity, peer-review agreement	
(Fazal vd., 2013)	An innovative approach for automatically grading spelling in essays using rubric-based scoring	2013	Rubric-based AES	NAPLAN literacy dataset	Presents a rubric-based method for spelling evaluation in essays, focusing on categorical word classification accuracy.	Word classification accuracy, rubric consistency	Heuristic-based rules and custom algorithms

(ElNaka vd., 2021)	AraScore: Investigating response-based Arabic short answer scoring	2021	Arabic short answer scoring	AraScore dataset, AR-ASAG, Hewlett SAS datasets	Introduces a novel dataset for Arabic short answers and evaluates models for unbiased and scalable scoring.	Performance metrics comparison, scoring bias	Supervised learning, dataset-specific analysis
(Lee vd., 2024)	Applying large language models and chain-of-thought for automatic scoring	2024	GPT-based LLM scoring	Science assessments (1,650 student responses)	Evaluates GPT-3.5/4 with Chain-of-Thought techniques, showing improvements in accuracy and interpretability.	Scoring accuracy, interpretability measures	Chain-of-Thought (CoT), Few-shot/Zero-shot learning
(Blake ve Gutierrez, 2011)	A semantic analysis approach for assessing professionalism using free-form text endered online	2011	Semantic analysis-based AES	Professional development program text responses	Measures professionalism in free-form responses using Latent Semantic Analysis (LSA).	Dimensional construct analysis	Latent Semantic Analysis (LSA)
(Wilson & Czik, 2016)	Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality	2016	Automated Essay Evaluation (AEE)	PEG Writing® platform dataset	Examines the impact of AEE on teacher feedback efficiency, student motivation, and writing quality.	Feedback time reduction, student motivation impact	PEG Writing®
(Liang vd., 2021)	A text GAN framework for creative essay recommendation	2021	Generative Adversarial Network (GAN)	Adapted ASAP dataset	Uses GANs to recommend creative essays by evaluating hidden patterns and originality in essay datasets.	Creativity classification metrics, GAN accuracy	Generative Adversarial Networks (GAN)
(Zupanc ve Bosnić, 2017)	Automated essay evaluation with semantic analysis	2017	Semantic-based AEE	Various high-stakes assessment datasets	Extends AEE systems with semantic coherence and consistency measures for improved grading accuracy.	Coherence and consistency metrics	Semantic coherence models, information extraction
(Stephen vd., 2021)	Automated essay scoring (AES) of constructed responses in nursing examinations: An evaluation	2021	Constructed-response AES	Nursing exam responses	Evaluates AES for higher-level thinking in nursing education, comparing human and automated scoring reliability.	Inter-rater agreement, scoring reliability measures	Natural Language Processing, keyword matching
(Dikli ve Bleyle, 2014)	Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback?	2014	AES (Criterion®)	37 essays from 14 ESL students	Investigates the discrepancies between automated and instructor feedback for ESL students in grammar, usage, and mechanics.	Feedback accuracy, student perception analysis	NLP-based Criterion scoring engine
(Li vd., 2015)	Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction	2015	AWE (Criterion®)	University ESL classroom essays	Examines Criterion's impact on student revisions and writing accuracy improvements in ESL classrooms.	Correlation coefficients, mixed-methods analysis	Criterion's NLP-based scoring
(Westera vd., 2018)	Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training	2018	NLP-based AES	173 Dutch-language student reports	Evaluates AES to reduce teacher workload while maintaining scoring precision in an educational simulation game.	Precision (95%), ROC analysis	NLP-based textual complexity indices
(Wilson vd., 2017)	Automated formative writing assessment using a levels of language framework	2017	Formative AWE	480 essays (Grades 6 and 8)	Develops a model using word, sentence, and discourse-level measures to predict state writing test performance.	Multigroup structural equation modeling, DAW scores	Coh-Metrix, latent factor analysis

(Klobucar vd., 2013)	Automated scoring in context: Rapid assessment for placed students	2013	Criterion AES	1,482 first-year university students	Validates Criterion AES for identifying at-risk students and improving course success outcomes.	Weighted kappa, correlation with human scoring	NLP-based statistical aggregation
(Tate vd., 2024)	Can AI provide useful holistic essay scoring?	2024	GPT-based AES	Essays from three student corpora	Compares ChatGPT's scoring outputs to human scores for low-stakes formative assessments.	Quadratic Weighted Kappa, adjacent agreement	ChatGPT (LLM-based)
(Wilson ve Rodrigues, 2020)	Classification accuracy and efficiency of writing screening using AES	2020	Project Essay Grade (PEG)	Essays from Grades 3–5 students	Assesses AES for classifying at-risk students and compares with word count methods for efficiency and accuracy.	ROC curve analysis, classification accuracy	PEG scoring engine
(Andersen vd., 2023)	Automatic proficiency scoring for early-stage writing	2023	Rasch-based AES	Danish early-stage writing dataset	Develops fine-grained AES for early writing stages in Danish, focusing on sentence construction and modifiers.	Statistical Rasch modeling, feature reliability	Machine learning and NLP
(Wilson vd., 2016)	Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assesment	2016	AES and word-based scoring	Essays from Grade 6 students	Compares AES to holistic and word-based scoring methods for identifying students at risk of failing state writing tests.	Area under ROC curve, classification accuracy	Project Essay Grade (PEG), benchmark testing
(Matta vd., 2022)	Cost analysis and cost-effectiveness of hand-scored and automated approachs to writing screening	2022	Automated and manual scoring	Secondary datasets from WE-CBM studies	Analyzes cost-effectiveness of automated vs manual scoring approaches in elementary writing screening.	Cost-effectiveness ratio	Automated text evaluation tools
(Zhou, 2022)	Construction of English Writing Hybrid Teaching Model Based on Machine Learning Automatic Composition Scoring System	2022	Hybrid AES system	Bingguo Intelligent Compositin Scoring System	Proposes a hybrid teaching model integrating AES into English writing instruction to enhance motivation and autonomy.	Teaching improvement metrics	NLP, Hybrid Teaching Framework
(Bridgeman ve Ramineni, 2017)	Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings	2017	AWE (Automated Writing Evaluation)	194 graduate students' coursework writing	Evaluates AWE models' effectiveness in predicting real-world writing tasks and associated subgroup differences.	Correlation, subgroup performance analysis	Statistical modeling
(Wilson vd., 2021)	Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation	2021	MI Write AWE	14 schools, Grades 3–5 students	Explores teacher perceptions of AWE's impact on writing motivation, challenges, and classroom integration.	Qualitative thematic coding, mixed-methods analysis	MI Write scoring engine
(James, 2008)	Electronic scoring of essays: Does topic matter?	2008	IntelliMetric AES	ACCUPLACER® WritePlacer® essays	Investigates the impact of essay prompts on AES scores, revealing minimal topic-based score variation.	Agreement rates, Pearson correlations	IntelliMetric, holistic scoring
(Weigle, 2013)	English language learners and automated scoring of essays: Critical considerations	2013	E-rater, TOEFL AES	TOEFL iBT tasks	Explores challenges and biases in using AES for second language learners, highlighting validity and construct definition.	Validity framework, score agreement rates	E-rater scoring engine

