www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 9, No: 4, 2022 (1484-1494)

El-Cezerî Journal of Science and Engineering
Vol: 9, No: 4, 2022 (1484-1494)
DOI : 10.31202/ecjse.1135868

ECJSE

Makale / Research Paper

Öznitelik Seçiminde Genetik Algoritma Kullanılarak Kur'an-ı Kerim Ayetlerinin Otomatik Sınıflandırılması

Fatih MERT[1,2a], Muhammed Ali AYDIN[1b], Zeynep ORMAN[1c]

[1]İstanbul Üniversitesi-Cerrahpaşa, Bilgisayar Mühendisliği Ana Bilim Dalı, İstanbul/TÜRKİYE
[2]Huawei Türkiye Ar-Ge Merkezi, İstanbul/TÜRKİYE
fatih.mert@ogr.iuc.edu.tr, aydinali@iuc.edu.tr, ormanz@iuc.edu.tr

Received/Geliş: 25.06.2022 **Accepted/Kabul:** 16.11.2022

Öz: Metin etiketleme olarak da bilinen metin sınıflandırması verilen bir metni organize gruplara ayırma işlemidir. Metin sınıflandırıcılar, Doğal Dil İşleme yöntemlerini kullanarak metni otomatik olarak analiz edebilir ve ardından içeriğine göre bir dizi önceden tanımlanmış etiket veya kategori ataması yapabilir. Söz konusu bir Kur'an ayeti ise, etiketlenmedeki temel amaç ayetin ilgili olduğu temanın belirlenmesidir. Ancak mevcuttaki ayet etiketleme yaklaşımları öncelikli olarak Arapça dilinde ve Kur'an tefsirinde derin uzmanlığa sahip alimlerin mevcudiyetine bağlıdır. Bu çalışmada metin sınıflandırma algoritmalarını kullanarak Kur'an ayetlerinin etiketlenmesi görevinin otomatikleştirilmesi önerilmektedir. Sınıflandırma algoritmaları ile gerçekleştirdiğimiz deneylerde ayetlerin İngilizce çevirilerinin ait oldukları önceden tanımlanmış 15 kategori öznitelik olarak kullanılmıştır. Literatürdeki benzer çalışmalardan farklı olarak öznitelik seçimi aşamasında diğerlerine kıyasla daha yüksek performans gösteren Genetik Algoritma tercih edilmiştir. Böylece gerçekleştirilen bu ara adımın nihai performansa olumlu etki etmesi amaçlanmıştır. Çalışmanın sonunda çeşitli performans değerlendirme metrikleri kullanılarak sınıflandırma modellerinin başarım değerleri karşılaştırılmalı olarak verilmiştir.

Anahtar Kelimeler: Kur'an, Metin Sınıflandırma, Genetik Algoritma, Makine Öğrenmesi, Öznitelik Seçimi

Automatic Classification of Quran Verses Using Genetic Algorithm for Feature Selection

Abstract: Text classification, also known as text tagging, is the process of dividing a given text into organized groups. Using Natural Language Processing methods, text classifiers can automatically analyze text and then assign a set of predefined tags or categories based on its content. If it is a verse of the Holy Qur'an, the main purpose of labeling is to determine the theme of the verse. However, current approaches to verse tagging depend primarily on the availability of scholars with deep expertise in the Arabic language and Qur'anic exegesis. In this study, it is suggested to automate the task of tagging Qur'anic verses using text classification algorithms. In the experiments we carried out with the classification algorithms, the 15 predefined categories to which the English translations of the verses belong were used as features. Unlike similar studies in the literature, Genetic Algorithm which outperforms the other methods was used in the feature selection stage. Thus, it is aimed that this intermediate step will have a positive effect on the final performance. At the end of the study, the performance values of the classification models are given comparatively by using various performance evaluation metrics.

Keywords: Quran, Text Classification, Genetic Algorithm, Machine Learning, Feature Selection.

Bu makaleye atıf yapmak için
Mert, F., Aydin, M.A, Orman, Z., "*Öznitelik Seçiminde Genetik Algoritma Kullanılarak Kur'an-ı Kerim Ayetlerinin Otomatik Sınıflandırılması*" El-Cezerî Journal of Science and Engineering, 2022, 9(4); 1484-1494.

How to cite this article
Mert, F., Aydin, M.A, Orman, Z., "*Automatic Classification of Quran Verses Using Genetic Algorithm for Feature Selection*" El-Cezerî Journal of Science and Engineering, 2022, 9(4); 1484-1494.

