T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DOĞAL DİL İŞLEME VE DERİN ÖĞRENME YÖNTEMLERİ KULLANILARAK FİNANSAL VERİLERİN ANALİZİ

MUSTAFA SAMİ KAÇAR

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Ocak-2024
KONYA

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Mustafa Sami KAÇAR

Tarih:

ÖZET

DOKTORA TEZİ

DOĞAL DİL İŞLEME VE DERİN ÖĞRENME YÖNTEMLERİ KULLANILARAK FİNANSAL VERİLERİN ANALİZİ

MUSTAFA SAMİ KAÇAR

Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Halife KODAZ

2024, 101 Sayfa

Jüri
Prof. Dr. Halife KODAZ
Prof. Dr. Harun UĞUZ
Prof. Dr. Mesut GÜNDÜZ
Dr. Öğr. Üyesi Onur İNAN
Dr. Öğr. Üyesi Murat KARAKOYUN

Son yıllarda, dünyadaki hemen her alanda dijital veri üretiminin her geçen gün büyük bir hızla artması, karar alma aşamasındaki kullanımını önemli ölçüde artmıştır. Bu rüzgâr, finans alanında da kendini ciddi bir şekilde göstermiştir. Ancak, geleneksel istatistiksel yöntemler, ham verinin kontrolsüz genişlemesi ve karmaşıklığı nedeniyle artık işlevini günden güne yitirmektedir. Bu nedenle, finansal verilerin temizlenmesi ve analiz edilmesi için modern makine öğrenimi yöntemlerinin kullanılması son derece önemlidir. Bu tez çalışmasında, şirketlerin paylaşmış oldukları yıllık ve dönemsel finansal bilgilerin yer aldığı raporlardan yeni veri setleri üreten ve üretilen setleri makine öğrenmesi yöntemleriyle analiz eden yenilikçi yaklaşımlar sunulmuştur. Çalışma kapsamında, finansal 10K yıllık raporları toplanarak analiz edilebilir veri setlerine dönüştürülmüş ve makine öğrenmesi yöntemleriyle sınıflandırma işlemi gerçekleştirilmiştir. Elde edilen başarılı sonuçlarla (%92 doğruluk değeri), 10K raporlarının veri setine dönüşümü için önemli bir katkı sağlanmıştır. Doğal dil işleme tekniklerinin oldukça karmaşık ve hatalı veriler içeren 10K raporlarına uygulanması da yine tez kapsamında gerçekleştirilerek, benzersiz yeni yaklaşımlar sunulmuştur. 10Q çeyrek raporlarının analizini mümkün kılan bir hibrit yöntem, tez kapsamında gerçekleştirilen çalışmalarla üretilmiştir. Yöntemle, şirketlerin metinsel içeriğe sahip çeyrek raporları, Doc2Vec ve K Means kümeleme algoritmaları kullanılarak verimli veri setlerine dönüştürülmüştür. Şirketlerin sonraki finansal çeyrekteki fiyat güçlerini gösteren ve bir ile on arasında değerler alan 'PriceRank' metriği, düşük, orta ve yüksek olacak şekilde üçe ayrılarak, veri setine sınıf değerleri olarak eklenmiştir. Daha sonra, derin öğrenme yöntemi olan Evrişimsel Sinir Ağı ile gerçekleştirilen sınıflandırma işlemiyle başarılı sonuçlar (%84 doğruluk değeri) üretilmiştir. Son olarak, Doc2Vec ve K Means algoritmalarıyla üretilen veri setine, veri üzerindeki hem kısa hem de uzun vadeli bağlantıları daha iyi ortaya çıkarabilmek ve özellik çıkarımı adımını iyileştirmek için Tek Boyutlu Evrişimsel Sinir Ağı ve Uzun Kısa Süreli Bellek algoritmalarından meydana gelen hibrit bir yöntem uygulanmıştır. Elde edilen başarılı sonuçlar (%88 doğruluk değeri), bu yeni hibrit yöntemin, veri analizinde geleneksel derin öğrenmesi yöntemlerinden daha iyi sonuçlar üretebileceğini göstermiştir.

Anahtar Kelimeler: Derin Öğrenme, Doğal Dil İşleme, Finans, Veri Madenciliği, XBRL

ABSTRACT

PhD THESIS

ANALYSIS OF FINANCIAL DATA USING NATURAL LANGUAGE PROCESSING AND DEEP LEARNING METHODS

MUSTAFA SAMİ KAÇAR

Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering

Advisor: Prof. Dr. Halife KODAZ

2024, 101 Pages

Jury
Prof. Dr. Halife KODAZ
Prof. Dr. Harun UĞUZ
Prof. Dr. Mesut GÜNDÜZ
Asst. Prof. Dr. Üyesi Onur İNAN
Asst. Prof. Dr. Murat KARAKOYUN

In recent years, with the rapid increase in digital data production at worldwide, its usage for decision-making has significantly grown. This trend has also made its mark in the finance. However, traditional statistical methods are no longer effective due to the uncontrolled expansion and complexity of raw data. Therefore, it is crucial to employ modern machine learning methods for cleansing and analyzing financial data. This thesis presents innovative approaches that generate new datasets from annual and periodic financial reports shared by companies and analyze these generated datasets using machine learning techniques. Within the scope of the study, 10K annual reports were collected, transformed into analyzable datasets, and subjected to classification processes using machine learning methods. The successful results obtained (92% accuracy) contribute significantly to the transformation of 10K reports into a dataset. The application of natural language processing techniques to complex and error-prone 10K reports was also performed within the thesis, presenting unique approaches. A hybrid method that enables the analysis of 10Q quarter reports was produced by the studies carried out within the scope of the thesis. With the method, companies' quarterly reports with textual content were transformed into efficient data sets using Doc2Vec and K Means Clustering algorithms. The 'PriceRank' metric, which shows the price power of companies in the next financial quarter and takes values between one and ten, was divided into three groups as low, medium, and high and added to the data set as class values. Later, successful results (84% accuracy) were produced by the classification process with the deep learning method, Convolutional Neural Network. Finally, a hybrid method consisting of 1D Convolutional Neural Network and Long Short-Term Memory algorithms was applied to the data set produced with Doc2Vec and K Means algorithms to better reveal both short- and long-term connections on the data and to improve the feature extraction step. The successful results obtained (88% accuracy) showed that this new hybrid method can produce better results than traditional deep learning methods in data analysis.

Keywords: Data Mining, Deep Learning, Finance, Natural Language Processing, XBRL

ÖNSÖZ

Doktora tez çalışmam sürecinde çok değerli yönlendirmelerine ve desteklerine mazhar olduğum, yardımını asla esirgemeyen ve bana karşı her zaman anlayışlı olan danışmanım Sayın Prof. Dr. Halife KODAZ' a ve ikinci danışmanım Dr. Öğr. Üyesi Semih YUMUŞAK' a, görüşmelerimizde şahsımın ve tez çalışmamın gelişimi için doğru sorularına ve yönlendirmelerine şahit olduğum tez izleme komitesi üyeleri Prof. Dr. Harun UĞUZ' a ve Dr. Öğr. Üyesi Onur İNAN' a, birlikte çalıştığımız Arş. Gör. Emir Ali DİNSEL' e, tüm doktora sürecimde, ilgilerini ve desteklerini gördüğüm, mesleki ve akademik çok önemli bilgiler öğrendiğim Konya Teknik Üniversitesi öğretim elemanlarına, her zaman ve her şartta yanımda olan, değerli eşim Melek ve aileme çok teşekkür ederim.

Mustafa Sami KAÇAR
KONYA-2024

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Kısaltmalar

XBRL	:	eXtensible Business Reporting Language
S&P	:	Standard and Poor’s
SEC	:	Security and Exchange Commision
EDGAR	:	Electronic Data Gathering, Analysis and Retrieval
CNN	:	Convolutinoal Neural Network
HTML	:	Hyper Text Markup Language
CIK	:	Central Index Key
XML	:	Extensible Markup Language
FASB	:	Financial Accounting Standards Board
GAAP	:	Generally Accepted Accounting Principles
LSTM	:	Long-Short Term Memory

ŞEKİLLER LİSTESİ

ÇİZELGELER LİSTESİ

1. GİRİŞ

Finansal verilerin analizi, işletmelerin performansını değerlendirmek, riskleri belirlemek ve gelecekteki eğilimleri tahmin etmek için kritik bir öneme sahiptir. Geleneksel yöntemlerle finans verilerinin analizi genellikle sayısal verilerin istatistiksel yöntemlerle incelenmesini içerirken, son yıllarda doğal dil işleme ve derin öğrenme tekniklerinin kullanımı bu alanda yeni bir bakış açısı sağlamıştır. Doğal dil işleme, metin tabanlı verileri anlamak ve çıkarılan bilgileri kullanmak için kullanılan bir dizi teknik ve algoritmayı içerirken, derin öğrenme ise karmaşık desenleri otomatik olarak öğrenmek için çok katmanlı sinir ağlarını kullanmaktadır. Bu tez çalışmasında doğal dil işleme ve derin öğrenme yöntemleri kullanılarak finansal verilerden daha önce benzerine rastlanmamış veri setleri üretilmiş ve üretilen veri setleri makine öğrenmesi yöntemleriyle analiz edilmiştir.

Tez çalışmasının amacı, doğal dil işleme gibi makine öğrenmesi yöntemlerinden faydalanarak, metin tabanlı finansal raporlardan makinelerle analiz edilebilir, anlamlı veri setleri oluşturulması ve finansal performansı değerlendirme, riskleri tahmin etme ya da gelecekteki eğilimleri belirleme gibi birçok finansal çıkarımın bu veri setlerinin analizi ile gerçekleştirilmesidir. Bu çalışma, geleneksel finansal analiz yöntemlerinin sınırlamalarını aşmak ve daha kapsamlı bir anlayış sağlamak amacıyla yeni ve yenilikçi bir yaklaşım sunmaktadır.

Tez çalışması için temel yönelimlerden birisi finansal raporlarda yaygın olarak kullanılan XBRL (eXtensible Business Reporting Language) standardının kullanılmasıdır. XBRL, finansal verileri yapılandırılmış bir formatta sunarak veri toplama ve işleme sürecini kolaylaştırmaktadır. Bu çalışmada, XBRL standartlarına göre hazırlanmış dosyalardaki bilgileri kullanarak finansal verilerin alınması, ön işleme adımlarının gerçekleştirilmesi ve doğal dil işleme yöntemlerinin uygulanmasıyla veri setlerinin oluşturulması sağlanmıştır.

Tez çalışmasının bir diğer önemli bileşeni, analizi gerçekleştirilen veriler üzerinde yine doğal dil işleme ve dahası derin öğrenme tekniklerinin kullanılmasıdır. Bu tekniklerin kullanımıyla, finansal metinlerin anlaşılması, duyarlılık analizi, veri özeti, metin sınıflandırması, gelecek tahmini gibi işlemlerin gerçekleştirilmesi hedeflenmiştir.

Dünyada bu alandaki çalışmalara ilginin oldukça arttığı günümüzde, tez kapsamında yürütülen çalışmaların literatüre önemli bir katkı sağlayacağı düşünülmektedir. Özellikle, finansal verilerden elde edilen, makine öğrenmesi

yöntemleriyle analiz edilebilir veri setlerinin eksikliği göze çarpmaktadır. Bu yüzden, çalışmada önerilen veri setlerinin, bu alanda yapılacak sonraki çalışmalar için de önemli bir materyal olacağı düşünülmektedir.

Gerçekleştirilen tüm çalışmalar dört ana bölümün alt bileşenlerini oluşturmuştur. Çalışmanın birinci kısmında, Amerika' da faaliyet gösteren ve Amerikan borsalarında listelenen şirketlerin yıllık tüm faaliyetlerinin yer aldığı finansal 10K raporları toplanmış, önerilen yöntemlerle veri setlerine dönüştürülerek makine öğrenmesi yöntemleriyle analiz edilmiştir. Doğal Dil İşleme teknikleri, finansal raporlara uygulanarak, verilerdeki bağlamların çıkarılması, veri özetlerinin oluşturulması, veri analizlerindeki verimliliğin artırılması hedeflerine uygun tekniklerin geliştirilmesi çalışmanın ikinci kısmını oluşturmuştur. Üçüncü kısımda, çeyrek mali dönemlik 10Q finansal raporları anlamlı veri setlerine dönüştürülmesi ve üretilen veri setleri üzerinde derin öğrenme sürecinin gerçekleştirilmesi için hibrit bir sistem önerilmiştir. Son olarak, üçüncü kısımda üretilen veri setine, finansal verilerdeki uzun ve kısa vadeli bağlantıları daha iyi yakalayabilmek için iki ayrı derin öğrenme algoritması birleştirilerek üretilen yeni bir hibrit yöntem uygulanmıştır.

10K raporlarının toplanması ve indirilebilmesi için bir web arama motoru (web crawler) geliştirilmiştir. Geliştirilen yöntemle veriler elde edilirken anlamsız ve gereksiz birçok veriden arındırılmıştır. Ancak, verilerin yapısı oldukça karmaşık ve problemli olduğu için, XBRL formatlı 10K raporlarına özel birçok veri ön işleme ve temizleme yöntemi geliştirilmiştir. Ardından verilerin düzenlenerek veri setlerine dönüştürülmesi için önerilen metotlarla makine öğrenmesi yöntemlerinde çalıştırılabilecek formatlı veriler elde edilmiştir. Son olarak, elde edilen veri setleri makine öğrenmesi yöntemlerinde çalıştırılarak önerilen yöntemlerden elde edilen veri setleri testlere tabi tutulmuştur. Sonuç çıktısı olarak, firmaların yıllık raporlarından Amerikan borsalarındaki en değerli şirketlerin endeksi olan S&P 500 endeksinde olup olmadıkları yüksek oranda doğruluk (accuracy) değeri (%92) ile tahmin edilmiştir. Elde edilen sonuçlarla, önerilen yöntemlerin finansal raporların analizinde önemli bir katkı sağlayacağı, yeni çalışmalara ön ayak olabileceği, farklı bakış açıları getirebileceği kanıtlanmıştır.

Tez kapsamında yapılan doğal dil işleme çalışmalarının temel amacı, finansal raporlarda yer alan metinsel ifadelerin özgün, sade, gerçek anlamını yitirmeden, olabildiğince genel ve anlaşılır normlarda sunulması için yeni teknikler geliştirmek olmuştur. Bu bağlamda, finansal raporlarda yer alan metinsel ifadeler, doğal dil işleme tekniklerine ve çalışmada önerilen diğer yöntemlerle üretilen veri setleri, bu çalışma için

ve bu alanda yapılacak sonraki çalışmalar için önemli bir materyal olarak sunulmuştur. Veri karmaşıklığının, veri seyrekliğinin, hatalı verilerin oldukça fazla olduğu böylesi raporlarda yapılan çalışmalarla üretilen sonuçların ve önerilen yöntemlerin literatüre önemli katkı sağlayacağına inanılmaktadır. Finansal raporların temel problemlerinden biri de genel standartlara sahip olması gerekirken, paylaşıldığı firma özelinde oldukça fazla özgünleşmesidir. Bu durum, böylesi raporların makine öğrenmesi yöntemleri ile analizini ciddi derecede zorlaştırmaktadır. Bu yüzden, farklı firmaların paylaştığı verilerdeki, finansal kalemlerdeki benzerliklerin, aynı anlama gelebilecek açıklamaların eşleştirilmesi ve sadeleştirilmesi önem arz etmektedir. Böylece, gerçekleştirilen doğal dil işleme çalışmalarının yoğunluğu, firma paylaşımlarındaki benzerliklerin, ortak noktaların ortaya çıkarılması üzerine yapılmıştır.

XBRL formatlı dosyalarda veriler, etiket ve etiketin taşıdığı değer biçiminde ikili olarak saklanmaktadır. Örneğin, bir firma eğer 100 dolar vergi ödemişse, bu raporlarda 'ÖdenenVergiler' ismiyle ve '100' değeriyle sunulmaktadır. Yapılan çalışmalarda, vergi ödemesi gibi mali kalemleri farklı firmaların farklı isimlerde sunduğu, 'VergilerÖdenmiş', 'VergiKalemi', 'ÖdenenVergiler' gibi tespit edilmiştir. Ortak etiketlerin tespiti, etiketlerdeki ortak kelimelerin tespiti, birbirini içeren etiketler, aynı anlama gelen kelimeleri kullanan etiketler, kelime köklerinin çıkarılması gibi yöntemlerle farklı firmaların paylaştığı farklı etiketlerdeki bağlantılar ortaya çıkarılmıştır. Ayrıca, metin değerli etiketler, uygulanan yöntemlerle sayısallaştırılarak da ham verilerden veri seti üretiminin yolu açılmıştır.

Tez çalışması kapsamında gerçekleştirilen faaliyetlerden biri de 10Q çeyrek dönem raporları üzerinedir. 10Q raporları, Amerika'da borsalarda işlem gören tüm firmaların, Menkul Kıymetler ve Borsa Komisyonu'na (Security and Exchange Commision, SEC) her yılın dört aylık periyotlarının sonunda paylaşmakla yükümlü oldukları bir rapordur. Bu raporların da XBRL formatlı versiyonları SEC tarafından yönetilen Elektronik Veri Toplama Analiz ve Geri Alma (Electronic Data Gathering, Analysis and Retrieval, EDGAR) veri tabanında sunulmaktadır. Yatırımcılar, analistler, finansal denetçiler bu raporları inceleyerek sonuç çıkarımı yapmaktadırlar. Tez çalışması kapsamında üretilen hibrit yöntemle, firmaların çeyrek raporlarının anlamlarını yitirmeden birleştirilerek bir veri setine dönüştürülmesi ve derin öğrenme yöntemlerinden Evrişimsel Sinir Ağı (CNN) (LeCun ve ark., 1998) ile analizi gerçekleştirilmiştir. Hibrit algoritma iki aşamadan oluşmaktadır. İlk aşaması, verilerin vektöre dönüştürülmesi ve bağlamsal ve semantik olarak birleşiminden veri seti üretilmesi üzerinedir. İkinci

aşamasında ise üretilen veri seti derin öğrenme ağında çalıştırılarak firmaların çeyrek raporlarından fiyat gücünün tespit edilmesi olmuştur. Elde edilen sonuçlar, diğer yöntemlerle kıyaslanarak bölüm 4.3' te sunulmuştur. Sonuçlara göre, önerilen hibrit yöntemin, tüm aşamalarında önemli derecede daha iyi sonuçlar ürettiği tespit edilmiştir.

Sıralı verilerdeki desenleri ve ilişkileri ortaya çıkarma konusunda başarılı bir yöntem olan Tek Boyutlu Evrişimsel Sinir Ağı (1D Convolutional Neural Network, 1D-CNN) (Kim, 2014) yöntemi ve özellikle veri üzerindeki dolaylı bağlantıları yakalama konusunda güçlü bir algoritma olan Uzun Kısa Süreli Bellek (Long Short Term Memroy, LSTM) (Hochreiter ve Schmidhuber, 1997) yöntemi ile yeni bir hibrit yöntem önerilmiştir. Genellikle veri karmaşıklığının ve veri seyrekliğinin yüksek olduğu finansal veriler için, bilhassa özellik çıkarımı adımının iyileştirmesiyle, önerilen hibrit yöntemin geleneksel derin öğrenmesi yöntemlerine kıyasla daha başarılı sonuçlar üreteceği düşünülmüştür. Gerçekleştirilen testlerle, bu hibrit yöntemin evrişimsel sinir ağına ya da diğer makine öğrenmesi yöntemlerine oranla daha iyi sonuçlar ürettiği gösterilmiştir. Detaylı sonuçlar, Bölüm 4.4' te verilmiştir.

1.1. Tezin Organizasyonu

Tez çalışması 5 bölümden oluşmaktadır. Tüm bölümler, aşağıda açıklandığı üzere organize edilmiştir.

İlk bölümde, tez çalışmasının ana fikri, yapılan çalışmalar ve literatüre katkıları hakkında giriş yapılarak, tezin genel çerçevesi açıklanmıştır. İkinci bölümde de tez çalışmasında konu olan veriler, çalışma alanı ve tezde kullanılan yöntemlere yönelik literatür araştırmaları yer almaktadır. Finansal verilerin analizinde yaklaşım tarzları, bu alandaki gelecek öngörüleri, makine öğrenmesi ve finans konularının birleştirildiği çalışmalar, tez konusunun ve sonrasında yapılan çalışmaların gerekçelendirilmesiyle ilgili bölümde sunulmuştur.

Üçüncü bölümde, tez kapsamında yapılan tüm çalışmalar ayrıntıları ile sunulmuştur. Tezde kullanılan verilerin elde edilmesinden bilgiye dönüşüm sürecine kadar tüm evreleri belirtilmiştir. 10K ve 10Q raporları, XBRL formatlı dosya yapısı, EDGAR veri tabanı ve kullanımı açıklanmıştır. Çalışmalarda kullanılan yöntemlerin ayrıntıları, ana mimari temelleri, tüm aşamaları, şekil ve çizelgelerle desteklenerek tüm açıklığı ile verilmiştir.

Dördüncü bölümde, gerçekleştirilen çalışmaların sonuç çıktıları yer almıştır. Önerilen yöntemlerin elde ettiği başarımlar, literatüre sağlayacağı katkılar, diğer yöntemlerden ayıran, öne çıkan özellikleri, kabul görmüş derecelendirme yöntemleri ile elde edilen sonuç değerleriyle destekli bir biçimde sunulmuştur. Sonuçların ne anlam ifade ettiği, değeri ve önemi de yine yer verilen değerlendirmeler ile bu bölümde yer almaktadır. Son bölüm olan beşinci bölümde de tez çalışmasında varılan sonucun özeti ve gelecekte ne tür çalışmalara ön ayak olabileceği aktarılmıştır.

2. KAYNAK ARAŞTIRMASI

Bu tez çalışmasının amaçlarına ulaşabilmesi için literatürdeki ilgili çalışmalar incelenmiştir. Finansal verilerin yer aldığı, analizi için doğal dil işleme ve derin öğrenme yöntemlerinin uygulandığı çalışmalar incelenerek bu bölümde sunulmuştur. Bu tezin konu seçimi, ilerleme süreci ve kullanılan yöntemlerin seçimleri yapılan çalışmalarla desteklenerek gerekçelendirilmiştir. Bu çalışmada, finansal verilerin analizinde yeni bir perspektif sunmak ve ilgili araştırmalara temel oluşturmak hedeflenmiştir. Bu hedef doğrultusunda da yapılan tüm çalışmaların literatür ile bağlantısına bu bölümde yer verilmiştir. Tez kapsamında yapılan çalışmalar dört ana bölümden oluştuğu için kaynak araştırması bölümü de çalışılan alana dair genel bakış açılarının sunulduğu, geleceğinin tartışıldığı, potansiyel çalışma konularının verildiği inceleme makalelerinin yer aldığı ilk başlıkla birlikte beş alt başlıkla sunulmuştur. İkinci başlıkta, 10K raporlarına yönelik çalışmalar ve analizinde makine öğrenmesi yöntemlerinin kullanıldığı çalışmalar incelenmiştir. Üçüncü başlıkta finansal veriler üzerinde gerçekleştirilen doğal dil işleme çalışmaları yer almıştır. Dördüncü başlıkta ise 10Q raporlarının analiz edildiği çalışmalar, bu çalışmada önerilen hibrit yöntemin bileşenlerini oluşturan algoritmaların finans alanında karşılık bulduğu çalışmalar yer almıştır. Son başlıkta ise, iki derin öğrenmesi yöntemiyle üretilen hibrit yöntemin uygulandığı diğer çalışmalara, çalışma alanlarıyla birlikte yer verilmiştir.

2.1. Genel Bakış

Bu başlıkta özellikle inceleme çalışmaları derlenmiştir. Finansal sektör ile bilgisayar bilimlerinin birleştiği, sektörün yapay zekâ alanına evrimi ve potansiyel geleceğine dair ipuçları içeren analizler ve öngörülere yer verilmiştir.

Brown ve arkadaşları çalışmalarında (Brown-Liburd ve ark., 2015) büyük verinin denetim ortamında kullanımını ele almıştır. Çalışmada, bilgi işleme zayıflıkları ve sınırlamaları göz önünde bulundurarak, büyük verinin denetim kararları üzerindeki davranışsal etkilerini psikoloji ve denetim literatürüne dayanarak incelenmiştir. Bilgi yüklenmesi, bilgi ilgisi, desen tanıma ve belirsizlik gibi konular üzerinde durulmuştur. Ayrıca, denetçilerin büyük veriyi entegre etme zorlukları ve şirketlerin büyük veri analizinde kullandığı analitik araçlar hakkında bilgi verilmiştir. Yazarlar, gelecekteki araştırmaların ele alabileceği konuları da tartışarak, büyük verinin denetimde kullanımıyla ilgili önemli soruları ortaya koyarak çalışmalarını tamamlamıştır. Alles ve

Gray’ in yazmış olduğu makalede (Alles ve Gray, 2016) büyük veri' ye olan yatırımların hızla arttığı ve muhasebe firmalarının da büyük veriyi denetim süreçlerinde önemli bir unsura dönüştürmeye çalıştığı belirtilmektedir. Ancak, büyük verinin denetim alanında diğer alanlardaki kadar net bir uygulama alanı bulunmadığı ifade edilmektedir. Yazarlar, yazmış oldukları makalenin amacının büyük veri kavramının finansal tablo denetimlerine entegrasyonunu engelleyen faktörleri tartışmak ve bu engellerin giderilmesi için araştırma yaklaşımlarını sunmak olduğunu belirtmiştir. Büyük veri ve analitiğinin denetime uygulanması konusunun hem araştırmada hem de uygulamada büyük ilgi gördüğüne Kriger ve Drews’ in çalışmasında (Krieger ve Drews, 2018) değinilmiştir. Bu konuda yapılan birçok kullanım senaryosu ve literatür incelemesi bulunmasına rağmen, kullanım senaryolarını sınıflandırmak için kullanılan kategorilerin hala parçalanmış durumda olduğu sonucuna varmışlardır. Gerçekleştirdikleri çalışmada, sistematik bir taksonomi geliştirme süreci kullanılarak, büyük veri ve analitik kullanım senaryolarını yapılandırmak için bir taksonomi oluşturulmuştur. Bu taksonomi, denetimde büyük veri ve analitiğin kullanımını yapılandırılmış bir şekilde sınıflandırmaya yardımcı olan boyutlar ve özellikler sunmaktadır. Earley çalışmasında (Earley, 2015) büyük verinin yükselişi ve veri analitiği alanının muhasebe sektöründe önemli bir konu haline geldiği belirtmiştir. Vergi ve danışmanlık alanlarında veri analitiğinin yaygın olarak kullanıldığı, ancak denetim alanında benimsenmesinin diğer alanlara göre daha yavaş olduğu ifade edilmektedir. Denetimin benzersiz zorluklarının, veri analitiğinin denetimde kullanımını engellediğine değinilmektedir. Bununla birlikte, muhasebe firmalarının denetim odaklı veri analitiği geliştirmek için önemli yatırımlar yaptığı ve bu çabaların dönüştürücü etkisini yakında görülmeye başlanabileceği vurgulanmaktadır. Makalenin amacı, veri analitiğinin finansal tablo denetimlerine nasıl uygulanabileceğini açıklamak ve araştırmacılara bu alanda çözülmesi gereken sorunlar hakkında bir perspektif sunmaktır.

Şirketlerin gönüllü olarak çeşitli finansal bilgileri internet üzerinden açıklama imkanının güçlü bir araç olduğu ve bu konuda geniş bir akademik literatür bulunduğu Bonson’ un yaptığı çalışma (Bonson ve Escobar, 2002) sonucu ortaya çıkarılmıştır. Araştırmalar, açıklanan finansal bilgilerin muhasebe düzenlemeleri tarafından normalde gereken bilgilerden daha geniş kapsamlı olduğunu göstermiştir. Makalede, farklı Avrupa ülkelerindeki önde gelen şirketler tarafından şu anda internet üzerinde sağlanan bilgiler karşılaştırmalı bir analiz yapmak amacıyla incelenmiştir. Bu amaçla, Avrupa Birliği ülkelerinin en büyük 20 şirketinden (piyasa değeri bazında) veriler toplanmıştır. Ardından, şirketlerin şeffaflığı (bağımlı değişken) ile sektörleri, ülke kökenleri ve

büyüklükleri (bağımsız değişkenler) arasındaki ilişkileri belirlemek için istatistiksel testler yapılmıştır. Sonuçlar, bu üç değişken ile internet üzerindeki gönüllü açıklama (şeffaflık) arasında istatistiksel olarak anlamlı bir ilişki olduğunu göstermiştir.

Kang ve arkadaşları çalışmalarında yatırımcıların önemli yatırım kararları verirken şirketlerin zorunlu olarak her mali yıl sonunda açıkladığı yıllık raporlardan faydalandığını belirtmektedir (Kang ve ark., 2018). Yazarlar, Amerika Menkul Kıymetler ve Borsa Komisyonu, şeffaf ve anlaşılabilir bir yıllık rapor oluşturmak için Düz İngilizce Kuralı' nı (Plain English Rule) uyguladığını vurgulamışlardır. Ancak çalışmada, yıllık raporların açıklanmasıyla ilgili genel bir kılavuzun mevcut olup, hacim ve format konusunda spesifik düzenlemeler bulunmadığına dikkat çekilmiştir. Bu nedenle, iş yöneticilerinin olumlu ifadelerini yorumlamanın riskli olduğu çünkü verilerin, iş yöneticilerinin subjektif görüşlerini ve şirketlerin bakış açılarını içerebildiği gerçeğine değinilmiştir. Yazarlara göre çalışmanın metodolojisi metin madenciliğine dayanmaktadır. Bu ilişkileri analiz etmek için, ABD'deki tüm halka açık şirketlerin 10K raporlarının toplanması ve metin madenciliği yöntemi kullanarak bu 10K anlatılarının tonlarının belirlenmesi ile mevcut kazanç seviyelerine uyumlu bir şekilde değişip değişmediği tespit edilmiştir. Ayrıca, raporlar arasında ton esnekliğine yol açabilecek faktörlerin araştırılmasıyla mevcut performanslarına göre tonu daha olumlu olan şirketler incelenmiş ve gelecekteki performansın mevcut performanstan nasıl farklı olabileceği belirlenmiştir.

Gunn çalışmasında (Gunn, 2007), XBRL'in kullanımının hem ulusal hem de uluslararası düzeyde sağladığı faydaları ve fırsatları açıklamıştır. Ayrıca, XBRL'in yaygın olarak kabul görmesi ve benimsenmesi için karşılaşılabilecek zorluklara da değinmiştir. Aynı zamanda çalışma, muhasebecilerin (işletme yönetimi, kamu hizmeti veya akademide yer alanlar) XBRL'in ne anlama geldiğini ve finansal bilgi kullanıcıları için nasıl bir etki yarattığını tartışmaktadır. XBRL finansal raporlama alanında önemli bir unsura dönüştüğü gerçeği Vasarhelyi ve arkadaşları tarafından yazılan makalede (Vasarhelyi ve ark., 2012) vurgulanmıştır. Çalışmada, XBRL'nin raporlanan finansal bilgilerin kullanışlılığı üzerindeki etkileri tartışılmaktadır. Teknoloji kabul modeli (Technology Acceptance Model, TAM) (Davis, 1989) teorik çerçevesini kullanarak, XBRL'nin raporlanan finansal bilgilerin kullanışlılığını geliştirmek için potansiyeli incelenmiştir. XBRL'nin daha geniş kabul görmesiyle birlikte, XBRL standardizasyonunun yayılacağı beş eksen üzerinde durulmaktadır: mevcut veri, açıklama formatı, tarihsel veri, veri doğruluğu/güvencesi ve üçüncü taraf verisi. Ayrıca, XBRL ve

ilgili teknolojilerin olgunlaşmasıyla ortaya çıkabilecek yeni araştırma fırsatlarına da dikkat çekilmiştir. Efendi ve arkadaşları XBRL raporlama formatının, önceden HTML formatında sunulan aynı 10K/10Q bildirimlerinin ötesinde artı bir bilgi değeri sağlayıp sağlamadığını araştırmıştır (Efendi ve ark., 2016). XBRL Gönüllü Bildirim Programı'ndan bir örnek kullanarak, gönüllü XBRL raporlarının bildirildiği gün hisse fiyatı varyansında önemli bir artış olduğu çalışmada kanıtlanmıştır. İçeriğin aynı gün içinde daha fazla bildirildiği durumlarda piyasa tepkisi daha güçlü olduğu sonucuna varılmıştır. Göreceli bilgi değerini değerlendirmek için, kazançlarla ilgili üç temel haber duyurusu için üç ayrı dönem getiri varyansını değerlendirilmiştir; kazançlar duyurusu, HTML bildirimi ve XBRL bildirimi. XBRL bildirimlerinin, HTML bildirimlerinden daha büyük bir göreceli bilgi değerine sahip olduğunu sonucuna vararak, XBRL raporlama formatının artı bir bilgi içeriği sağladığını göstermişlerdir.

