YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

FİNANSAL HABER MAKALELERİ KULLANILARAK HİSSE SENETLERİ FİYAT DEĞİŞİMLERİNİN TAHMİN EDİLMESİ

Bilgisayar Müh. Mustafa İdris Yasef KAYA

FBE Bilgisayar Mühendisliği Anabilim Dalında Hazırlanan

YÜKSEK LİSANS TEZİ

Tez Danışmanı : Yrd. Doç. Dr. M. Elif KARSLIGİL

İSTANBUL, 2010

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

$a_r(x_i)$	i. x örneğinin r. özellikteki değeri
b	Ofset değeri
$BK(i,c)$	i özelliğinin c sınıfındaki bilgi kazanım değeri
d	Polinom kernel fonksiyonu derece parametresi
$d(x_i, x_j)$	x_i ve x_j örnekleri arasındaki uzaklık
DN	Nagatif olarak tahmin edilen negatif örnek sayısı
DP	Pozitif olarak tahmin edilen pozitif örnek sayısı
$DS(i,c)$	i özelliğinin c sınıfına ait doküman sayısı
$EO(i,c)$	i özelliğinin c sınıfındaki eşitsizlik oranı
GCO	Geri çağırma oranı
$GSS(i,c)$	i özelliğinin c sınıfındaki GSS katsayısı değeri
κ	Sigmoid kernel fonksiyonu parametresi
$K(x, y)$	Kernel fonksiyonu
KO	Kesinlik oranı
$KK(i,c)$	i özelliğinin c sınıfındaki korelasyon katsayısı değeri
L	Lagrangian değeri
m	Marj değeri
N	Toplam doküman sayısı
$OB(i,c)$	i özelliğinin c sınıfındaki ortak bilgi değeri
$P(c)$	c sınıfına ait doküman sayısı
$P(\bar{c})$	c sınıfına ait olmayan doküman sayısı
$P(i)$	i özelliğini bulunduran doküman sayısı
$P(\bar{i})$	i özelliğini bulundurmayan doküman sayısı
$P(i,c)$	i özelliğini bulunduran c sınıfına ait doküman sayısı
$P(i,\bar{c})$	i özelliğini bulunduran c sınıfına ait olmayan doküman sayısı
$P(\bar{i},c)$	i özelliğini bulundurmayan c sınıfına ait doküman sayısı
$P(\bar{i},\bar{c})$	i özelliğini bulundurmayan ve c sınıfına ait olmayan doküman sayısı
P_{t-1}	(t-1). güne ait hisse senedi fiyatı
P_t	t. güne ait hisse senedi fiyatı
$TG(i,c)$	i özelliğinin c sınıfındaki terim gücü değeri
w	Normal vektör
x	Sınıflandırma örneği
x_i	i. sınıflandırma örneği
$X^2(i,c)$	i özelliğinin c sınıfındaki ki-kare istatistik değeri
y_i	i. sınıf değeri
YN	Negatif olarak tahmin edilen pozitif örnek sayısı
YP	Pozitif olarak tahmin edilen negatif örnek sayısı
α	Lagrange çarpanı
σ	Radyal tabanlı kernel fonksiyonu genişlik parametresi
θ	Sigmoid kernel fonksiyonu parametresi
ΔP	Hisse senedi günlük fiyat değişimi

KISALTMA LİSTESİ

HTML	Hyper Text Markup Language
KDD	Knowledge Discovery in Databases (Veri Tabanlarında Bilgi Keşfi)
KNN	k-Nearest Neighbor (k-En Yakın Komşuluk)
MSFT	Microsoft Şirketi'nin NASDAQ borsasındaki hisse senedi kodu
NewsCATS	News Categorization and Trading System (Haber Kategorilendirme ve Yatırım Sistemi)
OpenNLP	Open Natural Language Processing
ROC	Receiver Operating Characteristic (Alıcı İşletim Karakteristiği)
SVM	Support Vector Machines (Destek Vektör Makineleri)
tf*idf	Term Frequency Inverse Document Frequency (Kelime Frekansı Ters Doküman Frekansı)
URL	Uniform Resource Locator

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖNSÖZ

Bu çalışmanın konusu finansal haber makalelerinin analiz edilerek hisse senetleri fiyatlarında ileriye dönük gerçekleşebilecek değişikliklerin tahmin edilmesidir.

Bu tezin hazırlanması esnasında bilgi birikimini ve zamanını sonuna kadar benimle paylaşan değerli hocam ve tez danışmanım Yrd. Doç. Dr. M. Elif KARSLIGİL'e en içten duygularımla teşekkür ediyorum. Ayrıca yüksek lisans eğitimim boyunca bana sağladığı katkılarından dolayı Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı öğretim üyelerine teşekkürlerimi sunuyorum.

Son olarak eğitim hayatım boyunca desteklerini benden esirgemeyen aile bireylerime teşekkür ederim.

ÖZET

Hisse senetleri fiyatlarında gerçekleşebilecek değişikliklerin tahmin edilmesi gerek akademik gerekse finansal çevrelerde üzerinde en çok durulan konulardan birisidir. Bu problemin çözümü amacıyla yapılan çalışmalarda veri madenciliği tekniklerinden sıklıkla faydalanılmaktadır. Bu çalışmalarda zengin içerikli finansal haber bilgileri ile hisse senedi fiyatları bilgilerinin birlikte kullanılması ile başarılı sonuçlar elde edilmiştir.

Bu çalışmada metin formatındaki finansal haber makalelerinin analiz edilerek hisse senetleri fiyatlarında ileriye dönük gerçekleşebilecek değişikliklerin tahmin edilmesi amaçlanmıştır. Geçmişe yönelik bilgilerden faydalanılarak finansal haber makalelerinin içerikleri ile hisse senetleri fiyatları arasındaki ilişki analiz edilerek ileriye yönelik tahminler yapılması üzerine bir model geliştirilmiştir.