ORCID ID: [a]0000-0002-2896-5475; [b]0000-0002-1846-6090; [c]0000-0002-0205-4198

1. Giriş

Çevrimiçi ya da çevrimdışı olması fark etmeksizin metinsel verilerin varlığı bizleri bilginin aşırı yüklenmesi sorunuyla karşı karşıya getirmektedir. Bu nedenle, otomatik metin sınıflandırma çalışmaları yapay zeka ile ilgili pek çok çalışma alanından araştırmacıları kendine çekmiştir. Metin sınıflandırma problemi etiketlenmemiş metinlere ya da belgelere önceden tanımlanmış sınıf üyeliğinin otomatik olarak atanması olarak ifade edilmektedir ve haberlerin kategorize edilmesi, web aramaları, tıbbi döküman tasnifi, duygu analizi ve e-posta filtrelenmesi gibi pek çok konuda bilginin aşırı yüklenmesi ile ilgili sorunun ortadan kaldırılması için metin sınıflandırma yöntemleri kullanılmaktadır. Bahse konu problem, Kur'an-ı Kerim'in içeriğini anlama hususunda da mevcuttur. Kutsal kitapların sonuncusu olan Kur'an, Müslümanlar tarafından yüce Allah'ın insanlığa mesajlarını ilettiğine inandıkları ve sure ve ayetlerden oluşan kapsamlı bir metindir. Her ne kadar indirildiği dönemin ve toplumun yapısı itibarıyla Arapça olarak ortaya çıkmış olsa da günümüzde İslam dininin dünya çapında yüzyıllar boyunca oldukça yaygınlaşması hasebiyle dini alimler, müfessirler, Kur'an ile ilgilenen aydınlar, akademisyenler, uzmanlar ve Kur'an okuyucuları tarafından pek çok dilde yoğun bir şekilde çevirileriyle karşılık bulmuştur. Aynı zamanda bugüne kadar aynı dildeki çevirileri üzerinden dahi pek çok kısmında farklı yorumları beraberinde getirmiştir ve getirmeye devam etmektedir. Bu gibi karışıkların ortadan kalkması için Kur'an diline ve ayetlerle surelerin indirildiği bağlama ciddi manada hakim uzmanlarla çalışılması gerekmektedir.

Kur'an; cüz denilen toplam 30 ana bölümden ve sure denilen 114 alt bölümden müteşekkildir [1]. Her bir sureyi ayet adı verilen metinler oluşturmaktadır. Toplamdaki ayet sayısı ise 6236'dır. Bu çalışma, temelde metin sınıflandırma yöntemlerinin Kur'an-ı Kerim üzerinde uygulanması esasına dayanmaktadır. Daha önceden tanımlanmış 15 sınıf etiketi kullanılarak "Mekki" ve "Medeni" ayetler diye tabir edebileceğimiz kategorinin verilen bir ayete makine öğrenmesi algoritmaları ve genetik algoritma kullanılarak otomatik olarak atanması niyetlenmiştir. Mekke'de indirilen ayetler Mekki, Medine'de indirilen ayetler ise Medeni ayet olarak adlandırılmaktadır. Mekki ayetlerin indirilme dönemi son Peygamber Hz. Muhammed (S.A.V)'e ilk vahyin geldiği andan Medine'ye hicret edildiği ana kadarki dönemi kapsamakta olup yaklaşık 13 yıla tekabül etmektedir. Kur'an'ın yaklaşık 2/3'ü bu zaman aralığında nazil olmuştur. Medeni ayetlerin indirilme dönemi ise Hz. Muhammed'in Medine'ye gelişinden ölümüne kadar olan süreyi anlatır. Ayetlerin Mekki mi yoksa Medeni mi olduğunun ayrımının yapılması, Müslüman toplumunun ilerleme tarihini anlamak, İslam hukukunun gelişimini ve Peygamberin farklı gayrimüslim gruplarıyla nasıl ilgilendiğini görmek açısından önemlidir ve ayetleri doğru şekilde anlamamız ve onun rehberliğinden doğru şekilde istifade edilebilmesi noktasında bizlere yardımcı olmaktadır. Bunların yanısıra ayetlerin veya surelerin Mekki ve Medeni olarak sınıflandırılmasında bir diğer ayırt edici özellik ise ihtiva ettikleri temalardır. Mekki ayet veya sureler genel olarak tevhid, gayb inancı, geçmiş peygamlerin hayatlarından alıntılanmış hikayeler ve cennet ile cehennem gibi konulara yoğun şekilde odaklanmışken; Medeni ayet veya sureler daha çok yönetmelikler, kurallar, hicret, sosyal hayat ve toplumsal düzenlemelere ilişkin detayları anlatır.

2. Literatür Özeti

Kur'an-ı Kerim sadece dil bilimcileri açısından değil bilgisayar bilimleri alanında da pek çok çalışmaya konu olmuştur [2-5]. Bunların arasında Kur'an ayetlerinin sınıflandırılmasına ilişkin çalışmalar da yer almaktadır. Karar Ağaçlarını kullanan [6], Yapay Sinir Ağları'ndan istifade eden [7], ve K-En Yakın Komşuluk yöntemini deneyen [8] bunlara örnek olarak gösterilebilir. Benzer şekilde [9]'da SVM sınıflandırma modeli kullanılmıştır. Ak-Kabi vd. [10] tarafından Arapça veri seti (Modern Standard Arabic – MSA) kullanılarak yapılan çalışma Kur'an ayetlerinin 14 farklı sınıfa tasnif edilmesi esasına dayanmış olup bu çalışmada Naïve Bayes'in diğer sınıflandırma modellerine göre daha iyi sonuçlar verdiği gözlenmiştir.

