



T.C.
KAHRAMANMARAŞ SÜTÇÜ İMAM ÜNİVERSİTESİ
BİLİMSEL ARAŞTIRMA PROJELERİ KOORDİNASYON BİRİMİNE

Proje Numarası :	2021/2-38 M
Projenin Adı :	Büyük Veri Mimarisinde Yeni Bir Yazar Analiz ve Doğrulama Modeli
Projenin Türü :	Münferit Proje (M)
Proje Yön. Unv.-Adı-Soyadı :	Dr.Öğr.Üyesi Pelin CANBAY
Görev Yeri :	Mühendislik ve Mimarlık Fakültesi
Bölümü :	Bilgisayar Mühendisliği
Sözleşmeye Göre Projenin	
Başlangıç Tarihi :	20-05-2021
Bitiş Tarihi :	19-07-2023
Alınan Ek Süre :	12 Ay
Sonuç Raporu Tarihi :	02-07-2023
Proje Bütçesi :	44.993,40 TL
Alınan Ek Ödenek :	0,00 TL
Projenin Toplam Bütçesi :	44.993,40 TL
Harcanan Miktar :	43.999,84 TL
Proje Kalan Bütçesi :	993,56 TL
İşyeri Telefon Numarası :	(344)300-1701
Cep Telefon Numarası :	(534)548-5632
Proje Yöneticisinin İş Adresi :	
Demirbaş[Taşınır Alımları	Proje Kapsamında Demirbaş Niteliğinde Taşınır Alınmıştır. Sayıştay Başkanlığı'nın Raporu ve BAP Komisyonu Gereği, Taşınır GÖREV Yaptığım Birimin Ambarına, Harcama Yetkilisinin Bilgisi Dahilinde Teslim Edilmiştir.

Projenin sonuç raporunun kabulünü arz ederim. 02/07/2023

Dr.Öğr.Üyesi Pelin CANBAY
(İmza)

Projenin Sonuç Raporunun, Üniversitemiz BAP/Kütüphane İnternet Sayfasında;

- | | |
|---|-------------------------------------|
| 1.Yayınlanmasına İzin Vermiyorum. | ☐ |
| 2.Sadece Özetin Yayınlanmasına İzin Veriyorum. | ☒ |
| 3.Tam Metin Yayınlanmasına İzin Veriyorum. | ☐ |
| 4. Tarihinden İtibaren Tam Metin Yayınlanmasına İzin Veriyorum. | ☐ |



T.C.
KAHRAMANMARAŞ SÜTÇÜ İMAM ÜNİVERSİTESİ
BİLİMSEL ARAŞTIRMA PROJELERİ KOORDİNASYON BİRİMİ
SONUÇ RAPORU

F-6

Proje Adı

Büyük Veri Mimarisinde Yeni Bir Yazar Analiz ve Doğrulama Modeli

Proje No	Başlama Tarihi	Bitiş Tarihi	
2021/2-38 M	20-05-2021	19-07-2023	
Rapor Türü	Raporun Kapsadığı Dönemin Tarihleri		
Sonuç Raporu	20-11-2022	19-07-2023	
Bilimsel Araştırmalar Birimi Tarafından Sağlanan Destek Miktarı	Gelişme Raporu Döneminde Harcanan	Şimdiye Kadar Harcanan	Kalan
44.993,40 TL		43.999,84 TL	993,56 TL
Destek Sağlayan Diğer Kuruluşların Katkısı	Gelişme Raporu Döneminde Harcanan	Şimdiye Kadar Harcanan	Kalan

SAĞLANAN DESTEKTEN ŞİMDİYE KADAR YAPILAN HARCAMALARIN FASILLARA GÖRE DAĞILIMI

Fasıl	Harcanan	Kalan
MAMUL MAL ALIMLARI	43.999,84 TL	993,56 TL

RAPOR DÖNEMİNE AİT İLGİLİ YAYINLAR

No	Yayın
1	Pelin CANBAY, Zübeyde Demircioğlu, Melting of Privacy with Machine Learning, Big Data, and Social Media, Acta INFOLOGICA, 7, 153 - 164 (2023).