Çalışma kapsamında son bir yıl içerisinde Microsoft şirketi hakkında yayınlanmış finansal haber makaleleri ve yine aynı dönem içerisinde Microsoft şirketinin günlük hisse senetleri fiyatları bilgilerinden bir veritabanı oluşturulmuştur. Haber makaleleri yayınlandıkları tarihten itibaren hisse senedi fiyatında gerçekleşen değişikliğe göre olumlu veya olumsuz olarak etiketlenmiştir. Olumlu olarak etiketlenen haberlerin hisse senedi fiyatında artış etkisi, olumsuz olarak etiketlenen haberlerin ise düşüş etkisi yarattığı kabul edilmiştir. Haber makaleleri analiz edilirken kelimeleri ayrı ayrı değerlendirmek yerine, haberlerin özelliklerini belirlemek için aynı cümlede geçen biri isim diğeri fiil türünden iki kelime birlikte kullanılmıştır. Özellik olarak belirlenen isim ve fiil kelime ikililerinden ki-kare istatistik yöntemi ile sınıflandırmada en ayırt edici olanlar seçilmiştir. Veritabanındaki haber makaleleri, seçilen özellikler üzerinden destek vektör makinesi ve k-en yakın komşuluk yöntemleri ile sınıflandırılmıştır. Sınıflandırma sonucunda elde edilen tahminler ile sistemin yatırım stratejileri üretmesi sağlanmıştır.

Sistemin doğrulamasını yapmak amacı ile çeşitli senaryolar denenerek sistemin çıktıları değerlendirilmiştir. Elde edilen sonuçlar, finansal haber makalelerinin içerikleri ile hisse senetleri fiyatları arasında güçlü bir ilişki olduğunu ispat etmektedir.

Anahtar kelimeler: Hisse senedi fiyatı tahmini, Veri madenciliği, Metin madenciliği, Metin sınıflandırma.

ABSTRACT

Stock price prediction is one of the most important issues to be investigated in academic and financial researches. Data mining techniques are frequently involved in the works aimed to achieve this problem. When rich financial news information is used together with stock price information in these works, significant successful results have been obtained.

In this thesis, to predict stock price movements in the future by analyzing textual financial news articles is aimed. A prediction model to find and analyse correlation between contents of news articles and stock prices and then make predictions for future prices was developed.

Financial news articles published in last year about Microsoft Company has been retrieved. Also past stock prices of the Microsoft Company has been obtained. All articles has been labeled as positive or negative according to their effects on stock price. While positive labeled articles cause rising in stock price, negative labeled ones drop the price. While analyzing textual data, we used word couples consisting a noun and a verb as features instead of using individual words. Most discriminative features for classification has been selected according to chi-square statistics algorithm. Support vector machines and k nearest neighbor classifiers have been trained with labeled train articles. Then classes of test articles have been predicted with using the model resulted from train phase. Then investment recommendations have been generated according to the prediction results.

In order to validate the prediction system, various scenarios has been exploited and outputs of the system has been evaluated. Obtained results show that there is a strong relationship between financial news and stock price movements.

Keywords: Stock price prediction, Data mining, Text mining, Text categorization.

1. GİRİŞ

Son yıllarda bilgi teknolojilerindeki hızlı gelişmeler sayesinde çok büyük miktardaki veriler sayısal ortamlarda saklanabilmektedir. Bu büyük miktardaki verilerin insanlar tarafından işlenip anlamlı bilgiler çıkarılması işlemi oldukça zahmetli ve zaman alan bir iştir. Bu işlemin bilgisayarlar tarafından yapılmasının amaçlanması veri madenciliği disiplinini ortaya çıkarmıştır. Sayısal ortamlardaki verilerin büyük bir kısmı metin formatında saklanmaktadır. Özellikle internet teknolojilerindeki yenilikler ile birlikte metin formatındaki sınırsız veri internet üzerinden erişilebilir durumdadır. Metin formatındaki yapısal olmayan bu verilerin analiz edilerek anlamlı bilgiler çıkarılması işlemine metin madenciliği denilmektedir.

Hisse senetleri fiyatlarında gerçekleşebilecek değişikliklerin tahmin edilmesi konusunda şimdiye kadar birçok çalışma gerçekleştirilmiştir. Bu çalışmalarda genellikle hisse senetlerinin geçmiş dönem fiyatlarının analiz edilerek geleceğe yönelik tahminler yapılması amaçlanmıştır. Son yıllarda geçmiş dönem fiyatları ile birlikte finansal haber makaleleri de bu çalışmalarda girdi olarak kullanılmaya başlamıştır. Bu şekilde metin formatındaki finansal haber makalelerinin zengin içeriğinden de faydalanılmıştır.

İnternet üzerinde hisse senetleri yatırımcılarına yön vermek amacı ile finansal haberler yayınlayan bir çok kaynak yer almaktadır. Bu kaynaklarda hem geçmişe yönelik hem de güncel olarak şirketlerden haberler, ulusal ve global ekonomik gelişmeler ile hisse senetleri fiyatları sunulmaktadır. Bu kadar çok bilginin bilgisayar ortamında genel erişime açık olması bu konuda yapılacak veri madenciliği çalışmalarına elverişli bir zemin hazırlamıştır.

Hisse senedi fiyatları doğası gereği birçok gelişmeden etkilenmektedir. Çünkü hisse senetleri fiyatlarını yaptıkları işlemler ile yatırımcılar belirlemektedir. Yatırımcılar da hisse senetleri üzerindeki işlemlerini yaparken finansal gelişmelerin yer aldığı haberlerden faydalanmaktadır. Fakat birçok kaynakta yer alan çok fazla sayıdaki haberi analiz etmek ve bu doğrultuda yatırım kararı vermek oldukça zor ve vakit alan bir işlemdir. Üstelik finansal gelişmelere ait haberlerin hisse senetleri fiyatları üzerindeki etkisi çok kısa süreli olmaktadır. Bu nedenle yatırımcıların tüm gelişmeleri analiz edecek vakitleri bulunmamaktadır. Bu çalışmada internet üzerinde yer alan finansal veriler bilgisayar yardımı ile analiz edilerek hisse senedi yatırımcılarına yön vermek amaçlanmıştır. Bu kapsamda finansal haberlerin metin madenciliği yöntemleri kullanılarak hisse senedi fiyatlarına etkisinin olumlu veya olumsuz olarak sınıflandırılması amaçlanmıştır.

