



**T.C. İSTANBUL T CARET
 NİVERSİTESİ**

FEN BİLİMLERİ ENSTİT S 

Kİ İSELLE TİRİLMİ  HABER TAVSİYE SİSTEMİ

Melis  ZKARA

**Dan şman
Dr.  ğr.  yesi Metin TURAN**

**Y KSEK LİSANS TEZİ
BİLGİSAYAR M HENDİSLİĐİ ANABİLİM DALI
İSTANBUL - 2023**

KABUL VE ONAY SAYFASI

Melis ÖZKARA tarafından hazırlanan “**Kişiselleştirilmiş Haber Tavsiye Sistemi**” adlı tez çalışması 13/01/2023 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**’nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman **Dr. Öğr. Üyesi Metin TURAN**
İstanbul Ticaret Üniversitesi

Jüri Üyesi **Dr. Öğr. Üyesi Ali AKMAN**
İstanbul Ticaret Üniversitesi

Jüri Üyesi **Dr. Öğr. Üyesi Murat ORHUN**
İstanbul Bilgi Üniversitesi

Onay Tarihi : 14.04.2023

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsünün 14.04.2023 tarih ve 2023/382 numaralı Yönetim Kurulu Kararının 2. maddesi gereğince, ders yüklerini ve tez yükümlülüğünü yerine getirdiği belirlenen “Melis ÖZKARA” adlı öğrencinin mezun olmasına oy birliği ile karar verilmiştir.

Prof.Dr. Doğan KAYA
Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmada,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

14.04.2023

Melis ÖZKARA

İÇİNDEKİLER

	Sayfa
İÇİNDEKİLER.....	i
ÖZET	ii
ABSTRACT	iii
TEŞEKKÜR.....	iv
ŞEKİLLER.....	v
ÇİZELGELER.....	vi
SİMGELER VE KISALTMALAR.....	vii
1 GİRİŞ.....	1
2 LİTERATÜR ÖZETİ.....	4
3 TAVSİYE SİSTEMLERİ.....	8
3.1 Tavsiye Sistemlerinde Kullanılan Veri toplama Yaklaşımları.....	9
3.1.1 Dolaylı geribildirimler	9
3.1.2 Doğrudan geribildirimler.....	10
3.1.3 Hibrit geribildirimler	10
3.2 Tavsiye Sistemlerinde Kullanılan Yöntemler	10
3.2.1 İçerik tabanlı tavsiye sistemi.....	10
3.2.2 İşbirlikçi tavsiye sistemi.....	11
3.2.3 Popülarite tabanlı tavsiye sistemi	13
3.2.4 Hibrit tavsiye sistemi.....	14
3.3 Tavsiye Sistemlerinde Karşılaşılan Zorluklar	15
3.4 Haber Tavsiye Sistemi	15
4 MATERYAL VE YÖNTEM.....	16
4.1 Kullanılan Araçlar ve Kitaplıklar	16
4.2 Eğitim Veri Seti.....	17
4.3 Veri Ön İşleme	19
4.4 Tarayıcı Geçmişi ile Veri Çekme.....	19
4.5 WordCloud'un Kullanımı	20
4.6 Terim Ağırlıklandırma	21
4.7 Model Eğitimi.....	23
5 BULGULAR.....	25
5.1 Anlık Haberleri Gösterme	25
6 ARAŞTIRMA SONUÇLARI	27
7 SONUÇ VE ÖNERİLER	30
KAYNAKLAR	31
ÖZGEÇMİŞ.....	34

ÖZET

Yüksek Lisans Tezi

KİŞİSELLEŞTİRİLMİŞ HABER TAVSİYE SİSTEMİ

Melis ÖZKARA

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Metin TURAN
2023, 34 sayfa

Tavsiye Sistemleri veya diğer adıyla Öneri Sistemleri kullanıcının geçmişte tercih etmiş olduğu ürünlere bağlı olarak bir sonraki tercihini öngörülebilir hale getirip önermesini yapan bir yöntemdir. Bu yöntem gün geçtikçe daha popüler hale gelmeye başlamıştır. Elde ettiğimiz verilere göre ileriye yönelik önerme ve tahmin gerektirecek her türlü alanda kullanılabilir bir fayda sağlar. Bilgiyi en doğru şekilde çıkarmak gereklidir. Amazon şirketi gelirinin %35'ini tavsiye sistemlerinden elde etmesi bu yöntemin önemliliğini bize gösterir. Benzer olarak haber tavsiye sistemleri diğer uygulama alanları kadar yaygın bir şekilde kullanılmamaktadır. Geçmiş çalışmalara bakıldığında bu konu denenmemiş özgün bir çalışmadır. Bu çalışma haber tavsiye sistemi alanında bir uygulama olup kullanıcın arama motoru, girdiği siteler ve yer işaretleri dikkate alınarak bir sistem tasarlanmak amaçlanmıştır.

Kullanıcıya tavsiye edilecek haberleri sunabilmek için makine öğrenmesi modeli ile haber kategorileri ve haber içerikleri olan bir veri seti eğitilir. Kullanıcının girdiği siteler, arattığı kelimeler ve yer işaretleri verileri çekilerek eğitilen modele verilir ve sonucunda çıkan bilgiler dahilinde kullanıcının ilgilendiği haber kategorileri RSS (Rich Site Summary) tarafından anlık olarak işlenir. RSS'de seçilen haberler gündemdeki başlıklara göre öncelik sırasında kullanıcıya gösterilir. Model eğitiminde %86 doğruluk oranıyla gerçek kullanıcı testiyle %55 gibi bir doğruluk göstermiştir. Bu çalışma içerik tabanlı bir tavsiye sistemi çözümünü sunmayı amaçlamaktadır.

Anahtar Kelimeler: Kişiselleştirilmiş haber tavsiye sistemi, makine öğrenimi, NLP, tavsiye sistemleri.

ABSTRACT

M.Sc. Thesis

PERSONALIZED NEWS RECOMMENDATION SYSTEM

Melis ÖZKARA

**İstanbul Commerce University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering**

**Supervisor: Assoc. Prof. Dr. Metin TURAN
2023, 34pages**

Recommendation Systems or Suggestion Systems it is a method that makes the next choice predictable and makes its recommendation based on the products it has preferred in the past. This method is getting more and more popular day by day. According to the data we have obtained, it provides a useful benefit in all areas that require forward-looking suggestions and predictions. It is necessary to extract the information in the most accurate way. The fact that the Amazon company generates 35% of its revenue from referral systems shows us the importance of this method. Similarly, news recommendation systems are not as widely used as other application areas. Looking at the past studies, it is an untested original study on this subject. This study is an application in the field of news recommendation system and it is aimed to design a system by taking into account the user's search engine, sites and bookmarks.

In order to present the news to be recommended to the user, a data set with news categories and news contents is trained with the machine learning model. The sites the user has entered, the words he has searched for and the bookmarks data are drawn and given to the trained model, and the news categories that the user is interested in are instantly processed by the RSS (Rich Site Summary). The news selected in the RSS is displayed to the user in priority order according to the topics on the agenda. It showed an accuracy of 86% in model training and 55% in real user testing. This study aims to present a content-based recommendation system solution.

Keywords: Machine learning, NLP, personalized news recommendation system, recommendation systems.

TEŞEKKÜR

Bu araştırma boyunca beni değerli bilgi ve tecrübesi ile yönlendiren, karşılaştığım zorlukları aşmamda yardımcı olan değerli danışman hocam Dr. Metin TURAN'a teşekkürlerimi sunarım.

Melis ÖZKARA
İSTANBUL, 2023



ŞEKİLLER

	Sayfa
Şekil 3.1. Tavsiye sistemlerini aktif şekilde kullanan şirketler.....	8
Şekil 3.2 İçerik tabanlı tavsiye sistemleri	11
Şekil 3.3 İşbirlikçi tavsiye sistemleri	12
Şekil 3.4 Popülerite tabanlı tavsiye sistemi	13
Şekil 3.5 Hibrit tavsiye sistemleri	14
Şekil 4.1 Uygulamada kullanılacak kategoriler.....	18
Şekil 4.2 Eğitim seti için oluşturulan WordCloud	20



ÇİZELGELER

	Sayfa
Çizelge 4.1 Çalışmada kullanılan Python kütüphaneleri.....	17
Çizelge 4.2 Spor kategorisi için örnek veriler	18
Çizelge 4.3 Tf-Idf vectorizer metodu	21
Çizelge 4.4 Unigram ve Bigram metodu	22
Çizelge 4.5 Sınıflandırma sonuçları	24
Çizelge 5.1 Herhangi bir kullanıcı için ilgili kategori yüzde değerleri	25
Çizelge 5.2 BeautifulSoup kütüphanesi kullanımı	26
Çizelge 6.1 Kullanıcı#1'in tarayıcı geçmişi.....	27
Çizelge 6.2 15 Gerçek kullanıcı için ilk 3 kategori ve eğitim sonucu	28
Çizelge 6.3 Eğitim sonucunun kategorilenmiş hali	28
Çizelge 6.4 Tek tek kayıtların doğruluk başarısı.....	29

SİMGELER VE KISALTMALAR

HTMLKNN	Hypertext Markup LanguageK-Nearest Neighbors
MNB	Multinomial Naive Bayes
RSS	Rich Site Summary
URL	Uniform Resource Locator



1 GİRİŞ

Tavsiye sistemlerinin önemi Spotify, Netflix ve Amazon gibi şirketlerin yükselişiyle birlikte arttı diyebiliriz. Genel olarak e-ticaret, müzik film gibi alanlarla beraber reklam, video, sağlık gibi birçok alanda erişilebilir olan bu sistemler gelecek zamanda da hayatımızı etkilemeye ve şekillendirmeye devam edecek. Bu sistemler kullanan kişinin tercihlerini öngörmeye çalışarak, kişinin eylemi için de çözüm önermeyi sağlar.