Öznitelik seçimi veri ön işlemesinde en sık kullanılan ve en önemli aşamalardan birisidir. Bu sebeple makine öğrenmesi ile ilgili yürütülen pek çok çalışma için ayrılmaz bir parça halini almıştır. Özünde bir boyut azaltma yöntemi olup orijinal bir öznitelik kümesinden bir kısmının seçilmesi ve böylece belirli bir değerlendirme kriterine göre öznitelik uzayının en optimum şekilde azaltılması amaçlanmaktadır. [11] ve [12]'te öznitelik seçimiyle ilgili kapsamlı bir derleme ve deneysel çalışma örnekleri bulunmaktadır. Normalde ihtiyaç olunan en uygun öznitelikleri ihtiva eden öznitelik alt kümesini bulma işlemi 2^N muhtemel alt kümenin oluşturulması gerektiğinden oldukça maliyetlidir. Bu sebeple bizim çalışmamızda Genetik Algoritma kullanımı oldukça makul görünmektedir.

3. Metot

3.1. Veri Seti

Bu çalışma kapsamındaki çalışmaların gerçekleştirilmesi için önceden etiketlenmiş Kur'an ayetlerine ihtiyaç bulunmaktaydı ancak literatür taraması yaparken Türkçe ya da İngilizce etiketlenmiş veri setine rastlanmamıştır. Araştırmalar neticesinde [13] sitesinde Endonezce diline çevirisi yapılan Kur'an ayetlerinin 15 adet kategoriye göre tasnif edilerek etiketlendiği görülmüştür. Bu veri seti baz alınarak İngilizce Kur'an ayetleri Mekki veya Medeni olma kategorisi de eklenerek yeni bir veri setine dönüştürülmüştür. Bu işlem sonucunda bir ayetin Mekki ya da Medeni olması üzerindeki etkilerini gözlemlemeye çalıştığımız öznitelikler şu şekildedir: "pillars_of_islam", "faith", "al_quran", "science", "actions", "call_to_God", "jihad", "human_and_social_relations", "morals", "property_regulations", "legal_matters", "state_and_society", "agriculture_and_trade", "history_and_stories", "religion". Aşağıdaki tabloda yer alan son sütun, ayetin herhangi bir sınıfa ait olup olmadığı durumunu göstermektedir. Eğer bir ayet önceki sınıflardan hiçbirine ait değilse 1, aksi halde 0 olarak işaretlenmiştir.

Tablo 1. Kur'an ayetleri veri setinden örneklem

Sure Adı	Mekki Medeni	İngilizce Ayet	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Al-Fatiha	0	In the name of Allah, Most Gracious, Most Merciful.	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Al-Mu'minoon	0	"Far, very far is that which ye are promised!	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Al-Kahf	1	Then followed he (another) way,	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Al-Insan	1	A fountain there, called Salsabil.	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Şekil 1. Ayet etiketlerinin veri temizleme sonrası kelime bulutu gösterimi

3.2. Öznitelik Seçimi İçin Genetik Algoritma Kullanılması

Bu çalışmada performans değerlendirilmesi açısından farklı yöntemlere başvurulmuştur. Öncelikle hiçbir şekilde öznitelik seçimi yapılmadan ilgili deneyler yapılmıştır. Sonra ise SVM, Naive Bayes, Lojistik Regresyon ve Random Forest gibi yöntemlerle Genetik Algoritma'nın birleştirilerek kullanılması; daha sona da Recursive Feature Elimination ve Recursive Feature Elimination + Random Forest yönteminin kullanılması üzerine deneyler yapılmıştır. Bunların arasında performansa en olumlu katkıyı Genetik Algoritma sağladığı için bu yöntem ile devam edilmesine karar verilmiştir. Özünde Genetik Algoritma; Darwin'in evrim teorisini taklit eden önemli bir sezgisel algoritmadır. Genetik Algoritma'ları herhangi bir probleme uygularken ele alınması gereken ana konular uygun bir reprezentasyon ve yeterli bir değerlendirme fonksiyonudur. Genetik algoritmaların süreci aşağıdaki gibidir:

- Genetik Algoritmalar, farklı yöntemlerle kodlanan bir başlangıç popülasyonu var edebilmek için rastgele ilk bireyleri oluşturur.
- Her birey, verilen problemin çözümünü temsil eden değişken bir Gen'den oluşur ve Kromozom tarafından kodlanır.
- Genetik Algoritma tasarımı şu üç önemli operatörü içermelidir: Seçim, Çaprazlama ve Mutasyon.
- Seçim operatöründe, yüksek Fitness değerlerine sahip en iyi bireyler yeni bir nesil yetiştirmek için seçilir. Çaprazlama operatörü, daha iyisini üretmek amacıyla iki ebeveynin kromozomlarının birleştirilmesi işlemidir.
- Mutasyon, lokal olarak belirli bir bireyi genetik çeşitliliğin sağlanması adına değişikliğe uğratır.
- Öznitelik seçiminde asıl sorun çözümün reprezentasyonuyla ilgilidir.