(E. L. Wang vd., 2020)	eRevis(ing): Students' revision of text evidence use in an AWE system	2020	AWE	143 students' essays across Grades 5 and 6	Evaluates student revisions prompted by AWE feedback, focusing on argument and evidence use improvements.	Revision quality analysis, student response tracking	eRevise platform
(Mizumoto ve Eguchi, 2023)	Exploring the potential of using an AI language model for automated essay scoring	2023	GPT-based AES	TOEFL11 Corpus, 12,100 essays	Assesses GPT-3's effectiveness in AES and explores the role of linguistic features in enhancing score accuracy.	Accuracy, reliability, linguistic analysis	GPT-3 (text-davinci-003)
(Abdi Tabari & Johnson, 2023)	Exploring new insights into the role of cohesive devices in written academic genres	2023	Computational linguistic analysis	270 essays, narrative and argumentative genres	Analyzes cohesive device use in academic writing, linking cohesion features to essay quality across genres.	Regression models, human ratings	TAACO computational analysis
(Bonthu vd., 2024)	Framework for automation of short answer grading based on domain-specific pre-training	2024	Automated Short Answer Grading (ASAG)	SPRAG dataset	Proposes a domain-specific pretraining framework for ASAG using unlabeled corpora to improve grading accuracy.	Accuracy, F1-score (88.77%, 91.59%)	Pretrained Language Models (BERT, fine-tuning)
(Parker vd., 2024)	Graduate instructors navigating the AI frontier: The role of ChatGPT in higher education	2024	AI-based assessment	AI-generated and human-generated assessments	Explores ChatGPT's performance in assessments, comparing it to human grading and GTA recognition accuracy.	TurnItIn AI detector accuracy, human comparison	ChatGPT AI capabilities
(B. Liu vd., 2023)	Global information-aware argument mining based on a top-down multi-turn QA model	2023	Argumentative writing scoring	AbstRCT, SciARG, PE datasets	Proposes a multi-turn QA model for argument structure analysis in scientific papers and essays.	F1 score, ACC, ARI, ARC tasks	Multi-turn QA, NLP-based techniques
(Filho vd., 2019)	Imbalanced Learning Techniques for Improving Performance of Statistical Models in AES	2019	Imbalanced learning AES	Brazilian ENEM essays	Addresses class imbalance in AES using statistical learning and balancing methods to improve predictive accuracy.	Accuracy, average error	Imbalanced learning, statistical algorithms
(Vo vd., 2023)	Human scoring versus automated scoring for English learners in a statewide evidence-based writing assesment	2023	DIF and AES	Iowa Statewide Assessment (EL and non-EL test takers)	Investigates AES validity for English learners, comparing human and machine scores with DIF analysis.	Differential Item Functioning (DIF), effect size	Propensity score matching
(Ruipérez-Valiente vd., 2022)	Large-scale analytics of global and regional MOOC providers: Differences in learner' demographics, preferences, and perceptions	2022	MOOC-based AES	8M learners from 6,000 MOOCs	Analyzes demographic and regional impacts on MOOC outcomes, focusing on local vs global AES performance.	Demographic analytics, regional performance trends	Learning analytics, equity-based metrics
(Jansen vd., 2021)	Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language	2021	Teacher AES comparisons	Essays in English as a foreign language	Compares machine scores with teacher evaluations for English learners, focusing on diagnostic competence differences.	Teacher judgment accuracy	Machine and expert scoring