2.2. 10K Verilerinin Analizi

Finansal veriler, tablo veri yapısıyla birlikte birçok farklı formatta temsil edilebilir (ör. XBRL, pdf, HTML, xlsx). Dijitalleştirilmiş finansal raporlar, araştırmacıların şirketler, sektörler ve piyasalar arasında finansal analiz yapmalarını sağlamaktadır. Bununla birlikte, genellikle finansal raporları EDGAR veritabanı üzerinden toplamak ve analizde kullanmak çok yaygın değildir. Genellikle, dünya genelinde faaliyet gösteren küresel şirketlerin istatistiksel, finansal ve piyasa bilgilerini depolayan COMPUSTAT gibi veritabanları kullanılır. Örneğin, Chychyla ve Kogan (Chychyla ve Kogan, 2015), 10K verilerini COMPUSTAT ile karşılaştırmıştır. 5000 farklı şirketin 30 farklı muhasebe kalemi karşılaştırılmış ve analiz edilmiştir. Şirketler arasında paylaşılan bilgi açısından önemli farklılıklar olduğu ortaya çıkarılmıştır. Finansal raporlardaki veri işleme verimliliği, Rao ve Guo (Rao ve Guo, 2022) tarafından yaptıkları çalışmada hesaplanmıştır. Bunun için EDGAR ve COMPUSTAT veritabanlarını karşılaştırmışlardır. Şirketin boyutu, yaşı ve endüstrisi gibi bilgilerin veri paylaşımında etkili olmadığı sonucuna varmışlardır. Farklı bir çalışmada, Cunningham ve Leidner (Cunningham ve Leidner, 2019), yatırımcılar için değerli bilgiler sağlamak amacıyla SEC'de mevcut olan verilerin kalitesini artırmak üzere çalışmışlardır. Bu bağlamda, kaliteyi düşüren bazı eksiklikler belirlenmiştir. Muhasebe ve finansta kullanılan nicel analizlerin yanı sıra, Loughran ve McDonald (Loughran ve Mcdonald, 2016) tarafından metin analizinin etkinliği araştırılmıştır. Farklı bir yaklaşım olarak, Hoitash ve Hoitash

(Hoitash ve Hoitash, 2018), XBRL etiketleri üzerinde muhasebe raporlama karmaşıklığı kriteri önermişlerdir. Bu kriter, verinin güvenilirliğini etkileyen hataları, eksiklikleri ve uygulamadaki yanlışlıkları belirlemeyi amaçlamıştır. Peterson ve arkadaşları (Peterson ve ark., 2015), 10K dosyalarında açıklanan muhasebe politikaları dipnotlarındaki metin benzerliğini analiz ederek, aynı firmada zaman içinde ve farklı firmalar arasında muhasebe tutarlılık ölçütlerini belirlemek için bir çalışma yapmıştır. Bu çalışmanın sonucunda, muhasebe tutarlılığının ve doğru analist tahminlerinin daha güçlü hisse getirilerini tetiklediği sonucuna varılmıştır. Henselmann ve arkadaşları (Henselmann ve ark., 2015), XBRL dosyalarında anormal sayılara sahip şüpheli şirketleri tespit etmeyi amaçlayan bir sistem tasarımı sunmuştur. Genel olarak, işletmelerin yatırımcıları teşvik etmek için daha yüksek kazançlar gösterdikleri tespit edilmiştir. Chen ve arkadaşları (X. Chen ve ark., 2021), şirketler tarafından paylaşılan XBRL formatındaki 10K dosyalarını kullanarak yıllık kazançları tahmin etmek için bir makine öğrenimi yaklaşımı sunmuştur. SEC tarafından finansal raporların XBRL formatında paylaşılmasının gerekliliğinin finansal raporların karşılaştırılmasına etkisi Dhole ve arkadaşları (Dhole ve ark., 2015) tarafından araştırılmıştır. Bu çalışmada, son yıllarda finansal raporların ve firma özel taksonomisinin karşılaştırılmasının daha zor hale geldiği sonucuna varılmıştır. Öte yandan, Li (Li, 2010), Naive Bayes algoritmasını kullanarak 10K (yıllık) ve 10Q (çeyrek) raporlarının Yönetim Tartışması ve Analizi bölümündeki ileriye dönük beyanların içeriğini analiz etmiştir. Araştırma, içeriğin genellikle olumlu olduğunu ve birçok faktörün zamanla değişmesine rağmen içeriğin buna göre değiştirilmediğini ortaya koymuştur. XBRL formatlı raporlardaki etiket yapısı üzerine çalışmalar da yapılmıştır. Plumlee ve Plumlee (Plumlee ve Plumlee, 2008), etiketler aracılığıyla finansal raporlama sürecinde XBRL güvencesini incelemiştir. Mevcut durumda birçok eksiklik ve hata tespit edilmiştir. Loukas ve arkadaşları (Loukas ve ark., 2022), XBRL etiketlerini inceleyerek etiketlerden oluşan cümlelerden oluşan bir veri tabanı önermişlerdir. Bunun için etiketlerin anlamsal doğrulama ve ilişkilendirme işlemlerini doğal dil işleme ve makine öğrenimi yöntemleriyle sağlamışlardır. Chen ve arkadaşları (X. Chen ve ark., 2022), karar ağacı yöntemini kullanarak gelecekteki kazançları tahmin etmeye çalışmıştır. Bu çalışmada olduğu gibi, XBRL formatında 10K dosyalarını kullanmıştır. Ayrıca, etiketleri düzenlemek ve etiketler aracılığıyla ilişkiler bulmak üzerine çalışmışlardır.

2.3. Finansal Verilerde Doğal Dil İşleme

Doğal dil işleme, makine öğrenmesinin en eski araştırma alanlarından biridir ve günümüzde birbirinden bağımsız pek çok alanda farklı cevaplar ve çözümler üretmeye devam etmektedir. Finans alanı da bilgi doğruluğu ve özetleme gibi birçok konuda doğal dil işleme yöntemlerinin sıklıkla tercih edildiği bir sektördür. Jain ve arkadaşları gerçekleştirdikleri çalışmada (Jain ve ark., 2018), doğal dil işleme için kullanılan çeşitli algoritmalar ve onların çalışma prensipleri örneklerle açıklayarak farklı alanlara entegrasyonuna vurgu yapmıştır. Son on yılda bu alanda yapılan gelişmeler ve farklı algoritmalar arasındaki farklar incelenmiş, ayrıca bu algoritmaların uygulama alanları da belirtilmiştir. Doğal dil işleme henüz mükemmelliğe ulaşmadığı, ancak sürekli olarak yapılan iyileştirmelerle gelecekte daha da gelişeceğinin öngörüldüğü belirtilmiştir. Boritz ve arkadaşları (Boritz ve ark., 2013)denetçilerin raporlarında kullandıkları terimlere odaklanan otomatik içerik analizi yöntemini kullanarak, bilgi teknolojisi zayıflıklarını tanımlayan bir çalışma gerçekleştirerek 14 kategoride Bilgi Teknolojileri Zayıflıklarını belirlemiştir. Sun ve Vasarhelyi yaptıkları araştırmayla, denetçilerin metin verilerinden elde edilen bilgilerin kullanışlılığını inceleyerek derin öğrenme teknolojisiyle denetim süreçlerini geliştirmeyi amaçlamıştır (Sun ve Vasarhelyi, 2018). Metin verilerinin etkili bir şekilde kullanılmasını engelleyen teknoloji çözümlerinin eksikliğini göz önüne alarak, derin öğrenme yaklaşımıyla bu sorunun üstesinden gelmenin mümkün olduğu vurgulanmıştır. Yaptıkları araştırma, denetçilere derin öğrenme tekniklerini kullanmak için hazır araçlar ve açık kaynaklı kütüphaneler sağladığını belirterek, bu sayede daha kaliteli denetim delilleri elde etmek ve iş içgörülerini artırmanın mümkün olacağı sonucuna varmışlardır. Muhasebe araştırmacılarının büyük metin kümelerini analiz etmek için bilgisayar destekli içerik analizi tekniklerini kullanma olasılığı Bogaerd ve Aerts tarafından gerçekleştirilen çalışmada incelenmiştir (Bogaerd ve Aerts, 2011). Manuel olarak yapılan kodlama işleminin zaman alıcı, maliyetli ve doğruluk açısından sınırlamaları olduğu belirtilmiştir. Bu nedenle, bilgisayar destekli yöntemlerin doğruluğu üzerinde durulmuş ve LPU (Pozitif ve Etiketsiz Öğrenme) yönteminin yaklaşık %90 doğruluk oranına sahip olduğu bulunmuştur. Araştırmada, doğruluk değerlendirmesinin önemine vurgu yapılmış ve bu alandaki araştırmaların daha fazla yapılması gerektiği ifade edilmiştir.

Kearney ve Liu çalışmalarında metin duygusu literatürünü incelemiş, farklı bilgi kaynakları, içerik analizi yöntemleri ve deneysel modelleri karşılaştırmıştır. Metin

duygusunun bireyler, firmalar ve piyasalar üzerindeki etkisi ve bunlardan nasıl etkilendiği üzerine önemli bulgular özetlenmiştir. Ayrıca hem üzerinde uzlaşılan noktalar hem de tartışmalı konular belirtilmiştir. Gelecekteki araştırmalar için umut verici yönelimler, giderek daha karmaşık metin içerik analizi ve kapsamlı sözlüklerin kullanılmasıyla daha doğru ve verimli duygu ölçümlerinin elde edilebilmesinin imkânı vurgulanmıştır. Muhasebe alanındaki bilgi ve standart yükünün çok fazla olmasıyla birlikte, muhasebe dilinin evrensel doğası ve kullanıcı çeşitliliği nedeniyle otomatik olarak ilgili muhasebe kavramlarını gruplandırmak için istatistiksel yöntemlerin kullanılabilirliği Garnsey tarafından incelenmiştir (Garnsey, 2006). Frekanslarına dayalı olarak kelime belgelerinde ağırlıklandırma yapılarak Latent Semantic Indexing (LSI) (Papadimitriou ve ark., 1998) ve birleşik kümeleme yöntemleri kullanılarak ilgili muhasebe kavramlarının kümeleri türetilmiştir. Sonuçlar, ağırlıklandırılmış terim-belge matrisinden elde edilen kümelerin ve LSI matrisinden elde edilen kümelerin önemli sayıda ilgili terim içerdiğini göstermiştir. Özellikle LSI kümelerinde, kütüphane içinde daha az frekansa sahip terimler bulunmuştur.

Chen ve arkadaşları (C. L. Chen ve ark., 2011) finansal tablolardaki öznel ifadelerin, şirketlerin değerini ve karlılığını değerlendiren piyasa katılımcılarının kararlarını etkilediğini belirtmiştir. Yaptıkları çalışmada, yapay zekâ temelli bir strateji kullanılarak finansal durum ile öznel ifadeler arasındaki bağlantı araştırılmıştır. Koşullu rastgele alan (CRF) (Lafferty ve ark., 2001) modelleri kullanılarak öznel ifade desenleri tespit edilmiş ve bu yöntem önceki çalışmalarda kullanılan modellere göre daha başarılı olduğu sonucuna varılmıştır. Ayrıca, yeni algoritmalar sayesinde finansal tablolardaki yazılı ifadelerin gerçeklikle uyumsuzluk gösterdiği kanıtlanmıştır. Beklenmedik negatif kazançlar genellikle belirsiz ve hafif ifadelerle birlikte gelirken, bazen de parlak bir geleceğe dair vaatlerle desteklenmektedir. Kamaruddin ve arkadaşları metin belgelerini madencilik yaparak sapmaları veya anormallikleri keşfetme çabalarının son yıllarda arttığını çalışmalarında (Kamaruddin ve ark., 2015) belirtmişlerdir. Günümüz veri havuzlarında bulunan metin verilerinin artması nedeniyle, metin madenciliği gizli bilgi içeriklerini ortaya çıkarmada yardımcılığı, çeşitli metin madenciliği uygulamalarının mevcut olduğu ancak genellikle veri özetleme veya belge sınıflandırma konusunda yardımcı olmaya odaklanıldığı yine çalışmada vurgulanmıştır. Yöntemlerin faydalı olduğu kanıtlanmış olsa da belgelerdeki anormallikleri tespit etmek için yeterli bir semantik analiz ve titiz metin karşılaştırması sunmadığı kanaatine varmışlardır. Çalışmada, finansal belgeler koleksiyonunda bulunan cümle sapmalarını tespit edebilen

bir metin madenciliği sistemi önerilmiştir. Sistem, cümleleri grafik olarak temsil ederek birbirleriyle karşılaştırmak için bir benzerlik fonksiyonu uygulamaktadır. Önerilen sistem üzerinde yapılan değerlendirme, bir bankanın mali tablolarını kullanarak yapılan deneyler etrafında dönmektedir. Elde edilen bulgular, önerilen sistemin belgelerde meydana gelen sapmaları tespit etme yeteneğine sahip olduğunu göstermiştir. Tespit edilen sapmaların, yetkililere iş kararlarını iyileştirmek için faydalı olabileceği kanaatine varılmıştır.

2.4. Doc2Vec K Means CNN Hibrit Algoritmasıyla İlişkili Çalışmalar

Finansal raporlamayı Genişletilebilir İşletme Raporlama Dili (XBRL) standartlarına uygun hale getirmek büyük bir devrim olarak kabul edilmektedir. Bu, finansal verilerin paylaşımını ve yönetimini çok daha verimli hale getirerek yeni fırsatlar sunmuştur. Bununla birlikte, yalnızca XBRL, güvenilir, tutarlı ve hatasız bilgi sağlayamamaktadır. Özellikle raporlarda kullanılan etiketlerin doğruluğu ve geçerliliği bilgi kalitesini doğrudan etkilemektedir (Efendi ve ark., 2016). Ayrıca, etiketlerin standartlaştırılması ve gereksiz detaylardan arındırılması, araştırma konularının gelişmekte olan bir alanı olarak kabul edilmektedir (Vasarhelyi ve ark., 2012).

Loughran, veri biriktiricilerin göremediği temel ipuçlarını 10K ve 10Q raporlarının yatırımcılara sağlayabileceğine inanmaktadır (Loughran ve McDonald, 2017). Balsam ve arkadaşları 10Q raporlarını inceleyerek beklenmedik öngörüye dayalı giderler ile hisse senedi getirileri arasında negatif bir ilişki bulmuşlardır (Balsam ve ark., 2002). Çalışmalarının sonuçlarını istatistiksel çıkarımlarla göstermişlerdir. Griffin' in çalışması, yatırımcıların 10K ve 10Q raporlarına verdiği tepkiyi incelemiştir. Sonuçlarını, yayınlanma günleri etrafındaki ortak hisse senetlerinin getirileri olarak ölçülen yatırımcı tepkisini ölçen bir istatistiksel analizle sonuçlandırmıştır.

Makine öğrenimi yöntemlerini kullanarak finansal verileri analiz etmeye olan ilgi gün geçtikçe artmaktadır. Kotuza ve arkadaşları tarafından listelenen şirketlerin finansal raporlarında metinsel veya sayısal verilere yönelik doğal dil işleme, duygu analizi ve zaman serisi analizi gibi birçok konu yatırımcıların dikkatini çekmektedir (Fisher ve ark., 2016; Kamaruddin ve ark., 2015; Kearney ve Liu, 2014; Li, 2010). Şirketin mevcut mali durumunu veya diğer yönlerini ortak dönemlik ve yıllık raporlardan tahmin etmek (Liu, 2013; Magnusson ve ark., 2005), raporların hisse senedi fiyatına etkisi (Kang ve ark., 2018; Kim ve ark., 2019; Lee, 2012) ve raporlarda kullanılan dilin analizi gibi konular incelenmiştir. Özellikle muhasebe alanında büyük bir potansiyel olduğu bulunmuştur

(Fisher ve ark., 2016). Haider ve arkadaşları BBC haber veri setini kullanarak metin özetleme gerçekleştirmiştir (Haider ve Mahi, 2020). Bunun için metin verilerini vektörleştirmek için Gensim kütüphanesini kullanarak K Means kümeleme yöntemini kullanmışlardır. En iyi sonuçları işletme makalelerinde elde etmişlerdir. Literatürde K Means ve CNN yöntemlerinin birlikte kullanıldığı farklı alanlarda yapılan çalışmalar bulunmaktadır. Dong ve arkadaşları çalışmalarında, bu iki yöntemi birleştirdikleri bir kısa vadeli yük tahmin modeli sunmuşlardır (Xishuang ve ark., 2017). Plaka tanıma yöntemi Pustokhina ve arkadaşları tarafından önerilmiştir (Pustokhina ve ark., 2020). Tanıma ve bölütleme için K Means algoritmasını, plaka üzerindeki numarayı tanımak için CNN algoritmasını kullanmışlardır. Myokardit teşhisi konusundaki bir başka çalışma ise Sharifrazi ve arkadaşları tarafından gerçekleştirilmiş ve hibrit bir CNN K Means modeli kullanmışlardır (Sharifrazi ve ark., 2020). Myokardit teşhisinin doğru tahmininde %92.3 isabet oranına ulaşmışlardır. Literatür araştırması sonuçları incelendiğinde, bu çalışmada önerilen Doc2Vec, K Means ve CNN yöntemlerini birleştiren hibrit bir yaklaşım daha önce kullanılmamıştır.

2.5. Evrişimsel Sinir Ağı ve Uzun Kısa Süreli Bellek Hibrit Algoritması Çalışmaları

İki güçlü derin öğrenme algoritması olan evrişimsel sinir ağı ve uzun kısa süreli bellek yöntemleri çok farklı veri özelliklerine ve türlerine adapte edilebilmektedir. Bu yüzden, kullanım alanları da oldukça geniştir. Literatürde bu iki geleneksel derin öğrenmesi ağından üretilen birçok farklı hibrit yönteme rastlanmıştır. Farklı çalışmalarda bu iki algoritmanın seçim sebebi olarak, iki algoritmanın analizde veriyi ele alış şekilleri, verilerdeki özellikleri daha iyi kavramaları ve veri üzerindeki bağlantıları keşfetme güçleri en başta değinilen noktalar olmuştur.

Uzun kısa süreli bellek ve evrişimsel sinir ağı mimarilerinin yeni bir derin öğrenme modeline entegrasyonu, fotovoltaik enerji üretimini tahmin etmede, standart makine öğrenmesi yöntemlerini ve tek derin öğrenme modellerini geride bırakmada üstün performans sergilediği, Fas' tan gerçek dünya veri seti kullanılarak yürütülen çalışmayla yazarlar tarafından kanıtlanmıştır (Agga ve ark., 2022). Abdallah ve arkadaşları ağ sistemlerinde izinsiz girişleri tespit etmek için CNN ve LSTM ağlarından yeni bir hibrit yöntem önermiştir (Abdallah ve ark., 2021). Elde edilen sonuçlarla, hibrit yöntemin, derin öğrenme yöntemlerinin tek başına kullanılmasına göre izinsiz girişlerin tespitinde daha

başarılı sonuçlar ürettiği kanıtlanmıştır. Shang ve arkadaşları, çalışmalarında, metalik boru hatlarındaki ultrasonik kılavuzlu dalgaların çözülebilmesi için geliştirdikleri CNN-LSTM hibrit modelinin, bireysel CNN ve LSTM modellerine kıyasla çeşitli kusurları tespit etmede oldukça üstün sonuçlar ürettiğini kanıtlamıştır (Shang ve ark., 2023). CNN ve LSTM yöntemleri birleştirilerek üretilen hibrit yöntemin, Parkinson hastalığının tespitinde, geleneksel makine öğrenmesi yöntemlerine kıyasla oldukça başarılı sonuçlar ürettiği Lilhore ve arkadaşları tarafından gerçekleştirilen çalışmada sunulmuştur (Lilhore ve ark., 2023). Farklı dış etkenlere bağımlılığı ve sürekliliği zorlayıcı olan hisse fiyatlarının tahmini için de CNN ve LSTM ağlarından üretilen hibrit bir yöntem Lu ve arkadaşları tarafından önerilmiştir (Lu ve ark., 2020).

3. MATERYAL VE YÖNTEM

Tez çalışması kapsamında yapılan tüm araştırmalar, finansal verilerin elde edilmesi ve analizi, kullanılan ve üretilen tüm yöntemler ve çalışmada faydalanılan araçlar bu bölümde açıklanmıştır.

3.1. 10K Verilerinin Analizi

Son 20 yılda özellikle yapay zekâ yöntemlerinin kullanılarak finansal verilerin çeşitli amaçlarla analiz edildiği birçok çalışma gerçekleştirilmiştir. Dünya borsalarında faaliyet gösteren şirketlerin ve kamu kurumlarının tüm faaliyetlerine dair bilgi paylaşım zorunlulukları, bu bilgilerin paylaşımı sırasında önceden belirlenmiş yasa ve düzenlemelere uyma kısıtlamaları, paylaşılan bilgilerin şirketlere ve piyasalara dair önemli veriler içermesi ve özellikle yatırımcıların, hissedarların ve denetleyicilerin bu verilere ve bu verilerden elde edilecek üst bilgilere oldukça ilgi göstermeleri sebebiyle, finansal alanda yapılan çalışmaların nitelik ve niceliği ile öneminin her geçen gün artacağı öngörülmektedir.

Finans alanı, bilişim teknolojilerinin ilk uygulamalarının üretildiği alanlardan biri olması nedeniyle uzun yıllardır dijital verinin biriktiği ve önceleri analitik yöntemlerle daha sonra özellikle makine öğrenmesi yöntemleriyle verilerin analiz edilerek önemli bilgilerin çıkarıldığı bir alan olmuştur.

Bu bölümde, Amerika'da faaliyet gösteren ve yıllık raporlarını düzenli olarak halka açık platformlarda paylaşan şirketlerin finansal tablolardaki bilgilerin toplanması, veri temizleme adımlarından geçirilmesi, sonraki çalışmalara uygun olarak hazırlanması ve piyasa değerleri üzerinden sınıflandırılması hedeflenmektedir.

3.1.1. EDGAR

Bu tez çalışmasında kullanılan veriler, Amerikan Menkul Kıymetler ve Borsa Komisyonu (SEC) tarafından oluşturulan EDGAR (Electronic Data Gathering, Analysis, and Retrieval) veri tabanından elde edilmiştir (SEC, 2017). EDGAR veri tabanı, 1996 yılından bu yana tüm halka açık şirketlerin şirkete dair tüm raporlarını sunmakla yükümlü oldukları bir platformdur. Bu veri tabanından, şirketlerin isimleri veya Amerika piyasalarında kullanılan Merkezi Endeks Anahtarı (CIK) kodları veya kısa isimleri olan "ticker" (hisse senedi sembolü) değerleri ile şirketin paylaştığı tüm raporlara ve finansal verilere erişmek mümkündür.

Paylaşılan verilerin kullanımında veya analizinde herhangi bir yasal kısıtlama bulunmamaktadır. Bu veriler, öncelikle şirketlerin kendi muhasebecileri tarafından iç denetim amacıyla incelenmekte ve daha sonra şirketlerin hizmet aldığı dış denetçiler tarafından denetlenmektedir. Veriler, yasa ve düzenlemeler tarafından belirtilen şartlara uyumu kontrol edildikten sonra şirket tarafından EDGAR üzerinde paylaşılmakta ve doğruluğu şirket yöneticileri tarafından imzalanarak taahhüt edilmektedir. Bu yükümlülükler ve zorunluluklar, paylaşılan verilerin güvenilirliğini sağlamaktadır.

3.1.2. XBRL

EDGAR veri tabanında, şirket verileri web tarayıcılarında görüntülemeye uygun formatta yazılı metinler olarak sunulmaktadır. Bunun yanı sıra şirketler, verilerin bilgisayarlar tarafından analizini kolaylaştırmak ve keskinleştirmek amacıyla xml tabanlı bir format olan XBRL ile verileri EDGAR veri tabanına yüklemektedir. XBRL dosyaları, şirketin sayısal ve metinsel tüm bilgilerini içermektedir ve bilgi etiket-değer düzeniyle paylaşılmaktadır. Şekil 3.1 de bir XBRL dosyasında verilerin nasıl paylaşıldığı örneği yer almaktadır. Şirketler, XBRL dahil olmak üzere paylaştıkları tüm raporları, Amerika'da finans alanında bağımsız olarak faaliyet gösteren Finansal Muhasebe Standartları Kurumu (FASB) tarafından ilan edilen ve Amerika Menkul Kıymetler ve Borsa Komisyonu tarafından kabul edilen Genel Muhasebe Prensipleri (Generally Accepted Accounting Principles, GAAP) standartlarına uygun olarak hazırlamakla yükümlüdür. Bu sayede, şirketlerin paylaştığı bilgilerin aynı etiketle paylaşılması, bilgi kirliliğinin azaltılması ve bilginin kural ve yasalara uygunluğunun daha iyi denetlenmesi hedeflenmektedir. XBRL dosyalarında kullanılan etiketler genellikle 'us-gaap:' ön ekiyle veya şirketlerin ticker sembolleriyle başlamaktadır. Etiketlerdeki 'us-gaap:' veya şirketin ticker sembolünden sonraki ifade, paylaşılan bilginin ne olduğunu belirtmektedir. Örneğin, 'us-gaap:assets', GAAP standartlarına uygun olarak şirketin varlıklarını paylaştığı bir etikettir. Ayrıca, 'XYZ' ticker sembolüne sahip bir şirket, 'xyz:informationaboutsomething' şeklinde bilgi paylaşabilmektedir.

```
<us-gaap:AdvertisingExpense contextRef="D2010Q4YTD" decimals="-6" id="Fact-24AD09F1D250CF157AEE666DA846F82A" unitRef="usd">8000000</us-gaap:AdvertisingExpense>
<us-gaap:AdvertisingExpense contextRef="D2011Q4YTD" decimals="-6" id="Fact-30179BCA18672251A16C666DA84314FF" unitRef="usd">28000000</us-gaap:AdvertisingExpense>
<us-gaap:AdvertisingExpense contextRef="D2012Q4YTD" decimals="-6" id="Fact-16F7BC763ED4809A84F0666DA846FC77" unitRef="usd">67000000</us-gaap:AdvertisingExpense>
<us-gaap:AdvertisingRevenue contextRef="D2010Q4YTD" decimals="-6" id="Fact-3FF2EEF4B4ADCBCD27E46229481643F5" unitRef="usd">1868000000</us-gaap:AdvertisingRevenue>
<us-gaap:AdvertisingRevenue contextRef="D2012Q4YTD" decimals="-6" id="Fact-D1E8F02B0896E948BE5962294813F9D8" unitRef="usd">4279000000</us-gaap:AdvertisingRevenue>
<us-gaap:AdvertisingRevenue contextRef="D2011Q4YTD" decimals="-6" id="Fact-33B0AEFFF78EFE530EC262294817BF27" unitRef="usd">3154000000</us-gaap:AdvertisingRevenue>
<us-gaap:Assets contextRef="I2012Q4" decimals="-6" id="Fact-638CF29B999FAE7A18E0D7A597972E3D" unitRef="usd">15103000000</us-gaap:Assets>
<us-gaap:Assets contextRef="I2011Q4" decimals="-6" id="Fact-8542237C512A1D0E4F53D7A597645558" unitRef="usd">6331000000</us-gaap:Assets>
<us-gaap:AdditionalPaidInCapital contextRef="I2012Q4" decimals="-6" id="Fact-F413B324D31ED949BF96D7A597E13203" unitRef="usd">10094000000</us-gaap:AdditionalPaidInCapital>
<us-gaap:AdditionalPaidInCapital contextRef="I2011Q4" decimals="-6" id="Fact-45C65E46E7D83D53CE1CD7A59771CEEA" unitRef="usd">2684000000</us-gaap:AdditionalPaidInCapital>
<us-gaap:AssetsCurrent contextRef="I2011Q4" decimals="-6" id="Fact-DD14AB12A33F5398E8A7D7A597883565" unitRef="usd">4604000000</us-gaap:AssetsCurrent>
<us-gaap:AssetsCurrent contextRef="I2012Q4" decimals="-6" id="Fact-59E0BDF92021BE1F123ED7A5979B6E9B" unitRef="usd">11267000000</us-gaap:AssetsCurrent>
```


Şekil 3.1 Bir şirketin paylaşmış olduğu XBRL dosyasından bir bölüm

EDGAR veri tabanında, şirketler SEC tarafından belirlenen zamanlarda, paylaştıkları bilginin niteliğine uygun olan form başlığı altında bilgileri paylaşırlar. Bazı özel durumlarda ise belirli bilgiler paylaşılır. Örneğin, 8K formları, şirketlerde yönetimsel değişiklik gibi önemli değişimleri bildirmek için paylaşılır. Form 144 ise şirket içindeki hisse sahiplerinin hisselerini elden çıkarmak istediklerinde şirket tarafından yayınlanması zorunlu olan bir formdur. 10Q ve 10K formları ise sırasıyla her finansal çeyrek ve yıl sonunda paylaşılması gereken formlardır. 10Q formları çeyrek raporlarını, 10K formları ise yıl sonu raporları olarak paylaşılmaktadır. İlgili finansal yıldaki tüm işlevsel faaliyetler ve finansal bilgiler bu raporlarda paylaşılmaktadır.

3.1.3. 10K yıllık raporları

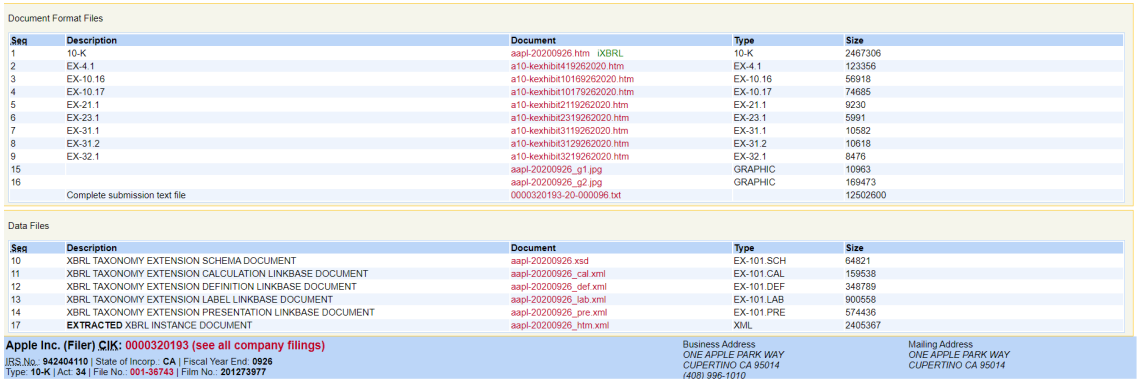
Şirketler tarafından paylaşılan raporlar arasında en geniş kapsamlı olanları 10K raporlarıdır. Bu raporlar, 4 ayrı bölüm ve 15 farklı programdan oluşmaktadır, bu detaylar Çizelge 3.1 de sunulmuştur. İlk bölüm, İş (Business) bölümüdür ve burada şirketler, faaliyet alanlarına bağlı olarak yıl boyunca gerçekleştirilen tüm operasyonları bildirmekle sorumludur. Birinci fıkrada, A bendinde risk faktörleri, B bendinde ise önceki yıl raporlarında bağımsız denetçilerin sonraki raporlarda düzeltme talep ettikleri uyarılar ve bu konuda yapılan düzeltmeler yer almaktadır. İkinci fıkrada, şirketlerin sahip oldukları önemli imkanlar, özellikler ve fiziksel varlıklar açıklanmaktadır. Üçüncü fıkrada, şirketin dahil olduğu yasal işlemler ve mahkemelerle ilgili bilgiler yer almaktadır. Dördüncü ve son fıkrada ise şirketlerin varsa iş sağlığı ve güvenliği ihlalleri veya diğer mevzuat maddeleri hakkında bilgiler bulunmaktadır.

İkinci bölümde, şirketin işlem gördüğü menkul kıymetler borsasındaki hisse değerleri, konsolide finansal veriler, yönetimin şirketin mevcut durumuyla ilgili yorumları, beklentileri, faaliyet gösterdikleri alandaki pazar riskleri, finansal tablolar ve diğer ilgili bilgiler yer almaktadır. Üçüncü ve dördüncü bölümlerde ise paylaşılan bilgilerin doğruluğunu taahhüt etme, şirket yetkililerinin imzaları ve muhasebeci

değerlendirmeleri bulunmaktadır. Bu çalışma kapsamında, tezin ilgili bölümünde 10K tablolarından yararlanılacaktır. Şekil 3.2, Amerika'da faaliyet gösteren APPLE şirketinin EDGAR veri tabanında paylaştığı 10K raporlarının bir bölümünü içeren bir web sayfasını göstermektedir.