Şirketler, bu şekilde çok büyük gelirler elde edebileceklerini bildikleri ve diğer şirketlerin önüne geçip avantaj sağlamak istedikleri için bu sistemlere önem vermektedirler (Dwivedi, 2020). Kullanıcı ve satıcı da doğrudan bu avantajdan etkilenir. Kullanıcı kendisi için özel olarak önerilmiş ürünleri tercih edip satın alabilir. Öte yandan satıcı da satışlarını arttırma fırsatı elde eder ve müşteri memnuniyeti de sağlayarak uzun vadeli bir ilişki kurulması sağlanır (Özkok, 2021).

Tavsiye sistemi türleri aşağıdaki gibi kategorize edilmiştir:

- İçerik Tabanlı Tavsiye Sistemleri
- İşbirlikçi Tavsiye Sistemleri
- Popülerite Tabanlı Tavsiye Sistemleri
- Hibrit Tavsiye Sistemler

İçerik Tabanlı Tavsiye Sistemlerinde, kullanıcının geçmişte satın aldığı veya baktığı ürünlere göre benzerlerini göstermeye çalışır. Kullanıcı hakkında kişisel bir bilgiye ihtiyaç duyulmaz. Öge ve kullanıcı arasında kurulan bütünlüğe duyarlıdır (Beel vd., 2013).

1992 yılında İşbirlikçi tavsiye terimi, “insanlar filtreleme sürecine dahil olduğunda bilgi filtrelemenin daha etkili olabileceği” fikriyle Goldberg ve arkadaşları tarafından önerildi.

İşbirlikçi yöntemde, kullanıcının tercih ettiği ürünler hakkında bilgi toplayıp benzer tercihlerden benzer kullanıcıların da neleri beğenebileceğini tahmin ederek önerilerde bulunan bir sistemdir. Kullanıcıdan kullanıcıya ve öğeden öğeye ortak öneri algoritmaları uygulanabilir. Benzer kullanıcılar tarafından seçilen ürünler kullanıcıya sunulur ve benzer ürünler öğedeki ürün benzerliğine göre görüntülenir.

Popülerite tabanlı sistemde, trend olan ürünler kullanıcıya gösterilir. Hibrit tavsiye sistemi ise işbirlikçi ve içerik tabanlı sistemlerin birleştirilmesiyle oluşturulur. Son yıllarda bilgi teknolojilerindeki gelişmelerle birlikte öneri sistemleri akademik alanda da ilerleme kaydetmiştir. Literatürdeki bilimsel tavsiye sistemi çalışmalarına bakıldığında en çok içerik tabanlı tavsiye türünün uygulandığı görülmektedir. İçerik tabanlı tavsiye türünü, işbirlikçi tavsiye ve karma tavsiye türleri izler.

Modeli oluştururken kullanıcıları uygun şekilde kategorize etmek ve kabul edilebilir başarıyı sağlamak için doğru verileri seçmek ve toplamak çok önemlidir. Sonuç olarak modelin kullanıcının ilgi alanlarını ve sevdiği şeyleri öğrenmesi ve bu yönde makul bir sonuç vermesi beklenir.

Bu çalışmada, eğitimli haber kategorileri modeli kullanılarak kullanıcının tarayıcı geçmişi, yer imleri ve Google aramalarından oluşan veriler, ilgi alanları belirlenmiş ve RSS haberlerinden gündem içerikleri seçilerek ilgili kategorilere göre haber tavsiye sistemi oluşturulmuştur. Bu çalışma, içerik tabanlı bir tavsiye sistemi olmanın ötesinde, kullanıcının ilgilendiği kategorilerin ve gündem haberlerinin ağırlıklarının belirlenmesi, güncel ve tercih edilen haberlerin tavsiye sisteminde seçilmesi açısından literatüre farklı bir bakış açısı getirmektedir. Bildiğimiz kadarıyla yerel bilgisayar bilgilerini haber tavsiye sistemi için kullanan ilk girişimdir.

7 farklı kategoride 4900 haber makalesi ön işleme tabi tutulmuş ve test verileri elde edilmiştir. Daha sonra kullanıcı bilgisayarından toplanan verilerden elde

edilen kelime torbası kullanılarak, eşik değerini aşanlar eğitilen model kullanılarak kullanıcının tercih kategorileri belirlenmiştir.

Son olarak gündeme konu olan haberler arasından (etiketlerine bakılarak) kullanıcının tercih kategorilerine uygun olanlar öncelik sırasına göre kullanıcıya sunulmaktadır.

Bu çalışma 7 bölümden oluşmaktadır. 2. bölümde literatürde yapılmış benzer çalışmalara, 3. bölümde tavsiye sistemlerinde veri toplama yaklaşımları ve kullanılan yöntemlere değinilmiştir. 4. bölümde projede kullanılan materyaller ve metotlar açıklanmaktadır. 5.bölümde projeyle alakalı bulgulara değinilmiştir. 6. bölümde uygulanan projenin sonuçları irdelenmektedir. 7. bölümde ise nihai sonuç değerlendirilmiş ve çalışmanın geliştirilmesi ile ilgili tavsiyelere yer verilmiştir.

2 LİTERATÜR ÖZETİ

Tavsiye sistemleri ile ilgili birçok çalışma bulunmaktadır. Jonnalagedda ve arkadaşları haber makalelerini oluş sırasına göre sunmak yerine, kullanıcının ilgi alanlarına göre sunmayı amaçlamıştır. Kullanıcılar ilgi alanlarına ve haberlere göre profiller oluşturmuşlardır. “Twitter” yardımıyla kişiselleştirilmiş bir haber öneri sistemi geliştirme araştırmasında, makaleler Twitter'ın genel zaman çizelgesinden tespit edilen makalenin popülerliğine göre sıralandı ve sonuçta hibrit yapılı bir sistem ortaya çıktı. Sonuç olarak popülerlik temelli sistemin katkı sağladığı gözlemlenmiştir. Genel olarak, $\alpha=0,5$ olan hybrid2, hybrid1'den %7,8, kişiselden %2,4 ve popülerlikten %6,0 daha iyi performans göstermiştir (Jonnalagedda vd., 2016).

Benzer araştırmalara bir başka örnek de Seven ve Alpkoçak tarafından yürütülen metin tabanlı haber tavsiye sistemidir. KNN (K-Nearest Neighbors) yöntemi kullanılarak oluşturulan algoritma, tüm metin arama mimarisinde kullanılmıştır. Ortak tavsiye sistemi kullanıcıların eşleşme puanının 10 üzerinden 6 olduğu, anahtar kelime tabanlı bir sistem kullanılarak geliştirildi (Seven ve Alpkoçak, 2020).

Taşçı (Taşçı, 2015), çalışmasında içerik tabanlı bir tavsiye sistemi oluşturmuş ve kullanıcıya önerilecek öğelerin ilişkisine göre nesne-kullanıcı matrisi kullanan bir öneri sistemi önermiştir. Sonuçların değerlendirilmesi için sistem için uygulanan süre uzun tutulmuş ve kullanıcılara geri bildirim sağlanmıştır. Elde edilen sonuçlara göre haber önerirken de farklılıkları dikkate almalıyız. Anahtar kelimeyi de dikkate almak gerekir. Bir nesne-güven ilişkisi kurulmalıdır. Makalede kaynak güven değerinin kullanıcılardan 60 olarak alınabileceği belirtilmektedir. Doğru haber kaynaklarının belirlenmesi ve kullanıcı güven durumunun belirlenmesi gerektiği sonucuna varmıştır.

Liu ve arkadaşları, insanların haberleri okuduğu süreyi ve haber metninin uzunluğunu dikkate alarak farklı bir bakış açısıyla zamana dayalı bir yöntem uygulamışlardır. 10.000 kullanıcı yaklaşık olarak aynı sayıda kontrol ve test

grubuna rastgele atandı. Kontrol grubundaki kullanıcılar işbirlikçi yöntemi, test grubundaki kullanıcılar ise hibrit yöntemi denemiştir. Bu çalışmada Google News kullanılmıştır. Araştırma sonucunda test grubundaki kullanıcıların, kontrol grubuna göre Google Haberleri %14,1 daha fazla ziyaret ettiği görülmüştür (Liu vd., 2010).

Pazzani, çalışmasında hibrit bir model kullanmıştır. Modelde yer alan kullanıcılar tarafından okunan haberlere göre kısa süreli ve uzun süreli olmak üzere iki farklı algoritma birleştirilerek sonuçlar elde edilmiştir. Kısa dönemli algoritmada KNN, uzun dönemli algoritmada ise Naive Bayes kullanılmıştır. Bu iki algoritmayla beraber kısa vadede hızlı reaksiyon verilmesi sağlanmıştır. Kısa vadede anlık eğilimlere göre haber tavsiye edilmesi sağlanmış, uzun vadede büyük çaplı eğilimlere göre haber tavsiyesinde bulunulmuştur. 10 kullanıcı üzerinden sistem 4 ila 8 gün boyunca kullanıcı başına 300 haber üzerinden eğitildi. Sonuca göre hibritsistem için doğruluk puanı 75-80, F1 oranı ise 55-60 oldu. Bu iki algoritma karşılaştırıldığında kullanıcıların kısa vadede daha hızlı tepki verdiği gözlemlendi (Pazzani, 1999).