Bu çalışma, başta portföy yöneticileri olmak üzere finans sektörü çalışanlarının veya

birikimlerini hisse senetleri piyasasında değerlendirmek isteyen yatırımcıların yatırım portföylerini belirlemesine yardımcı olacaktır. Yatırımcıların hisse senetlerini doğru zamanda alması veya elinde bulundurdukları hisse senetlerini doğru zamanda ellerinden çıkarması getiri elde edebilmeleri açısında büyük önem arz etmektedir. Hisse senetleri piyasasında getiri elde edebilmek için temel prensip belirli bir fiyattan alınan hisse senedinin fiyatının belirli bir artış kaydettikten sonra satılmasıdır. Hisse senedi fiyatı artışa geçtikten sonra bu artış eğilimin ne kadar süreceği bilinmemektedir. Yatırımcılar fiyatının artacağını düşündüğü hisse senedini almak veya elinde bulunduruyorsa bu durumu korumak ister. Diğer taraftan yatırımcılar fiyatının düşeceğini düşündüğü hisse senedine yatırım yapmamak veya elinde bulunduruyorsa fiyatı düşmeden satmak ister. Hedeflenen sistemin gerçekleştirilmesi ile birlikte yatırımcılar belirli bir hisse senedi fiyatının artış veya azalış olarak ileride nasıl bir yön izleyeceği konusunda önceden fikir sahibi olabileceklerdir. Ayrıca yatırımcıların risklerini minimize ederken muhtemel getirilerinin artması yönünde karar vermelerine yardımcı olacak bilgi sağlanabilecektir.

1.1 Veri Madenciliği ve Metin Madenciliği

Son birkaç yıla kadar bilgi sistemleri uzmanları bilgi çıkarım işlemini yapısal formda yer alan sayısal verilerin saklandığı veritabanları ve veri ambarları üzerinden gerçekleştirmekteydi. Fakat işletmelere ait verilerin çoğu yapısal olmayan formda metin dosyaları olarak saklanmaktadır. (Kroeze vd., 2003)

Wen'e (2001) göre metin madenciliği veri içerisinde yer alan ilişkilerin keşfedilmesi açısından veri madenciliğine benzemektedir. Fakat veri madenciliğinden farklı olarak metin formatında yer alan veriler üzerinde çalışır. Hearst (2003) metin madenciliğinin veri madenciliğinden temel farkının örüntülerin çıkarılması işleminin gerçeklere ait sonuçların yapısal olarak saklandığı veritabanları yerine yapısal olmayan metin verileri üzerinden yapılması olarak belirtmiştir. Dorre vd., (1999) metin madenciliği işlemini veri madenciliğinde kullanılan analiz fonksiyonlarının özel metin analiz teknikleri uyarlanarak metin verisine uygulanması olarak tarif etmiştir.

Sonuç olarak metin madenciliği ve veri madenciliği büyük miktardaki veriler üzerinde çalışması ve bilgi çıkarımı alanınında yer alması bakımından birbirine benzemektedir. Metin madenciliği yapısal olmayan metin verisinden örüntü çıkarmaya odaklanırken, veri madenciliği yapısal verilerden örüntü çıkarır. Veri madenciliği çok daha olgun bir yaklaşım iken metin madenciliği henüz yeni gelişmekte olan bir alandır. (Wen, 2001) Buna karşın

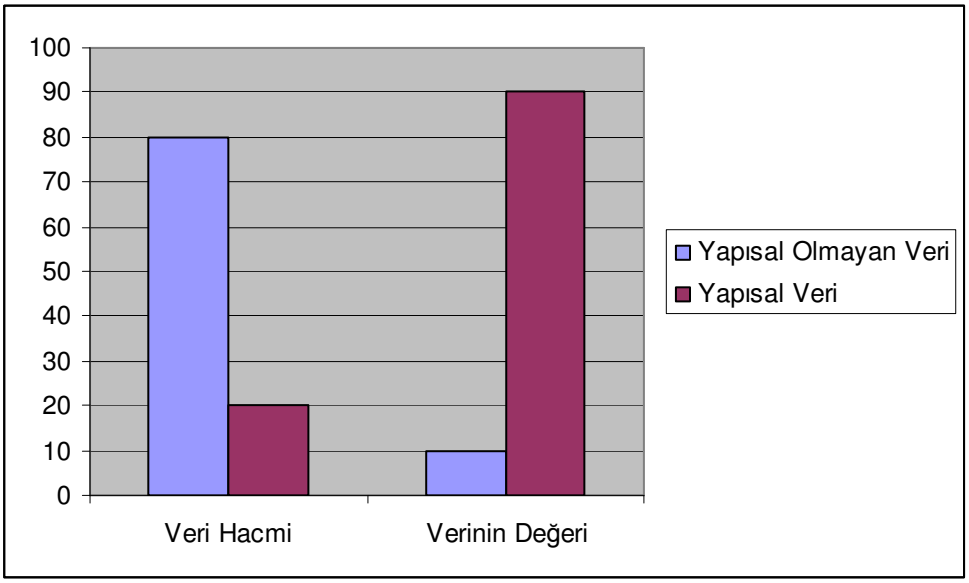
metin madenciliği belirsizliği yüksek yapısal olmayan metin verileri ile çalıştığı için daha karmaşık bir işlemdir.

1.2 Metin Madenciliğinin Önemi

Metin madenciliği Web üzerinde ve dosya tabanlı çeşitli sistemlerde yer alan çok büyük miktardaki metin verilerinin analiz edilmesi ihtiyacından ötürü hızla gelişmekte olan bir alandır. (Karanikas ve Theodoulidis, 2002)

Grobelnik'e (2000) göre World Wide Web teknolojisinin ortaya çıkışı ile birlikte çevrimiçi haberler, kurumsal arşivler, araştırma yayınları, finansal raporlar, tıbbi kayıtlar, e-posta mesajları gibi birçok yapısal olmayan veri kaynağının madencilik işlemine tabi tutulması ihtiyacı doğmuştur.