3.3. Çok Etiketli Metin Sınıflandırması İçin Sözde Kod

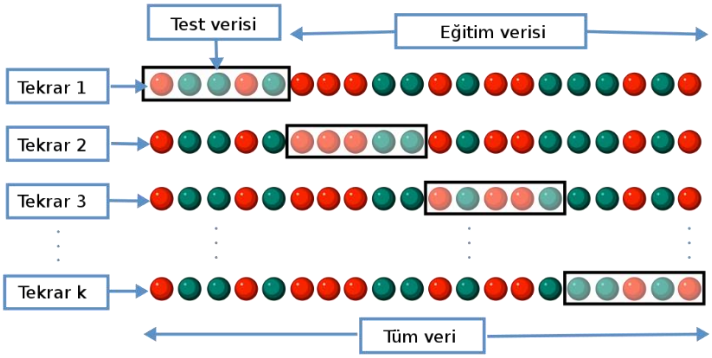
Çalışmamız kapsamında gerçekleştirdiğimiz çok etiketli metin sınıflandırması için uygulanan algoritmanın sözde kodu aşağıda verilmiştir.

Tablo 2. Metin sınıflandırma algoritmasının sözde kodu

1. genetik_algoritma_modelleri : [Naive Bayes, SVM, Lojistik Regresyon, Random Forest]
2. başlangıç_etiket_listesi : ['pillars_of_islam', 'faith', 'al_quran', 'science','actions', 'call_to_God','jihad','human_and_social_relations','morals','property_regulations','legal_matters,'state_and_society', 'agriculture_and_trade', 'history_and_stories', 'religion']
3. genetik_algoritma_modelleri içindeki her bir **model** için
   a. Başlangıç popülasyonunu (initial population) oluştur
   b. Uygunluk (fitness) değerini belirle
   c. Uygunlu tekrarla
      i. Seleksiyon
      ii. Çaprazlama
      iii. Mutasyon
      iv. Uygunluk (fitness) değerini hesapla
      v. Uygunluk değeri ile sonlandırma kriterine ulaşıldıysa bitir
4. En yüksek performans değeri veren **model**i öznitelik seçiminde kullan
5. seçilen_öznitelikler : En yüksek performans değeri veren modelin seçtiği öznitelik değerleri
6. df_train : Train (Eğitim) veri seti, **df_test :** Test veri seti
7. eğitim_seti_boyutu : 0.80**, test_seti_boyutu :** 0.20
8. sınıflandırma_modelleri : [Naive Bayes, SVM, Lojistik Regresyon, K-NN, AdaBoost]
9. df_train ve **df_test içindeki her bir ayet için** : #Veri temizleme işlemlerini başlat
   a. Noktalama işaretlerini kaldır
   b. Alfanümerik olmayan karakterleri kaldır
   c. Stopword (etkisiz kelime)leri kaldır
   d. Kök oluşturma (Stemming) ve Gövdeleme(Lemmatization) uygula
   e. En sık kullanılan 25 kelimeyi kaldır
   f. En az kullanılan 25 kelimeyi kaldır
10. train, test: df_train veri setini **test_size** ile orantılı olacak şekilde iki parçaya böl
11. sınıflandırma_modelleri içindeki her bir **model** için:
   a. **X_train**: train veri setinin vektörize edilmiş hali
   b. **X_test**: test veri setinin vektörize edilmiş hali
   c. **ngram_aralığı:** (1,2)
   d. **seçilen_öznitelikler** listesindeki her bir **öznitelik** için:
      i. model.fit(X_train, train[label])
      ii. **tahmin** : model.predict(X_test)
12. tahmin ve **test[label]** değerlerini kıyaslayarak performans ölçümlerini göster

3.4. Performans Değerlendirme Metrikleri

Bu çalışmada yer alan başarım ölçümleri, literatürde sıklıkla rastlanan performans değerlendirme metriklerinden Accuracy (Doğruluk), Precision (Kesinlik), Recall (Hassasiyet), F1 Skoru, AUROC (ROC Eğrisi Altında Kalan Alan) ve Hamming Loss (Hamming Kaybı) kullanılmıştır. Bunlara ek olarak Kur'an ayetleriden oluşan ilk veri setimizi eğitim ve test veri setleri olarak ayırdığımız için dağılımdan kaynaklanan sapma ve hataların asgariye indirilebilmesi ve objektif bir ölçüm yapılabilmesi için K-Fold Cross Validation (K-Katlamalı Çapraz Doğrulama) yöntemi de kullanılmıştır.