Li ve arkadaşlarının çalışması, iki aşamalı kişiselleştirilmiş bir haber sistemi olan SCENE'yi (SCalable two-stage pErsonalized News rEcommendation) önermektedir. Sistem üç ana unsurdan oluşmaktadır. Bu ögeler haber kümeleme, kullanıcı profillemesi ve haber sıralamadır. Kümeleme için Yerelliğe Duyarlı Karma (LSH) ve hiyerarşik kümeleme kullandılar. Sistemi test etmek için spor, siyaset ve film olmak üzere 9 farklı kategoride haberler toplandı. Veriler kullanıcılara sunulmadan önce ön işleme tabi tutulmuştur. Ön işlemeden sonra 4.630 kullanıcı ile her gün ortalama 1.221 haber depolanmıştır. Karşılaştırma için Goo, ClickB, Bilinear ve Bandit yöntemlerini seçtiler. Elde edilen sonuca göre SCENE 0.6930 ile aralarında en başarılı yöntem olarak bulunmuştur (Li vd., 2011). Saranya ve Sudha tarafından yürütülen bir başka çalışmada çeşitli haber sitelerinden elde edilen sporla ilgili haberlerin bir derlemesi üzerinde yapılan deneyde, haber ajanslarından otomatik olarak veriler elde edilmiştir. Alınan veriler veri tabanına yazılmaktadır. Sınıflandırma haber ajansları tarafından yapılmıştır. Çalışmada iki tür kullanıcı profili

oluşturulmuştur. Biri statik kullanıcı profili diğeri ise dinamik kullanıcı profilidir. Kayıt sırasında kullanıcıdan toplanan bilgiler kullanılarak statik kullanıcı profilleri oluşturulurken, sıklıkla değişen kullanıcı ilgi alanlarına yönelik dinamik kullanıcı profilleri kullanılmaktadır. Spor kategorisi için 500 adet haber kullanılmıştır. Bu kategori altında kriket, tenis, futbol, atletizm ve hokey gibi alt kategorilere ait 25 farklı sorgunun sonuçları elde edilmiştir. The Hindu, Times of India, Indian Express, India Times gibi haber siteleri ile önerilen sistem karşılaştırıldı. Sonuç olarak önerilen sistem 0,9-0,95 oranıyla başarılı olmuştur. Deneysel sonuçlardan önerilen yöntemin iyi bir öneri doğruluğu sağladığı gözlemlenmiştir. Önerilen yöntem önerilen öğelerin sayısı az olsa bile, kullanıcı tercihlerine daha uygun içerik sağlayabilmektedir (Saranya ve Sudha, 2012).

Tan ve Teo, Kişiselleştirilmiş Bilgi Ağı (PIN) adını verdikleri sisteme kullanıcılardan ilgilendikleri kelimeleri girmelerini istedi ve sistem girilen kelimeye göre en uygun tavsiyeyi vermeye çalıştı. World Wide Web'de haber aramak ve almak için bir erişim aracı, profesyonel öğrenme ve bilgi filtreleme için kişisel bir öğrenme aracı ve kişiselleştirilmiş bir haber tarayıcısı içeren PIN, üç ana alt sistemden oluşmaktadır. Yirmi iki günlük öğrenme ve tahmin süresi boyunca deneysel sonuçlara bakıldığında, eğitim MSE (Mean squared error) her zaman test MSE'sinden daha küçüktür ve genellikle sonuç sıfırdır. Bu sonuca göre sistem çevrimiçi öğrenmeyi etkin bir şekilde gerçekleştirebilmektedir (Tan ve Teo, 1998).

Dhruv ve arkadaşları seçilen sanatçıyı aynı tarzdaki diğer sanatçılarla andıran melez bir yaklaşım kullanarak kullanıcıların arkadaşları tarafından dinlenen sanatçılar dikkate alınarak geliştirilen bir model oluşturdu.

Bu sistemde UBCF (User based Collaborative Filtering) algoritması, benzerlik matrisleri, içerik tabanlı filtreleme ve hibrit filtreleme kullanılmaktadır. Veri temizleme ve eksik değerlerin doldurulması gibi aşamalar uygulandı. Sonuç olarak, UBCF yöntemini kullanan sistemin doğruluğu %95, hibrit modeli kullanan sistemin doğruluğu ise %69 olmuştur (Dhruv vd., 2019)

Yu ve ark. tarafından 2015 yılında yapılan çalışmada, odak nokta hastalıkların erken teşhisi için bir sistem oluşturmaktır. Sağlık alanındaki verileri güvenli bir ortamda yararlı bilgiler üretmek için kullanmak amaçlanmaktadır. Yapılan uygulamada kullanıcıların sağlık koşullarını belirlemek, tedavi optimizasyonunu sağlamak için Apache Mahout üzerinde çalışan Naive Bayes sınıflandırma algoritması kullanılmıştır. Hastalık teşhisi konduktan sonra doğru bakım verilmesi tedavi maliyetini de optimize etmelidir. Tedaviler erken dönemden itibaren uygulanırsa insanların ortalama yaşam süresi de artabilir (Yu vd., 2015)

2014 yılında Meng ve arkadaşları tarafından yapılan çalışmada KASR (Keyword-Aware Service Recommendation) adı verilen anahtar kelimeye dayalı kişiselleştirilmiş bir tavsiye sistemi geliştirilmiştir. Kullanıcıların tercihleri anahtar kelimeler ile belirtilmiş ve doğru öneriler sunabilmek için işbirlikçi tavsiye sistemi yöntemi kullanılmıştır. KASR sistemi MapReduce paradigması kullanılarak Hadoop platformunda geliştirilmiştir. Geliştirilen bu sistemde iki veri yapısı olarak aday listesi ve uzmanlaşmış alan adı kullanılmıştır. Sistem yakın kullanıcıların incelemelerinden çıkarılan anahtar kelimeler, kullanıcı tercihlerini belirlemek için kullanılmıştır.

Anahtar kelime kümeleri eşleştirilirken aktif kullanıcı ve yakın kullanıcı tercihleri dikkate alınarak sağlanır. Jaccard benzerliği ile aktif kullanıcıya yakın kullanıcılar hesaplanır. Gerçek veri kümeleri üzerinden geniş kapsamlı deneyler gerçekleştirilir ve sonuçlar KASR'nin mevcut yaklaşımlara göre öneri sistemlerinin doğruluğunu ve ölçeklenebilirliğini önemli ölçüde iyileştirdiğini gösterir (Meng vd., 2014)

Literatürde ilgili çalışmalar incelendiğinde halihazırda var olan bir veri seti ile çalışılabildiği gibi özgün veri setleri ile de çalışıldığı görülmüştür. Mevcut araştırma hazır veri seti ile ele alınmıştır. Bununla birlikte araştırma içerik tabanlı bir tavsiye sistemi sunmaktadır.

3 TAVSİYE SİSTEMLERİ

Teknolojinin gelişmesiyle birlikte alışkanlıklarımızın da değiştiği bir gerçektir. Bu gelişmeler neticesinde büyük miktarda veriler oluşmuştur. Tavsiye sistemleri, 1990'ların ortalarında bağımsız araştırmalarla ortaya çıkmış; kullanıcının tercihlerini tahmin etmeye ve kullanıcı eylemi için doğru çözümü önermeye çalışan yazılım ve tekniklerdir. Ne okuyabileceğin, hangi film izleyebileceğin, hangi müziği dinleyebileceğin gibi şeyleri tavsiye sistemi teknikleriyle çözümlmek mümkündür. Özellikle çevrimiçi platformlarda (Amazon, Spotify, Netflix, Google vb.) yoğun olarak kullanıldığı görülmektedir. Şekil 3.1'de tavsiye sistemlerini aktif kullanan şirketler gösterilmiştir.



Şekil 3.1. Tavsiye sistemlerini aktif şekilde kullanan şirketler

Derin öğrenme, veri madenciliği, makine öğrenmesi ve büyük veri tekniklerinden yararlanan tavsiye sistemleri çevrimiçi platformlara ek olarak ürün arama motorları, sağlık uygulamaları ve sosyal ağlarda (Facebook, Twitter, Instagram) kullanılır.

Aşırı fazla bilginin olduğu bu dönemde tavsiye sistemi kendisinin değerli bir araç olduğunu göstermiştir. Kullanıcının gereksiz bilgileri görmemesi ve doğru

seime y6nlendirilmesi 6nemlidir. Bundan dolayı kullanıcıyı fazla bilgiden kurtarıp 6neriye g6re karar vermesi saėlanır.

Tavsiye sistemlerinde 6r6n satışı arttırmak ve buna baėlı kar arttırmak ilk amalardan biridir. Bu ama iin gereken genel teknikler benzerlik, yenilik, keşfetme ve 6neri eşitliliėi artırma diyebiliriz. Bunlar olduėu takdirde kullanıcının beėenme olasılıėı da artacaėı iin bu da kar arttırımını paralel olarak saėlamış olur (Durukan, 2018).

3.1 Tavsiye Sistemlerinde Kullanılan Veri toplama Yaklaşımları

Kullanıcılara 6neri sunulacaėı zaman bu bilginin tahmin ařamasında kiřisel profilini ıkarmak iin kullanıcının eriřtiėi kaynak ierikleri ve davranışları gibi bilgiler kullanılır.

6rnek olarak m6zik uygulamaları hořumuza giden m6ziėi anlamak iin ilk bařta kullanıcıdan bilgi toplayıp kullanıcı profili oluřturulur. Sonrasında sevdiėimiz, sevmediėimiz, en ok dinlediėimiz m6ziklere g6re kullanıcı profili daha da řekillenip doėru m6ziklere ulařmamızı saėlar.

Kullanıcı profilleri tavsiyelerin kalitesini de arttırır. M6mk6n olduėunca kullanıcı hakkında fazla bilgiye sahip olmak en doėru ve beklenen cevabı vermede yardımcı olur.