Sayısal ortamlarda saklanan ve erişime açık metin formundaki veri miktarı hızla artarken insanların birim zamanda okuyup anlayabileceği veri miktarı değişmemektedir. Tan'a (1999) göre organizayonlara ait verilerin %80'i yapısal olmayan rapor, e-posta gibi formlarda saklanmaktadır. Şekil 1.1'de gösterildiği üzere yapısal olmayan verilerin hacim olarak daha fazla olmasına rağmen yeteri kadar işlenememesi dolayısı ile organizasyonlara kattığı değer çok daha düşük olmaktadır. (Raghavan, 2002)



Şekil 1.1 Yapısal olmayan ve yapısal verinin organizasyonlardaki dağılımı (Raghavan, 2002)

1.3 Önceki Çalışmalar

Literatür taraması kapsamında genel olarak metin madenciliği konusunda gerçekleştirilen çalışmalar ile birlikte, finansal konular üzerinde özellikle hisse senedi fiyatlarının tahmin edilmesi konusu üzerinde gerçekleştirilen çalışmalar incelenmiştir. Daha önce benzer konularda çeşitli çalışmaların yapıldığı gözlemlenmiş fakat bu çalışmalarda hisse senetleri fiyatlarının tahmin edilmesinde arzu edilen başarının elde edilemediği gözlemlenmiştir. Bunun sebebinin hisse senedi fiyatlarının artışının birçok parametreden etkilenmesi veya veri setlerinin yetersiz olduğu düşünülebilir. Fakat tahmin işlemi finansal haberlerin karakteristiğine uygun olarak gerçekleştirildiğinde önemli bir başarı oranının yakalanabileceği düşünülmektedir.

Hisse senetleri fiyatlarında olabilecek değişikliklerin yayınlanan haberler ışığında tahmin edilmesi konusunda şimdiye kadar birçok çalışma gerçekleştirilmiştir. Bu çalışmalarda haberlerin olumlu ve olumsuz olarak sınıflandırılması amacı ile birçok farklı sınıflandırma yönteminden yararlanılmıştır. Fakat finansal haberlerin karakteristiği ve haberlerin olumlu veya olumsuz içerikli gibi anlamsal olarak daha derin bir seviyeden ayrılması gerekliliği gibi faktörlerden dolayı mevcut yöntemlerin direkt uygulanmasının istenen başarıyı sağlamadığı gözlenmiştir. Ön işleme ve özellik seçimi adımlarının bahsedilen bu karakteristiklere uygun olarak gerçekleştirilmesinin başarıyı çok arttıracağı düşünülmektedir. Bu çalışmada, daha önce yapılan çalışmalara nazaran yenilik olarak bu adımların problemin karakteristiğine göre gerçekleştirilmesi ve sonuçların değerlendirilmesi amaçlanmıştır.

Literatür taraması kapsamında metin formatındaki verilerin sınıflandırılması işlemi için en uygun yöntem araştırılmıştır, en çok kullanılan yöntemler ve başarıları değerlendirilmiştir. Değerlendirilen yöntemler aşağıdaki gibidir.

- Destek vektör makinesi (Support vector machine)
- k en yakın komşuluk (k nearest neighbor)
- Naive bayes
- Maximum entropy
- Karar ağaçları (Decision trees)

Bu yöntemlerin bazılarının belirli alt alanlarda çok daha iyi sonuç verdiği gözlemlenmiştir. Örneğin Naive Bayes reklam amaçlı maillerin belirlenmesinde veya bir makale veya şiirin yazarının belirlenmesinde çok başarılı sonuçlar vermektedir. Benzer konulardaki çalışmalar için başarı oranları karşılaştırıldığında destek vektör makineleri yönteminin diğer yöntemlere

göre daha başarılı olarak öne çıktığı tespit edilmiştir. Günümüzde bu sınıflandırma yöntemlerinin gerçeklemelerini içeren birçok yazılım kütüphanesi bulunmaktadır. Bu kütüphaneler bu konularda gerçekleştirilen akademik çalışmalarda ücretsiz olarak kullanılabilmektedir. Bu tür araçların kullanılması aynı çalışmada birden fazla yöntemin kullanılabilmesi ve farklı yöntemlerle elde edilen sonuçlarının değerlendirilebilmesine olanak sağlamaktadır.

Hisse senedi fiyatlarının haber makaleleri kullanılarak tahmin edilmesi üzerine gerçekleştiren çalışmalarda genellikle metin sınıflandırma yaklaşımı takip edilmektedir. Bu yaklaşım probleme uygulanırken haber makalelerinin hisse senedi fiyatı üzerinde gösterdiği etkiye göre olumlu veya olumsuz olarak sınıflandırılması amaçlanmaktadır. Bu kapsamda olumlu veya olumsuz olarak etiketlenmiş eğitim makalelerinden yola çıkılarak test makalelerinin sınıfı tahmin edilmektedir. Bu tür bir çalışmayı gerçekleştirebilmek için çözülmesi gereken problemlerden birisi eğitim setindeki haber makalelerinin önceden olumlu veya olumsuz olarak etiketlenmesidir. Mevcut çalışmalarda makalelerin olumlu veya olumsuz olarak etiketlenmesinde genel olarak iki yaklaşım kullanılmıştır. Bu yaklaşımlar elle ve otomatik etiketleme yaklaşımlarıdır. Elle etiketleme yaklaşımında eğitim makaleleri kişiler tarafından okunup değerlendirilerek olumlu veya olumsuz olarak etiketlenmektedir. Bu yaklaşım makalelerin daha doğru etiketlenmesini sağladığı için genel olarak hedef sistemin sınıflandırma başarısını artırmaktadır. Elle etiketleme yaklaşımının olumsuz tarafı ise etiketlemenin zaman alıcı olmasından ötürü veri setinde kullanılabilecek makale sayısının sınırlı oluşudur. Bir diğer dezavantaj ise eğitim setinin büyütülmesi amacı ile sisteme eklenecek yeni makaleler için yeniden elle işlem yapma gerekliliğidir. Otomatik etiketleme yaklaşımında makalelerin etiketlenmesi makalenin yayınlandığı günden itibaren hisse senedi fiyatında oluşan değişime göre gerçekleştirilmektedir. Otomatik etiketleme yaklaşımı daha fazla makale içeren veritabanları ile çalışma imkanı tanıdığı için daha elverişli bir yaklaşım olmaktadır. Ayrıca sisteme sürekli yeni makaleler eklenerek veritabanı kolaylıkla genişletilebilmektedir. Fakat bu yöntemde makalelerin tamamen doğru etiketlenmesi mümkün olamamaktadır. Çünkü gerçekte olumlu olan bir makale yayınlandıktan sonra hisse senedi fiyatında değişik nedenlerden dolayı düşüş yaşanabilmektedir. Ya da tam tersi olarak olumsuz bir makale yayınlandıktan sonra hisse senedi fiyatında başka bir nedenden ötürü artış yaşanabilmektedir. Örneğin bir şirketin başarılı finansal sonuçlarının bir haber makalesinde yayınlanmasından sonra yatırımcıların yine de global krizden ötürü bu hisseye ilgi göstermemesi hisse senedi fiyatında düşüşe sebebiyet verebilmektedir. Bu durumda makale olumlu içeriğe sahip olmasına rağmen olumsuz olarak etiketlenecektir. Bir diğer örnek ise bir

şirketin zarar açıkladığını bildiren bir haber makalesinin yayınlanmasından sonra zararın beklenenden daha düşük olmasından dolayı yatırımcıların hisseye ilgi göstermesi ve hisse senedi fiyatının artmasıdır. Bu durumda ise aslında olumsuz içeriğe sahip bir makale olumlu olarak etiketlenecektir. Otomatik etiketleme yaklaşımının getirdiği bu yanlış etiketlenebilme olasılıkları sistemin genel başarı oranında bir miktar düşüklüğe sebep olmaktadır.

