KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

BANKACILIK BAKIŞ AÇISIYLA YAPAY ZEKA DESTEKLİ HABER ANALİZ SİSTEMİ

DOKTORA TEZİ

Hasan AMANET

EKİM 2024
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

BANKACILIK BAKIŞ AÇISIYLA YAPAY ZEKA DESTEKLİ HABER ANALİZ SİSTEMİ

Hasan AMANET

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"DOKTOR (İSTATİSTİK)"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 17 / 09 / 2024
Tezin Savunma Tarihi : 31 / 10 / 2024

Tez Danışmanı : Dr. Öğr. Üyesi Tolga BERBER

Trabzon 2024

ÖNSÖZ

“Bankacılık Bakış Açısıyla Yapay Zeka Destekli Haber Analiz Sistemi” isimli bu tez Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik ve Bilgisayar Bilimleri Anabilim Dalı, Doktora Programı’nda hazırlanmıştır.

Tez sürecimde rehberlik eden ve bilimsel gelişimime katkı sağlayan danışman hocam Dr. Öğretim Üyesi Tolga BERBER’e, değerli tavsiyeleriyle yol gösteren Prof. Dr. Asiye Mevhibe ÇOŞAR ve Dr. Öğretim Üyesi Uğur ŞEVİK’e; akademik duruşuyla ilham veren Prof. Dr. Zafer KÜÇÜK’e ve tüm eğitim öğretim hayatım boyunca katkısı olan kıymetli hocalarıma teşekkür ederim. Ayrıca, tez sürecimdeki desteklerinden dolayı arkadaşım Nisanur ŞAHİNER'e teşekkürlerimi sunarım.

Ayrıca, bu süreçte bana destek olan Albaraka Tech Global ailesine; çalışmanın hayata geçmesinde büyük katkıları olan Hasan LAÇİN Bey, Ufuk ŞENGEL Bey ve Ömer KAYA Bey başta olmak üzere tüm iş arkadaşlarıma teşekkürlerimi sunarım.

Bana her zaman koşulsuz sevgi ve destek veren, aldığım tüm kararlarda arkamda duran ve bugünlere gelmemde büyük pay sahibi olan canım annem Hanımzar AMANET ve babam Ali AMANET; sabrınız, sevginiz ve dualarınızla her zaman yanımda oldunuz, sizlere ne kadar teşekkür etsem azdır. Ayrıca, sevgili kardeşim Kadriye PATAN’a, bana her zaman inandığı ve destek olduğu için teşekkür ederim.

Sevgili eşim Gülümser AMANET, her zaman yanımda oldun ve bu zorlu süreçte verdiğin destekle en büyük dayanağım oldun. Sabrın, anlayışın ve desteğin olmasaydı bu yolculuk çok daha zor olurdu.

Son olarak, sevgili çocuklarım Ali Çınar ve Alp Çağlar, sizlerin neşesi ve varlığı bana her daim güç verdi. Bu süreçte size ayıramadığım zamanları, bir gün bu çalışmamın meyvelerini gördüğünüzde anlayışla karşılayacağınızı umuyorum. Her şeyden çok, size layık bir ebeveyn olmayı diliyorum.

Hepinize sonsuz teşekkürlerimle...

Hasan AMANET

Trabzon 2024

TEZ ETİK BEYANNAMESİ

Doktora Tezi olarak sunduğum “Bankacılık Bakış Açısıyla Yapay Zeka Destekli Haber Analiz Sistemi” başlıklı bu çalışmayı baştan sona kadar danışmanım Dr. Öğr. Üyesi Tolga BERBER’in sorumluluğunda tamamladığımı, verileri/örnekleri açık kaynak olarak kamuya sunulan internet ortamından toplandığını, deneyleri/analizleri yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 31/10/2024

Hasan AMANET

İÇİNDEKİLER

Sayfa No

Doktora Tezi

ÖZET

BANKACILIK BAKIŞ AÇISIYLA YAPAY ZEKA DESTEKLİ HABER ANALİZ SİSTEMİ

Hasan AMANET

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik ve Bilgisayar Bilimleri Anabilim Dalı
Danışman: Dr. Öğr. Üyesi Tolga BERBER
2024, 79 Sayfa, 5 Sayfa Ek

Banka ve finans sektörleri, sürekli artan veri akışını yönetebilmek için yenilikçi çözümlere ihtiyaç duymaktadır. Bu doğrultuda geliştirilen sistem, ulusal ve yerel haber kaynaklarını düzenli olarak tarayarak firma istihbaratını otomatikleştirmektedir. Sistem, firmalarla ilgili haberleri tespit edip analiz ederek, metinlerde yer alan firmalara skor ataması yapmaktadır. Böylece, haber içeriğinin hangi firmaya ait olduğu firma ağırlık skoruna göre belirlenmektedir. Sistemin duygu analizi modelinde %86'lık bir başarı oranı elde edilmektedir. Bu model, haber içeriklerini pozitif, negatif ve nötr olarak sınıflandırmaktadır. Sistem analiz sonuçları incelendiğinde yerel ve ulusal haber kaynaklarının bankacılık açısından farklı kelimelere ağırlık verdiği tespit edilmiştir.

Önerilen sistem, bankacılık perspektifinde önemli olan pozitif ve negatif kelimelere ve duygu analizi modeline dayalı olarak firmaların piyasadaki algısını ve kamuoyundaki duruşunu belirlemektedir. Bu sayede bankalar ve finans kuruluşları, risk yönetimi süreçlerine katkı sağlayacak stratejik bir firma istihbarat kaynağına sahip olmaktadır. Modüler yapıda tasarlanan sistem, gerçek zamanlı işlem kabiliyeti ile çalışmakta ve her modül bağımsız bir mikro servis olarak görev yapmaktadır. Bu özellikleri sayesinde, bankacılık sektörünün dinamik ihtiyaçlarına uyum sağlayarak, zamanında ve doğru firma istihbaratı sunarak ve karar alma süreçlerini desteklemektedir.

Anahtar Kelimeler: İnternet kazıma, Haber analizi, Duygu analizi, Yapay zekâ, Makine öğrenmesi

PhD. Thesis

SUMMARY

ARTIFICIAL INTELLIGENCE-BASED NEWS ANALYSIS SYSTEM FROM A BANKING PERSPECTIVE

Hasan AMANET

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Statistics and Computer Sciences Graduate Program
Supervisor: Asst. Prof. Dr. Tolga BERBER
2024, 79 Pages, 5 Pages Appendix

The banking and finance sectors require innovative solutions to manage the ever-increasing flow of data. In line with this need, the developed system continuously scans national and local news sources to automate company intelligence. The system identifies and analyzes news related to companies and assigns scores to the companies mentioned in the texts. This allows the determination of which company the news content pertains to, based on the company's weight score. The sentiment analysis model of the system achieved an 86% success rate. This model classifies news content as positive, negative, or neutral, and it was found that national and local news sources emphasize different terms and keywords relevant to the banking sector.

The proposed system determines the market perception and public stance of companies based on positive and negative keywords that are significant from a banking perspective. In this way, banks and financial institutions gain access to a strategic company intelligence resource that contributes directly to their risk management processes. The system, designed with a modular structure, operates with real-time processing capabilities, with each module functioning as an independent microservice. These features provide a significant advantage by offering timely and accurate company intelligence that meets the dynamic needs of the banking sector, thereby supporting decision-making processes.

Key Words: İnternet scraping, News analysis, Sentiment analysis, Artificial intelligence, Machine learning

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER VE KISALTMALAR DİZİNİ

DA : Duygu Analizi

DDİ : Doğal Dil İşleme

KA : Karar Ağaçları

MÖ : Makine Öğrenmesi

DÖ : Derin Öğrenme

AVT : Adlandırılmış Varlık Tanıma (Named Entity Recognition)

RO : Rastgele Orman (Random Forest)

SVM : Destek Vektör Makineleri (Support Vector Machine)

LR : Lojistik Regresyon

NB : Naive Bayes

$UKSB$: Uzun Kısa Süreli Bellek (Long Short-Term Memory)

TSA : Tekrarlayan Sinir Ağı (TSA)

$KDTV$: Küresel Olaylar, Dil ve Ton Veritabanı (Global Database of Events, Language, and Tone)

KİAÖ : Küresel İçerik Analizi Ölçütü (Global Content Analysis Measure)

$TF-TDF$: Terim Frekansı-Ters Doküman Frekans (Term Frekansı-Ters Doküman Frekansı)

OKS : Olay Kayıt Sistemi (Event Registery)

$BÇ$: Bilgi Çıkarımı

İDB : İç-Dış-Başlangıç (Inside-Outside-Beginning)

TSA : Tekrarlayan Sinir Ağlarının (Recurrent Neural Networks)

1. GENEL BİLGİLER

“News” kelimesinin İngilizce'de coğrafi yönlerin baş harflerinin (North, East, West, South) bir araya getirilmesiyle oluştuğuna dair bir inanış bulunmaktadır. Bu kavramsal çerçeve, haberlerin toplumun farklı bölgelerinden gelen olayları değerlendirerek toplumsal algıları şekillendirdiğini vurgulamaktadır (Stephenson vd., 2013)

Dijital çağın getirdiği veri bolluğu ve hızlı bilgi akışı, bilgiye erişimde büyük bir dönüşüm yaratmıştır. Günümüzde bankalar, rekabet avantajını korumak ve piyasa dinamiklerine hızlı bir şekilde adapte olabilmek için büyük veri analitiği ve yapay zekâ gibi teknolojilere daha fazla yatarım yapmaktadır. Finansal piyasaların hızla değişen dinamikleri ve sürekli artan veri hacmi yeni bilgi işleme tekniklerinin kullanımı zorunlu kılmaktadır. Doğal Dil İşleme (DDİ) ve Makine Öğrenmesi (MÖ) tekniklerinin finansal istihbarat ve risk yönetimi süreçlerinde kullanımı gün geçtikçe daha fazla önem kazanmaktadır. Haber metinlerinin firmalar hakkında önemli bilgiler içermesi, haber metinlerinin finans alanından kullanımını gündeme getirmiştir.

İnternet haber metinlerinden elde edilen verilerin işlenmesi, analiz edilmesi ve yorumlanması süreçleri, bankacılık sektöründeki karar verme mekanizmalarını destekleyecek stratejik bilgileri ortaya çıkarmayı hedeflemektedir. Bankalar ve diğer finans kurumları için firmalar hakkında doğru ve güncel bilgiye hızlı bir şekilde ulaşabilmek, kredi verme, yatarım yapma ve risk değerlendirme süreçlerinde kritik öneme sahiptir. Dolayısıyla internet üzerinden erişilen haber metinlerinin, bu tür kararları destekleyecek şekilde analiz edilmesi, bankalar için stratejik bir avantaj sağlamaktadır.

Bu potansiyele rağmen, izlenmesi gereken haber makalelerinin çok fazla olması nedeniyle pratik zorluklar ortaya çıkmakta ve haber analizini manuel olarak yapmayı imkansız hale getirmektedir. Bu nedenle, doğal dil işleme tabanlı otomatik yöntemler hem uygulamada hem de akademide önerilmektedir. Bazı ticari ürünler belirli ülkeler ve şirketlerle ilgili ortaya çıkan konuları tespit etmek için çevrimiçi haber makaleleri arama araçları sunmaktadır (Waldron, 2021).

Finansal haber metinlerinde doğal dil işleme ve makine öğrenmesi tekniklerinin kullanımı araştırmacıların ilgilisini çekmektedir. Firmalarla ilgili haber metinlerinin doğru ve hızlı analizi, firmaya dair kapsamlı ve derinlemesine bilgi edinmeyi mümkün kılmaktadır.

Şirketlere ilişkin haber metinleri ile finansal değerler arasındaki ilişkilerin incelenmesi için metin madenciliği ve duygu analizi teknikleri kullanılmaktadır (Atan ve Çınar, 2019).

Haber metinlerinin analizi şirketlerin, karşılaşacağı olası risklerin ve problemlerin önceden tespit edilmesinde önemli bir araç olarak öne çıkmaktadır. Yapılan farklı çalışmalar haber metni analizinin firmalar hakkında risk ve sorunları tespit etmek için veri kaynağı olarak potansiyeli olduğunu ortaya koymaktadır (Chung vd., 2023).

Haber metinlerinin analizinde, doğal dil işleme tekniklerinin doğru ve etkin bir şekilde kullanılması kritik bir öneme sahiptir. Bu analizlerde, metin içerisinden derinlemesine bilgi edinmek amacıyla Adlandırılmış Varlık Tanıma (AVT) gibi teknikler kullanılmaktadır. AVT tekniği, metin içeriğindeki kurum, kişi, tarih ve konum gibi kavramsal bilgilerin doğru bir şekilde tespit edilmesi ve metinden detaylı bilgi edinmek için kullanılmaktadır (Tu, 2024).

Haber metinlerinden bilgi edinme sürecinde önemli bir yere sahip olan diğer doğal dil işleme tekniği ise Duygu Analizi (DA) tekniğidir. DA, haber metinlerinde var olan duygu tonlarını ve yönelimleri analiz ederek, haberlerin olumlu, olumsuz ve nötr bir bakış açısına sahip olup olmadığını tespit etmektedir. Bu analiz sayesinde, firmalarla ilgili kamuoyu algısı ve kriz süreçlerinde ortaya çıkabilecek erken uyarı sinyalleri tespit edilerek, firmalar hakkında alınacak stratejik kararlar daha etkin bir şekilde desteklenmektedir. Bu nedenle DA, haber metinlerin içerisinde sadece nitel verilerin değil, duygusal ve psikolojik faktörlerin de dikkate alınmasına imkan tanımaktadır (Jacobs ve Hoste, 2022).

Bankacılık alanında kredi tahsisi ve risk yönetimi ekipleri tarafından müşterisi olan veya potansiyel müşterilerle ilgili haber ve bilgilerin sürekli izlenilmesi büyük önem arz etmektedir. İnternet ortamında her gün büyük miktarda metin verisi üretilmekte ve çok farklı kanallar üzerinden yayılmaktadır. Büyük miktardaki bu veriyi manuel olarak analiz etmek ve yorumlamak çok fazla iş gücü gerektirmektedir. Bu bağlamda, DDİ ve makine öğrenmesi teknikleri kullanılarak haber metinlerinden firma istihbaratı elde etmek, bankacılık operasyonlarının etkinliğini arttırmak için önemli bir fırsat sunmaktadır.

Bu tez kapsamında, firmalar hakkında toplanan haberlerin analiz edilmesi ve bankacılık bakış açısıyla değerlendirilmesi amacıyla çeşitli makine öğrenmesi teknikleri ve doğal dil işleme yöntemlerini kullanan bir sistem önerilmektedir. Çalışma kapsamında geliştirilen sistem haber metinlerini otomatik olarak toplayarak kredi tahsis ve risk yönetimi ekiplerine güvenilir ve anlık bilgi akışı sağlamayı amaçlamaktadır. Bu sayede bankalar, mevcut müşteri ve potansiyel müşterileriyle ilgili risk oluşturabilecek durumları takip

edebileceklerdir. Örneğin, müşteri olan bir firma hakkında “kara para aklama” konusuyla ilgili haber çıktığında sistem kullanıcıya bu firmayla ilgili uyarıda bulunacaktır. Kullanıcı bu uyarıyla birlikte firmayla ilgili alması gereken tüm aksiyonları hızlı bir şekilde alma fırsatına sahip olacaktır. Geliştirilen bu sistemle, bankacılık açısından oluşacak bilecek mali kayıplar ve itibar kaybı gibi olumsuz durumların önenmesi ve iş gücü verimliliğinin artırılması hedeflenmektedir.

1.1. Literatür Taraması

Dijital çağın hızla değişen dinamikleri, bilgiye ulaşma ve veri analizi alanlarında köklü değişikliklere yol açmıştır. Özellikle finans alanında bu dönüşüm belirgin bir şekilde hissedilmektedir. İnternetin yaygınlaşması, haber kaynağı ve içerik sayısında büyük bir artışa neden olmuştur. Bu durum, banka müşterilerinin haber konusu olması halinde, haberlerin ayrıntılı ve hızlı bir şekilde analiz edilmesini zorunlu kılmaktadır.

Günümüzde hızla tüketilen ve tekrarlanmayan gündemlerin kolayca unutulması, potansiyel risklerin gözden kaçmasına neden olmaktadır. Bundan dolayı özellikle banka müşterilerinin haberlere konu olduğu durumlarda, oluşan haber yığınlarının kapsamlı bir şekilde analiz edilmesi ve sınıflandırılması gerekmektedir. Haberler üzerinde yapılacak bu analiz, bankaların risk yönetimi ve müşteri ilişkileri açısından büyük önem taşımaktadır. Yapılan literatür incelemesinde haber metinlerinin bankacılık bakış açısıyla analiz edilmesiyle ilgili çalışmaların kısıtlı olduğu görülmüştür.

Haber metinlerinin analizi, bankacılık sektöründe müşterilerle ilgili karar alma süreçlerini destekleyebilecek önemli sonuçlar sunma potansiyeline sahiptir. Bu yaklaşım banka müşterisi veya potansiyel müşterisi olabilecek firmaların, finansal durumları ve risk profilleri hakkında daha derinlemesine bilgi edinme imkanı sağlamaktadır.

Makine öğrenmesi ve metin analitiği alanındaki teknolojik gelişmelerin haber analiz süreçlerini olumlu yönde etkilediği görülmektedir. Metin analitiği alanındaki olanakların sınırlı olduğu dönemlerde, haber analizi işlemi, araştırmacıların haberleri bizzat okuyarak değerlendirmesi şekilde yapılmaktaydı. Ancak geliştirilen algoritmaların, haber analizi süreçlerinde kullanılmasıyla manuel olarak incelenmesi mümkün olmayan, büyük miktardaki haber metinlerinin analizine imkan sağladığı görülmektedir.

Atan ve Çınar, BIST30 endeksinde işlem gören firmalar hakkında çıkan haberlerin, firmaların piyasa değerine olan etkisini incelemiştir. Bu çalışma metin madenciliği ve duygu

analizi tekniklerini kullanarak önemli bulgular ortaya koymuştur. Yapılan analizler, finans haberlerinin içerdiği duygu tonlarıyla, firmaların piyasa değerleri arasında istatistiksel olarak anlamlı bir ilişki olduğunu ortaya koymaktadır. Elde edilen bulgular Türkçe haber kaynaklarının finansal piyasaların değerlendirilmesinde etkili bir araç olarak kullanılabileceğini göstermektedir (Atan ve Çınar, 2019).

Jacobs ve Hoste, çalışmalarında ekonomik ve finansal haberlerde şirketlere özgü olayların çıkarımını sağlayan SENTiVENT adında bir sistem sunmuşlardır. Çalışma, haber analizinin, şirketlerin piyasa değerleri üzerindeki etkilerini anlamada kritik role sahip olduğunu göstermektedir (Jacobs ve Hoste, 2022).

Zheng ve arkadaşları, internet ortamındaki finans haberlerinin duyu analizi ile metin içerisindeki varlıkların tanımlanması için uçtan uca bir analiz sistemi önermiştir. Haber metni içinde yer alan anahtar varlıkların ve bu varlıklar üzerindeki duyguların doğru tespit edilmesi kritik öneme sahiptir olduğu görülmektedir (Zheng vd., 2021).

Panagiotou ve arkadaşları, gerçek zamanlı haber analizi için geliştirdikleri "News Monitor" uygulaması ile haberler metinlerin toplanması, işlenmesi ve analiz edilmesinde yenilikçi bir yaklaşım sunmaktadırlar. News Monitor uygulaması Zengin Site Özeti (ZSÖ) kaynağından sürekli haber metinlerini toplayarak, bu haberleri çeşitli doğal dil işleme teknikleriyle analiz etmektedir. Bu çalışmada haberlerin sadece metinsel içeriklerini değil, aynı zamanda sosyal medya etkileşimlerini de analiz ederek, kullanıcıların haberi çok boyutlu değerlendirmesine imkan sunmaktadır (Panagiotou vd., 2022).

Leban ve arkadaşları tarafından geliştirilen Olay Kayıt Sistemi (OKS) sistemi, dünya çapındaki olayları haber metinlerinden tespit etmek ve analiz etmek için tasarlanmış bir araçtır. Bu sistem, farklı dillerdeki haber metinlerini toplayarak, aynı olayı anlatan grupları tespit edebilmekte ve bu grupları tek bir olay olarak temsil edebilmektedir. OKS, haber metinlerinden olayların temel bilgilerini (örneğin, olayın yeri, tarihi, olayda yer alan kişiler ve olayın içeriği) çıkarıp bu bilgileri bir veri tabanında saklamaktadır. Kullanıcılar, bu olayları çeşitli arama kriterleri (örneğin, belirli bir konu, yer veya tarih) kullanarak arayabilir ve sonuçları görselleştirerek detaylı bir şekilde inceleyebilir (Leban vd., 2014).

Chung ve arkadaşları, doğal dil işleme tekniklerini kullanarak haber metinlerinden ülke risklerini belirleyen faktörleri, tespit etmeyi amaçlayan bir sistem sunmuşlardır. Yapılan çalışma, uluslararası inşaat projelerinde ülke riskini önemli bir değişken olarak dikkate almaktadır. Ülkenin siyasi, ekonomik ve sosyal durumlarının projenin tamamlanmasında önemli bir faktör olduğunu ifade etmektedir. Bu kapsamda doğal dil

işleme teknikleriyle haber metinlerinin sayısallaştırılması ve riskle ilgili konuların belirlenmesi amaçlanmaktadır (Chung vd., 2023).

Agarwal ve arkadaşları makine öğrenimi algoritmalarını kullanarak haber metinlerinin sınıflandırılmasıyla ilgili bir çalışma gerçekleştirmişlerdir. Bu çalışma, kişilerin yalnızca ilgi duydukları konulara odaklanabilmeleri için haber metinlerini otomatik olarak kategorilere ayrılmasını amaçlamaktadır. Çalışmada, 2012-2022 yılları arasındaki yaklaşık 210.000 haber başlığından oluşan bir veri seti kullanılmaktadır. Bu veri seti üzerinde Rastgele Orman, Karar Ağaçları, K-NN Sınıflandırıcısı ve Naive Bayes gibi dört farklı makine öğrenimi algoritması kullanılmaktadır. Elde edilen sonuçlar incelendiğinde, haber metinlerinin sınıflandırılmasında bu algoritmaların başarısını ortaya koymaktadır (Agarwal vd., 2023).

Kumar ve arkadaşları haber metinleri ve sosyal medya metinlerinin duygu analizini yaparak hisse senedi fiyatlarının tahminiyle ilgi bir model geliştirmişlerdir. Yapılan çalışmada, özellikle finansal haber metinlerinin ve sosyal medya platformlarında yayımlanan metinlerin, hisse senedi piyasası üzerindeki etkileri incelenmektedir. Bu kapsamda metinlerden elde edilen duygusal ipuçlarını kullanarak hisse senedi fiyatlarının tahmin edilmesi amaçlanmaktadır. Geçmişteki veriler ve hisse senedi fiyatlarının tarihsel eğilimleri ile birleştirilen duygu analizinin, hisse senedi fiyatlarını tahmin edebildiğini göstermektedir (Kumar vd., 2022).

Nurrahmat ve Sunindyo firmaların dış çevre analizi başlığı altındaki metriği kullanmak amacıyla internet ortamındaki haber başlıklarından yararlanarak bir yöntem geliştirmişlerdir. Yapılan çalışmada, PEST (Politik, Ekonomik, Sosyal, Teknolojik) analizi ve metin madenciliği tekniklerini kullanarak, internet haber başlıkları otomatik olarak toplanmaktadır. Toplanan başlıklar fırsat veya risk olarak sınıflandırılmaktadır (Nurrahmat ve Sunindyo, 2019).

Seungwon ve arkadaşları doğal dil işleme tekniklerini kullanarak internet haber metinlerinden, inşaat sektörüne giriş yapan firmalar hakkında bilgileri çıkaran bir sistem önermektedir. Geliştirilen Adlandırılmış Varlık Tanıma (AVT) modeliyle %85'lik bir F1 skoru elde edildiği görülmektedir (Seungwon vd., 2023).

Piccioni ve arkadaşları çevresel, sosyal ve yönetişim haberlerinin, Brezilya borsasında işlem gören şirketlerin hisse senedi fiyatları üzerindeki etkilerini incelemiştir. Bu çalışma, çevresel, sosyal ve yönetişim haberlerinin piyasa değerleri üzerindeki etkilerini analiz etmek için yenilikçi bir Çevresel, Sosyal ve Yönetişim Terimleri Sözlüğü kullanarak, haber

içerikleri sınıflandırılmaktadır. Sonuçlar, olumlu Çevresel, Sosyal ve Yönetişim haberlerinin hisse senedi fiyatlarında önemli pozitif reaksiyonlara neden olduğunu, olumsuz haberlerin ise negatif etkiler yarattığını göstermektedir (Piccioni vd., 2024).

Nguyen ve arkadaşları çevrimiçi haber siteleri ve sosyal medya ortamlarından metin verisi toplama, işleme ve analiz etme yeteneğine sahip gerçek zamanlı bir metin analiz sistemi geliştirmişlerdir. Bu sistem, haber portalları ve sosyal medya gibi açık kaynaklardan alınan uzun ve kısa metin belgelerini işleyebilmekte ve bu metinlerden adlandırılmış varlık çıkarımı, varlık bağlantısı, varlık izleme ve olay çıkarımı yapabilmektedir (Nguyen vd., 2023).

Balcı ve arkadaşları, Türk ekonomi haberlerinin Borsa İstanbul'daki fraktal yapı üzerindeki etkilerini incelemekte ve bu haberlerdeki duygu unsurları ile finansal korelasyon ağlarının fraktal boyutlarını analiz etmektedirler. Çalışma, ekonomi haberlerinin duygusal içeriğinin ve piyasa ilişkilerinin, Borsa İstanbul'un fraktal yapısını nasıl etkilediğini anlamayı amaçlamaktadır (Balcı vd., 2024).

Nyman ve arkadaşları, finansal sistemlerdeki risk değerlendirme işlemi için haber metinlerinin önemini inceleyen bir çalışma sunmuşlardır. Çalışmada, finansal verilerinden elde edilen büyük miktarda metin verisi üzerinde çeşitli algoritmik analizler uygulanmaktadır. Yapılan bu analiz sonuçlarında haber metnin duygusal durumunun piyasa gelişmeleri üzerindeki etkileri değerlendirilmektedir (Nyman vd., 2021).

Zhang ve arkadaşları, internet ortamında finans platformlarının risklerini belirlemek için metin verilerini kullanan bir derin öğrenme yaklaşımı geliştirmişlerdir. Çalışmada, yasa dışı ticaret, yüksek faiz vaatleri, piramit şeması şüpheleri gibi riskleri içeren bir risk endeksi sistemi oluşturulmaktadır. Bu kapsamda internet ortamındaki finans platformları hakkındaki haber metinleri, forumlardaki yorumlar ve platformların resmi internet sitelerinden toplanan metinlerle birlikte farklı metin verileri kullanılmaktadır (Zheng vd., 2021).

Atan ve arkadaşları Türkiye'deki hastaneler hakkında medyada yer alan haberlerin hastanenin kurumsal imajı üzerindeki etkilerini metin madenciliği teknikleriyle analiz etmişlerdir. Yapılan çalışmada, 2013-2017 yılları arasında ulusal bir gazetede yayınlanan 3.117 hastane haberi kullanılmıştır. Sonuçlar incelendiğinde, hastanelerle ilgili çıkan haberlerin çoğunluğunun bilgilendirici ve nötr duygu tonlara sahip olduğunu ortaya koyulmaktadır. En sık tekrarlanan konular incelendiğinde, hastane inşaatları ve hastane yangınları olmuştur. Bu çalışma, medya içeriklerinin hastanelerin kurumsal imajını nasıl şekillendirebileceğine dair önemli içgörüler sunmaktadır (Atan, 2018).

Chen ve arkadaşları portföy risk tahmini için haber analiz sistemi geliştirmişlerdir. Bu sistem, geleneksel sözlük tabanlı metin madenciliği yöntemlerinden farklı olarak, yatırımcıların haber analiz yeteneklerini simüle etmek amacıyla vaka tabanlı bir çıkarım yöntemi kullanmaktadır. Sistem, olay düzeyinde haber analizine odaklanarak, haber olayları ve şirket performansları arasındaki dolaylı ilişkileri açığa çıkarmakta ve bu bilgileri portföy risk tahmini için kullanmaktadır (Chen vd., 2011).

Startseva ve arkadaşları, bankacılık işlemlerinin amacını tespit etmek için metin etiketlerinin analizine dayalı bir sistem geliştirmişlerdir. Kullanılan metin etiketleri bankacılık işlemlerinin amacını tespit etmeye yönelik belirlenen anahtar kelimeler ve terimlerden oluşmaktadır. Çalışma, bu metin etiketlerinin ön işleme, gövdeleme ve anahtar kelime tespiti gibi yöntemlerle analiz edilmesini sağlamaktadır. Bu yöntem, dinamik olarak müşteri risk profillerinin güncellenmesini sağlamaktadır. Geliştirilen analiz sistemi, finansal suistimal ve para aklama faaliyetlerini önlemede önemli bir adım olarak değerlendirilmektedir (Startseva vd., 2020).