Şekil 2. K-katlamalı çapraz doğrulama

Literatürde sıklıkla kullanılan bu yöntemde amaç, veri setini k adet kümeye bölmek ve her bir aşamada k-1 sınıfı eğitim için kullanırken 1 sınıfı da test için kullanmaktır. Böylece veriyi homojen olarak eğitim ve test kümelerine ayırma şansı olacaktır. Bu çalışmada seçilen algoritmalarda k=5, k=10, k=15 ve k=20 değerleri kullanılmıştır.

4. Bulgular ve Tartışma

Bu çalışma kapsamında Kur'an ayetlerinin otomatik olarak sınıflandırılması için makine öğrenmesi yöntemlerinden yararlanılmıştır. Naive Bayes, SVM (Karar Destek Makinesi), Lojistik Regresyon ve Rastgele Orman algoritmaları Genetik Algoritma ile birleştirilerek kullanılmıştır. Her bir kolektif (ensemble) öğrenme modelinin sergilediği performans değerleri kontrol edilerek içlerinden en yüksek AUROC (Area Under Receiving Operator Curve) skorunu veren model ve o modele ait öznitelik kümesi bir sonraki adımda asıl hedeflenen sınıflandırma işleminin gerçekleştirilmesi açısından girdi vazifesi üstlenmiştir. Hem zaman maliyeti hem de karmaşıklığın düşürülmesi ve gereksiz özniteliklerin hesaba katılmaması için Genetik Algoritma öznitelik seçimi aşamasında katkıda bulunmuştur. Tablo 2'de de görüleceği üzere 0.991 başarım oranı ile Lojistik Regresyon + Genetik Algoritma tarafından üretilen 8 tane öznitelik bir sonraki aşamada kullanılmış olup başlangıçta verisetinde yer alan 15 etiketten 7' sini çıkarmamıza sebep olmuştur.

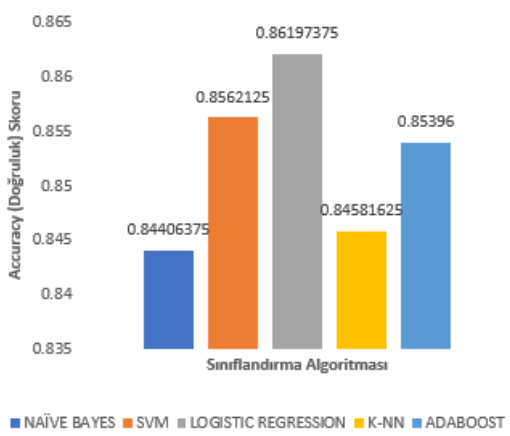
Tablo 3. En iyi özniteliklerin seçilmesi için farklı algoritmalarının performans karşılaştırma tablosu

Algoritma	Çapraz Doğrulama (k)	Öznitelik Sayısı	Seçilen Öznitelikler	Skor
Random Forest + Genetik Algoritma	k = 5	8	pillars_of_islam,faith,al_quran,actions,jihad,human_and_social_relations,morals, history_and_stories	0.989
SVM + Genetik Algoritma	k = 5	9	pillars_of_islam,faith,al_quran,science,human_and_social_relations,morals,property_regulations,agriculture_and_trade, history_and_stories	0.988
Random Forest + Genetik Algoritma	k = 15	10	pillars_of_islam,faith,al_quran,science,actions, call_to_God, morals,legal_matters, history_and_stories,religion	0.986
SVM + Genetik Algoritma	k = 15	9	pillars_of_islam,faith,al_quran,call_to_God,human_and_social_relations,morals,property_regulations,state_and_society, history_and_stories	0.985