Hisse senedi fiyatlarının tahmin edilmesi üzerinde gerçekleştirilen çalışmaların sonuçları incelendiğinde başarı oranlarının genellikle %40 ve %60 arasında değiştiği gözlemlenmektedir. Bu çalışmalarda başarı oranlarının göreceli olarak düşük olması hisse senetleri fiyatlarının doğasından kaynaklanmaktadır. Çünkü hisse senetlerinin fiyatları yatırımcıların alım satım taleplerine göre belirlenmektedir. İnsan davranışlarının tahmin edilmesinin güçlüğü bu çalışmaların sonuçlarına da yansımaktadır. Diğer taraftan başarı oranlarının rasgele tahminlere göre yüksek olması hisse senetleri fiyatları ile haber makaleleri arasında önemli bir ilişki olduğunu kanıtlamaktadır. Hisse senetleri alım satımı yapmak isteyen yatırımcıların yatırımlarına yön verirken bu haberleri dikkate alması bu ilişkinin kurulmasını sağlamaktadır.

Finansal haber makalelerinin olumlu-olumsuz veya artış-düşüş etkili şeklinde sınıflandırılması ile ilgili çalışmalarda genellikle makalelerin içeriğinde yer alan kelimeler özellik olarak kullanılmaktadır. Her bir kelime ayrı ayı özellik olarak kabul edilmektedir. Kelimeler özellik olarak belirlenirken yalnızca sınıflandırmada anlamlı olacak kelimeler seçilmektedir. Sınıflandırmada etkisi olmayan edat, bağlaç, zarf vb. türden kelimeler ayıklanmaktadır.

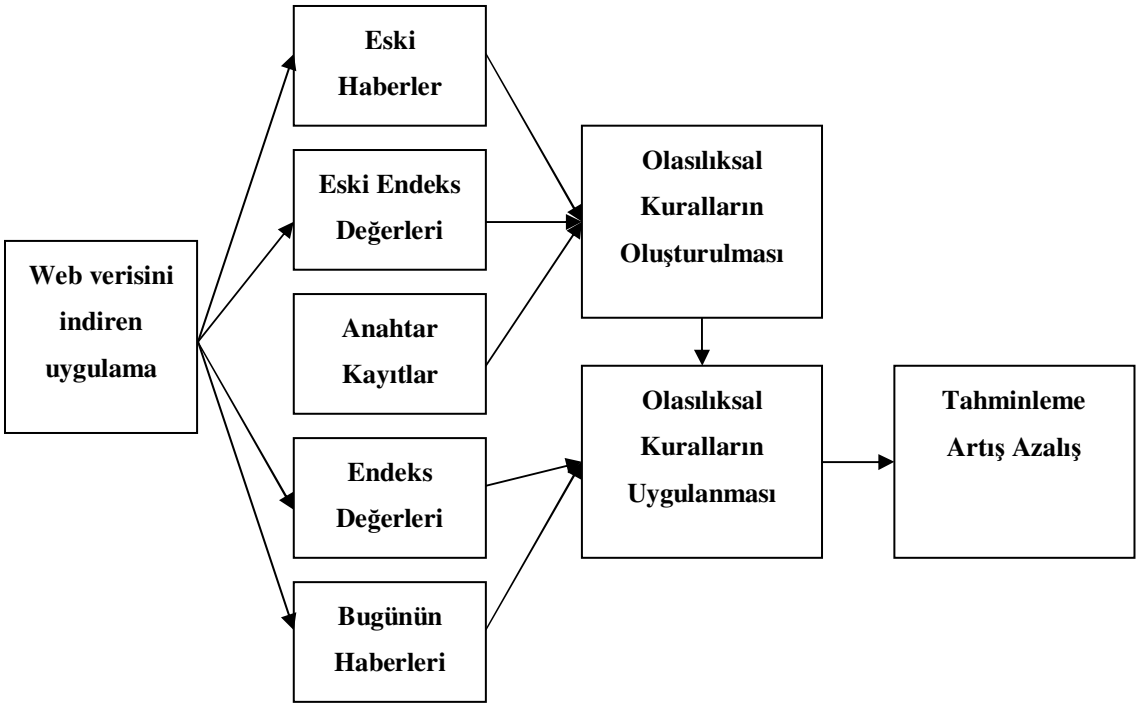
Benzer konularda şimdiye kadar gerçekleştirilen çalışmalardan en önemlileri Çizelge 1.1'de yayın yılı ve yazar bilgileri ile listelenmiştir. Bu bölümün ilerleyen kısımlarında bu çalışmalardan kısaca bahsedilmiştir.