Ahbali ve arkadaşları, finansal haber metinlerinden kurumsal kredi risk duyarlılıklarını tespit edebilmek için derin öğrenme tabanlı bir yöntem geliştirmişlerdir. Bu çalışmada, firmalarla ilgili olumsuz kredi olaylarını otomatik olarak tespit etmeyi ve bu durumlara bağlı olarak kredi duyarlılık skorunun oluşturulması amaçlanmaktadır. Geliştirilen analiz sistemi, haberlerin toplanması, verinin zenginleştirmesi ve belirlenen duyarlılık varlık tanımlaması gibi süreçleri içeren doğal dil işleme analizlerinden oluşmaktadır (Ahbali vd., 2022).

Makeeva ve Sinilschikova, yaptıkları çalışmada iflas tahmin modellerinde haber metinlerinden elde edilen duygusal değişkenlerin kullanımının önemini vurgulamaktadır. Özellikle Rusya'daki perakende sektörüne yönelik yapılan çalışmada, 190 şirket analiz edilmiş ve 95 iflas eden şirket ile iflas etmeyen benzer şirketler karşılaştırılmıştır. Bu analizde, finansal verilerin yanı sıra 4877 haber makalesinden elde edilen duygusal değişkenler kullanılmıştır. Sonuçlar, duygusal değişkenlerin iflas tahmin modellerinde istatistiksel olarak anlamlı olduğunu ve modelin performansını artırdığını göstermektedir (Makeeva ve Sinilshchikova, 2020).

Ayrıca, makine öğrenimi yöntemleri kullanılarak yapılan analizlerde, haber metinlerinde kullanılan olumsuz kelimelerin iflas olasılığı üzerinde daha güçlü bir etki yarattığı tespit edilmiştir. Bu bulgu, haberlerde kullanılan kelimelerin, şirketlerin finansal durumu hakkında ipucu verdiğini ve olumsuz haberlerin finansal istikrarsızlıkla, güçlü bir şekilde ilişkili olduğunu ortaya koymaktadır (Makeeva & Sinilshchikova, 2020).

Mai ve arkadaşları tarafından yapılan çalışmada, derin öğrenme tekniklerinin metin tabanlı iflas tahmin modellerine entegre edilmesinin etkisi incelenmiştir. Çalışmada iflas tahmin modellerinde derin metin tabanlı girdilerin kullanılmasının doğruluğu arttırdığı ortaya koyulmaktadır. Bu çalışma 11.827 ABD kamu şirketinin verilerini kullanarak gerçekleştirilmiştir. Derin öğrenme modelleri, metinsel açıklamalarla birlikte geleneksel muhasebe ve piyasa tabanlı değişkenlerin bir araya getirilmesiyle daha yüksek tahmin doğruluğu elde ettiği görülmektedir. Araştırma, basit kelime gömme yöntemlerinin, metinsel verilerle çalışan derin öğrenme modellerinde daha etkili olduğunu da ortaya koymuştur (Mai vd., 2019).

1.2. Yapay Zekâ

Yapay zekâ (YZ), makinelerin insan zekâsı gerektiren görevleri yerine getirme yeteneğini ifade eder. Bu yetenek, öğrenme, akıl yürütme ve kendini düzeltme gibi bilişsel süreçleri kapsar. Yapay zekâ sistemleri, verilerden anlam çıkarma, karar verme ve bu kararları optimize etme yeteneklerini içermektedir. Bu süreç, model tabanlı ve veri tabanlı yaklaşımlar olmak üzere iki ana kategoriye ayrılmaktadır. Model tabanlı yaklaşımlar, uzmanlar tarafından tanımlanan kurallara ve modellere dayanırken, veri tabanlı yaklaşımlar, doğrudan veriden öğrenerek modellerin oluşturulmasını sağlamaktadır.

YZ, doğal dil işleme (DDİ) gibi birçok alt alanda kullanılmaktadır. DDİ, insan dillerini anlama, üretme ve analiz etme sürecini içermektedir. DDİ, yazılı veya sözlü dili işleyebilmek için YZ tekniklerini kullanmaktadır. Bu bağlamda, yapay zekâ, dil modellerinin geliştirilmesi, duygu analizi, konuşma tanıma ve makine çevirisi gibi çeşitli DDİ görevlerini yerine getirmektedir. Günümüzde derin öğrenme (DÖ) tekniklerinin, büyük dil modellerinin oluşturulmasında ve dil işlemede önemli gelişmeler sağladığı gözlemlenmektedir. Derin öğrenme, YZ'nin, insanlar gibi düşünme ve öğrenme yeteneklerini artırarak, karmaşık sorunlara etkili çözümler sunmasını sağlamaktadır (Pillai ve Tedesco, 2023).

1.3. Makine Öğrenmesi (MÖ)

Makine öğrenmesi, bilgisayar sistemlerinin veri desenlerini analiz ederek, anlamlı ilişkiler çıkartma ve bu bilgileri kullanarak karar verme yeteneği geliştirme sürecidir. Bu süreçte, algoritmalar büyük veri kümeleri üzerinde çalışarak, kendi performanslarını optimize etmek amacıyla deneyimden öğrenmektedir. Makine öğrenmesi, istatistiksel

yöntemler ve optimizasyon tekniklerini bir araya getirerek, bilgisayarlar sistemlerinin karmaşık ve zor görevleri otomatik olarak gerçekleştirmesine imkan tanımaktadır.

Bu disiplin, metin madenciliği, doğal dil işleme (DDİ), görüntü tanıma ve karar destek sistemleri gibi alanlarda yaygın kullanılmaktadır. Makine öğrenmesi, denetimli, denetimsiz ve pekiştirmeli öğrenme gibi farklı öğrenme yöntemlerini içermektedir. Bu yöntemler, amaç fonksiyonları kullanarak kendi performanslarını iyileştirmeye çalışır ve veri içindeki desenleri öğrenmektedir. Böylece, veri odaklı karar alma süreçleri ve karmaşık problem çözme yetenekleri geliştirilmektedir (Jackson ve Moulinier, 2007).

1.4. Doğal Dil İşleme (DDİ)

Doğal dil işleme (DDİ), bilgisayarların insan dillerini (doğal diller) işlemek veya anlamak için kullanıldığı bir alandır. DDİ, yapay zeka, bilişim bilimi, bilişsel bilim ve dilbilim gibi birçok disiplinin birleşiminden oluşan bir alandır. Bilimsel açıdan bakıldığında, DDİ, insan dilinin anlaşılması ve üretilmesine yönelik bilişsel mekanizmaları modellemeyi amaçlamaktadır. Mühendislik perspektifinden ise DDİ, bilgisayarlar ile insan dilleri arasındaki etkileşimleri kolaylaştıracak yeni uygulamalar geliştirmeye odaklanmaktadır.

DDİ'nin yaygın uygulamaları arasında konuşma tanıma, konuşulan dilin anlaşılması, diyalog sistemleri, dil çözümleme, makine çevirisi, bilgi grafikleri, bilgi alma, soru yanıtlaması, duygu analizi, sosyal bilişim, doğal dil üretimi ve özetleme yer almaktadır. Bu alanlar, DDİ'nin insan dilini anlama ve işlemeye yönelik geniş yelpazedeki uygulama alanlarını temsil etmektedir.

DDİ hem insan dilinin yapısını hem de bilgisayarların bu yapıları nasıl işleyebileceğini anlamak için çok disiplinli yaklaşımlardan faydalanmaktadır. Bu süreç, bilgisayarlara daha insancıl ve anlamlı yanıtlar verme yeteneği kazandırmayı hedeflemektedir (Deng & Liu, 2018).

1.5. Adlandırılmış Varlık Tanıma (AVT)

Adlandırılmış varlıklar (AVT), belirli bireyleri, kuruluşları, kişileri, tarihleri ve benzeri kategorileri ifade eden özel isim tamlamalarıdır. Adlandırılmış Varlık Tanıma (AVT) sisteminin amacı, metindeki bu adlandırılmış varlıkların tüm metinsel referanslarını belirlemektir. Bu süreç iki alt göreve ayrılır: adlandırılmış varlığın sınırlarını belirlemek ve

türünü tanımlamak. Örneğin, “Mustafa Kemal ATATÜRK” isminin “KİŞİ” (PERSON) türünde bir adlandırılmış varlık olarak tanımlanması gerekmektedir.

AVT genellikle Bilgi Çıkarımı (BÇ) sürecinde ilişkilerin tanımlanmasına başlangıç olarak kullanılmaktadır. Ancak başka görevlerde de önemli katkılar sunabilir. Örneğin, Soru Cevaplama sistemlerinde, bilgi alma işleminin hassasiyetini artırmak için, yalnızca kullanıcı sorusunun cevabını içeren sayfaların belirli bölümlerini geri getirmek amacıyla kullanılabilmektedir. Örneğin, “Türkiye'nin ilk cumhurbaşkanı kimdir?” sorusunu yanıtlamak için, “Mustafa Kemal ATATÜRK, Türkiye’nin ilk cumhurbaşkanı olmuştur.” şeklindeki bir metin AVT sistemi tarafından analiz edilerek doğru cevabın “Mustafa Kemal ATATÜRK” olduğu belirlenebilmektedir.

Şekil 1. Örnek AVT analiz sonucu

AVT, çok kelimeli adlar ve iç içe geçmiş adlar gibi zorluklarla karşılaşabilmektedir. Bu nedenle, çoklu belirteç dizilerinin başlangıç ve bitiş noktalarını doğru bir şekilde belirlemek önemlidir. Bu tür görevler, adlandırılmış varlık tanımada sıklıkla İç-Dış-Başlangıç (İDB) formatı kullanılarak, her bir kelimeyi etiketleyen bir sınıflandırıcıyla ele alınabilmektedir. Bu sınıflandırıcı, eğitim verilerine dayanarak yeni cümleleri etiketlemek için kullanılmaktadır. Bu etiketler, etiket dizilerini bir kelime grubu ağacına dönüştürmekte yardımcı olmaktadır (Bird vd., 2009).

1.6. Duygu Analizi (DA)

“Duygu analizi”, ya da diğer adıyla “görüş madenciliği” insanların belirli varlıklar ve onların nitelikleri hakkındaki düşüncelerini, duygularını, değerlendirmelerini, tutumlarını ve duygusal tepkilerini inceleyen bir alandır. Bu alan, metin madenciliği ve doğal dil işleme yöntemleri kullanarak duyguları ve tutumları doğal dil metinlerinden çıkarmayı amaçlamaktadır (Liu, 2015).

Duygu analizi, genellikle pozitif, negatif veya nötr duyguların belirlenmesine odaklanmaktadır. Bu analiz, kullanıcıların ürünler, hizmetler, olaylar veya genel olarak

sosyal konular hakkındaki görüşlerini anlamak için kritik öneme sahiptir. Bu alandaki araştırmalar, genellikle belge düzeyinde, cümle düzeyinde ve daha ince ayrıntılar sunan özellik düzeyinde olmak üzere farklı gruplarda yürütülmektedir. Her seviyede, duyguların hedeflerinin belirlenmesi (örneğin, bir ürünün tadı veya hizmetin kalitesi) ve bu duyguların yönünün (pozitif veya negatif) tanımlanması hedeflenmektedir (Liu, 2015).

1.7. İnternetten Veri Toplama (İnternet Kazıma)

İnternet scraping, Türkçe'ye "internet kazıma" olarak çevrilen, internet sitelerinden otomatik olarak bilgi toplama işlemidir. Akademik ve endüstriyel uygulamalarda sıklıkla kullanılan bu yöntem, büyük miktarda veri toplamak ve bu verileri işlemek için kullanılmaktadır. İnternet kazıma işlemi genellikle HTML yapısında yer alan verilerin tespit edilmesi, düzenlenmesi ve anlamlı hale getirilmesi süreçlerinden oluşmaktadır (Naing vd., 2024).

Veri toplama süreçlerinde çeşitli internet kazıma araçları ve kütüphaneleri kullanılmaktadır. Bu araçlar, bir internet sitesinin yapısını analiz ederek belirlenen kriterlere uygun verileri otomatik olarak toplayabilmektedir. Ancak bu süreçte dikkat edilmesi gereken bazı etik ve yasal hususlar bulunmaktadır. Birçok internet sitesinin kullanım koşulları, otomatik veri toplama işlemlerine izin vermemektedir. Bu nedenle, internet kazıma yapılmadan önce hedef sitenin gizlilik ve kullanım politikalarının dikkatlice incelenmesi gerekmektedir (Boyapati ve Aygun, 2023).

İnternet kazımanın en önemli avantajlarından biri, büyük miktarda veriyi hızlı ve otomatik bir şekilde toplayabilmektedir. Özellikle akademik araştırmalarda, araştırmacılar belirli anahtar kelimelere dayalı olarak hızlı bir şekilde makaleleri toplamak ve bu makalelerin erişilebilirlik durumlarını değerlendirmek için internet kazıma kullanılmaktadır (Glez-Peña vd., 2014).

Bununla birlikte, internet kazıma işlemi sırasında karşılaşılan zorluklar bulunmaktadır. Özellikle dinamik internet sitelerinde veya veri güvenliği nedeniyle belirli sayfalara erişim kısıtlaması olan platformlarda, internet kazıma işlemi zor olmaktadır. Dinamik içeriklerin kazınması için tarayıcı otomasyonu araçları kullanılarak internet siteleri ile etkileşim sağlanabilmektedir. Ancak bu durum, tarayıcı tabanlı işlemlerin hızını düşürmekte ve kaynak tüketimini arttırmaktadır. Sonuç olarak, internet kazıma, büyük veri çağında bilgi toplama süreçlerini otomatikleştiren ve verilerin hızlı bir şekilde analiz edilmesini sağlayan

güçlü bir tekniktir. Ancak, bu süreçte yasal ve etik sınırların gözetilmesi, veri güvenliğine dikkat edilmesi gerekmektedir (Munzert vd., 2014).

1.8. Derin Öğrenme (DÖ)

Derin öğrenme (DÖ), özellikle son yıllarda doğal dil işleme (DDİ) ve yapay zeka alanlarında devrim niteliğinde ilerlemeler sağlayan bir makine öğrenmesi paradigmasıdır. Geleneksel makine öğrenmesi yöntemlerinden farklı olarak, derin öğrenme, özellik mühendisliği gibi insan müdahalesi gerektiren adımları en aza indirmektedir. Verilerden anlamlı özelliklerin otomatik olarak çıkarılmasına olanak tanımaktadır. Bu yaklaşım, çok katmanlı yapay sinir ağları kullanarak verilerden hiyerarşik temsiller öğrenmektedir. Böylece karmaşık veri düzeneklerini çözme kapasitesine sahip olmaktadır.

Derin öğrenmenin temelinde, biyolojik sinir sistemlerinden esinlenerek tasarlanmış çok katmanlı sinir ağları yer almaktadır. Bu ağlar, düşük seviyeli özelliklerden yüksek seviyeli kavramlara doğru bir dizi doğrusal olmayan dönüşüm gerçekleştirmektedir. Her bir katman, önceki katmandan aldığı bilgiyi daha soyut bir seviyeye taşır ve böylece, dil, görsel veya işitsel veriler gibi çeşitli veri türleri üzerinde etkili bir şekilde çalışabilmektedir.

Bu yeni öğrenme paradigması, geniş veri kümelerinin ve güçlü hesaplama kaynaklarının mevcudiyetiyle birlikte gelişmiş, özellikle de büyük ölçekli doğal dil işleme görevlerinde (örneğin, makine çevirisi, konuşma tanıma) üstün başarılar göstermektedir. Derin öğrenme, dildeki anlamlı kalıpları ve ilişkileri yakalayarak, dilin karmaşıklığını daha önceki tekniklerin yapamadığı bir şekilde anlamayı mümkün kılmaktadır (Deng & Liu, 2018).

1.9. Çalışmada Kullanılacak Makine Öğrenmesi Yöntemleri

Bu çalışmada haber metinlerin duygu durumlarına göre sınıflandırılması için farklı makine öğrenmesi modelleri kullanılmıştır. Haber metinlerinin duygu durumuna göre sınıflandırılması için “pozitif”, “negatif” ve “nötr” şeklinde üç kategoride sınıflandırma işlemi yapılmaktadır. Bu sınıflandırma işlemi için makine öğrenmesi algoritmalarının yanı sıra derin öğrenme yaklaşımları da kullanılmıştır.

Çalışma kapsamında klasik makine öğrenmesi algoritmalarından Destek Vektör Sınıflandırıcısı (DVS), Lojistik Regresyon (LR), Naive Bayes (NB), Karar Ağaçları ve Rastgele Orman algoritmaları kullanılmaktadır. Derin öğrenme algoritmalarından ise “Long

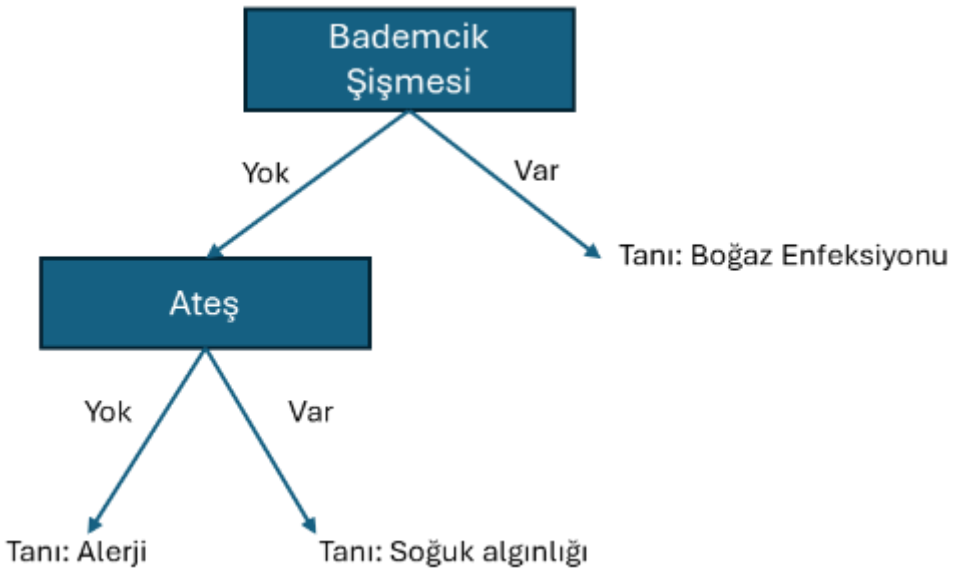
Short-Term Memory" (LSTM), Türkçeye Uzun Kısa Süreli Bellek (UKSB) olarak çevrilen algoritmasının en başarılı algoritma olduğu görülmektedir.

Bu bölümde çalışmada kullanılacak olan yöntemler anlatılacaktır. Bulgular ve Tartışma Bölümü'nde modellerin performans sonuçları değerlendirilecektir.

1.9.1. Karar Ağaçları (KA)

Karar Ağacı, ağaç yapısını andıran bir düzen ile belirli kurallar dizisi kullanarak hedef değişken üzerinde tahminlerde bulunmaktadır. Bu algoritma, genellikle insanların karar alma sürecinde izlediği yönteme benzer bir yaklaşımla çalışmaktadır. Bu sebeple, KA modelleri sezgisel olup anlaşılması ve açıklanması kolay olmaktadır.

Karar Ağacı algoritması, hem sınıflandırma hem de regresyon problemleri için yaygın olarak kullanılan bir yöntemdir. Bu teknik, veri kümesindeki basit karar kurallarını ve ilgili özellikleri kullanarak bir hedefi tahmin etmeye dayanmaktadır. Kolay yorumlanabilir olması ve çeşitli çıktılara sahip sorunları çözebilme kapasitesi, bu yöntemin başlıca avantajlarını oluşturmaktadır. Bununla birlikte, aşırı karmaşık ağaçlar oluşturarak aşırı öğrenme sorununa yol açabilmesi önemli bir dezavantajlarından biridir. Karar Ağacı algoritmasının genel yapısı Şekil 2'te sunulmaktadır.



Şekil 2. Örnek bir Karar Ağacı yapısı ve kuralları

KA algoritmaları, kümeleme ve tahmin problemlerinde sıklıkla tercih edilen bir algoritmadır. KA algoritmaları, veri eğitimi sırasında genelden özele doğru ilerleyen bir

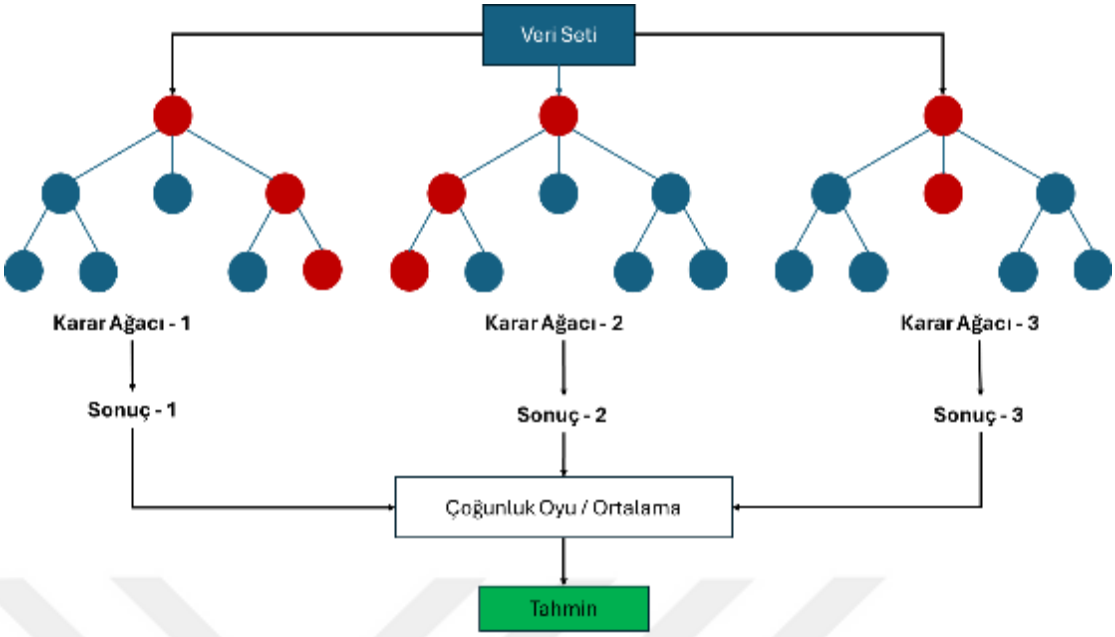
yapıya sahiptir. Akış şemasına benzer yapıda bu yapıda her düğümün öznitelik değeri hesaplanarak elde edilen değerlerle, dallar meydana gelmektedir. Verinin boyutu büyüdükçe algoritma içindeki dallanma işlemi zor olmaktadır. Aşırı öğrenmenin önlenmesi için yapak düğümleri kaldıran budama algoritmaları kullanılmaktadır (Hüseyin, 2024).

KA algoritmalarının en büyük avantajı karar verme sürecinin açıklanabilir ve yorumlanabilir olmasıdır. KA algoritmaları, veri dönüştürme veya veri normalleştirme işlemlerine gerek kalmadan hem kategorik hem de sürekli değişkenlerde çalışmaktadır. Bunun yanında kayıp değerlerde ve gürültülü veride de çalışabilmektedirler. Ayrıca çok özellikli karar kuralları oluşturarak verideki doğrusal olmayan ve hiyerarşik olan ilişkileri bulabilmektedir (Hüseyin, 2024).

1.9.2. Rastgele Orman (RO) Algoritması

Rastgele Orman algoritması, birden fazla karar ağacının birlikte kullanıldığı bir sınıflandırma yöntemidir. Sınıflandırma işlemlerinde, ağaç topluluğu çoğunluk kararı ile bir sonuca ulaşmaktadır. Bu algoritma, verilerin aşırı öğrenmesini (overfitting) önlemek ve tahminlerin doğruluğunu artırmak amacıyla kullanılmaktadır. Algoritma, her bir ağaç için rastgele alt örnekler ve özellikler seçerek bir "bagging" (bootstrap aggregation) yöntemi uygulamaktadır (Žižka vd., 2019).

RO'nın en önemli avantajlarından bir tanesi, aşırı öğrenme problemine karşı dirençli olmasıdır. Bu yöntemde her ağaç, verilerin sadece bir alt kümesine dayanarak öğrenme gerçekleştirmekte ve bu ağaçların her biri üzerinde rastgele özellik seçimi yapılmaktadır. Bu yöntem, karar ağaçları arasındaki korelasyonu azaltmakta ve modelin genel performansını olumlu yönde etkilemektedir (Qamar ve Raza, 2024).



Şekil 3. Rastgele Orman algoritması karar yapısı

1.9.3. Destek Vektör Makineleri Sınıflandırıcısı (DVMS)

Destek Vektör Makineleri (DVM), hem sınıflandırma hem de regresyon problemleri için yaygın olarak kullanılan güçlü makine öğrenimi algoritmalarından biridir. Bu algoritmalar, veriyi iki sınıfa en iyi şekilde ayıran hiperdüzlemi belirleyerek çalışmaktadır. Hiperdüzlem, veri uzayını ikiye bölen bir düzlem ya da çizgi olarak tanımlanmaktadır. Bu algoritmanın amacı, iki sınıf arasındaki marjin mesafesini maksimize eden hiperdüzlemi bulmaktır. Bu mesafe, hiperdüzlem ile her iki sınıfın en yakın örnekleri arasındaki uzaklık olarak ölçülür ve bu mesafenin büyük olması, sınıfların daha iyi ayrıldığı anlamına gelmektedir (Hersh, 2008).

DVM, parametrik olmayan bir modeldir; dolayısıyla, veri dağılımının şekline ilişkin herhangi bir varsayımda bulunmamaktadır. Bu sayede geniş bir yelpazede problem çözmek için kullanılmaktadır. DVM'lerin başlıca avantajları şunlardır:

Doğruluk: Büyük veri setleri üzerinde eğitildiğinde yüksek doğruluk seviyelerine ulaşabilmektedir.

Yorumlanabilirlik: Sonuçların anlaşılması ve hata ayıklama açısından kolay yorumlanabilir modellerdir.

Ölçeklenebilirlik: Büyük veri setleriyle başa çıkabilecek şekilde ölçeklenebilir, bu da SVM'ni gerçek dünya uygulamaları için ideal bir model yapmaktadır.

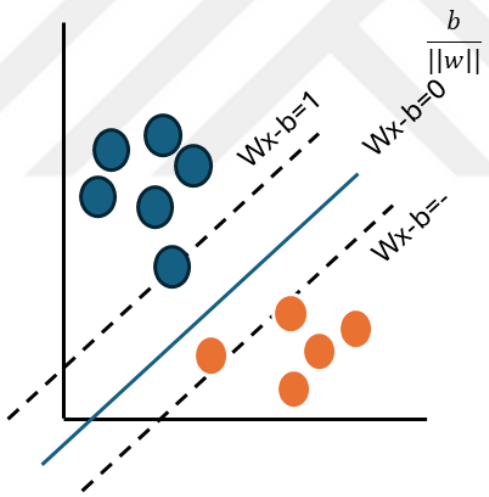
Sağlamlık: Gürültüye ve aykırı değerlere karşı dirençli olmaları nedeniyle temiz olmayan verilerde sıklıkla tercih edilmektedir.

Bununla birlikte, algoritmanın bazı zorlukları da bulunmaktadır. Bu zorluklar şunlardır:

Hesaplama maliyeti: Özellikle büyük veri setleriyle çalışırken eğitilmesi yoğun hesaplama gücü gerektirmektedir.

Hiperparametre ayarı: Performansını artırmak için çeşitli hiperparametrelerin optimize edilmesi gerekmektedir. Bu süreç zaman alıcı olabilmektedir (Qamar ve Raza, 2024).

DVM'ler, çeşitli alanlarda etkili bir şekilde kullanılmaktadır. Yüksek boyutlu verilerle çalışırken DVM'nin başarısı, özelliklerin birbirleriyle korelasyonlu olduğu ve az sayıda özelliğin gereksiz olduğu durumlarda ortaya çıkmaktadır. Bu nedenle, DVM metin verisi gibi seyrek ve yüksek boyutlu veri setleri için sıkça tercih edilmektedir (Demner-Fushman vd., 2009).



Şekil 4. Destek Vektör Makineleri

1.9.4. Lojistik Regresyon (LR)

Lojistik regresyon, makine öğrenimi ve istatistik alanında yaygın olarak kullanılan bir yöntemdir. LR, bağımlı bir değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi incelemek amacıyla kullanılmaktadır. Özellikle bağımlı değişkenin kategorik olduğu durumlarda tercih edilen bir modelleme yöntemidir. Bu model, sınıflandırma problemlerinde, bir olayın meydana gelip gelmeyeceğini tahmin etmek için yaygın olarak kullanılmaktadır (Isaac ve Harikumar, 2016).

LR'nin temel amacı, bağımlı değişkenin olasılığını tahmin etmek için bir lojistik fonksiyon (sigmoid fonksiyon) kullanmaktır. Bu fonksiyon, sonsuz bir aralıkta yer alan bağımsız değişken değerlerini sınırlı bir aralıkta (0 ve 1) ifade etmektedir. Bundan dolayı lojistik regresyon sınıflandırma problemlerinde kullanılabilir hale gelmektedir. Bu fonksiyon Formül 1'deki gibi tanımlanmaktadır.

$$P(Y = 1|X) = \frac{1}{1 + e^{(b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n)}} \quad (1)$$

Formül 1, bağımsız değişkenlerin değerlerine göre belirli bir sınıfa ait olma olasılığını hesaplamaktadır. Buradaki $b_1, b_2, \ldots, b_n$bağımsız değişkenlerin katsayılarını temsil etmektedir.