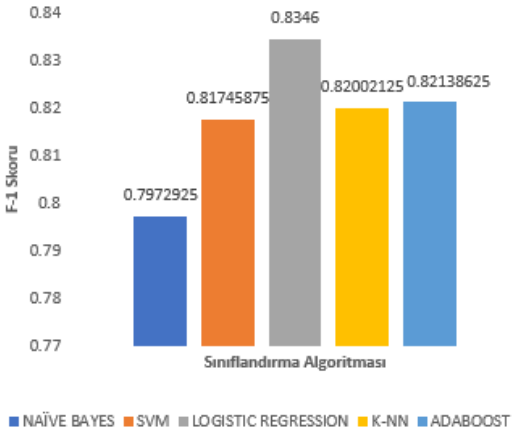
SVM + Genetik Algoritma	k = 20	9	pillars_of_islam,faith,science,actions,jihad,human_and_social_relations,morals,property_regulations,history_and_stories	0.985
Random Forest + Genetik Algoritma	k = 10	9	pillars_of_islam,faith,al_quran,science,actions,human_and_social_relations,property_regulations,agriculture_and_trade,history_and_stories	0.984
Random Forest + Genetik Algoritma	k = 20	10	pillars_of_islam,faith,actions,call_to_God,jihad,human_and_social_relations,morals,property_regulations,state_and_society,history_and_stories	0.982
SVM + Genetik Algoritma	k = 10	8	pillars_of_islam,faith,al_quran,jihad,human_and_social_relations,property_regulations,legal_matters,history_and_stories	0.981
Lojistik Regresyon + Genetik Algoritma	k = 20	10	pillars_of_islam,faith,call_to_God,jihad,human_and_social_relations,morals,property_regulations,state_and_society,agriculture_and_trade,history_and_stories	0.979
Lojistik Regresyon + Genetik Algoritma	k = 5	7	pillars_of_islam,faith,science,jihad,human_and_social_relations,morals,history_and_stories	0.978
Lojistik Regresyon + Genetik Algoritma	k = 10	9	pillars_of_islam,faith,science,call_to_God,jihad,human_and_social_relations,state_and_society,history_and_stories,religion	0.975
Lojistik Regresyon + Genetik Algoritma	k = 15	7	pillars_of_islam,faith,call_to_God,jihad,property_regulations,legal_matters,history_and_stories	0.96
Naive Bayes + Genetik Algoritma	k = 5	3	pillars_of_islam,faith,agriculture_and_trade	0.897
Naive Bayes + Genetik Algoritma	k = 20	4	pillars_of_islam,faith,state_and_society,agriculture_and_trade	0.89
Naive Bayes + Genetik Algoritma	k = 10	3	pillars_of_islam,faith,legal_matters	0.87
Naive Bayes + Genetik Algoritma	k = 15	6	pillars_of_islam,faith,call_to_God,jihad,state_and_society,agriculture_and_trade	0.85
Lojistik Regresyon (Öznitelik Seçimi Yok)	k = 15	15	pillars_of_islam', 'faith', 'al_quran', 'science', 'actions', 'call_to_God','jihad','human_and_social_relations', 'morals', 'property_regulations', 'legal_matters','state_and_society', 'agriculture_and_trade', 'history_and_stories', 'religion'	0.656
Lojistik Regresyon (Öznitelik Seçimi Yok)	k = 5	15	pillars_of_islam', 'faith', 'al_quran', 'science', 'actions', 'call_to_God','jihad','human_and_social_relations', 'morals', 'property_regulations', 'legal_matters', 'state_and_society', 'agriculture_and_trade', 'history_and_stories',	0.654

			'religion'	
Lojistik Regresyon (Öznitelik Seçimi Yok)	k = 10	15	pillars_of_islam', 'faith', 'al_quran', 'science', 'actions', 'call_to_God', 'jihad','human_and_social_relations', 'morals', 'property_regulations', 'legal_matters', 'state_and_society', 'agriculture_and_trade', 'history_and_stories', 'religion'	0.654
Lojistik Regresyon (Öznitelik Seçimi Yok)	k = 20	15	pillars_of_islam', 'faith', 'al_quran', 'science', 'actions', 'call_to_God', 'jihad','human_and_social_relations', 'morals', 'property_regulations', 'legal_matters', 'state_and_society', 'agriculture_and_trade', 'history_and_stories', 'religion'	0.654
Random Forest + Recursive Feature Elimination	k = 5	4	faith', 'al_quran', 'call_to_God', 'human_and_social_relations'	0.629
Recursive Feature Elimination	k = 20	8	pillars_of_islam', 'al_quran', 'science', 'actions', 'call_to_God', 'jihad', 'state_and_society', 'agriculture_and_trade'	0.628
Random Forest + Recursive Feature Elimination	k = 20	5	pillars_of_islam', 'faith', 'al_quran', 'call_to_God', 'human_and_social_relations'	0.627
Random Forest + Recursive Feature Elimination	k = 15	4	faith', 'al_quran', 'call_to_God', 'human_and_social_relations'	0.62
Recursive Feature Elimination	k = 15	8	pillars_of_islam', 'al_quran', 'science', 'actions', 'call_to_God', 'jihad', 'state_and_society', 'agriculture_and_trade'	0.62
Random Forest + Recursive Feature Elimination	k = 10	5	faith', 'al_quran', 'actions', 'call_to_God', 'human_and_social_relations'	0.619
Recursive Feature Elimination	k = 10	6	pillars_of_islam', 'al_quran', 'science', 'call_to_God', 'jihad', 'agriculture_and_trade'	0.617