RAPOR DÖNEMİNE AİT İLGİLİ FİKRİ MÜLKİYETLER

No	Fikri Mülkiyet Adı	Fikri Mülkiyet No	Alındığı Ülke	Tescil Tarihi
----	--------------------	-------------------	---------------	---------------

RAPOR DÖNEMİNDE YAPILAN ÇALIŞMALAR VE ELDE EDİLEN SONUÇLAR

Sonuç Rapor Özeti :

Dijital dünyanın hızlı ve kontrolsüz gelişimi, suç örgütleri veya terör organizasyonları için iletişim kurulabilecek anonim bir ortam sağlamaktadır. Dijital ortamların anonimliğinden faydalanan suçluların tanımlanabilmesi büyük bir zorluk olarak adli bilişim uzmanlarının önlerine çıkmaktadır. Bu proje kapsamında geliştirilen büyük veri temelli yazar analiz modeli ile dijital ortamlarda anonim olarak paylaşımda bulunan kişilerin profil analizinin yapılabilmesi amaçlanmıştır. Günümüz teknolojik haberleşme ortamlarında her gün milyonlarca dijital metin üretilmektedir. Yazar analizini de kapsayan Doğal Dil İşleme çalışmalarında verinin büyüklüğü arttıkça elde edilecek başarı da artmaktadır. Öte yandan makine öğrenmesi yöntemlerinin kullanıldığı analiz çalışmalarında verinin büyüklüğü arttıkça standart uygulama ortamları işlem gücü bakımından yetersiz kalmaktadır. Bu yetersizlik karşısında, yazar analizi işlemlerinde büyük veri platformlarının kullanılması gerekliliği ortaya çıkmıştır. Bu gereklilikler doğrultusunda, projede amaçlanan yazar analiz uygulaması büyük veri platformunda gerçekleştirilmiştir. Türkçe dijital metinler kullanılarak büyük veri platformunda gerçekleştirilen ilk yazar analizi çalışması olan bu projede, yazar yaş ve cinsiyet bilgisine dayalı profil çıkarımı yapılmaktadır. Makine öğrenmesi yöntemleri kullanılarak dijital metinlerin yazarı ile ilgili yaş ve cinsiyet bilgisinin tahminlendiği projede elde edilen sonuçların başarısı, kullanılan veri kümesi ve metinlerin özelliklerine bağlı olarak kabul edilebilir aralıktadır. Proje kapsamında ayrıca, yazar analiz çalışmalarını da içeren yapay zeka çalışmalarının mahremiyet duyarlılığı ele alınmıştır. Yazar analizi çalışmaları kişisel bilginin analizini içerdiğinden proje kapsamında mahremiyet konusunun ele alınması bir gereklilik olmuştur. Bu gereklilik sebebi ile projede geniş çaplı bir, yapay zeka, büyük veri ve dijital metinleri kapsayan mahremiyet araştırması yapılmıştır. Yapılan teknik ve sosyolojik araştırmalar sonucunda dijital dünyada veri paylaşmanın ve paylaşılan bu veriler üzerinde analiz yapmanın önüne geçilemeyeceği, bu veri ve analizler sonucunda birçok profil bilgisinin analiz edilebileceği sonucuna ulaşılmıştır.

Rapor içeriği için lütfen eke bakınız.

EK: [bap_final_PC.pdf](#)

Proje Kapsamında Demirbaş Niteliğinde Taşınır Alınmıştır. Sayıştay Başkanlığı'nın Raporu ve BAP Komisyonu Kararı Gereği, Taşınır G rev Yaptıėım Birimin Ambarına, Harcama Yetkilisinin Bilgisi Dahilinde Teslim Edilmiřtir.

S.No	DemirbaşTařınır adı
1.	İř İstasyonu-1 (1 Adet) - 830.06.01 kodlu kalemden alım yapılmıřtır.
2.	İř İstasyonu-2 (1 Adet) - 830.06.01 kodlu kalemden alım yapılmıřtır.