(Condon, 2013)	Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings?	2013	Commercial AES and local measures	Various datasets for large-scale testing	Discusses AES's limitations in construct representation and local measures as richer alternatives for student evaluation.	Validation metrics, construct representation analysis	Locally-developed and standardized AES tools
(Bernius vd., 2022)	Machine learning based feedback on textual student answers in large courses	2022	ML-based assessment feedback	Technical University of Munich dataset	Introduces CoFee framework for feedback on textual answers, reducing instructor grading workload significantly.	Precision, feedback acceptance rate	Topic modeling, NLP, hierarchical clustering
(Garner vd., 2019)	N-gram measures and L2 writing proficiency	2019	N-gram analysis for proficiency	Korean L1 English learner essays	Uses bigram and trigram indices to predict writing proficiency, emphasizing productive phraseological knowledge.	Correlation, regression analysis	N-gram modeling
(Vojak vd., 2011)	New Spaces and Old Places: An Analysis of Writing Assessment Software	2011	Holistic and diagnostic AES	Various automated and human assessment datasets	Critiques existing AES tools for narrow construct definitions and promotes multimodal approaches to writing assessment.	Holistic analysis of assessment frameworks	Intelligent Essay Assessor, Project Essay Grader
(Bejar vd., 2014)	On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration	2014	e-rater AES	GRE essays	Evaluates vulnerabilities in automated scoring systems, testing strategies like lexical alterations for construct validity.	Score comparison between altered and original essays	Lexical analysis, scoring engine simulation
(Deane, 2013)	On the relation between automated essay scoring and modern views of the writing construct	2013	AES	TOEFL, GMAT, GRE datasets	Explores AES alignment with writing constructs and cognitive-based assessments, addressing criticism of construct validity.	Correlation, construct analysis	Regression-based, construct-aligned scoring
(Y. Peng vd., 2023)	Predicting Chinese EFL Learners' Human-rated Writing Quality in Argumentative Writing Through Multidimensional Computational Indices of Lexical Complexity	2023	Computational indices-based AES	ICNALE (220 argumentative essays)	Uses lexical complexity measures to predict human-rated writing quality, validating dimensions like diversity and density.	Structural Equation Modeling (SEM), Confirmatory Factor Analysis (CFA)	Computational indices, Coh-Metrix
(Roscoe vd., 2017)	Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation	2017	AWE	Criterion, Writing Pal systems	Investigates students' perceptions of AWE tools, their impact on writing performance, and future adoption intentions.	User feedback analysis, improvement metrics	Human-computer interaction analysis
(Pino Castillo vd., 2023)	Profiling support in literacy development: Use of natural language processing to identify learning needs in higher education	2023	NLP-based literacy profiling	University students' writing samples	Diagnoses students' writing needs using NLP to identify text complexity and specialized language skills.	Factor analysis, complexity indices	Natural Language Processing (NLP), statistical models
(Humphry ve Heldsinger, 2019)	Raters' perceptions of assessment criteria relevance	2019	Rubric-based AES	Narrative writing samples	Examines how raters perceive criteria relevance in narrative writing, exploring implications for rubric development.	Rubric reliability, qualitative gradation	Rubric-based assessment

(Ahmed ve Sorour, 2024)	Classification-driven intelligent system for automated evaluation of higher education exam paper quality	2024	Evaluation of exam paper quality based on Bloom's taxonomy	Exam questions categorized by Bloom's taxonomy	Evaluates both the formal and content-based quality of exam papers. Ensures adherence to Bloom's taxonomy levels by classifying exam questions into knowledge, comprehension, and higher cognitive skills.	Average accuracy (98.5%)	Logistic Regression, CountVectorizer
(Alsanie vd., 2022)	Automatic scoring of Arabic essays over three linguistic levels	2022	Arabic essay scoring at lexical, syntactic, and semantic levels	Real dataset with Arabic language learners' essays	Focuses on scoring essays at three linguistic levels: lexical (word choice), syntactic (sentence structure), and semantic (meaning). Aimed at Arabic learners with beginner and intermediate proficiency.	Quadratic Weighted Kappa	Linear and Non-linear combination methods
(Ulum, 2020, s. 2)	A critical deconstruction of computer-based test application in Turkish State University	2020	Evaluation of AI-based Versant English Test in EFL settings	Turkish EFL students' and instructors' feedback on Versant Test	Analyzes the reliability and validity of the AI-driven Versant English Test. Highlights issues such as repetitive test questions, focus on memory over language skills, and misalignment with curriculum objectives.	Subjective feedback analysis (qualitative metrics)	No ML techniques used (qualitative analysis)
(R. Zhao vd., 2023)	AI-assisted automated scoring of picture-cued writing tasks for language assessment	2023	Grading of picture-cued writing tasks using AI	4,000 responses from 279 K-12 students	Assesses how students describe images in picture-cued writing tasks. Combines AI techniques like cross-modal matching to evaluate alignment between text and visuals, reducing subjectivity in grading.	Mean Absolute Error (0.479), Adjacent Agreement (90.64%)	Cross-modal matching, NLP algorithms
(Sharma ve Jayagopi, 2024)	Modeling essay grading with pre-trained BERT features	2024	Essay grading using BERT-based embeddings	ASAP-AES dataset (English essays)	Uses pre-trained BERT embeddings to capture essay structure, coherence, and topic relevance. Employs robust evaluation methods against adversarial samples, such as off-topic essays.	Quadratic Weighted Kappa (QWK = 0.81)	Pre-trained BERT embeddings, SVR, LSTM
(Schneider vd., 2023)	Towards Trustworthy AutoGrading of Short, Multi-lingual, Multi-type Answers	2023	Multi-lingual autograding for short answers	10M multi-lingual question-answer pairs	Addresses the challenges of grading short multi-lingual answers across diverse disciplines. Incorporates human validation for ambiguous cases and achieves trustworthiness through adaptive thresholds.	Accuracy (86.5%)	Fine-tuned Transformer models (BERT, LaBSE)
(Ruseti vd., 2024)	Automated Pipeline for Multi-lingual Automated Essay Scoring with ReaderBench	2024	Automated multi-lingual essay scoring with AutoML	English, Portuguese, and French multilingual datasets	Develops an AutoML pipeline for essay scoring across languages. Integrates linguistic feature analysis with Transformer models for consistent and scalable evaluation.	Comparison to state-of-the-art results	AutoML hybrid pipeline (Transformers + traditional ML)
(Myers ve Wilson, 2023)	Evaluating the Construct Validity of an Automated Writing Evaluation System with a Randomization Algorithm	2023	Validation of automated scoring with randomized essays	100 randomized persuasive essays (7th-8th graders)	Investigates the validity of MI Write's scoring system by randomizing essays at the sentence level. Evaluates traits like idea development and organization to ensure reliability.	Trait-based validity evaluation with randomized data	Automated Essay Scoring (AES), Randomization-based validation