Çizelge 3.1 10K Rapor formatı ve Bölümleri

1. Bölüm	1. İş
	1A Risk Faktörleri
	1B Çözülmeyen Denetleyici Yorumları
	2. Özellikler
	3. Yasal Takibatlar
	4. İş Sağlığı ve Güvenliği Açıklamaları
2. Bölüm	5. Pazar
	6. Konsolide Finansal Veriler
	7. Yönetimin Finansal Durum ve Faaliyet Sonuçlarına Yorumları ve Analizi
	7A. Piyasa Risklerine İlişkin Niceliksel ve Niteliksel Açıklamalar
	8. Finansal Tablolar
	9. Muhasebe ve Mali Bilgilendirme Konusunda Muhasebecilerle Yapılan Değişiklikler ve Anlaşmazlıklar
	9A. Kontrol ve Prosedürler
	9B. Diğer Bilgiler
3. Bölüm	10. Direktörler, Üst Düzey Yöneticiler ve Kurumsal Yönetim
	11.Yönetici Ücretleri
	12. Bazı İntifa Hakkı Sahiplerinin ve Yönetimin Menkul Kıymet Mülkiyeti ve İlgili Hissedar Konuları
	13. Belirli İlişkiler ve İlgili İşlemler ve Direktör Bağımsızlığı
	14. Ana Muhasebe Ücretleri ve Hizmetleri
4. Bölüm	15. Anlaşma ekleri, Mali Tablo Onay İmzaları



Document Format Files

Seq	Description	Document	Type	Size
1	10-K	aapl-20200926.htm iXBRL	10-K	2467306
2	EX-4.1	a10-kexhibit419262020.htm	EX-4.1	123356
3	EX-10.16	a10-kexhibit10169262020.htm	EX-10.16	56918
4	EX-10.17	a10-kexhibit10179262020.htm	EX-10.17	74685
5	EX-21.1	a10-kexhibit2119262020.htm	EX-21.1	9230
6	EX-23.1	a10-kexhibit2319262020.htm	EX-23.1	5991
7	EX-31.1	a10-kexhibit3119262020.htm	EX-31.1	10582
8	EX-31.2	a10-kexhibit3129262020.htm	EX-31.2	10618
9	EX-32.1	a10-kexhibit3219262020.htm	EX-32.1	8476
15		aapl-20200926_g1.jpg	GRAPHIC	10963
16		aapl-20200926_g2.jpg	GRAPHIC	169473
	Complete submission text file	0000320193-20-000096.txt		12502600

Data Files

Seq	Description	Document	Type	Size
10	XBRL TAXONOMY EXTENSION SCHEMA DOCUMENT	aapl-20200926.xsd	EX-101.SCH	64821
11	XBRL TAXONOMY EXTENSION CALCULATION LINKBASE DOCUMENT	aapl-20200926_cal.xml	EX-101.CAL	159538
12	XBRL TAXONOMY EXTENSION DEFINITION LINKBASE DOCUMENT	aapl-20200926_def.xml	EX-101.DEF	348789
13	XBRL TAXONOMY EXTENSION LABEL LINKBASE DOCUMENT	aapl-20200926_lab.xml	EX-101.LAB	900558
14	XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE DOCUMENT	aapl-20200926_pre.xml	EX-101.PRE	574436
17	**EXTRACTED** XBRL INSTANCE DOCUMENT	aapl-20200926_htm.xml	XML	2405367

Apple Inc. (Filer) CIK: 0000320193 (see all company filings)
IRS No.: 942404110 | State of Incorp.: CA | Fiscal Year End: 0926
Type: 10-K | Act: 34 | File No.: 001-36743 | Film No.: 201273977

Business Address
ONE APPLE PARK WAY
CUPERTINO CA 95014
(408) 996-1010

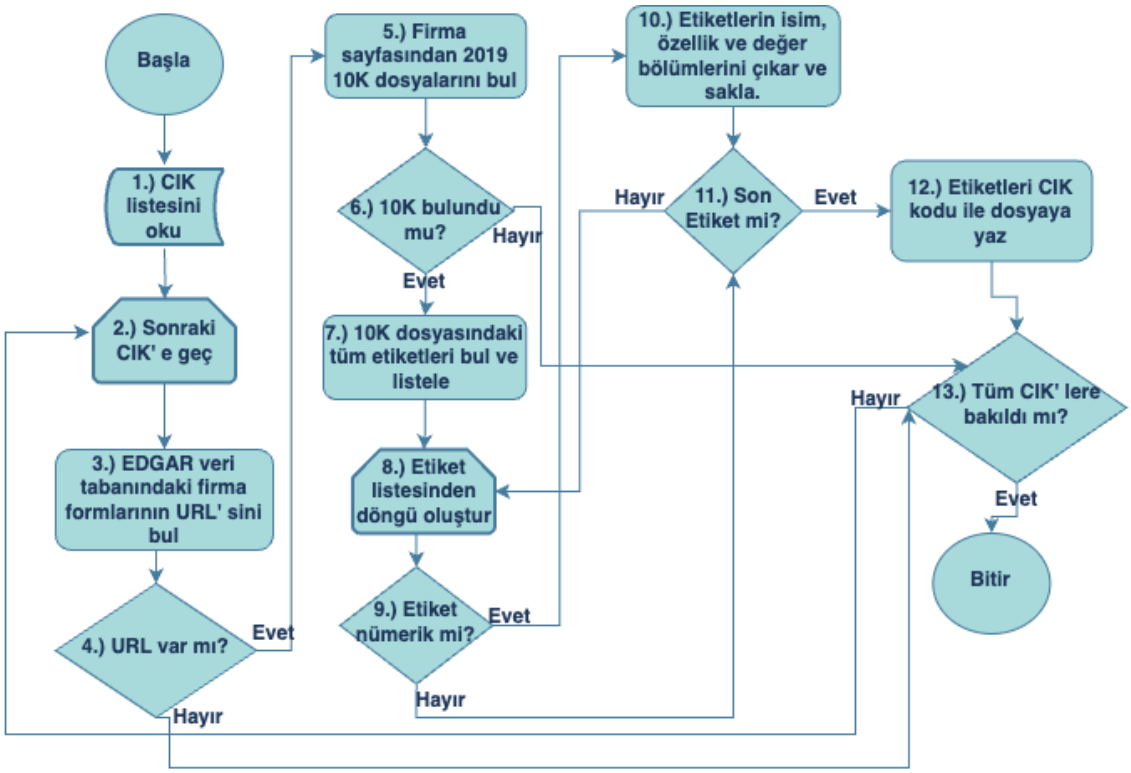
Mailing Address
ONE APPLE PARK WAY
CUPERTINO CA 95014

Şekil 3.2 EDGAR veri tabanı APPLE Şirketi 10K Formlarının yer aldığı sayfa

3.1.4. Verilerin toplanması

Tez kapsamında gerçekleştirilen çalışmalar Python[1] dilinde Pycharm[2] program geliştirme ortamında yürütülmüştür. Verilerin EDGAR veritabanından çekilmesi için BeautifulSoup[3] çalışma ortamından yararlanılmıştır. Amerika'da faaliyet gösteren en büyük özel borsa olan National Association of Securities Dealers Automated Quotations (NASDAQ) borsasında 2020 yılı itibariyle faaliyet gösteren 10711 şirketin EDGAR veri tabanına yüklemiş oldukları 10K tablolarının taranması ve indirilmesi hedeflenmiştir. Bu bağlamda, şirketlerin veri tabanı adreslerine CIK ile ulaşan, şirketlerin geçmiş 10K raporlarına erişen ve bu raporları ayrıştırarak metin dosyası olarak çeken bir arama motoru (web crawler) tasarlanmıştır. Bu motor, elde edilen bağlantı ile şirketlerin 01.01.2011 ile 31.12.2019 tarihleri arasında paylaşmış oldukları XBRL formatlı 10K raporlarını taramıştır. Bu on yıllık süreçte, kapanan veya yeni açılan şirketlerin olması, bazı şirketlerin raporları hatalı formatta yüklemiş olmaları, ya da eksik veri paylaşmış olmaları gibi durumlar göz önünde bulundurulduğunda, her yıl için, 3000 ile 4000 arasında değişen sayılarda 10K dosyaları elde edilmiştir. Bu çalışmada şirketlerin finansal tabloları analiz edileceği için yalnızca değer bilgisi sayısal olan XBRL etiketleri, kaynak referansları (contextref) ve ilgili etiketin tuttuğu sayısal değerler, metin dosyaları olarak indirilmiştir. Şekil 3.4' te elde edilen dosyalardan bir örnek gösterilmektedir. Toplamda 33628 dosya, şirketlerin CIK kodları ile yıl bazlı ayrılarak saklanmıştır. Üretilen arama motoru algoritmasının akış diyagramı Şekil 3.3' te verilmiştir.

[1] Açık kaynak kodlu bir programlama dili. Web sayfası: https://www.python.org/
[2] https://www.jetbrains.com/pycharm/
[3] https://shorturl.at/fkRY3



Şekil 3.3 10K yıllık raporlarını tarayıp indiren web arama motoru akış diyagramı


```
{'scheme': 'http://www.sec.gov/CIK'};identifier;0001308606
{'contextref': 'P01_01_2019To12_31_2019'};dei:entitycentralindexkey;0001308606
{'contextref': 'P01_01_2019To12_31_2019'};dei:documentfiscalyearfocus;2019
{'contextref': 'P01_01_2019To12_31_2019_SpiritRealtyLpMemberdeiLegalEntityAxis'};dei:documentfiscalyearfocus;2019
{'contextref': 'PAsOn02_21_2020', 'decimals': 'INF', 'id': 'hidden11118253', 'unitref': 'Unit_shares'};dei:entitycommonstocksharesoutstanding;102522792
{'contextref': 'P01_01_2019To12_31_2019'};dei:entityaddresspostalzipcode;75201
{'contextref': 'P01_01_2019To12_31_2019'};dei:cityareacode;972
{'contextref': 'PAsOn06_28_2019', 'decimals': '-8', 'unitref': 'Unit_USD'};dei:entitypublicfloat;3800000000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:landandlandimprovements;1910287000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:landandlandimprovements;1632664000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:buildingsandimprovementsgross;3840220000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:buildingsandimprovementsgross;3125053000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:realestateinvestmentpropertyatcost;5750507000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:realestateinvestmentpropertyatcost;4757717000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:realestateinvestmentpropertyaccumulateddepreciation;717097000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:realestateinvestmentpropertyaccumulateddepreciation;621456000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};src:netrealestateheldforinvestment;5033410000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};src:netrealestateheldforinvestment;4136261000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:loansandleasesreceivablenetreportedamount;34465000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:loansandleasesreceivablenetreportedamount;47044000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:finitelivedintangibleassetsnet;385079000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:finitelivedintangibleassetsnet;294463000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:directfinancingleasenetinvestmentinlease;14465000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:capitalleasesnetinvestmentindirectfinancingleases;20289000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:realestateheldforsale;1144000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:realestateheldforsale;18203000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:realestateinvestmentpropertynet;5468563000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:realestateinvestmentpropertynet;4516260000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:cashandcashequivalentsatcarryingvalue;14492000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:cashandcashequivalentsatcarryingvalue;14493000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:deferredcostsandotherassets;124006000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:deferredcostsandotherassets;156428000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:heldtomaturitysecurities;33535000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:investmentownedatcost;150000000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:goodwill;225600000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:goodwill;225600000
{'contextref': 'PAsOn12_31_2019', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:assets;5832661000
{'contextref': 'PAsOn12_31_2018', 'decimals': '-3', 'unitref': 'Unit_USD'};us-gaap:assets;5096316000
```


Şekil 3.4 Bir şirketin paylaştığı XBRL dosyalarından çekilen Kaynak Referansı-Etiket-Değer formatlı bilgiler

3.1.5. Veri ön işleme

Çalışmanın bu adımında XBRL formatlı 10K yıllık raporlarının makine öğrenmesi yöntemlerine dönüştürülebilmesi için önemli veri ön işleme ve temizleme

yöntemleri önerilmiştir. Önceki adımda veriler elde edilirken her ne kadar uygulanan filtrelerle çalışma için gereksiz bilgilerin büyük kısmından ayrılmış olsa da hala ham veri olma özelliğini korumaktadır. XBRL formatlı raporların en büyük problemi, şirketlerin paylaştıkları raporların bilgisayar ortamında analizi için paylaşılmış olmalarına rağmen standartlardan uzak, şirket bazlı özelleşen, çok fazla hatalı ya da eksik veri içeren, çoğunlukla gelişigüzel bir şekilde hazırlanan raporlar olmasıdır. Bu yüzden, bu bölümde önerilen yöntemlerin literatüre önemli katkı sağlayacağı düşünülmektedir. Önceki adımda elde edilen veri, bu adımda önerilen yöntemler uygulandığında, csv[4] formatlı, analizlere hazır veri setlerine dönüştürülmüştür. Bunun için aşağıdaki adımlar sırasıyla gerçekleştirilmiştir.

- Kaynak Referansı bilgilerinden yıl bilgisinin elde edilmesi
- Yıl bilgisi olmayan veya uygun formatta olmayan dosyaların elenmesi
- Değeri olmayan satırların dosyalardan silinmesi
- Aynı etiket için farklı yıl aynı değeri olan satırların dosyalardan silinmesi
- Değeri 1 veya 0 olan satırların dosyalardan silinmesi
- Etiket tekrarlarının silinmesi
- Etiket ön eklerinin çıkarılması
- Aynı anlama gelebilecek etiketlerin tespiti
- Birleştirilen etiketlerin birlikte kullanıldığı tüm dosyalarda aynı olup olmadığının kontrolü
- Etiket Birleşimlerinin Çapraz Eşleştirilerek Gruplandırılması
- Verilerin Sınıflandırma için Etiket Grupları Üzerinden Vektörizasyonu
- Verilerin Şirketlerin Piyasa Değerine Göre Sınıflara Ayrıştırılması

3.1.5.1. Kaynak referansı bilgilerinden yıl bilgisinin elde edilmesi

XBRL dosyalarında yer alan kaynak referansı etiketinde şirketler, ilgili etiketin şirkete özel kimlik kodu, ilgili etiketin değerinin para biriminin ne olduğu, değerin kaç rakamdan oluştuğu ve etiketin yıl veya dönem olarak hangi yıl veya dönem için açıklandığı bilgilerini paylaşmaktadırlar. Çalışmada etiketlerin kimlik kodu kullanılmayacağı, incelen tüm şirketlerin paylaştığı mali değerlerin Amerikan doları (USD) cinsinden oluşu ve paylaşılan bilgilerdeki mali değerlerin bin, milyon cinsinden

[4] CSV: Virgülle Ayrılmış Değerler (Comma seperated values) anlamına gelen bir dosya saklama türü

gösterimi gibi herhangi bir gösterime tabi olmadığı için tarih bilgileri hariç tüm bilgiler elimine edilmiştir. Tarih bilgileri de 10K raporları dönemlik değil yıllık raporlar olduğu için yalnızca yıl bilgileri kalacak şekilde düzenlenmiştir.

3.1.5.2. Yıl bilgisi olmayan veya uygun formatta olmayan dosyaların elenmesi

Şirketler XBRL dosyalarının kaynak referansı kısmında yıl veya dönem bilgilerini girmekle yükümlüdür. Ancak, bazı şirketlerin bilgileri, bu bilgileri eksik veya hatalı girdikleri için sonraki analiz adımına geçilmeden elimine edilmiştir. Şirketler dönem veya yıl bilgisini paylaşma zorunlulukları olsa da bu bilgileri istedikleri formatta paylaşabilmektedir. Bu yüzden, dosyalarda farklı şirketler yyyy/aa/gg, aa/gg/yyyy, gg/ay adı/yyyy gibi farklı tarih formatlarını tercih etmiş, bazı şirketler 'Ocak ayından aralık ayına' gibi ifadelerle yıl bilgisi olmadan yalnızca bilgilerin hangi ay aralığını kapsadığını belirtmiş bazıları da 10K raporları yıllık raporlar olmasına rağmen birden fazla yılı kapsayacak veya bir yılı tamamlamayan aralıklar belirtmiştir. Tablo 3.2' de kaynak referansında yer alan farklı dönem-tarih bilgi örnekleri verilmiştir. Farklı tarih formatlı bilgilerden yıl bilgisi çekilmiş ancak örneklerle açıklanan diğer şirket verileri herhangi bir belirsizliğe yer vermemek için analizden çıkarılmıştır. Son olarak elde edilen yıl bilgisi ile incelenen dosyadaki bilgiler etiketin ilgili yılı, etiketin adı ve etiketin değerinden oluşan satırlara çevrilmiştir.

Çizelge 3.2 Farklı Şirketlerde farklı formatta belirtilen dönem veya tarih bilgileri

Dönem veya Tarih Bilgisi
FD2019Q4YTD
D20190101-20191231
Duration_1_1_2019_To_12_31_2019
As_Of_12_31_2018_us-gaap_Statement
iee3fce9cc0c446609a421a12d16_I20200117
From2019-01-01to2019-12-31
From_January_to_December
Duration_01_January_2019_To_31_December_2019

3.1.5.3. Değeri olmayan satırların dosyalardan silinmesi

Bu aşamada, önceki aşamada elde edilen satırlar incelenmiştir. Dosyadaki satırlarda, ilgili yılda her bir etiket için bir değer olmak zorundadır. Örneğin varlık etiketine karşılık gelen şirketin varlıklarının durumunu belirten sayısal bir değer olması gerekmektedir. Ancak, eksik veya hatalı veri girişlerinden veya paylaşılması zorunlu

olmayan bilgilerin etiketleri girildiği halde değerleri boş bırakıldığı için bazı satırlarda eksik bilgiler yer almaktadır. Bilgi eksikliği nedeniyle böyle satırlar dosyalardan çıkarılmıştır.

3.1.5.4. Aynı etiket için farklı yıl aynı değeri olan satırların dosyalardan silinmesi

Önceki aşamalardan geçen satırlar bu aşamada tarih bazlı karşılaştırılmıştır. Bir şirket dosyasının aynı etiket ismi için aynı değere ancak farklı yıl bilgisine sahip satırları silinmiştir. Şirketin farklı yıllar için aynı değere sahip varlık veya gider kalemleri söz konusu olabilir ancak dosyalar incelendiğinde bunun anlam karmaşasına yol açtığı, verinin tutarlılığını azalttığı sonucu elde edilince söz konusu bilgilerin yanlış veya hatalı girilmiş olabileceği göz önüne alınarak ilgili satırlar dosyalardan çıkarılmıştır.

3.1.5.5. Değeri 1 veya 0 olan satırların dosyalardan silinmesi

İlgili etiketler için verilen değer bilgilerinde 0 veya 1 içeren satırlar bu aşamada dosyalardan çıkarılmıştır. Dosyalar incelendiğinde 0 veya 1 değerlerinin doğrulama değeri olarak da kullanılabildiği, söz konusu 0 veya 1 değerlerinin eksik veri girişleri ile ilişkili olması ve gerçekleştirilecek çalışmada finansal tablolarda yer alan bilgilerin analiz edileceği değerlendirilerek ilgili satırlar elimine edilmiştir. Böylece yıl, etiket adı ve etiketin değerinden oluşan satırlar ön temizleme adımlarından geçirilmiş ve sonraki aşamalara hazır hale getirilerek csv formatlı dosyalar olarak kaydedilmiştir. Şekil 3.5' te bu aşamalardan geçirilip kaydedilen bir dosyadan bir bölüm gösterilmiştir.

```
2019;us-gaap:comprehensiveincomenetoftax;57453000000
2018;us-gaap:comprehensiveincomenetoftax;58037000000
2017;us-gaap:comprehensiveincomenetoftax;56505000000
2020;us-gaap:cashandcashequivalentsatcarryingvalue;38016000000
2019;us-gaap:cashandcashequivalentsatcarryingvalue;48844000000
2020;us-gaap:marketablesecuritiescurrent;52927000000
2019;us-gaap:marketablesecuritiescurrent;51713000000
2020;us-gaap:accountsreceivablenetcurrent;16120000000
2019;us-gaap:accountsreceivablenetcurrent;22926000000
2020;us-gaap:inventorynet;4061000000
2019;us-gaap:inventorynet;4106000000
2020;us-gaap:nontradereceivablescurrent;21325000000
2019;us-gaap:nontradereceivablescurrent;22878000000
2020;us-gaap:otherassetscurrent;11264000000
2019;us-gaap:otherassetscurrent;12352000000
2020;us-gaap:assetscurrent;143713000000
2019;us-gaap:assetscurrent;162819000000
```

Şekil 3.5 Şirket Dosyalarının Veri Temizleme Adımlarından Sonra Yıl-Etiket-Değer Düzenli Satırları

3.1.5.6. Etiket tekrarlarının silinmesi

Şekil 3.5' te de gösterildiği üzere dosyalarda genellikle önceki veya sonraki yıllara ait bilgiler de paylaşılmaktadır. Genellikle şirketler finansal tablolarda karşılaştırma tanımlayabilmek için son 3 yılın verilerini paylaşmaktadır. Bu yüzden incelenen yıldan üç yıl öncesine kadar veriler yer alabilmektedir. Ayrıca, 10K raporları geçerli olduğu yıl bitimini takip eden 3 ay içinde açıklanabildiği için sonraki yıla ait bazı bilgiler de yine raporlarda yer almaktadır. Bu çalışmada, yıl bazlı analizler gerçekleştirileceği için öncelikle incelenen dosya hangi yıla aitse o yılın etiket ismi ve değeri tercih edilmiş diğerleri dosyalardan çıkarılmıştır. Ancak, dosyalar incelendiğinde bazı etiketlerin yalnızca geçmiş dönemdeki bir kalemi temsil ettiği, bazı etiketlerin ise gelecek beklentileri gibi sonraki yıla dair tahmin bilgilerini içerdiği görülmüştür. Bu yüzden, etiket değeri için incelenen yılı içeren satır değeri yoksa, geçmiş veya gelecek yılı içeren satırlar tercih edilmiştir. Bu aşamadan sonra, etiket tekrarları dosyalardan çıkarılmış, tekrarsız etiketler elde edilmiştir.

3.1.5.7. Etiket ön eklerinin çıkarılması

XBRL dosyalarındaki etiketlerde üç farklı tür ön eke rastlanmıştır. Bunlar 'us-gaap:', şirketin borsadaki hisse senedi kodu (ticker sembolü) veya 'dei' ekleridir. 'us-gaap:' eki GAAP kurumu tarafından belirlenen her biri özel anlam ifade eden yaklaşık 17.000 farklı etikette kullanılan bir ön ektir (Bragg, 2004). Şirketler kendi oluşturdukları etiketleri hisse senedi kodlarını ön ek olarak tanımlayarak paylaşmaktadır. 'dei' kelimesi veri giriş yönergesi (data entry instructions) anlamına gelmektedir. Şirketler XBRL dosyalarında yıl, şehir kodu, merkezi indeks anahtarı gibi bilgileri bu ön ekle paylaşmaktadır. Bu çalışmada, şirketlerin finansal tablolarındaki bilgiler inceleneceği için söz konusu ön eke sahip etiketleri içeren satırlar da dosyalardan çıkarılmıştır.

Tüm dosyalar için farklı ön eke sahip aynı anlama gelen etikete rastlanmamıştır. Bu yüzden tüm etiketlerden 'us-gaap:' veya söz konusu şirketin hisse senedi kodlarını içeren ön ekler çıkarılmıştır.

3.1.5.8. Aynı anlama gelebilecek etiketlerin tespiti

Önceki aşamalardan geçirilen dosyalar incelendiğinde, birbirinden bağımsız yaklaşık bir buçuk milyon farklı etiket tespit edilmiştir. Bunun sebebi şirketlerin kendi etiketlerini oluşturabilmeleridir. Farklı şirketler için aynı anlamlara gelebilecek değerler

farklı isimlerde etiketlerle sunulmuştur. Bu durum, şirket verilerinin analizini oldukça zor hale getirmektedir. Toplanan 10 yıllık verilerde istisnasız tüm dosyalarda kullanılan etiketler tespit edilmeye çalışılmış ve sadece 1 etiketin tüm dosyalarda geçtiği görülmüştür. Şirket verilerinin yarısında, %60'ında veya en az 200 satır bilgisi olan dosyalarda geçen ortak etiketlere bakıldığında da sayı 15'i geçmemiştir. Böylece, tez kapsamında sonraki aşamalarda, farklı isme sahip aynı anlam taşıyan etiketlerin tespiti için metinsel analizler gerçekleştirilmesi de hedeflenmiştir.

İncelenen şirket dosyalarında aynı dosya içinde farklı etiket ismiyle aynı değerlerin tekrar ettiği tespit edilmiştir. Bazı örnekler incelendiğinde bu etiketlerin aynı değeri işaret ettiği tespit edilmiştir. Örneğin, genel muhasebe kuralları gereği 'assets' etiketi ile 'liabilitiesandstockholdersequity' etiketi şirketin varlıklarını ve borçlarını ifade ederken, birbirine eşitlenerek paylaşılmaktadır. Söz konusu etiketlerin tespiti için bir dosyada tekrar eden değerlerin etiketleri eşleştirilmiştir. Bu bağlamda yaklaşık 770.000 tekrarda 2 veya daha fazla etiket birleştirilmesi elde edilmiştir.

3.1.5.9. Birleştirilen etiketlerin birlikte kullanıldığı tüm dosyalarda aynı olup olmadığının kontrolü

Önceki aşamada elde edilen etiket birleştirmeleri birlikte kullanıldıkları tüm dosyalara bakılarak kontrol edilmiştir. Bunun sonucunda, eşleşen etiketler arasında birçok etiketin geçtiği tüm yerlerde aynı olduğu, bazılarının çoğunlukla aynı olduğu ve bazılarının ise çoğunlukla farklı olduğu tespit edilmiştir. Geçtiği tüm dosyalarda aynı değeri işaret eden etiketlerin farklı dosyalarda kullanım sıklığı az olduğu için çoğunlukla farklı değeri tutan etiket eşleşmeleri atılmış tümünde aynı olanlar ve çoğunlukla aynı olan etiketler üzerinden analize devam edilmiştir.

3.1.5.10. Birleştirilen etiketlerin tüm dosyalardaki kullanım sıklığına bakılması

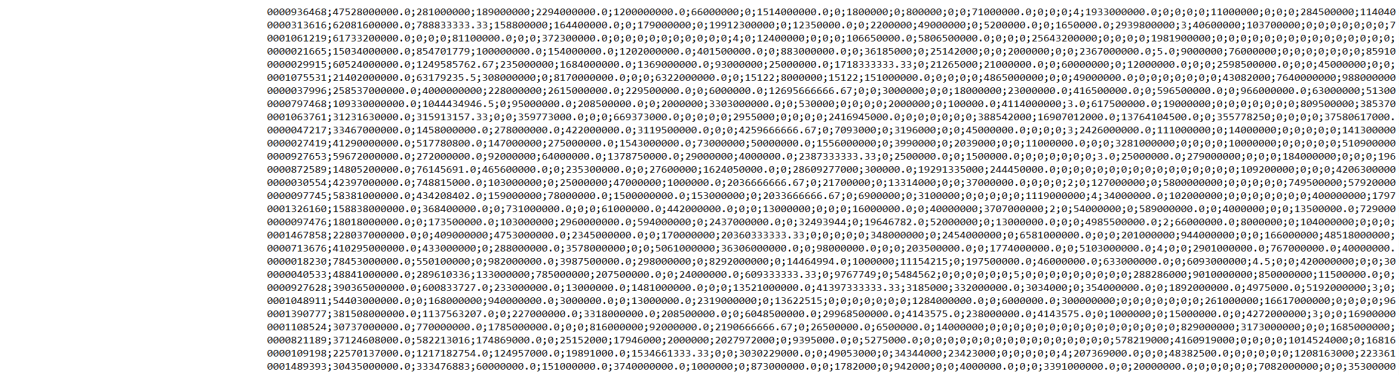
Şirketler kendi etiketlerini oluşturabilmeleri sebebiyle etiketler birleştirildikten sonra etiket birleşimlerinin büyük çoğunluğunun (yaklaşık %79'u) 2 veya 1 dosyada kullanıldığı tespit edilmiştir. Bu durum, sonraki aşamalarda etiket gruplarını oluştururken soruna yol açmıştır. Bu yüzden, bir şirketin 10 yıl için en fazla 10 farklı dosyası olacağı için en az farklı iki firmada da etiket birleşimlerinin aynı anlama gelebileceğini tespit için 10 veya daha fazla yerde aynı değeri tutan etiket birleşimleri filtre edilmiştir.

3.1.5.11. Etiket birleşimlerinin çapraz eşleştirilerek gruplandırılması

Bu aşamada etiket birleşimleri, farklı dosyalarda aynı anlamda kullanımlarına göre çapraz eşleştirilerek etiket grupları oluşturulmuştur. Örneğin; A ile B, B ile C, A ile D etiketleri önceki aşamalarda eşleştirmeye tabi tutulmuş ise, burada yapılan işlemlerle A, B, C, D etiketleri birleştirilerek tek bir etiket grubu haline getirilmiştir. Böylelikle, aynı etiketlerin farklı dosyalardaki tekrarlarının düşük olması nedeniyle zorlaşan ortak noktaların çıkarımının, özellik vektörü oluşturulabilecek bir düzenin kurgulanabilmesi açısından bu işlemle kolaylaşması hedeflenmiştir. Örneğin, üç farklı şirket varlıklarını belirtirken üç farklı isimde etiket kullandıysa, bu aşamada bu üç etiket birleştirilerek tek bir grup haline getirilmiştir. Gruplandırmadan sonra, 89 etiket grubu oluşturulmuştur. Her grupta en az 2 etiket yer aldığı için, grupların 200'den fazla etiket değerinin özniteliği olduğu bir durum oluşturulmuştur. Bu durum, hem 10K raporlarındaki verilerden daha fazla özellik çıkarımına hem de tekli etiketlerin veri özniteliği seçildiği duruma göre veri eksikliğini önemli ölçüde azaltmasıyla verinin daha iyi temsil edilmesine olanak sağladığı, elde edilen sonuçlarda görülmüştür.

3.1.5.12. Verilerin sınıflandırma için etiket grupları üzerinden vektöre dönüşümü

Önceki aşamalarda elde edilen etiket gruplarının veriyi temsil edecek özellikler olarak belirlenmesi ile 10K raporlarındaki bilgilerle şirket bazlı vektörler oluşturulmuştur. Bunun için, her bir şirketin dosyasına ayrı ayrı bakılarak etiket gruplarındaki etiketlerin dosyalardaki değerleri dikkate alınarak şirketin özellik vektörüne karşılık gelen değerleri ile veri seti için satırlar elde edilmiştir. Eğer şirketin dosyalarında etiket grubundaki etiketlerle birden fazla eşleşme varsa ilgili etiketlere karşılık gelen değerlerin ortalaması alınmıştır. Böylece, her bir satırı şirketin etiket gruplarına karşılık gelen değerlerini temsil eden bir matris elde edilmiştir. Son olarak, bazı verilerin birçok dosyada aynı değere sahip olduğu dikkate alınarak elde edilen matriste farklı satırlarda çoğunlukla aynı değere sahip olan etiket grupları ve değerleri de veri analizini etkileyebileceği için matristen çıkarılmıştır. Satır başları şirketlerin merkezi indeks anahtarları olmak üzere, Şekil 3.6' da bu aşamadan sonra elde edilen matristen bir bölüm gösterilmektedir. Elde edilen matris, sonraki aşamada sınıflandırma algoritmaları ile test edilmiştir.