LR, bağımlı değişkenin sürekli değil, ikili (binary) olduğu durumlarda kullanılmaktadır. Bu nedenle, modelin sonuçları bir sınıfın meydana gelme olasılığına dayanmaktadır. Örneğin, bir hastanın hastalığa sahip olma olasılığını veya bir ürünün tercih edilip edilmeyeceğini belirlemek için kullanılmaktadır (Prabhat ve Khullar, 2017).

LR, makine öğrenimi uygulamalarında sıklıkla kullanılmakta ve büyük veri kümelerinde başarılı bir şekilde çalışmaktadır. Özellikle doğal dil işleme, tıp, finans gibi alanlarda çeşitli sınıflandırma problemlerini çözmek için tercih edilmektedir. LR, ikili sınıflandırma problemlerinde etkin bir yöntem olup, özellikle doğal dil işleme ve makine öğrenimi alanlarında geniş bir kullanım alanına sahiptir. Bu modelin performansı, veri seti ve problem tipiyle doğrudan ilişkilidir. Bundan dolayı lojistik regresyon, birçok sınıflandırma probleminde başarılı sonuçlar vermektedir (Wen vd., 2022).

1.9.5. Naive Bayes (NB) Algoritması

Naive Bayes algoritması, olasılıksal bir öğrenme yöntemi olup, her bir özelliğin bağımsız olduğunu varsayarak sınıflandırma işlemini yapmaktadır. Bu algoritma, bir metnin ya da veri örneğinin belirli bir kategoriye ait olma olasılığı, bağımsız özelliklerin koşullu olasılıklarının çarpımı ile hesaplanmaktadır. Ancak, bu varsayım pratikte her zaman doğru olmasa da, özellikle metin sınıflandırma gibi bazı uygulamalarda oldukça başarılı sonuçlar verebilmektedir (Oğul vd., 2017).

Bayes Teoremi

NB algoritması, istatistiksel bir sınıflandırma yöntemidir. Özellikle büyük boyutlu veri kümelerinde hızlı ve etkili sonuçlar sunmasıyla bilinmektedir. NB algoritması, Bayes teoremine dayanarak her özelliğin sınıfa bağımsız katkı yaptığı varsayımına dayanmaktadır. Bu nedenle, "naive" yani “saf” olarak adlandırılır, çünkü bu varsayım gerçek dünyadaki verilerde geçerli değildir. Fakat birçok uygulamada dikkate değer performans göstermektedir.

NB sınıflandırıcısı, her özelliğin sınıf etiketine koşullu olasılığını hesaplayıp, bu olasılıkları birleştirerek en yüksek olasılığa sahip sınıfı tahmin etmektedir. Bayes Teorisi formülü Formül 2’de ifade edilmektedir.

$$P(A|B) = \frac{P(A).P(B|A)}{P(B)} \tag{2}$$

P(A|B) : B olayı gerçekleştiğinde A olayının da gerçekleşmesi olasılığı

P(A): A olayının gerçekleşme olasılığı

P(B|A) : A olayı B olayı gerçekleştiğinde A olayının da gerçekleşmesi olasılığı

P(A) : B olayının gerçekleşme olasılığı

NB, her bir özelliğin değerinin, belirli bir kategori için nasıl bir olasılıkla gerçekleşeceğini hesaplamaya dayanmaktadır. Özellikler $x_1, x_2, \dots, x_n$ olarak ifade edilmektedir. Bu özelliklerin belli bir kategori C_k altında gerçekleşmesi olasılığı Formül 3’deki gibi tanımlanmaktadır.

$$P(x|C_k) = P(x_1|C_k).P(x_2|C_k).\dots.P(x_d|C_k) = \prod_{i=1}^{d} P(x_i|C_k) \tag{3}$$

Formül 3’de NB’nin temel varsayımı olan bağımsızlık ilkesine ifade edilmektedir. Her bir özellik x_i, diğer özelliklerden bağımsız olarak değerlendirilmektedir. Bu varsayım genellikle gerçek dünyada tam olarak sağlanmasa da, pratikte başarılı sonuçlar üretmektedir.

Bir metnin belirli bir kategoriye ait olma olasılığı, eğitim verilerinde o kategoriye ait örneklerin frekansları kullanılarak hesaplanmaktadır. Örneğin, bir eğitim veri setinde N_k

adet C_k kategorisinde örnek bulunuyorsa ve bu örneklerden N_{ki} tanesinde x_i özelliği gözlemleniyorsa, bu durumda $P(x_i|C_k)$ olasılığı Formül 4'deki gibi ifade edilmektedir.

$$P(x_i|C_k) = \frac{N_{ki}}{N_k} \tag{4}$$

Fakat bir özellik değeri bazı durumlarda hiç gözlemlenmemektedir. Bu durumda NB algoritmasında sıfır olasılık problemi ortaya çıkmaktadır. Bu problemin önlenmesi için "smooth" olarak ifade edilen bir düzeltme yöntemi kullanılmaktadır. Bu düzeltme yöntemiyle her olasılığa bir sabit değer eklenerek sıfır olasılık problemi önlenmektedir. Bu düzeltme işlemi Formül 5'deki gibi ifade edilmektedir.

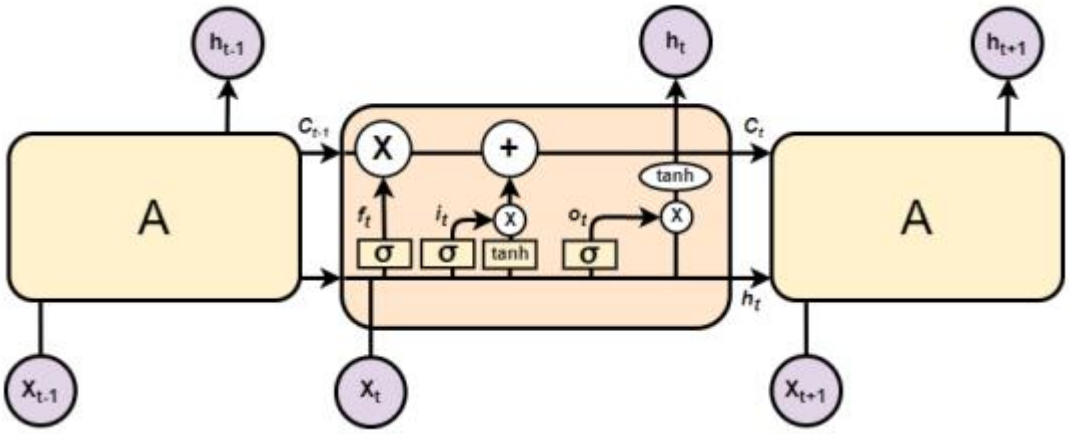
$$P(x_i|C_k) = \frac{B+N_{ki}}{B+N_k} \tag{5}$$

Formül 5'deki B ifadesi sabit bir değeri ifade etmektedir. Bu düzeltme işlemiyle, özellik değeri hiç gözlenmeyen bir değer olması durumunda, olasılığın sıfır gelmesi engellenmektedir. NB algoritması, metin sınıflandırma ve duygu analizi gibi alanlarda yaygın şekilde kullanılmaktadır (Jo, 2019).

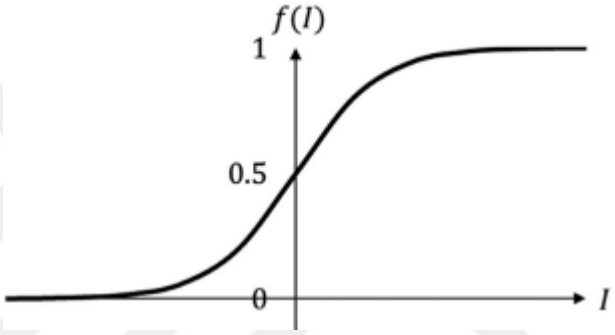
1.9.6. Uzun Kısa Süreli Bellek Algoritması (UKSB)

Uzun Kısa Süreli Bellek algoritması ilk olarak Sepp Hochreiter ve Jürgen Schmidhuber tarafından 1997 yılında tanıtılmıştır. Bu model, geri yayılım yöntemlerinin uzun zaman aralıklarında bilgi depolama ve öğrenme problemlerini çözmek amacıyla geliştirilmiştir. UKSB, Tekrarlayan Sinir Ağlarının (TSA) bir varyasyonu olup, özellikle ardışık veriyle çalışırken oldukça etkilidir. UKSB algoritması, zamanla bilgiyi saklama kapasitesi sayesinde uzun bağımlılıkları öğrenebilmektedir (Hochreiter ve Schmidhuber, 1997).

UKSB'nin ortaya çıkışında temel sorun, TSA algoritmalarının öğrenme sürecinde karşılaşılan "kaybolan gradyan" ve "patlayan gradyan" problemleridir. Bu problemler, zaman içinde hata sinyallerinin ya çok hızlı bir şekilde sıfıra doğru azalmasına ya da aşırı büyüyerek dengesiz hale gelmesine yol açmaktadır. Bu durum, uzun süreli bağımlılıkların öğrenilmesini zorlaştırmaktadır. UKSB, bu sorunu çözmek için özel bir hücre yapısı ve kapı mekanizmaları kullanmaktadır.



Şekil 5. UKSB Mimarisi (Nergiz vd., 2019)



Şekil 6. Sigmoid Fonksiyonu (Kılınç vd., 2022)

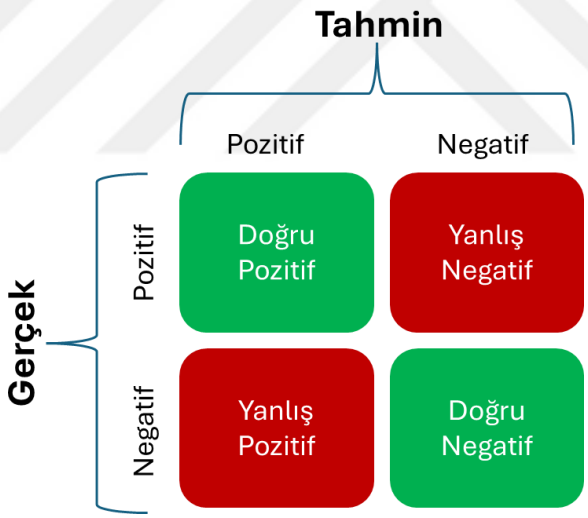
UKSB'nin en önemli bileşeni, "hafıza hücreleri" olarak adlandırılan ve bilgiyi uzun süre hafızada tutabilen hücrelerdir. Bu hücreler, geri yayılım sırasında bilginin kaybolmasını önlemek için tasarlanmıştır ve hata sinyallerinin sabit kalmasını sağlamaktadır. UKSB hücreleri, üç temel kapı birimi içermektedir. Bunlar giriş kapısı (input gate), çıkış kapısı (output gate) ve unutma kapısı (forget gate). Bu kapılar, bilginin ne zaman hücreye gireceğine, ne zaman hücrede saklanacağına ve ne zaman çıkış olarak kullanılacağına karar vermektedir. Bu sayede model, gereksiz bilgiyi unutarak önemli bilgileri uzun süreli bellekte saklayabilmektedir (Kılınç vd., 2022).

Geleneksel geri yayılım algoritmalarına kıyasla UKSB, uzun zaman aralıklarındaki bağımlılıkları daha iyi bir şekilde öğrenmektedir. Bundan dolayı, özellikle dil işleme, zaman serisi analizi ve konuşma tanıma gibi ardışık veri gerektiren uygulamalarda sıklıkla tercih edilmektedir (Küçük ve Arıcı, 2018).

1.10. Model Performansı İçin Değerlendirme Metrikleri

Geliştirilen makine öğrenimi modellerinin performansını değerlendirmek, modelin başarılı olup olmadığını ve gerçek dünyadaki verilere ne kadar iyi genellenebileceğini anlamak için kritik bir öneme sahiptir. Modelin performansını ölçmek için farklı metrikler kullanılmaktadır. Kullanılan bu metrikler, sınıflandırma algoritmalarının başarısını belirlemektedir. Geliştirilen modelin kalitesi, doğruluk (accuracy), kesinlik (precision), geri çağırma (recall) ve F1-skoru gibi metriklerle ölçülmektedir (Žižka vd., 2019).

Karmaşıklık Matrisi (Confusion Matrix): Karmaşıklık matrisi çok sınıflı sınıflandırma problemlerinde, model performansını sınıf bazında gösteren önemli bir göstergedir. Matris, model tarafından her sınıfa ait doğru ve yanlış tahmin edilen kayıt sayısından oluşmaktadır. Karmaşıklık matrisi, model tarafından tahmin edilen durumları ifade eden tahmin edilen sınıf ve gerçek durumları ifade eden gerçek sınıf olmak üzere iki eksenden oluşmaktadır. Bu eksenler sınıf sayısı kadar sütun ve satırdan meydana gelmektedir (Tosun, 2021).



Şekil 7. Karmaşıklık Matrisi

Doğru Pozitif (TP: True Positive): Gerçek pozitif değeri, model tarafından doğru tahmin edilen pozitif sınıfına ait veri sayısını ifade etmektedir.

Doğru Negatif (TF: True Negative): Doğru negatif değeri, model tarafından tahmin edilen negatif sınıfına ait veri sayısını ifade etmektedir.

Yanlış Negatif (FN: False Negative): Yanlış Negatif değeri, gerçekte pozitif sınıfa ait olan fakat model tarafından negatif olarak tahmin edilen verilerin sayısını ifade etmektedir.

Yanlış Pozitif (FP: False Positive): Yanlış Pozitif değeri, gerçekte negatif sınıfa ait olan fakat model tarafından pozitif olarak tahmin edilen verilerin sayısını ifade etmektedir (Alrefaai, 2021).

Sınıflandırma algoritmalarının kalitesi doğruluk (accuracy), keskinlik (precision), duyarlılık (recall) ve F1-skoru gibi metriklerle ölçülmektedir.

Accuracy (Doğruluk): Doğruluk, bir modelin tüm sınıflandırma örnekleri arasında doğru sınıflandırdığı örneklerin oranıdır. Bu metrik, özellikle dengeli veri kümelerinde anlamlı sonuçlar sağlar, ancak sınıflar arasında dengesizlik olduğunda yanıltıcı olabilmektedir. Doğruluk metriğinin hesaplanması için kullanılan formül, Formül 6'da ifade edilmektedir.

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision (Kesinlik): Kesinlik, pozitif sınıfa ait olduğu tahmin edilen örneklerden, kaçı gerçekten pozitif sınıfa aittir sorusunu cevaplamaktadır. Bu metrik, modelin yanlış pozitif tahminlerde bulunma eğilimini ölçer ve özellikle yanlış pozitiflerin maliyeti yüksek olduğunda önemli bir göstergedir. Kesinlik metriğinin hesaplanması için kullanılan formül, Formül 7'de ifade edilmektedir.

$$Kesinlik = \frac{TP}{TP + FP} \tag{7}$$

Recall (Duyarlılık): Duyarlılık metriği, gerçekte pozitif olan örneklerin kaç tanesini doğru şekilde pozitif olarak tahmin edildiğini göstermektedir. Duyarlılık, yanlış negatiflerin önemli olduğu durumlarda (yani, pozitif olup yanlış şekilde negatif tahmin edilen örnekler) kritik bir ölçüt olmaktadır. Duyarlılık metriğinin hesaplanması için kullanılan formül, Formül 8'de ifade edilmektedir.

$$Duyarlılık = \frac{TP}{TP + FN} \tag{8}$$

Micro F1-Skoru: Bu metrik çoklu sınıflandırma problemlerinde her bir sınıf için hesaplanan F1 skorlarının birleştirilerek elde edildiği bir ölçüttür. Micro F1-Skoru tüm sınıflar üzerinden elde edilen genel performansı yansıtır ve özellikle dengesiz veri setlerinde

faydalı olmaktadır. Micro F1 skoru metriğinin hesaplanması için kullanılan formül, Formül 9'da ifade edilmektedir.

$$Micro\ F1 - Skoru = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{9}$$

Makro F1-Skoru: Makro F1 skoru, çoklu sınıflandırma problemlerinde her bir sınıf için ayrı ayrı hesaplanan F1 skorlarının aritmetik ortalamasını alarak hesaplanan bir performans ölçüsüdür. Bu skorda, her bir sınıf eşit ağırlıkla değerlendirilmekte ve sınıf büyüklüğünden bağımsız olarak her sınıfın F1 skoru, toplam skorun hesaplanmasına katkıda bulunmaktadır. Makro F1 skoru metriğinin hesaplanması için kullanılan formül, Formül 10'da ifade edilmektedir.

$$Macro\ F1 - Skoru = \frac{1}{|C|}\sum_{i=1}^{|C|} F1_i \tag{10}$$

Burada;

$|C|$: Toplam sınıf sayısıdır.

$F1_i$: her sınıf i için hesaplanan F1 skorunu ifade etmektedir.

Makro F1 skoru, her bir sınıfın eşit ağırlığa sahip olduğu bir ölçüttür. Bundan dolayı, özellikle veri setinde dengesizlik varsa, bu sınıfların performansları makro F1 skoruna daha fazla etki eder. Mikro F1 skoru ise, sınıflar arası farkları dikkate alarak tüm doğru pozitifler, yanlış pozitifler ve yanlış negatifler üzerinden hesaplandığı için, büyük sınıfların performansı üzerinde daha fazla durur. Bu nedenle, veri dengesizliği durumlarında makro F1 skoru küçük sınıfları daha fazla temsil ederken, mikro F1 skoru genel model performansını daha iyi yansıtır (Žižka vd., 2019).

2. YAPILAN ÇALIŞMALAR

Bu tez çalışması, Albaraka Tech Global firması tarafından firma içi proje olarak desteklenmeye uygun bulunmuştur. Sistem için gerekli altyapı ile teknik destek firma tarafından sağlanmaktadır. Çalışma kapsamında, öncelikle geliştirilecek sistemin çalışabilmesi için gerekli donanım altyapısı belirlenmesi gerekmektedir. Bu doğrultuda, sistemin verimli çalışmasını sağlamak amacıyla sistem, üç ana bileşen şeklinde yapılandırılmıştır. Bu bileşenler; "İnternet Gezgini ve Analiz Sunucusu", "Veri Tabanı Sunucusu" ve "Arayüz Sunucusu".

İnternet Gezgini ve Analiz Sunucusunun temel işlevi, belirlenen haber kaynaklarından internet haber metinlerini toplamak ve bu metinlerden gerekli bilgileri çıkarmaktır. Veri Tabanı Sunucu'sunun görevi, toplanan haber metinleri ile analiz sonuçlarını depolamaktır. Arayüz Sunucusu ise kullanıcıların firma arama motorunu kullanmalarını sağlamak ve analiz sonuçlarını görüntülemek amacıyla hizmet vermektedir.

Sistem içindeki bileşenlerin özellikleri Tablo 1'deki gibidir.

Tablo 1. Sistem bileşenlerine ait teknik özellikler

Sistem Bileşeni	Özellikleri
İnternet Gezgini ve Analiz Sunucusu	4 CPU,16 GB RAM, 250 GB Disk, RedHat
Veri Tabanı Sunucusu	2 CPU, 8 GB RAM, 500 GB Disk (MongoDB)
Arayüz Sunucusu	2 CPU, 4 GB RAM, 60 GB Disk, Ubuntu

Şekil 8. Sistemin temel bileşenleri

Sistem bileşenlerinin geliştirilmesi sürecinde kullanılan teknolojik tercihlerin projeye uygunluğu dikkatlice değerlendirilmiştir. Şekil 8'de sistem bileşenlerinin genel mimarisine yer verilmiştir. Yazılım geliştirme aşamasında Python 3.9 programlama dili tercih edilmiştir.

Python, geniş kütüphane desteği, esneklik ve veri işleme yetenekleri nedeniyle bu projede önemli bir rol oynamaktadır. Python dilinin sunduğu esneklik, veri toplama, işleme ve analiz süreçlerinin etkili bir şekilde yürütülmesini sağlamaktadır.

Sistemin ilk bileşeni olan İnternet Gezgini ve Analiz Sunucu'su, belirlenen haber kaynaklarından veri toplama ve analiz yapma işlevini yerine getirmektedir. Haberlerin internet üzerinden toplanması için Python'ın "BeautifulSoup" ve "Requests" gibi kütüphaneleri kullanılmaktadır. Bu kütüphaneler, haber kaynaklarının taranarak içeriklerin elde edilmesine olanak tanımaktadır. Toplanan haber metinlerinin analiz edilmesi ve anlamlı bilgilere dönüştürülmesi amacıyla doğal dil işleme süreçleri kullanılmaktadır. Bu aşamalar, haber metinlerinin derinlemesine analizini sağlarken, metinlerden bankacılık açısından kritik bilgilerin çıkarılmasına da imkan tanımaktadır.

İnternet Gezgini ve Analiz bileşeni için güvenlik riskleri dikkate alınarak Red Hat Linux işletim sistemi tercih edilmiştir. Red Hat, özellikle banka gibi güvenlik açısından hassas sistemlerde kullanılan, kurumsal seviye güvenlik ve destek suFnan bir işletim sistemi olması nedeniyle tercih edilmektedir. Banka sistemlerinde daha güvenli hizmet verebilmek ve veri toplama süreçlerinin güvenlik ihlali riski olmadan yürütülebilmesi için Red Hat tercih edilmektedir. Bununla birlikte, Red Hat'in sağladığı güvenilir performans ve sertifikalı güvenlik özellikleri, sistemin ihtiyaçlarını karşılamak açısından ideal bir çözüm sunmaktadır.

Toplanan verilerin güvenli ve verimli bir şekilde saklanması için Veri Tabanı Sunucusu kurulumu gerçekleştirilmiştir. Bu aşamada sistem için en ideal veri tabanı olarak MongoDB veri tabanı edilmiştir. MongoDB, özellikle yapılandırılmamış ve yarı yapılandırılmış büyük veri kümelerini depolama konusunda güçlü bir NoSQL veri tabanı çözümü olarak tercih edilmektedir. MongoDB'nin sağladığı esnek yapı, haber verileri ile analiz sonuçlarının sorunsuz bir şekilde saklanmasına ve hızlı erişimine imkan tanımaktadır. Veri tabanı yapısı, sistemin ölçeklenebilirliğini artırırken, aynı zamanda büyük hacimli verilerle çalışmayı da kolaylaştırmaktadır.

Sistemin üçüncü bileşeni olan Arayüz Sunucusu, kullanıcıların sisteme erişim sağlaması ve analiz sonuçlarını görselleştirmesi için kullanılmaktadır. Bu amaçla Python'ın "Streamlit" kütüphanesi kullanılmaktadır. Streamlit, interaktif internet arayüzleri geliştirme konusunda kolaylık sağlayan bir araç olup, veri analiz sonuçlarının kullanıcılarla etkileşimli bir şekilde paylaşılmasını sağlamaktadır. Kullanıcı arayüzünün basit ve etkili bir şekilde

sunulması, sistemi kullanacak olan banka çalışanlarının hızlı ve doğru bilgilere erişimine imkan sunmaktadır

Arayüzün arka plan işlemlerini yöneten servis tabanlı mimari için ise Python tabanlı FastAPI framework'ü kullanılmaktadır. FastAPI, Python tabanlı modern bir framework olup, asenkron yapısıyla yüksek performanslı internet servisleri geliştirmeyi kolaylaştırmaktadır. Bu framework, sistemin farklı bileşenleri arasında hızlı ve güvenilir bir iletişim sağlamaktadır. Aynı zamanda kullanıcıların taleplerini verimli bir şekilde karşılayarak sistemi daha hızlı ve esnek hale getirmektedir.

Bu teknik altyapı sayesinde, sistemin veri toplama, işleme, depolama ve kullanıcı etkileşimi süreçleri optimize edilmekte ve yüksek performanslı ve ölçeklenebilir bir yapı oluşturulmaktadır. Geliştirilen sistem, büyük veri kümeleriyle çalışmayı sağlamaktadır. Kullanıcı dostu bir arayüz aracılığıyla bankacılık analizlerini etkin bir şekilde sunulmaktadır.

Sistemin en iyi performansla çalışması için gerekli olan teknik gereksinimler belirlendikten sonra sistemin geliştirilmesi için ilk olarak verilerin toplanması işlemine başlanılmıştır.

Çalışma kapsamında geliştirilecek sistemde en kritik rol, İnternet Gezgini ve Analiz Bileşenine aittir. Bu bileşenin etkili bir şekilde geliştirilebilmesi amacıyla, üç farklı modül oluşturulmasına karar verilmiştir: İnternet Gezgini Modülü, Analiz Modülü ve Veri Tabanı Modülü. Bu modüller, sistemin veri toplama, analiz etme ve depolama süreçlerini ayrı ayrı optimize etmek amacıyla yapılandırılmıştır. Buradaki her bir modül, kendi işlevsel gereksinimlerine uygun olarak geliştirilmiştir. Bu bağlamda, sistemin verimli bir şekilde çalışabilmesi için verilerin toplanmasını sağlayacak olan bu yapının geliştirilmesi büyük önem arz etmektedir.

Tez kapsamında dijital ortamda yayın yapan haber siteleri veri kaynağı olarak kullanılmaktadır. Bu haber kaynaklarından haber içeriklerinin otomatik olarak toplanabilmesi için bir "internet gezgini" yapısının geliştirilmesine ihtiyaç duyulmaktadır. İnternet gezgini, internet üzerinde belirli haber sitelerinden veri toplama, işleme ve analiz yapma süreçlerinde temel bir araçtır.

Çalışmanın geliştirme sürecinde, veri toplama işlemlerine başlamadan önce, kapsamı belirlemek amacıyla bir pilot çalışma yapılmıştır. Pilot çalışma, toplam 20 haber sitesini kapsayacak şekilde planlanmıştır. Bu süreçte, her bir haber sitesinin kaynak kodları incelenmiş ve siteler arasındaki tasarım farklılıkları tespit edilmiştir. İncelenen 20 haber

sitesinin tasarımlarının birbirinden önemli ölçüde farklı olduğu görülmüştür. Türkiye genelinde yayın yapan tüm haber siteleri göz önüne alındığında, bu tasarım farklılıklarının sayısının ve çeşitliliğinin çok daha fazla olacağı öngörülmektedir. Bu farklılıkların veri toplama sürecinde zorluk yaratmaması için standart bir veri süpürme yapısı geliştirilmesine ihtiyaç duyulmuştur.

Pilot çalışma kapsamında, her bir haber sitesinin tasarımına özgü farklılıkların üstesinden gelebilmek amacıyla standart bir içerik süpürme veri rehberi oluşturulmuştur. Bu standart içerik süpürme veri rehberi, her haber sitesinin kod içeriği incelenerek, veri toplama işlemlerini tutarlı bir hale getirilmiştir. Oluşturulan bu yapı sayesinde, sitelerin farklı tasarımsal özelliklerinden kaynaklanan zorluklar, belirlenen veri yapılarıyla aşılmış ve veri toplama işlemi daha sistematik hale getirilmiştir.