Çizelge 1.1 Hisse senedi fiyatı tahmin etmeye yönelik gerçekleştirilen bazı önemli çalışmalar

Çalışmanın Başlığı	**Yazar ve Yayın Tarihi**
Daily Stock Market Forecast from Textual Web Data	Wuthrich vd., 1998
Electronic Analyst of Stock Behavior	Lavrenko vd., 1999
Language Models for Financial News Recommendation	Lavrenko vd., 2000
Mining of Concurrent Text and Time Series	Lavrenko vd., 2000
Using News Articles to Predict Stock Price Movements	Gidofalvi, 2001
Stock Broker P – Sentiment Extraction for the Stock Market	Khare vd., 2004
Good News or Bad News? Let the Market Decide	Koppel vd., 2004
Classification of Stock Exchange News	Kroha vd., 2004
Forecasting Intraday Stock Price Trends with Text Mining Techniques	Mittermayer, 2004
The Predicting Power of Textual Information on Financial Markets	Fung vd., 2005
Stock Trend Prediction Using News Articles: A Text Mining Apporach	Falinouss, 2007
Predicting the Trend of Taiwan Weighted Stock Index with Text Mining Techniques	Wu, 2007

Wuthrich vd. (1998) web üzerinden yayınlanan makalelerde barındırılan bilgileri kullanarak hisse senetleri endeks değerlerinin tahmin edilmesi üzerine bir çalışma gerçekleştirmiştir. En etkili ve saygın gazetelerden alınan finansal haber makaleleri sistemin girdisi olarak alınmaktadır. Bu makalelerden Asya, Avrupa ve Amerika'daki önemli borsaların hisse senetleri endeksleri tahmin edilmektedir. Tahmin etme işlemi makaleler içerisinde geçen anahtar kayıtlar üzerinden gerçekleştirilmektedir. Bir anahtar kayıt bir veya daha fazla kelimeden oluşmakta ve hisse senedi endeks değerleri üzerinde etkisi olduğu

düşünülmektedir. Makaleler anahtar kayıtları içerip içermedikleri bilgilerinden faydalanılarak hisse senedi endeksi üzerinde artış, azalış veya durağan etkili olarak sınıflandırılmaktadır. Finansal bir uzman tarafından belirlenen 400 kadar anahtar kayıt sistemin tahminleme işleminde kullanılmaktadır. Kural tabanlı, k en yakın komşuluk, yapay sinir ağları gibi farklı teknikler kullanılmış ve bu farklı tekniklerden elde edilen sonuçlar birbirleri ile karşılaştırılmıştır. Sistemin tahmin ettiği değerlere göre bir yatırım stratejisi sistemin kullanıcılarına sunulmaktadır. Şekil 1.2'de bu sistemin genel mimarisine yer verilmiştir.

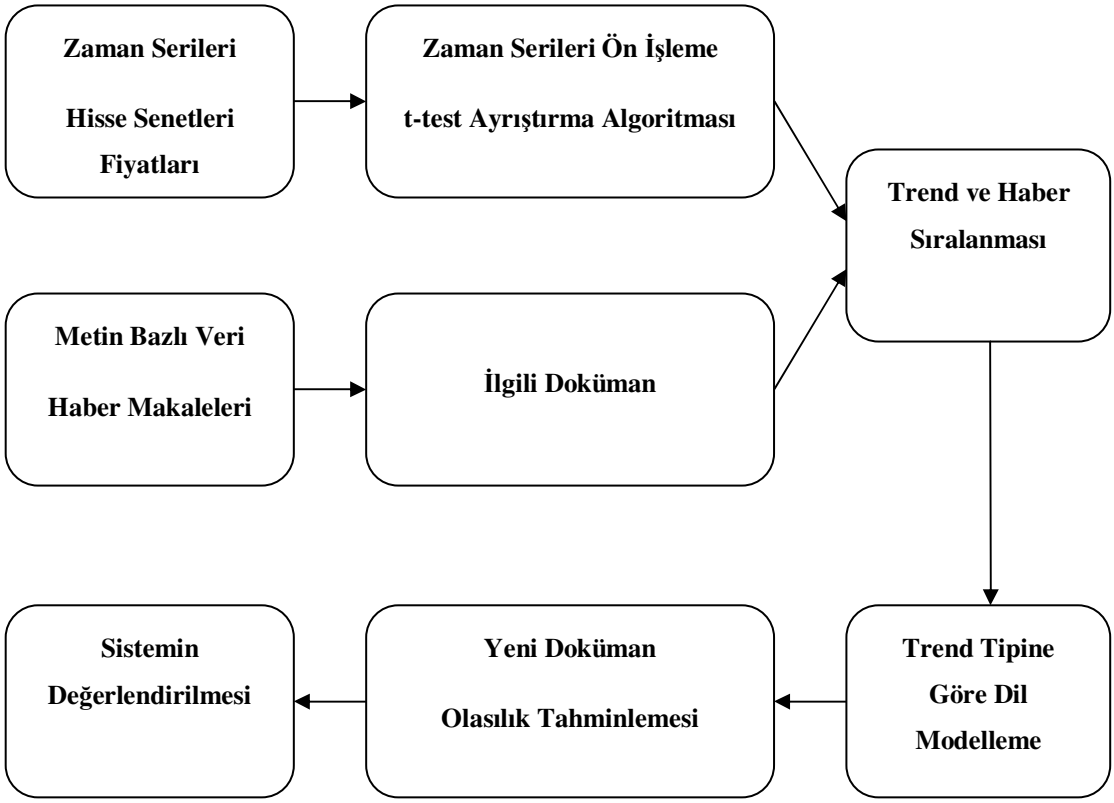


Şekil 1.2 Wuthrich vd. sisteminin genel mimarisi (Wuthrich vd., 1998)

Her ne kadar sistemin başarısı rasgele yapılan tahminlere göre yüksek olsa da sisteme yönelik bazı eleştiriler yapılabilmektedir. Örneğin sistemin finansal makaleleri sınıflandırmak için kullandığı anahtar kayıtlar önceden belirlenmektedir. Bu durumda önceden tanımlanmayan fakat makaleler içerisinde bulunabilecek diğer anahtar kayıtlar ihmal edilmektedir. Diğer bir eleştiri de sistemin belirli hisse senetlerine ait fiyatlar yerine hisse senetleri endeksleri üzerinde tahminde bulunmasıdır. Fakat yatırımcılar daha çok hisse senedi fiyatları üzerinde oluşabilecek değişiklikler ile yakından ilgilenmektedir .

Lavrenko vd. (1999, 2000) haber makalelerinden hisse senetleri fiyatlarının tahmin edilmesi üzerine kapsamlı bir çalışma gerçekleştirmişlerdir. Ænalyst adını verdikleri sistem gerçek zamanlı haber makalelerinin içeriklerini analiz ederek hisse senetleri fiyatlarında oluşabilecek

değişiklikleri tahmin etmektedir. Ænalyst dil modelleme yaklaşımı üzerine geliştirilmiştir ve zaman serileri ve haber makaleleri arasındaki bağımlılığı modellemeye çalışmaktadır. Bu sistem zaman serileri ve haber makaleleri şeklindeki iki farklı tipteki datayı toplamakta, işlemekte ve arasında ilişkiler kurmaktadır. Ænalyst sisteminin tasarımı Şekil 1.3'te görülmektedir.