(Morris vd., 2024)	Automated Scoring of Constructed Response Items in Math Assessment Using Large Language Models	2024	Large Language Model AES	NAEP Math Scoring Challenge	Explores using LLMs like DeBERTa for scoring math constructed responses with fine-tuning for specific item types.	Quadratic Weighted Kappa (QWK)	DeBERTa, Data Augmentation, Input Modification
(Martínez, 2024)	Re-evaluating GPT-4's Bar Exam Performance	2024	Legal Exam Analysis	Simulated Bar Exam Dataset	Challenges GPT-4's 90th percentile bar exam claim by analyzing percentile calculation methodologies and prompting techniques.	Percentile Distribution Analysis	GPT-4, Few-Shot Prompting, Chain-of-Thought Prompting
(Mardini G. vd., 2024)	A Deep-Learning-Based Grading System (ASAG) for Reading Comprehension Assessment by Using Aphorisms as Open-Answer-Questions	2023	Deep Learning ASAG	Aphorism Dataset (Spanish)	Develops ASAG for reading comprehension in Spanish using aphorisms with BERT and Skip-Thought embeddings.	Pearson Correlation, RMSE	BERT, Skip-Thought, Multilayer Perceptron
(Gaheen vd., 2021)	Automated Students Arabic Essay Scoring Using Trained Neural Network by E-Jaya Optimization	2021	Optimization-Based AES	Custom Arabic Essays Dataset	Introduces an optimized neural network (eJaya-NN) for scoring Arabic essays, improving personalized instruction systems.	Correlation (0.92), Accuracy	E-Jaya Optimization, Neural Networks
(Filighera vd., 2024)	Cheating Automatic Short Answer Grading with the Adversarial Usage of Adjectives and Adverbs	2024	Short answer grading with adversarial inputs	SciEntsBank dataset	Tests robustness of short answer grading models against adversarial attacks by inserting adjectives and adverbs.	Accuracy drop analysis (10–22% loss in adversarial settings)	BERT, T5 adversarial robustness testing
(Conijn vd., 2022)	Early Prediction of Writing Quality Using Keystroke Logging	2022	Writing quality prediction during process	126 English academic summarizations with keystrokes	Analyzes keystroke data for early prediction of writing quality in English learners during a timed summarization task.	AUC (highest = 0.57)	Regression and classification models on keystroke features
(Ahmed ve Sorour, 2024)	A Novel Automated Essay Scoring Approach for Reliable Higher Educational Assessments	2021	Essay scoring in higher education	Kaggle's ASAP dataset	Transformer-based scoring model using Bi-LSTM and RoBERTa for essay coherence in higher education.	QWK score comparison with human raters	Bi-LSTM over RoBERTa embeddings
(Beseiso vd., 2021, s. 202)	Feedback Sources in Essay Writing: Peer-Generated or AI-Generated Feedback?	2024	Comparison of AI and peer feedback in essay writing	74 graduate students' essays	Examines AI-generated and peer feedback on essays, finding complementary strengths in descriptive and problem-focused feedback.	MANOVA and Spearman correlation	Descriptive and inferential analysis with ChatGPT
(Banihashem vd., 2024)	Applying Natural Language Processing and Hierarchical Machine Learning Approaches to Text Difficulty Classification	2020	NLP-based text difficulty classification	iSTART library text datasets	Combines NLP indices and hierarchical machine learning to classify text difficulty, improving accuracy over traditional metrics by 10%.	Accuracy improvement over traditional readability metrics	Hierarchical machine learning and NLP models
(Balyan vd., 2020)	Automatic creation of Human-Competitive Programs and Controllers by Means of Genetic Programming	2000	Designing competitive programs using genetic methods	Simulated genetic program datasets	Develops competitive programs for complex tasks using evolutionary algorithms.	Performance comparison against human benchmarks	Genetic programming algorithms