Oluşturulan veri yapısı, haber sitelerinden veri çekebilmek için gerekli olan temel HTML bilgilerini içermektedir. Bu yapı üç farklı alt veri bileşenine ayrılmaktadır, Bunlar birincisi haber kaynağına ait genel bilgiler, ikincisi haber kaynağındaki haber linklerine ilişkin bilgiler ve üçüncüsü haber içeriklerine ilişkin bilgilerdir. Bu veri yapıları, her bir haber kaynağı için özelleştirilmiş ve veri toplama süreçlerinin daha etkin bir şekilde yönetilmesine olanak sağlamaktadır.

```
news_source_dict = [
    {"_id": "milliyet",
    "agency_name": "Milliyet",
    "links": [
        "https://www.milliyet.com.tr/ekonomi/?page=1",
        "https://www.milliyet.com.tr/ekonomi/?page=2",
        "https://www.milliyet.com.tr/ekonomi/?page=3",
        "https://www.milliyet.com.tr/ekonomi/?page=4"
    ]
}
]
```

Şekil 9. Haber kaynaklarına ait bilgilerin yer aldığı veri yapısı

Haber kaynağı veri yapısı, geliştirilen sistemin internet üzerinden haber toplayabilmesi için gereken internet adreslerini içermektedir. Bu yapı, her haber kaynağına ait benzersiz bir kimlik ve ilgili haber ajansının ismini içeren verilerden oluşmaktadır. Ayrıca, sistemin veri toplama işlemini gerçekleştirebilmesi için gerekli olan haber sayfalarının internet adresleri

de bu yapıda yer almaktadır. Örneğin, Şekil 10'daki veri yapısında görüldüğü gibi, belirli bir haber kaynağından (örneğin; Milliyet) ekonomi kategorisindeki dört farklı sayfaya ait bağlantılar tanımlanmaktadır. Bu şekilde, sistem ilgili sayfalara erişerek haber içeriklerini çekmekte ve işlemektedir. Bu veri yapısı, sistemin düzenli ve tutarlı bir şekilde farklı kaynaklardan veri toplamasını sağlamak amacıyla oluşturulmuştur.

```
news_links_guide_dict = [
    {"_id": "milliyet",
    "link_key": "/ekonomi",
    "link_key_location": "stars",
    "source_url": "https://www.milliyet.com.tr"}
]
```

Şekil 10. Haber kaynağı içerisindeki linklere ait bilgiler örneği

Haber kaynağı içindeki linklere ait bilgiler, sistemin belirli bir haber sitesinden veri toplama işlemlerini doğru bir şekilde gerçekleştirebilmesi için büyük öneme sahiptir. Bu yapıda, her haber kaynağına özgü bir kimlik (örneğin, "milliyet") ile birlikte, ilgili haber kategorisine veya bölüme işaret eden bir "link_key" (örneğin, "/ekonomi") bulunmaktadır. Ayrıca, bu anahtar kelimenin internet sayfası adresinin başlangıcında mı, ortasında mı yer aldığını belirten "link_key_location" alanı tanımlanmıştır (bu örnekte, "starts"). Bu bilgi, sistemin haber içeriklerini çekmek için hangi internet sayfalarının kullanılacağını ve ilgili sayfaları nasıl bulacağını belirlemede yardımcı olmaktadır. Son olarak, haber kaynağının temel internet adresi "source_url" alanında belirtilmiştir. Bu yapı sayesinde, sistem farklı haber kaynaklarının bağlantılarını etkili bir şekilde tespit edip analiz edebilir.

```
news_content_guide_dict = [
    {"agency_name": "milliyet",
    "target_name": "div",
    "target_arg": "class",
    "target_arg_value": ["news-content readingTime"],
    "target_date_name": "p",
    "target_date_arg": "arg",
    "target_date_arg_value": ["news-detail-text"]}
]
```

Şekil 11. Haber içeriğinin tespiti için kullanılan bilgiler örneği

Haber içeriğine ait bilgilerin yer aldığı bu yapı, sistemin haber sitelerinden doğru ve tutarlı bir şekilde içerik çekebilmesi için belirli HTML etiketlerini hedef alacak şekilde oluşturulmuştur. Bu yapı, her bir haber kaynağının HTML yapısına özgü olarak haber içeriğinin hangi HTML etiketlerinde yer aldığını tanımlar. Örneğin Şekil 11'de, "agency_name" "milliyet" olarak belirtilmiş olup, haber içeriği div etiketindeki "class" özelliği "news-content readingTime" olan bir alan içerisinde yer almaktadır. Aynı şekilde, haberin tarihi de "p" etiketi içinde, class değeri "news-detail-text" olan bir alanda bulunmaktadır. Bu bilgiler, sistemin haber içeriği ve tarihi gibi kritik verileri doğru bir şekilde çekebilmesi için gerekmektedir. Bu sayede, farklı haber kaynaklarının çeşitli HTML yapılarından veri toplama işlemi standart bir biçimde gerçekleşmekte ve sistem tarafından işlenebilir hale getirilmektedir.

Oluşturulan veri yapısı, JSON dosya formatında depolanmaktadır. Bu tercih, ilgili haber kaynaklarında meydana gelebilecek tasarımsal değişikliklerin kolayca güncellenebilmesine olanak sağlamaktadır. JSON formatı, esnek ve yapılandırılmış bir veri modeli sunduğu için, haber sitelerindeki değişikliklerin hızlı ve etkili bir şekilde veri yapısına entegre edilmesine imkan tanımaktadır. Böylece, sistemin güncelliği ve işlevselliği sürdürülebilir hale getirilmektedir.

Bu yapının oluşturulabilmesi amacıyla, yerel, ulusal ve uluslararası olmak üzere toplamda 220 haber kaynağı incelenmiştir. Özellikle yerel haber sitelerinde firmalar hakkında çıkan haberlerin yakalanması büyük önem taşımaktadır. Zira küçük ölçekli firmalara ilişkin haberler, genellikle ana akım medya organlarında yer bulamamaktadır. Buna karşılık, yerel haber kaynaklarında küçük ölçekli firmalarla ilgili haberler daha sık yayınlanmaktadır. Bu durum, yerel haber sitelerinin sistemde dikkate alınmasını gerekli kılmaktadır.

Veri toplama işleminden önce belirlenen tüm kaynak siteler için bu çalışma tamamlanarak, süpürme işlemine hazırlık süreci tamamlanmıştır. Bu yapının kurulmasının ardından, analiz sürecinin en kritik aşaması olan metin sınıflandırma işlemi üzerinde çalışmalar tamamlanmıştır.

Bu noktada, modelleme sürecinin en önemli aşamalarından biri olan, haber kaynaklarından elde edilen içeriklerin doğru bir şekilde etiketlenmesi gündeme gelmektedir. Veri etiketleme süreci, veri hazırlama sürecinin en çok zaman alan ve modelleme başarısına doğrudan etki eden aşamalarındandır. Etiketleme işleminin kalitesi, geliştirilecek modelin performansını belirlemede büyük bir önem arz etmektedir. Bu nedenle, elde edilen verilerin

doğru ve tutarlı bir şekilde etiketlenmesi için farklı yöntemler ve veri kaynakları üzerinde kapsamlı çalışmalar yapılmıştır.

Veri etiketleme çalışmasının ilk aşamasında, Türkçe diline yönelik daha önce geliştirilmiş önceden eğitilmiş modeller kullanılmıştır. İlgili modellerin performansları değerlendirilmiştir. Bununla birlikte, bu modellerin performansını test etmek ve geliştirilecek modelin başarısını değerlendirmek amacıyla, bir etiketli test veri seti oluşturulmasına karar verilmiştir. Bu kapsamda, toplamda 300 haber metni, üç farklı kişi tarafından bağımsız olarak olumlu ve olumsuz şeklinde etiketlenmiştir. Bu etiketli veri seti, önceden eğitilmiş modellerin test edilmesinde kullanılmaktadır. Bu veri setiyle modelin haber içeriklerini doğru etiketleme başarısı %90 olarak belirlenmiştir. Bu sonuçlar, modelin genel olarak başarılı bir performans sergilediğini göstermektedir. Bu modelle etiketlenmemiş haber içeriklerini etiketleme işlemi gerçekleştirilmiştir.

Etiketleme işlemi için kullanılan model, Türkçe diline yönelik daha önce eğitilmiş bir model olan "Bert-base Turkish Sentiment Model" olmuştur. Bu model, film ve ürün incelemelerinden oluşan veri setleri ile Twitter verilerini kullanmaktadır (Yildirim, 2020). Ancak, bu modelin eğitildiği veri setlerinin haber metinlerinden farklı türde içerikler olması nedeniyle, haber metinleri üzerinde özel olarak eğitilmiş bir modelin daha iyi sonuçlar verebileceği öngörülmüştür. Özellikle, haber metinlerinin yapısal özellikleri ve dilsel farklılıkları göz önüne alındığında, bu alana özel bir modelin geliştirilmesinin performansı artıracağı düşünülmektedir. Bu bağlamda, haber metinleri üzerine odaklanacak bir model geliştirme ihtiyacı doğmuştur. Bunun için, haber metinlerinin sistematik bir şekilde etiketlenmesi ve ardından modelleme sürecine geçilmesi gerekmektedir (Demirtas ve Pechenizkiy, 2013; Hayran ve Sert, 2017).

Ancak, sistemde arşivlenen haber sayısının çok fazla olması, etiketleme işleminin manuel olarak yapılmasını imkânsız hale getirmektedir. Manuel etiketleme işleminin büyük zaman ve emek gerektireceği göz önüne alındığında, bu süreci hızlandıracak alternatif yöntemler araştırılmıştır. Küresel Olaylar, Dil ve Ton Veritabanı **(**KDTV) Projesi verilerinin farklı akademik çalışmalarda kullanıldığı görülmüştür (Mertoğlu, 2020; Sağlam, 2019).

KDTV Projesi (URL-1, 2024), Kalev H. Leetaru tarafından başlatılmış olup, dünya genelinde geniş çaplı bir veri tabanı oluşturmayı amaçlayan bir projedir. KDTV, küresel ölçekte gerçekleşen sosyal olayları ve insan hareketlerini kaydeden bir mekânsal olay veri tabanı olarak tanımlanmaktadır. Proje, dünya genelindeki tüm ana akım haber kaynaklarını, internet sitelerini, televizyon yayınlarını, akademik veri tabanlarını (CORE, DTIC, JSTOR

gibi) ve haber videolarını otomatik olarak taramaktadır. Ayrıca KDTV, yalnızca İngilizce değil, 65 farklı dili de desteklemekte olup, haber metinlerini gerçek zamanlı olarak İngilizce'ye çevirerek doğal dil işleme süreçlerine tabi tutmaktadır. Bu sistematik tarama ve analiz işlemleri, her 15 dakikada bir tekrarlanarak veri tabanına kaydedilmektedir. KDTV, bu özellikleriyle küresel ölçekte sürekli güncellenen ve genişleyen bir veri kaynağı sunmaktadır.

Bu çalışma kapsamında, Google BigQuery platformunda yer alan KDTV veri tabanı üzerinde araştırmalar yapılmış ve tez çalışmasına uygun veri kümelerinin elde edilmektedir. Yapılan araştırmalar ve uygun SQL sorgularının yazılması sonucunda, KDTV veri tabanından toplamda 85 bin etiketlenen Türkçe haber metni elde edilmektedir. Bu haber metinleri, duygu analizi yapabilmek için yeterli veriyi sağlamaktadır. Bunun yanı sıra, tez çalışmasına doğrudan katkı sağlayacak, firmalarla ilgili 2985 Türkçe haber metni de elde edilmektedir. Bu haberlerin, özellikle firmalar hakkındaki bilgi akışını ve haber sitelerindeki algıyı tespit etmeye yönelik olarak kullanılmasına karar verilmiştir.

KDTV projesinde, "Tone" değeri adı verilen bir değer hesaplanmaktadır. Ton değeri metinlerin duygusal eğilimlerini ölçen bir analiz yöntemi olarak kullanılmaktadır. Bu değer, bir metindeki olumlu ve olumsuz kelimelerin oranı dikkate alınarak hesaplanmaktadır. İlk aşamada, metindeki olumlu ve olumsuz kelimeler belirlenmektedir. Metindeki genel ton, olumlu kelimelerin toplamı ile olumsuz kelimelerin toplamının farkının, metindeki toplam kelime sayısına bölünmesi ile elde edilmektedir (GLOBE, 2020).

KDTV'in tone değerleri -100 ile +100 arasında bir aralıkta olup, -100 son derece olumsuz bir tonu, +100 ise son derece olumlu bir tonu göstermektedir. Çoğu metin için bu değer genellikle -10 ile +10 arasında yoğunlaşmaktadır ve 0 değeri nötr bir tonu ifade etmektedir. Tone değeri sıfıra yakın olduğunda, ya metnin düşük duygusal bir tepki içerdiği ya da olumlu ve olumsuz kelimelerin birbirini dengelediği anlamına gelmektedir (GLOBE, 2020).

KDTV'in bu hesaplama yönteminde, Harvard Üniversitesi'nin "General Inquire IV" gibi çeşitli duygu sözlüklerini kullandığı görülmektedir. General Inquirer IV, Harvard Üniversitesi tarafından geliştirilen ve doğal dil işleme ve metin analitiği alanlarında kullanılan önemli bir duygusal sözlüktür. Bu sözlük, metinlerdeki kelimelerin anlamlarını ve taşıdıkları duygusal ağırlığı analiz etmek amacıyla geliştirilmiştir. General Inquirer IV'nin, özellikle duygusal analiz ve metinlerdeki duygusal eğilimleri belirlemek amacıyla kullanıldığı görülmektedir. Bu sözlük, metinlerde yer alan kelimeleri çeşitli kategorilere

ayırarak bunların duygusal tonlarını tanımlamaktadır. Örneğin, bir kelimenin olumlu ya da olumsuz olup olmadığını belirlemektedir. Bu sayede, bir metnin genel duygusal tonu ölçülebilir ve pozitif ya da negatif eğilimler sayısal olarak ifade edilmektedir (URL-2, 2024).

KDTV gibi projelerde, General Inquirer IV gibi sözlükler, metinlerde yer alan kelimeleri olumlu veya olumsuz olarak sınıflandırmada ve bu kelimelerden yola çıkarak metnin tonunu hesaplamada kritik rol oynamaktadır. Bu tür sözlükler, metin madenciliği, duygu analizi ve dil analitiği projelerinde yaygın olarak kullanılmaktadır.

KDTV Projesinde, General Inquirer IV duygu sözlüğünün yanı sıra metinlerdeki duygusal eğilimleri yakalamak için gelişmiş içerik analizi araçları da kullanmaktadır. KDTV'in Global Content Analysis Measure (KİAÖ) algoritması, bu araçları bir araya getirerek metinlerdeki 2.230'dan fazla gizli boyutu rapor etmektedir. Böylece metinlerdeki ton, yer, tema ve duygusal eğilimler doğru bir şekilde analiz edilerek ve kodlanmaktadır (URL-3, 2024).

Küresel İçerik Analizi Ölçütü (KİAÖ), KDTV projesi tarafından kullanılan gelişmiş bir içerik analiz algoritmasıdır. Bu algoritma, dünya genelinde yayımlanan haberlerdeki duygusal, tematik ve içeriksel eğilimleri analiz etmek amacıyla geliştirilmiştir. KİAÖ, metinlerdeki gizli duygusal ve içeriksel boyutları tanımlayarak her bir haberin tonunu, temalarını ve diğer dilsel özelliklerini sayısal olarak değerlendirmektedir. Bu algoritma, binlerce dilsel ve duygusal kategoriyi işleyebilecek kapasiteye sahiptir.

KİAÖ algoritmasının çalışma prensibi incelendiğinde temel olarak dört adımdan oluştuğu görülmüştür.

Kelime Duyarlılığı ve Sınıflandırma: KİAÖ ilk olarak, metinlerdeki kelimeleri bir çok duygusal ve tematik kategorilere ayırmaktadır. Bu kategoriler, önceden belirlenmiş duygu sözlükleri (örneğin, General Inquirer IV gibi) kullanılarak kelimelerin olumlu, olumsuz ya da nötr olup olmadığını belirlemektedir.

Gizli Boyutların Yakalanması: KİAÖ, metinlerde bulunan 2.230'dan fazla gizli (latent) duygusal ve içeriksel boyutu tanımlayarak bunları ölçmektedir. Bu gizli boyutlar, metindeki karmaşık dil yapıların detaylı analiz edilmesini sağlamaktadır. Örneğin, bir haberin tonu, duygusal yükü ve içeriğin karmaşıklığı bu boyutlardan bazılarıdır.

Duygu ve Ton Hesaplaması: KİAÖ, kelime düzeyinde belirlenen duygu kategorilerini kullanarak her bir metindeki genel duygusal tonu (pozitif, negatif, nötr) hesaplamaktadır. Bu işlem yapılırken, metindeki olumlu ve olumsuz kelimeler sayılır ve bu kelimelerin oranına

göre genel bir ton skoru atanmaktadır. Skorlar incelendiğinde, her bir metin -100 ile +100 arasında bir ton değeri ile skorlanmaktadır.

Veri Çıktısı: Algoritma, bu analiz sonucunda elde edilen duygusal ve içeriksel verileri, metinle ilgili diğer parametrelerle (temalar, kişiler, olaylar, yerler vb.) birleştirmektedir. KİAÖ, bu bilgileri kullanarak her bir metin için kapsamlı bir analiz raporu sunmaktadır. Bu saye veriler hem niceliksel hem de niteliksel analizlerde kullanılabilir hale gelmektedir.

KİAÖ, metinlerde birden fazla duygusal boyutu aynı anda ölçmektedir. Bu, sadece olumlu ya da olumsuz duyguların değil, metindeki daha ince duygusal ve tematik unsurların yakalanmasına olanak tanımaktadır. Ayrıca algoritma, dünya çapındaki haber kaynaklarından gelen verileri her 15 dakikada bir güncelleyerek sürekli olarak yeni veri akışı sağlamaktadır. KİAÖ, büyük boyutlu metin analizi ve duygu analitiği gibi alanlarda güçlü bir araç olarak kullanılmaktadır. Özellikle, haber akışlarını analiz ederek dünya genelinde yaşanan olayların toplumsal ve duygusal yansımalarını ölçmekte yaygın olarak kullanılmaktadır (URL-3, 2024).

Bu süreç, KDTV'in dünya çapında yayımlanan haberleri analiz etmesinde önemli bir role sahiptir ve özellikle toplumsal olaylar, politik gelişmeler ve kriz durumları gibi konularda geniş bir veri kaynağı oluşturmaktadır. Sağlam, (2019) ve Mertoğlu, (2020) tarafından yapılan doktora çalışmaları ve KDTV projesinin detaylı incelemeleri sonucunda, bu veri kaynağının güvenilir olduğu ve modelleme sürecinde kullanılabileceği sonucuna varılmaktadır. Bu kapsamda, KDTV veri tabanından elde edilen etiketli haber içeriklerinin, geliştirilecek modellerin eğitilmesi ve test edilmesi sürecinde kullanılabileceği değerlendirilmektedir.

Bu süreçte hem önceden eğitilmiş modellerin test edilmesi hem de KDTV veri tabanından elde edilen etiketli haber verileriyle özel bir modelin geliştirilmesi hedeflenmektedir. Elde edilen verilerin doğru ve güvenilir olması, geliştirilecek modelin performansını olumlu yönde doğrudan etkilemesi beklenmektedir. KDTV veri tabanının sunduğu geniş veri yelpazesi, bu tez çalışmasına büyük katkı sağlayarak, haber içeriklerinin analizinde yüksek doğrulukta sonuçlar elde edilmesine imkan tanımaktadır.

Veri etiketleme sürecinin otomatikleştirilebilmesi amacıyla, veriye özgü bir algoritma geliştirilmiştir. Bu algoritma kullanılarak verilerin etiketlenmesi işlemi gerçekleştirmektedir. KDTV veri tabanından elde edilen verilerde, her bir haber metnine ilişkin tone (ton) değeri bulunmaktadır. KDTV projesi kapsamlı bir şekilde incelendiğinde, tone değerinin -10 ile +10 arasında değiştiği görülmektedir. Bu bilgiler doğrultusunda,

Türkçe haber metinleri tone değerleri dikkate alınarak etiketlenmektedir. Bu süreçte, eğer tone değeri -10'a yakınsa haber metni negatif, +10'a yakınsa pozitif; tone değeri 0'a 3 birim yakınsa nötr olarak etiketlenmiştir.

KDTV gibi uluslararası bir projede kullanılan Türkçe haberler birlikte Hugging Face platformunda yer alan "Turkish Sentiment Analysis Dataset" adlı veri seti çalışma kapsamında kullanılmaktadır. Bu veri seti Türkçe metinlerde duygu analizi gerçekleştirmek amacıyla geliştirilmiş zengin bir veri kümesidir. Bu veri seti kullanılarak Türkçe metinlerdeki duygusal eğilimleri analiz edebilen bir model geliştirilmiştir (URL-4, 2024).

Tablo 2. Geliştirilen sınıflandırma modelinde kullanılan veriler

Kaynak	Veri Adedi
GEDLT	60306
HUMIR	65000
Mağaza Yorumları	8489
Tweet-pn	11055
Ürün Yorumlar	234180
Wikipedia	170413

Veri seti, çeşitli kaynaklardan toplanan Türkçe metinlerin etiketlenmiş halini içermekte olup, duygu analizine yönelik pozitif, negatif ve nötr kategorilerde sınıflandırma yapmayı amaçlayan modellerin eğitimi için ideal bir yapıya sahiptir. Bu veri setindeki etiketli metinler ve KDTV projesinden alınan ve sonrada pozitif, negatif ve nötr olarak etiketlenen veri seti birlikte kullanılarak bir duygu analizi modeli oluşturulmuştur. Bu model, Türkçe dilindeki metinleri başarılı bir şekilde sınıflandıracak şekilde optimize edilmiştir.

Veri seti, modelin geliştirilme sürecinde önemli katkılar sağlamıştır. Özellikle, sosyal medya gönderileri, kullanıcı yorumları ve haber içerikleri gibi gerçek dünya örneklerini içermesi, modelin geniş kapsamlı ve genelleştirilebilir bir performans göstermesine olanak tanımaktadır. Bu bağlamda, duygu analizi modelimizin eğitimi sırasında, veri setindeki etiketli metinler yardımıyla modelin doğruluğu ve güvenilirliği arttırmaktadır. Modelin test aşamasında da veri seti, performans değerlendirme metrikleri açısından kullanılarak, sonuçların doğrulanması sağlanmaktadır.

Kullanılan veri seti, duygu analizi amacıyla geliştirdiğimiz modelin temel eğitim verilerini sağlayarak, Türkçe metinlerde yüksek doğruluk oranlarıyla pozitif, negatif ve nötr

sınıflandırma yapabilen bir sistemin oluşturulmasına olanak tanımaktadır. Bu kapsamlı veri seti, projemize büyük katkılar sağlamakta ve Türkçe duygu analizi alanında güvenilir bir model geliştirilmesine yardımcı olmaktadır

Veri seti hazırlığı sürecinin tamamlanmasının ardından, sınıflandırma modeli geliştirme çalışmalarına başlanmıştır. Metin analitiği teknikleri kullanılarak haber içeriklerinin analiz edilmesi, günümüzde firmaların finansal durumu ve işleyişi hakkında önemli çıkarımlar yapmayı mümkün kılan bir yaklaşımdır. Bu analizlerde en kritik aşama, haber içeriklerinin pozitif, negatif ve nötr olmak üzere sınıflandırılmasıdır. Sınıflandırma, firmalarla ilgili haberlerin olumlu ya da olumsuz bir eğilim sergileyip sergilemediğini belirlemeye yönelik ilk adımı oluşturmaktadır. Haberlerin pozitif, negatif ya da nötr olarak kategorize edilmesi, firmaların finansal durumları hakkında öngörü sağlamanın yanı sıra, risk analizi ve geleceğe yönelik stratejik kararlar alınması açısından da önemli bir veri kaynağı olmaktadır.

Sınıflandırma modeli geliştirme sürecinde, Karar Ağacı, Rastgele Orman, Destek Vektör Makineleri (SVM), Lojistik Regresyon ve Naive Bayes gibi klasik makine öğrenmesi algoritmalarının yanı sıra, doğal dil işleme çalışmalarında başarılı sonuçlar veren derin öğrenme algoritmalarından UKSB kullanılmaktadır. Modelleme kapsamında geliştirilen tüm modellerin performansları karşılaştırılmıştır. Elde edilen sonuçlar incelendiğinde, en yüksek başarı oranı UKSB modeli ile elde edilmektedir. Bu durum, özellikle dizisel veri ve zaman serisi analizlerinde güçlü bir performans sergileyen UKSB'in, haber metinlerinin sınıflandırılmasında da etkili olduğunu göstermektedir.

Modelleme aşamasında, metin verilerinin sayısallaştırılması için Terim Frekansı-Ters Doküman Frekansı (TF-TDF) dönüşümü uygulanmıştır. TF- TDF yöntemi, her bir kelimenin ilgili dokümandaki önemini belirlemek için kullanılan bir tekniktir. Bu süreçte, TF- TDF dönüşümüne ait n-gram parametreleri değiştirilmiş ve bu parametrelerin farklı değerleri kullanılarak modelin performansı optimize edilmeye çalışılmıştır. Her parametre değişiminden sonra model yeniden eğitilmiş ve sonuçlar değerlendirilmektedir. En yüksek başarıya ulaşmak amacıyla, n-gram aralıkları, minimum ve maksimum kelime frekansları gibi sayısal dönüşüm parametrelerinde farklı denemeler yapılmıştır. Bu deneysel süreçte, performans ölçütü olarak doğruluk, kesinlik, geri çağırma ve F1 skoru gibi metrikler kullanılarak modellerin etkinliği değerlendirilmiştir.

Sınıflandırma modelinin geliştirilmesinin ardından, haber metinlerinin bankacılık perspektifinden analiz edilmesi çalışmalarına başlanmıştır. Bu aşamada, ilk olarak banka

bünyesinde risk değerlendirme ekipleriyle kapsamlı görüşmeler gerçekleştirilmiştir. Bu görüşmelerde, bankacılık bakış açısıyla olumlu ve olumsuz anlamlar taşıyan kelimelerin tespit edilmesine yönelik bir liste oluşturulması hedeflenmiştir. Aynı zamanda, literatürde bankacılık sektörü açısından olumlu ve olumsuz kabul edilen kelimeler üzerine yapılan çalışmalar detaylı olarak incelenmiştir. Bu literatür araştırmaları sonucunda belirlenen kelimeler, iş uzmanlarına sunulmuş ve her bir kelimenin pozitif veya negatif anlamda taşıdığı önem, 0 ile 10 arasında bir skor ile değerlendirilmiştir. Böylece, uzman görüşleri doğrultusunda, bankacılık açısından anlamlı kabul edilen olumlu ve olumsuz kelimeler ile bu kelimelere atfedilen skorların bulunduğu bir kelime listesi oluşturulmuştur (EK-1 ve EK-2).

Oluşturulan bu kelime listesi, metin içeriklerinin bankacılık bakış açısından analiz edilmesi ve içerik skorlarının hesaplanmasında kullanılmaktadır. Bu skorlama sistemi, sınıflandırma modelinin sonuçlarına dayalı olarak işlemektedir. Örneğin, sınıflandırma modeli tarafından bir haber metni pozitif olarak etiketlendiğinde, bu metinde bankacılık bağlamında olumlu kabul edilen kelimeler listesinde yer alan kelimelerin bulunması hedeflenmektedir. Metin içinde tespit edilen bu kelimelerin her biri için, önceden belirlenmiş skorlar ve bu skorların ağırlıklandırılmasıyla metnin genel içerik skoru hesaplanmaktadır. Bu yaklaşım, bankaların, haberlerde yer alan firmaların durumu hakkında daha derinlemesine içgörüler elde etmelerini sağlayacak önemli bir analiz süreci sunmaktadır. Aynı zamanda, bu skorlama yöntemi, firmalarla ilgili risk analizleri yapılmasına ve stratejik kararlar alınmasına katkı sağlayacak önemli bir veri kaynağı oluşturmaktadır.

Haber kaynaklarından elde edilen haber metinlerinin otomatik olarak toplanmasının ardından, bu metinlerin uygun bir şekilde depolanması ve analiz edilebilmesi için bir veri tabanı sistemine kaydedilmesi gerekmektedir. Bu süreçte, proje kapsamında, veri depolama ve yönetim ihtiyaçlarına en uygun çözümü sunan bir veri tabanı sunucusu kurulmuştur. Bu sunucu, doküman tabanlı depolama yapısıyla esneklik sağlayan ve büyük hacimli verilerin hızlı bir şekilde işlenmesine olanak tanıyan MongoDB sistemi üzerine inşa edilmiştir.

MongoDB'nin tercih edilmesinin başlıca sebepleri, esnek veri yapılarıyla çalışabilmesi ve yüksek performans sunmasıdır. MongoDB, koleksiyon temelli bir yapı kullanmakta olup, veri depolama ve erişim süreçlerinde geniş bir esneklik ve hız sağlamaktadır. Bu çalışma kapsamında, MongoDB veri tabanında haber kaynaklarından toplanan verilerin çeşitli kategorilerde organize edilmesi amacıyla bir dizi koleksiyon oluşturulmuştur. Bu koleksiyonlar arasında, haber adreslerinin saklandığı “haber adresleri koleksiyonu”, haber

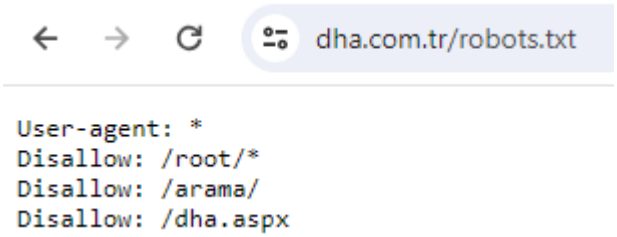
içeriklerinin kaydedildiği “haber metni koleksiyonu” ve yapılan analizlerin sonuçlarının yer aldığı “analiz sonuçları koleksiyonu” bulunmaktadır.

Veri toplama sürecinde kullanılan internet gezgini servisi, haber kaynaklarından otomatik olarak haber adreslerini tespit etmekte ve bu adreslerdeki haber metinlerini çekmektedir. Elde edilen bu veriler, sistematik bir şekilde ilgili veri tabanındaki koleksiyonlara kaydedilmektedir. Haber adresleri koleksiyonunda, süpürme işlemi sırasında tespit edilen her bir haberin adresi saklanmakta; haber metni koleksiyonunda ise bu adreslerden elde edilen içerikler depolanmaktadır. Ayrıca, metin analitiği ve doğal dil işleme teknikleri kullanılarak yapılan analizler sonucunda elde edilen bulgular, analiz sonuçları koleksiyonunda yer almaktadır. Bu yapı sayesinde, haber verileri üzerinde gerçekleştirilen tüm işlemler, düzenli ve erişilebilir bir formatta depolanmakta ve daha sonra yapılacak analizler için kolay erişim sağlanmaktadır.