Şekil 3.6 Etiket grupları ile oluşturulan şirket bazlı matris

3.1.5.13. Verilerin şirketlerin piyasa değerine göre sınıflara ayrıştırılması

Önceki aşamalarda gerçekleştirilen çalışmalarla başlangıçta toplanan ham veri, makine öğrenmesi yöntemleriyle analiz edilebilen veri setine dönüştürülmüştür. Ardından, elde edilen veri setinin veriyi temsil niteliğinin sonuç çıkarım başarımının çeşitli sınıflandırma yöntemleriyle test edilmesi hedeflenmiştir. Sınıflandırmanın gerçekleştirilebilmesi için gereken sınıf etiketleri, şirketlerin NASDAQ borsasında belirtilen güncel piyasa değerlerine göre çıkarılmıştır. Şirketlerin 10K raporlarında girmiş olduğu bilgiler üzerinden piyasa değerlerine göre ayrıştırılması hedeflenmiştir. Böylece, bir şirketin piyasa değeri ile yıllık raporları arasındaki bağlantının ortaya çıkarılması hedeflenmiştir. Veriler ile piyasa değerleri, merkezi indeks anahtarlarına göre eşleştirilmiş, eşleştirme sonucunda 2281 örnek elde edilmiştir. Sınıf sayısının belirlenmesi için veri, Expectation Maximization (Moon, 1996) kümeleme algoritmasına tabi tutulmuştur. Bunun sonucunda algoritma veriyi en fazla 3 kümeye ayırmıştır. Küme sayısı 3'ten fazla girildiğinde de yine 3 küme dışındaki kümelerde toplanan örnek sayısı 0 olmuştur. Bu yüzden şirketlerin piyasa değerleri en yüksek piyasa değeri olan şirket ile en düşük piyasa değeri olan şirket arasındaki fark üzerinden 3'e ayrılmıştır. Bu ayrıştırma sonucu şirketler, piyasa değeri en alt bölümde olan sınıfta %80 oranında yoğunlaşmıştır. Ayrıştırma sonrası çeşitli sınıflandırma algoritmalarında teste tabi tutulan verinin analizinde başarı elde edilmemiş, sınıflandırma sonucu doğruluk değerleri %50'ye yakın veya %50'nin altında bulunmuştur. Bu yüzden, şirket verileri piyasa değeri en yüksek olan şirket ile en düşük olan şirket arasında alınan ortalamaya göre 2 farklı sınıfa bölünmüştür. Bunun sonucunda şirketlerin %82'si ortalamanın altında kalan sınıfa, geri kalanı üzerinde kalan sınıfa dahil olmuştur.

3.1.6. Sınıflandırma çalışması

Önceki bölümde hazırlanan veriler 6 farklı sınıflandırma algoritmasına tabi tutularak, algoritmaların sınıflandırma başarımları farklı ölçütlerle test edilerek elde edilen veri setinin veriyi temsil yeteneğinin, tutarlılığının ve verimliliğinin ölçümü hedeflenmiştir. Dünya borsalarında endeksler belirlenirken belirli sayıdaki en değerli şirketlerin hisselerinin ortalaması baz alınır (S&P Global, 2020). Örneğin, İstanbul Borsası BIST100 endeksi veya Amerikan S&P500 endeksi bu bağlamda oluşturulan endekslere örnek olarak verilebilir. Bu yüzden test sonucunda, 10K raporları üzerinden şirketlerin paylaşmış oldukları bilgilere bakarak tüm şirketler içinde işlem gördükleri diğer şirketlere oranla ortalamanın üzerindeki en değerli şirketlerin tahmin edilmesi hedeflenmiştir. Tahmin edebilmek için kullanılan makine öğrenmesi yöntemlerinden olan sınıflandırma için Rastgele Orman (Random Forest), K En Yakın Komşu (K Nearest Neighbourhood, KNN), Destek Vektör Makineleri (Support Vector Machines, SVM), Naive Bayes, Karar Ağacı-C4.5 (Decision Tree, DT), AdaBoost, Lojistik Regresyon (Logistic Regression) (Aggarwal, 2554) algoritmaları kullanılmıştır. Verinin eğitim ve test olarak ayrılmasında K-Katlı Çapraz Doğrulama yöntemi, K sayısı 10 seçilerek kullanılmıştır. Doğruluk, dengelenmiş doğruluk, kesinlik ve duyarlılık (Lavesson ve ark. Davidsson, 2007) skorları hesaplanarak algoritmaların veri üzerindeki tahmin yetenekleri ölçülmüştür. Korelasyon yöntemiyle özellik seçimi adımı gerçekleştirilerek testlerin iyileştirilmesi hedeflenmiştir. Tüm verilerin 0-1 aralığına ayrıştırılmasıyla da verilerin skala edilmesi amaçlanmıştır. Son olarak, sınıflara ait veri setindeki örnek sayısı eşitlenerek tüm testler tekrar edilmiştir. Tüm sonuçlar detaylı bir şekilde Bölüm 4.1' de verilmiştir. Ayrıca, çıkan sonuçların yorumlanması da Bölüm 4.1.4' te yer almaktadır.

3.2. Doğal Dil İşleme Çalışmaları

Öncelikle, tez çalışmalarının bu aşamasında, önceki bölümde açıklanan çalışmalar geliştirilmiştir. Özellikle, elde edilen şirket dosyalarından 2019 yılına ait olanlar üzerinde çalışmalar ve analizler gerçekleştirilmiştir. Önceki çalışmalarda analiz için üretilen veri seti değiştirilerek sekiz farklı veri seti, yeni bakış açılarıyla oluşturulmuştur. Sınıflandırma için farklı parametreler kullanılmıştır. Analizde kullanılan yöntemlerden bazıları ve ölçüm metrikleri de yine değiştirilmiştir. Önceki bölümde açıklanan analiz yöntemlerinden üretilen sonuçlarda da iyileştirmeler elde edilmiştir. Dahası, veri toplama sürecinde elde edilen dosyalarda yer alan her bir satırda finansal bir değeri temsil eden

etiketlerin oluşturulma ve kullanılma biçimlerindeki farklılıkları anlamlandırabilmek, ortak yönlerini ortaya çıkarabilmek ve iyileştirmeler önerebilmek için etiketlerin semantik analizinde doğal dil işleme yöntemleri uygulanmıştır. Yapılan çalışmalar, bu bölümde alt başlıklarda anlatılmaktadır.

3.2.1. Veri setlerinin yeniden oluşturulması

Bu aşamada, önceki bölümde elde edilen veri setinde iyileştirmeler yapılması hedeflenmiştir. Özellikle raporlar üzerindeki etiketlerin veri setine etkisinin tespit edilerek ölçekleme çalışmaları gerçekleştirilmiştir. Böylece, daha tutarlı veri setlerinin oluşturulmasıyla daha sağlıklı sonuçların üretilmesi amaçlanmıştır. Bu amaçla, aşağıdaki adımlar gerçekleştirilmiştir.

3.2.1.1. Filtreli çoğunlukla aynı etiketler

Önceki çalışmalarda belirlenen çoğunlukla aynı etiketler için aynı olduğu dosyaların sayısına göre tekrar eşleştirmeler yapılmıştır. Belirlenen sayılardaki tekrar sayısına göre etiket grupları yeniden düzenlenmiştir. Buna göre, 10 ile 200 arasında belirlenen filtrelerde tekrar eden aynı değere sahip farklı etiketler gruplandırılmıştır. Filtrelerin kullanılmasının sebebi, bazı etiketlerin az sayıda dosyada aynı değerlere sahip iken bazılarının daha fazla sayıda dosyada aynı değere sahip olmalarıdır. Beklenildiği üzere, etiket grupları için belirlenen filtrelerde sayı büyüdükçe her bir şirket satırındaki değerler de düşmektedir. Örneğin filtre olarak 10 seçildiğinde eksik veri hücresi sayısı 14.347 iken 100 belirlendiğinde ise sayı 30.854'e çıkmıştır. Farklı filtrelerle elde edilen veri setleri analiz edildiğinde sonuçlarda iyileştirme görülmemiştir. Ayrıca, filtrenin tek başına etiketlerin aynı anlama gelip gelmediğini belirlemek için yetersiz olduğu gözlemlenmiştir.

3.2.1.2. Etiketlerin kullanım yüzdeleri

Önceki dönemde elde edilen çoğunlukla aynı olan etiketlerin gruplandırılması ile elde edilen veri setinden başka veri setleri oluşturulmuştur. Veri seti oluştururken şirketlerin ortak etiketleri ile vektöre dönüştürülmesi amaçlandığı için tüm dosyalarda geçen etiketlerin ortaya çıkarılması hedeflenmiştir. Ancak, sadece iki etiketin tüm dosyalarda geçtiği tespit edilmiştir. Ardından, belirli limitlerle tekrar arama gerçekleştirilmiştir. Örneğin, en az 50-100-200 satır bilgi yer alan dosyalar arasında tüm

dosyalarda geçen etiketler, dosya boyutu ortalamanın üzerinde olan dosyalar arasında tüm dosyalarda geçen etiketler gibi yöntemlerle arama yapılmıştır ancak yine de etiket sayısı artırılamamıştır. Bu yüzden, etiketlerin dosyalardaki kullanım sıklıklarının tespit edilmesine karar verilmiştir. Buna göre, tüm dosyalarda olmayan ancak kullanım sıklığı diğer etiketlerden fazla olan etiketlerin tespit edilmesi hedeflenmiştir. Elde edilen etiketler ile de şirket bilgilerinin vektöre dönüştürülmesi amaçlanmıştır. Tüm etiketler yüzde 100 ile yüzde 0 arasında kullanılma sıklığına göre sıralanmıştır. Yüzde 60'a kadar en çok tekrar eden etiketler ve tüm dosyalarda yer alma yüzdeleri Çizelge 3.3' te verilmiştir.

Çizelge 3.3 Etiketler ve etiketlerin dosyalar arasındaki kullanım yüzdeleri

Etiket Adı	**%**
entitycommonstocksharesoutstanding	97.31
entitypublicfloat	95.08
liabilitiesandstockholdersequity	94.95
assets	94.82
stockholdersequity	87.76
cashandcashequivalentsatcarryingvalue	85.95
propertyplantandequipmentnet	84.66
accumulateddepreciationdepletionandamortizationpropertyplantandequipment	81.48
operatingleaserightofuseasset	80.54
operatingleaseliability	79.31
sharebasedcompensation	77.24
propertyplantandequipmentgross	77.18
liabilities	76.69
netcashprovidedbyusedinoperatingactivities	76.63
incometaxexpensebenefit	76.37
commonstocksharesoutstanding	75.79
liabilitiescurrent	74
weightedaveragenumberofsharesoutstandingbasic	73.94
assetscurrent	73.94
commonstocksharesissued	73.91
weightedaveragenumberofdilutedsharesoutstanding	73.62
netincomeloss	72.68
lesseeoperatingleaseliabilitypaymentsdue	71.58
interestpaidnet	71.06
sharebasedcompensationarrangementbysharebasedpaymentawardequityinstrumentsotherthanoptionsgr	70.18
deferredtaxassetsgross	69.76
sharebasedcompensationarrangementbysharebasedpaymentawardequityinstrumentsotherthanoptionsve	69.25
sharebasedcompensationarrangementbysharebasedpaymentawardequityinstrumentsotherthanoptionsno	68.89
deferredtaxassetsvaluationallowance	68.86
lesseeoperatingleaseliabilitypaymentsdueyeartwo	68.86
entitycentralindexkey	68.7
lesseeoperatingleaseliabilityundiscountedexcessamount	68.05
documentfiscalyearfocus	67.89
paymentstoacquirepropertyplantandequipment	67.82
lesseeoperatingleaseliabilitypaymentsdueyearthree	67.53
interestexpense	66.62
netcashprovidedbyusedinfinancingactivities	66.2
allocatedsharebasedcompensationexpense	66.04
lesseeoperatingleaseliabilitypaymentsdueyearfour	65.3
sharebasedcompensationarrangementbysharebasedpaymentawardoptionsoutstandingnumber	64.88
sharebasedcompensationarrangementbysharebasedpaymentawardequityinstrumentsotherthanoptionsfo	62.8
currentstateandlocaltaxexpensebenefit	62.22
operatingleasepayments	61.86
antidilutivesecuritiesexcludedfromcomputationofearningspershareamount	61.51
lesseeoperatingleaseliabilitypaymentsdueyearfive	61.48
lesseeoperatingleaseliabilitypaymentsduenexttwelvemonths	60.89
accountspayablecurrent	60.8

3.2.2. Kullanım yüzdelerine göre veri setlerinin oluşturulması

Önceki aşamada elde edilen etiket yüzdeleri üzerinden farklı veri setlerinin oluşturulması hedeflenmiştir. Buna göre farklı yüzdelerde eşik değerleri belirlenerek farklı özellik sayılarına sahip veri setleri oluşturulmuştur. Daha düşük değerlerde tutarsız sonuçlar ürettiği için en düşük yüzde olarak %30, daha yükseğinde birden fazla etiket

kalmadığı için de en yüksek değer olarak de %90 seçilmiştir. %30-90 arasında her 10 birimlik artışla bir veri seti oluşturmak kaydıyla toplam 7 farklı veri seti oluşturulmuştur. Buna göre eşik değeri olarak %30 seçildiğinde, tüm dosyaların en az %30 ve üzerinde yer alan etiketler ile şirket dosyalarının vektörlere dönüşümü gerçekleştirilmiştir. Oluşturulan veri setlerinde özellik olarak yer alan etiketlerin sayılarının yüzdeye göre değişimi Çizelge 3.4' te verilmiştir. Bu bölümde toplamda yedi farklı veri seti elde edilmiştir.

Çizelge 3.4 Eşik değeri olarak seçilen yüzdelere bağlı değişen etiket sayısı

Yüzde	Etiket Sayısı
30	183
40	126
50	90
60	50
70	27
80	13
90	5

3.2.2.1. Veri setlerinin çoğunlukla aynı olarak tespit edilen etiket üzerinden düzenlenmesi

Veri tekrarlarının önüne geçmek için önceki çalışmada elde edilen çoğunlukla aynı etiketlerin veri setlerinden çıkarılması sağlanmıştır. Aynı değere sahip farklı etiketlerin varlığı sebebiyle böyle değerlerin veri setlerinin analizinde ezbere veya yanlış sonuçlara sebep olduğu tespit edilmiştir. Buna göre, aynı olarak tespit edilen etiketlerden daha az tekrar yüzdesine sahip etiketler çıkarıldığında elde edilen etiket sayıları Çizelge 3.5' te verilmiştir.

Çizelge 3.5 Çoğunlukla aynı tespit edilen etiketlerin çıkarılması sonrası yüzdelere bağlı etiket sayı

Yüzde	**Etiket Sayısı**
30	167
40	119
50	83
60	43
70	22
80	8
90	3

3.2.2.2. Dosyalarda geçen tüm etiketlerle veri seti oluşturulması

Tüm dosyalarda kullanılan etiketler çıkarılmıştır. Buna göre, 2019 yılı dosyaları arasında 125.000'in üzerinde farklı etiketin kullanıldığı tespit edilmiştir. Etiketler üzerinde herhangi bir işlem yapmadan tamamından oluşan bir özellik vektörü ile yeni bir veri seti oluşturulmuştur. Bu veri setinin avantajı bir şirketin paylaştığı tüm finansal bilgilerin veri setinde yer alması iken dezavantajı ise eksik veriye sahip hücre sayısının çok yüksek olmasıdır. Şirketlerde ortalama 100 farklı etiket kullanıldığı düşünülürse her bir satırda yaklaşık 124.900 hücrenin eksik veri içerdiği tahmin edilmektedir. Sonuç olarak, bölümün sonunda 3.089 şirket için aynı sayıda satıra ve 125.000 sütuna sahip bir veri seti elde edilmiştir.

3.2.2.3. Sınıf etiketlerinin oluşturulması

Önceki çalışmada şirketlerin yıllık gelirlerine göre sınıf etiketi oluşturulmuştur. Ancak, bazı şirketlerin veri paylaşımında gelir kalemlerini de paylaşmaları ile özellik olarak verilen bir değer üzerinden sınıflandırma yapılması önyargıya neden olabilmektedir. Bu yüzden, sınıf etiketleri için farklı bir yaklaşım geliştirilmiştir. Paylaşımları analiz edilen tüm şirketler Amerikan borsalarında işlem gören şirketler olduğu için sınıf belirlemede S&P 500 borsa endeksinden yararlanılmıştır. Böylece, ilgili endekste kayıtlı olan şirketler bir sınıfa; diğerleri başka bir sınıfa ayrıştırılmak üzere ikili sınıflandırma hedeflenmiştir. Önceki aşamalarda elde edilen tüm veri setlerine sınıf etiketleri eklenmiştir. Bu aşamadan sonra, toplamda sekiz farklı veri seti, makine öğrenmesi algoritmalarıyla analiz edilebilir düzende elde edilmiştir. Bu aşamalarla elde edilen veri setleri farklı algoritmalarla test edilmiştir. Test sonuçları ve yorumları Bölüm 4' te verilmiştir.

3.2.3. Veri setlerinin iyileştirilmesi

Çalışmanın bu aşamasında, yıllık raporların gereksiz veya hatalı bilgilerden temizlenmesi ve sonraki analizlerde kullanılmak üzere çıkarımların yapılması için çeşitli teknikler uygulanmıştır. Yöntemlerin seçiminde ve uygulanmasında verilerin asıl anlamını yitirmemesi üzerinde durulmuştur. Özellikle metinsel verilerin sayısallaştırılması için ve veriler arasında bağlantılar kurmak için çalışmalar yürütülmüştür.

3.2.3.1. Şirketlerin yıllık raporlarının genişletilmesi

EDGAR veri tabanından çekilen veri, 2011 ile 2021 yılları arasını kapsayacak şekilde 10 yıllık şirket verileri ile genişletilmiştir. 2.248 şirkete ait toplamda 17.520 yıllık rapor EDGAR veri tabanından indirilmiştir. Kaç farklı yıla ait kaç farklı şirketin raporunun indirildiğinin sayıları Çizelge 3.6' da verilmiştir.

Çizelge 3.6 Her yıl için raporu olan şirket sayısı

Yıl Sayısı	Şirket Sayısı
1	819
2	659
3	102
4	87
5	111
6	80
7	76
8	93
9	104
10	117
TOPLAM	17520

Verilerin sayısının ve niteliğinin artırılması, şirketlerin daha iyi analiz edilebilmesi, yıllık değişimler üzerinden tahlillerin yapılması, sürenin uzamasıyla farklı ekonomik dalgalanmaların (krizler, seçimler vb. sebebiyle) şirket verilerine etkisinin en az indirilmesi gibi sebeplerle 10 yıllık verinin toplanması sağlanmıştır. Bazı şirketler, ilgili yıllık raporları bir yıl içinde birden fazla kez paylaştığı için (çeşitli güncellemeler yapmak için) ilgili yıla ait dosyaların son versiyonları hariç diğerleri silinmiştir. Önceki dönemde olduğu gibi yine veriler yıl, etiket ve değer şeklinde csv formatlı dosyalara indirilmiştir. Ayrıca şirket verilerindeki etiketler, analizlerde kullanılmak üzere kelimelere ayrıştırılmıştır. Ayırma işlemi öncesi ve sonrası etiket Şekil 3.7 ve 3.8' de verilmiştir. Kelimelerin ayrıştırılması için python wordninja[5] kütüphanesi ile NLTK[6] kütüphanelerinden faydalanılmıştır. Bu kütüphanelerdeki yöntemler değiştirilmiş ve kelime kıyaslamaları için ayrıca finansal kelimeler kütüphanelerin kelime havuzuna eklenmiştir.

[5] https://github.com/keredson/wordninja
[6] https://www.nltk.org/

```
abovemarketleaseamortizationexpenseafteryearfive
abovemarketleaseamortizationexpenseremainderoffiscalyear
abovemarketleaseamortizationexpenseyearfive
abovemarketleaseamortizationexpenseyearfour
abovemarketleaseamortizationexpenseyearthree
abovemarketleaseamortizationexpenseyeartwo
abovemarketleaseamortizationincomenexttwelvemonths
abovemarketleaseamortizationincomeremainderoffiscalyear
abovemarketleaseamortizationincomeyearfive
abovemarketleaseamortizationincomeyearfour
abovemarketleaseamortizationincomeyearthree
abovemarketleaseamortizationincomeyeartwo
abovemarketleaseamortizationrentalexpensedecreaseafterfifthfiscalyear
abovemarketleaseamortizationrentalexpensedecreaseyeartwo
abovemarketleasegross
abovemarketleasenet
abovemarketleasenetcurrent
abovemarketleases
abovemarketleasesnet
abovemarketleasesnetnoncurrent
abovemarketleasesnoncurrent
abovemarketrents
```

Şekil 3.7 2011 yılında raporlarda kullanılan etiketlerin ayrıştırma işlemi öncesi görünümü

```
above market lease amortization expense after year five
above market lease amortization expense remainder of fiscal year
above market lease amortization expense year five
above market lease amortization expense year four
above market lease amortization expense year three
above market lease amortization expense year two
above market lease amortization income next twelve months
above market lease amortization income remainder of fiscal year
above market lease amortization income year five
above market lease amortization income year four
above market lease amortization income year three
above market lease amortization income year two
above market lease amortization rental expense decrease after fifth fiscal year
above market lease amortization rental expense decrease year two
above market lease gross
above market lease net
above market lease net current
above market leases
above market leases net
above market leases net non current
above market leases non current
above market rents
```

Şekil 3.8 2011 yılında raporlardaki etiketlerin ayrıştırma işlemi sonrası görünümü

Tüm şirket verileri arasındaki bağımsız tüm etiketler de çıkarılmıştır. Toplamda 254.805 farklı etiketin firmalar tarafında ilgili 10 yıllık süreçte kullanıldığı tespit edilmiştir. Bu etiketlerin yarısından fazlası yalnızca birkaç firma tarafından kullanılırken

az bir kısmı firmaların büyük çoğunluğu tarafından kullanılmıştır. Şekil 3.9' da 2011 yılı için 843 firma tarafından en çok kullanılan etiketlerden bazıları verilmiştir. Önceki dönemde yapılan çalışmalardan olan kelime köklerinin bulunması işlemi de yine her firma raporundaki etiketler için gerçekleştirilmiştir. Alt bölümlerde yapılan diğer işlemler detaylı olarak anlatılmıştır.

```
dei entity public float;815
dei entity common stock shares outstanding;784
us-gaap liabilities and stockholders equity;770
us-gaap cash and cash equivalents at carrying value;749
us-gaap net cash provided by used in operating activities;749
us-gaap cash and cash equivalents period increase decrease;715
us-gaap property plant and equipment net;705
us-gaap income tax expense benefit;677
us-gaap assets current;654
us-gaap weighted average number of shares outstanding basic;629
us-gaap liabilities current;625
us-gaap weighted average number of diluted shares outstanding;602
us-gaap other assets non current;587
us-gaap interest expense;566
us-gaap operating income loss;555
us-gaap stockholders equity;550
us-gaap share based compensation;549
```

Şekil 3.9 2011 yılında raporlarda en sık kullanılan etiketler

3.2.3.2. Raporlarda en sık kullanılan kelimeler

Firmaların kullanımına sunulan veya firmalar tarafından oluşturulan etiketlerdeki kelimelerin ayrıştırılarak hangi kelimelerin ilgili yıllarda daha çok tercih edildiği tespit edilmiştir. Bağlaç veya sıfatların dışında özellikle finansal kelimeler üzerinden firmaların kullandığı etiketlerin ortak noktalarının tespit edilebilmesine katkı sunması hedeflenmiştir. Ayrıca, firmaların kelime kullanımlarının sonrasında analizler için oluşturulacak veri setlerinde niteleyici bir özellik olarak alınması da hedeflenmektedir. Çizelge 3.7' de 2011 yılında en çok kullanılan kelimeler ve kullanım sayıları verilmiştir. Buna ek olarak, on yıllık süreçte firmaların en çok tercih ettiği kelimeler de çıkarılmıştır. Apple firmasının en sık tercih ettiği bazı kelimeler Çizelge 3.8' de verilmiştir.

Çizelge 3.7 2011 yılında raporlarda en sık kullanılan kelimeler

Kelime	Kullanım Sayısı
tax	35529
income	34299
net	31692
current	28471
other	27436
assets	25986
benefit	21188
stock	20321
expense	20065
share	19714
based	19213
from	18621
loss	18589
deferred	17300
liabilities	16944
value	16768
non	16284
cash	15937
compensation	14696

Çizelge 3.8 Apple şirketinin raporlarında en sık kullandığı kelimeler

Kelime	Kullanım Sayısı
tax	1507
income	1032
of	791
and	736
share	720
based	690
expense	650
other	631
net	554
from	534
current	509
compensation	504
for	448
in	440
loss	432
assets	384
deferred	373
benefit	351
payments	330

3.2.3.3. Ortak kelimelere sahip etiketlerin çıkarılması

Firmaların etiketlerinde tercih ettikleri kelimeler üzerinden etiketlerin gruplandırılması amaçlanmıştır. Buna göre, etiketlerin kullanım amacına göre ayrıştırılması hedeflenmiştir. Örneğin, vergiler, gelirler, giderler gibi finansal tablolarda sıkça yer alan ve firmaların paylaşımlarının ana kalemlerini oluşturan değerlerin kullanıldığı farklı etiketlerin tespit edilmesi ve en genel paylaşımların bulunması ile paylaşılan finansal kalemlerin nasıl detaylandırıldığının çıkarılması hedeflenmiştir. Şekil 3.10' da Microsoft firmasına ait raporlarda içinde gider anlamına gelen 'expense' kelimesinin geçtiği etiketler, noktalı virgülle ayrılmış şekilde verilmiştir.

advertising expense;allocated share based compensation expense;costs and expenses;current federal tax expense benefit;current foreign tax expense benefit;current income tax expense benefit;current state and local tax expense benefit;deferred federal income tax expense benefit;deferred foreign income tax expense benefit;deferred income tax expense benefit;deferred state and local income tax expense benefit;deferred tax assets tax deferred expense compensation and benefits share based compensation cost;deferred tax assets tax deferred expense other;deferred tax assets tax deferred expense reserves and accruals impairment losses;deferred tax assets tax deferred expense reserves and accruals other;deferred tax assets tax deferred expense reserves and accruals restructuring charges;employee service share based compensation tax benefit from compensation expense;finite lived intangible assets amortization expense;finite lived intangible assets amortization expense after year five;finite lived intangible assets amortization expense next twelve months;finite lived intangible assets amortization expense year five;finite lived intangible assets amortization expense year four;finite lived intangible assets amortization expense year three;finite lived intangible assets amortization expense year two;future amortization expense after year five;future amortization expense year five;future amortization expense year four;future amortization expense year one;future amortization expense year three;future amortization expense year two;general and administrative expense;impairment integration and restructuring expenses;income tax expense benefit;integration and restructuring expenses;interest expense;lease and rental expense;non operating income expense;operating expenses;other non operating income expense;research and development expense;segment operating expense excluding impairment integration and restructuring;selling and marketing expense;tax cuts and jobs act of 2017 income tax expense benefit;tax cuts and jobs act of 2017 measurement period adjustment income tax expense benefit;unrecognized tax benefits income tax penalties and interest expense

Şekil 3.10 İçinde 'expense' kelimesi geçen etiketler

Sonrasında, ortak kelimeler içeren etiketler arasında bağlantıların tespit edilebilmesi için 2,3 ve 4 ortak kelimeye sahip etiketler gruplandırılmıştır. Böylece, özellikle en çok kullanılan kelimeler arasında, paylaşılan finansal kalemlerdeki kelimelerin veya kelime gruplarının artması ile farklı firmalar için farklı etiketler olsa da benzer anlam taşıyabilecek etiketlerin tespit edilebilmesi hedeflenmiştir. Google firması için ortak kullanılan 2,3 ve 4 kelimesi ortak etiketlerden bazıları sırası ile Şekil 3.11, 3.12 ve 3.13' te verilmiştir.

2 Kelimesi Ortak Bazı Etiketler

Birinci Grup
accounts payable current
accounts receivable net current
disposal group including discontinued operation accounts payable
increase decrease in accounts payable

İkinci Grup
share based payment award options outstanding number
share based compensation arrangement by share based payment award options vested and expected to vest exe rc
share based compensation arrangement by share based payment award options vested in period fair value 1
share based compensation excluding discontinued operations and cash settled awards
share based payment arrangement non cash expense including liabilities settled
stock repurchase program additional shares authorized to be repurchased
tax benefit from stock based award activity
tax cuts and jobs act of 2017 incomplete accounting change in tax rate provisional income tax expense benefit
tax withholding related to vesting of restricted stock units
unrecognized tax benefits decreases resulting from prior period tax positions
unrecognized tax benefits decreases resulting from settlements with taxing authorities
unrecognized tax benefits increases resulting from current period tax positions
unrecognized tax benefits increases resulting from prior period tax positions

Şekil 3.11 Google firmasına ait raporlardaki etiketlerden 2 kelimesi ortak olanlar

3 Kelimesi Ortak Bazı Etiketler
allocated share based compensation expense
deferred tax assets tax deferred expense compensation and benefits share based compensation cost
employee service share based compensation non vested awards total compensation cost not yet recognized
employee service share based compensation tax benefit from compensation expense
excess tax benefit from share based compensation operating activities
share based compensation
share based compensation arrangement by share based payment award equity instruments other than option sex pec
share based compensation arrangement by share based payment award equity instruments other than options fo
share based compensation arrangement by share based payment award equity instruments other than options gr
share based compensation arrangement by share based payment award equity instruments other than options no
share based compensation arrangement by share based payment award equity instruments other than options sh
share based compensation arrangement by share based payment award equity instruments other than options ve
share based compensation arrangement by share based payment award options exercisable intrinsic value 1
share based compensation arrangement by share based payment award options exercisable number

Şekil 3.12 Google firmasına ait raporlardaki etiketlerden 3 kelimesi ortak olanlar

4 Kelimesi Ortak Etiketler

Birinci Grup
available for sale securities current
available for sale securities debt maturities after five through ten years fair value
available for sale securities debt maturities after one through five years fair value
available for sale securities debt maturities after ten years fair value

İkinci Grup
derivative fair value of derivative asset
derivative fair value of derivative liability
derivative liability fair value gross asset
derivative liability fair value offset against collateral net of not subject to master netting arrangement

Şekil 3.13 Google firmasına ait raporlardaki etiketlerden 4 kelimesi ortak olanlar

3.2.3.4. Birbirini içeren etiketlerin çıkarılması

Etiketler incelendiğinde, bazı etiketlerin diğer etiketler içerisinde tekrar ettiği ve diğer etiketler tarafından detaylandırıldıkları görülmüştür. Bu yüzden, ilgili etiketi içeren diğer etiketlerin hem firma özelinde hem de tüm firmalar için listelenmesi amaçlanmıştır. Örneğin, 'us-gaap:assetacquisition' etiketini içeren 329 farklı etiket bulunmuştur. Çizelge 3.9' da tüm firmalar arasında bu etiketi içeren bazı etiketlerin listesi verilmiştir.

Çizelge 3.9 ‘assetacquisition’ etiketini içeren etiketlerden bazıları

asset acquisition accounts payable
asset acquisition accounts payable accrued liabilities and other liabilities
asset acquisition accounts payable
asset acquisition accounts payable accrued liabilities and other liabilities
asset acquisition accounts receivable and other assets
asset acquisition accounts receivable net
asset acquisition accrued expenses and other liabilities
asset acquisition acquire doff market leases
asset acquisition acquired intangible assets amount
asset acquisition acquired patterns
asset acquisition acquired property and equipment amount
asset acquisition acquisition costs capitalized
asset acquisition acquisition related costs capitalized
asset acquisition additional consideration
asset acquisition additional consideration if earned
asset acquisition adjustments accounts payable
asset acquisition adjustments accounts receivable net

Özellikle ilgili etiket ile başlayan diğer etiketlerin, bu etiketin alt açıklamaları olduğu tespit edilmiştir. Böylece, firmaların kullanmış olduğu ortak etiket sayısı düşük olduğu için, bu etiketler arasında firmalarla bağlantı kurulması hedeflenmiştir. Çizelge 3.10’ da diğer etiketlerin içinde en çok yer alan etiketlerden bazıları verilmiştir.

Çizelge 3.10 Etiketler ve etiketi içeren diğer etiketlerin sayısı

Etiket	Geçme Sayısı
deferred	9394
business combination	6063
other	5864
share based compensation	5175
debt	4226
deferred tax asset	4203
share based compensation arrangement by share based payment	4171
income	3943
deferred tax assets	3696
fair value	3258
operating	3039
income tax	2807
equity	2646
stock issued	2485
cash	2472
other comprehensive income	2451
business combination recognized identifiable assets	2284
business combination recognized identifiable assets acquired	2261
total	2258
business combination recognized identifiable assets acquired and liabilities assumed	2140

3.2.3.5. Etiketlerdeki anlamsız kelimelerin tespit edilmesi

Finansal raporların okunabilirliği şirketlere yapılacak olan yatırımları veya şirkete olan bakış açısını birinci dereceden ilgilendirmektedir (Mcdonald, 2009). Bu yüzden, firmaların raporlarında kullandığı cümle ve kelime seçimlerinin dikkate alınması gerekmektedir. Firmaların kullanmış olduğu etiketlerde İngilizce' de herhangi bir anlama gelmeyen kelimelerin tespit edilmesi hedeflenmiştir. Bunun için, Doğal Dil İşleme çalışmalarında kullanılan İngilizce kelime veya özel ifadeler ile Amerikan borsalarından biri olan NASDAQ' dan da alınan finansal kelimelerin birleştirilmesi ile oluşturulan dört yüz binden fazla kelime ile etiketlerdeki kelimeler karşılaştırılmıştır. NASDAQ' dan alınan kelimelerde finansal metinlerde yer alan kısaltmalar da mevcuttur. Kıyaslama sonucu yaklaşık bin kelimenin, oluşturulan finansal veri setinde olmadığı tespit edilmiştir. Kelimelerden bazıları aşağıda listelenmiştir.