3.1.1 Dolaylı geribildirimler

Dolaylı geribildirimler kullanıcının yaptıėı farklı eylemlerin takip edilmesiyle elde edilir. Bu eylemler arasında baėlantı gemiři, hangi web sayfasında ne kadar kaldıėı, tıklanan linkler, aray6z 6zerindeki tıklamalar gibi 6rnekler verilebilir. Direkt olarak kullanıcıyla alakalı davranışların analiz edilmesi ve kullanıcı tarafından uėrařtırmadan kolaylıkla elde edilebilmesi bu yaklařımı objektif bir yaklařım olarak ele alınır.

3.1.2 Doğrudan geribildirimler

Doğrudan geribildirimlerde bir kullanıcı arayüzü üzerinden kullanıcının o öge hakkında puan verip değerlendirmesi beklenir. Değerlendirme ve puan verme ne kadar fazla olursa sunulan öneriler de daha doğru bir şekilde kullanıcıya gösterilir.

Kullanıcının çaba gösterip değerlendirmesi ve yetersiz bilgi vermesi yöntemin dezavantajlarından biridir. Elde edilen veriler dolaylı geribildirimde göre daha güvenilir olup şeffaflık sağlar (Utku ve Akçayol).

3.1.3 Hibrit geribildirimler

En iyi performans ve sonuç elde etmek için doğrudan ve dolaylı geribildirim birlikte kullanılarak hibrit geribildirim elde edilir.

Doğrudan geribildirimde kontrol mekanizması ekleyerek veya dolaylı geribildirimde kullanıcıların puanları veya değerlendirmeleri ele alınarak sonuçlanır.

3.2 Tavsiye Sistemlerinde Kullanılan Yöntemler

Tavsiye sistemlerinde en çok kullanılan yöntemler aşağıdaki gibidir.

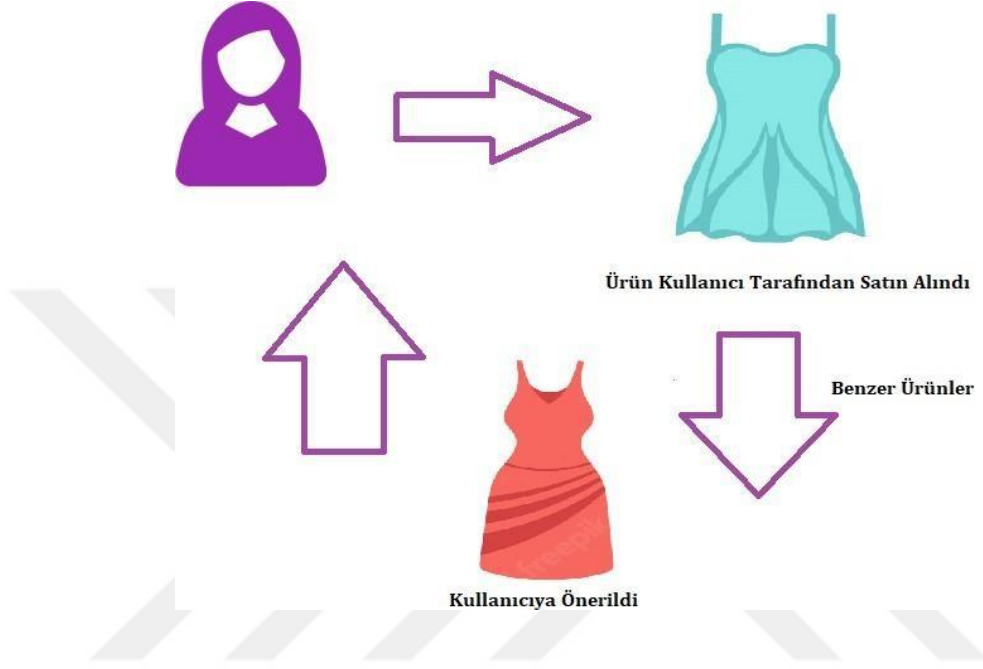
- İçerik Tabanlı Tavsiye Sistemleri
- İşbirlikçi Tavsiye Sistemler
- Popülerite Tabanlı Tavsiye Sistemleri
- Hibrit Tavsiye Sistemler

3.2.1 İçerik tabanlı tavsiye sistemi

İçerik Tabanlı Tavsiye Sistemlerinde kullanıcının geçmişte tercih ettiği ürünlere göre benzer ürünler gösterilmeye çalışılır. Bu yöntemde diğer kullanıcılar

hakkında herhangi bir bilgiye ihtiyacımız yoktur. Öge ve kullanıcı arasında kurulan bütünlüğe çok duyarlıdır (Beel vd., 2013).

Şekil 3.2’de içerik tabanlı bir sistem gösterilmiştir.



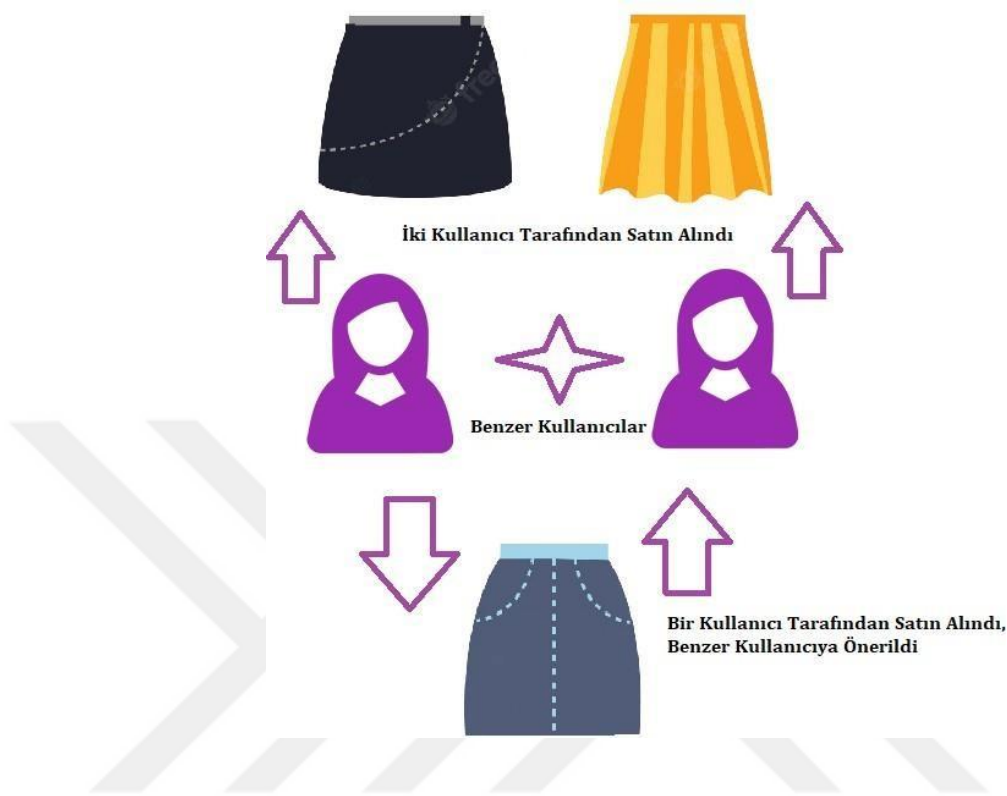
Şekil 3.2 İçerik tabanlı tavsiye sistemleri

Bu yöntemde ürünü tanımlamak için anahtar kelimeler oluşturulur. Algoritmalar sayesinde kullanıcıya benzer ürünler tavsiye etmeye çalışılır. Bu sistemin fikri aslında örneğin bir elbiseyi beğendiğinizde bu o elbiseye benzer başka bir elbiseyi beğeneceğiniz anlamına gelir. İçerik Tabanlı Sistemlerde en büyük sorun ürünlerin içerik kaynağından kullanıcı tercihlerinin öğrenip öğrenilemeyeceğidir.

3.2.2 İşbirlikçi tavsiye sistemi

İşbirlikçi yöntemde kullanıcının tercih ettiği ürünler hakkında bilgi toplayan ve benzer tercihlerden diğer kullanıcıların neleri beğenebileceğini tahmin ederek önerilerde bulunan bir sistemdir. Kullanıcıdan kullanıcıya ve öğeden öğeye ortak öneri algoritmaları uygulanabilir. Benzer kullanıcılar tarafından seçilen

ürünler kullanıcıya sunulur ve benzer ürünler öğedeki ürün benzerliğine göre görüntülenir. Şekil 3.3'te bu yöntem gösterilmiştir.



Şekil 3.3 İşbirlikçi tavsiye sistemleri

Kullanıcı öğe benzerliklerini bulur. Benzer kullanıcıların benzer tercihler yapacağı fikrini benimser. Benzer ürünleri bulmak için Kosinüs Benzerliği ve Pearson Korelasyon Katsayısı gibi teknikler yaygın kullanılır. Model tabanlı düşündüğümüzde Kümeleme, Karar Ağaçları, Yapay Sinir Ağları gibi tekniklerden yararlanır.

Bu sistem kullanıcı ve öğe bazlı olarak ayrılır:

- Kullanıcı bazlı filtreleme: Bu filtrelemede kullanıcının beğendiklerinin ona göre benzerliğini dikkate alan bir yöntemdir. Örneğin iki kullanıcı aynı elbiseyi almışsa bu elbiseyi diğer kullanıcının da satın alma olasılığına dayanır. Pearson Korelasyonu ya da Kosinüs Benzerliği yöntemleri sonucunda kullanıcıların puanlarına göre bu yöntem hesaplanır.

- Öğe bazlı filtreleme: Bu filtreleme algoritmasında ise kullanıcı skorları üzerinden gelen ürünlerin benzerlikleri bulunur. Tıklama ve görüntüleme kullanıcı skor yerine kullanılabilir. Kullanıcıların görüntülediği aldığı veya tıkladığı ürünün benzerliğini analiz edip diğer benzer kullanıcılara da önerilmesini sağlar. Bu yöntemde de Pearson Korelasyonu ya da Kosinüs Benzerliği yöntemleriyle ürün benzerliklerinin bulunması sağlanmış olur.