Kullanılan veri tabanı mimarisi, haber metinlerinin hızlı ve verimli bir şekilde toplanması, depolanması ve analiz edilmesine olanak tanıyan esnek bir çözüm sunmaktadır. MongoDB’nin koleksiyon temelli yapısı, çalışmada kullanılan veri kaynaklarının ve analiz sonuçlarının etkin bir şekilde yönetilmesine katkı sağlamakta ve haberlerin hızlıca işlenmesini mümkün kılmaktadır.

Gerekli alt yapının oluşturulmasından sonra çalışmanın ilk aşamasını oluşturan süpürme modülünün geliştirilmesine başlanmıştır. Metin süpürme modülü, haber kaynaklarından metin içeriklerini çekerek veri toplama işlemini gerçekleştirmektedir. Bu süreç, internetten veri toplama yöntemleri arasında en yaygın olanlardan biridir. Ancak, bir internet sitesinden veri toplarken göz önünde bulundurulması gereken çeşitli etik ve teknik faktörler bulunmaktadır.

İlk olarak, süpürme işleminin gerçekleştirilmesi planlanan internet sitelerinin kullanım politikaları dikkatle incelenmeli ve bu sitelerin internet kazıma işlemlerine izin verip vermediği kontrol edilmelidir. İnternet sitelerinin “robots.txt” dosyaları veya kullanım şartları, hangi içeriklerin süpürülmesine izin verildiği konusunda rehberlik sağlayabilir. Bunun yanı sıra, süpürme işlemi sırasında internet sitelerinin performansını olumsuz etkilememek için sitenin teknik altyapısına ve bant genişliği sınırlarına saygı göstermek gerekmektedir. Yüksek frekansta veri çekme işlemleri, sitenin işleyişine zarar verebilir ya da site tarafından engellenmeye neden olabilir. Bu nedenle, geliştirilen süpürme modülünde her internet sitesi için bu tür kısıtlamalar göz önünde bulundurularak yapılandırılmış bir tasarım tercih edilmiştir.



Şekil 12. Demirören Haber Ajansı (DHA) robots.txt dosyası içeriği

Örnek olarak, Şekil 12'de Demirören Haber Ajansı'na (DHA) ait "robots.txt" dosyasında bazı kısıtlamalar ifade edilmektedir. Bu kısıtlamalar dikkate alınarak veri süpürme işlemi yapılması gerekmektedir. DHA'ya ait "robots.txt" dosyasındaki başlıca direktiflerin açıklamaları aşağıda ifade edilmektedir.

"User-agent: *": Bu satır, tüm arama motoru robotlarının (crawler'larının) belirtilen kurallara uyması gerektiğini belirtir. "*", herhangi bir arama motoru veya veri süpürme robotunu ifade eder.

"Disallow: /root/": Bu direktif, root dizini altındaki tüm içeriklerin arama motorları ve veri süpürme robotları tarafından taranmasını engeller. Örneğin, www.dha.com/root/ veya www.dha.com/root/alt-dizin/ gibi adresler taranamaz.

"Disallow: /arama/": Bu direktif, arama dizini altındaki tüm içeriklerin taranmasını engeller. Örneğin, www.dha.com/arama/ veya www.dha.com/arama/sonuclar/ gibi adresler süpürülemez.

"Disallow: /dha.aspx": Bu direktif, "dha.aspx" dosyasının taranmasını engeller. Örneğin, www.dha.com/dha.aspx adresi taramaya kapalıdır.

Özetle, DHA'nın "robots.txt" dosyası, tüm arama motoru robotlarına ve veri süpürme yazılımlarına, /root/ dizini altındaki tüm içerikleri, /arama/ dizini altındaki tüm içerikleri ve "dha.aspx" dosyasını taramalarına izin vermediğini belirtmektedir. Bu kısıtlamalar dikkate alınarak, ilgili internet sitesinin bu bölümlerinin arama motorları tarafından dizine eklenmesini ve arama sonuçlarında görünmesini engellemeyi amaçlamaktadır. İfade edilen kısıtlamalar haricindeki adreslerde süpürme işlemi yapılabilmektedir.

Süpürme modülünün geliştirilmesi aşamasında Python programlama dili kullanılmış ve veri tabanı işlemleri için PyMongo kütüphanesi tercih edilmiştir. PyMongo, MongoDB veri tabanına erişim sağlayarak verilerin daha esnek bir şekilde koleksiyonlar halinde saklanmasını mümkün kılmaktadır. İnternet sitelerinden içerik çekmek için ise, yaygın olarak kullanılan "BeautifulSoup" kütüphanesi kullanılmıştır. BeautifulSoup, haber

ajanslarına ait internet sitelerindeki haber bağlantılarının tespit edilmesini ve metin içeriklerinin ayrıştırılmasını sağlayan etkili bir araçtır.

Haber kaynaklarından toplanan verilerin doğru bir şekilde işlenebilmesi için çeşitli veri ön işleme ve temizleme işlemleri gerçekleştirilmektedir. Elde edilen internet adreslerinden haber başlıkları, haber metinleri ve haberlerin yayınlanma tarihleri başarılı bir şekilde çıkarılmaktadır. Bu bilgiler, uygun formatlarda veri tabanına kaydedilerek saklanmıştır. Ayrıca, bu süreç boyunca veri kalitesinin yüksek tutulması amacıyla haber metinlerinden gereksiz veya bozuk içerikler temizlenmiş, sadece analiz için gerekli olan bölümler veri tabanına aktarılmıştır.

Çalışma kapsamında geliştirilen süpürme modülü, belirlenen haber kaynaklarından veri çekme ve bu verilerin depolanması süreçlerini başarılı bir şekilde tamamlamıştır. Bu süreçte, modül hem teknik hem de etik gereksinimlere uygun olarak yapılandırılmış, haber ajanslarından toplanan haber içerikleri başarıyla veri tabanına entegre edilmiştir. Süpürme modülünün geliştirilmesi, ihtiyaç duyulan haber verilerinin sistematik ve güvenilir bir şekilde toplanmasını sağlamış ve veri analiz süreçlerine temel teşkil etmektedir.

Bu aşamada haber metinlerinin sınıflandırma modeli içeri süpürme modülünün başarı bir şekilde çalışmaktadır. Bu aşamadan sonra, metin içeriklerinin hangi firma ya da firma yöneticisi ile ilgili olduğunu belirlemek amacıyla metin içerisindeki varlık ifadelerinin tespit edilmesi gerekmektedir. Bu işlem, doğal dil işleme tekniklerinden biri olan Adlandırılmış Varlık Tanıma (AVT) tekniğiyle yapılmaktadır. Çalışmanın bu aşaması, özellikle “Organizasyon Adı” (şirket veya kurum) ve “Kişi Adı” gibi varlıkların doğru şekilde tanımlanmasına odaklanmaktadır. Haber içeriklerinin firma adı veya firma yöneticileri temelinde analiz edilmesi, çalışma için kritik bir öneme sahip olup, yapılan analizlerin doğruluğu ve güvenilirliği açısından temel aşamalardan birini oluşturmaktadır.

Varlık tanıma süreci için, mevcut literatürde yaygın olarak kullanılan ve büyük veri setleriyle eğitilmiş modeller test edilmiş ve bu modellerin performansları değerlendirilmiştir. Ancak, bu çalışmada Türkçe dilindeki haber metinleri üzerinde çalışıldığı için, dilin morfolojik yapısına ve Türkçe'ye özgü özelliklere uyum sağlayacak bir AVT modeli geliştirilmesi veya kullanılması gerekmektedir. Türkçe doğal dil işleme görevlerinde, dilin karmaşık yapısı, eklemeli morfolojisi ve sözcüklerin anlamlarının bağlama göre değişebilmesi gibi faktörler nedeniyle, mevcut uluslararası dil modelleri her zaman istenen başarıyı gösterememektedir.

Bu kapsamda önceden eğitilmiş Türkçe haber metinlerinde varlık tanıma işlemleri için "Turkish NLP Suite Python" kütüphanesi tercih edilmiştir (Altinok, 2023). Bu kütüphane, doğal dil işleme görevlerinde SpaCy dil kütüphanesi baz alınarak geliştirilmiş bir model içermektedir. SpaCy, Python ve Cython tabanlı, ileri düzey doğal dil işleme analizleri yapabilen ve yüksek performans sunan açık kaynaklı bir yazılım kütüphanesidir. SpaCy'nin bu çalışmada tercih edilmesinin temel sebeplerinden biri, büyük veri setleriyle eğitilmiş olması ve hızlı ve verimli doğal dil işleme yeteneklerine sahip olmasıdır.

Çalışmanın bu aşamasında, Turkish NLP Suite tarafından geliştirilen "tr_core_news_trf" adlı önceden eğitilmiş model kullanılmıştır. Bu model, Türkçe metinler üzerinde çeşitli doğal dil işleme görevlerini gerçekleştirmek üzere eğitilmiş bir modeldir. "tr_core_news_trf" modeli, Türkçe metinler üzerinde yüksek doğrulukta doğal dil işleme yetenekleri sunan "Transformer" tabanlı bir modeldir. SpaCy kütüphanesi üzerine inşa edilen bu model, özellikle Türkçe dilinin morfolojik yapısına uygun şekilde tasarlanmış ve büyük veri setleri üzerinde eğitilmiştir.

Bu model, projenin ihtiyaçları doğrultusunda kullanılarak, haber metinlerindeki organizasyon ve kişi adlarının başarılı bir şekilde tespit edilmesini sağlamaktadır. "tr_core_news_trf", SpaCy ile entegre çalışarak varlık tanıma işlemlerinin doğruluğunu artırmış ve haber içeriklerinin firma ve yönetici bazlı analizlerinin daha etkin bir şekilde yapılmasına katkı sunmuştur.

Tablo 3. Haber içeriğinde model tarafından tespit edilen varlıklar

ORG NAME	BEST SCORE ORG NAME	DATE	GPE	ORG	ORG SCORES	QUANTITY	LABEL
Akfen', 'Akfen Yenilenebilir', 'Akfen Yenilenebilir Enerji'	'Akfen'	Nisan 2023'te, '2023 yılının ikinci yarısında'	Hasanoba', 'Denizli', 'Kocalar', 'Denizli', 'Çanakkale', 'Çanakkale'	TEİAŞ', 'Bakanlık', 'Akfen Yenilenebilir', 'EPDK', 'Akfen Yenilenebilir Enerji', 'Akfen'	Akfen': 26, 'Akfen Yenilenebilir': 21, 'Akfen Yenilenebilir Enerji': 21, 'Bakanlık': 5, 'EPDK': 10, 'TEİAŞ': 10	95,42 MWe', '40,02 MWe', '12,68 Mwe', '6,36 MWe', '4,95 MWe',	'Pozitif'

Metin içerisindeki varlık tespit işlemlerinin tamamlanmasının ardından, aynı metin içerisinde birden fazla kişi ya da organizasyon adının geçtiği durumlarla karşılaşılmaktadır. Bu tür durumlarda, haberin hangi firma ya da organizasyonla ilgili olduğunun doğru bir şekilde belirlenmesi gerekmektedir. Örneğin, "Şikayetvar'dan Yurtiçi Kargo'ya başarı ödülü" başlıklı bir haberde, hangi firmanın ana konu olduğunun belirlenmesi ve buna bağlı olarak firma istihbarat skorlamasının yapılması gerekmektedir. Bu tespit işlemi, haber metinlerinin yazım süreçlerinin teknik olarak incelenmesini gerektirmektedir.

Bu amaçla, haber içeriklerinde kullanılan dil yapıları ve anlatım biçimleri analiz edilerek, metnin hangi organizasyon veya kişiye odaklandığı tespit edilmeye çalışılmıştır. Bu inceleme sürecinde, genellikle haber metinlerinde odak firma veya kişi isimlerinin başlıkta ya da metnin başında yer aldığı, ödüller, başarılar veya eleştiriler gibi ana konuların bu varlıklarla ilişkilendirildiği gözlemlenmektedir. Bu bulgular doğrultusunda, metin içerisindeki en önemli varlığın belirlenmesine yönelik bir strateji geliştirilmiştir.

Haber yazımı, hem hızlı ham de etkili bir şekilde bilginin sunulmasını sağlayan belirli yapılar üzerine inşa edilmiştir. Her haber metninde "manşet", "özet metin", "haber girişi" ve detaylar bölümü bulunmaktadır. Bu bileşenlerin okuyucunun dikkatini çekecek şekilde sunulması haber yazımı açısından kritik öneme sahiptir. Bu bileşenlerin her biri, okuyucuya haberin neyle ilgili olduğunu hızla aktarırken aynı zamanda okuyucunun ilgisini devam ettirmek için stratejik bir dil ve yapı kullanılmaktadır. Manşet, haberin genel içeriğini özetleyen kısa, çarpıcı ve haberin ana temasını en net şekilde sunan bölümdür. Özet metin, haber detaylarına girmeden okuyucuya haber temel unsurlarının sunulduğu bölümdür. Bu bölümde, haberde yanıtlanacak temel soruların (ne, kim, ne zaman, nerede, nasıl ve neden) bazılarına kısa cevaplar verilerek okuyucunun habere dair daha fazla bilgi edinmesi sağlanmaktadır. Giriş bölümü, haberin giriş kısmını ifade etmektedir ve haberde yer alan en önemli bilgilerin özetlenmesi amacını taşır ve haberin akışını başlatır. Haber içeriği, haberle ilgili ayrıntılı bilgilerin ifade edildiği bölümdür (Durna, 2020).

Haber metnin yazımındaki detaylar dikkate alınarak metin içerisinde birden çok organizasyon ve kişi ifadesi tespit edilmesi durumunda bir ağırlıklandırma metodolojisi uygulanmaktadır. Bu metodoloji, metinlerdeki organizasyon isimlerinin geçiş sıklığını ve konumsal önemini temel alarak bir puanlama sistemi uygulamaktadır. Haber metinleri, oluşturulma şekillerine göre analitik bir yaklaşımla giriş, gelişme ve sonuç olmak üzere üç bölüme ayrılmaktadır.

Her bölümdeki organizasyon isimlerinin geçiş sıklığı, bölümün içeriğindeki konumsal ağırlığına göre farklı puanlama katsayıları ile çarpılarak değerlendirilir. Bu aşamada ilgilenilen metin n adet bölüme ayrılır ve her bölümde algılanana varlıkların frekansları hesaplanmaktadır. Nihai olarak elde edilecek olan S skoru Denklem 11 ile hesaplanmaktadır.

$$S = \sum_{i=1}^{n} w_i f_i \tag{11}$$

Burada n toplam bölüm sayısını, w_i i. bölüm için frekans ağırlıklarını ve f_i i. bölümdeki varlık frekansını göstermektedir. Toplanan bu puanlar değerlendirildikten sonra en yüksek puana sahip organizasyon, haberin odaklandığı ana organizasyon olarak belirlenmektedir. Bu ağırlıklandırma metodolojisi, metnin bölümleri arasında bilginin nasıl dağıldığını ve organizasyonların haber metni içerisindeki önemini objektif tespit etmeyi sağlamaktadır. Çalışmada yaptığımız denemeler sonucunda optimum başarı n=3, ve w_0=10, w_1=5 ve w_2=1 değerleri ile elde edilmiştir.

Tablo 4. Bölümleme temelli kurum skorlama deney parametreleri

Bölüm Sayısı	Metin Bölümü	Katsayı (w)
3	Metin bölümü 1	10
	Metin bölümü 2	5
	Metin bölümü 3	1
5	Metin bölümü 1	10
	Metin bölümü 2	7
	Metin bölümü 3	5
	Metin bölümü 4	3
	Metin bölümü 5	1
7	Metin bölümü 1	10
	Metin bölümü 2	8
	Metin bölümü 3	6
	Metin bölümü 4	4
	Metin bölümü 5	3
	Metin bölümü 6	2
	Metin bölümü 7	1

- 3 bölüme ayrıldığında: Birinci bölümde kurum adı frekansı çarpı 10, ikinci bölümdeki kurum adı frekansı çarpı 5, üçüncü bölümdeki kurum frekansı çarpı 1 şeklinde katsayılar denenmiştir.
- 5 bölüme ayrıldığında: Birinci bölümdeki kurum adı frekansı çarpı 10, ikinci bölümdeki kurum adı frekansı çarpı 7, üçüncü bölümdeki kurum adı frekansı çarpı 5, dördüncü bölümdeki kurum adı frekansı çarpı 3, beşinci bölümdeki kurum adı frekansı çarpı 1 şeklinde katsayılar denenmiştir.
- 7 bölüme ayrıldığında: Birinci bölümdeki kurum adı frekansı çarpı 10, ikinci bölümdeki kurum adı frekansı çarpı 8, üçüncü bölümdeki kurum adı frekansı çarpı 6, dördüncü bölümdeki kurum adı frekansı çarpı 4, beşinci bölümdeki kurum adı frekansı çarpı 3, altıncı bölümdeki kurum adı frekansı 2, yedinci bölümdeki kurum adı frekansı çarpı şeklinde katsayılar denenmiştir.

Tablo 5. Bölümleme temelli kurum skorlama deney sonuçları

Bölüm Sayısı	Başarılı Metin Sayısı
3	81
5	9
7	10

Bu üç farklı yaklaşımla yapılan deneylerde toplamda 100 metin kullanıldı. Deney sonuçlarına göre, 100 metinden 81'inde 3 bölüme ayırma yaklaşımı en yüksek kurum skorunu elde ederek en başarılı yöntem olarak belirlendi. 5 bölüme ayırma yaklaşımı, yalnızca 9 metinde en yüksek kurumsal skoru sağladı. 7 bölüme ayırma yaklaşımında ise sadece 10 metinde yüksek kurum skoru elde edildi. Sonuç olarak, optimum başarıyı 3 bölüme ayırma işleminde, katsayıların w_0=10, w_1=5 ve w_2=1 olarak belirlendiği durumda sağlandığı görülmüştür.

Haber içeriği analiz işlemlerinin son aşaması haber içeriklerinin bankacılık perspektifiyle incelenmesi sürecidir. Bu aşamada daha önce iş uzmanları tarafından önerilen ve yapılan araştırmalar sonucunda tespit edilen olumlu ve olumsuz kelimeler listesi kullanılmaktadır. Oluşturulan bu listedeki olumlu ve olumsuz kelimeler risk ekiplerindeki uzman kişiler tarafından incelenip onaylanan ve 0-10 arasında skorlanan kelimelerdir.

Metin analizi sürecinde, belirlenen ve onaylanan kelimelerin metin içerisindeki dağılımı dikkatlice incelenmiş, her bir kelimenin ağırlık skoru, kelimenin metindeki geçiş

frekansı ile çarpılarak toplam puan hesaplamasına katkıda bulunmuştur. Bu toplam puanlar, ilgili firmanın genel bilgi skorunu oluşturmak için kullanılmaktadır. Bu skor, firmanın piyasadaki pozisyonunu ve müşteri algısını objektif bir biçimde değerlendirmek için kritik bir ölçüttür olmaktadır. Sonuç olarak, bu ağırlıklandırma ve puanlama sistematiği, metin bazlı verilerden elde edilen bilgilerin stratejik karar verme süreçlerinde etkin bir şekilde kullanılmasını sağlamaktadır. Böylece firmaların rekabetçi avantajlarını artırmalarına ve piyasa dinamiklerine hızla adapte olmalarına olanak tanımaktadır. Skor hesaplaması için Denklem 12, Denklem 13 ve Denklem 14'deki formül kullanılmıştır.

$$S_{pks} = \sum_{k \in K_{pos}} w_k \tag{12}$$

$$S_{nks} = \sum_{k \in K_{neg}} w_k \tag{13}$$

$$S_{all} = \begin{cases} S_{pks}, & P_{pozitif} > P_{negatif} \\ S_{nks}, & P_{negatif} > P_{pozitif} \end{cases} \tag{14}$$

Burada S_{pks} ve S_{nks} sırasıyla pozitif ve negatif haber skorlarını, K_{pos} ve K_{neg} sırasıyla pozitif ve negatif kelime kümelerini, w_k kelimenin uzmanlar ve ChatGPT aracılığı ile belirlenmiş ortalama ağırlığını, S_{all} haberin genel skorunu, $P_{pozitif}$ ve $P_{negatif}$ sırasıyla haberin pozitif ve negatif olma olasılıklarını göstermektedir.

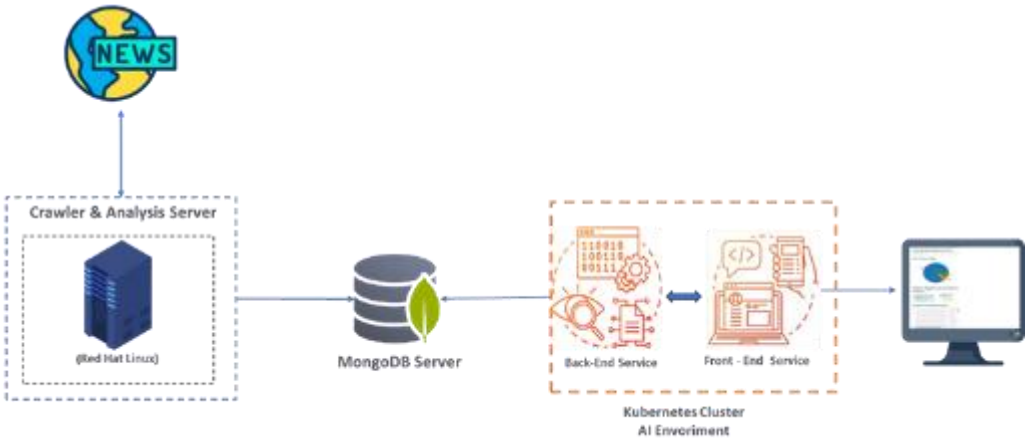
Duygusallık Skoru haber metninin duygu seviyesini belirlemek için kullanılmaktadır.

Analiz aşamasının tamamlanmasının ardından, sistemin kullanıma sunulabilmesi için ön yüz ve arka yüz servislerinin geliştirilmesi de başarılı bir şekilde gerçekleştirilmiştir. Python programlama dilinin "Streamlit" kütüphanesi kullanılarak önyüz servisinin geliştirme süreci tamamlanmıştır. Streamlit, interaktif internet arayüzleri oluşturma sürecinde sunduğu kolaylıklarla, veri analizi sonuçlarının kullanıcılarla etkileşimli bir biçimde paylaşılmasını yardımcı olmaktadır.

Şekil 13. Geliştirilen arayüz servisi ve firma analiz sonucu örneği

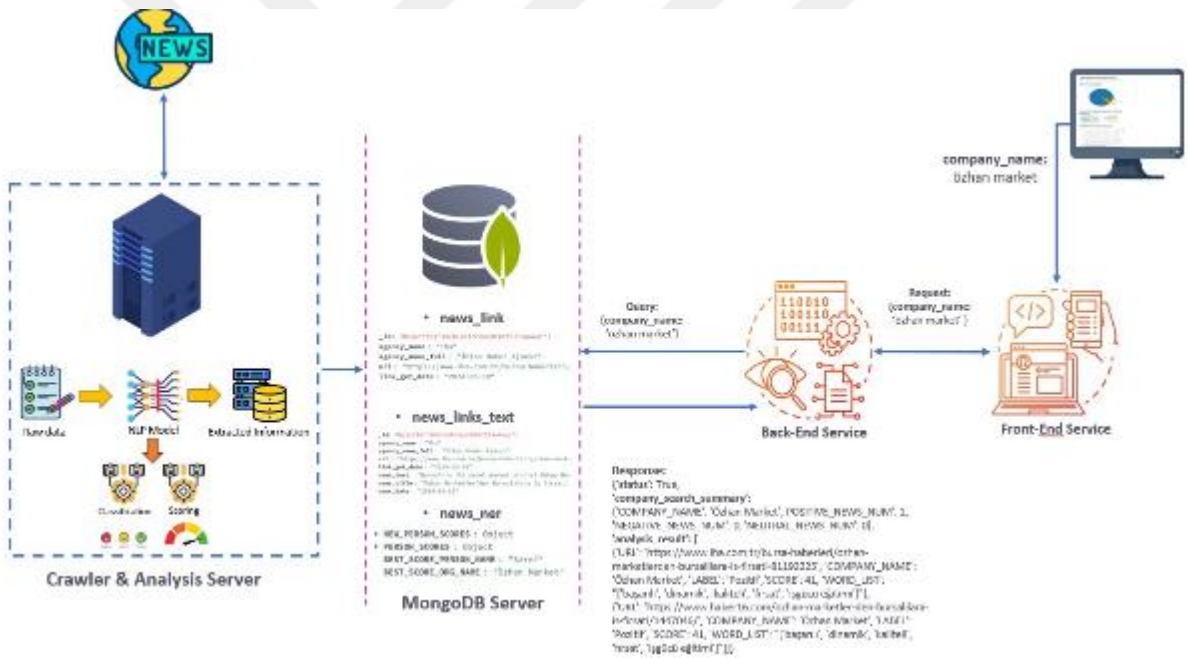
Arka plan işlemlerini yöneten ve servis tabanlı bir mimari yapıyı destekleyen altyapının oluşturulmasında ise FastAPI framework'ü kullanılmaktadır. Bu framework, kullanıcı taleplerinin verimli bir şekilde karşılanmasına olanak tanıyarak sistemin genel performansını ve esnekliğini artırmıştır. Geliştirilmesi tamamlanan iki servis yapay zeka uygulamalarının çalıştığı Kubernetes Cluster üzerinde konteynır mantığında ayağa kaldırılarak hizmete sunulmuştur.

Bankacılık Bakış Açısıyla Firma İstihbarat Sisteminin aktif olarak kullanılan sistem mimarisi Şekil 14'de ifade edilmektedir.



Şekil 14. Çalışan sistem mimarisi

Geliştirilen sistemin analiz sisteminin çalışma yapısı Şekil 15'de yer almaktadır.



Şekil 15. Geliştirilen sistemin çalışması yapısı

Şekil 15 incelendiğinde sistemin çalışma şekli şu şekilde olmaktadır. İlk olarak kullanıcı geliştirilen arayüz ekranına girerek sorgulamak istediği firma ismini arama alanı yazıp arama butona tıklamaktadır. Sonrasında arayüz ekranı, arka yüz servisiyle iletişime geçerek kullanıcının arama istediği firma bilgisini veri tabanına iletmektedir. Veri tabanı ilgili koleksiyonda arama işlemi yaparak firmayla ilgili bulduğu analiz sonuçları önyüz servisine ileterek kullanıcıyla paylaşmaktadır. Sistemin en kritik bileşeni olan internet gezgini ve analiz sunucusu şu an için günlük veri toplama ve analiz yapacak şekilde

ayarlanmıştır. Bu sunucu üzerindeki internet gezgini ve analiz servisi tüm analiz işlemleri yaparak sonuçları veri tabanında depolamaktadır. Geliştirilen sistem banka tarafında canlı ortamda aktif olarak kullanımdadır.

3. BULGULAR VE TARTIŞMA

Bu bölümde ilk olarak geliştirilen sistemin bir parçası olan metin sınıflandırma modeline ait bulgular paylaşılmaktadır. Sonrasında sistemin uçtan uça çalışmasıyla elde edilen analiz sonuçlarıyla ilgili bulgular üzerinde değerlendirmeler yapılmaktadır.

Model geliştirme sürecinde hem klasik makine öğrenmesi algoritmaları hem de derin metin sınıflandırma problemlerinde başarısını kanıtlamış olan UKSB algoritması kullanılmıştır. UKSB ve klasik makine öğrenmesi algoritmalarının performansları değerlendirilmiştir. Tablo 6'daki performans değerleri incelendiğiden UKSB modelinin en iyi performansa sahip olan model olduğu görülmektedir.

Tablo 6. Model performanslarının karşılaştırılması

Model	F1-Micro	F1-Macro
Lojistik Regresyon	0,83	0,77
Naive Bayes	0,72	0,62
Karar Ağaçları	0,81	0,74
Destek Vektör Makineleri	0,83	0,76
Rastgele Orman	0,83	0,78
UKSB	0,86	0,79

Bu çalışmada kullanılan makine öğrenmesi ve derin öğrenme modellerinin performansları, F1-Micro ve F1-Macro skorlarıyla değerlendirilmektedir. F1-Macro skoru her bir sınıfın F1 skorunun ortalamasını alarak sınıflar arasındaki dengesizliği dikkate alırken, F1-Micro skoru, tüm örnekler üzerinden hesaplanan doğru sınıflandırmaların ağırlıklı ortalamasını dikkate almaktadır.

Lojistik Regresyon modeli, F1-Micro skorunda %83, F1-Macro skorunda ise %77 performans göstermiştir. Bu sonuçlar, modelin genel doğruluk performansının yüksek olduğunu ancak farklı sınıflar arasında bir dengesizlik olduğunu göstermektedir.

Naive Bayes modeli ise F1-Micro'da %72, F1-Macro'da %62 skorlarıyla, diğer modellere kıyasla daha düşük performans göstermiştir. Bu durum, özellikle sınıflar arası dengesiz verilerde Naive Bayes modelinin düşük bir başarı sağladığını ortaya koymaktadır.