Proje Y�r�t�c� Unvanı, Adı SOYADI	İmzası	Tarih
Dr.�ėr.�yesi Pelin CANBAY		02-07-2023



**T.C.
KAHRAMANMARAŞ SÜTÇÜ İMAM
ÜNİVERSİTESİ**

**BÜYÜK VERİ MİMARİSİNDE YENİ BİR YAZAR
ANALİZ VE DOĞRULAMA MODELİ**

**BİLİMSEL ARAŞTIRMA PROJELERİ
KOORDİNASYON BİRİMİ**

2023 KAHRAMANMARAŞ

BÜYÜK VERİ MİMARİSİNDE YENİ BİR YAZAR ANALİZ VE DOĞRULAMA MODELİ

Proje No: 2021/2-38 M

Dr. Öğr. Üyesi Pelin CANBAY

2023 KAHRAMANMARAŞ

ÖNSÖZ

Bu bilimsel araştırma projesi, kullanıcılara gizlilik imkanı sunan dijital ortamlar üzerinden zararlı içerikler yayınlayan kişilerin profil analizinin yapılabilmesi amacı ile geliştirilmiştir. Yazar analizi kapsamında gerçekleştirilen bu çalışmada, dijital ortamda üretilen metinlerin analiz edilmesi ile söz konusu metnin yazarı ile ilgili yaş, cinsiyet gibi yazara ait bilgiler doğal dil işleme teknikleri kullanılarak tahmin edilmektedir. Türkçe dijital metinlerin büyük veri platformunda analizini içeren bu çalışma ile dijital ortamlarda zararlı içerik üreten kişilerin tespitinin adli bilişim kapsamında daha kolay ve hızlı yapılabilmesi ön görülmektedir. Bu çalışmada ayrıca, yazar analizini de içeren yapay zeka çalışmalarına yönelik mahremiyet endişeleri ve yaklaşımları ayrıntılı bir şekilde araştırılmış, araştırma sonuçları elde edilen bulgular doğrultusunda yorumlanmıştır. Bu çalışmada kullanılan Türkçe dijital veriler Hacettepe Üniversitesi Bilgisayar Mühendisliği Bölümü Multimedya Information Retrieval laboratuvarında toplanmıştır. Veri toplama aşamasında yönlendirmeleri ile destek olan Sayın Prof. Dr. Ebru Akçapınar Sezer'e ve bu sürecin sağlıklı ilerlemesinde katkılarını esirgemeyen laboratuvar elemanlarına teşekkürlerimi sunarım.

Bu bilimsel araştırma projesi Kahramanmaraş Sütçü İmam Üniversitesi tarafından proje numarası 2021/2-38 M ile desteklenmiştir. Destekleri için Kahramanmaraş Sütçü İmam Üniversitesi Rektörlüğü Bilimsel Araştırma Projeleri Koordinasyon Birimine teşekkür ederim.

Projenin başlama tarihi 20/05/2021 ve bitiş tarihi 19/07/2023'tür.

Proje Yürütücüsü
Dr. Öğr. Üyesi Pelin CANBAY

İÇİNDEKİLER

ÖNSÖZ	i
İÇİNDEKİLER	ii
TABLolar DİZİNİ	iii
ŞEKİLLER DİZİNİ	iv
SİMGELER VE KISALTMALAR DİZİNİ	v
ÖZET	vi
ABSTRACT	vii
1. GİRİŞ	1
2. MATERYAL VE YÖNTEM (METOD)	3
2.1. BÜYÜK VERİ İLE YAZAR PROFİL ÇIKARIMI	3
2.1.1. Büyük Veri Özellikleri ve Araçları	3
2.1.2. Kullanılan Veri Kümesi	9
2.1.3. Kullanılan Algoritmalar ve Makine Öğrenmesi Yöntemleri	10
2.2. MAKİNE ÖĞRENMESİ, BÜYÜK VERİ VE MAHREMİYET	13
2.2.1. Mahremiyet ve Sosyal Bir Varlık Olarak Birey	16
2.2.2. Makine Öğrenimi ve Mahremiyet	19
2.2.3. Büyük Veri ve Mahremiyet	21
2.2.4. Mahremiyet ile Teknoloji	22
3. ARAŞTIRMA BULGULARI VE TARTIŞMA	26
3.1. DEĞERLENDİRME	29
4. SONUÇLAR VE ÖNERİLER	32
KAYNAKLAR	34