3.2.3 Popülerite tabanlı tavsiye sistemi

Popülerite tabanlı sistemlerde en trend olan ürünler kullanıcıya gösterilir. Örneğin X uygulaması yeni kaydolan bir kullanıcıya daha önce yeni kaydolup o kullanıcıların aldığı ürünleri önerebilir. Şekil 3.4'te bu yöntem gösterilmiştir.

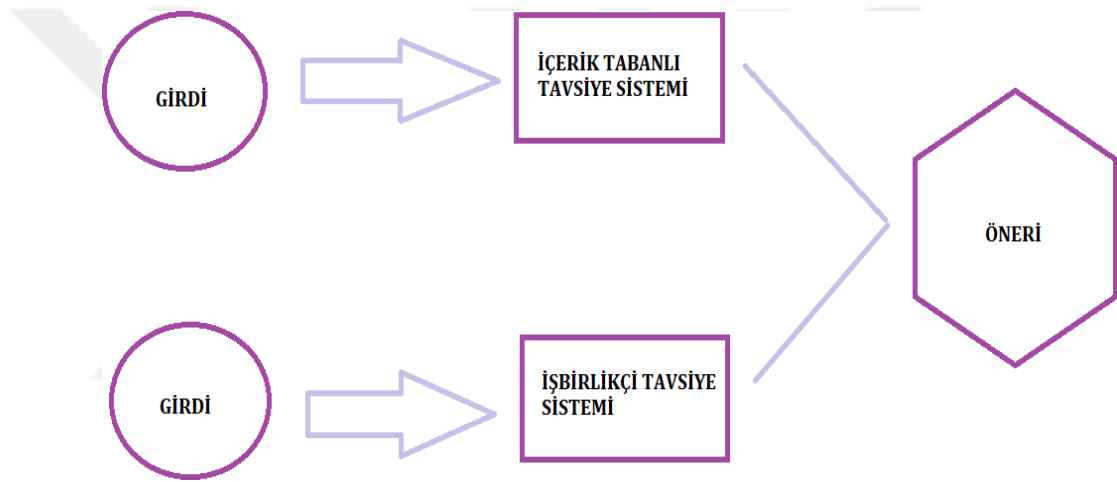


Şekil 3.4 Popülerite tabanlı tavsiye sistemi

Bu sistem bazı sorunları da beraberinde getirir. Bu sistemde en büyük sorun veriler kişiselleştirme için kullanılmaz. Bundan dolayı da kullanıcının davranışı veya ne sevip sevmediği bilinse bile tavsiye edilen ürünler bu değerlendirmelere göre önerilmez.

3.2.4 Hibrit tavsiye sistemi

Hibrit tavsiye sistemi adından da anlaşılacağı üzere iki sistemin birleştirilmesiyle oluşan hibrit bir yöntemdir. İki sistemin de dezavantajlarını en aza indirerek en iyi performansı elde etmek amaçlanmıştır. Bu yöntem ayrı ayrı yapılarak uygulandığı gibi işbirlikçi yaklaşıma içerik tabanlı yöntemi ekleyerek ya da içerik tabanlı yöntemle işbirlikçi özellikler ekleyerek tek model üzerinden yürütülebilir. Veri azlığında üstesinden gelebilecek bir sistem olarak düşünülebilir. Şekil 3.5'te bu yöntem gösterilmiştir.



Şekil 3.5 Hibrit tavsiye sistemleri

Hibrit tabanlı sisteme verilecek en iyi örnek Spotify ve Netflix'tir. Spotify 'da belirli aralıklarla öneri listeleri çıkartılır. Kullanıcının dinlediği ve o kullanıcıya benzer kullanıcıların dinledikleri şarkıları bir araya getirerek en iyi tavsiyeyi vermeyi amaçlamaktadır.

Aynı şekilde Netflix'te kullanıcılara benzer kullanıcıların izleme geçmişini tıklamasını ve görüntülemesini karşılaştırarak kullanıcın yüksek skor verdiği içeriklere benzer olanları diğer kullanıcılara önerir.

3.3 Tavsiye Sistemlerinde Karşılaşılan Zorluklar

Tavsiye sistemlerin avantajları olduğu gibi birtakım zorlukları da beraberinde getirir.

Soğuk Başlangıç problemi yeni bir kullanıcıya tavsiye sunarken yaşanır. Kullanıcıyla alakalı bilgi ve veri olmadığından doğru tavsiyelerde bulunmak pek mümkün olmaz. Bu sorunun çözümü için popülerite tabanlı sistem kullanılmasının yanı sıra başlangıçta kullanıcıdan sevdiği şeyleri seçmesini isteyen bir yöntem de izlenebilir. Örneğin Spotify 'da ilk kayıt açtıktan sonra sevdiğiniz müzik türü ve birkaç şarkı seçmeniz istenir.

Seyreklik probleminde kullanıcının aldığı ve kullandığı ürünlere geri bildirim vermemesi ve değerlendirmemesi düşünülebilir. Bu da kullanıcı-öge analizini seyrek hale getireceğinden yapılan tavsiyelerdeki başarı oranını azaltır.

Gizlilik sorunu önemli bir sorundur. Tavsiye sistemlerinde en önemli şey veriler olduğu için kullanıcıların gizlilik ve güvenliklerine dair tereddüt yaşayabilmeleri çok normaldir.

3.4 Haber Tavsiye Sistemi

Her sistemin kendine has özellikleri vardır. Haber tavsiye sistemleri anlık değişebilen bir sistemken film müzik gibi tavsiye sistemlerinin anlık değişmesi çok önemli değildir. Haber tavsiye sistemlerinin bazı özellikleri aşağıdaki gibidir (Saranya ve Sudha, 2012);

- Haber tavsiye sistemi kısa ömürlüdür; aynı zamanda çok hesaplama gerektirir.
- Her haber sitesinin haber formatı aynı değildir.
- Kullanıcıların haber türlerine ilgisi diğer müzik film gibi alanlara göre daha kısa sürede değişiklik gösterir.
- Haber makalesi fazla olduğundan bu haberleri kategorilere ayırmak karmaşık algoritmalar ve hesaplamalar getirir.

4 MATERYAL VE YÖNTEM

Çalışmada gerek veri seti gerekse kullanıcılardan alınan veriler ve RSS'den alınan haberler üzerinde sırasıyla belli başlı adımlar yapılmıştır. Bu adımlar sayesinde verilerin daha doğru okunması ve daha iyi bir sonuç alınması amaçlanmaktadır. Çalışmada yapılan işlem adımları sırasıyla ve genel olarak aşağıdaki gibidir;

- Arama geçmişinden verileri alma
- Alınan verilerden URL yol adresini ve başlıklarını çıkartma
- Gereksiz kelimeleri (stop-words) kaldırma
- Noktalama işaretleri ve sayıların temizlenmesi
- Eğitimde kullanılacak veri kümesi için;
 - Verinin temizlenmesi
 - Veri normalizasyonu
 - Sözcüklere ayırma ve kökleri bulma (tokenization ve lemmazation)
 - Terim ağırlıklandırma (term frequency)
- Sınıflandırma
- Sonuca göre haber kategorilerinin yüzdelik değerini belirleme
- Kategorilere uygun haberlerin RSS'den çekilmesi
- Gündeme uygun haberlerin öncelik sırasına konulması ve kullanıcı tercihağırlığına göre sunulması

4.1 Kullanılan Araçlar ve Kitaplıklar

Bu çalışmada aşağıdaki sebeplerden ötürü geliştirme ortamı olarak PyCharm ve programlama dili olarak Python tercih edilmiştir.

- Python, NLP için daha kapsamlı kütüphaneler içerir ve C-tabanlısözdizimi kodlamayı kolaylaştırır.
- PyCharm, Python programlama dili için kullanıcı dostu geliştirmeortamıdır.

Çalışmada browserhistory, feedparser, numpy, pandas, nltk, sklearn, urlparse, matplotlib, seaborn, wordcloud, request, beautifulsoup kütüphaneleri kullanılmıştır. Kütüphane açıklamaları kısaca Çizelge 4.1'de verilmiştir.

Çizelge 4.1 Çalışmada kullanılan Python kütüphaneleri

Browser History	Chrome'dan verileri almak için kullanılmıştır.
Feedparser	RSS'den gelen verileri ayrıştırmak amacıyla kullanılmıştır.
Numpy	Çalışmadaki eğitim ve test verilerini düzenlemek için kullanılmıştır.
Pandas	CSV formatındaki verilerin işlenmesinde kullanılmıştır.
Nltk	İnsan dilleri üzerinde çalışmak için oluşturulan açık kaynak kodlu bir kütüphanedir.
Urlparse	Chrome'dan alınan verilerin Url kısmını ayrıştırmak amacıyla kullanılmıştır.
Matplotlib	Verileri görselleştirmek için kullanılmıştır.
Seaborn	İstatiksel tabanlı verileri görselleştirmek için kullanılmıştır.
Wordcloud	Sık kullanılan verileri görsel olarak göstermek için kullanılmıştır.
Requests	Gündem etiketlerini almadan önce Url sayfasını çağırmak için kullanılmıştır.
BeautifulSoup	Sayfanın HTML (Hypertext Markup Language) etiketleri arasında verileri almak için kullanılmıştır.