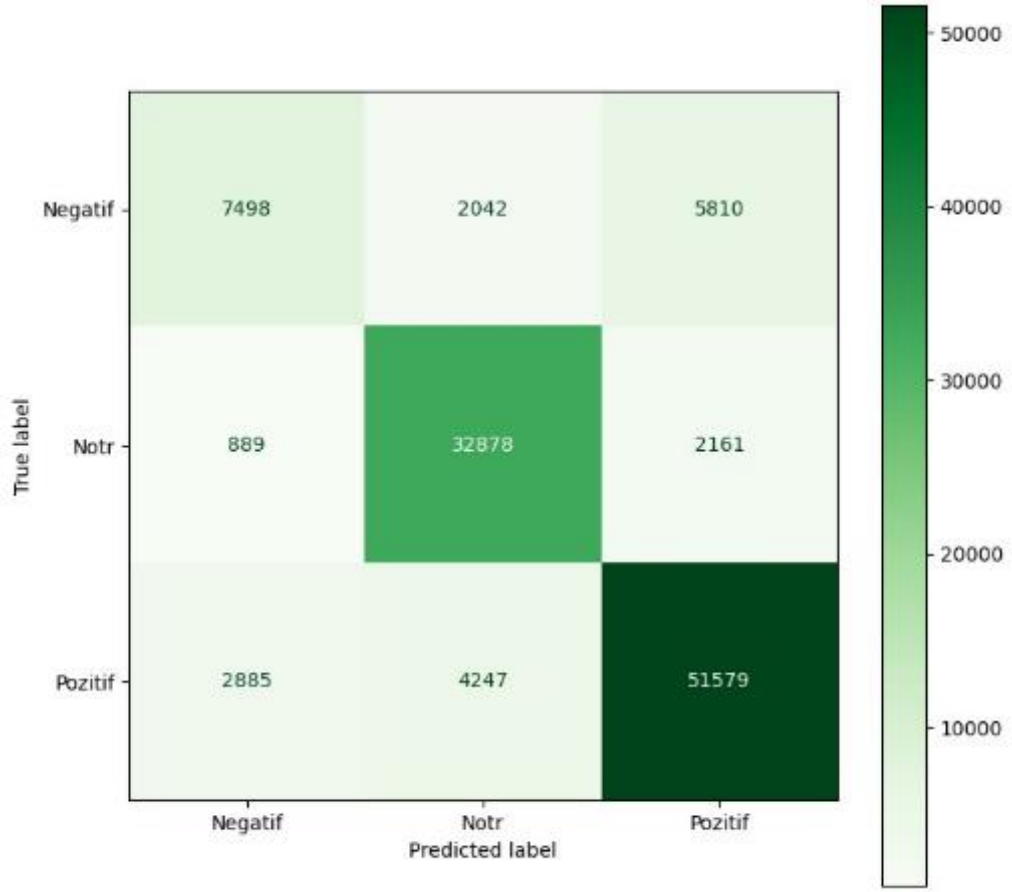
Karar Ağaçları ve Rastgele Orman modelleri benzer şekilde yüksek performans göstermiştir. Karar Ağaçları %81 F1-Micro ve %74 F1-Macro skoru elde ederken, Rastgele Orman modeli %83 F1-Micro ve %78 F1-Macro skorlarıyla daha dengeli bir performans göstermektedir. Bu bulgular, ağaç tabanlı algoritmaların veriyi başarılı şekilde sınıflandırabildiğini göstermektedir.

Destek Vektör Makineleri (DVM), %83 F1-Micro ve %76 F1-Macro skoru ile hem doğruluk hem de denge açısından güçlü bir performans göstermiştir. Bu sonuçlar, DVM'nin özellikle sınıflar arası ayrımı başarılı bir şekilde yapabildiğini göstermektedir.

Derin öğrenme yaklaşımı olan UKSB modeli ise, hem %86 F1-Micro hem de %79 F1-Macro skorlarıyla en yüksek performansı göstermiştir. Bu durum, UKSB'nin veri içerisindeki uzun süreli bağıntıları başarılı bir şekilde öğrenebildiğini ve sınıflar arası dengesizliklerde daha başarılı sonuçlar verdiğini göstermektedir.

Genel olarak, elde edilen bulgular değerlendirildiğinde, UKSB modelinin diğer makine öğrenmesi modellerine göre daha üstün bir performans sergilediği tespit edilmiştir. Özellikle karmaşık ve uzun vadeli bağıntılar içeren verilerde daha başarılı olduğunu ortaya koymaktadır. Naive Bayes modelinin düşük performansı ise, bu modelin sınıf dengesizliklerine karşı duyarlı olduğunu ve daha gelişmiş modellerin tercih edilmesi gerektiğini göstermektedir.

Çalışma kapsamında kullanılan tüm modeller değerlendirirken karmaşıklık matrisi sonuçları da değerlendirilmektedir.



Şekil 16. Lojistik Regresyon modeline ait karmaşıklık matrisi

Lojistik Regresyon modeli için elde edilen karmaşıklık matrisi Şekil 16'da yer almaktadır. Bu matris modelin sınıflandırma performansını detaylı bir şekilde göstermektedir. Karmaşıklık matrisi, her bir sınıf için doğru sınıflandırılan örneklerin yanı sıra modelin yanlış sınıflandırdığı örnekleri de gösterir. Aşağıda, verilen matris temel alınarak modelin performansı yorumlanmaktadır.

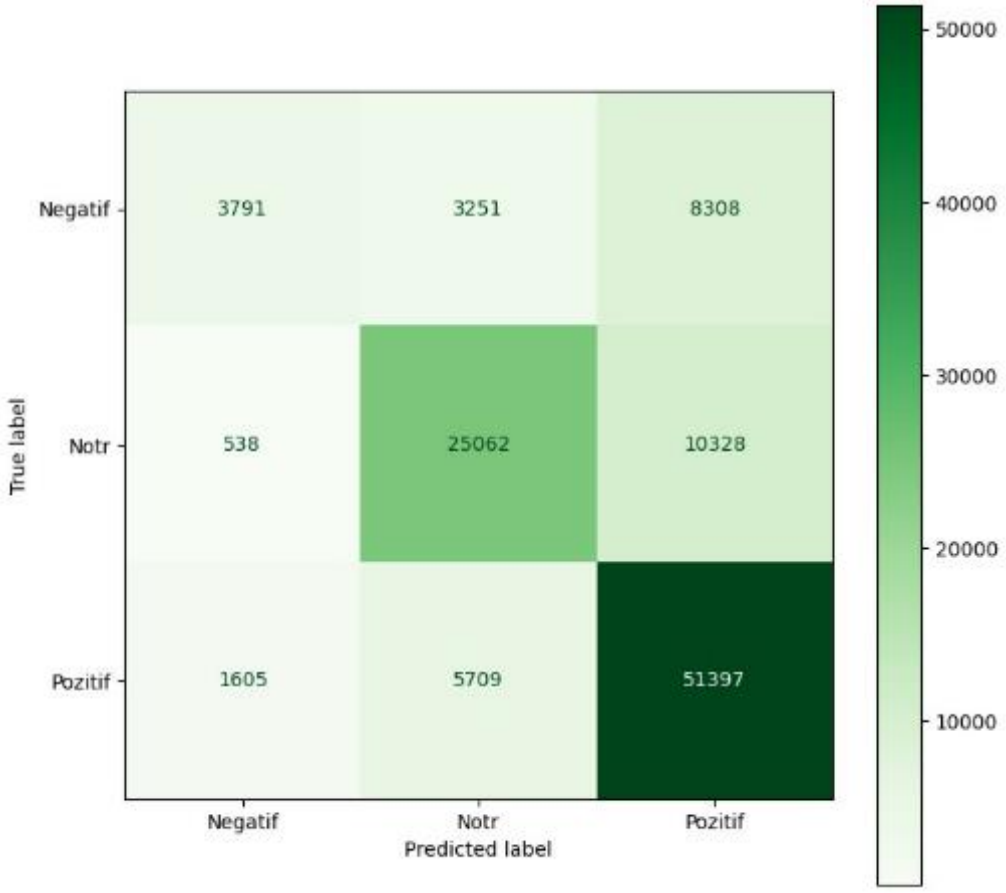
Negatif sınıf: Model, negatif sınıfa ait 7498 örneği doğru bir şekilde sınıflandırmıştır. Ancak, 2042 negatif örnek notr olarak, 5810 negatif örnek ise pozitif olarak yanlış sınıflandırılmıştır. Bu durum, modelin negatif örnekleri doğru bir şekilde sınıflandırma konusunda bazı zorluklar yaşadığını ve negatif sınıf ile diğer sınıflar arasında ayrım yapmada güçlük çektiğini göstermektedir.

Notr sınıf: Model, notr sınıfa ait 32.878 örneği doğru bir şekilde sınıflandırmıştır, ancak 889 notr örnek negatif, 2161 notr örnek ise pozitif olarak yanlış sınıflandırılmıştır. Notr sınıfında modelin genel doğruluk oranı oldukça yüksek görünmektedir, bu da modelin notr örnekleri ayırt etmede oldukça başarılı olduğunu göstermektedir.

Pozitif sınıf: Model, pozitif sınıfa ait 51.579 örneği doğru bir şekilde sınıflandırmıştır. Bununla yanı sıra, 2885 pozitif örnek negatif, 4247 pozitif örnek ise notr olarak yanlış sınıflandırılmıştır. Pozitif sınıf için genel doğruluk yüksek olmakla birlikte, notr ve negatif sınıflar ile pozitif sınıf arasında bazı yanlış sınıflandırmalar gözlemlenmektedir.

Genel olarak, modelin en iyi performansı notr sınıfında gösterdiği söylenebilir. Pozitif sınıfta da genel olarak iyi bir performans sergilemesine rağmen, negatif sınıfın diğer sınıflarla karıştırılma oranı daha yüksektir. Bu durum, modelin negatif sınıfı doğru bir şekilde ayırt etmede diğer sınıflara göre daha zorlandığını göstermektedir. Negatif örneklerin sınıflandırma performansını artırmak için modelin daha fazla optimize edilmesi veya farklı özellik çıkarım yöntemlerinin kullanılması faydalı olabilir.

Naive Bayes modeli için elde edilen karmaşıklık matrisi Şekil 17'de yer almaktadır.



Şekil 17. Naive Bayes modeline ait karmaşıklık matrisi

Naive Bayes modeli için elde edilen karmaşıklık matrisi, detaylı bir şekilde incelenmiştir.

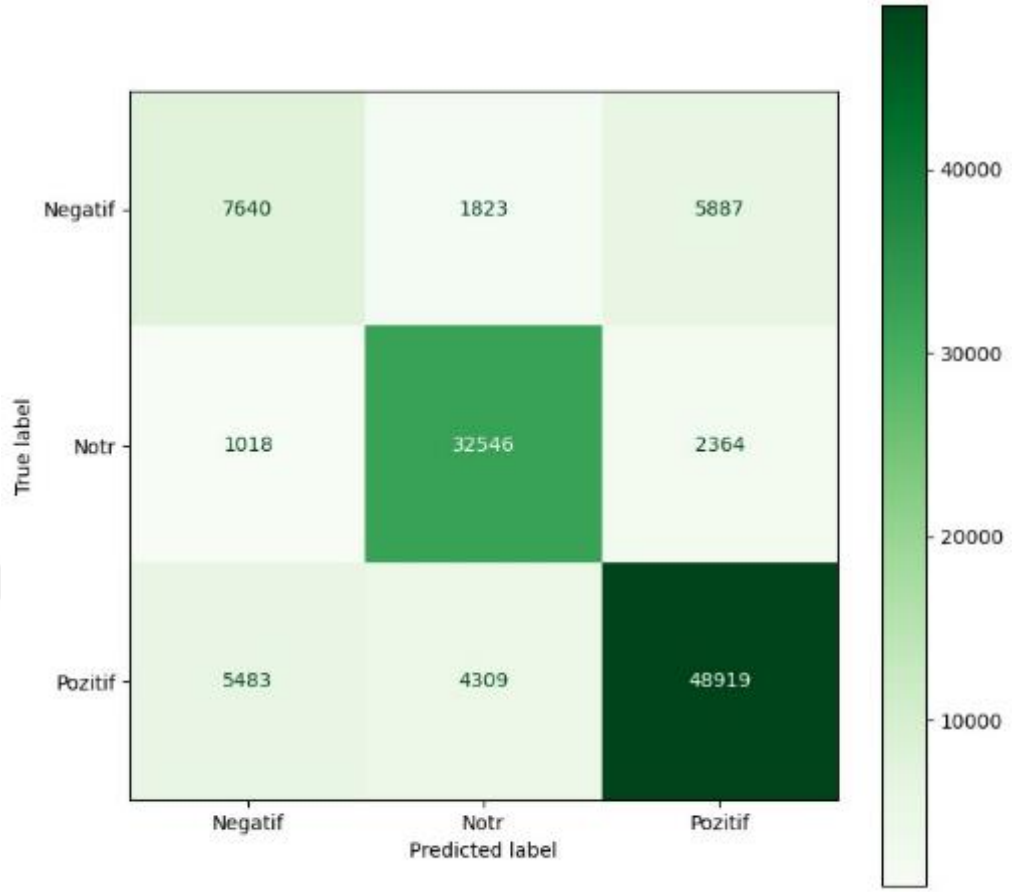
Negatif sınıf: Negatif sınıfa ait 3791 örnek doğru sınıflandırılmış, buna karşılık 3251 negatif örnek notr, 8308 negatif örnek ise pozitif olarak yanlış sınıflandırılmıştır. Bu durum, modelin negatif sınıfı doğru ayırt etmekte zorlandığını ve negatif örneklerin büyük bir kısmını pozitif ya da notr sınıfa yanlış olarak sınıflandırdığını göstermektedir.

Notr sınıf: Notr sınıfa ait 25.062 örnek doğru sınıflandırılmıştır. Buna karşılık, 538 notr örnek negatif olarak, 10.328 notr örnek ise pozitif olarak yanlış sınıflandırılmıştır. Modelin notr sınıf için genel doğruluğu yüksek olsa da, pozitif sınıf ile karışıklık yaşandığı dikkat çekmektedir. Bu da modelin notr ve pozitif sınıfları ayırt etmede bazı zorluklar yaşadığını göstermektedir.

Pozitif sınıf: Pozitif sınıfa ait 51.397 örnek doğru sınıflandırılmıştır. Buna karşılık 1605 pozitif örnek negatif, 5709 pozitif örnek ise notr olarak yanlış sınıflandırılmıştır. Modelin pozitif sınıfı oldukça iyi ayırt ettiği görülmekte, ancak pozitif örneklerin bir kısmı notr sınıfa yanlış atanmıştır.

Naive Bayes modeli genel olarak değerlendirildiğinde, notr ve pozitif sınıflarda nispeten iyi performans gösterdiği, buna karşılık negatif sınıfta ciddi sınıflandırma hatalarının olduğu söylenebilir. Özellikle negatif sınıfın diğer sınıflarla olan karışıklık oranının yüksek olduğu görülmektedir. Bu durum modelin negatif sınıfı ayırt etme yeteneğinin zayıf olduğunu göstermektedir. Naive Bayes modelinnin performansını artırmak için, negatif sınıf için daha fazla veri işleme veya özellik mühendisliği uygulanması düşünülebilir.

Karar Ağacı algoritmasına ait elde edilen karmaşıklık matrisi Şekil 18’de yer almaktadır.



Şekil 18. Karar Ağacı modeline ait karmaşıklık matrisi

Karar Ağacı algoritması için elde edilen karmaşıklık matrisi dikkate alındığında modelin performansı şu şekilde yorumlanabilir:

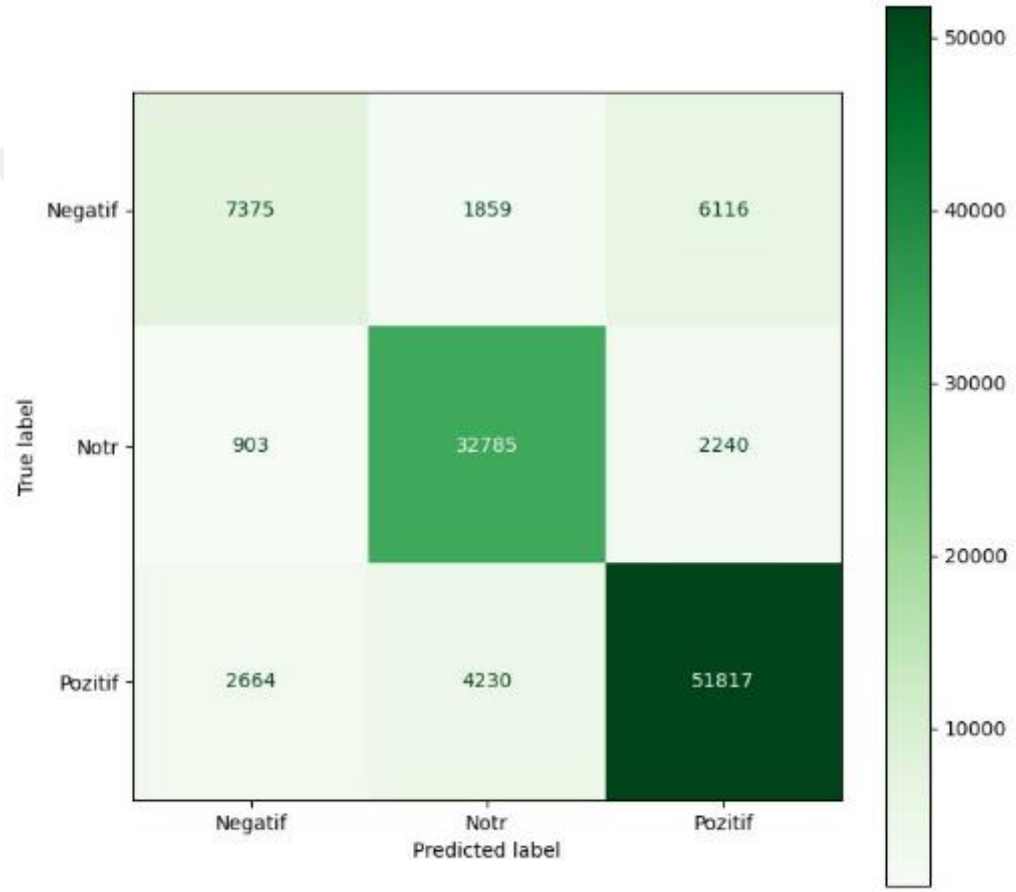
Negatif sınıf: Model, negatif sınıfa ait 7640 örneği doğru bir şekilde sınıflandırdığı görülmektedir. Buna karşılık, 1823 negatif örnek notr olarak, 5887 negatif örnek ise pozitif olarak yanlış sınıflandırılmıştır. Bu sonuçlar, modelin negatif sınıfı diğer sınıflardan ayırt etmede zorluklar yaşadığını göstermektedir. Yanlış sınıflandırılan negatif örneklerin büyük bir kısmı pozitif sınıfa atanmıştır.

Notr sınıf: KA modeli, notr sınıfa ait 32.546 örneği doğru sınıflandırmıştır. 1018 notr örnek negatif sınıfa, 2364 notr örnek ise pozitif sınıfa yanlış sınıflandırılmıştır. Notr sınıfı için modelin genel doğruluk oranı yüksek olup, bu sınıfta modelin ayırt etme yeteneğinin güçlü olduğu söylenebilir.

Pozitif sınıf: Model, pozitif sınıfa ait 48.919 örneği doğru sınıflandırmıştır. Buna karşılık, 5483 pozitif örnek negatif sınıfa, 4309 pozitif örnek ise notr sınıfa yanlış atanmıştır. Pozitif sınıf için modelin doğruluk oranı yüksek olmakla birlikte, negatif ve notr sınıflar arasında belirli bir karışıklık gözlenmektedir.

Genel olarak, model en iyi performansı notr ve pozitif sınıflarda göstermektedir. Bununla birlikte, negatif sınıf diğer sınıflarla karıştırılmaya daha yatkın görünmektedir. Bu durum, modelin negatif örnekleri doğru sınıflandırma konusunda sorun yaşadığını göstermektedir. Modelin negatif sınıf üzerindeki performansını artırmak için, veri seti içindeki negatif örnek sayısı arttırabilir.

Destek Vektör Makineleri için elde edilen karmaşıklık matrisi Şekil 19'da yer almaktadır.



Şekil 19. Destek Vektör Makineleri modeli için elde edilen karmaşıklık matrisi

Destek Vektör Makineleri (DVM) algoritması için elde edilen Şekil 19'da yer alan karmaşıklık matrisine dikkate alınarak modelin performansını şu şekilde değerlendirebiliriz:

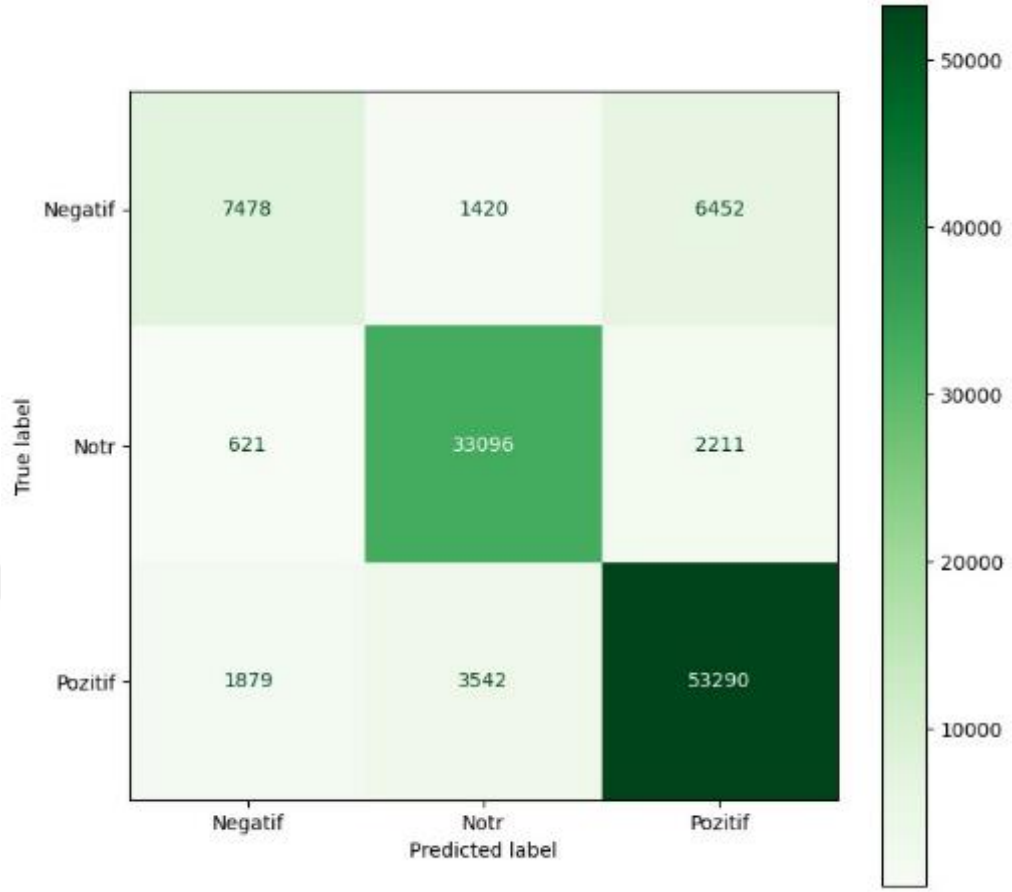
Negatif sınıf: Model, negatif sınıfa ait 7375 örneği doğru bir şekilde sınıflandırmıştır. Buna karşılık, 1859 negatif örnek notr olarak, 6116 negatif örnek ise pozitif olarak yanlış sınıflandırılmıştır. Negatif sınıfta, modelin diğer sınıflarla karıştırma oranının yüksek olduğu görülmektedir. Bu durum modelin negatif örnekleri ayırt etmede bazı zorluklar yaşadığını göstermektedir.

Notr sınıf: Notr sınıfa ait 32.785 örnek doğru sınıflandırılmıştır. Ancak, 903 notr örnek negatif sınıfa, 2240 notr örnek ise pozitif sınıfa yanlış olarak sınıflandırılmıştır. Notr sınıfında modelin performansı yüksek olduğu görülmektedir. Ancak pozitif sınıfla karışıklık göze çarpmaktadır.

Pozitif sınıf: Model, pozitif sınıfa ait 51.817 örneği doğru sınıflandırmıştır. Bununla birlikte, 2664 pozitif örnek negatif olarak, 4230 pozitif örnek ise notr olarak yanlış sınıflandırılmıştır. Pozitif sınıfta genel olarak yüksek bir doğruluk gözlemlenmekle birlikte, notr sınıfla karışıklıklar bulunmaktadır.

Genel olarak, Destek Vektör Makineleri algoritması en iyi performansı pozitif ve notr sınıflarda göstermektedir. Ancak, negatif sınıfta sınıflar arası karışıklık diğer iki sınıfa göre daha fazladır. Negatif örneklerin yanlış sınıflandırılması modelin genel performansını olumsuz etkilemektedir. DVM algoritmasının bu performansı, özellikle negatif sınıf için daha fazla veriye ihtiyaç olduğunu işaret etmektedir.

Rastgele Orman algoritması için elde edilen karmaşıklık matrisi Şekil 20'de yer almaktadır.



Şekil 20. Rastgele Orman modeli için elde edilen karmaşıklık matrisi

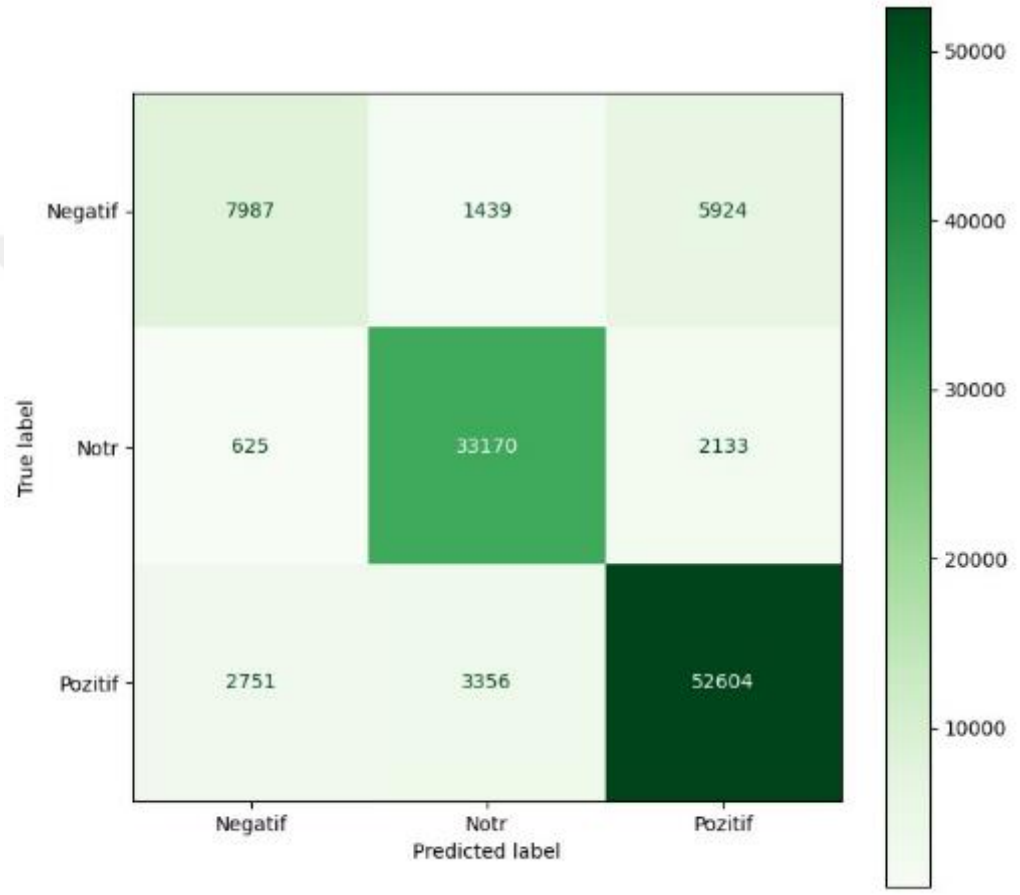
Rastgele Orman algoritması için elde edilen karmaşıklık matrisi incelendiğinde modelin performansı şu şekilde yorumlanabilir:

Negatif sınıf: Model, negatif sınıfa ait 7478 örneği doğru bir şekilde sınıflandırmıştır. Buna karşılık 1420 negatif örnek notr olarak, 6452 negatif örnek ise pozitif olarak yanlış sınıflandırılmıştır. Negatif sınıfta diğer sınıflarla karışıklık gözlenmekte olup, özellikle pozitif sınıfa önemli bir sayıda yanlış sınıflandırma yapılmıştır.

Notr sınıf: Notr sınıfa ait 33.096 örnek doğru sınıflandırılmıştır. 621 notr örnek negatif sınıfa, 2211 notr örnek ise pozitif sınıfa yanlış atanmıştır. Notr sınıf için modelin performansı yüksektir olduğu görülmektedir. Fakat pozitif sınıfa karışma oranının yüksek olduğu göze çarpmaktadır.

Pozitif sınıf: Model, pozitif sınıfa ait 53.290 örneği doğru bir şekilde sınıflandırmıştır. Buna karşılık, 1879 pozitif örnek negatif olarak, 3542 pozitif örnek ise notr olarak yanlış sınıflandırılmıştır. Pozitif sınıfta genel doğruluk yüksek olmasına rağmen, notr sınıfla olan karışıklık oranının yüksek olduğu gözlenmektedir.