- recieved
- organisations
- nederlandsch

- measurment
- maturi
- forres
- definately
- authoritys

Yanlış kullanılan kelimelerin de hem rapor kalitesini düşüreceği hem de firmaların kullanmış olduğu ortak etiketlerin farklı etiketmiş gibi algılanmasına yol açmasıyla analizi güçleştireceği bir gerçektir. Bu yüzden, analiz için oluşturulacak finansal sözlük veri setine firmaların kullandığı anlamsız kelime sayılarının da eklenmesi düşünülmüştür.

3.2.3.6. En sık kullanılan kelimeler üzerinden k means clustering yöntemi ile etiketlerin kümelenmesi

Etiketleri oluşturmak için kullanılan kelimeler arasında en sık kullanılan kelimelerin çok fazla kullanılması, bu kelimelerin yine birlikte çok defa tekrarlanması, etiketlerin gruplandırılmasında olumlu sonuç üretebileceği düşünülmüştür. Bunun için etiketlere K Means Kümeleme yöntemi uygulanmıştır. Öncelikle, tüm şirketlerde kullanılan eşsiz etiketler çıkarılmıştır. Ardından bu etiketler üzerinden TF-IDF vektörü oluşturulmuştur. Sonraki adımda, algoritma birlikte sık kullanılan kelimeleri küme merkezi olarak seçerek etiketleri gruplandırmıştır. Küme merkezlerinde 5 ve 10 kelime seçilerek yine 5 ve 10 farklı kümeye ayrıştırılması sağlanarak toplamda 4 farklı sonuç üretilmiştir. Etiketler önce karıştırılarak (Shuffle) %20'si test, kalanı eğitimde kullanılacak şekilde ayrılmıştır. Toplamda 100 tekrar çalıştırılmıştır. Farklı seçimlerle oluşturulan kümelerdeki etiket sayıları karşılaştırması Çizelge 3.11' de verilmiştir. Küme merkezleri olarak seçilen kelimeler de aşağıda verilmiştir.

5 Küme, 5 Kelime:

1. debt-term-long-instrument-maturity
2. tax-deferred-income-asset-benefit
3. loss-lease-income-gain-operating
4. asset-number-cost-value-liability
   stock-share-based-compensation-common

5 küme, 10 Kelime:

1. asset-cost-number-debt-payment-value-lease-liability-business-non

2. based-share-compensation-award-arrangement-payment-option-hare-stock-vested
3. stock-common-issued-hare-period-preferred-value-issuance-option-share
4. loss-gain-income-comprehensive-net-unrealized-security-tax-investment-operating
5. tax-deferred-income-asset-benefit-liability-expense-reconciliation-unrecognized-current

10 Küme, 5 kelime

1. lease-number-debt-payment-cost
2. discontinued-operation-disposal-group-including
3. tax-income-benefit-reconciliation-expense
4. business-acquisition-combination-consideration-price
5. investment-realest-ate-method-equity
6. stock-common-issued-hare-period
7. plan-benefit-defined-pension-contribution
8. asset-deferred-tax-liability-intangible
9. loss-security-gain-available-sale
10. based-share-compensation-award-arrangement

10 Küme, 10 Kelime

1. debt-term-long-instrument-maturity-repayment-principal-year-security-issuance
2. stock-payment-lease-expense-asset-cash-loan-investment-net-non
3. loss-gain-income-net-comprehensive-unrealized-security-investment-tax-operating
4. business-combination-acquisition-acquired-assumed-asset-identifiable-liability-recognized-consideration
5. number-ofs-hare-property-stock-contingency-unit-agreement-employee-facility
6. cost-current-non-accrued-liability-deferred-asset-expense-related-net
7. based-share-compensation-award-arrangement-payment-option-hare-vested-stock
8. discontinued-operation-disposal-group-including-asset-current-income-expense-net
9. value-fair-stock-disclosure-asset-security-issued-air-recurring-basis
10. tax-deferred-income-asset-benefit-liability-expense-reconciliation-unrecognized foreign

Çizelge 3.11 Seçilen küme ve kelime sayılarına göre küme içerisindeki etiket sayıları

	5 Kelime	**10 Kelime**
5 Küme	133121	151441
	23577	17177
	18478	16662
	16976	12962
	11604	5514
10 Küme	118444	105269
	14990	15602
	12590	14927
	12590	14311
	11236	13069
	10359	12768
	9534	9750
	5443	9091
	5177	5639
	3393	3330

Çizelge 3.11' de yer alan kümelerden özellikle çok sayıda etiket içeren kümelerdeki kelimelerin ve küme ayrıştırmalarının sonraki analizlerde girdi olarak kullanılması düşünülmektedir. Farklı küme ve kelime sayılarında aynı gruplarda yer alan zıt anlamlı ve benzer anlamlı kelimeler de yine analizlerde kullanılacaktır. Örneğin 'asset' ve 'liability' kelimeleri varlık ve borç anlamı taşırken aynı kümelerde yer almıştır ve yine vergi anlamına gelen 'tax' ile 'liability' aynı kümelerde yer almıştır. Küme merkezinde kelime seçimlerinde bağlaç, sıfat gibi değer ifade etmeyen kelimelerin yer almamasının da yine kümelemenin sağlığı açısından önemli olduğu düşünülmektedir.

3.3. Doc2Vec-K Means-CNN Hibrit Yöntemi ile Fiyat Gücü Tahmini

Finansal raporların kullanımı, verilerin dijitalleşmesiyle birlikte son yıllarda önemli ölçüde artmıştır. Bununla birlikte, geleneksel istatistiksel yöntemler, ham verinin kontrolsüz genişlemesi ve karmaşıklığı nedeniyle artık işe yaramamaktadır. Bu nedenle, finansal verilerin temizlenmesi ve analiz edilmesi için modern makine öğrenimi yöntemlerinin kullanılması önem taşımaktadır. Bu uygulamada, ABD'deki halka açık şirketlerin üç aylık finansal raporları (yani 10Q bildirimleri), veri madenciliği yöntemleri kullanılarak analiz edilmiştir. Çalışmada, 2019-2022 yılları arasında şirketlerin 8.905 adet üç aylık raporu kullanılmıştır. Önerilen yöntem, üç farklı makine öğrenimi yönteminin kombinasyonuyla iki aşamadan oluşmaktadır. İlk iki yöntem, 10Q bildirimlerinden yeni özellikler çıkararak bir veri kümesi oluşturmak için kullanılmış, son yöntem ise

sınıflandırma problemi için kullanılmıştır. 10Q bildirimlerindeki metin etiketlerinden vektörler oluşturmak için Gensim çalışma ortamındaki Doc2Vec yöntemi kullanılmıştır. Oluşturulan vektörlere, K Means algoritması kullanılarak etiketleri anlamsal olarak birleştirmek üzere kümeleme işlemi gerçekleştirilmiştir. Bu şekilde, farklı finansal unsurları temsil eden 94.000 etiket, bu etiketlerden oluşan 20.000 kümeye dönüştürülmüş ve analiz daha verimli ve yönetilebilir hale getirilmiştir. Veri kümesi, kümelerdeki etiketlere karşılık gelen değerlerle oluşturulmuştur. Ayrıca, sınıf etiketi olarak, bir sonraki finansal çeyrek için şirketlerin geçmiş dönemdeki finansal piyasa fiyatı gücünü gösteren Fiyat Gücü (PriceRank) (SEC, 2022) metriği veri kümesine eklenmiştir. Böylece, bir şirketin üç aylık raporlarının şirketin piyasa fiyatı üzerindeki etkisinin belirlenmesi amaçlanmıştır. Son olarak, sınıflandırma problemi için Evrişimsel Sinir Ağı modeli kullanılmıştır. Sonuçları değerlendirmek için önerilen hibrit yöntemin tüm aşamaları Terim Frekansı-Ters Doküman Frekansı (Term Frquency-Inverse Document Frequency, TFIDF) (Sparck Jones, 1972) , Bert Temelli Nli Jetonları (Bert Base Nli Tokens) (Reimers ve a Gurevych, 2019), DBSCAN (Ester ve ark., 1996), Geriye Yayılım Sinir Ağı (Back Propagation Neural Network), Karar Ağacı, K En Yakın Komşu ve İkinci Derece Ayırma Analizi (Quadratic Discriminant Analysis, QDA) (Stone ve ark., 2004) teknikleriyle karşılaştırılmıştır. Bu yeni yaklaşım, yatırımcılara şirketleri bir arada incelemelerine ve yeni, önemli bilgiler çıkarmalarına yardımcı olmasının yanı sıra, sonraki çeyrek döneminde raporu paylaşılan şirketin finansal piyasalardaki fiyat gücüne dair de bilgi çıkarımı yapılmasına olanak sağlaması hedeflenmiştir. Alt bölümlerde, çalışmada kullanılan veri ve önerilen hibrit yöntem detaylıca açıklanmıştır.

3.3.1. Verilerin toplanması

Bu çalışmada kullanılan veriler, SEC’ den elde edilen finansal raporlardan toplanmıştır. ABD'de, halka açık şirketlerin standart, doğrulanmış formlarda ve belirli zamanlarda SEC' ye faaliyetlerine ve yapısal durumuna dair bilgilerini paylaşma zorunluluğuna çalışmanın bu bölümünde kullanılan 10Q çeyrek raporları da dahildir. Herhangi bir kişi, SEC' nin EDGAR veri tabanından paylaşılan tüm bildirimlere ve formlara erişerek kullanma hakkına sahiptir.

3.3.2. 10Q formları

10Q, şirketlerin üç aylık raporlarının içeriğini ve yükümlülüklerini belirleyen bir bildirim standardıdır. Hem sayısal hem de metinsel bilgileri içermektedir. Şirketler, XBRL de dahil olmak üzere çeşitli formatlarda 10Q formlarını erişime sunmaktadır. XBRL, finansal bilgilerin tüm ilgili taraflar için analiz ve takasını mümkün kılmaktadır. Tez çalışmasının bu bölümünde, her yıl SEC (SEC, 2021) tarafından üç aylık olarak özetlenen 10Q bildirimlerinin sayısal kısımlarını incelenmiştir. 10K raporları ile aynı standartlara ve yazım düzenine sahip olan bu raporların temel farkı, her üç ayda bir olmak üzere yılda dört defa paylaşılıp güncellenerek, daha kısa bir zaman dilimini kapsayan bilgiler içermesidir. Çalışmada, SEC tarafından sunulan, şirketlerin üç aylık raporlarından (10Q) oluşan veri grupları kullanılmıştır. Her üç aylık grup, dört csv formatlı dosyadan oluşmaktadır. Dosyaların isimleri aşağıda listelenmiştir.

- Num
- Pre
- Sub
- Tag

3.3.2.1. Num dosyası

Her bir 'num' dosyası, şirketlerin ait olduğu çeyrekte paylaştığı tüm finansal kalemlerin bilgilerini içermektedir. Her bir satır, şirketlerin paylaştığı etiketlerin 'adsh', 'tag', 'version', 'ddate', 'qtrs', 'value' başlıklarına karşılık gelen bilgilerini içermektedir. 'adsh', SEC tarafından, verilerini paylaşan şirkete verilen etiketlenmiş verinin erişim numarasını, 'tag' mali kalemin adını, 'version' paylaşılan etiketin sürümünü, 'ddate' şirketin bilgiyi paylaştığı günü, ' qtrs', çeyrek sayısı olarak ilgili finansal yılın hangi çeyreğine ait bilginin paylaşıldığını belirtmek için verilen ve 1-4 arasında değer alan bir rakamı, ‘value’ ise ilgili şirkete ait paylaşılan etikete dair değeri göstermektedir. Bu çalışmada dosyaların ‘adsh’, 'tag', 'value' bölümleri kullanılmıştır. Örneğin erişim numarası ‘000008858-21-000048’ olan ‘NetIncomeLoss’ etiketinin dolar bazlı değeri ‘-18889000’ bilgisi, bu veriyi paylaşan şirketle ilişkilendirilerek kullanılmıştır.

3.3.2.2. Pre dosyası

'Pre' dosyaları, ilgili veri kümesinin sunumunu ifade etmektedir. Bu dosyada mali tablolarda yer alan etiket ve numaraların nasıl sunulduğu belirtilmiştir. Bu dosyalar çalışmada kullanılmamıştır.

3.3.2.3. Sub dosyası

Firmaların XBRL gönderim bilgileri 'Sub' dosyalarında verilmektedir. Bilgilerini paylaşan şirketler hakkında detaylı bilgiler içermektedir. Örneğin, şirketin erişim numarası, adı, adresi, sunum türü ve merkezi dizin anahtarı numaraları (CIK) alt dosyalarda yer almaktadır. CIK numaraları, SEC tarafından halka açık şirketlere özel olarak verilen kimlik numaralarıdır. Bu çalışmada, erişim numaraları ayrı çeyrek veri setlerinde tekrarlanabileceğinden, şirketlerin mali tablolarını ayrıştırmak için CIK numaraları kullanılmıştır. Örneğin, '000008858-21-000048' numaralı paylaşımı yapan şirketin CIK numarasının '8858' olduğu bilgisi Sub ve Num dosyaları kullanılarak ortaya çıkarılmıştır.

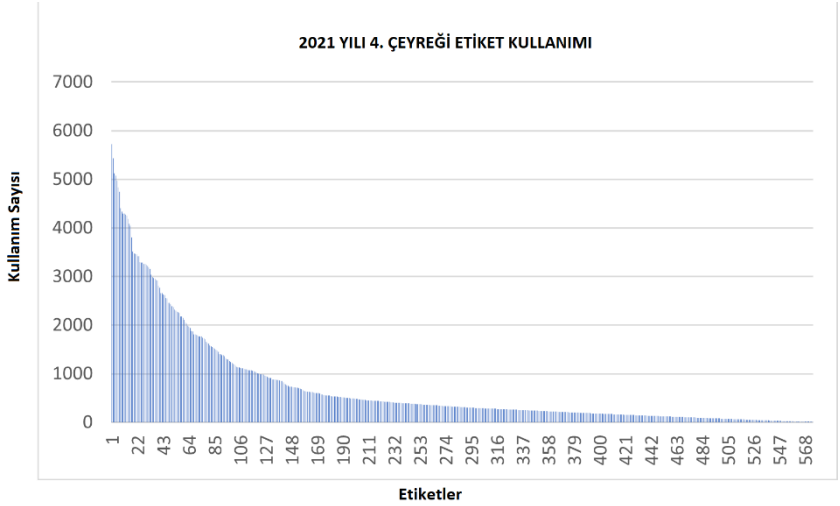
3.3.2.4. Tag dosyası

'Tag' dosyaları, finansal paylaşımlarda kullanılması gereken standartlaştırılmış (çoğunlukla GAAP tarafından) tüm etiketleri içerir. Etiketlerin adlarını, oluştuğu kelimeleri ve açıklamalarını içerir. Çalışmada etiketlerin gruplandırılmasında bu dosyanın içeriğinden yararlanılmıştır.

3.3.3. Veri seti ön işlemleri

Bu bölümde, önceki bölümde anlatılan ham verilerden elde edilen veri setinin oluşturma süreci açıklanmıştır. Çok sayıda hata içeren, eksik bilgi barındıran ve en önemli problemi seyreklik olan verinin, uygulanan temizleme ve ön işleme teknikleriyle makine öğrenmesi yöntemlerinde verimli bir şekilde çalıştırılabilecek veri setine dönüştürülme süreci açıklanmıştır. Birçoğu yalnızca bir veya iki kez kullanılan etiketler 10K ve 10Q raporları üzerinde anlamlı analizler yapmanın önünde duran en önemli zorlukların başında gelmektedir. Şekil 3.14' te 2021 yılı dördüncü çeyrek raporlarında kullanılan etiketlerin kullanım dağılım grafiği verilmiştir. Grafikten de görüldüğü üzere etiketlerin çok büyük çoğunluğu çok az şirket tarafından kullanılmıştır. Bu yüzden, şirketlerin toplu olarak analiz edilebilmesi için etiketlerin ve etiketlere karşılık gelen değerlerin kullanımı

oldukça zorlaşmaktadır. Veri üzerinde gerçekleştirilen işlemlerle problemlerin en aza indirgenmesi hedeflenmiştir.



Şekil 3.14 2021 yılının dördüncü çeyreğinde etiketlerin farklı şirketler tarafından kullanım sayıları

3.3.4. 10Q verilerinin temizlenmesi

Çalışmanın bu adımında toplu 10Q dosyaları kullanılarak iki veri seti oluşturulmuştur. Veri setlerinde satırlar bir şirketin çeyrek dönem paylaşımlarına karşılık gelmektedir. Her paylaşım, şirketin erişim numarası ve CIK değerleri iki ayrı dosyada eşleştirilerek benzersiz hale getirilmiştir. Daha sonra verilerdeki eksik veya hatalı bilgilerin giderilmesi için veri temizleme işlemleri yapılmıştır. Yinelenen girişler, eksik CIK, etiket veya karşılık gelen değere sahip girişler de veri setinden kaldırılmıştır. Temizleme işlemlerinden sonraki verilerin bir örneği Çizelge 3.12' de yer almaktadır.

Çizelge 3.12 Veri temizleme işlemleri sonrası elde edilen verilerden bir görünüm

adsh	tag	value	cik
0001654954-19-012972	Goodwill	17203000	839087
0001654954-19-012906	ShareBasedCompensation	87033	314227
0001558370-19-009160	ShareBasedCompensation	46400000	820313
0001213900-19-020268	NonoperatingIncomeExpense	-1834	747540
0001424929-19-000078	CommonStockSharesOutstanding	38524	1424929
0001101215-19-000224	DepositsFairValueDisclosure	1,3692E+10	1101215
0001711269-19-000063	ProceedsFromPaymentsForOtherFinancingActivities	-5300000	1711269
0001505155-19-000079	GainsLossesOnExtinguishmentOfDebt	-2317000	1505155
0001409970-19-001368	IncomeTaxExpenseBenefit	-341000	1409970
0001493152-19-019758	ProfitLoss	-2348829	1701756

3.3.5. Veri seti oluşturma

Bu bölümde, önceki aşamalarda anlatılan ham verinin makine öğrenmesi yöntemleriyle çalıştırılmasına uygun biçimde veri setlerine dönüşümü açıklanmıştır. Bunun için, 2019' dan 2022' ye kadar paylaşılan toplam 12 toplu çeyrek 10Q formu kullanılmıştır. Nihai veri setinde, veri kümesinin her bir satırı, bir şirketin üç aylık dönemdeki açıklamasını temsil etmiştir. Özellik vektörü olarak 2019 yılının dördüncü çeyreğinde kullanılan etiketler seçilmiştir. Bu nedenle veri setinde aynı şirket tarafından farklı çeyreklerde açıklanan finansal tablolar farklı örneklem olarak seçilmiştir. Veri setinde toplam 8905 farklı şirkete ait 10Q raporları yer almıştır. Veri seti, farklı çeyrekler için paylaşılan raporlar birleştirildikten sonra 34686 satırdan oluşmuştur. Özellik sayısı ise, önerilen yöntemdeki özellik seçimi adımlarının uygulanmasıyla veri setlerinde farklılık göstermiştir.

3.3.6. Özellikler

Finansal kalemlerin 10Q formlarında etiket değeri formatında paylaşılması, veri seti yaratımında önemli avantajlar sunar. Örneğin, etiket-değer paylaşım yapısı, bir şirketin çeyrek finansal dönemdeki finansal bir kaleminin bir özellik olarak kullanılmasına imkân sağlamaktadır. Ancak, şirketler kurumlar tarafından önerilen etiketleri kullanmayı daha az tercih ettiği için, şirketlerin etiket kullanımı oldukça özelleşerek farklı şirketler arasındaki aynı finansal kalem üzerindeki bağlamı ortaya çıkarmayı zorlaştırmaktadır. Bu nedenle şirketlerin paylaştığı ortak etiketlerin yüzdesi, tüm etiketler kullanıldığında düşük kalmaktadır. Bu da şirketler arasında paylaşılan ortak

verilerin çok düşük olmasına neden olmaktadır. Örneğin, Çizelge 3.13, farklı şirketler tarafından ‘Stock Issuance’ (hisse senedi ihracı) ile ilgili olarak kullanılan etiketleri ve bunların SEC üzerindeki kelimelere ayrıştırılmış versiyonlarını göstermektedir. 2019 dördüncü çeyrek verilerine bakıldığında 5767 farklı firma toplam 94000 farklı etiket kullanmıştır. Ancak etiketlerin 80312 adedi sadece bir firma tarafından kullanılmıştır. Şirketlerin yüzde 99' u ile en sık kullanılan etiket 'Assets' (varlıklar) olduğu belirlenmiştir. 2021 4. Çeyrek' te iki veya daha fazla şirket tarafından kullanılan etiketlerin kullanım sıklığı Şekil 3.14' de gösterilmektedir. Buna göre x ekseni etiketleri gösterirken, y ekseni kaç farklı şirketin kullanıldığını göstermektedir. Veri seti oluşturulurken, verilerin oldukça seyrek olması nedeniyle, etiketlerin doğrudan kullanımının makine öğrenimi araçlarını kullanarak şirketleri analiz etmek için tek başına yeterli olmadığı, sağlıklı sonuçlar üretmediği gözlemlenerek, bu problemin giderilmesi hedeflenmiştir.

Çizelge 3.13 Raporlarda kullanılan etiketler ve ayrıştırılmış versiyonları

Etiket	Tokenize Hali
StockIssuanceCost	Stock Issuance Cost
StockIssuanceCostAccrued	stock issuance cost accrue
StockIssuanceCosts	Stock issuance cost
StockIssuanceCostsAccruedAndAmortized	Stock Issuance Costs accrue and amortize
StockIssuanceCostsAndDiscounts	Stock Issuance Costs and discount
StockIssuanceCostsIncurred	Stock Issuance Costs incurred
StockIssuanceCostsIncurredButNotYetPaid	Stock Issuance Costs Incurred but not yet pay
StockIssuanceCostsNotPaid	Stock Issuance Costs not pay
StockIssuanceCostsPaidInKind	Stock Issuance Costs Paid in kind
StockIssuanceCostsNotPaid	Stock Issuance Costs not pay
StockIssuanceCostsPaidInKind	Stock Issuance Costs Paid in kind

3.3.7. Sınıflandırma metriği

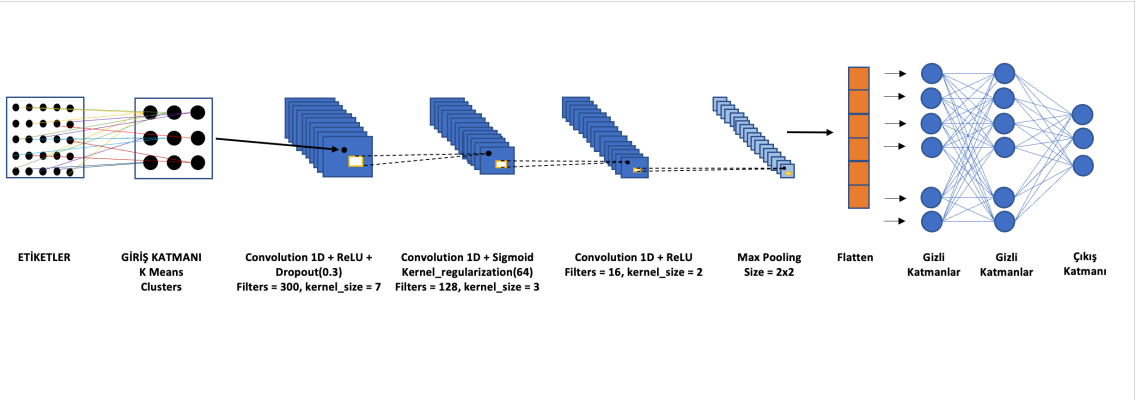
Çalışmada önerilen algoritmalarda oluşturulan veri setlerini test etmek için SEC tarafından paylaşılan 'PriceRank' metrik değerleri kullanılmıştır. Metrik, SEC tarafından bir şirkete verilen, piyasa fiyatı oynaklığı seviyesi ve finansal piyasalardaki işlem hacmi gibi faktörleri hesaba katan bir derecelendirme olarak tanımlanabilir. Dolayısıyla, 'PriceRank', bir şirketin son mali çeyrek içerisindeki mali durumuna göre piyasa fiyat gücünü gösteren, 1 ile 10 arasında değer alabilen bir göstergedir. Buna göre 10, fiyat gücü en yüksek firmaları, 1 ise piyasa fiyatı çok riskli ve zayıf olan firmaları göstermektedir. Bu çalışmada sınıflandırma ölçütleri olarak düşük, orta ve yüksek olmak üzere 3 sınıf

oluşturulmuştur. Paylaşılan verilere göre üç sınıf, alt sınıfta 1,2,3, orta sınıfta 4,5,6,7 ve üst sınıfta 8,9,10 değerlerine sahip firmalar olacak şekilde oluşturulmuştur. 2019-2022 yılları arasındaki 12 çeyrekte yapılan tüm çeyrek paylaşımlarına şirket bazında sınıf bilgisi eklenmiştir.

3.3.8. Hibrit model

Firmaların finansal raporlamalarında aynı finansal kalemleri farklı etiketlerle kullanması ve finansal kalemin detaylarına göre etiketlerin özelleştirilebilmesi, 10Q dosyalamalarından veri seti oluştururken sorun yaratmaktadır. Bu nedenle bu çalışmadaki amaç, etiketleri anlamlarına göre kümelemek ve veri seti oluşturulurken etiketler yerine grup tabanlı kümeler kullanmaktır. Bu sayede şirketlerin raporları arasındaki ortaklıklar arttırılması ve veri seyrekliği anlamlı bir şekilde azaltılması hedeflenmiştir. Böylece, Doc2Vec ve K Means kümeleme yöntemleri kullanılarak, önceki adımlarda yaratılan veri setlerindeki özelliklerden yeni özellikler anlamlı bir şekilde yaratılarak analizlerin verimli hale gelmesi amaçlanmıştır. Önerilen yöntemlerin uygulandığı, farklı yöntemlerin uygulandığı ve herhangi bir özellik seçiminin uygulanmadığı veri setleri, hibrit algoritmanın ikinci kısmını oluşturan CNN ağı da dahil olmak üzere farklı algoritmalarla test edilerek sonuçların anlam kazanması düşünülmüştür.

Önerilen yöntemde, öncelikle etiketler ve açıklamalarının yer ‘Tag’ dosyalarındaki açıklamalar, Doc2Vec yöntemi kullanılarak vektörize edilmiştir. Daha sonra, üretilen vektörler, girdiler için bir özellik vektörü oluşturmak üzere K Means kümeleme yöntemiyle kümelenmiştir. Elde edilen vektöre göre girdiler belirlenerek, firmaları fiyat güçlerine göre sınıflandırmak için oluşturulan CNN'de eğitim gerçekleştirilmiştir. Önerilen yöntemin mimarisi Şekil 3.15' te gösterilmiştir. Model daha önce görmediği verilerle test edilmiş ve sonuçlar elde edilmiştir. Önerilen yöntemin performansını değerlendirmek için, aynı CNN modelinde ve farklı makine öğrenme yöntemlerinde etiketlerin özellik vektörleri olarak seçildiği veri seti de eğitilmiş ve test edilmiştir. Ayrıntılı sonuçlar, farklı metriklerle karşılaştırmalı olarak 4. bölümde verilmiştir.



Şekil 3.15 Önerilen hibrit algoritmanın mimari tasarımı

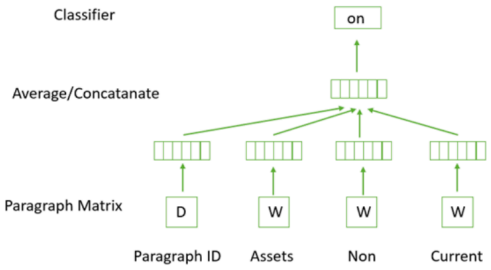
3.3.8.1. Doc2Vec

Etiket analizi için 2021'in dördüncü çeyreği veri grubuna ait tag.txt dosyası kullanılmıştır. Dosya, etiketlerin orijinal, simgeleştirilmiş sürümünü ve açıklamalarını içermektedir. Dosyadan bazı satırlar 3.14 çizelgesinde örnek olarak listelenmiştir. Daha sonra etiketlerin ve açıklama sütunları kelimelere ayrıştırılmış ve tüm harfler küçük harfe dönüştürülerek cümlelerdeki noktalama işaretleri kaldırılmıştır. Ardından, dosya Spacy[7] yöntemi kullanılarak kelime köküne indirgeme (lemmatization) işlemi gerçekleştirilmiştir. Bu adımlardan sonra açıklaması olmayan etiketler veya hatalı satırlar silinerek toplam 120412 olan etiket sayısı 67862' ye düşürülerek etiket anlam benzerliği en üst düzeye çıkarılmanın yanı sıra etiket belirsizliğinin de ortadan kaldırılması sağlanmıştır.

Daha sonra birden çok cümleden oluşan etiketlerin açıklamaları seçilerek etiketlerin kümelenmesi sağlanmış ve Doc2Vec kullanılarak vektöre dönüştürülmüştür. Gensim, kelimeleri (word2vec) veya paragrafları (Doc2vec) başarıyla vektörlere dönüştürebilen çok popüler bir denetimsiz öğrenme çerçevesidir. Özellikle, küçük veri kümeleri için standart kelime torbası (BOW) veya n-gram tekniklerinden daha iyi çalıştığı kanıtlanmıştır (Haider ve ark. Mahi, 2020). Bunun nedeni, kelimeler standart Bag of Words modeli kullanılarak vektörlere dönüştürüldüğünde, kelime sırası ile ilgili tüm bilgilerin kaybolmasıdır. Örneğin, “Liabilities and Stockholder’ Equity” ve “Equity Liabilities and Stockholder” cümleleri, sırası ile “Borçlar ve Hissedar Özkaynakları” ve “Öz Sermaye Yükümlülükleri ve Hissedarlar” anlamlarına sahip olmalarına rağmen aynı vektöre sahip olacaktır. Bunun üstesinden gelmek için Bag of n-gram yöntemi

[7] https://spacy.io/

kullanılabilir, ancak bu yöntem de sonrasında veri seyrekliği ve yüksek boyutluluk sorunlarının ortaya çıkmasına neden olmaktadır. Çalışmada Doc2vec yöntemi kullanılmıştır. Tekniğin mimarisi Şekil 3.16' da verilmiştir. Buna göre paragraf matrisinin adımında D harfi ile gösterilen sütunda tüm paragraflar benzersiz vektörlere eşlenir ve W ile belirtilen sütunlardaki her kelime de benzersiz vektöre eşlenir. Kısaca modelin daha önce görmediği paragrafı vektöre dönüştürebilmesi için, paragraf ve kelime vektörlerinin parametreleri ile çıktı için kullanılacak softmax algoritması parametrelerinin hesaplanması gerekmektedir. Bunun için Stochastic Gradient Descent (Mendelson ve ark. Smola, 2003) yöntemi kullanılmıştır.