(Koza vd., 2020)	Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features	2018	Scoring non-native essays using linguistic features	TOEFL, FCE datasets	Analyzes lexical, syntactic, and discourse-level linguistic features to predict essay scores for non-native learners.	Pearson correlation and prediction accuracy	Regression modeling
(Perin ve Lauterbach, 2018)	Assessing Text-Based Writing of Low-Skilled College Students	2018	Automated scoring of developme ntal college writing	Writing samples scored with Coh-Metrix	Uses linguistic features to evaluate writing quality of low-skilled college students.	Correlation analysis with human ratings	Discrimina nt analysis with linguistic variables
(Rahimi vd., 2017)	Assessing Students' Use of Evidence and Organization in Response-to-Text Writing: Using NLP for Rubric-Based Automated Scoring	2017	Rubric-based scoring for evidence and organizatio n	RTA dataset (grades 5-8)	Scores essays on evidence and organizational quality using source material and specific rubrics.	Cross-validation and rubric-specific performance metrics	NLP-based feature extraction for task-dependent models
(Husni vd., 2022)	A Bi-modal Automated Essay Scoring System for Handwritten Essays	2021	Bi-modal AES	ASAP, ChnStd	Proposes VisualAES, integrating text and handwritten image features to address OCR errors and improve scoring accuracy.	Accuracy, OCR Error Reduction	Faster R-CNN for image features, Transforme r-based models (Visual-BERT, Visual-RoBERTa, Visual-ALBERT), Stacking Ensemble
(Uto vd., 2023)	Integration of Prediction Scores from Various Automated Essay Scoring Models Using Item Response Theory	2023	Score Integration AES	Benchmark AES dataset	Combines predictions from multiple AES models using Item Response Theory for enhanced interpretability and accuracy.	Scoring Accuracy, Interpretabilit y	Item Response Theory, Hybrid Regression Models
(Sethi ve Singh, 2022)	NLP-based Automated Essay Scoring with Parameter-Efficient Transformer Approach	2022	Transforme r-based AES	ASAP	Leverages pre-trained language models with adapter modules to reduce trainable parameters while boosting performance.	Quadratic Weighted Kappa (QWK)	Transforme r models (BERT with Adapters), BiDAF (Bidirection al Attention Flow)
(Catulay vd., 2021)	Neural-Network Architecture Approach: An Automated Essay Scoring Using Bayesian Linear Ridge Regression Algorithm	2021	Neural Network AES	ASAP	Utilizes neural networks and Bayesian ridge regression to enhance the correctness and reliability of automated scoring.	Mean Absolute Error (MAE), Accuracy	Deep Neural Networks, Bayesian Ridge Regression
(Tashu, 2020)	Off-Topic Essay Detection Using C-BGRU Siamese	2020	Off-Topic Detection	Kaggle AEE datasets	Implements a Siamese network with convolutional and bidirectional GRU layers for off-topic essay detection.	Accuracy, False Positive Rate	C-BGRU Siamese Architectur e, Deep Learning

(Ajetunmobi ve Daramola, 2017)	Ontology-Based Information Extraction for Subject-Focussed Automatic Essay Evaluation	2017	Subject-Focused AES	Domain-specific Ontology	Employs ontology-based semantic matching to provide feedback and domain-specific grading for essays.	Usability Rating (4.17/5), Correlation	Ontology Development, Semantic Information Extraction
(Idhom vd., 2022)	Performance Evaluation of Automated Essay Scoring Online System for Competency Assessment of Community Academy	2022	Competency Assessment AES	Community Academy data	Evaluates AES performance using cosine similarity and human correlation for vocational competency assessments.	Correlation (0.90), MAE (0.13)	Cosine Similarity, Statistical Evaluation
(M. Chen ve Li, 2018)	Relevance-Based Automated Essay Scoring via Hierarchical Recurrent Model	2018	Relevance-based AES	ASAP	Proposes using essay-topic relevance as an auxiliary feature in hierarchical Bi-LSTM models for improved scoring.	Accuracy, Relevance Correlation	Bi-LSTM, Attention Mechanisms, Sentence- and Document-Level Encoding
(Z. Chen ve Zhou, 2019)	Research on Automatic Essay Scoring of Composition Based on CNN and OR	2019	Deep Learning AES	Not specified	Combines CNN and Ordinal Regression to improve the fairness and accuracy of scoring essays.	Accuracy, Mean Squared Error	Convolutional Neural Networks (CNN), Ordinal Regression
(Xue vd., 2022)	Robust Automated Essay Scoring by Using Attentive Capsule	2022	Capsule Network AES	ASAP	Employs a capsule network with attention mechanisms for robust essay representation and scoring.	State-of-the-Art Performance (ASAP)	Capsule Networks, Attention Mechanisms
(Ejima ve Takeuchi, 2022)	Statistical Learning Models for Japanese Essay Scoring Toward One-shot Learning	2022	One-shot Learning AES	GoodWriting (Japanese dataset)	Uses bag-of-words and BERT embeddings to evaluate essays with minimal training data using one-shot learning.	F1-Score, Regression Coefficients	Bag-of-Words Encoding, BERT, Support Vector Regression (SVR)
(Qiu vd., 2022)	Tapping the Potential of Coherence and Syntactic Features in Neural Models for AES	2022	Coherence and Syntactic AES	Custom dataset	Combines coherence and syntactic embeddings with BERT for enhanced scoring, especially for long essays.	Performance Metrics, Statistical Analysis	Coherence Modeling, Syntactic Embeddings, BERT
(Nakamoto vd., 2024)	Towards Human-Level Evaluation: Assessing the Potential of GPT-4 in Automated Evaluation and Feedback Generation on Japanese Essays	2024	Generative AI AES	Shonan Seminar Japanese Essays Dataset	Examines GPT-4's capability to score Japanese essays and generate feedback based on a rubric, comparing with human evaluators. Achieved high agreement with humans, but challenges like hallucination and slight scoring bias were noted.	RMSE (0.449–0.967), Concordance Rate (91%)	GPT-4, Reinforcement Learning from Human Feedback (RLHF), Prompt Engineering
(X. Li vd., 2017)	Unsupervised Off-topic Essay Detection Based on Target and Reference Prompts	2017	Off-Topic Detection	Kaggle AEE datasets	Proposes an unsupervised system using semantic differences between target and reference prompts for detection.	Accuracy, Semantic Similarity	Semantic Similarity Metrics, Reference Prompts
(Jin ve He, 2015)	Utilizing Latent Semantic Word Representations for Automated Essay Scoring	2015	Semantic Feature AES	ASAP	Explores latent semantic features through CBOW and Recursive Autoencoder to align scores with human ratings.	Root Mean Squared Error, Pearson's Correlation	Continuous Bag-of-Words (CBOW), Recursive