Genel olarak, Rastgele Orman algoritması pozitif ve notr sınıflarda iyi performans göstermektedir. Buna karşılık negatif sınıfta ise karışıklık daha yüksektir. Bu, özellikle negatif örneklerin pozitif sınıf ile karıştırılmaya eğilimli olduğunu göstermektedir. Modelin genel performansı iyi olmakla birlikte, negatif sınıfın diğer sınıflarla olan karışıklığını azaltmak için veri setinde dengeleme işlemi yapılarak sorun giderilebilir.



Şekil 21. UKSB modeli için elde edilen karmaşıklık matrisi

UKSB modeli için elde edilen karmaşıklık matrisi Şekil 21'den incelendiğinde elde edilen sonuçlar modelin performansı için şu şekilde yorumlanabilir:

Negatif sınıf: UKSB modeli, negatif sınıfa ait 7987 örneği doğru sınıflandırmıştır. Bununla birlikte, 1439 negatif örnek notr olarak, 5924 negatif örnek ise pozitif olarak yanlış sınıflandırılmıştır. Negatif sınıf ile diğer sınıflar arasında belirli bir karışıklık gözlemlenmekte olup, özellikle pozitif sınıfa hatalı sınıflandırmalar yapılmaktadır.

Notr sınıf: Notr sınıfa ait 33.170 örnek doğru bir şekilde sınıflandırılmıştır. Ancak 625 notr örnek negatif olarak, 2133 notr örnek ise pozitif sınıfa yanlış sınıflandırılmıştır. Notr sınıfı için modelin genel doğruluğu oldukça yüksektir ve karışıklıklar nispeten azdır.

Pozitif sınıf: Pozitif sınıfa ait 52.604 örnek doğru sınıflandırılmıştır. Bununla birlikte, 2751 pozitif örnek negatif sınıfa, 3356 pozitif örnek ise notr sınıfa yanlış sınıflandırılmıştır. Pozitif sınıf için modelin doğruluk oranı oldukça yüksektir, ancak özellikle notr sınıfla karışıklıklar dikkate değerdir.

Genel olarak, UKSB modeli en iyi performansı pozitif ve notr sınıflarda sergilemektedir. Bununla birlikte, negatif sınıf diğer sınıflarla daha fazla karıştırılmaktadır. Özellikle pozitif sınıf ile negatif örneklerin yanlış sınıflandırılması sıkça karşılaşılan bir durumdur. Bu sonuçlar, UKSB modelinin dil verilerinde uzun vadeli bağıntıları öğrenme kabiliyetini ortaya koyarken, negatif sınıfın doğru sınıflandırılmasında bazı iyileştirme gerektirdiğini göstermektedir. Modelin performansını artırmak için, negatif sınıfa ait örneklerin sayısının artırılmasının performansı artıracağı beklenmektedir.

En yüksek performansa sahip olan UKSB modelinin model bilgileriyle ilgili detaylar Tablo 7'de ifade edilmektedir.

Tablo 7. UKSB modeline ait bilgiler

Katman	Tür	Çıkış Şekli	Parametre Sayısı	Açıklama
Embedding	Embedding	(None, 100, 128)	640,000	Kelime indekslerini 128 boyutlu vektörlere dönüştürmektedir. Model, embedding matrisi olarak 5000 farklı kelimeyi temsil edebilen bir matris kullanmaktadır.
SpatialDropout1D	SpatialDropout1D	(None, 100, 128)	0	Model overfitting hatasına düşmesini engellemek için eğitim esnasında %20 oranında rastgele özellik sıfırlama yapılmıştır. Bu işlem aynı zamanda modelin genelleme yeteneği arttıran bir yaklaşımdır.

Tablo 7’nin devamı

UKSB	UKSB	(None, 100)	91,600	UKSB katmanı, sıralı veriler içindeki uzun süreli bağımlılıkları öğrenmektedir. Bu katman, 100 UKSB birimine sahiptir ve dropout ile recurrent dropout oranları %20'dir.
Dense	Dense	(None, len (le.classes))	303	fully connected layer, aşamasında softmax aktivasyonu kullanılarak çok sınıflı bir sınıflandırma işlemi gerçekleştirilmektedir. Bu işlem katman sonucunda çıkış boyutu, sınıf sayısına eşittir.
Toplam Parametre Sayısı			731,903	
Eğitilebilir Parametre			731,903	
Eğitilemez Parametre			0	

Kelime girdilerinin sayısal vektörlere dönüştürülmesi için kullanılan Embedding katmanı, kelimeler arasındaki semantik ilişkilerin yakalanmasını sağlamaktadır. Bu çalışmada, 5000 farklı kelimeyi temsil eden bir embedding matrisi kullanılmış ve her bir kelime, dilin bağlamsal ilişkilerini temsil eden 128 boyutlu vektörlerle ifade edilmiştir. Embedding katmanı, modelin dilin semantik yapılarını öğrenmesi ve kelimeler arasındaki anlamsal benzerlikleri yakalaması açısından temel bir bileşen olarak görev yapmaktadır (Zhang vd., 2020).

Modelin aşırı öğrenme (overfitting) problemine karşı daha dayanıklı hale getirilmesi amacıyla SpatialDropout1D katmanı kullanılmıştır. Eğitim esnasında %20 oranında rastgele özellik sıfırlaması yaparak modelin genelleme yeteneği arttırmaktadır. Sıralı veri ile çalışıldığından, dropout işleminin sıralı veri yapısına uygun hale getirilmesi gerekmektedir. Bundan dolayı SpatialDropout1D katmanı tercih edilmektedir. Bu katman sayesinde model, eğitim verilerine aşırı uyum sağlama eğiliminden korunarak daha genellenebilir bir performans sergilemektedir (Nayoga vd., 2021).

UKSB katmanı ise sıralı verilerde uzun süreli bağımlılıkları öğrenme kapasitesine sahip olup, dilin bağlamını koruma konusunda önemli bir rol üstlenmektedir. Bu katmanda, 100 UKSB birimi bulunmakta olup giriş ve dönüş bağlantıları arasındaki %20 oranında

dropout ve recurrent dropout uygulanmıştır. Bu yapı, modelin yalnızca kısa vadeli bağımlılıkları değil, aynı zamanda uzun süreli bağlamları da başarılı bir şekilde öğrenmesini sağlamaktadır. UKSB katmanı, metin verileri arasındaki anlamlı ilişkileri ve bağıntıları modelleyerek, metin sınıflandırma görevlerinde modelin başarısını artırmaktadır (Hochreiter ve Schmidhuber, 1997).

Modelin çıkış katmanı olan Dense katmanı, çoklu sınıflandırma görevini gerçekleştirmek için Softmax aktivasyon fonksiyonu kullanmaktadır. Bu katman, modelin öğrenilen özellikler doğrultusunda metinleri belirlenen sınıflara doğru bir şekilde atamasını sağlayan önemli bir bileşendir.

Modelin eğitim sürecinde "sparse_categorical_crossentropy" kayıp fonksiyonu ve "adam" optimizasyon algoritması kullanılmaktadır. Eğitim esnasında farklı parametrelerle fine-tuning işlemi yapılmış ve bu parametreler arasında en iyi sonuçların "adam" optimizasyon algoritması ve "sparse_categorical_crossentropy" kayıp fonksiyonu ile elde edildiği görülmüştür.

Model, 10 epoch boyunca 64'lük batch size ile eğitilmiş ve validasyon veri seti üzerinde %86 doğruluk oranı elde edilmektedir. Bu sonuç, modelin metin sınıflandırma ve duygu analizi görevlerinde yüksek performans göstereceğini ortaya koymaktadır. Modelin elde ettiği başarı, doğal dil işleme alanında kullanılan UKSB temelli modellerin etkili bir çözüm sunduğunu göstermektedir.

Model performanslarının detaylı bir şekilde incelenmesi, çalışmanın temel unsurlarından biridir. Bu değerlendirmelerden sonra, sistemin genel analiz sonuçlarına odaklanılmaktadır. Geliştirilen sistemin uçtan uca analiz sonuçları ele alınarak, elde edilen bulguların kapsamlı bir değerlendirmesi yapılmaktadır.

Sistemin devreye alındığın günden itibaren yapmış topladığı haberle ilişkin bilgiler Tablo 8'de ifade edilmektedir.

Tablo 8. Haber toplama ve analiz işlemlerine ait veriler

Negatif Haber Sayısı	Pozitif Haber Sayısı	Nötr Haber Sayısı	Ortalama (Günlük)	Toplam
36.294	99.492	43.456	1493	179.242

Haber toplama ve analiz süreçlerine ait veriler, toplanan haberlerin duygu dağılımı hakkında önemli bulgular sunmaktadır. Toplamda 36.294 haber negatif, 99.492 haber pozitif

ve 43.456 haber nötr olarak sınıflandırılmıştır. Bu dağılım, analiz edilen içeriklerin yarısından fazlasının pozitif haberlerden oluştuğunu, geri kalan kısmın ise nötr ve negatif haberlerden meydana geldiğini göstermektedir. Günlük ortalama olarak 1.493 haber işlenmiş ve toplamda 179.242 haber analiz edilmiştir. Burada elde edilen sınıflandırılmış haber metinleri, duygu analizi modelinin performansını arttırmak amacıyla tekrar model eğitiminde kullanılması planlanmaktadır.

Bu kapsamda ilk olarak ulusal haber kaynaklarından elde edilen ve içerik olarak negatif sınıflandırılmış haberlerin durumu değerlendirilmiştir. Negatif olarak sınıflandırılan haberlerin, bankacılık açısından önem taşıyan hangi anahtar kelimeleri içerdiği incelenmiştir. Analiz sonuçları incelendiğinde, ulusal haber kaynaklarında şirketler hakkında kullanılan negatif ilk on kelime Tablo 9'daki gibi elde edilmektedir.

Tablo 9. Ulusal haber ajanslarından elde edilen negatif haberlerin bankacılık açısından sahip olduğu negatif anahtar kelimeler dağılımı

Negatif Kelimeler	Yüzde (%)
ceza	0,12
enflasyon	0,09
sorun	0,07
düşüş	0,06
zarar	0,05
iddia	0,05
mahkeme	0,05
gerileme	0,05
kayıp	0,04
dava	0,04

Elde edilen sonuçlar incelendiğinde %12'lik kullanım sıklığıyla en yüksek negatif kelimenin "ceza" kelimesi olduğu anlaşılmaktadır. "Enflasyon" kelimesinin %9'luk kullanım oranıyla en çok kullanılan ikinci negatif kelime olduğu anlaşılmaktadır. "Sorun" kelimesinin kullanım oranın %7, "düşüş" kelimesinin %6, "zarar", "iddia", "mahkeme" ve "gerileme" kelimelerinin %5, "kayıp" ve "dava" kelimelerinin ise %4'lük kullanım oranına sahip olduğu görülmektedir. Haber kaynaklarından elde edilen negatif içerikler firmalarla ilgili risk faktörlerinin tespit edilmesinde önemli bir rol oynamaktadır. Aşağıda, ulusal haber

kaynaklarından negatif olarak tespit edilen terimlerin bankacılık bakış açısıyla incelenmesi yapılmıştır.

Ceza: “Ceza” kelimesi, kanuni yaptırımları ve ceza gerektiren işlemleri ifade etmektedir (Gökcan, 2015). Bankacılık açısından değerlendirildiğinde bu tarz haberlerin ilgili firmaların kanuni risklere maruz kalabileceğini ifade etmektedir. Bir firmanın hukuki yaptırımlara maruz kalması, firmanın finansal durumunu olumsuz etkileyecektir. Aynı zamanda bu durum kredi riskinin artmasına neden olabilmektedir.

Enflasyon: Enflasyon, ekonomik dalgalanmalarla doğrudan ilişkili bir kavramdır. Haber içeriklerinde çokça yer alması, bir istikrarsızlık göstergesi olabilmektedir. Enflasyonun yüksek olması, satın alma gücünde düşüşe, faiz oranlarında ise artışa neden olmaktadır. Aynı zamanda tüketici güveninin düşmesine neden olmaktadır (Oktar & Dalyancı, 2011). Bankacılık açısından değerlendirildiğinde, enflasyonla alakalı haberlerin çok olması, ekonomide durgunluk dönemi olduğuna ve kredi verme süreçlerinde daha dikkatli olunması gerektiğini göstermektedir.

Sorun: “Sorun” kelimesi şirket veya sektör de yaşanan sıkıntıları ifade etmektedir. Bankalar bu tür haberlere konu olan firmalara karşı dikkatli olmalıdır. Bu habere konu olan firmaların özellikle kredilerin geri ödemelerinde aksaklıklar yaşanabileceği göz önünde bulundurulmalıdır.

Düşüş: “Düşüş” kelimesi bankacılık açısından ekonomik gerilemeyi ve potansiyel kredi risklerini ifade etmektedir. Belirli sektörlerle ilgili çok fazla düşüş haberlerinin olması, o sektörde faaliyet gösteren firmalarla ilgili kredi verme kararlarının tekrar gözden geçirilmesine neden olabilir.

Zarar, iddia, mahkeme: Bu üç kelime incelendiğinde, firmaların yasal ve finansal bazı zorluklarla karşı karşıya kaldıkları sonucu çıkarılabilir. “Zarar” kelimesi, firmaların finansal açıdan zor bir durumda olduğunu belirtirken, “iddia” ve “mahkeme” kelimeleri, yargı süreçlerini belirtmektedir. Bankalar açısından bu tespitler, kredi riski oluşturan firmaların tespit edilmesine imkan sağlayabilir.

Gerileme: “Gerileme” kelimesi bankacılık açısından potansiyel olarak kredi risklerinin habercisi olabilir. Firma ve faaliyet gösterdiği sektörde ekonomik gerileme yaşanması durumunda, bankanın bu firmaya kredi verme konusunu dikkatlice değerlendirmesi gerekebilir.

Kayıp ve Dava: “Kayıp” ve “dava” kelimeleri maddi ve hukuki sorunları belirtmektedir. Firmalarla ilgili bu tarz haberlerin çokça görülmesi, ilgili firmaların yasal

veya finansal sorunlar yaşadığını göstermektedir. Bankaların bu tarz haberlere konu olan firmaların kredi risklerini dikkatlice değerlendirmesi gerekebilir.

Benzer şekilde ulusal haber kaynaklarından elde edilen ve pozitif olarak sınıflandırılan haberlerin, bankacılık açısından önem taşıyan hangi anahtar kelimeleri içerdiği incelenmiştir. Analiz sonuçları incelendiğinde, ulusal haber kaynaklarında şirketler hakkında kullanılan pozitif ilk on kelime Tablo 10'daki gibi elde edilmektedir.

Tablo 10. Ulusal haber ajanslarından elde edilen pozitif haberlerin bankacılık açısıyla sahip olduğu pozitif anahtar kelimeler dağılımı

Pozitif Kelimeler	Yüzde (%)
önemli	0,09
artış	0,06
yatırım	0,06
güçlü	0,04
gelir	0,04
yükselme	0,03
hızlı	0,03
fırsat	0,03
başarı	0,03
büyüme	0,03

Elde edilen sonuçlar incelendiğinde %9'luk kullanım sıklığıyla en yüksek pozitif kelimenin "önemli" kelimesi olduğu anlaşılmaktadır. "artış" ve "yatırım" kelimelerinin %6'lık kullanım oranıyla en çok kullanılan ikinci pozitif kelimeler olduğu anlaşılmaktadır. "güçlü" ve "gelir" kelimelerinin kullanım oranın %4, "yükselme", "hızlı", "fırsat", "başarı" ve "büyüme" kelimelerinin %3'lük kullanım oranına sahip olduğu görülmektedir. Aşağıda, ulusal haber kaynaklarından pozitif olarak tespit edilen terimlerin bankacılık bakış açısıyla incelenmesi yapılmıştır.

Önemli: "önemli" kelimesinin çokça kullanılması, bankalar açısından firmalarla ilgili geleceğe dönük olumlu beklentileri ifade edebileceği düşünülmektedir.

Artış: "artış" kelimesi, bankalar açısında değerlendirildiğinde firmayla ilgili kredi riskinin azalmasını ifade edilebileceği düşünülmektedir.

Yatırım: “yatırım” kelimesi, bankalar açısından değerlendirildiğinde, firmanın gelecek dönemlerdeki potansiyel büyüme beklentilerini yansıtmaktadır.

Güçlü: “güçlü” kelimesi, bankacılık bakış açısıyla değerlendirildiğinde bu tür haberlere konu olan firmaların, kredi riskinin düşük olacağı ve firmanın kredi geri ödeme kapasitesinin yüksek olduğunu ifade etmektedir.

Yükselme: “yükselme” kelimesi, finansal göstergelerde olumlu bir değişimi ve iyileşmeyi ifade etmektedir. Bankalar için bu tarz haberlerin, yatırım stratejilerini belirlemede faydalı olacağı düşünülmektedir.

Hızlı: “hızlı” kelimesi, bankalar için piyasada haraketliliğin ve talebin artışının bir göstergesi olabilir.

Fırsat: “fırsat” kelimesi, bankacılık açısından değerlendirildiğinde yeni yatırım alanlarının ve ekonomik kazançların göstergesi olarak yorumlanabilir. Bankalar bu tarz haberlere konu olan firmalarla kredi hacimlerini genişletme yoluna gidebilir.

Başarı: “başarı” kelimesi, firmaların belirledikleri hedeflere ulaştığı şeklinde yorumlanabilir. Bankalar açısında değerlendirildiğinde, ilgili firmanın kredi geri ödeme kapasitesini arttırdığının göstergesi olabilir.

Büyüme: “büyüme” kelimesi, ilgili firmanın faaliyetlerinde veya finansal performansında artış olduğu şeklinde yorumlanabilir. Büyüme odaklı haberlere konu olan firmalar bankalar için yeni fırsatlar yaratabilir.

Sonuç olarak negatif ve pozitif haber içeriklerine ait kelimeler dikkate alındığında, bankacılık sektörünün bu analizlerden elde ettiği veriler, risk ve fırsat yönetimi açısından büyük önem taşımaktadır. Negatif kelimeler arasında yer alan "ceza", "enflasyon", "sorun", "düşüş" ve "zarar" gibi ifadeler, firmaların karşı karşıya kaldıkları ekonomik ve hukuki riskleri işaret etmektedir. Bu tür durumlar bankalar için kredi riskinin artmasına neden olabilecek faktörlerdir. Özellikle, cezai yaptırımlar, ekonomik gerileme ve finansal kayıplar, firmaların kredi geri ödeme kapasitelerini zayıflatmaktadır.

Diğer yandan, pozitif kelimeler arasında yer alan "önemli", "yatırım", "güçlü", "gelir" ve "büyüme" gibi ifadeler, bankaların kredi sağlayabileceği potansiyel firmaların büyüme eğiliminde olduğunu göstermektedir. Bu firmalar, artan gelirleri ve güçlü finansal yapıları sayesinde bankalar açısından düşük riskli müşteriler olarak değerlendirilebilir. Dolayısıyla, haber içeriklerinden elde edilen bu veriler, bankaların hem olası riskleri belirlemesinde hem de fırsatları değerlendirmesinde kritik bir role sahiptir. Bankalar, bu analizlerle kredi

politikalarını şekillendirerek, hem ekonomik durgunluk dönemlerinde riskleri minimize edebilir hem de büyüme fırsatlarını yakalayabilir.

Geliştirilen sistemle ulusal haber kaynaklarından elde edilen negatif ve pozitif içerikler bankacılık açısından detaylı bir şekilde analiz edilmiştir. Bu analiz firmaların karşılaştıkları risk ve fırsatlar hakkında önemli bilgiler sunmaktadır. Ancak ulusal haber kaynaklarının yanı sıra yerel haber kaynakları da firmalarla ilgili önemli bilgiler içermektedir. Yerel haber kaynakları, bankaların bölgesel seviyedeki risk ve fırsatları değerlendirmesinde önemli bir rol oynamaktadır.

Yerel haber kaynaklarında tespit edilen kelimelerin analiz edilmesi, yerel firmaların durumunu daha yakından incelenmesine olanak sağlamaktadır. Bu bağlamda, çalışmanın devamında yerel haber kaynaklarındaki negatif ve pozitif içerikler analiz edilerek, bankacılık bakış açısıyla yerel firmalarla ilgili risk ve fırsatların belirlenmesine yönelik kapsamlı bir değerlendirme yapılacaktır. Bunun yanı sıra ulusal haber kaynakları analiz sonuçlarıyla yerel haber kaynakları analiz sonuçları karşılaştırılarak elde edilen sonuçlar değerlendirilecektir.

Yerel haber kaynaklarından elde edilen ve negeatif olarak sınıflandırılan haberlerin, bankacılık açısından önem taşıyan hangi anahtar kelimeleri içerdiği incelenmiştir. Analiz sonuçları incelendiğinde, yerel haber kaynaklarında şirketler hakkında kullanılan negatif ilk on kelime Tablo 11'deki gibi elde edilmektedir.

Tablo 11. Yerel haber ajanslarından elde edilen negatif haberlerin bankacılık açısıyla sahip olduğu negatif anahtar kelimeler dağılımı

Yerel Negatif Kelime	Yüzde (%)
ceza	0,03
soruşturma	0,03
sorun	0,03
iddia	0,03
hkeme	0,03
kayıp	0,02
zarar	0,02
enflasyon	0,02
dava	0,02
borç	0,02

Elde edilen sonuçlar incelendiğinde %3'lük kullanım sıklığıyla "ceza", "soruşturma", "sorun", "iddia" ve "mahkeme" kelimelerinin en yüksek kullanıma sahip kelimeler olduğu anlaşılmaktadır. "kayıp", "zarar", "enflasyon", "dava" ve "borç" kelimelerinin ise %2'lik bir kullanıma sahip olduğu anlaşılmaktadır. Aşağıda, yerel haber kaynaklarından negatif olarak tespit edilen terimlerin bankacılık bakış açısıyla incelenmesi yapılmıştır.

Ceza: "ceza" kelimesi, yerel işletmelerin yasal yaptırımlarla karşı karşıya kalma riskini ifade etmektedir. Küçük ve orta ölçekli firmalarda, cezai yaptırımlar, firmaların finansal yapısını olumsuz etkileyebilir.

Soruşturma: "soruşturma" kelimesi, firmaların yasal inceleme altında olduğunu göstermektedir. Bu tür haberler bankalar için önemli ipuçları verebilir. Soruşturma altında olan bir işletme kredi ödemelerinde zorluk yaşayabilir.

Sorun: "sorun" kelimesi, bankacılık açısından değerlendirildiğinde, firmanın kredi riskinin artma ihtimaline sahip olduğunu gösterebilir.

İddia: "iddia" kelimesi, firmalar hakkında ortaya atılan yasal veya finansal şüpheleri yansıtır. Bu tür iddialar, firmaların itibarını ve finansal yapısını olumsuz etkileyebilir. Bankalar açısından, iddia konusu olan firmaların kredi riski yüksek olabilir ve bu durum geri ödeme performanslarını olumsuz etkileyebilir.

Mahkeme: Bankalar için mahkemeye konu olan firmalar, yüksek riskli müşteriler olabilir. Bankaların bu müşterilere kredi verme konusunda dikkatli olunması gerekebilir. Mahkeme süreçleri, firmanın kredi geri ödeme kapasitelerini olumsuz yönde etkileyebilir.

Kayıp: "kayıp" kelimesi, bankacılık açısından değerlendirildiğinde, firmanın finansal kayıplarını ifade etmektedir. Bu tür haberler konu olan firmaların kredi geri ödeme kapasitelerinde düşüş olabilir.

Zarar: "zarar" kelimesi, bankacılık bakış açısıyla değerlendirildiğinde, zrar haberleri, finansal risklerin bir işareti olabilir

Enflasyon: "enflasyon" kelimesi, bankalar açısından değerlendirildiğinde, yerel işletmelerin kârlılığını azaltarak kredi geri ödeme kapasitelerini zayıflatabilir

Dava: "dava" kelimesi, bankacılık bakış açısıyla değerlendirildiğinde, dava süreçleri içinde olan firmalar, yüksek riskli firmalar olabilir. Dava süreçleri firmaların finansal durumlarına olumsuz etki edebilir. Bu nedenle bu firmaların daha detaylı incelenmelidir.

Borç: "borç" kelimesi, bankacılık bakış açısıyla değerlendirildiğinde, borç altında olan firmaların kredi riski dikkatle değerlendirmelidir.

Benzer şekilde yerel haber kaynaklarından elde edilen ve pozitif olarak sınıflandırılan haberlerin, bankacılık açısından önem taşıyan hangi anahtar kelimeleri içerdiği incelenmiştir. Analiz sonuçları incelendiğinde, yerel haber kaynaklarında şirketler hakkında kullanılan pozitif ilk on kelime Tablo 12'teki gibi elde edilmektedir.

Tablo 12. Yerel haber ajanslarından elde edilen pozitif haberlerin bankacılık açısıyla sahip olduğu pozitif anahtar kelimeler dağılımı

Pozitif Kelime	Yüzde (%)
önemli	0,03
işgücü	0,02
başarı	0,02
artış	0,02
güzel	0,02
yatırım	0,02
gelir	0,02
güçlü	0,02
fırsat	0,02
yükselme	0,02

Elde edilen sonuçlar incelendiğinde %3'lük kullanım sıklığıyla "önemli" kelimesinin en çok kullanılan pozitif kelime olduğu anlaşılmaktadır. "işgücü", "başarı", "artış", "güzel", "yatırım", "gelir", "güçlü", "fırsat" ve "yükselme" kelimelerinin %2'lik kullanıma sahip olduğu görülmektedir. Aşağıda, yerel haber kaynaklarından negatif olarak tespit edilen terimlerin bankacılık bakış açısıyla incelenmesi yapılmıştır.

Önemli: "önemli" kelimesi, bankalar açısından geleceğe dönük fırsatların veya olumlu gelişmelerin işareti olarak yorumlanabilir.

İşgücü: "işgücü" kelimesi, bankacılık bakış açısıyla değerlendirildiğinde, bu habere konu olan firmaların büyüme ve gelişme potansiyeline sahip olduğu düşünülebilir.

Başarı: "başarı" kelimesi bankacılık bakış açısıyla değerlendirildiğinde, firmaların mali açıdan sağlıklı olduğunu ve kredi geri ödeme kapasitelerinin güçlü olduğu söylenebilir.

Artış: "artış" kelimesi, bankacılık bakış açısıyla değerlendirildiğinde, firmaların büyüme ve gelişim potansiyeline ifade etmektedir. Bu tür haberlere konu olan firmalar, kredi riskinin azalacağı firmalar olarak değerlendirilebilir.

Güzel: “güzel” kelimesi, bankacılık bakış açısıyla değerlendirildiğinde, genel ekonomik göstergelerin veya sosyal koşulların olumlu seyrettiğini gösterebilir. Bu durum bankaları firma için kredi stratejisini destekleyebilir.

Yatırım: “yatırım” kelimesi bankacılık bakış açısıyla değerlendirildiğinde, büyük bir fırsat olarak değerlendirilmektedir. Yatırım yapan firmalar, uzun vadeli büyüme potansiyeline sahiptir ve bankalar için kredi sağlanabilecek güvenilir müşteriler arasında yer almaktadır.

Gelir: “gelir” kelimesi bankacılık bakış açısıyla değerlendirildiğinde, firmaların finansal performansının güçlendiğini gösterir. Gelir artışları, kredi geri ödeme kapasitesini artıran bir güçtür ve bankalar bu firmalara daha fazla kredi verebilir.

Fırsat: “fırsat” kelimesi bankacılık bakış açısıyla değerlendirildiğinde, bölgesel ekonomideki olumlu gelişmeleri ve bankalar açısından yatırım yapılabilir yeni alanları ifade edebilir.

Yükselme: “yükselme” kelimesi bankacılık bakış açısıyla değerlendirildiğinde, firmaların finansal performansındaki pozitif değişimi ifade eder. Bu tür haberlere konu olan firmaların büyüme trendinde olduğu değerlendirilir. Bu firmalara sağlanan kredi riskinin düşük olacağı düşünülmektedir.

Yerel haber kaynaklarından elde edilen negatif ve pozitif içeriklerin değerlendirilmesi, bankacılık sektörü açısından bölgesel riskler ve fırsatlar hakkında önemli ipuçları sunmaktadır. Negatif kelimeler arasında yer alan "ceza", "soruşturma", "sorun", "mahkeme" ve "borç" gibi terimler, yerel düzeydeki firmaların hukuki ve mali zorluklarla karşı karşıya olduğunu işaret etmektedir. Bu tür haberler, firmaların kredi geri ödeme kapasitelerini olumsuz etkileyebilecek risk faktörleri olarak değerlendirilmektedir. Özellikle soruşturma ve dava süreçlerine dahil olan firmalar, bankalar için yüksek risk teşkil etmektedir.

Diğer yandan, pozitif kelimeler arasında yer alan "önemli", "işgücü", "başarı", "yatırım" ve "gelir" gibi ifadeler, bölgesel ekonomide olumlu gelişmeleri işaret etmektedir. Yerel firmaların yatırım yapması, güçlü işgücü ile büyüme göstermesi ve gelir artışı gibi unsurlar, bankalar açısından düşük riskli ve kredi verilebilir müşteri profilleri oluşturmaktadır. Özellikle başarı ve fırsat kelimeleri, bankaların yerel firmalarla uzun vadeli işbirlikleri kurabilmesi ve yeni kredi olanakları yaratabilmesi için olumlu bir ortam sunmaktadır.