4.2 Eğitim Veri Seti

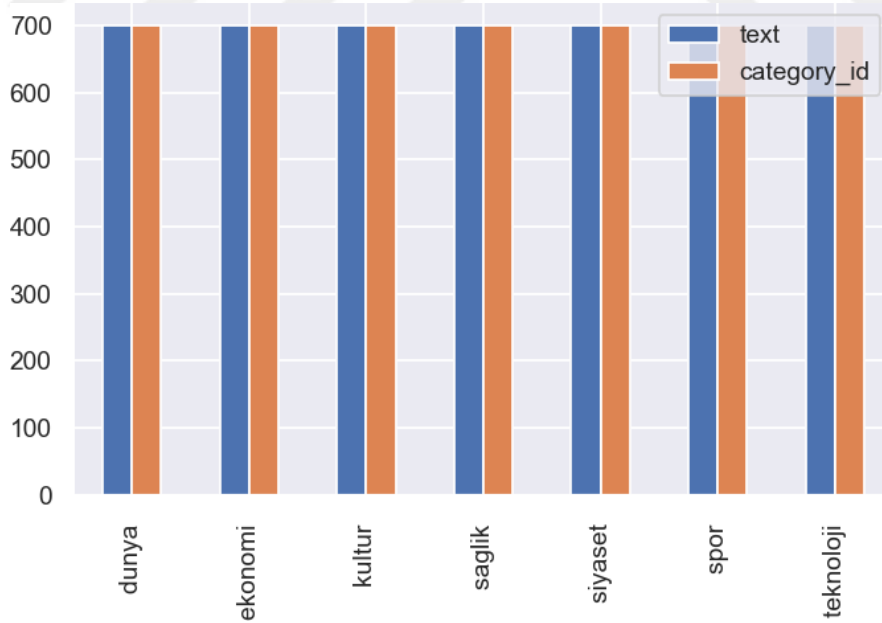
Veri seti iki sütundan oluşmaktadır. Birincisi kategori başlığını (sınıf bilgisi) ifade ederken, ikinci sütun metin başlığı altındaki verileri içermektedir. Çizelge 4.2'de spor kategorisindeki örnek verileri göstermektedir.

Çizelge 4.2 Spor kategorisi için örnek veriler

spor	oklahoma harden ın boşluğunu dolduramadı nba de geçtiğimiz sezonun finalistlerinden olan Oklahoma city thunder yaz sezonunda takımdan ayrılarak Houston rockets a takas olan James harden yokluğunu hissediyor
spor	masa tenisinde 3 etap izmir de masa tenisinde süperlig 3 etap karşılaşmalarının izmir de yapılacağı bildirildi masa tenisi federasyonundan yapılan açıklamaya göre müsabakalar 10 11 kasım tarihlerinde fuar celal atik spor salonu nda gerçekleştirilecek

Haber metinleri 7 kategoriye ayrılmıştır: spor, dünya, ekonomi, kültür, teknoloji, siyaset ve sağlık. Her kategoride eşit olarak dağılmış 700 adet veri bulunmaktadır (Yıldırım, 2017). Eğitilecek veri seti için; Pandas kullanılarak her kategori adı sayısal bir dizinle eşleştirilir ve kategori değerlerine göre sıralanır. Kategori id ve kategoriler aşağıdaki gibidir.

0: Politika, 1: Dünya, 2: Ekonomi, 3: Kültür, 4: Sağlık, 5: Spor, 6: Teknoloji



Şekil 4.1 Uygulamada kullanılacak kategoriler

4.3 Veri Ön İşleme

Veri seçimi yapıldıktan sonra yapılacak ilk şey veri ön işleme adımlarını uygulamaktır. Veriler işlenmeye hazır forma dönüştürülür. Ön işleme adımları çalışmanın sonucunda elde edilecek başarıyı doğrudan etkilemektedir. Böylelikle ön işleme aşamasının başarılı olması doğru ve kesin sonuca ulaşmaya imkân sağlamış olur (Olson vd., 2008). Çalışmada yapılan ön işleme adımları kısaca şu şekildedir:

- 1) Tokenization: Metni istenilen şekilde kelime kelime parçalayarak dizilerek kaydetme işlemidir.
- 2) Removing Stop Words: Metin içerisinde geçen anlamda herhangi bir değişiklik yapmayan kelimelerin atılması işlemidir.
- 3) Removing punctuation and digits: Metin içerisindeki noktalama işaretleri ve sayıların çıkartılması işlemidir.
- 4) Normalization: Metindeki büyük küçük harf ayrımının ortadan kaldırılması işlemidir.
- 5) Stemming: Metinde geçen kelimelerdeki eklerin atılarak kelime köklerinin kaydedilmesi işlemidir.

4.4 Tarayıcı Geçmişi ile Veri Çekme

BrowserHistory kullanıcının yerel bilgisayarından tarayıcı geçmişini çıkaran ve CSV dosyalarına veri yazan Linux, MacOS, Windows platformları için kullanılan bir kütüphanedir. Tarayıcı açısından Firefox, Google Chrome ve Safari'yi destekleyen basit bir Python modülüdür. URL bağlantısı, sayfanın adını ve tarihini içerir.

Kullanıcının kullandığı bilgisayardaki geçmiş veriler alınır ve CSV formatında kaydedilir. Sözde kod bloğu aşağıda verilmiştir. BrowserHistory ile her bir site bağlantısı, yer işaretleri ve aranan kelimeler için tarayıcıdan alınan veriler tahmin edilerek kategoriye eklenmiştir (Pypi, Browser History).

4.6 Terim Ağırlıklandırma

Veriler artık frekans için analiz edilmeye hazırdır. Ardından Python'da CountVectorize metodu kullanılarak bir metin belgesi terim sayısı vektörüne dönüştürülür (Uslu ve Akyol, 2021).

Terim ağırlıklandırma, terimin verilerdeki önemini belirtmek için ağırlık atama işlemidir. Bu çalışmada problemin doğasına uygun görünen TF-IDF terim ağırlıklandırma yöntemi kullanılmıştır. TF-IDF ile her kategoriye özel kelimeler bulunur. TF-IDF terimin verilerdeki önemini gösteren ağırlık faktörüdür (Kumaş, 2021).

TF, bir belgede o terimin kaç kez geçtiğini hesaplar. IDF ise bu terimin metinlerde kaç kez geçtiği bilgisini vermektedir. Bu iki değer çarpılarak TF-IDF değeri elde edilir. Bu çalışmada terim olarak veri setinde yer alan haberde geçen kelimeler ve belge olarak haber içeriği kullanılmıştır. TF-IDF değeri, tfidfvectorizer sınıfı kullanılarak bulundu. TF-IDF ile her kategoride belirlenen önemli kelimeler bulunmaktadır. Tfidfvectorizerın kullanıldığı sözde kod bloğu aşağıdaki gibidir.

Çizelge 4.3 Tf-Idf vectorizer metodu

Adım 1: TfidfVectorizer methodunu oluştur
Adım 2: Subliner_tf'yi True olarak ayarla
Adım 3: Min_df'yi 5 olarak ayarla
Adım 4: Normunu 12 olarak ayarla
Adım 5: 1-2 olacak şekilde ngramları ayarla
Adım 6: Stop words listesini methoda ekle
Adım 7: Sonucu tranform et

TF-IDF vectorizer işleminde Sublinear_tf, min_df, norm, tokenizer, encoding, ngram_range ve stop_words parametreleri kullanılmaktadır. Burada sublinear_tf parametresi seçilerek sublinear tf ölçekleme (bir belgedeki bir terimin öneminin normalize edilmesi için logaritmik tabanlı frekans dönüşümü) uygulanır. Min_df parametresi ise kelime dağarcığı oluşturulurken verilen

eşikten kesinlikle daha düşük bir belge frekansına sahip terimlerin göz ardı edilmesine izin verir.

Norm parametresi ile her çıkış satırı l1 ve l2 formlarından birine sahiptir. Tokenizer parametresi, ön işleme ve n-gram oluşturma adımlarını korur.

Kodlama parametresi, verilerdeki karakterler için algılama standardını belirler. ngram_range parametresi ile min ve max olarak verilen ngram aralığını hesaplamak mümkündür. (1,2) ile modelde hem 1 gram (unigram) hem de 2 gram (bigram) kullanılır (eğitim için kullanılan veri boyutu küçük olduğundan daha büyük gramlar iyi çalışmayacaktır).

Stop_words parametresi, verilerden anlamsız ve değeri olmayan kelimeleri çıkarmak içindir (Çelik ve Koç, 2021).

Sklearn kütüphanesindeki chi2 sınıfı kütüphane olarak eklenerek her haber kategorisi için 1-gram ve 2-gram değerleri elde edilmiştir. Bu bölümün sözde kod bloğu aşağıdaki gibidir.

Çizelge 4.4 Unigram ve Bigram metodu

Adım 1: Her haber kategorisi için
Adım 2: Chi2 kitaplığını kullan
Adım 3: Bigram ve unigram kelimeleri bulun
Adım 4: Yazdır

Her haber kategorisi için en sık kullanılan kelimeler ve ifadeler listelenir. Bu ifadeler azalan sıklık sırasına göre listelenmiştir. Listede yüksek frekans değerine sahip olanlar uygun eş dizimler olarak kabul edilir. Örneğin dünya ve ekonomi kategorileri için unigram ve bigram çıktıları aşağıdaki gibidir.

#'Dünya':

#'Ekonomi':

-En ilişkili unigram ifadeler:
Suriye
İsrail

- En ilişkili unigram ifadeler:
yatırım
yüzde

-En ilişkili bigram ifadeler:
kişi hayatını
kişi öldü

- En ilişkili bigram ifadeler:
endeksi yüzde
yılın aynı

4.7 Model Eğitimi

Verileri analiz ederken hangi öğrenme modelinin en doğru sonucu vereceğine karar vermek çok önemlidir. Bu çalışmada MNB, GradientBoosting, AdaBoost ve Decision Tree algoritmaları aynı veri seti için ayrı ayrı denenmiş (en uygun algoritma olarak seçilmiş) ve sınıflandırma yapılmıştır. GradientBoosting başarı açısından iyi bir sonuç vermesine rağmen en iyi sınıflandırma ölçütleri MNB ile elde edilmiştir. Sonuç olarak testler MNB üzerinden gerçekleştirilmiştir. MNB, olasılık ve istatistik tabanlı bir Bayes sınıflandırmasıdır (Kaşıkçı ve Gökçen, 2014). Formül 1, çok sınıflı Bayes modelinin formülünü gösterir. Burada c sınıfları (kategoriler), X ise sınıfının bulunması istenen örneği ifade eder. Küçük x 'ler öznitelik vektörleridir ve bu problemde her kategori için belirlenmiş önemli kelimeleri içerir.