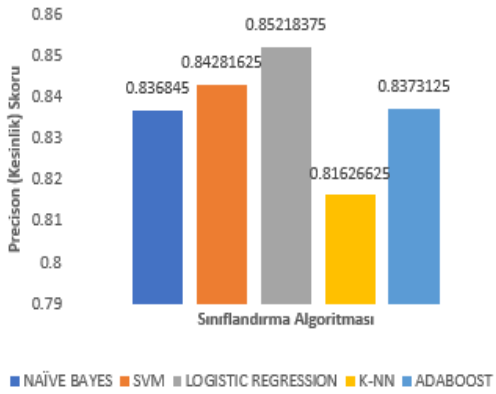
Öznitelik seçimi tamamlandıktan sonra test veri seti üzerinde Naive Bayes, SVM, Lojistik Regresyon, K-NN ve AdaBoost algoritmaları kullanılarak çok etiketli sınıflandırma probleminin çözümü için Doğruluk, F1 Skoru, Kesinlik, Duyarlılık ve AUROC (ROC Eğrisinin Altında Kalan Alan) gibi farklı performans metrikleri ışığında performans ölçümlemeleri yapılmıştır. Toplamda seçilen 8 etiketin her biri için öğrenme oranını gösteren ölçüm sonuçları Şekil 3'te ayrıntılı olarak gösterilmiştir.



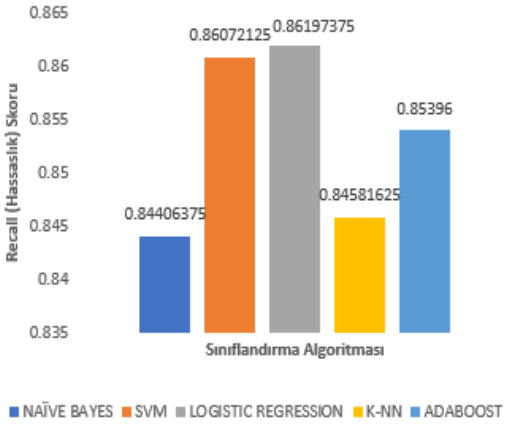
(a)



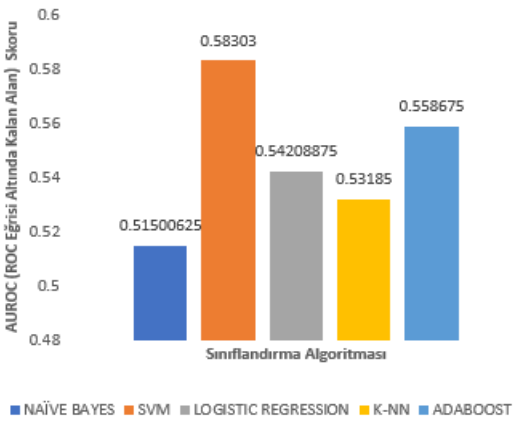
(b)



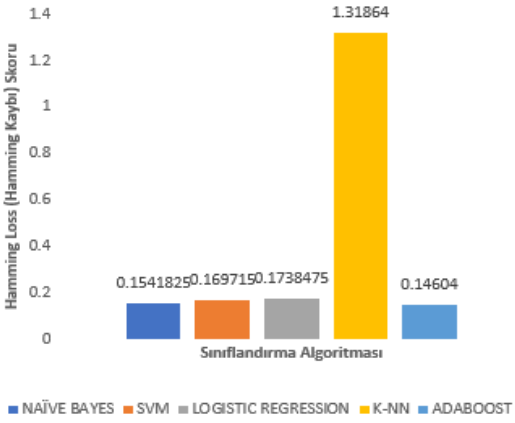
(c)



(d)



(e)



(f)

Şekil 3. Metin sınıflandırma modellerinin performans karşılaştırması

Şekil 3'teki veriler doğrultusunda Doğruluk (Şekil 3a), F1 Skoru (Şekil 3b), Kesinlik (Şekil 3c), ve Hassasiyet (Şekil 3d) değerleri bakımından performans ölçümleri yapıldığında Lojistik Regresyon, SVM ve AdaBoost'un Naive Bayes ve K-NN'e göre daha iyi sonuçlar verdiği görülmektedir. ROC Eğrisi Altında Kalan Alan (Şekil 3e) bakımından kıyaslandığında ise SVM'in daha iyi bir başarım gösterdiği sonucuna varılabilir. Hamming Kaybı (Şekil 3f) açısından bakıldığında ise en kötü sonuçlar K-NN modeli için gözlenmiş olup AdaBoost en iyi sonucu vermiştir. Şekil 3'te yer alan metrikler baz alındığında çalışmanın bir sonraki adımı olan seçilen sınıflandırma modelinin test veri setinde kullanılması kısmında Lojistik Regresyon modelinin tercih edilebileceği görülmektedir.

Ortalamadaki değerler baz alındığında kaybı en düşük olan Lojistik Regresyon ya da SVM modellerinin tercih edilebileceği görülmektedir. Bu çalışmada ise literatürdeki diğer çalışmalarda da çoğulukla rastlandığı üzere F1 Skoru referans alınmış olup bu metrik özelinde en yüksek performansı sergileyen SVM tercih edilmiştir.