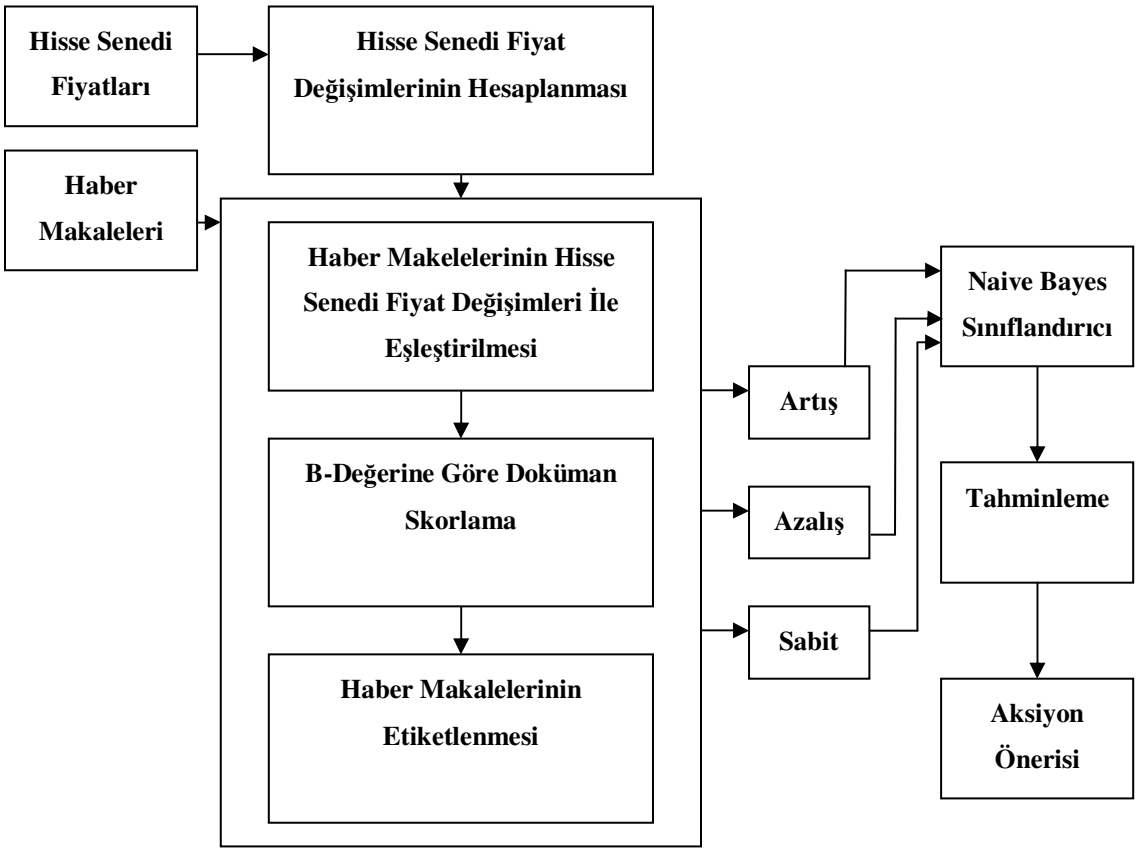


Şekil 1.3 Ænalyst sisteminin genel mimarisi (Lavrenko vd., 1999, 2000)

Zaman serileri trendlerinin tanımlanmasında t-test ayrıştırma algoritması kullanılmıştır. Zaman serilerine ait trendler ayrıştırılmış ve uzunluk, eğim, diğer trendler ile kesişme gibi karakteristik özellikleri kullanılarak etiketlere atanmıştır. Bu işlem birikimli kümeleme algoritması (Everitt, 1993) ile gerçekleştirilmiştir. Bu etiketler haber makaleleri ile trendlerin ilişkilendirilmesinde temel bilgiyi oluşturmaktadır. Trendleri belirli bir tarihe ait haber makaleleri ile ilişkilendirilirken haber makalesinin trendin başlangıcından belirli bir eşik değerinden daha az süre önce yayınlanması gerekmektedir. Denemeler sonucunda eşik değeri için 5 ila 10 saatin en uygun sonuçları verdiği ortaya konmuştur. Belirli bir trend ile ilişkili haber makalelerinin dil modelleri (Ponte ve Croft, 1998; Wallis vd., 1999) öğrenilmiştir. Bu şekilde eğitim seti makalelerinde geçen kelimelerin kullanım istatistikleri hesaplanmıştır.

Sistemin performansını ölçümlemek amacı ile market simülasyonu kullanılmıştır.

Gidofalvi (2001) finansal haber makalelerinden hisse senedi fiyatları değişiminin tahmin edilmesi üzerine bir çalışma gerçekleştirmiştir. Bu çalışmada kısa dönem hisse senedi fiyat değişimlerinin finansal haber makaleleri üzerinden tahmin edilmesi amaçlanmıştır. Sistem hisse senedi için kesin bir fiyat tahmini yapmak yerine kullanıcılara al veya sat önerileri getirmektedir. Şekil 1.4'te Gidofalvi sisteminin tasarımı görülmektedir.

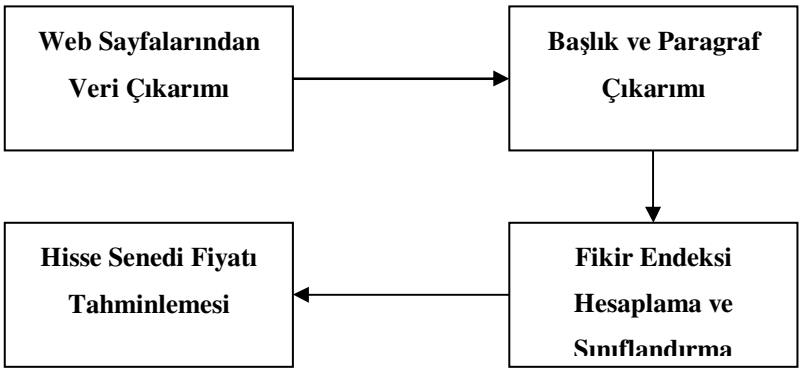


Şekil 1.4 Gidofalvi sisteminin genel mimarisi (Gidofalvi, 2001)