							Autoencoder
(Husni vd., 2022, s. 20)	Word Ambiguity Identification using POS Tagging in Automatic Essay Scoring	2022	NLP-based AES	Custom dataset (Indonesian)	Focuses on handling word ambiguity in Indonesian essays using POS tagging and cosine similarity.	Accuracy (79%)	Part-of-Speech Tagging, Cosine Similarity
(Das vd., 2022)	FACToGRADE: Automated Essay Scoring System	2022	Hybrid AES	Custom dataset	Integrates LSTM and entity detection for scoring essays while providing feedback and structural analysis.	Feedback Analysis, Accuracy	Long Short Term Memory (LSTM), Entity Detection
(Contreras vd., 2021)	Essay Question Generator based on Bloom's Taxonomy for Assessing AES System	2021	Question Construction	Custom dataset	Implements Bloom's taxonomy in essay question generation and evaluates AES effectiveness for different cognitive levels.	F1-Measure, Frequency Distribution	Support Vector Machine (SVM), Natural Language Processing (NLP)
(J. Huang vd., 2023)	Enhancing Essay Scoring with Adversarial Weights Perturbation and Metric-specific Attention Pooling	2023	Advanced AES for ELLs	ELL-focused datasets	Explores DeBERTa and adversarial training with metric-specific attention pooling for scoring English language learners' essays.	Adversarial Learning Impact	DeBERTa, Adversarial Training, Attention Pooling
(Sendra vd., 2016)	Enhanced Latent Semantic Analysis by Considering Mistyped Words in Automated Essay Scoring	2016	Feature Enhanced AES	119 manually graded essays	Proposes enhanced LSA that accounts for mistyped words and compares its performance to standard LSA and GLSA.	Closeness to Human Judgment	Enhanced LSA, Singular Value Decomposition (SVD)
(Wu, 2023)	English Composition Assisted Writing Training Based on Big Data Text Mining	2023	Big Data AES	English writing datasets	Develops an AES model using big data text mining techniques, achieving high accuracy and aiding English composition training.	Accuracy (93%)	Big Data Text Mining, Random Forest, SVM
(Riyanto vd., 2024)	Employing SBERT for Essay Assessment	2024	Semantic Similarity AES	Indonesian school essays	Utilizes SBERT for essay assessment, achieving low mean absolute error and high correlation with human evaluators.	Mean Absolute Error (0.26)	SBERT, Dependency Parsing, Part-of-Speech Tagging
(Johnsi ve Kumar, 2024)	Digital Essay Assessment using Machine Learning Algorithms	2024	Regression-based AES	Varying datasets	Implements multiple regression techniques, with XGBoost achieving the best results for essay scoring tasks.	Quadratic Weighted Kappa (0.68)	XGBoost, Elastic Net Regression, Gradient Boosting
(Lahitani vd., 2016)	Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment	2021	Similarity-based AES	Indonesian essay datasets	Applies TF-IDF and cosine similarity to assess essays, focusing on similarity with reference texts.	Similarity Ranking Accuracy	Cosine Similarity, TF-IDF
(Mim vd., 2021, s. 202)	Corruption Is Not All Bad: Incorporating Discourse Structure Into Pre-Training via Corruption for Essay Scoring	2021	Discourse AES	Custom discourse-annotated datasets	Introduces discourse corruption pre-training to improve coherence and cohesion metrics in essay evaluation.	Essay Organization Score	Longformer, Discourse Corruption Pre-Training