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c) \quad (4.1)$$

1-gram ve 2-gram ilişkilendirmelerden sonra TF-IDF dönüşümü ve CountVectorizer dönüşümü yapılmıştır. CountVectorizer, metin verilerini bir terim/sembol sayım vektörüne önceden işler. TF-IDF değerlerine göre vektörleştirilmiş metinler %25 test ve %75 eğitim verisine bölünerek MNB sınıflandırıcısına verilmiştir. Böylece eğitim tamamlanmış olur.

Eđitilen veriler MNB, Gradient Boosting, AdaBoost ve Decision Tree algoritmaları ile ayrı ayrı deęerlendirilmiřtir. Sonular izelge 4.3'te verilmiřtir.

izelge 4.5 Sınıflandırma sonuları

	Test Accuracy	Precision	Recall	F1
MNB ()	86.07	0.86	0.86	0.86
DecisionTreeClassifier ()	63.74	0.63	0.63	0.63
GradientBoostingClassifier ()	82.28	0.82	0.82	0.82
AdaBoostClassifier ()	59.36	0.59	0.59	0.59

5 BULGULAR

5.1 Anlık Haberleri Gösterme

Bu çalışmada, her kategori için yüzde değerleri bulunduktan sonra, bu değerlerden eşik değerini aşan haber türleri için RSS'den anlık veri çıkarmak için FeedParser kitaplığı kullanılmıştır. Eşik değeri ortalamayı ifade eden tüm kategorilerin yüzde sayısının yediye bölünmesiyle elde edilen değerdir.

Örnek bir kullanıcı ortamından elde edilen veriler kullanılarak kategorilere göre tahmin edildikten sonra kullanıcının ilgilendiği kategorilerin yüzde değerleri hesaplanarak Çizelge 5.1'de gösterilmiştir. Belirli bir eşik değerinin üzerindeki kategoriler tercih olarak seçilir. Bu çalışmada haber sunumunda yüzde değeri en yüksek olan ilk 3 kategori dikkate alınmıştır.

Çizelge 5.1 Herhangi bir kullanıcı için ilgili kategori yüzde değerleri

	Yüzde Değerleri
Ekonomi	33.333
Teknoloji	23.404
Politika	14.184
Spor	13.475
Dünya	5.673
Kültür	4.964
Sağlık	4.964

Feedparser kitaplığı, haberleri anında RSS'den alır; haberin başlığını, bağlantısını ve açıklamasını yakalamak için bir kütüphanedir (Pypi, Feed Parser).

Bu haberlerin önem sırasını belirlemek için istekler modülü ile istenilen HTML sayfası elde edilmiştir. Request modülü; web sayfası için al, gönder, koy, sil isteklerini yapan bir modüldür.

Web sayfası için bir alma isteği yapıldıktan sonra BeautifulSoup kütüphanesi kullanıldı. Bu kitaplık bir sayfanın HTML etiketlerini ayrıştırmak ve bu etiketler arasındaki verileri almak için kullanılır (Erdinç, 2017). Uygulamanın sözde kod örneği aşağıda verilmiştir.

Çizelge 5.2 BeautifulSoup kütüphanesi kullanımı

Adım 1: Url sayfasını al ve response değişkenine ekle
Adım 2: Response değişkeninin content içeriğini al
Adım 3: BeautifulSoup kütüphanesini kullanarak html parser ile içeriği ayrıştır ve soup değişkenine ekle
Adım 4: Soup değişkenine find all kullanarak içinde htmden alacağın tagları result değişkenine ekle
Adım 5: Sonucu yazdır

Haber sitelerinde gündem sayfasından gündemdeki önemli etiketler çekilmekte ve bu kütüphane sayesinde haberlerin öncelik sırası belirlenmektedir. Kullanıcının kategori sonucuna göre o haber başlığında gündemde olan kelimenin olup olmadığı kontrol edilir. Gündem etiketli haberler en üst satırda kullanıcıya gösterilir. Örneğin 18/03/2022 tarihli RSS ajanda etiketlerinin en basit hali aşağıdaki gibidir.

Gündem Etiketleri: Ukrayna, hava durumu, sadık çiftpınar, putin, 1915 Çanakkale

Gündemdeki sözler elde edildikten sonra RSS'den gelen anlık haber başlığında, gündemde ve kullanıcının ilgili kategorisine ait eşik değer üzerinde herhangi bir kelime olup olmadığı kontrol edildi. Son olarak haberler öncelik sırasına konularak kullanıcıya sunulmuştur.

6 ARAŞTIRMA SONUÇLARI

15 gerçek kullanıcı ile yapılan testlerde, kullanıcıya ilk olarak 7 kategoriden en çok ilgi duyduğu üç kategori sorulmuştur. Bir sonraki aşamada kendi kişisel bilgisayarlarından kullanıcının geçmiş verileri toplanır ve yukarıda açıklandığı şekilde işlenir. Çizelge 6.1'de, Kullanıcı#1'in tarayıcı geçmişinden toplanan verilerden bir kısım sunulmaktadır.

Çizelge 6.1 Kullanıcı#1'in tarayıcı geçmişi

	Tarayıcı Geçmişi
Kullanıcı#1	x"https://www.haberturk.com/borsa-gune-yukselisle-basladi-3379626-ekonomi","Borsa güne yükselişle başladı"
Kullanıcı#1	"https://www.dunya.com/finans/haberler/rusya-mb-yaptirimlara-ragmen-faizi-yuzde-20de-sabit-tuttu-haberi-652356","Rusya MB, yaptırımlara rağmen faizi yüzde 20'de sabittuttu - Dünya Gazetesi"
Kullanıcı#1	"https://www.webtekno.com/dunyada-yilin-otomobili-odulleri-2022-finalistler-h121852.html","Dünyada Yılın Otomobili Ödülleri Finalistleri Açıklandı"
Kullanıcı#1	"https://www.cnnturk.com/spor/futbol/bulent-korkmaz-kaybetsek-kirilma-macimiz-olurdu","Bülent Korkmaz: Kaybetsek kırılma maçımız olurdu - Son Dakika Futbol Haberi"
Kullanıcı#1	"https://www.bloomberght.com/abd-nin-kripto-adimi-bitcoin-i-vurabilir-2301609","""ABD'nin kripto adımı Bitcoin'i vurabilir"" - Bloomberg HT"

Eğitilen modelin ilk 3 kategori çıktısı Çizelge 6.2'de gösterilmektedir. Eşleşen kategorilerin sayısının artması bize daha iyi bir model vermektedir. Modelin doğruluğu eğitilen model kategorilerinin kullanıcı tercih kategorileri ile eşleşme yüzdesi açısından değerlendirilmektedir. Eğitilen modelin kullanıcının öngördüğü kategorilerle genel test sonucu kategori bazında %89 oranında eşleştiği (başarılı olduğu) görülmektedir (doğru tahmin edilen kategori sayısının toplam kategori sayısına oranı, 40/45).

Çizelge 6.2 15 Gerçek kullanıcı için ilk 3 kategori ve eğitim sonucu

	Kullanıcı Kategorileri	Eğitilmiş Model Çıktısı
Kullanıcı#1	Ekonomi, Dünya, Teknoloji	Ekonomi, Teknoloji,Dünya
Kullanıcı #2	Dünya,Spor, Politika	Spor, Teknoloji, Politika
Kullanıcı #3	Teknoloji,Spor, Kültür	Teknoloji,Spor, Kültür
Kullanıcı #4	Sağlık,Kültür, Teknoloji	Sağlık, Teknoloji, Kültür
Kullanıcı #5	Teknoloji,Ekonomi,Sağlık	Ekonomi,Teknoloji,Sağlık
Kullanıcı #6	Ekonomi,Politika, Sağlık	Politika, Teknoloji, Sağlık
Kullanıcı #7	Politika, Teknoloji,Ekonomi	Politika, Teknoloji,Ekonomi
Kullanıcı #8	Dünya, Ekonomi,Kültür	Ekonomi, Dünya, Teknoloji
Kullanıcı #9	Spor, Ekonomi,Teknoloji	Spor, Teknoloji,Ekonomi
Kullanıcı#10	Teknoloji,Ekonomi,Sağlık	Teknoloji,Ekonomi,Sağlık
Kullanıcı#11	Teknoloji,Sağlık, Kültür	Sağlık, Teknoloji,Ekonomi
Kullanıcı#12	Kültür, Sağlık, Ekonomi	Sağlık, Kültür, Ekonomi
Kullanıcı#13	Spor, Ekonomi,Dünya	Spor, Ekonomi,Dünya
Kullanıcı#14	Teknoloji, Spor, Kültür	Teknoloji, Spor, Ekonomi
Kullanıcı#15	Spor, Ekonomi,Teknoloji	Spor, Ekonomi,Teknoloji

En son haliyle etiketlemeler yapılmış, hangi kategoride olduğu belirlenmiştir. Kullanıcı#1 için eğitilmiş model çıktısı sonucunun bir kısmı aşağıda görüldüğü gibidir.