TABLÖLAR DİZİNİ

TABLO 1. FARKLI ALANLAR İÇİN KULLANILAN BÜYÜK VERİ TEKNOLOJİLERİ.	3
TABLO 2. KULLANILAN VERİ KÜMESİNİN İSTATİSTİKSEL ÖZELLİKLERİ.....	10
TABLO 3. YAZAR ANALİZİNDE CİNSİYET TAHMİNLEME RAPORU	29
TABLO 4. YAŞ TAHMİNLEME SINIFLANDIRMA RAPORU	30

ŞEKİLLER DİZİNİ

ŞEKİL 1. TEMEL HADOOP BİLEŞENLERİ.....	5
ŞEKİL 2. MAPREDUCE TABANLI ÖRNEK BİR SİSTEM MİMARİSİ	6
ŞEKİL 3. TEMEL SPARK BİLEŞENLERİ	7
ŞEKİL 4. SPARK TABANLI ÖRNEK BİR SİSTEM MİMARİSİ.....	9
ŞEKİL 5. TEMSİLİ KARAR AĞACI GÖSTERİMİ.....	12
ŞEKİL 6. KNN ALGORİTMASI ÇALIŞMA PRENSİBİ.....	13
ŞEKİL 7. DİJİTAL ORTAMA VERİLEN VEYA DİJİTAL ORTAMDAN ELDE EDİLEN BAZI KİŞİSEL BİLGİLER	17
ŞEKİL 8. BAZI GÜNCEL UYGULAMALARIN SOSYAL MEDYA VE METİN ÜZERİNDEN SÜREÇ AKIŞI	20
ŞEKİL 9. TEKNOLOJİK VE ALGISAL DEĞİŞİMLERİN ETKİSİYLE MAHREMİYETİN ERİMESİNİN KRONOLOJİK SIRALAMASI.....	24
ŞEKİL 10. CİNSİYET TAHMİNLEMEDE GERÇEK VE TAHMİN DAĞILIMI	30
ŞEKİL 11. YAŞ TAHMİNLEMEDE GERÇEK VE TAHMİN DAĞILIMI.....	31

SİMGELER VE KISALTMALAR DİZİNİ

SİMGELER

t	: terim
d	: doküman
D	: Derlem
N	: Derlemdeki farklı kelime sayısı
Tf	: terim frekansı
Idf	: Inverse Document Frequency (ters doküman frekansı)
Y	: Sınıf etiketi
X	: Öznitelik kümesi
x_i	: i. öznitelik

KISALTMALAR

HDFS	: Hadoop File System
RDD	: Resilient Distributed Datasets (Esnek Dağıtılmış Veri Kümeleri)
NLP	: Natural Language Processing
LR	: Logistic Regression
SVM	: Support Vector Machine
MNB	: Multinomial Naive Bayes
RF	: Random Forests
kNN	: k Nearest Neighbor
SparkMllib	: Spark machine learning library
GDPR	: General Data Protection Regulation (Genel Veri Koruma Tüzüğü)

ÖZET

Dijital dünyanın hızlı ve kontrolsüz gelişimi, suç örgütleri veya terör organizasyonları için iletişim kurulabilecek anonim bir ortam sağlamaktadır. Dijital ortamların anonimliğinden faydalanan suçluların tanımlanabilmesi büyük bir zorluk olarak adli bilişim uzmanlarının önlerine çıkmaktadır. Bu proje kapsamında geliştirilen büyük veri temelli yazar analiz modeli ile dijital ortamlarda anonim olarak paylaşımında bulunan kişilerin profil analizinin yapılabilmesi amaçlanmıştır.