Şekil 3.16 Doc2Vec yönteminin çalışma mantığı

Çizelge 3.14 Tags dosyasından etiket, ayrıştırılmış etiket ve etiket açıklamasına dair bir görünüm

Etiket	Tokenize Etiket	Etiketin Açıklaması
RedemptionsOfNotesPayable	redemption of Notes payable	redemption of Notes payable
UtilitiesOperatingExpensePurchasedPowerFromRe	utility operating Expense Purchased	the amount of purchase power from relate
latedParties	Power from Related party	party charge against earning for the period
StockUnitsIssuedDuringPeriodSharesNewIssues	Stock Units issue during Period Shares New	number of new stock unit issue during the
	Issues	period
AdjustmentToRedemptionValue	adjustment to Redemption Value	adjustment to redemption value
SeriesaConvertiblePreferredStockMember	Seriesa Convertible Preferred Stock Member	Represents information relate to Series A
		Convertible Preferred stock
NonAffiliatesMember	Non Affiliates Member	Non affiliate
ProvisionForInvestmentRelatedCreditLossBenefit Ex-	provision for Investment Related Credit Loss	provision for investment relate credit loss
pense	Benefit Expense	benefit expense
CurrentNonrecourseFinancialLiabilitiesOfVari-	current nonrecourse financial liability of	current nonrecourse financial liability of
ableInterestEntities	variable interest entity	variable interest entity
AdjustmentsToAdditionalPaidInCapitalDeferredTax-	adjustment to additional Paid in Capital De-	adjustment to additional Paid in Capital De-
Asset	ferred Tax Asset	ferred Tax Asset
IncreaseDecreaseInBrokerDealerServicingFees-	increasedecreaseinbrokerdealerservicing-	the amount of increase decrease in broker
PayableGeneralPartner	feespayablegeneralpartn	dealer servicing fee payable general part- ner
IncreaseDecreaseInSalesTaxAndNetValueAddedTax	increase Decrease in Sales Tax and Net Value	increase Decrease in Sales Tax and Net Value
	Added tax	Added tax
FivePointFivePercentSeniorNotesDueTwenty Twen-	five Point five Percent Senior Notes due	five point five percent senior note due
tyFourMember	Twenty Twenty Four Member	twenty twenty four
GainLossOnInvestmentsAndDividends	gain Loss on Investments and dividend	gain loss on investment and dividend

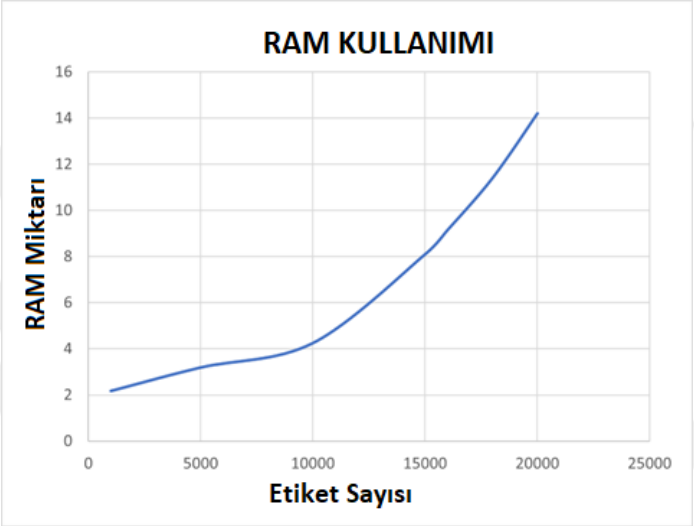
3.3.8.2. K means kümeleme yöntemi

XBRL formatlı finansal verilerdeki etiket-değer çiftleri, makine öğrenimi yöntemlerine uygun veri setlerini kolayca oluşturmak için kullanılabilir. Ancak, şirkete göre özelleşen ve ayrıntılanan çok sayıda etiket, analizi çok zorlaştırmaktadır. Bu çalışmada önerilen yöntem eğitilirken, farklı sayıda etikette oluşturulan veri kümelerinin analizine bağlı RAM kullanımı 3.17' de gösterilmiştir. 2019 yılının dördüncü çeyreğindeki tüm etiketlerle üretilen bir veri setinin çalıştırılması için gereken RAM 200 GB olacağı hesaplanmıştır. Bu nedenle, çalışmanın bu bölümünde, etiketleri önemlerine göre birleştirerek finansal veri analizi için özellik olarak kullanılması hedeflenmiştir.

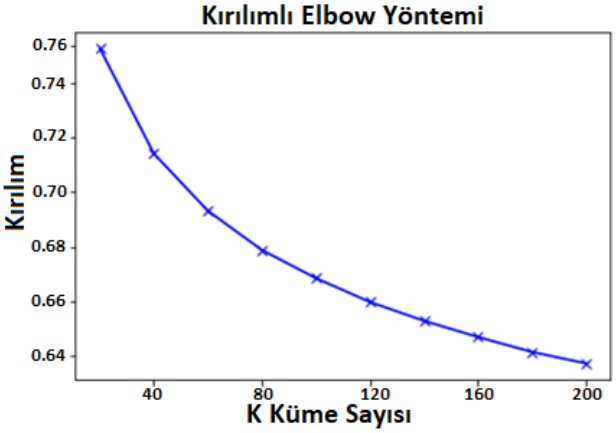
Doğru küme sayısını tahmin etmek için kullanılan Elbow tekniği (Bholowalia ve ark. Kumar, 2014), bir önceki adımda oluşturulan vektörler kullanılarak gerçekleştirilmiştir. Girdilerin varyans yüzdelerine bağlı olarak küme sayısının etkinliğini tahmin etmek için bu yöntem tercih edilmiştir. Varyans için yanlılıkların kırılma noktası, doğru küme sayısını vermektedir. Etiketlerin açıklamaları için küme sayısının bir

fonksiyonu olarak yanlılığın varyasyon grafiği Şekil 3.18' de gösterilmektedir. Böylece, küme sayısı girdi sayısının yaklaşık %30' u olarak hesaplanmıştır. Bu yüzden, analiz edilen 67862 benzersiz etikete karşılık küme sayısı 20000 olarak seçilmiştir.

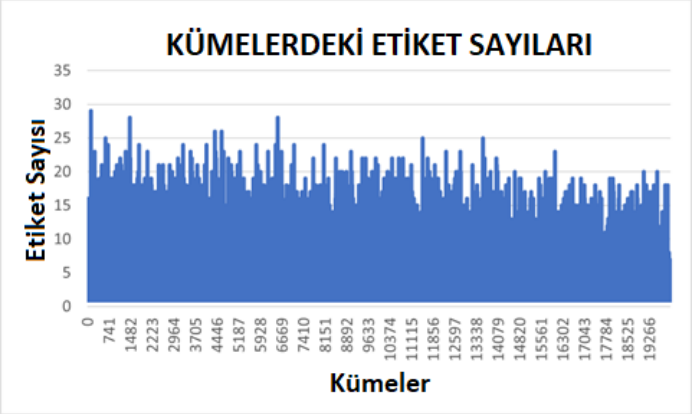
K Means kümeleme işlemi sonucunda 67832 etiketten 20000 küme oluşturulmuştur. Kümelerin yaklaşık yarısı tek bir etiketten oluşurken, en fazla elemana sahip küme 29 etiket içermektedir. Kümelerin dağılım diyagramı Şekil 3.19' da gösterilmiştir. Ayrıca ortaya çıkan kümelerden üçü içerdikleri etiketlerle örnek olarak Çizelge 3.15' te verilmiştir.



Şekil 3.17 Etiket sayısına bağlı olarak analizde gereken RAM miktarı



Şekil 3.18 Elbow yöntemiyle kırılma varyansının K küme sayısına bağlı değişimi



Şekil 3.19 Kümeleme işlemi sonrası kümelerdeki etiket sayıları

Çizelge 3.15 Aynı küme içerisinde kümelenen etiketlerden örnekler

Kümeler	Etiketler
1	payment to proceed from effect of Merger Net of
	Transaction cost
	principal payment on Securities sell under agreement
	to Repurchase financing activity
	proceed from Payment of Finance lease Mortgages
	and other net
2	financing fee
	Incentive fee
	offer Expenses
	Office Expense
	product expense
	licensing fee
3	issuance of common stock for account payable share
	issuance of common stock for warrant exercise value
	issuance of Common Stock from Warrant Exercise
	Value
	issuance of Common Stock in Follow on offer share

3.3.8.3. Evrişimsel sinir ağı

Evrişimsel Sinir Ağı (CNN), 1998' de Lecun ve arkadaşları tarafından bir derin sinir ağı modeli olarak geliştirilmiştir (LeCun ve ark., 1998). Özellikle görüntü işleme odaklı çalışmalarda kullanılmasına rağmen ses işleme, doğal dil işleme ve zaman serileri analizi için de kullanılmaktadır (Ayyadevara, 2018) . Önceki adımda gerçekleştirilen K Means kümeleme yöntemi, parametreleri %70 oranında azaltmıştır. Ancak, analiz edilen finansal verilerdeki çok sayıda özel etiket, yani yalnızca bir veya birkaç şirket tarafından

kullanılan etiketler, algoritma başarımını zorlayacak derecededir. Bu nedenle, uygulanacak algoritmanın seçimi ve ayarları önem arz etmektedir. Şekil 3.15' te gösterilen CNN tabanlı model, parametreleri önemli ölçüde azaltarak ve ayarlayarak öğrenme modellerinin etkinliğini artırmak için geliştirilmiştir. Modelin ilk evrişimsel katmanında filtre olarak 300, çekirdek boyutu olarak 7 ve aktivasyon fonksiyonu olarak ReLU seçilmiştir.

Parametreler seçilirken giriş verilerinin türü ve boyutu dikkate alınmıştır. Ara testlerde, düşük filtre seçimi algoritmanın performansını düşürürken, yüksek filtre seçimi önemli ölçüde iyileştirmemiştir. Ezberleme riskini azaltmak için, 0.3 parametresiyle bırakma (Dropout) gerçekleştirilmiştir. Bırakma, belirtilen parametreye sahip özellikleri rastgele kaldırır. Örneğin, 0.3 parametresi seçilirse, özelliklerin %70' i rastgele göz ardı edilir. Filtre 128' e düşürülürken çekirdek boyutu da 3'e düşürülmüştür. İkinci evrişim katmanında aktivasyon fonksiyonu olarak Sigmoid seçilmiştir. Önceki katmanlarda çok belirgin ve şişirilmiş parametreleri cezalandırmak için çekirdek düzenlileştirme (Kernel Regularization) uygulanmıştır. Böylece aşırı uyum sorununun önüne geçmek hedeflenmiştir (Rustam ve ark., 2020). Çekirdek düzenlemesi, bir katmandaki her parametre için cezaların toplanması ve bir sabitle çarpılmasıyla elde edilen kayıp fonksiyonu kullanılarak gerçekleştirilir. Deneylerde alt değerler başarısız, yüksek değerler verimsiz olduğu için düzenlileştirme birimlerinin değeri 64 olarak alınmıştır.

Üçüncü evrişim katmanı, ReLU aktivasyon fonksiyonu, çekirdek boyutu 2 ve filtre 16 ile oluşturulmuştur. Ardından, filtre boyutu 2 olan havuzlama (Maximum Pooling) katmanı eklenmiştir. Havuzlama katmanı, özellik haritasının boyutunu azaltmak için kullanılır. Böylece, öğrenilecek parametreler ve hesaplanacak hesaplamalar azaltılarak ağ daha verimli hale getirilir. Ayrıca havuzlama, önceki katmanlarda seçilen özellikler yerine özelliklerin özetlerini oluşturarak ağı daha sağlam hale getirir. Bir sonraki adımda evrişimli katmanlarda üretilen öznitelikler düzleştirme (Flatten) işlemi uygulanarak gizli katmanlara giriş için uygun hale getirilmiştir. İki gizli katman kullanılmıştır. İlkinde 32 nöron, ikincisinde ise girdideki özellik sayısı kadar nöron üretilmiştir. Son olarak, sınıflandırma çıktısı için Softmax algoritması kullanılmıştır.

Çıkış katmanında, çalışmalarda ağırlıklı olarak sigmoid veya softmax algoritmaları kullanılmaktadır. Özellikle bir sınıfta sınıflandırılması gereken örnekler için softmax algoritması tercih edilmektedir. Bunun nedeni, sigmoid aktivasyon fonksiyonunun her çıkış için 0 ile 1 arasında bir değer üretmesidir. Ancak birkaç sigmoid fonksiyonun birleştirilmesiyle oluşturulan softmax algoritmasında çıkışlara, çıkışların

toplamı 1 olacak şekilde ağırlık değerleri verilir. Böylece softmax algoritması, her bir çıkışın tam sınıfını bulmak için daha uygun bir yol sağlar. Adam optimizeri, 0.0001 öğrenme oranı, 0.9 beta1, 0.99 beta2 ve 1e-10 epsilon değerleri ile model optimize edici olarak seçilmiştir. Döngü değeri olarak da 20 belirlenmiştir. Oluşturulan modelin değerlendirmesi bir sonraki bölümde yapılmıştır.

Hibrit modelin performansını ölçmek ve karşılaştırmak için modelin her iki kısmı üzerinde de testler yapılmıştır. Doc2Vec ve K Means kümeleme yöntemi ile oluşturulan veri setine ek olarak, karşılaştırmalar için dört farklı veri seti de oluşturulmuştur. Etiketlerin vektöre dönüşüm aşamasını test etmek için Bert Base Nli Mean Tokens ve TFIDF yöntemleri Doc2Vec yöntemini karşılaştırmak için kullanılmıştır. Kümeleme için, K Means kümeleme yöntemi DBSCAN yöntemi ile karşılaştırılmıştır. Bu aşamada algoritmalarda kullanılmak üzere dört farklı veri seti oluşturulmuştur. Ayrıca çalışmada önerilen öznitelik çıkarım yönteminin karşılaştırılabilmesi için özellik olarak firmaların kullandığı etiketlerin direkt kullanıldığı bir veri seti üretilmiştir. Böylece önerilen hibrit yöntemin ilk kısmı için toplam beş farklı veri seti üretilmiştir. Veri setindeki şirketlerin finansal oranları çok değişken olduğundan standart ölçekleme (Standard Scaler) yöntemi kullanılarak normalize edilmiştir. Ek olarak, veri setleri yeniden karıştırılarak yüzde 90 eğitim, yüzde 10 test olacak oranda ayrılmıştır.

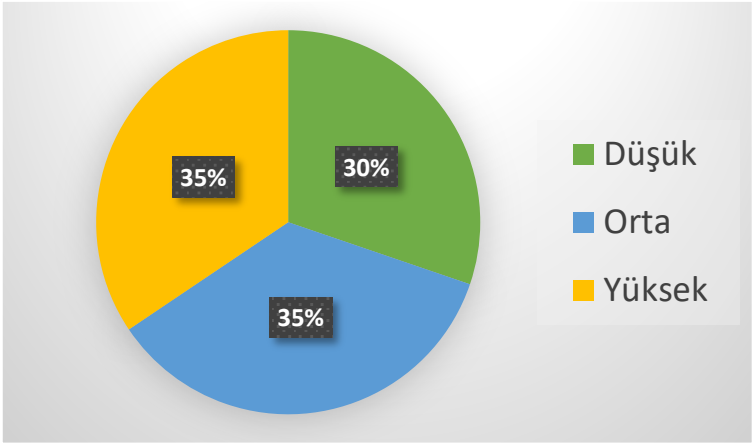
Üretilen veri setleri, şirketlerin fiyat gücünü tahmin etmek için beş farklı makine öğrenme yöntemiyle analiz edilmiştir. Oluşturulan CNN ağına ek olarak Geriye Yayılım Sinir Ağı, Karar Ağacı, K En Yakın Komşu ve İkinci Derece Ayırma Analizi algoritmaları da çalıştırılmıştır. Oluşturulan CNN ağının özellikleri önceki bölümde verilmiştir. Geri yayılım yöntemi için ara testler yapılmıştır. En iyi sonucu veren ağın parametreleri şu şekildedir; giriş katmanı (Dense 200, fonksiyon Relu), birinci orta katman (Dense 100, fonksiyon Relu), ikinci orta katman (Dense 50, fonksiyon Relu), aktivasyon katmanı (Dense 3, fonksiyon Sigmoid) ile dört katman. Optimizer olarak Adamax seçilmiş ve diğer ayarlar; öğrenme oranı 0.01, epsilon 1.0 E-07, parça büyüklüğü 25, döngü değeri 100 şeklinde belirlenmiştir. Karar Ağacı algoritmasında maksimum derinlik 100 ve kriter olarak da entropi seçilmiştir. K En Yakın Komşu Algoritması' nda komşu değeri 25 olarak belirlenmiş ve İkinci Derece Ayırma Analizi algoritmasında kovaryans değil ve tolerans değeri 0.0001 olarak seçilmiştir. Bu üç algoritmanın açıkça belirtilmeyen diğer özellikleri için varsayılan ayarlarla Scikit-Learn (Buitinck ve ark., 2013) çerçevesi kullanılarak yürütülmüştür. Tüm sonuçlar, karşılaştırmalar ve yorumlar sonraki bölümde yer almaktadır.

3.4. Tek Boyutlu Evrişimsel Sinir ağı ve Uzun Kısa Süreli Bellek Hibrit Yöntemiyle Finansal Verilerin Sınıflandırılması

Önceki bölümde açıklanan ve Doc2Vec ile K Means kümeleme yöntemlerinin birleştirilmesiyle üretilen veri setinin analizi için yeni bir hibrit yöntem geliştirilmiştir. Finansal raporlardan üretilen Doc2Vec_KMeans_10Q_Dataset[8] veri seti, iki güçlü derin öğrenme algoritmaları olan evrişimsel sinir ağı ve uzun kısa süreli bellek yöntemlerinin birleştirilmesiyle oluşturulan yeni bir hibrit yöntemle sınıflandırılmıştır. Hibrit yöntemi oluşturan her iki algoritmanın da güçlü yönleri öne çıkarılarak, özellikle özellik seçimi adımının çok daha verimli olması ve daha iyi sonuçların elde edilmesi hedeflenmiştir. Bunun için, tek boyutlu evrişimsel sinir ağı ile standart uzun kısa süreli bellek ağlarından geçirilen verinin sınıflandırılması için çıkış fonksiyonu olarak da softmax algoritması seçilmiştir.

3.4.1. Doc2Vec_KMeans_10Q_Dataset veri seti

Finansal 10Q raporlarından elde edilen veri seti önceki bölümde detaylıca açıklanmıştır. 34.686 satır ve 20.000 sütundan oluşan veri seti üç ayrı sınıfla ayrışmaktadır. Veri seti, sınıf dağılım ağırlıkları açısından homojen bir yapıya sahiptir. Sınıfların ağılıkları Şekil 4.1' de verilmiştir.



Şekil 3.20 Doc2Vec_KMeans_10Q_Dataset veri seti sınıf dağılımı

[8] https://github.com/samikacar/Doc2Vec_KMeans_10Q_Dataset

3.4.2. Tek boyutlu evrişimsel sinir ağları

Tek boyutlu evrişimsel sinir ağları, zaman serileri gibi sıralı verilerin analizi için geliştirilmiş bir derin öğrenme ağıdır. Tek boyutlu evrişimsel sinir ağları diziler içindeki desenleri ve özellikleri tanımlamada oldukça başarılıdır. Boyutları dışında geleneksel iki boyutlu evrişimsel sinir ağları ile birçok yönden aynı özelliklere sahiptir. Evrişim katmanları, havuzlama katmanları, aktivasyon fonksiyonları ve tam bağlı katmanların birleşimiyle oluşturulur. Bırakma ve toplu normalleştirme (Batch Normalization) yöntemleri de aşırı öğrenmeyi engellemek için kullanılır.

3.4.3. Uzun kısa süreli bellek sinir ağları

1997 yılında Hochreiter ve Schmidhuber tarafından tanıtılan uzun kısa süreli bellek ağları, geleneksel tekrarlayan sinir ağlarının (Recurrent Neural Network, RNN) (Rumelhart ve ark., 1986) eğitiminde sıklıkla sorunlara neden olan yok olan gradyan probleminin üstesinden gelmek için özel olarak tasarlanmıştır (Hochreiter ve Schmidhuber, 1997). LSTM ağlarının merkezinde, hücre durumunu ve giriş kapısı, unutma kapısı ve çıkış kapısı olmak üzere üç geçit mekanizmasını koruyan bir bellek hücresi bulunur. Bu kapılar, bilgi akışının bellek hücresinin içine, dışına ve hücre içindeki akışını kontrol ederek uzun kısa süreli belleklerin genişletilmiş diziler boyunca bilgiyi seçici olarak hatırlamasını veya unutmasını sağlar. Bu ayırt edici mimari, uzun kısa süreli bellek algoritmasının sıralı verilerdeki uzun vadeli bağımlılıkları kavramasını ve bunlardan yararlanmasını sağlayarak onları doğal dil işleme, konuşma tanıma ve zaman serisi tahmini gibi görevlerde oldukça etkili olmasını sağlar (Hochreiter ve Schmidhuber, 1997). Uzun kısa süreli bellek sinir ağları, finansal marketlerin analizi (Fisher ve Klaus, 2018), hisse senetleri fiyat tahminleri (Thakkar ve Chaudhari, 2021) gibi birçok finansal alanda da kullanılmıştır.

3.4.4. Hibrit yöntem

Tek boyutlu evrişimsel sinir ağlarının ve uzun kısa süreli bellek ağlarının güçlü yönlerinden yararlanan ve sıralı verilerin analizinde sağlam bir model oluşturan hibrit bir yöntem önerilmiştir. Mimaride, verilerdeki uygun özellikleri çıkarmak için bir dizi evrişim katmanına yer verilmiştir. Uzun vadeli bağlılıkları ve zamansal bağlamı çıkarmak için de LSTM katmanları kullanılmıştır. Bu birleşim, verideki doğrudan ve dolaylı olan bağlamları çıkarmak için verimli bir yapı oluşturmuştur. Ağın eğitiminde, zaman içinde

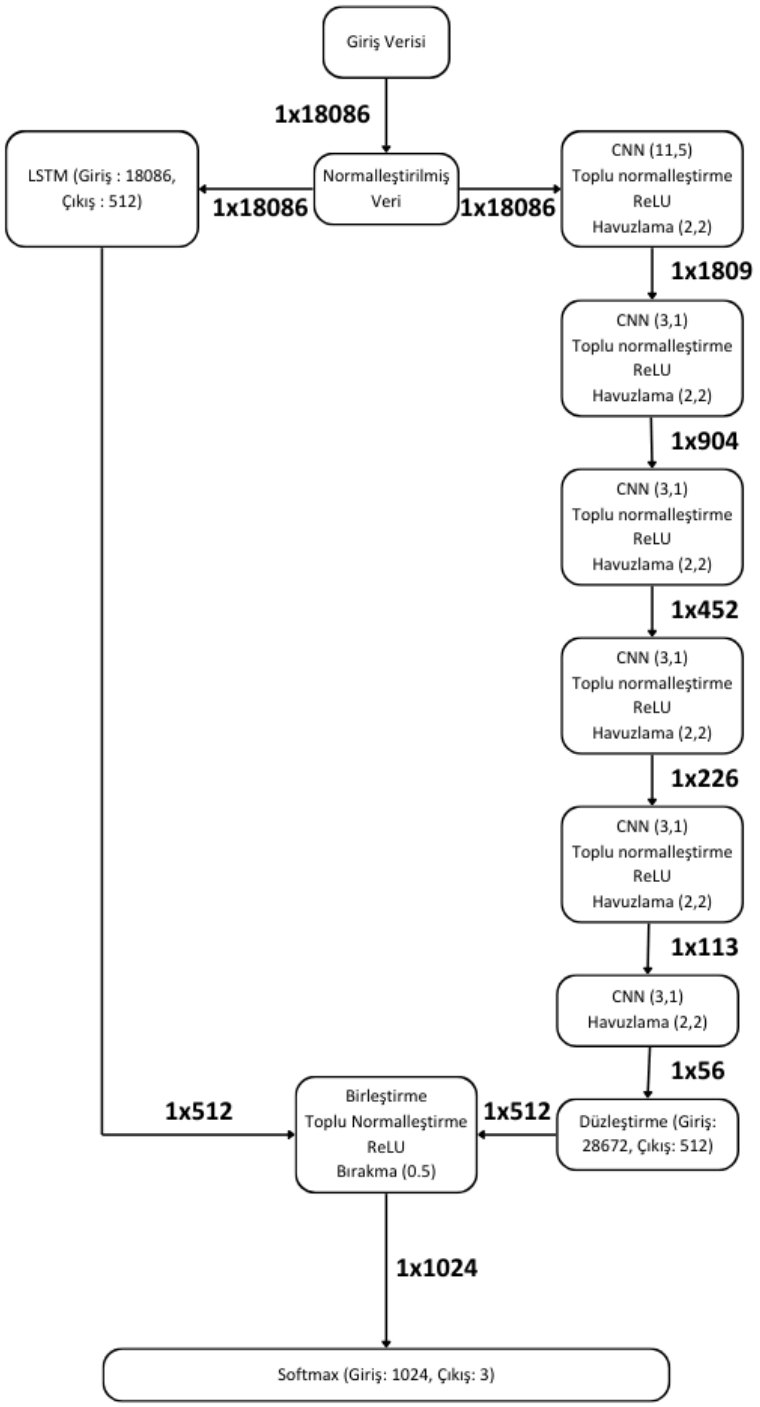
geri yayılım (Back Propagation Through Time, BPTT) ve dereceli azalma (Gradient Decent) yöntemleri kullanılarak algoritmanın optimizasyonu gerçekleştirilmiştir. Modelin performansını artırmak için normalizasyon dahil veri ön işleme teknikleri de kullanılmıştır.

Hibrit model, altı tek boyutlu evrişim katmanı ve bir uzun kısa vadeli bellek katmanından meydana gelmiştir. Bunun dışında, toplu normalleştirme, havuzlama, bırakma yöntemleri de modele dahil edilmiştir. Çıkış katmanında softmax, diğer tüm katmanlarda ReLU aktivasyon fonksiyonları kullanılmıştır. Modelin tek boyutlu evrişimsel sinir ağı bölümünün ayrıntıları Çizelge 3.16' da verilmiştir.

Çizelge 3.16 Tek boyutlu evrişimsel sinir ağı mimari özellikleri

Tür	**Çıkış**	**Çekirdek Boyutu**	**Adım**	**Çerçeve**
Evrişim	3618 x 32	1 x 11	1 x 5	1 x 5
Havuzlama	1809 x 32	1 x 2	1 x 2	-
Evrişim	1809 x 64	1 x 3	1 x 1	1 x 1
Havuzlama	904 x 64	1 x 2	1 x 2	-
Evrişim	904 x 128	1 x 3	1 x 1	1 x 1
Havuzlama	452 x 128	1 x 2	1 x 2	-
Evrişim	452 x 256	1 x 3	1 x 1	1 x 1
Havuzlama	226 x 256	1 x 2	1 x 2	-
Evrişim	226 x 512	1 x 3	1 x 1	1 x 1
Havuzlama	113 x 512	1 x 2	1 x 2	-
Evrişim	113 x 512	1 x 3	1 x 1	1 x 1
Havuzlama	56 x 512	1 x 2	1 x 2	-
Düzleştirme	1x28672	-	-	-
Tam Bağlantı	1x512	-	-	-

1 x 18086 uzunluğundaki giriş verileri hem uzun kısa süreli bellek hem de evrişimsel sinir ağlarına gönderilmiştir. Bu ağların çıktıları birleştirilmiş ve sonuçta 1 x 1024 uzunluğunda özellik matrisi elde edilmiştir. Çıkarılan bu özellikler, aşırı öğrenmeyi engellemek için toplu normalleştirme ve bırakma işlemleri uygulanarak, sınıflandırma işlemini tamamlamak üzere Softmax aktivasyon fonksiyonlu çıkış katmanına iletilmiştir. Tasarlanan hibrit modelin ayrıntılı özellikleriyle mimarisi Şekil 3.21' de gösterilmiştir.



Şekil 3.21 Önerilen hibrit modelin mimarisi

4. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

Tez çalışması kapsamında, finansal raporların makine öğrenmesi ve doğal dil işleme teknikleri kullanılarak analizi ile önemli sonuçlara ulaşılmıştır. Bu bölümde, gerçekleştirilen çalışmaların sonuçları, bu sonuçların değerlendirilmesi ve literatüre katkısı tartışılmıştır. Sonuçların değerlendirilmesinde, yapay zekâ alanında kabul görmüş metriklerle, hazırlanan veri setleri üzerinde çalıştırılan yöntemler değerlendirilmiştir. Önerilen yöntemler, geçmişteki birçok çalışmada kullanılmış, önemi kanıtlanmış diğer makine öğrenmesi yöntemleriyle kıyaslanarak karşılaştırmalı olarak analizlerin gerçekleştirilmesi sağlanmıştır. Yapılan tüm çalışmaların sonuçları, farklı başlıklar altında alt bölümlerde sunulmuştur.

4.1. 10K Raporlarının Analiz Sonuçları

Bölüm 3.2' de hazırlanan veriler, 6 farklı sınıflandırma algoritmasına uygulanmış, algoritmaların sınıflandırma başarımları farklı ölçütlerle test edilmiştir. Tüm sonuçlar aşamalı olarak farklı çizelgelerde gösterilmiştir. Çizelge 4.1' de üretilen veri setinin 6 farklı makine öğrenmesi yöntemiyle sınıflandırma işlemi sonucunda elde edilen değerler yer almaktadır. Sonraki bölümlerde de eklenen yeni yöntemlerin sonuçlara etkisi gözlemlenebilmektedir.

Çizelge 4.1 10K raporlarından üretilen veri setinin 6 farklı makine öğrenmesi yönteminde sınıflandırma testi performansları

	10-Katlı Çapraz Doğrulama				**%60 Eğitim %40 Test Ayrımı**
Kullanılan Algoritma	**Doğruluk**	**Dengelenmiş Doğruluk**	**Kesinlik**	**Duyarlılık**	**Doğruluk**
Rastgele Orman	93.47	85.61	73.66	86.23	94.00
KNN	88.91	76.17	56.78	72.55	90.00
SVM	67.12	52.98	21.46	34.53	47.00
Naive Bayes	87.72	80.12	68.54	63.06	86.00
Karar Ağacı (C4.5)	90.75	81.94	68.54	75.28	92.00
AdaBoost	91.10	79.92	62.92	80.92	92.00
Lojistik Regresyon	91.98	82.79	68.80	81.52	91.00

4.1.1. Korelasyon değeri ile özellik seçimi

Veriyi temsil eden en iyi öznitelikleri bulmak için her bir özniteliğin korelasyon değeri hesaplanmıştır. Veriler arasında düşük korelasyondan (en az 0.30 korelasyon

değeri) daha az ilişki bulunan öznitelikler veriden çıkarılmıştır. Bu işlemin sonucunda 89 öznitelikten 11'i en az düşük korelasyon değerine sahip olmak üzere seçilmiştir. Gerçekleştirilen özellik seçimi adımından sonra veri tekrar sınıflandırma algoritmalarına tabi tutulmuştur. Ölçüm metrikleri tekrar hesaplanarak algoritmaların tahmin yeteneklerindeki başarımları tespit edilmiştir. Korelasyona bağlı özellik seçimi sonrası elde edilen sonuçlar Çizelge 4.2' de sunulmuştur.

Çizelge 4.2 Korelasyon işlemi sonrası test sonuçları

	10-Kat Çapraz Doğrulama				**%60 Eğitim %40 Test Ayrımı**
Kullanılan Algoritma	**Doğruluk**	**Dengelenmiş Doğruluk**	**Kesinlik**	**Duyarlılık**	**Doğruluk**
Rastgele Orman	93.56	86.37	75.45	85.26	93.96
KNN	89.92	79.72	64.19	73.61	90.24
SVM	55.77	48.76	38.11	16.27	80.15
Naive Bayes	91.98	82.79	68.80	81.52	92.87
Karar Ağacı (C4.5)	91.01	81.70	67.52	77.19	92.98
AdaBoost	90.44	77.30	57.29	81.45	91.22
Lojistik Regresyon	92.50	82.29	66.75	86.42	93.00

4.1.2. Verilerin 0-1 aralığında ayrıştırılması

Veri incelendiğinde şirketlerin finansal tabloları arasında ciddi farklılıklar tespit edilmiştir. Piyasa değeri en yüksek olan şirketlerin verileri milyar dolarla ifade edilirken en küçüklerin de binlerle ifade edilebilmektedir. Bu yüzden, önceki adımda korelasyon değerlerine göre veriyi temsil eden en iyi 11 özniteliği seçilen veri üzerindeki örneklerin değerleri 0-1 Ayrıştırılması (Fernández ve ark., 2006) ile ayrıştırılmıştır. Ayrıştırma sonrası elde edilen veri tekrar aynı sınıflandırma algoritmalarına tabi tutulmuştur. Sonuçlar, sınıflandırma algoritmaları ve ölçüm metrikleri ile Çizelge 4.3' te sunulmuştur.

Çizelge 4.3 Ayrıştırma işlemi sonrası algoritma test performansları

	10-Katlı Çapraz Doğrulama				%60 Eğitim %40 Test Ayrımı
Kullanılan Algoritma	**Doğruluk**	**Dengelenmiş Doğruluk**	**Kesinlik**	**Duyarlılık**	**Doğruluk**
Rastgele Orman	93.69	86.56	75.70	85.80	94.18
KNN	89.92	79.72	64.19	73.61	90.24
SVM	89.87	71.78	44.25	93.01	89.00
Naive Bayes	91.98	82.79	68.80	81.52	92.87
Karar Ağacı (C4.5)	91.36	82.52	69.05	78.03	92.98
AdaBoost	90.44	77.30	57.29	81.45	91.22
Lojistik Regresyon	92.50	82.29	66.75	86.42	93.00

4.1.3. Sınıflara ait örnek sayısının eşitlenmesi

Sınıf etiketleri oluşturulurken örnekler %82'si ortalamanın altında %18'i üzerinde olacak biçimde ayrılmıştır. Dengesiz sınıf dağılımının olduğu verilerde algoritmanın performansını ölçmede sınıflandırma doğruluğu aldatıcı olabileceği için farklı performans metrikleri kullanılmıştır. Ancak, sınıfların dengeli dağıldığı bir veri setinde de sonuçları görmek adına, 732 farklı örnekten oluşan ve örneklerin sınıflara eşit dağıldığı bir alt veri grubunda önceki adımlarda kullanılan sınıflandırma algoritmaları çalıştırılmış ve ölçüm metrikleri hesaplanmıştır. Sonuçlar, Çizelge 4.4' te sunulmuştur.