(Syed Abuthahir vd., 2024)	Building a Deep Neural Network for Automated Essay Scoring in English Language Teaching	2024	Explainable AES	Curated English education datasets	Explores explainable AI techniques for AES, achieving high accuracy through DNN and SHAP for interpretability.	Accuracy (92.5%)	Deep Neural Networks, SHAP Interpretability Models
(Mahmoud vd., 2024)	Automatic Scoring of Arabic Essays: A Parameter-Efficient Approach for Grammatical Assessment	2024	Arabic AES	QALB-2014, QALB-2015, ZAEBUC datasets	Employs AraBART with parameter-efficient methods for scoring essays with a focus on grammar correction.	F-scores, Precision, Recall	AraBART, Parameter-Efficient Fine-Tuning
(Alqahtani ve Alsaif, 2019)	Automatic Evaluation for Arabic Essays: A Rule-Based System	2024	Rule-based AES	Custom Arabic essay dataset	Proposes a baseline rule-based system for scoring Arabic essays based on structural and surface-level features.	Correctness Percentage (73%)	Rule-Based System
(Arifin vd., 2021, s. 202)	Automatic Essay Scoring for Indonesian Short Answers using Siamese Manhattan Long Short-Term Memory	2021	Siamese LSTM AES	SIMPLE-O Dataset	Develops a model for short answer grading using Siamese Manhattan LSTM, comparing answers for semantic similarity.	Training Accuracy (92.82%), Validation Accuracy (84.03%)	Siamese Manhattan Long Short-Term Memory (LSTM), Word2Vec
(Chang ve Lee, 2009)	Automatic Chinese Essay Scoring Using Connections between Concepts in Paragraphs	2009	Conceptual Structure AES	Custom Chinese Essays Dataset	Uses semantic associations in paragraph structures to evaluate Chinese essays.	Conceptual Similarity Metrics	Paragraphic Conceptual Structure Analysis
(Thongyoo vd., 2016)	Automated Thai Online Assignment Scoring	2016	Text Mining AES	Thai Assignment Dataset	Implements text mining with clustering and classification for scoring Thai assignments.	Accuracy (86.51%), Correlation Coefficient (R=0.60)	K-Means Clustering, SVM
(Sharma ve Jayagopi, 2018)	Automated Grading of Handwritten Essays	2018	Handwriting Recognition AES	Hewlett Foundation AES Dataset	Combines Optical Handwriting Recognition (OHR) and GloVe embeddings for grading handwritten essays.	QWK Score (0.88)	MDLSTM, GloVe Word Embeddings
(Contreras vd., 2018)	Automated Essay Scoring with Ontology based on Text Mining and NLTK tools	2019	Ontology-Based AES	Essays on Human-Computer Interaction	Employs ontology generation and NLP for grading essays with domain-specific ontologies.	Ontology Accuracy	OntoGen, NLTK, Linear Regression
(Chang vd., 2008)	Automated Essay Scoring Using Set of Literary Sememes	2008	Semantic Feature AES	Chinese Literary Essays Dataset	Scores essays using sememes extracted from literary concepts for advanced semantic analysis.	Sememe Frequency Accuracy	Literary Sememes, HowNet
(Gunawansyah vd., 2020)	Automated Essay Scoring Using Natural Language Processing And Text Mining Method	2020	NLP and Text Mining AES	Custom Dataset	Uses preprocessing techniques like tokenization and stemming for effective AES, focusing on semantic similarity.	Similarity Accuracy	NLP, Text Mining, Synonym Matching
(Ajit Tambe ve Kulkarni, 2022)	Automated Essay Scoring System with Grammar Score Analysis	2022	Grammar-Focused AES	ASAP Dataset	Separates scoring into structure and grammar, leveraging BERT for contextual embeddings and grammar error detection.	QWK Score (Structure=0.75, Grammar=0.70)	BERT, Sequential Model
(Eid ve Wanas, 2017)	Automated Essay Scoring Linguistic Feature: Comparative Study	2017	Feature-Based AES	Custom (Lexical Feature Dataset)	Evaluates 22 linguistic features for essay scoring, combining lexical and surface-level features with regression models.	Feature Elimination Impact, Correlation	Linear Regression, Feature Engineering

(Nair ve Naga Malleswari, 2024)	Automated Essay Scoring for IELTS using Classification and Regression Models	2024	Classification ve Regression AES	Custom (IELTS Essays Dataset)	Implements classification and regression models for IELTS essay grading with linguistic and structural analysis.	Cohen's Kappa, Accuracy	Random Forest, Logistic Regression, Support Vector Machines
(Zainal ve Abu Hassan, 2022)	Automated Essay Scoring Using English Essay Question	2022	NLP-Based AES	WordNet Corpus, Custom Dataset	Combines spelling checking, semantic analysis, and LSA for evaluating English essay responses.	Correlation with Human Scores (70%)	Latent Semantic Analysis (LSA), Thesaurus Matching
(Yang, 2023)	Automated English Essay Scoring Based on Machine Learning Algorithms	2023	Machine Learning AES	Kaggle Dataset	Compares handcrafted and polynomial features with various regressors for essay scoring accuracy improvement.	MSE, MAPE, Feature Importance	XGBoost, Ridge Regression, Polynomial Features
(X. Peng vd., 2010)	Automated Chinese Essay Scoring Using Vector Space Models	2010	Vector Space Model AES	Chinese MHK Dataset	Compares four VSM approaches and introduces statistical surface features for scoring Chinese essays.	Human Correlation	Weighted VSM, Sequence VSM, SVR
(Obata vd., 2023)	Assessment of ChatGPT's Validity in Scoring Essays by Foreign Language Learners of Japanese and English	2023	AI-Powered AES	Custom (Japanese ve English Essays Dataset)	Evaluates ChatGPT's scoring performance compared to human evaluators and linguistic feature-based models.	GPT-Human Correlation	ChatGPT, Lexical ve Syntactic Feature Analysis
(Nyomi ve Moccozet, 2022)	Anatomy of a Large-Scale Real-Time Peer Evaluation System	2022	Peer Evaluation AES	Custom Open-Ended Questions Dataset	Uses NLP and clustering for semantic grouping of peer evaluations in large-scale educational settings.	Teacher Feedback Correlation	NLP Clustering, Peer Review Algorithms
(Chassab vd., 2024)	An Optimized LSTM-Based Augmented Language Model (FLSTM-ALM) Using Fox Algorithm for Automatic Essay Scoring Prediction	2024	Augmented LSTM AES	ASAP, ETS Datasets	Proposes FLSTM with Fox algorithm for feature selection and dimensionality reduction, achieving high accuracy.	Accuracy (98.92%), QWK, F1-Score	LSTM, Fox Optimization Algorithm, Neural Knowledge Retriever
(C. Li vd., 2022)	An Automated Essay Scoring Model Based on Stacking Method	2022	Ensemble Learning AES	ASAP Dataset	Uses ensemble learning with stacking method to combine multiple models, improving QWK values across subsets.	QWK Improvement (1%-2.8%)	Stacking Method, Neural Networks, Feature Engineering
(Rababah ve Al-Taani, 2017)	An Automated Scoring Approach for Arabic Short Answer Essay Questions	2017	Short Answer AES	Custom Arabic Dataset	Uses cosine similarity and word root analysis for Arabic short answers scoring with high accuracy.	Cosine Similarity	NLP, Cosine Similarity, Semantic Feature Matching
(Suresh vd., 2023)	AI-based Automated Essay Grading System using NLP	2023	Graph-based AES	Custom Dataset	Implements graph-based sentence similarity and NLP methods to evaluate essay structure, grammar, and content.	Accuracy, Sentence Similarity	Graph-based Features, NLP, Neural Networks
(Reddy Chavva vd., 2024)	A Transformer-Based Approach for Enhancing Automated Essay Scoring	2024	Transformer-Based AES	ASAP Dataset	Proposes a RoBERTa-based model fine-tuned for automated essay scoring with high QWK scores.	QWK (0.815)	Transformer Models (RoBERTa)