Günümüz teknolojik haberleşme ortamlarında her gün milyonlarca dijital metin üretilmektedir. Yazar analizini de kapsayan Doğal Dil İşleme çalışmalarında verinin büyüklüğü arttıkça elde edilecek başarı da artmaktadır. Öte yandan makine öğrenmesi yöntemlerinin kullanıldığı analiz çalışmalarında verinin büyüklüğü arttıkça standart uygulama ortamları işlem gücü bakımından yetersiz kalmaktadır. Bu yetersizlik karşısında, yazar analizi işlemlerinde büyük veri platformlarının kullanılması gerekliliği ortaya çıkmıştır. Bu gereklilikler doğrultusunda, projede amaçlanan yazar analiz uygulaması büyük veri platformunda gerçekleştirilmiştir. Türkçe dijital metinler kullanılarak büyük veri platformunda gerçekleştirilen ilk yazar analizi çalışması olan bu projede, yazar yaş ve cinsiyet bilgisine dayalı profil çıkarımı yapılmaktadır. Makine öğrenmesi yöntemleri kullanılarak dijital metinlerin yazarı ile ilgili yaş ve cinsiyet bilgisinin tahminlendiği projede elde edilen sonuçların başarısı, kullanılan veri kümesi ve metinlerin özelliklerine bağlı olarak kabul edilebilir aralıktadır.

Proje kapsamında ayrıca, yazar analiz çalışmalarını da içeren yapay zeka çalışmalarının mahremiyet duyarlılığı ele alınmıştır. Yazar analizi çalışmaları kişisel bilginin analizini içerdiğinden proje kapsamında mahremiyet konusunun ele alınması bir gereklilik olmuştur. Bu gereklilik sebebi ile projede geniş çaplı bir, yapay zeka, büyük veri ve dijital metinleri kapsayan mahremiyet araştırması yapılmıştır. Yapılan teknik ve sosyolojik araştırmalar sonucunda dijital dünyada veri paylaşmanın ve paylaşılan bu veriler üzerinde analiz yapmanın önüne geçilemeyeceği, bu veri ve analizler sonucunda birçok profil bilgisinin analiz edilebileceği sonucuna ulaşılmıştır.

Anahtar Kelimeler: Büyük veri, yazar analizi, makine öğrenmesi, mahremiyet

ABSTRACT

The rapid and uncontrolled development of the digital world provides an anonymous environment for criminal organizations or terrorist organizations to communicate. Identifying criminals who benefit from the anonymity of digital environments is a major challenge for forensic experts. With the big data-based author analysis model developed within this project's scope, it aims to analyze the profile of people who share anonymously in digital media.

Millions of digital texts are produced every day in today's technological communication environments. In Natural Language Processing studies, which also include author analysis, the success to be achieved increases as the size of the data increases. On the other hand, as the data size increases in analysis studies in which machine learning methods are used, standard application environments become insufficient in processing power. In the face of this inadequacy, the necessity of using big data platforms in author analysis processes has emerged. In line with these requirements, the author analysis application aimed at the project was carried out on the big data platform. In this project, which is the first author analysis study carried out on a big data platform using Turkish digital texts, profile extraction is made based on the author's age and gender information. The success of the results obtained in the project, in which the age and gender information about the author of the digital texts are estimated using machine learning methods, is in the acceptable range depending on the dataset used and the characteristics of the texts.

Within the scope of the project, the privacy sensitivity of artificial intelligence studies, which also includes author analysis studies, was discussed. Since author analysis studies involve the analysis of personal information, it was necessary to address the issue of privacy within the project's scope. Due to this necessity, large-scale privacy research was carried out in the project, covering artificial intelligence, big data, and digital texts. As a result of the technical and sociological research, it has been concluded that sharing data in the digital world and analyzing these shared data cannot be avoided, and much profile information can be analyzed as a result of these data and analyses.

Keywords: Big data, author analysis, machine learning, privacy