Çizelge 4.4 Sınıf ağırlıklarının eşitlenmesi sonrası test sonuçları

	10-Katlı Çapraz Doğrulama				%60 Eğitim %40 Test Ayrımı
Kullanılan Algoritma	**Doğruluk**	**Dengelenmiş Doğruluk**	**Kesinlik**	**Duyarlılık**	**Doğruluk**
Rastgele Orman	84.15	84.15	83.88	84.34	81.00
KNN	71.86	71.86	62.30	77.03	75.00
SVM	54.23	54.23	36.89	56.49	54.00
Naive Bayes	71.17	71.17	56.28	80.16	72.00
Karar Ağacı (C4.5)	78.42	78.42	75.68	80.06	78.00
AdaBoost	79.37	79.37	75.14	82.09	77.00
Lojistik Regresyon	78.42	78.42	72.95	81.90	78.00

4.1.4. Sonuçların değerlendirmesi

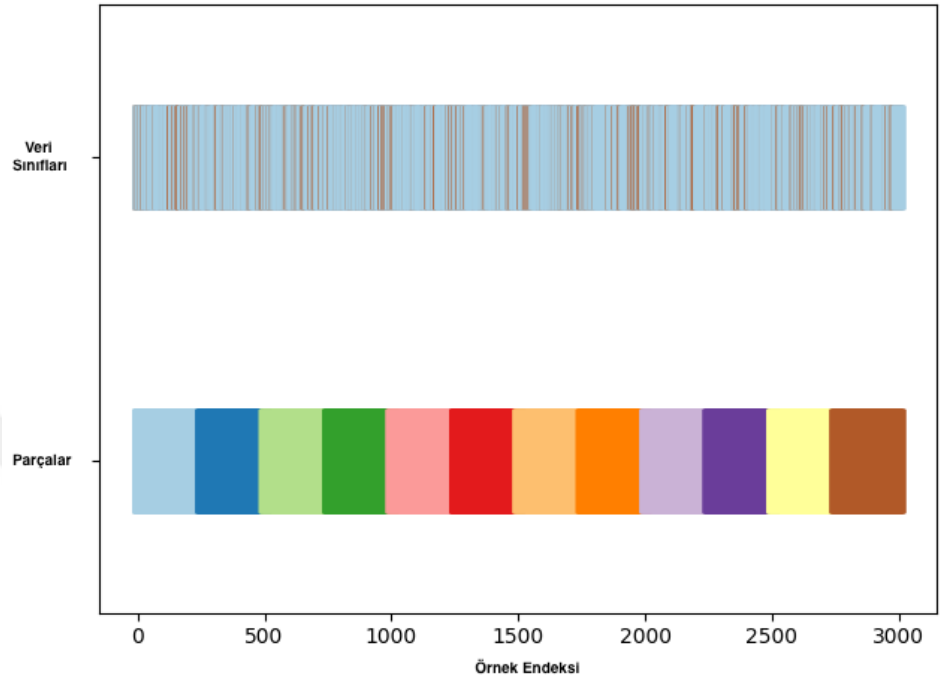
EDGAR veri tabanından alınan şirketleri ait 10K raporlarının finansal tablolarındaki veriler çeşitli ver ön işleme adımlarına tabi tutularak hazırlanmıştır. Şirketlerin 2019 yılında paylaşmış oldukları veriler ile NASDAQ borsasından alınan piyasa değerlerinden elde edilen sınıflar birleştirilmiş ve 10K raporlarındaki finansal

tablo verileri üzerinden en değerli şirketlerin tespit edilmesi hedeflenmiştir. Piyasa değerlerinin ortalamasının altında olan şirketlerin oluşturduğu sınıftaki örnek sayısının fazla olması nedeniyle doğruluk değerleri algoritmaların tahmin başarımını iyi temsil edememektedir. Diğer ölçüm metrikleri ile algoritmaların elde edilen veri üzerindeki başarımları tespit edilmiştir. Veri, özellik seçimi ve ayrıştırma öncesi, korelasyon ile özellik seçimi sonrası ve ayrıştırma işlemi sonrası ayrı ayrı 6 farklı sınıflandırma algoritmasına tabi tutulmuştur. Rastgele Orman algoritması 3 ayrı veride de en iyi sonuçları üretmiştir. Özellik seçimi sonrası, Rastgele Orman, KNN ve Naive Bayes algoritmalarının başarısı artmıştır. Destek Vektör Makineleri ve AdaBoost algoritmalarının başarısının düştüğü, diğer algoritmaların ise skorlarında bir değişim olmadığı tespit edilmiştir. Ayrıştırma işlemi sonucunda ise, Destek Vektör Makineleri algoritmasında ciddi bir iyileşme görülmüş, diğerlerinde ise önemli bir artış veya azalış elde edilememiştir. Son olarak, her iki sınıf için de eşit sayıda örneğe sahip alt veri grubuyla yapılan sınıflandırma çalışmasında en kayda değer sonucu yine Rastgele Orman üretmiştir. Destek Vektör Makineleri algoritması ise başarısız olmuştur. Kesinlik ve duyarlılık metriklerinde ise eşit sayıda örneğe sahip sınıfların kullanıldığı veri kümesi için ilgili sınıflandırma algoritmalarında önemli derece artış olduğu sonucu elde edilmiştir.

4.2. Doğal Dil İşleme İşlemleri Sonrası Analiz Sonuçları

Bölüm 3.1 de ve Bölüm 3.2 de verilerin toplanması, ön işleme teknikleriyle veri setine dönüşümü detaylı bir şekilde açıklanmıştır. Bu bölümde 10K dosyalamalarından elde edilen veriler, ait olduğu firmayı temsil etme yeteneği, ön işleme tekniklerinin ne kadar kullanışlı olduğu, uygulanan doğal dil işleme yöntemlerinin veriye etkisi, sadece 10K dosyalama bilgisi ile piyasadaki en değerli şirketlerin diğerlerinden ayırt edilip edilemeyeceği ve büyük miktarda veri ile bu verilerden elde edilen özet arasındaki farkın ortaya çıkarılması için gerçekleştirilen analizlerin sonuçları sunulmuştur. Bunun için önceki aşamalarda elde edilen sekiz veri setinin tamamı beş makine öğrenmesi algoritmasıyla analiz edilmiştir: KNN, Karar Ağacı, Rastgele Orman, Adaboost ve İkinci Derece Ayırma Analizi. Algoritma seçiminde, örneklemleri sınıflandırırken farklı matematiksel yöntemler kullanan algoritmaları seçmeyi ve böylece elde edilen verilerin farklı yöntem ve yaklaşımlardaki etkinliğini belirleme amaçlanmıştır. Verileri yeniden örneklemek için 10 katlı çapraz doğrulama (K fold cross-validation) kullanılmıştır ve

analizlerden önce veri kümeleri karıştırılmıştır (Fushiki, 2011). Bir veri setindeki veri örneklerinin sınıf etiketlerine göre dağılımı ve buna göre her kat için 10 eşit parçaya bölünmesi Şekil 4.1' de verilmiştir.



Şekil 4.1 10 Katlı çapraz doğrulama işleminde veri sınıflarının dağılım görünümü

Algoritmaları değerlendirmek için doğruluk, hassasiyet, geri çağırma, f-Skor ve ROC Eğrisi metrikleri kullanılmıştır. Sonuçlar, tüm metrikler açısından her kat için ayrı ayrı hesaplanmış ve nihai ortalama değerler ve güven aralığı sapmaları da çıkarılmıştır. Belirtilen sınıf etiketlerine göre veri seti dengesiz olduğu için tüm metriklerin makro ortalama sonuçları çıkarılmıştır. Çizelge 4.5' de verilen karışıklık matrisinden tüm metrikler çıkarılabilmektedir. Buna göre gerçek pozitif sayısı, S&P 500 endeksindeki şirketlerin kaçının algoritmayı doğru tahmin ettiğini göstermektedir. Gerçek negatiflerin sayısı, aslında bu endekste yer almayan şirketlerin kaçının algoritma tarafından doğru bir şekilde tanımlandığını gösterir. Yanlış pozitif ve negatif değerler, sırasıyla, gerçekte S&P 500 endeksine sahip olmadığı halde algoritmanın bu endekste olduğunu tahmin ettiği şirket sayısını ve algoritmanın aslında S&P 500 endeksinde olan şirketlerin sahip olmadığı tahmin ettiği şirket sayısını temsil eder. Özetle, şirketlerin 10K dosyalama yoluyla değerlendirilmesi için önerilen veri madenciliği yöntemleri ile elde edilen yedi veri seti, beş metrik açısından beş yöntemle analiz edilmiş ve sonuçlar elde edilmiştir. Herhangi bir işlem yapılmayan veri setiyle birlikte tüm analiz sonuçları Çizelge 4.6' da

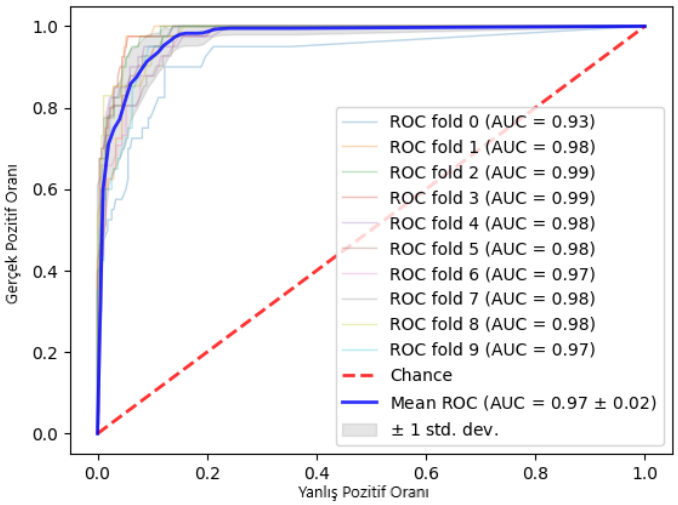
verilmiştir. Veri setleri için en iyi sonuçlar altı çizili olarak gösterilmiştir ve algoritmalar için en iyi sonuçlar koyu renkle gösterilmiştir. Ayrıca en iyi ve en kötü üç test sonucunun ROC eğrileri de Şekil 4.3-4.8' de verilmiştir.

Çizelge 4.5 Karışıklık matrisi yapısı

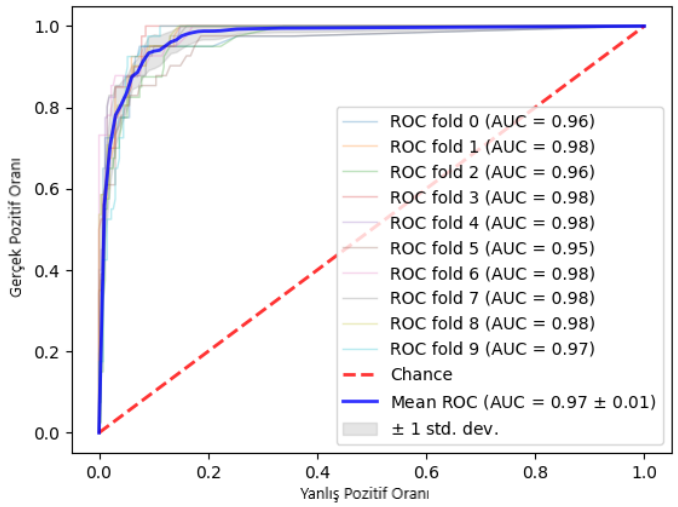
	Tahmin Edilen Sınıflar	
Gerçek Sınıflar	Gerçek Pozitif (TP)	Yanlış Negatif (FP)
	Yanlış Pozitif (FP)	Gerçek Negatif (TN)

Çizelge 4.6 Sekiz farklı veri setinin farklı algoritmalarla güven aralıklı test performansları

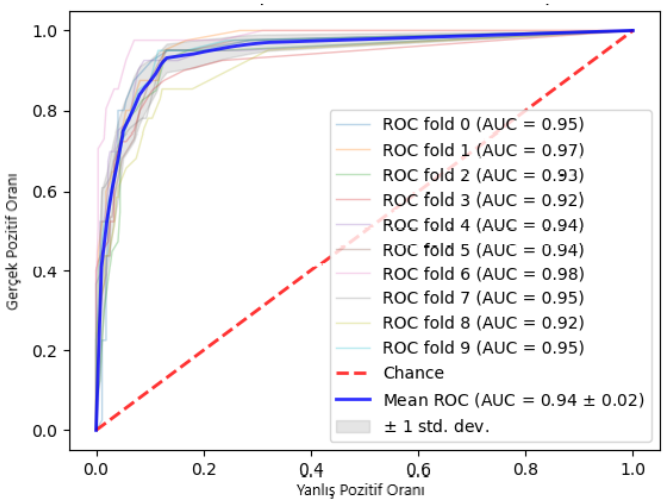
	Metrik	**KNN**	**GA %95**	**Karar Ağacı**	**GA %95**	**Rastgele Orman**	**GA %95**	**Adaboost**	**GA %95**	**QDA**	**GA %95**
30	Doğruluk	0,9236	0,0073	**0,9317**	0,0087	0,9414	0,0075	**0,9433**	0,0064	0,8990	0,0105
	Hassasiyet	**0,8589**	0,0244	**0,8457**	0,0247	0,8800	0,0237	**0,8829**	0,0138	0,7759	0,0236
	Geri Çağırma	0,7768	0,0211	**0,8516**	0,0246	**0,8511**	0,0288	**0,8591**	0,0193	**0,8792**	0,0216
	F-Skoru	0,8088	0,0204	**0,8480**	0,0237	0,8626	0,0231	**0,8698**	0,0152	0,8121	0,0228
40	Doğruluk	0,9223	0,0054	0,9249	0,0100	0,9414	0,0107	0,9408	0,0052	0,9058	0,0115
	Hassasiyet	0,8549	0,0118	0,8387	0,0267	0,8867	0,0277	0,8803	0,0118	0,7872	0,0222
	Geri Çağırma	0,7747	0,0159	0,8228	0,0250	0,8483	0,0233	0,8515	0,0208	0,8639	0,0209
	F-Skoru	0,8067	0,0132	0,8297	0,0243	0,8646	0,0233	0,8635	0,0138	**0,8176**	0,0222
50	Doğruluk	0,9246	0,0096	0,9230	0,0101	**0,9433**	0,0046	0,9398	0,0069	**0,9107**	0,0121
	Hassasiyet	0,8560	0,0254	0,8381	0,0250	**0,8911**	0,0207	0,8787	0,0206	0,7985	0,0260
	Geri Çağırma	0,7888	0,0212	0,8210	0,0250	0,8474	0,0166	0,8521	0,0125	0,8433	0,0215
	F-Skoru	0,8164	0,0214	0,8268	0,0201	**0,8663**	0,0151	0,8633	0,0117	0,8171	0,0235
60	Doğruluk	0,9243	0,0116	0,9126	0,0066	0,9408	0,0071	0,9401	0,0076	0,9081	0,0128
	Hassasiyet	0,8516	0,0247	0,8055	0,0220	0,8822	0,0172	0,8754	0,0193	0,8054	0,0255
	Geri Çağırma	0,7995	0,0166	0,8084	0,0187	0,8480	0,0094	0,8558	0,0158	0,7797	0,0242
	F-Skoru	0,8220	0,0189	0,8058	0,0185	0,8635	0,0116	0,8641	0,0147	0,7897	0,0227
70	Doğruluk	0,9242	0,0084	0,9142	0,0073	0,9382	0,0062	0,9343	0,0080	0,9071	0,0097
	Hassasiyet	0,8511	0,0146	0,8120	0,0156	0,8716	0,0143	0,8602	0,0150	**0,8227**	0,0287
	Geri Çağırma	0,7975	0,0186	0,8103	0,0198	0,8453	0,0199	0,8442	0,0275	0,7268	0,0176
	F-Skoru	0,8200	0,0156	0,8099	0,0153	0,8572	0,0165	0,8502	0,0201	0,7616	0,0200
80	Doğruluk	**0,9252**	0,0074	0,9126	0,0109	0,9388	0,0061	0,9298	0,0077	0,8980	0,0093
	Hassasiyet	0,8515	0,0211	0,8098	0,0312	0,8799	0,0177	0,8552	0,0167	0,8281	0,0274
	Geri Çağırma	0,8015	0,0229	0,8055	0,0227	0,8360	0,0204	0,8234	0,0249	0,6547	0,0277
	F-Skoru	0,8217	0,0193	0,8062	0,0260	0,8553	0,0179	0,8367	0,0197	0,6955	0,0339
90	Doğruluk	0,9252	0,0057	0,9077	0,0067	0,9255	0,0042	0,9301	0,0061	0,8705	0,0125
	Hassasiyet	0,8461	0,0122	0,7967	0,0185	0,8485	0,0169	0,8559	0,0201	0,7049	0,0973
	Geri Çağırma	**0,8056**	0,0157	0,7833	0,0162	0,8002	0,0184	0,8223	0,0116	0,5199	0,0093
	F-Skoru	**0,8234**	0,0134	0,7895	0,0168	0,8210	0,0165	0,8372	0,0137	0,5058	0,0180
100	Doğruluk	0,9132	0,0061	0,9149	0,0065	0,7941	0,0060	0,8708	0,0075	0,1327	0,0089
	Hassasiyet	0,8491	0,0183	0,8117	0,0211	0,6087	0,0174	0,8807	0,0239	0,3820	0,1343
	Geri Çağırma	0,7265	0,0172	0,8150	0,0207	0,6492	0,0148	0,5062	0,0197	0,4995	0,0037
	F-Skoru	0,7684	0,0174	0,8116	0,0169	0,6212	0,0155	0,4776	0,0176	0,1182	0,0069



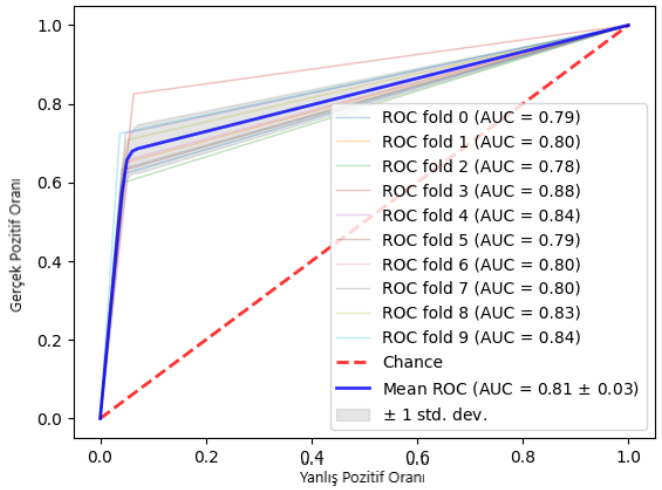
Şekil 4.2 30DS veri seti için Rastgele Orman algoritması test sonucu ROC eğrisi



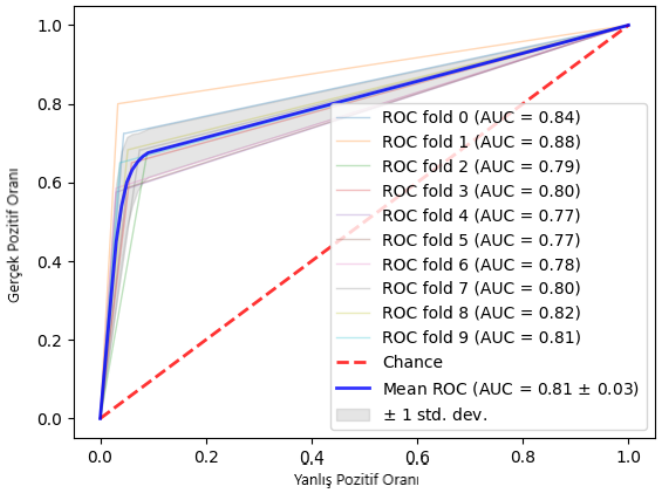
Şekil 4.3 50DS veri seti için Rastgele Orman algoritması test sonucu ROC eğrisi



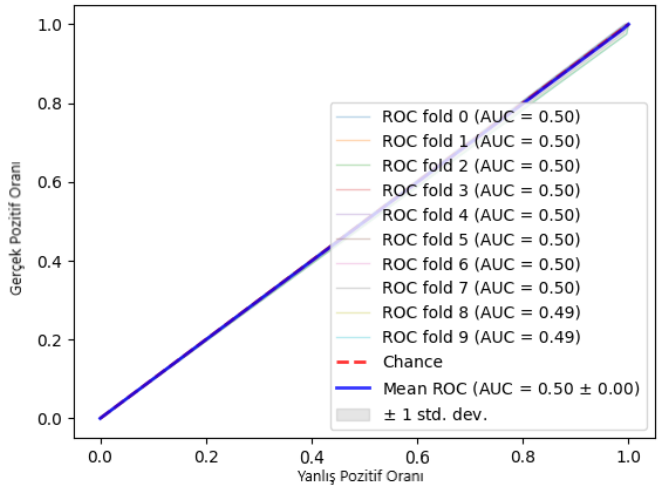
Şekil 4.4 80DS veri seti için K En Yakın Komşu algoritması test sonucu ROC eğrisi



Şekil 4.5 30DS veri seti için Karar Ağacı algoritması test sonucu ROC eğrisi



Şekil 4.6 90DS veri seti için Karar Ağacı algoritması test sonucu ROC eğrisi



Şekil 4.7 AllTags veri seti için İkinci Derece Ayırma Analizi algoritması test sonucu ROC eğrisi

4.2.1. Analiz sonuçlarının değerlendirilmesi

Şirketlerin her mali yılın ilk üç ayında tamamladıkları yıla ait bilgilerini açıkladıkları SEC tarafından sağlanan EDGAR veri tabanından indirilen 10K rapor dosyalarından elde edilen veri seti analiz edilmiştir. Farklı sayıda özellik vektörüne sahip sekiz veri seti, beş metrik üzerinden beş ML yöntemiyle değerlendirilmiştir. Analizde 16 GB Ram, Intel i7 işlemcili emtia bilgisayar kullanılmıştır. 10 katlı çapraz doğrulama yöntemiyle algoritmalar aracılığıyla analizlerde kullanılan tüm veri setlerinin ortalama çalışma süreleri Çizelge 4.7' de verilmiştir. Beklendiği gibi, sadece 125.000 sütundan

oluşan veri seti analiz edilirken çalışma süreleri uzamıştır, diğerleri ise çok daha kısa sürelerde analizi tamamlamıştır.

Çizelge 4.7 Tüm veri setlerinin algoritmalarda çalışma süreleri

Veri Seti	KNN	Karar Ağacı	Rastgele Orman	Adaboost	QDA	Toplam
30	0.28	0.219	1.36	0.15	0.15	2.16
40	0.18	0.175	1.45	1.39	0.10	3.30
50	0.126557	0.198	1.46	1.10	0.06	2.94
60	0.082812	0.080	0.81	0.61	0.03	1.62
70	0.0375	0.048	0.67	0.39	0.01	1.15
80	0.02344	0.023	0.45	0.22	0.01	0.72
90	0.01719	0.013	0.31	0.16	0.01	0.50
Tüm Etiketler	3.98	21.84	29.72	416.85	156.21	628.61

Çizelge 4.6 incelendiğinde Tüm Etiketler veri seti hariç tüm veri setlerinde başarılı sonuçların üretildiği görülmüştür. Ayrıca analizde kullanılan makine öğrenmesi yöntemleri arasında da birbirine çok yakın, tutarlı sonuçlar bulunmuştur. Ancak tüm bağımsız etiketlerin öznitelik vektörü olarak yer aldığı veri seti, hemen hemen tüm algoritmalar için başarısız sonuçlar vermiştir. Özellik seçim adımlarının kullanıldığı diğer yedi veri setinde ise sonuçlar birbirine oldukça yakın çıkmıştır. Tüm Etiketler veri setindeki algoritmaların doğruluk değerlerine bakıldığında en iyi performans ile en kötü performans arasındaki fark %78 iken, diğer yedi veri setinde fark sadece %5' tir. Algoritmaların uyum ve skor sürelerine bakıldığında Adaboost en yavaş olanıdır. Tüm Etiketler veri setinin her bir kat için ortalama toplam çalışma süresi Adaboost için 416 saniye iken, diğer gruptaki tüm sonuçlar arasında en yüksek ortalama toplam süre 2 saniyenin altında ölçülmüştür. Bu sonuçlara göre, tanımlanan etiketlerin büyük bir kısmı firmaya özel olduğu için, firmalar tarafından paylaşılan bazı, firmalar arasında ortak bağıntı bulunan, verilerin seçilmesi firmalar için ortak olan sınıflandırma problemlerinde performansı arttırdığı sonucuna varılmıştır. Rastgele Orman, tüm metrikler için yedi veri kümesinde en başarılı algoritmadır ve bunu Adaboost izlemiştir. Bu veri kümesi için en başarısız algoritma çifti K En Yakın Komşu ve İkinci Derece Ayırma Analizi olarak tespit edilmiştir. Tüm Etiketler veri setinde Karar Ağacı dışında hiçbir algoritma tüm metriklerde başarılı sonuçlara ulaşamamıştır. Güven aralığı değeri olarak %95 seçilmiştir. Bu aralıkta Çizelge 4.6' da tüm metrik ve algoritmalarda sapma değerlerinin düşük olduğu görülmektedir. Böylece 10 farklı kattaki değerler arasında anlamlı bir fark olmadığı

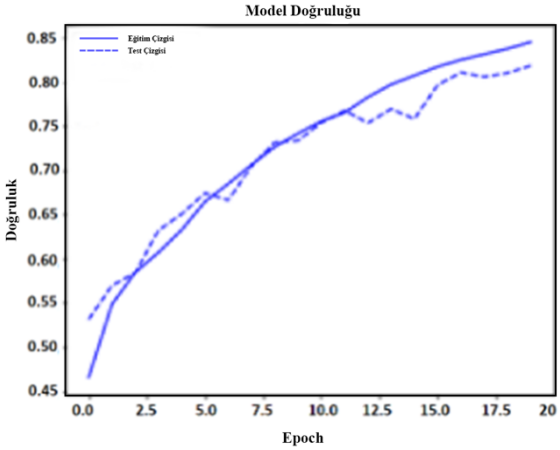
ortaya çıkmıştır. Öznitelik seçiminin uygulandığı yedi veri setine bakıldığında en başarılı sonuçlar %30 ve %50 veri setlerinde elde edilmiş ancak genel olarak anlamlı bir fark görülmemiştir. Çünkü yüzdeler arttıkça örnekleri temsil eden etiket sayısı azalsa da eksik bilgi sayısı da azalmaktadır. Böylece, düşük yüzdeli etiketler genellemeye izin verse de yüksek yüzdeli etiketler ayırt etmeyi kolaylaştırdığına kanaat getirilmiştir. Daha fazla veri ile daha iyi sonuç elde eden algoritmalarda, dosyalarda yüzdesi düşük olan etiketlerin kullanılması ve örnekler arasında güçlü ayırt edici özellikleri tercih eden algoritmalarda yüzdesi yüksek olan etiketlerin az da olsa seçilmesinin sınıflandırma doğruluğunu artıracağı sonucuna varılmıştır. Bu gruptaki en başarısız sonuçlar da çoğunlukla %90 veri setinden üretilmiştir. Bu aynı zamanda beklenen bir durumdur çünkü %90 veri kümesinde yalnızca üç etiket özellik vektöründe yer almıştır. Yine de bu üç etiket, incelenen 10K dosyalamaların en az %90'ında yer aldığı ve bu nedenle temsil yetenekleri yüksek olduğu için sınıflandırmada başarılı sonuçlar verebilmiştir. Sonuç olarak EDGAR veri tabanından indirilen ham verilerin bu çalışmada açıklanan yöntemlerle işlenerek makine öğrenmesi yöntemleriyle başarılı sonuçlar elde edilebilecek veri setlerine dönüştürülebileceği; Böylece, halka açık şirketlerin, düzenli olarak paylaştıkları 10K raporları aracılığıyla analiz edilebileceği kanıtlanmıştır.

4.3. Doc2Vec K Means CNN Hibrit Algoritma Sonuçları

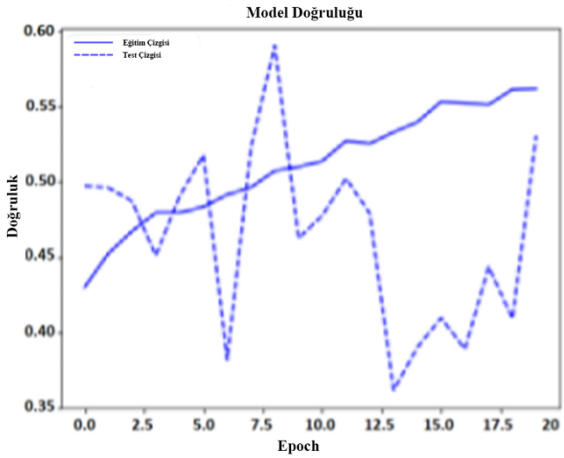
Halka açık, borsada işlem gören şirketlerin 2019-2022 yılları arasındaki toplamda on iki mali çeyreğe ilişkin yayınladıkları 10Q raporları, çalışmada oluşturulan model kullanılarak test edilmiştir. XBRL formatlı 10Q dosyalarında, 10K raporlarında olduğu gibi, şirket paylaşımları etiket-değer formatındadır. Hisse senetlerinden şirket bazında veri kümelerinin üretilmesindeki en büyük sorun, şirketler arasında tam etiket eşleştirmesidir. Bu nedenle oluşturulan modelin ilk bölümünde etiketlerin verimli bir şekilde kullanılabilmesi için birleştirilmesi ve etiketlerden yeni özelliklerin üretilmesi sağlanmıştır. Bu amaçla SEC tarafından verilen etiketler Doc2Vec yöntemi kullanılarak vektöre dönüştürülmüş ve elde edilen vektörler K Means kümeleme algoritması kullanılarak kümelenmiştir. Kümelenmiş etiketler ve şirketlerin fiyat gücünü gösteren PriceRank değerleri temel alınarak üretilen sınıflarla, her bir örnek bir şirket paylaşımını temsil edecek biçimde veri seti oluşturulmuştur. Hibrit modelin ikinci bölümünde, oluşturulan veri setini eğitmek için bir CNN modeli yaratılmıştır.

Tüm eğitim ve test süreçleri en az 10 kez çalıştırılmıştır. Eğitilen modeller doğruluk, kesinlik, geri çağırma ve f-skoru metrikleri ile değerlendirilmiştir. Tüm sonuçlar Çizelge 4.8' de verilmiştir. Ayrıca önerilen kümelenmiş veri setinin ve etiketlerle oluşturulan veri setinin CNN modelindeki doğruluk fonksiyon grafikleri Şekil 4.8 ve Şekil 4.9' da verilmiştir. Sonuçlar incelendiğinde, etiketler yerine önerilen yöntemle üretilen yeni özelliklerinin kullanımının algoritmaların performansında önemli farklılıklar yarattığı gözlemlenmiştir. Bu durum, finansal verilerin analizinde, mevcut paylaşımlardaki bilgilerden elde edilecek yeni çıkarımların önemli iyileştirmeler sağlayabileceğini göstermektedir. Etiketlerin vektöre dönüştürülmesi için önerilen Doc2Vec yöntemi, TFIDF ve Bert Base Nli Mean Token yöntemleriyle karşılaştırılmıştır. Kümeleme için kullanılan K Means yöntemi, DBSCAN yöntemi ile de karşılaştırılmıştır. Sonuçlara bakıldığında önerilen yöntemin diğerlerine göre çok daha başarılı sonuçlar ürettiği görülmektedir.