(Nagata vd., 2009)	A Topic-independent Method for Automatically Scoring Essay Content Rivaling Topic-dependent Methods	2009	Topic-Independent AES	Custom Essays Dataset	Introduces MIDF for topic-independent essay scoring, achieving competitive accuracy compared to topic-dependent methods.	Accuracy (0.848)	Mutual Information, MIDF-based Classification
(H. Chen vd., 2012)	A Ranked-based Learning Approach to Automated Essay Scoring	2012	Learning-to-Rank AES	GRE, GMAT, TOEFL Essays Dataset	Applies learning-to-rank algorithms for AES, improving performance compared to traditional regression models.	Ranking Precision	Learning-to-Rank, Feature Engineering
(Bojamma vd., 2024)	A Lexical Model for Automated Essay Evaluation System Using Ensemble Learning Techniques	2024	Ensemble Learning AES	Kaggle Essay Dataset	Uses Random Forest Classifier for essay scoring based on lexical and grammatical features.	QWK, Feature Importance	Random Forest, NLP
(Bhatt vd., 2020)	A Graph-Based Approach to Automate Essay Evaluation	2020	Graph-Based AES	Kaggle's ASAP Dataset	Evaluates semantic and syntactic consistency with graph-based models and statistical features.	Correlation with Human Scores	Random Forest, Semantic Graph Analysis
(Perera vd., 2015)	A Dynamic Semantic Space Modelling Approach for Short Essay Grading	2015	Semantic Space AES	CRESST, CSE Dataset	Uses VSM and NLP techniques for model answer-based scoring, emphasizing conceptual accuracy.	Human-Computer Correlation	Latent Semantic Analysis (LSA), VSM
(Suriyasat vd., 2023)	A Comparison of Machine Learning and Neural Network Algorithms for an Automated Thai Essay Scoring	2023	Thai Essay AES	DifferSheet Dataset	Compares ML and DL techniques, finding XGBoost and LSTM+CNN models perform best for Thai essay scoring.	F1-Score, Accuracy	XGBoost, LSTM+CNN, BERT
(Wangkriangkri vd., 2020)	A Comparative Study of Pretrained Language Models for Automated Essay Scoring with Adversarial Inputs	2020	Adversarial-Resistant AES	ASAP Dataset	Evaluates GloVe, ELMo, and BERT models for adversarial robustness and accuracy in AES tasks.	Robustness, QWK	GloVe, ELMo, BERT
(Stephen vd., 2021)	Exploring new insights into the role of cohesive devices in written academic genres	2021	Computational linguistic AES	narrative and argumentative student essays	Analyzes cohesive device use in academic writing, linking cohesion features to essay quality across genres	Regression models, human ratings	TAACO computational analysis
(Vajjala, 2018)	Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features	2018	Scoring non-native essays using linguistic features	TOEFL, FCE datasets	Analyzes lexical, syntactic, and discourse-level linguistic features to predict essay scores for non-native learners.	Pearson correlation and prediction accuracy	Regression modeling

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı **:** Çağrışım Helhel Bayraktaroğlu

Doğum Yeri ve Tarihi **:**

Eğitim Durumu

Lisans Öğrenimi **:** Boğaziçi Üniversitesi, Eğitim Fakültesi, Fizik Öğretmenliği Bölümü.

Yüksek Lisans Öğrenimi **:** Akdeniz Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitimde Ölçme ve Değerlendirme Tezli Yüksek Programı.

Bildiği Yabancı Diller **:** İngilizce

İletişim

E-Posta Adresi **:**

Tarih **:** 27/01/2025

BİLDİRİM

Hazırladığım tezin/raporun tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kâğıt ve elektronik kopyalarının Akdeniz Üniversitesi Eğitim Bilimleri Enstitüsü arşivlerinde aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım:

☐ Tezimin/Raporumun tamamı her yerden erişime açılabilir.

☐ Tezim/Raporum sadece Akdeniz Üniversitesi yerleşkelerinden erişime açılabilir. Tezimin/Raporumun **3** yıl süreyle erişime açılmasını istemiyorum. Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir.

27/01/2025

Çağrışım Helhel Bayraktaroğlu

18% Genel Benzerlik

Her veri tabanı için çakışan kaynaklar da dâhil tüm eşleşmelerin kombine toplamı.

Rapordan Filtrelenen

- Bibliyografya
- Alıntılanan Metin

Ön Sıradaki Kaynaklar

13% İnternet kaynakları
15% Yayınlar
7% Gönderilen çalışmalar (Öğrenci Makaleleri)

Bütünlük Bayrakları

İnceleme için 1 Bütünlük Bayrağı

Gizli Metin
1 sayfada 182 şüpheli karakter
Metin, belgenin beyaz arka planına karıştırılmak üzere değiştirilir.

Sistemimizin algoritmaları bir belgede, onu normal bir gönderiden ayırabilecek her türlü tutarsızlığı derinlemesine inceler. Tuhaf bir şey fark edersek incelemeniz için bayrak ekleriz.

Bir Bayrak mutlaka bir sorun olduğunu göstermez. Ancak daha fazla inceleme için dikkatinizi vermenizi öneririz.