Sonuç olarak, yerel haber kaynaklarından elde edilen bu analizler, bankaların kredi risklerini yönetmelerinde ve yatırım kararlarını şekillendirmelerinde önemli bir rol

oynamaktadır. Negatif içerikler, yerel firmaların karşılaştıkları risklere dikkat çekerken, pozitif içerikler bölgesel düzeydeki ekonomik fırsatları gözler önüne sermektedir. Bankalar, bu tür analizleri dikkate alarak kredi politikalarını daha etkin bir şekilde yönetebilir ve yerel düzeyde büyüme fırsatlarını değerlendirebilir.

Yukarıda ulusal haber kaynakları ve yerel haber kaynaklarından elde edilen negatif ve pozitif kelimelerin bankacılık açısından değerlendirilmesine yer verilmiştir. Fakat ulusal ve yerel haberlerde öne çıkan kelimeler arasındaki farklılıklar derinlemesine incelenmelidir.

Ulusal haberlerde sıkça rastlanan "ceza", "enflasyon", "sorun", "düşüş" ve "zarar" gibi kelimeler, genellikle büyük ölçekli ekonomik krizleri ya da genel ekonomik istikrarsızlıkları işaret etmektedir. Özellikle enflasyon ve düşüş gibi terimler, geniş çaplı ekonomik durgunlukların ve finansal piyasalardaki dalgalanmaların göstergesi olmaktadır. Bankalar açısından bu tür haberler, genel kredi riskini artıran unsurlar olarak değerlendirilmektedir. Ulusal düzeyde ekonomik olumsuzlukların yoğun olduğu haber içerikleri, bankaların kredi verme süreçlerini daha temkinli yürütmelerini gerektirir. Özellikle büyük ölçekli krediler ve uzun vadeli yatırımlar bankalar açısından önemli bir risk oluşturabilir.

Tablo 13. Ulusal ve yerel haber kaynaklarından elde edilen en çok kullanın 10 negatif kelimelerin dağılımı

Negatif Kelime (Ulusal)	Yüzde (%)	Negatif Kelime (Yerel)	Yüzde (%)
ceza	0,12	ceza	0,13
enflasyon	0,09	soruşturma	0,08
sorun	0,07	sorun	0,08
düşüş	0,06	iddia	0,08
zarar	0,05	mahkeme	0,07
iddia	0,05	kayıp	0,06
mahkeme	0,05	zarar	0,06
gerileme	0,05	enflasyon	0,06
kayıp	0,04	dava	0,05
dava	0,04	borç	0,04

Yerel haber kaynaklarında ise "ceza", "soruşturma", "sorun", "mahkeme", "borç" gibi terimler ön plana çıkmaktadır. Yerel düzeyde bu kelimeler, genellikle daha özel ve bölgesel ekonomik sorunlara ya da küçük ve orta ölçekli işletmelerin yaşadığı yasal ve finansal

sorunlara işaret etmektedir. Örneğin, borç ve mahkeme kelimeleri, yerel firmaların karşı karşıya kaldığı yasal süreçleri ve finansal zorlukları ifade edebilmektedir. Bankaların özellikle küçük ve orta ölçekli yerel işletmelere kredi verirken daha dikkatli bir risk analizi yapmasını zorunlu kılmaktadır.

Elde edilen sonuçlar incelendiğinde ulusal haber kaynakları, genellikle geniş çaplı ekonomik göstergelere ve ülke genelinde yaşanan makroekonomik sorunlara odaklanırken, yerel haber kaynakları daha çok bölgesel düzeydeki işletmelere ve onların karşılaştığı özel sorunlara odaklanmaktadır. Ulusal haberlerde sıkça geçen enflasyon ve düşüş gibi makroekonomik terimler, büyük ölçekli ekonomik risklerin işaretçisiyken, yerel haberlerdeki soruşturma, mahkeme ve borç gibi kelimeler, daha çok bireysel veya küçük ölçekli işletmelerin risklerini yansıtmaktadır. Dolayısıyla, bankaların ulusal haberlerdeki geniş kapsamlı riskleri yönetirken, yerel düzeyde daha mikro düzeyde, bölgeye özgü risk faktörlerini değerlendirmesi gerektiği sonucuna varılabilir.

Negatif içeriklerin bankacılık sektörüne yansıttığı risk unsurlarının ardından, ulusal ve yerel haber kaynaklarında pozitif içeriklerin de önemli ipuçları sunduğu görülmektedir. Bankacılık sektörü açısından pozitif içerikler, büyüme ve yatırım fırsatlarını ortaya çıkararak bankaların kredi portföylerini genişletmesine olanak tanımaktadır.

Tablo 14. Ulusal ve yerel haber kaynaklarından elde edilen en çok kullanın 10 pozitif kelimelerin dağılımı

Pozitif Kelime (Ulusal)	Yüzde (%)	Pozitif Kelime (Yerel)	Yüzde (%)
önemli	0,09	önemli	0,05
artış	0,06	işgücü	0,04
yatırım	0,06	başarı	0,04
güçlü	0,04	artış	0,03
gelir	0,04	güzel	0,03
yükselme	0,03	yatırım	0,03
hızlı	0,03	gelir	0,03
fırsat	0,03	güçlü	0,03
başarı	0,03	fırsat	0,03
büyüme	0,03	yükselme	0,03

Ulusal ve yerel haber kaynaklarında tespit edilen pozitif kelimeler, bankaların kredi verme ve yatırım kararlarında rehber niteliği taşımaktadır. Ulusal düzeyde öne çıkan "önemli" (0,09%), "artış" (0,06%), "yatırım" (0,06%) ve "gelir" (0,04%) gibi kelimeler, genel ekonomik büyümeyi ve firmaların finansal yapılarının güçlü olduğunu göstermektedir. Özellikle "yatırım" ve "artış" gibi terimler, geniş çaplı sermaye yatırımlarını ve ekonomik genişlemeleri işaret ederek bankalar için büyük ölçekli kredi fırsatları sunduğu düşünülmektedir.

Yerel haber kaynaklarında öne çıkan "işgücü" (0,04%), "başarı" (0,04%) ve "yatırım" (0,03%) gibi kelimeler ise, yerel düzeyde firmaların büyüme potansiyelini ve bölgesel ekonominin olumlu seyrettiğini göstermektedir. Özellikle "işgücü" kelimesi, yerel firmaların istihdam artışı sağladığını ve bölgesel düzeydeki ekonomik hareketliliğin arttığını gösterir. Bu tür firmalar, bankalar için uzun vadeli işbirlikleri ve sürdürülebilir kredi fırsatları yaratabilir.

4. SONUÇLAR VE ÖNERİLER

Bu tez çalışmasında, ulusal ve yerel haber kaynaklarından elde edilen veriler duygu analizi ve adlandırılmış varlık tanıma (AVT) teknikleri ile işlenerek firmaların finansal durumları ve risk analizleri yapılmasını sağlayan entegre bir sistem geliştirildi. Çalışma kapsamında geliştirilen bu sistem, bankacılık sektöründe karar destek süreçlerine katkı sunacak nitelikte yenilikçi bir bilgi yönetim aracıdır. Sistem, haber kazıma modülü, duygu analizi modeli, varlık tanıma modeli, bankacılık için özel olarak hazırlanan negatif ve pozitif kelime listesi ile haber içeriklerinin firma isimlerine göre ağırlıklandırılmasını sağlayan bir altyapıya dayanır.

Çalışmada geliştirilen haber kazıma modülü, ulusal ve yerel haber kaynaklarından geniş kapsamlı veri toplar. Toplanan veriler, duygu analizi ve firma ismi tespiti için işlenir. Haber toplama modülü, sürekli güncellenen veri akışı sağlayarak haber içeriklerinin anlık olarak sisteme aktarılmasını ve analiz edilmesini sağlar.

Geliştirilen sistemde duygu analizi için hem klasik makine öğrenmesi algoritmaları hem de derin öğrenme modelleri uygulandı. Klasik modellerin yanı sıra en yüksek performansı sergileyen UKSB modeli, duygu analizinde %86 F1-skoru ile öne çıktı. Duygu analizi, haber içeriklerinin pozitif veya negatif olarak sınıflandırılmasını sağlayarak bankaların riskli firmaları tespit etmesine ve fırsatları değerlendirmesine olanak tanır.

Haber metinlerindeki firma isimlerinin doğru tespiti, sistemin en önemli görevlerinden biridir. AVT modeli, haberlerde geçen firma isimlerini yüksek doğrulukla tespit ederek bu bilgileri duygu analiz sürecine aktarır. Bu sayede geliştirilen sistem, belirli firmalar için finansal risk değerlendirmesini daha sağlıklı bir şekilde gerçekleştirir.

Bankacılık sektörüne özgü olarak oluşturulan negatif ve pozitif kelime listesi, sistemin haber içeriklerini daha doğru analiz etmesine yardımcı olur. Bu liste, bankacılık açısından kritik öneme sahip “zarar”, “düşüş”, “ceza” gibi negatif ve “yatırım”, “büyüme”, “başarı” gibi pozitif kelimelerden oluşur. Liste, haberlerin duygu tonunun belirlenmesinde kilit rol oynar.

Metinlerde geçen firma isimleri ile haber içeriklerini ilişkilendiren bir ağırlıklandırma sistemi geliştirildi. Bu sistem, firma isimlerinin yer aldığı haberlerde hangi içeriğin hangi firmaya ait olduğunu daha doğru bir şekilde belirler. Bu sayede, firma bazlı duygu skorlamaları daha güvenilir ve tutarlı bir şekilde hesaplanır.

Çalışmada geliştirilen haber süpürme modülünün daha etkin ve güncel veri toplayabilmesi amacıyla 'Really Simple Syndication' sisteminin kullanılması planlanmaktadır. Really Simple Syndication ifadesi Türkçeye 'Zengin Site Özeti' (ZSÖ) olarak çevrilmiştir. ZSÖ sistemi, haber sitelerinden sürekli olarak güncel veri akışı sağlayarak haberlerin hızlı biçimde sisteme aktarılmasına olanak tanır. Bu sistem, haber süpürme işlemini daha düzenli ve sistematik hale getirirken veri güncelliğini korur ve bankaların anlık analiz yapmasına imkan tanır. Böylece, haber kaynaklarından veri toplama süreçleri hızlanmakta, bankaların karar süreçlerine daha güncel ve gerçek zamanlı bilgi sunulması hedeflenmektedir.

Geliştirilen duygu analizi modelinin ilerleyen aşamalarda performansını artırmak ve veri dengesizliğini gidermek amacıyla negatif haber sayısını artırmak gerekmektedir. Negatif haber sayısındaki artış, modelin hem pozitif hem de negatif haberler üzerinde dengeli sonuçlar vermesini ve genelleme yeteneğini geliştirmesini sağlar. Bu iyileştirme, sistemin daha doğru ve güvenilir duygu analizi sonuçları elde etmesine katkı sunar.

Bu çalışmada, haber metinlerinde firma isimlerinin doğru ve bağlamsal olarak tespit edilmesi amacıyla Spacy'nin Varlık Eşleştirme modelinin geliştirilmesi hedeflenmektedir. Spacy'nin mevcut adlandırılmış varlık tanıma modeli, firma isimlerini tespit etmekte güçlü bir araç sunmakla birlikte, bu isimlerin doğru firmalarla ilişkilendirilmesi için daha ileri bir eşleştirme modeline ihtiyaç duyulmaktadır. Bu modelin geliştirilmesi, aynı ismi taşıyan farklı firmalar arasında bağlamı analiz ederek doğru firmaya ait bilgilerin güvenilir bir şekilde çıkarılmasını sağlayacaktır. Böylece firma bazlı duygu analizinin doğruluğu artırılarak, bankaların firma tespiti süreçlerinde daha güvenilir ve ayrıntılı sonuçlar elde etmelerine katkıda bulunulacaktır.

İlerleyen çalışmalarda, Spacy NER modelinin banka sektörüne özel hazırlanmış bir ontoloji ile desteklenmesi de planlanmaktadır. Bu ontoloji, haber metinlerindeki varlıklar arasında anlamlı ilişkiler kurarak, sektörel bağlamda derinlemesine bilgi çıkarmayı mümkün kılacaktır. Firma isimleri ve olaylar arasındaki bağlamsal ilişki ağı, yalnızca yüzeysel bilgi çıkarmakla kalmayıp, sektöre özgü risk, finansal durum ve olaylar arasında stratejik analizler yapmayı sağlayacaktır. Ontoloji destekli Spacy NER modelinin bu entegrasyonu, haber içeriklerinden banka karar destek süreçlerine doğrudan aktarılabilecek nitelikli veri sunarak, analiz sürecine kapsamlı bir katkı sağlayacaktır.

Bu bütünsel yaklaşım, banka istihbaratında sektörel bilgiye dayalı olarak daha güvenilir ve derinlikli bir analiz imkanı tanıyacak, firma istihbaratı ve risk analizi süreçlerine stratejik bir destek sunacaktır.

Geliştirilen sistemin, bankadaki mevcut erken uyarı sistemleriyle entegre edilmesi hedeflenmektedir. Bu entegrasyon sayesinde bankalar, müşterileri ve potansiyel müşterileriyle ilgili risk ve fırsatları anlık olarak izleyebilecektir. Özellikle kredi risk yönetimi ve müşteri ilişkileri yönetiminde bankaların daha proaktif davranması mümkün hale gelecektir. Örneğin, negatif haberlerle ilişkilendirilen bir firma için anında devreye giren bir uyarı mekanizması, bankanın ilgili riskleri hızla değerlendirmesini sağlayacaktır.

Sistemin kapsamını genişletmek amacıyla Ticaret Sicil Gazetesi verilerinin entegrasyonu planlanmıştır. Firmalarla ilgili resmi bilgilere erişim sağlayan önemli bir kaynak olan Ticaret Sicil Gazetesi, firma kurulumları, adres değişiklikleri, ortaklık yapıları ve faaliyet alanlarındaki değişiklikler gibi güncel verilerin sisteme dahil edilmesine olanak tanıyacaktır. Bu bilgilerin otomatik olarak sisteme alınması, firma bazlı analizlerin ve duygu analizlerinin daha güvenilir sonuçlar sunmasını sağlayacaktır. Bu entegrasyon, bankaların firma risk analizlerini ve kredi değerlendirmelerini daha kesin ve güvenilir verilere dayanarak yapmalarına katkıda bulunacak; aynı zamanda firma tespitinde yaşanan belirsizlikleri azaltarak daha hassas ve güncel bir takip süreci sunacaktır.

5. KAYNAKLAR

Agarwal, J., Christa, S., Pai, A., Kumar, M. A., & Prasad, G. (2023). Machine learning application for news text classification. *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 463–466.

Ahbali, N., Liu, X., Nanda, A., Stark, J., Talukder, A., & Khandpur, R. P. (2022). Identifying corporate credit risk sentiments from financial news. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, 362–370.

Alrefaai, M. (2021). *Applying NLP Machine Learning for News Analysis and Classification*. Bahçeşehir University, Fen Bilimleri Enstitüsü.

Altinok, D. (2023). A diverse set of freely available linguistic resources for Turkish. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13739–13750.

Atan, S. (2018). Haberlerin Kurumsal İmaja Etkisi ve Türkiye'deki Hastaneler Hakkında Medyada Yer Alan Haberlerin Metin Madenciliği ile Analizi. *İletişim Kuram ve Araştırma Dergisi*, *0*(46).

Atan, S., & Çınar, Y. (2019). Borsa İstanbul'da Finansal Haberler ile Piyasa Değeri İlişkisinin Metin Madenciliği ve Duygu (Sentiment) Analizi ile İncelenmesi. *Ankara Üniversitesi SBF Dergisi*, *74*(1), 1–34.

Balcı, M. A., Akgüller, Ö., Batrancea, L. M., & Nichita, A. (2024). The Impact of Turkish Economic News on the Fractality of Borsa Istanbul: A Multidisciplinary Approach. *Fractal and Fractional*, *8*(1).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Boyapati, M., & Aygun, R. (2023). Phishing web page detection using web scraping. *SoutheastCon 2023*, 167–174.

Chen, K., Yang, Z., Wang, H., & Liu, L. (2011). Commonsense Knowledge Supported Intelligent News Analysis for Portfolio Risk Prediction. *2011 44th Hawaii International Conference on System Sciences*, 1–9.

Chung, S., Kim, J., Chi, S., & Kim, D. Y. (2023). Identifying the factors of country risk fluctuation from news text data using natural language processing. *Proceedings of the 2023 European Conference on Computing in Construction and the 40th International CIB W78 Conference*, *4*.

Demirtas, E., & Pechenizkiy, M. (2013). Cross-lingual polarity detection with machine translation. *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 1–8.

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What Can Natural Language Processing Do for Clinical Decision Support? *Journal of Biomedical Informatics*, *42*(5), 760–772.

Deng, L., & Liu, Y. (2018). *Deep Learning in Natural Language Processing*. Springer Nature Singapore Pte Ltd.

Durna, T. (2020). *Sivil Toplum Kuruluşları için Hak Temelli Gazetecilik Kılavuzu* (1st ed.). Um Vakfı Yayınları.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in Bioinformatics*, *15*(5), 788–797.

GLOBE. (2020). *The Empirical Use of KDTV Big Data in Academic Research (GA No. 822654)*.

Gökcan, H. T. (2015). Türk Ceza Kanunu Uygulamasında Kamu Görevlisi Kavramı. *Ceza Hukuku ve Kriminoloji Dergisi*, *3*(2).

Hayran, A., & Sert, M. (2017). Sentiment analysis on microblog data based on word embedding and fusion techniques. *2017 25th Signal Processing and Communications Applications Conference (SIU)*, 1–4.

Hersh, W. (2008). *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation MIT-Press*.

Hüseyin, B. (2024). *Makine Öğrenmesine Dayalı Hisse Senedi Değer Tahmini*. Yalova Üniversitesi, Lisansüstü Eğitim Enstitüsü.

Isaac, J. C., & Harikumar, S. (2016). Logistic regression within DBMS. *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 661–666.

Jackson, P., & Moulinier, I. (2007). *Natural language processing for online applications: Text retrieval, extraction, and categorization* (5th ed.). John Benjamins Publishing.

Jacobs, G., & Hoste, V. (2022). SENTiVENT: enabling supervised information extraction of company-specific events in economic and financial news. *Language Resources and Evaluation*, *56*(1), 225–257.

Jo, T. (2019). *Text Mining: Concepts, Implementation, and Big Data Challenge* (Vol. 45). Springer International Publishing.

Kılınç, M., Aydın, C., & Tarhan, Ç. (2022). KİTLE FONLAMASINDAKİ PROJE METİN İÇERİKLERİNİN LSTM İLE ANALİZİ. *Journal of Research in Business*, *7*(IMISC2021 Special Issue), 48–59.

Küçük, D., & Arıcı, N. (2018). DOĞAL DİL İŞLEMEDE DERİN ÖĞRENME UYGULAMALARI ÜZERİNE BİR LİTERATÜR ÇALIŞMASI. *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi*, *2*(2), 76–86.

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014). Event registry: learning about world events from news. *Proceedings of the 23rd International Conference on World Wide Web*, 107–110.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, *274*(2), 743–758.

Makeeva, E. Yu., & Sinilshchikova, M. (2020). News Sentiment in Bankruptcy Prediction Models: Evidence from Russian Retail Companies. *Journal of Corporate Finance Research / Корпоративные Финансы | ISSN: 2073-0438*.

Mertoğlu, U. (2020). *Türkçe İçin Sahte Haber Tespit Modelinin Oluşturulması*. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü.

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons.

Naing, I., Aung, S. T., Wai, K. H., & Funabiki, N. (2024). A Reference Paper Collection System Using Web Scraping. *Electronics*, *13*(14), 2700.

Nayoga, B. P., Adipradana, R., Suryadi, R., & Suhartono, D. (2021). Hoax Analyzer for Indonesian News Using Deep Learning Models. *Procedia Computer Science*, *179*, 704–712.

Nguyen, C. M., Thai, P. T., & Lam, D. K. (2023). A Real-Time Text Analysis System. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 340–345.

Nurrahmat, M. H., & Sunindyo, W. D. (2019). Determining External Factors Analysis Summary (EFAS) Metric for Company External Factors Using Online News Titles. *2019 5th International Conference on Science and Technology (ICST)*, *1*, 1–6.

Nyman, R., Kapadia, S., & Tuckett, D. (2021). News and narratives in financial systems: Exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, *127*, 104119.

Oğul, İ. Ü., Özcan, C., & Hakdağlı, Ö. (2017). Fast text classification with Naive Bayes method on Apache Spark. *2017 25th Signal Processing and Communications Applications Conference (SIU)*, 1–4.

Oktar, S., & Dalyancı, L. (2011). Türkiye ekonomisinde para politikası ve enflasyon arasındaki ilişkinin analizi. *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, *31*(2), 1–20.

Panagiotou, N., Saravanou, A., & Gunopulos, D. (2022). News Monitor: A Framework for Exploring News in Real-Time. *Data*, *7*(1).

Piccioni, C. A., Bastos, S. B., & Cajueiro, D. O. (2024). Stock Price Reaction to Environmental, Social, and Governance News: Evidence from Brazil and Financial Materiality. *Sustainability*, *16*(7).

Pillai, A. S., & Tedesco, R. (2023). *Machine learning and deep learning in natural language processing*. CRC Press.

Prabhat, A., & Khullar, V. (2017). Sentiment classification on big data using Naïve Bayes and logistic regression. *2017 International Conference on Computer Communication and Informatics (ICCCI)*, 1–5.

Qamar, U., & Raza, M. S. (2024). *Applied Text Mining*. Springer.

Sağlam, F. (2019). *Otomatik Duygu Sözlüğü Geliştirilmesi ve Haberlerin Duygu Analizi*. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü.

Seungwon, B., H, H. S., & Wooyong, J. (2023). Automated Identification of Active Players for International Construction Market Entry Using Natural Language Processing. *Journal of Management in Engineering*, *39*(5), 04023025.

Startseva, A., Vulfin, A., Vasilyev, V., Nikonov, A., & Kirillova, A. (2020). Analysis of financial payments text labels in the dynamic client profile construction. *2020 International Conference on Information Technology and Nanotechnology (ITNT)*, 1–10.

Stephenson, A., Reese, D., & Beadle, M. (2013). *Broadcast announcing worktext: A media performance guide*. Routledge.

Tosun, İ. B. (2021). *Makine Öğrenmesi İle Metin Sınıflandırma: Bakım Yönetim Sistemi Örneği*. Sakarya Üniversitesi, İşletme Enstitüsü.

Tu, M. (2024). Named entity recognition and emotional viewpoint monitoring in online news using artificial intelligence. *PeerJ Computer Science*, *10*.

URL-1. (2024). *Global Database of Events, Language, and Tone Project*.

URL-2. (2024). *General Inquirer*.

URL-3. (2024). *Introducing the Global Content Analysis Measures (KİAÖ)*.

URL-4. (2024). *Turkish Sentiment Analysis Dataset*.

Waldron, M. (2021). *RADAR, AYLIEN's New Risk Identification and Monitoring Solution, Wins 721 Deloitte Innovation Awards*. Aylien.

Wen, C., Wu, J., & Chen, D. (2022). Analysis of text emotion based on logistic regression model. *2022 IEEE 5th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, 891–895.

Yildirim, S. (2020). Comparing deep neural networks to traditional models for sentiment analysis in Turkish language. *Deep Learning-Based Approaches for Sentiment Analysis*, 311–319.

Zhang, F., Dvornek, N., Yang, J., Chapiro, J., & Duncan, J. (2020). Layer Embedding Analysis in Convolutional Neural Networks for Improved Probability Calibration and Classification. *IEEE Transactions on Medical Imaging*, *39*(11), 3331–3342.

Zheng, X., Li, L., & Zhang, W. (2021). An end-to-end hierarchical multi-task learning framework of sentiment analysis and key entity identification for online financial texts. *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 1245–1250.

Žižka, J., Dařena, F., & Svoboda, A. (2019). *Text mining with machine learning: principles and techniques*. Crc Press.

6. EKLER

Ek 1. Bankacılık Açısından Olumsuz Kelimeler Listesi

Olumsuz-Kelime	Olumsuz-Kelime-Skor
kara para	10
terör finansmanı	10
kaçakçılık	10
dolandırıcılık	10
sel	10
yangın	10
deprem	10
rüşvet	9
sahtekarlık	9
ceza	9
yolsuzluk	9
taciz	9
tazminat	9
sahtecilik	8
bataklık	8
çöküş	8
çöküntü	8
batış	8
batık	8
ihlal	8
resesyon	8
suistimal	8
istismar	8
hile	8
mahkumiyet	8
tahrip	7
korsanlık	7
iflas	7
soruşturma	7
çatışma	7
güvenilmezlik	7
temerrüt	7
kârsızlık	7
kriz	7
kırılganlık	7
riskli	7
borçlanma	7
krizde	7
rezil	7

Ek 1’in devamı

Olumsuz-Kelime	Olumsuz-Kelime-Skor
kapanış	7
devalüasyon	7
işsizlik	7
sarsıntı	7
enflasyon	7
şike	7
güvensizlik	7
uyuşmazlık	7
manipülasyon	7
kayıp	6
istifalar	6
düşüş	6
borç	6
istikrarsızlık	6
fiyasko	6
gerileme	6
istifa	6
sorun	6
zarar	6
ihmalkarlık	6
protesto	6
israf	6
hatalar	6
iddialar	6
tartışmalar	6
zayıflık	6
aksaklık	6
sorunlar	6
gecikme	6
skandal	5
skandallar	5
durgunluk	5
mahkeme	5
iddia	5
azalış	4
sıkıntı	4
yoksulluk	4
dava	4
kapatma	4

Ek 2. Bankacılık Açısından Olumlu Kelimeler Listesi

Olumlu-Kelimeler	Olumlu-Kelime-Skor
profesyonel	10
güvenilir	9
saygın	9
yetenekli	9
başarılı	9
etik	9
uzman	9
genişleme	9
liderlik	9
inovasyon	9
karlılık	9
ahenkli	9
dijitalleşmiş	9
tutumlu	9
eğitici	9
dayanışma	9
kaliteli	9
lider	9
işbirlikçi	9
bilgili	9
güçlü	9
zeki	9
yenilikçi	9
sorumlu	9
etkin	9
öncü	9
eğitimli	9
kararlı	9
başarı	9
ilerleme	9
rekabetçi	9
inanılır	9
çevreci	9
etkili	9
modern	9
içten	9
disiplinli	9
katılımcı	9
yetkilendirilmiş	9
bütünsel	9
özverili	9
işbirliği	9
eşitlik	9

Ek 2’nin devamı

Olumlu-Kelimeler	Olumlu-Kelime-Skor
Şeffaflık	9
adil	9
sürdürülebilir	9
akıllı	9
uyumlu	9
büyüme	9
stratejik	9
sezgisel	9
girişimci	9
özgün	9
şeffaf	9
tutkulu	9
bilinçli	9
saygılı	9
dürüst	9
huzurlu	9
yardımsever	9
borçsuz	9
tasarruf	9
inovatif	8
verimli	8
dinamik	8
deneyimli	8
gelişmiş	8
enerjik	8
iyimserlik	8
kârlı	8
esnek	8
tutarlı	8
hızlı	8
güzel	8
cömert	8
yaratıcı	8
gönüllülük	8
zenginlik	8
gelir	7
likidite	7
büyüyen	7
ilerleyen	7
gelişen	7
iyimser	7
zenginleşme	7
yükselme	7

Ek 2’nin devamı

Olumlu-Kelimeler	Olumlu-Kelime-Skor
Fon	7
analitik	7
etkileyici	7
yatırım	7
üretken	7
iyileşme	7
refah	7
fırsat	7
kazançlı	7
ödüllendirilen	7
verimlilik	7
iyileşen	7
istikrar	6
artış	6
sürdürülebilirlik	6
kârlılık	6
temettü	6
karlı	6
adalet	6
istikrarlı	6
artan	6
stabilite	6
önemli	6
kazanç	6

ÖZGEÇMİŞ

İlk ve ortaöğrenimini Atatürk İlköğretim Okulu'nda bitirdikten sonra lise eğitimini İstanbul Denizcilik ve Su Ürünleri Meslek Lisesi Bilgisayar Yazılım bölümünde tamamladı. 2010 yılında Giresim Üniversite, Tirebolu Mehmet Bayrak Meslek Yüksekokulu, Bilgisayar Teknolojileri ve Programcılığı Bölümü'nü okul birinci olarak bitirdi. 2011 yılında DGS ile girdiği KTÜ, Fen Fakültesi İstatistik ve Bilgisayar Bilimleri Bölümü'nden 2014 yılında mezun oldu. 2014 yılında KTÜ, Fen Bilimleri Enstitüsü, İstatistik ve Bilgisayar Bilimleri Bölümü'nde yüksek lisans eğitimine başladı ve 2017 yılında yüksek lisans eğitimini tamamladı. 2019 yılından itibaren bir teknoloji firmasının, yapay zeka uygulamaları geliştiren ekibinde kıdemli veri bilimci olarak çalışmaktadır.