Bir şirketin mevcut 10Q raporlarından sonraki dönemdeki fiyat gücünü tahmin etmede, yani açıklanan finansal raporlar ile geleceğin öngörülmesinde, önerilen hibrit yöntem tüm veri kümelerinde en iyi sonuçları vermiştir. Nitekim yöntemin ikinci kısmını oluşturan CNN' in ile böyle raporların analizinde önemli bir fark yaratabileceği gösterilmiştir. Ayrıca, önerilen hibrit modelin performansı dikkat çekici bir şekilde öne çıkmaktadır. Firmaların fiyat gücünün belirlenmesinde önerilen hibrit yöntem %84 doğruluk değerine ulaşmıştır. En yakın sonuç DBSCAN kümeleme algoritmasının Doc2Vec yöntemi ile birleşiminden bir veri seti oluşturup CNN üzerinde çalıştırarak elde edilen %77 değeridir. Tüm bu bilgiler ışığında önerilen hibrit modelin hem veri seti üretim aşamasında hem de belirlenen CNN modelinde ne kadar başarılı olduğu açıkça görülmektedir. Çalışmada önerilen model, metinsel ve anlamsal olarak yakın ifade edilen birçok etiket üzerinden üretilecek veri setlerinde önemli bir potansiyel vaat etmektedir. Seyrek verileri analiz etmek için verimli, ölçeklenebilir ve sağlam bir yol sunmuştur.



Şekil 4.8 Önerilen hibrit yöntemle üretilen veri setinin CNN' de çalıştırılmasında validasyon doğruluk eğrisinin döngü sayısına göre değişimi



Şekil 4.9 Temel veri setinin CNN' de çalıştırılmasında doğruluk değerinin döngü sayısına göre değişimi

Çizelge 4.8 Tüm veri setlerinin farklı makine öğrenmesi yöntemleriyle analizinin sonuçları

Önerilen Doc2Vec-K Means Veri Seti				
	Doğruluk	**Hassasiyet**	**Geri Çağırma**	**Fscore**
CNN	0.84 ± 0.02	0.83 ± 0.02	0.84 ± 0.02	0.84 ± 0.02
Geriye Yayılım	0.73 ± 0.01	0.73 ± 0.01	0.73 ± 0.01	0.73 ± 0.01
Karar Ağacı	0.73 ± 0.01	0.747 ± 0.006	0.725 ± 0.001	0.736 ± 0.005
K En Yakın Komşu	0.675 ± 0.01	0.681 ± 0.004	0.67 ± 0.01	0.675 ± 0.007
QDA	0.72 ± 0.01	0.76 ± 0.01	0.72 ± 0.01	0.74 ± 0.01
TFIDF K Means Veri Seti				
CNN	0.54 ± 0.03	0.53 ± 0.08	0.54 ± 0.05	0.53 ± 0.04
Geriye Yayılım	0.48 ± 0.04	0.46 ± 0.03	0.52± 0.02	0.48 ± 0.02
Karar Ağacı	0.52 ± 0.01	0.50 ± 0.05	0.53 ± 0.01	0.51± 0.05

K En Yakın Komşu	0.52 ± 0.01	0.52 ± 0.03	0.52 ± 0.04	0.52 ± 0.03
QDA	0.50 ± 0.01	0.47 ± 0.05	0.53 ± 0.03	0.49 ± 0.07
Bert Base Nli Mean Tokens K Means Veri Seti				
CNN	0.67 ± 0.01	0.68 ± 0.02	0.68 ± 0.02	0.68 ± 0.03
Geriye Yayılım	0.58 ± 0.02	0.57 ± 0.03	0.58 ± 0.01	0.58 ± 0.01
Karar Ağacı	0.59 ± 0.02	0.584 ± 0.008	0.59 ± 0.03	0.58 ± 0.02
K En Yakın Komşu	0.56 ± 0.02	0.56 ± 0.05	0.565 ± 0.05	0.56 ± 0.02
QDA	0.58 ± 0.02	0.63 ± 0.02	0.54 ± 0.01	0.58 ± 0.01
Doc2Vec DBSCAN Veri Seti				
CNN	0.77 ± 0.03	0.76 ± 0.02	0.77 ± 0.05	0.773 ± 0.005
Geriye Yayılım	0.63 ± 0.04	0.63 ± 0.05	0.62± 0.02	0.62 ± 0.05
Karar Ağacı	0.62 ± 0.01	0.625 ± 0.005	0.60 ± 0.01	0.612± 0.005
K En Yakın Komşu	0.64 ± 0.03	0.63 ± 0.06	0.65 ± 0.03	0.63 ± 0.08
QDA	0.61 ± 0.01	0.63 ± 0.02	0.58 ± 0.01	0.62 ± 0.01
Tags Veri Seti				
CNN	0.528 ± 0.037	0.53 ± 0.03	0.53 ± 0.02	0.53 ± 0.04
Geriye Yayılım	0.40 ± 0.04	0.35 ± 0.05	0.50± 0.02	0.40 ± 0.05
Karar Ağacı	0.51 ± 0.01	0.525 ± 0.005	0.50 ± 0.01	0.512± 0.005
K En Yakın Komşu	0.50 ± 0.05	0.51 ± 0.05	0.49 ± 0.05	0.50 ± 0.05
QDA	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.01	0.50 0.01

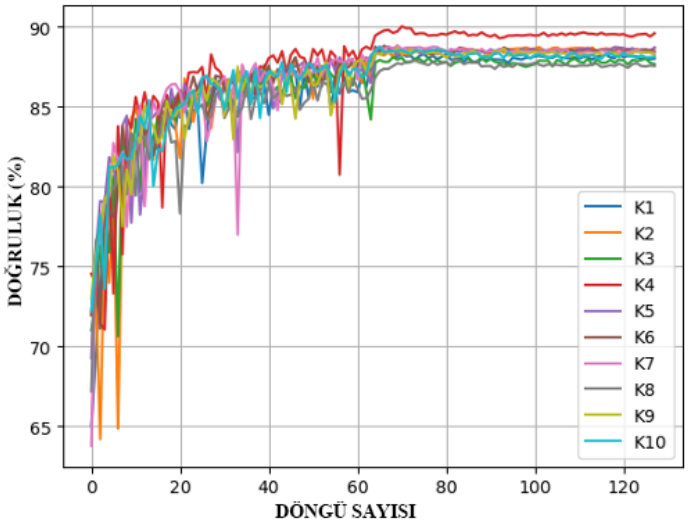
4.4. Tek Boyutlu CNN ve LSTM Temelli Hibrit Algoritma Sonuçları

Modelin değerlendirilmesinde 10 katlı çapraz doğrulama yöntemi uygulanmıştır. 10 parçaya bölünen veri setinin her bir parçası bir defa test kalanlarında eğitim setine dahil edilmiştir. Model bu yöntemle döngü değeri 128 olmak üzere, 10 ayrı zamanda çalıştırılmıştır. Her bir çalışmanın, sınıflandırma işlemi gerçekleştirilirken, her bir parça için ortalama doğruluk değerleri hesaplanmıştır. Ortalama değerler Çizelge 4.9' da verilmiştir. Birinci ve dokuzunca çalışmalar en iyi sonuçları üretmiştir. Şekil 4.10' da dokuzuncu çalışmanın her bir parça için ürettiği sınıflandırma doğruluk oranlarının döngüye göre değişim grafiği verilmiştir.

Çizelge 4.9 Her bir çalışmanın ortalama sınıflandırma doğruluk değerleri

Çalışma	Ortalama doğruluk (K = 10)
1	**88,338,203**
2	88,289,185

3	88,309,362
4	88,009,564
5	88,248,845
6	88,127,738
7	88,205,570
8	88,274,783
9	**88,338,203**
10	88,234,422



Şekil 4.10 Dokuzuncu çalışmanın her bir parça için doğruluk döngü değişimi

Her bir çalışmanın ortalama hassasiyet, geri çağırma ve F1-skor metrikleri açısından sınıflandırma başarımları da hesaplanmıştır. Ortalama değerler, sınıf ayrımları ile, Çizelge 4.10' da verilmiştir.

Çizelge 4.10 Sınıflandırma işlemi hassasiyet, geri çağırma ve F1-skor sonuçları

Çalışma	Sınıf	Hassasiyet	Geri Çağırma	F1-Skor	Örnek Sayısı
1	0	0.8943877	0.8958971	0.8951417	9481
	1	**0.85804**	0.8539865	**0.8560085**	13759
	2	0.9044802	**0.9083523**	0.9064121	11446
2	0	0.8944644	0.8930493	0.8937563	9481
	1	0.8544308	**0.857039**	0.8557329	13759
	2	**0.9076977**	0.9055565	0.9066258	11446
3	0	**0.8962634**	0.8930493	0.8946534	9481
	1	0.8548025	**0.857039**	0.8559193	13759
	2	0.9063265	0.9061681	0.9062473	11446
4	0	0.8940664	0.8884084	0.8912284	9481
	1	0.8508435	0.8540592	0.8524483	13759
	2	0.9038764	0.9045081	0.9041921	11446
5	0	0.8942755	0.8930493	0.8936619	9481
	1	0.8555023	0.8554401	0.8554712	13759
	2	0.9051483	0.9062555	0.9057016	11446
6	0	0.8960543	0.8910452	0.8935428	9481
	1	0.8538445	0.8547133	0.8542787	13759
	2	0.9020461	0.9051197	0.9035803	11446
7	0	0.8954435	0.8933657	0.8944034	9481
	1	0.8553647	0.8540592	0.8547114	13759
	2	0.9029506	0.9063428	0.9046436	11446
8	0	0.8936439	0.8942095	0.8939266	9481
	1	0.8549219	0.8552947	0.8551083	13759
	2	0.9072066	0.9062555	**0.9067308**	11446
9	0	0.8957566	**0.8972682**	**0.8965118**	9481
	1	0.8564983	0.8554401	0.8559689	13759
	2	0.9053901	0.9054692	0.9054296	11446
10	0	0.8930335	0.8964244	0.8947258	9481
	1	0.8541546	0.8547133	0.8544338	13759
	2	0.9074643	0.9038966	0.9056769	11446

Çizelge 4.11’ de, önerilen hibrit algoritmanın tüm çalışmalardaki sonuçlarının ortalamasının, önceki bölümde çalıştırılan diğer makine öğrenmesi yöntemleriyle doğruluk, hassasiyet, geri çağırma ve F1 skor metrikleri üzerinden karşılaştırılması verilmiştir.

Çizelge 4.11 Önerilen hibrit yöntemin ve diğer makine öğrenmesi yöntemlerinin ayrıntılı sonuçları

Yöntem	Hassasiyet	Geri Çağırma	F1-Skor	Doğruluk
QDA	0.76 ± 0.01	0.72 ± 0.01	0.74 ± 0.01	0.72 ± 0.01
KNN	0.681 ± 0.004	0.67 ± 0.01	0.675 ± 0.007	0.675 ± 0.01
Karar Ağacı	0.747 ± 0.006	0.725 ± 0.001	0.736 ± 0.005	0.73 ± 0.01
Geri Yayılım	0.73 ± 0.01	0.73 ± 0.01	0.73 ± 0.01	0.73 ± 0.01
CNN	0.83 ± 0.02	0.84 ± 0.02	0.84 ± 0.02	0.84 ± 0.02
CNN-LSTM	**0.8849 ± 0.0005**	**0.8848 ± 0.0007**	**0.8848 ± 0.0006**	**0.8823 ± 0.0006**

Sonuçlar incelendiğinde önerilen hibrit yöntemin diğer tüm yöntemlerden her açıdan üstün sonuçlar ürettiği görünmektedir. Hibrit yöntemin üretilmesinin altında yatan en önemli gerekçe, bir verideki hem kısa hem de uzun vadeli ilişkilerin tespit edilmesi için özellik çıkarımı adımının iyileştirilmesidir. Sonuçlara bakıldığında, yalnızca evrişimsel sinir ağı algoritmasına göre çok daha iyi sınıflandırma başarımı gösteren hibrit algoritma, veriyi tanıma ve tanımlama konusunda daha başarılı olduğunu kanıtlamıştır. Böylece, uzun vadeli bağlantıların bulunmasında LSTM ağından destek alan hibrit algoritma güçlü olduğunu göstermiştir. Özellikle, çalışmada kullanılan veri seti gibi seyrek veri setlerinde, önerilen hibrit algoritmanın fark yaratabileceği gösterilmiştir.

5. SONUÇLAR VE ÖNERİLER

5.1. Sonuçlar

Günümüzde, tüm sektörlerde veri üretimi, saklanması ve analizi oldukça önemli hale gelmiştir. Bu yüzden, dijital dönüşümle birlikte artık üretilen ve depolanan veri miktarı, her geçen gün büyük bir hızla artmaktadır. Firmaların işleyişini, işlevsel faaliyetlerini, rekabet ortamlarını, verimliliklerini oldukça etkiyen bu değişim ve dönüşüm birçoğu için kaçınılmaz hale gelmiştir. Finans alanı da bu trendden etkilenmiştir. Ancak, geleneksel istatistiksel yöntemler artık kontrolsüz genişleme ve karmaşıklık nedeniyle etkili olamamaktadır. Bu nedenle, finansal verilerin temizlenmesi ve analiz edilmesi için modern makine öğrenimi yöntemleri kullanmanın ne denli önemli olduğu birçok akademik çalışmayla kanıtlanmıştır. Bu tez çalışmasında, finansal raporlardan yeni veri setleri üreten ve bunları makine öğrenimi yöntemleriyle analiz eden yenilikçi yaklaşımlar sunulmaktadır. Çalışmada, 10K yıllık raporları üzerinde çalışılmış ve bunların analiz edilebilir veri setlerine dönüştürülmesi sağlanmıştır. Makine öğrenimi yöntemleri kullanılarak sınıflandırma işlemi gerçekleştirilmiş ve farklı algoritmalarla başarılı sonuçlar elde edilmiştir. Doğruluk, kesinlik, geri çağırma gibi kabul görmüş sınıflandırma metrikleriyle de sonuçların başarısı kıyaslanmış ve teyit edilmiştir. Algoritmalardan üretilen başarılı sonuçlar (%92 doğruluk) bu veri dönüşümünün önemini göstermiştir. Ayrıca, 10K raporlarındaki karmaşık ve hatalı veriler üzerinde doğal dil işleme teknikleri uygulanmış ve veri boyutunun azaltılması, seyrekliğin giderilmesi, hatalı verilerin tespit edilmesi gibi konularda benzersiz yeni yaklaşımlar sunulmuştur. Tez çalışmasının bir diğer ayağı ise 10Q çeyrek raporlarının analizine odaklanmıştır. Bu kapsamda, metinsel içeriğe sahip çeyrek raporlarının vektöre dönüştürülmesi için Doc2Vec ve K Means kümeleme algoritmaları birleştirilerek verimli veri setleri oluşturulmuş ve evrişimsel sinir ağı kullanılarak üretilen setlerde sınıflandırma işlemi gerçekleştirilmiştir. Sınıflandırma yöntemi ile, verileri kullanılan firmaların çeyrek raporlarına bakılarak fiyat gücünün tahmini gerçekleştirilmiştir. Hibrit algoritmanın tüm aşamaları literatürdeki diğer yöntemlerle kıyaslanarak sonuçlar anlamlandırılmıştır. Önerilen yöntemle elde edilen sonuçlar (%84 doğruluk, %83 hassasiyet, %84 geri çağırma) bu hibrit algoritmanın etkinliğini göstermiştir. Son olarak, evrişimsel sinir ağı ve uzun kısa süreli bellek algoritmalarının güçlü yanlarını ortaya çıkaran hibrit bir yöntem sunulmuştur. Seyrek ve karmaşık yapıya sahip finansal verilerin analizi için, güçlü, veri

içerisindeki doğru ya da dolaylı bağlantıları çözümleyen, desenleri ortaya çıkaran bir yöntemin üretilmesi hedeflenmiştir. Doc2Vec ve K Means yöntemleriyle üretilen veri seti, önerilen yöntemle sınıflandırılmıştır. Elde edilen sonuçlarla (%88 doğruluk, %88 hassasiyet, %88 geri çağırma), finansal verilerin analizinde, başarılı, güçlü bir yöntemin üretildiği ortaya konulmuştur.

5.2. Öneriler

Bu çalışmada sunulan yöntemler, gelir tabloları, bilançolar ve nakit akış tabloları gibi diğer finansal rapor analizleri için uygulanabilir. Ayrıca, önerilen yöntemler, metin içerikli verilerin olduğu diğer birçok alanda kullanılabilir. Üretilen başarılı sonuçlar, farklı yapay zekâ yaklaşımlarının finans alanında oldukça önemli bir potansiyele sahip olduğunu göstermektedir. Önerilen yöntemler, halka açık bir şekilde sunularak diğer çalışmalarda kullanımı da sağlanacaktır. Böylece, çalışmada sunulan modellerin iyileştirilmesi, hiper parametrelerinin değiştirilmesi ile de yine yeni ve daha başarılı sonuçların elde edilmesinin önü açılmıştır. Finans alanındaki büyük verinin anlamlı veri setine dönüşüm sürecinde, önemli bir potansiyelinin olduğu, ciddi tasarruflar saylayabileceği ve anlamlı, değerli sonuçların üretilebileceği kanıtlanarak bu konularda yeni ufukların açılabileceği öngörülmektedir.

Sonuç olarak, tez kapsamında gerçekleştirilen çalışmalar, makine öğrenmesi yöntemlerinin henüz yeni uygulanmaya başlandığı finans alanında yapılmıştır. Veri üretim sürecinde, temel hedef yapay zekâ uygulamaları ile analiz olmadığı için verilerin analiz edilebilir veri setlerine dönüşümü çalışmalarda karşılaşılan en büyük zorluk olmuştur. Bu yüzden, çalışmanın büyük bölümünde bu noktaya odaklanılmıştır. Çalışma kapsamında, metinsel verilerden veri seti üretimini ve derin öğrenme yöntemleriyle analizini mümkün kılan hibrit bir algoritma üretilerek literatüre katkı sağlanmıştır.

KAYNAKLAR

Abdallah, M., An Le Khac, N., Jahromi, H., and Delia Jurcut, A., 2021, A Hybrid CNN-LSTM Based Approach for Anomaly Detection Systems in SDNs, *Proceedings of the 16th International Conference on Availability, Reliability and Security*, 1–7. https://doi.org/10.1145/3465481.3469190

Agga, A., Abbou, A., Labbadi, M., Houm, Y. el, and Ou Ali, I. H., 2022, CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production, *Electric Power Systems Research*, *208*, 107908. https://doi.org/10.1016/j.epsr.2022.107908

Aggarwal, C. C., 2016, Data classification Algorithms and Applications. *Chapman and Hall/CRC*.

Alles, M., and Gray, G. L., 2016, Incorporating big data in audits: Identifying inhibitors and a research agenda to address those inhibitors. *International Journal of Accounting Information Systems*, *22*, 44-59. https://doi.org/10.1016/j.accinf.2016.07.004

Ayyadevara, V. K., 2018, Convolutional Neural Network, *Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python and R* (ss. 179-215)., https://doi.org/10.1007/978-1-4842-3564-5_9

Balsam, S., Bartov, E., and Marquardt, C., 2002, Accruals management, investor sophistication, and equity valuation: Evidence from 10-Q filings. *Journal of Accounting Research*, *40*(4), 987-1012, https://doi.org/10.1111/1475-679X.00079

Bholowalia, P., and Kumar, A., 2014, EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. , *International Journal of Computer Applications* (C. 105, Sayı 9).

Bonson, and Escobar., 2002, A Survey on Voluntary Disclosure on the Internet. Empirical Evidence from 300 European Union Companies, *The International Journal of Digital Accounting Research*, *2*(1), 27-51, https://doi.org/10.4192/1577-8517-v2_2

Boritz, J. E., Hayes, L., and Lim, J. H., 2013, A content analysis of auditors' reports on IT internal control weaknesses: The comparative advantages of an automated approach to control weakness identification, *International Journal of Accounting Information Systems*, *14*(2), 138-163, https://doi.org/10.1016/j.accinf.2011.11.002

Bragg, S. M., 2004, GAAP Implementation Guide, *John Wiley & Sons, Inc.*

Brown-Liburd, H., Issa, H., and Lombardi, D., 2015, Behavioral implications of big data's impact on audit judgment and decision making and future research directions, *Accounting Horizons*, *29*(2), 451-468, https://doi.org/10.2308/acch-51023

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., and Varoquaux, G., 2013, *API design for machine learning software: experiences from the scikit-learn project*.

Chen, C. L., Liu, C. L., Chang, Y. C., and Tsai, H. P., 2011, Exploring the relationships between annual earnings and subjective expressions in US financial statements, *Proceedings - 2011 8th IEEE International Conference on e-Business Engineering, ICEBE 2011*, 1-8, https://doi.org/10.1109/ICEBE.2011.47

Chen, X., Cho, Y. H., Dou, Y., and Lev, B., 2022, Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data, *Journal of Accounting Research*, *60*(2), 467-515, https://doi.org/10.1111/1475-679X.12429

Chen, X., Cho, Y. H., Dou, Y., and Lev, B. I., 2021, Fundamental Analysis of XBRL Data: A Machine Learning Approach. *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.3741015

Chychyla, R., and Kogan, A., 2015, Using XBRL to conduct a large-scale study of discrepancies between the accounting numbers in compustat and SEC 10-K filings, *Journal of Information Systems*, *29*(1), 37-72, https://doi.org/10.2308/isys-50922

Cunningham, L. M., and Leidner, J. J., 2019, The SEC Filing Review Process: Insights from Accounting Research, *SSRN Electronic Journal* (May), https://doi.org/10.2139/ssrn.3494830

Davis, F. D., 1989, Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology, *MIS Quarterly*, *13*(3), 319, https://doi.org/10.2307/249008

Dhole, S., Lobo, G. J., Mishra, S., and Pal, A. M., 2015, Effects of the SEC's XBRL mandate on financial reporting comparability, *International Journal of Accounting Information Systems*, *19*, 29-44, https://doi.org/10.1016/j.accinf.2015.11.002

Earley, C. E., 2015, Data analytics in auditing: Opportunities and challenges. *Business Horizons*, *58*(5), 493-500, https://doi.org/10.1016/j.bushor.2015.05.002

Efendi, J., Park, J. D., and Subramaniam, C., 2016, Does the XBRL Reporting Format Provide Incremental Information Value? A Study Using XBRL Disclosures During the Voluntary Filing Program, *Abacus*, *52*(2), 259-285, https://doi.org/10.1111/abac.12079

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., 1996, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226-231.

Fernández, M., Vallet, D., and Castells, P., 2006, Probabilistic score normalization for rank aggregation, *Lecture Notes in Computer Science (including subseries Lecture*

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *3936 LNCS*, 553-556, https://doi.org/10.1007/11735106_63

Fisher, I., Garnsey, M., and Hughes, M., 2016, Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research, *Intelligent Systems in Accounting, Finance and Management*, *23*, n/a-n/a. https://doi.org/10.1002/isaf.1386

Fischer, T., and Krauss, C., 2018, "Deep learning with long short-term memory networks for financial market predictions,"*European Journal of Operational Research, Elsevier*, vol. 270(2), 654-669.

Fushiki, T., 2011, Estimation of prediction error by using K-fold cross-validation, *Statistics and Computing*, *21*(2), 137-146, https://doi.org/10.1007/s11222-009-9153-8

Garnsey, M. R., 2006, Automatic Classification of Financial Accounting Concepts, *JOURNAL OF EMERGING TECHNOLOGIES IN ACCOUNTING* (C. 3).

Gunn, J., 2007, XBRL: Opportunities and Challenges in Enhancing Financial Reporting and Assurance Processes, *Current Issues in Auditing*, *1*(1), A36-A43, https://doi.org/10.2308/ciia.2007.1.1.a36

Haider, M. M., and Mahi, H. R., 2020, *Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm*, *April*.

Henselmann, K., Ditter, D., and Scherr, E., 2015, Irregularities in accounting numbers and earnings management-a novel approach based on SEC XBRL filings, *Journal of Emerging Technologies in Accounting*, *12*(1), 117-151, https://doi.org/10.2308/jeta-51247

Hochreiter, S., Schmidhuber, J., 1997, Long Short-Term Memory, *Neural Comput*, 9 (8): 1735–1780.

Hoitash, R., and Hoitash, U., 2018, Measuring accounting reporting complexity with XBRL. *Accounting Review*, *93*(1), 259-287, https://doi.org/10.2308/accr-51762

Jain, A., Kulkarni, G., and Shah, V., 2018, Natural Language Processing Aditya, *International Journal of Computer Sciences and Engineering*, *1*, 7, https://doi.org/10.11604/pamj.2014.17.246.2230

Kamaruddin, S. S., Bakar, A. A., Hamdan, A. R., Nor, F. M., Nazri, M. Z. A., Othman, Z. A., and Hussein, G. S., 2015, A text mining system for deviation detection in financial documents, *Intelligent Data Analysis*, *19*(S1), S19-S44, https://doi.org/10.3233/IDA-150768

Kang, T., Park, D. H., and Han, I., 2018, Beyond the numbers: The effect of 10-K tone on firms' performance predictions using text analytics, *Telematics and Informatics*, *35*(2), 370-381, https://doi.org/10.1016/j.tele.2017.12.014

Kearney, C., and Liu, S., 2014, Textual sentiment in finance: A survey of methods and models, *International Review of Financial Analysis*, *33*(Cc), 171-185, https://doi.org/10.1016/j.irfa.2014.02.006

Kim, Y., 2014, Convolutional neural networks for sentence classification. *arXiv preprint arXiv*:1408.5882.

Kim, C., Wang, K., and Zhang, L., 2019, Readability of 10-K Reports and Stock Price Crash Risk, *Contemporary Accounting Research*, *36*(2), 1184-1216, https://doi.org/10.1111/1911-3846.12452

Krieger, F., and Drews, P., 2018, Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy Agile Governance of Corporations in the light of Digital Transformation View Project*, Big Data in Auditing View Project*, https://www.researchgate.net/publication/328902212

Lafferty, J. D., McCallum, A., and Pereira, F. C. N., 2001, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of the Eighteenth International Conference on Machine Learning*, 282-289.

Lavesson, N., and Davidsson, P., 2007, Evaluating learning algorithms and classifiers, *International Journal of Intelligent Information and Database Systems*, *1*(1), 37-52, https://doi.org/10.1504/IJIIDS.2007.013284

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., 1998, Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, *86*(11), 2278-2323, https://doi.org/10.1109/5.726791

Lee, Y. J., 2012, The Effect of Quarterly Report Readability on Information Efficiency of Stock Priceś, *Contemporary Accounting Research*, *29*(4), 1137-1170, https://doi.org/10.1111/j.1911-3846.2011.01152.x

Li, F., 2010, The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach, *Source: Journal of Accounting Research* (C. 48, Sayı 5).

Lilhore, U. K., Dalal, S., Faujdar, N., Margala, M., Chakrabarti, P., Chakrabarti, T., Simaiya, S., Kumar, P., Thangaraju, P., and Velmurugan, H., 2023, Hybrid CNN-LSTM model with efficient hyperparameter tuning for prediction of Parkinson's disease, *Scientific Reports*, *13*(1), 14605. https://doi.org/10.1038/s41598-023-41314-y

Liu, X., 2013, Full-Text Citation Analysis: A New Method to Enhance, *Journal of the American Society for Information Science and Technology*, *64*(July), 1852-1863, https://doi.org/10.1002/asi

Loughran, T., and Mcdonald, B., 2016, Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research*, *54*(4), 1187-1230, https://doi.org/10.1111/1475-679X.12123

Loughran, T., and McDonald, B., 2017, The Use of EDGAR Filings by Investors, *Journal of Behavioral Finance*, *18*(2), 231-248, https://doi.org/10.1080/15427560.2017.1308945

Loukas, L., Fergadiotis, M., Chalkidis, I., Spyropoulou, E., Malakasiotis, P., Androutsopoulos, I., and Paliouras, G., 2022, *FiNER: Financial Numeric Entity Recognition for XBRL Tagging*. 4419-4431, https://doi.org/10.18653/v1/2022.acl-long.303

Lu, W., Li, J., Li, Y., Sun, A., and Wang, J., 2020, A CNN-LSTM-Based Model to Forecast Stock Prices, *Complexity*, *2020*, 1–10. https://doi.org/10.1155/2020/6622927

Magnusson, C., Arppe, A., Eklund, T., Back, B., Vanharanta, H., and Visa, A., 2005, The language of quarterly reports as an indicator of change in the company's financial status, *Information and Management*, *42*(4), 561-574, https://doi.org/10.1016/j.im.2004.02.008

Mcdonald, B., 2009, *Plain English, Readability , and 10-K Filings*. *September*.

Mendelson, S., and Smola, A. J. (Ed.)., 2003, Advanced Lectures on Machine Learning, *Springer Berlin Heidelberg*, https://doi.org/10.1007/3-540-36434-X

Moon, T. K., 1996, The expectation-maximization algorithm, *IEEE Signal Processing Magazine*, *13*(6), 47-60. https://doi.org/10.1109/79.543975

Papadimitriou, C. H., Tamaki, H., Raghavan, P., and Vempala, S., 1998, Latent semantic indexing. *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems -PODS '98*, 159-168, https://doi.org/10.1145/275487.275505

Peterson, K., Schmardebeck, R., and Wilks, T. J., 2015, The earnings quality and information processing effects of accounting consistency, *Accounting Review*, *90*(6), 2483-2514, https://doi.org/10.2308/accr-51048

Plumlee, R. D., and Plumlee, M. A., 2008, Assurance on XBRL for financial reporting, *Accounting Horizons* (C. 22, Sayı 3, ss. 353-368), https://doi.org/10.2308/acch.2008.22.3.353

Pustokhina, I. V., Pustokhin, D. A., Rodrigues, J. J. P. C., Gupta, D., Khanna, A., Shankar, K., Seo, C., and Joshi, G. P., 2020, Automatic Vehicle License Plate Recognition Using Optimal K-Means with Convolutional Neural Network for Intelligent Transportation Systems, *IEEE Access*, *8*, 92907-92917, https://doi.org/10.1109/ACCESS.2020.2993008

Rao, Y., and Guo, K. H., 2022, Does XBRL help improve data processing efficiency? *International Journal of Accounting and Information Management*, *30*(1), 47-60, https://doi.org/10.1108/IJAIM-07-2021-0155

Reimers, N., and Gurevych, I., 2019, *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*.

Rumelhart, D.E., Hinton, G.E., ve Williams, R.J., 1986, Learning internal representations by error propagation.

Rustam, Z., Hartini, S., Pratama, R. Y., Yunus, R. E., and Hidayat, R., 2020, Analysis of architecture combining Convolutional Neural Network (CNN) and kernel K-means clustering for lung cancer diagnosis, *International Journal on Advanced Science, Engineering and Information Technology*, *10*(3), 1200-1206, https://doi.org/10.18517/ijaseit.10.3.12113

SEC., 2021, *Securities Exchange Act Of 1934*.

SEC., 2022, *PriceRank*. SEC. https://www.sec.gov/opa/data/market-structure/market-structure-data-security-and-exchange [Ziyaret Tarihi: 14 Mayıs 2022].

Shang L., Zhang Z., Tang F., Cao Q., Pan H., and Lin Z., 2023, CNN-LSTM Hybrid Model to Promote Signal Processing of Ultrasonic Guided Lamb Waves for Damage Detection in Metallic Pipelines. *Sensors*, 23(16):7059. https://doi.org/10.3390/s23167059

Sharifrazi, D., Alizadehsani, R., Joloudari, J. H., Shamshirband, S., Hussain, S., Sani, Z. A., Hasanzadeh, F., Shoaibi, A., Dehzangi, A., and Alinejad-Rokny, H., 2020, CNN-KCL: Automatic Myocarditis Diagnosis using Convolutional Neural Network Combined with K-means Clustering, https://doi.org/10.20944/preprints202007.0650.v1

S&P Global., 2020, *Index Attributes*. https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview, [Ziyaret Tarihi: 22 Eylül 2021]

SParck Jones, K., 1972, A Statistical Interpretation Of Term Specificity And Its Application In Retrieval. *Journal of Documentation*, *28*(1), 11-21, https://doi.org/10.1108/eb026526

Stone, H., Sidel, J., Oliver, S., Woolsey, A., and Singleton, R. C., 2004, Sensory Evaluation by Quantitative Descriptive Analysis. *Descriptive Sensory Analysis in Practice* (ss. 23-34). Wiley, https://doi.org/10.1002/9780470385036.ch1c

Thakkar, A. and Chaudhari, K., 2021, A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions, *Expert Systems with Applications*, **177**: p. 114800.

Van Den Bogaerd, M., and Aerts, W., 2011, Applying machine learning in accounting research, *Expert Systems with Applications*, *38*(10), 13414-13424, https://doi.org/10.1016/j.eswa.2011.04.172

Vasarhelyi, M. A., Chan, D. Y., and Krahel, J. P., 2012, Consequences of XBRL standardization on financial statement data, *Journal of Information Systems*, *26*(1), 155-167, https://doi.org/10.2308/isys-10258

Xishuang, D., Lijun, Q., and Lei, H., 2017, Short-term load forecasting in smart grid: A combined CNN and K-means clustering approach, *2017 IEEE International Conference on Big Data and Smart Computing, BigComp 2017*, 119-125, https://doi.org/10.1109/BIGCOMP.2017.7881726