5. Sonuç ve Öneriler

Bu çalışma kapsamında bir çok etiketli metin sınıflandırma problemi olarak Kur'an-ı Kerim ayetlerinin önceden belirlenen sınıflara doğru bir şekilde atanabilmesi için farklı makine öğrenmesi yöntemleri denenmiştir. Bunun yanısıra ikili sınıflandırma örneği olarak bahse konu ayetlerin Mekki ya da Medeni olarak kategorize edilmesi için kullanılabilecek özniteliklerin seçimi noktasında Genetik Algoritma'dan yararlanılmış, farklı makine öğrenme algoritmalarıyla birleştirilerek kullanılmış ve gereksiz ya da nihai tahminlemeye yeterince katkısı bulunmayan özniteliklerin göz ardı edilebilmesi sağlanmıştır. Gelecek çalışmalarımız arasında ise metin sınıflandırma performansımızın iyileştirilmesi ve daha geniş bir karşılaştırma zemini tesis edilebilmesi açısından diğer evrimsel ve yarı-sezgisel algoritmaların veya hibrit çözümlerin de denenmesi yer almaktadır.

Teşekkür

İstanbul Üniversitesi-Cerrahpaşa Bilgisayar Mühendisliği Ana Bilim Dalı ve Huawei Türkiye Ar-Ge Merkezi'nin süreçteki katkılarından dolayı teşekkür ederiz.

Yazar(lar)ın Katkıları

Bu çalışmanın fikir, implementasyon ve makale yazım aşamaları Doç. Dr. Muhammed Ali AYDIN ve Doç. Dr. Zeynep Orman'ın danışmanlığında; İstanbul Üniversitesi-Cerrahpaşa Bilgisayar Mühendisliği Ana Bilim Dalı'nda doktora öğrencisi olarak eğitimine devam eden Fatih MERT tarafından gerçekleştirilmiştir. Her üç yazar da makalenin son halini okumuş ve onaylamıştır.

Çıkar Çatışması

Yazarlar, çıkar çatışması olmadığını beyan eder.

Kaynaklar

[1]. Shakir, M.H., The Holy Quran English Translation, Erişim : 13.11.2022, https://quran-archive.org/explorer/m-h-shakir

[2]. Adeleke, A.O., Samsudin, N.A., Mustapha, A. ve Nawi, N.M., Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses, International Journal on Advance Science, Engineering and Informational Technologies, 7(4), 2017, 1419-1427.

[3]. Adeleke, A.O., Samsudin, N.A., Mustapha, A. Ve Nawi N.M., A Group-Based Feature Selection Approach to Improve Classification of Holy Quran Verses, Recent Advances on Soft Computing and Data Mining, Advances in Intelligent Systems and Computing, 2018, 700, 282-297

[4]. Goudjil, M., Bedda, M., Koudil, M. ve Ghoggali, N., Using Active Learning in Text Classification of Quranic Sciences, International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, 2015, 209-213

[5]. Ibrahim, E.A.A, Ataelfadiel, M.A.M., Atwel, E.S., Provisions of Quran Tajweed Ontology (Articulations Points of Letters, UN Vowel Noon and Tanween), International Journal of Science and Research, 2017, 6(8), 756-761.

[6]. Zharmagambetov, A.S., Pak. A.A., Sentiment analysis of document using deep learning and decision trees, Twelve IEEE International Conference on Electronics Computer and Computation, 2015, 1-4.
[7]. Wang, J.H., Wang, H.Y., Incremental Neural Network Construction for Text Classification, IEEE International Symposium on Computer Consumer and Control, 2014, 254, 970-973.
[8]. Townsend, K.R., Sun. S., Johson, T., Attia, O.G., Jones, P.H., ve Zambreno, J., k-NN text classification using an FPGA-based sparse matrix vector multiplication accelerator, IEEE International Conference on Electronic/Information Technology, 2015, 257-263.
[9]. Sabbah, T., Selamat, A., Support Vector Machine based approach for Quranic words detection in online textual content, 8th IEEE Malaysian Software Engineering Conference, Malaysia, 2014, 325- 330.
[10]. Al-Kabi, M. N., Ata, B. M. A., Wahsheh, H. A., ve Alsmadi, I. M., A topical classification of Quranic Arabic text, Proceedings of the 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, 2013, 252-257.
[11]. Kalousis A., Prados J. ve Hilario M., Stability of Feature Selection Algorithms: a study on high dimensional spaces, Knowledge and Information System, 2007, 12(1), 95-116.
[12]. Sun, Y., Qu, W., Zhou, J., Tang, X., Di, Y., & Wu, W., An Improved Feature Selection Method in Chinese Text Categorization, International Journal of Knowledge, www.ijklp.org and Language Processing KLP International, 2011, 2(3), 48-55.
[13]. Telkom University Database, Erişim : 13.11.2022, https://dataverse.telkomuniversity.ac.id/