Çizelge 6.3 Eğitim sonucunun kategorilenmiş hali

hafta vizyona giren yeni filmler güncel, kultur
tesla otomobillerin fiyatında indirime gitti, spor
türkiyedeki ucuz elektrikli arabalar güncel, teknoloji
insanlığın çocuk sahibi olduğu ortalama yaş açıklandı, saglik
güncel bilim teknoloji haberleri webtekno, teknoloji
iphone ilgili şok iddia modellerde yer alacak teknoloji haberleri, teknoloji

Çizelgeden anlaşılacağı üzere bazı etiketlerin yanlış olarak verildiği anlaşılmaktadır. Kayıt bazında modelin göstermiş olduğu gerçek başarıyı ölçmek üzere, eğitilmiş MNB sınıflandırıcısı ile kullanıcının her geçmiş verisi tek tek doğruluk açısından ele alınmış, bu durumda başarının %55 ile sınırlı olduğu görülmüştür (Çizelge 6.4).

Çizelge 6.4 Tek tek kayıtların doğruluk başarısı

	Test Accuracy	Precision	Recall	F1
MultinomialNB()	55.74	0.55	0.55	0.55

7 SONUÇ VE ÖNERİLER

Bu çalışmada kullanıcının girdiği siteler, aradığı kelimeler ve yer imleri dikkate alınarak içerik tabanlı bir tavsiye sistemi geliştirilmiştir. Diğer modellerle ve sonucun doğruluğuyla karşılaştırıldığında MNB doğrulama için %86 başarı oranıyla en iyi tavsiyeyi vermiştir. Ayrıca model eğitiminde %86 doğruluk oranı, Kullanıcı tercihlerine göre kategori bazında %89, ve gerçek kayıt testiyle de %55 gibi bir doğruluk göstermiştir. Sonuçlara bakıldığında kullanıcı testinde eşik değerini geçen kategoriler alındığından dolayı ilk 3 kategori için yüksek bir başarı elde edilmiştir. Tüm kategorilerin doğruluk değerine baktığımızda ise bu oran düşmektedir. Veri sayısının az olduğu ve kategori sayısının az olduğu durumlar ayrıca ele alınarak çalışılarak, parametrik bir yapı oluşturulması düşünülebilir. Bu çalışmada, mevcut veriler gözetiminde MNB bize en yakın haber kategorisonucunu vermiştir.

Sistem genel olarak başarılı olmasına rağmen, bazı kullanıcıların hem Türkçe hem de farklı bir dil kullandığı düşünüldüğünde bu koşullarda başarı etkisini azalttığı gözlemlenmiştir. Oranın yükselmesi, işbirlikçi yöntem eklenerek kullanıcı benzerliklerine göre hibrit bir yaklaşımla arttırılabilir. TF-IDF ağırlıklandırmasına ek olarak literatürdeki farklı ağırlıklandırma yöntemlerinin performans etkilerinin de gözlemlenmesi gerekmektedir.

Haberler getirilirken haber sitelerinin gündem manşetleri elde edilerek gündemdeki manşetlere daha yakın olan haberlerin ilk başta çıkması sağlanmıştır. Böylece hem ilgi çeken hem de gündemde daha çok öne çıkan haberler kullanıcıya gösterilmektedir. Kullanıcının beğendiği bir şeyi sunmak hem okuma alışkanlığını arttırmakla kalmaz hem de zaman kaybını önler.

Tavsiye sistemlerinin günümüzde bile hayatımızda kalıcı bir role sahip olduğu düşünüldüğünde, gelecekte haber tavsiye sistemlerinin hızlı günlük verilere ihtiyaç duyan kişilere hizmet edeceği bir gerçektir.

KAYNAKLAR

- Beel J., Gipp B., Langer S., Breitinger C. 2016. Paper Recommender Systems: A Literature Survey. *International Journal on Digital Libraries*, 17(4), 305-338.
- Billsus D., Pazzani, M.J. 1999. A Hybrid User Model for News Story Classification. In: Kay, J. (eds) *UM99 User Modeling*. CISM International Centre for Mechanical Sciences 407, 99-108 Springer, Vienna.
- Çelik Ö., Koç B. C., 2021. Classification of Turkish News Texts with TF-IDF, Word2vec and Fasttext Vector Model Methods. *DEÜ FMD*. 23(67), 121-127
- Deniz E., Öz V. K., Bozkurt Keser S., Okyay S., Kartal Y. 2021. İçerik Tabanlı Bilimsel Yayın Öneri Sisteminde Benzerlik Ölçümlerinin İncelenmesi. *DUMF Journal of Engineering* 12(2), 221-228, doi:10.24012/dumf.838084.
- Dhruv A., Kamath A., Powar A., Gaikwad K. 2019. Artist Recommendation System Using Hybrid Method: A Novel Approach. In: Shetty N., Patnaik L., Nagaraj H., Hamsavath P., Nalini N. (eds) *Emerging Research in Computing, Information, Communication and Applications*. Advances in Intelligent Systems and Computing. (882). Springer, Singapore.
- Durukan, F., 2018, Öneri Sistemleri. Erişim Tarihi: 03.11.2022. <https://www.fatihdurukan.com/oneri-sistemleri/>
- Dwivedi, R., 2020, What Are Recommendation Systems in Machine Learning. Erişim Tarihi: 03.12.2022. <https://www.analyticssteps.com/blogs/what-are-recommendation-systems-machine-learning>,
- Erdinç, S., 2017. Python BeautifulSoup Modülü. Erişim Tarihi: 03.12.2022. <https://www.sinanerdinc.com/python-beautifulsoup-modulu>
- Jin, Y., 2016. Development of Word Cloud Generator Software Based on Python, 13th Global Congress on Manufacturing and Management 28-30 November, Zhengzhou, China.
- Jonnalagedda N., Gauch S., Labille K., 2016. Alfarhood S. Incorporating Popularity in a Personalized News Recommender System. *PeerJ Computer Science*, 2: e63.
- Kaşıkcı, T., Gökçen H. 2014. Determination of E-Commerce Sites with Text Mining. *Journal of Information Technologies*. 7(1). DOI:10.12973/bid.2014
- Kumaş, E. 2021. Comparison of Classifiers While Performing Sentiment Analysis from Turkish Twitter Data. *Journal of ESTUDAM Information*. 2(2), 1-5.

- Li, L., Wang, D., Li, T., Knox, D., Padmanabhan, B., 2011. Scene: A scalable twostagepersonalized news recommendation system, Proceedings of the 34th international ACM SIGIR conference on Resear. 125–134.
- Liu, J., Dolan P., Pedersen E.R. 2010. Personalized news recommendation based on click behavior. In: Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI, 31–40. ACM, New York.
- Meng, S., Dou, W., Zhang, X. Chen J., 2014. KASR: A Keyword-Aware Service Recommendation method on MapReduce for big data applications. IEEE Transactions on Parallel and Distributed Systems, 25(12), 3221-3231.
- Metin S.K., Karaoğlu B., 2010. Collocation extraction in Turkish texts using statistical methods. International Conference on Natural Language Processing. Springer, Berlin, Heidelberg.
- Olson, D.L. Delen, D., 2008. Advanced Data Mining Techniques, Springer Science & Business Media, Verlag Berlin Heidelberg.
- Özkok H. 2020. Recommendation Engine (Tavsiye-Öneri Sistemleri). Erişim Tarihi: 03.11.2022. <https://www.datascienceearth.com/recommendation-engine-tavsiye- oneri-sistemleri/>
- Pypi.Browser History. 2022. Erişim Tarihi: 01.12.2022. <https://pypi.org/project/browserhistory>
- Pypi.Feed Parser. 2022. Erişim Tarihi: 01.12.2022. <https://pypi.org/project/feedparser/>
- Saranya, K.G., Sudha, G., Sadhasivam. 2012. A Personalized Online News Recommendation System. International Journal of Computer Applications. 57(18), 6-14.
- Seven, S., Alpkoçak, A. 2020. Kişisel Haber Öneri Sistemi. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi. 22(64), 301-307.
- Tan, A. H., Teo, C. 1998. Learning user profiles for personalized information dissemination, Neural Networks Proceedings. IEEE World Congress on Computational Intelligence. 1, 183-188.
- Taşçı, S., 2015. Content Based Media Monitoring and News Recommendation System. Master Thesis. Hacettepe University, Computer Engineering, Ankara.
- Uslu, O., Özmen Akyol, S., 2021. Türkçe Haber Metinlerinin Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması, Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 2(1), 15-20.

Utku, A., Akcayol, M. A., 2021. Derin Öğrenme Tabanlı Model ile Bir Olayın Sonraki Olma Zamanının Tahmini, Politeknik Dergisi, 24(1), 1-15, doi:10.2339/politeknik.620613.

Weider, D., Yu, Pratiksha, C., Swati, S., Akhil, S., Sarath, M., 2015. A Modeling Approach to Big Data Based Recommendation Engine in Modern Health Care Environment. In Computer Software and Applications Conference (COMPSAC), 75-86.

Yildirim, S. 2017. Text Categorization for Turkish-Multi NB. Erişim Tarihi: 01.12.2022. <https://www.kaggle.com/savasy/text-categorization-for-turkish-multi-nb>,



ÖZGEÇMİŞ

Adı Soyadı : Melis ÖZKARA

Eğitim Durumu

Lise : Mustafa Kemal Anadolu ÖğretmenLisesi,2014

Lisans : KastamonuÜniversitesi,2018

Yüksek Lisans : İstanbul Ticaret Üniversitesi, Fen bilimleri
Enstitüsü, Bilgisayar Mühendisliği,2022

Yayınları

Özkara M., Turan M., 2023. Kişiselleştirilmiş Haber Tavsiye Sistemi, İstanbul
TicaretÜniversitesi Teknoloji ve Uygulamalı Bilimler Dergisi, Basımda.