Belirli bir zaman dilimi içerisinde hisse senedi fiyat değişimleri artış, azalış ve sabit olarak sınıflandırılmıştır. Eğitim setindeki her bir haber makalesi makalenin yayınlanma zamanındaki hisse senedi fiyat değişimine göre artış, azalış veya sabit olarak etiketlenmiştir. Naive Bayes yöntemi kullanılarak eğitim setindeki makaleler ile sistemin eğitimi gerçekleştirilmiştir. Ardından bir test makalesinin eğitilmiş sınıflandırıcı tarafından hangi fiyat değişim sınıfına dahil olduğu tahmin edilmiştir. Her ne kadar etkin piyasa hipotezi böyle

bir sınıflandırma işleminin tahmin etme başarısı olmayacağını öngörse de bu çalışmada bir haber makalesinin yayınlandığı andan itibaren 20 dakika içerisindeki hisse senetlerinin fiyatlarında olabilecek değişimlerin tahmin edilebildiği ortaya çıkarılmıştır.

Bu konuda bir diğer çalışma ise Khare vd. (2004) tarafından gerçekleştirilmiştir. Stock Broker P diğer bir deyişle Stock Broker Prediction adını verdikleri sistem Web üzerinde yayınlanan haber makalelerini madencilik işlemine tabi tutarak hisse senedi fiyat değişimlerini tahmin etmektedir. Haber makalelerinin sınıflandırılmasında uyarlanmış bir Naive sınıflayıcı kullanılmıştır. Çalışma sonucunda %60'ın üzerinde başarı oranı elde edildiği belirtilmiştir. Sistemin temel bileşenlerini Web sayfalarını işleyerek yatırımcı görüşleri ve haberleri toplayan bir modül ile görüşleri madencilik işlemine tabi tutan bir modül oluşturmaktadır. Her bir haber makalesi için bir fikir endeksi oluşturulmaktadır. Bu fikir endeksleri haber makalelerindeki her bir cümleden çıkarılan fikir değerlerinin toplanması ile elde edilmektedir. Yeni yayınlanan haber makalelerine ait bu fikir endeksi kullanılarak hisse senedi fiyatında oluşabilecek değişiklikler tahmin edilmektedir. Çeşitli finansal haber siteleri incelenerek önemli kelimeler bir sözlükte toplanmıştır. Ayrıca sistem zaman içerisinde yeni önemli kelimeleri de öğrenebilmektedir. Stock Broker P sisteminin mimarisi Şekil 1.5'te görülmektedir.



Şekil 1.5 Stock Broker P sisteminin genel mimarisi (Gidofalvi, 2001)

Koppel vd. (2004) finansal haber makalelerinin iyi haber veya kötü haber olarak sınıflandırılması üzerine bir çalışma gerçekleştirmiştir. Halka açık şirketler hakkında yayınlanan haberler ilgili şirketin hisse senedi fiyatının değişimine göre pozitif veya negatif olarak etiketlenmiştir. Sözcükler üzerinden analiz yapılmasını öngören modeller kullanılarak haber makalelerinin iyi haber veya kötü haber olarak ayrıştırılabilmesi %70 nispetinde başarı

ile gerçekleştirilmiştir.

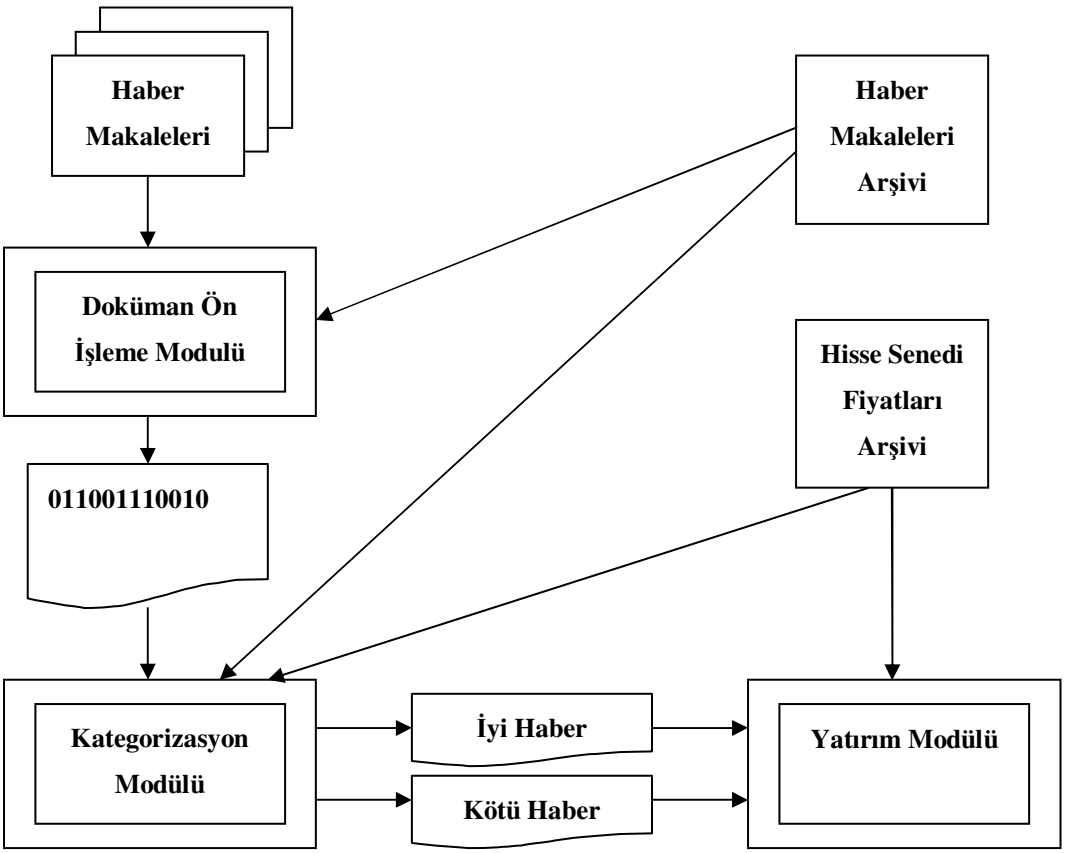
Eğitim setinde 60 defadan fazla geçen sözcükler özellik olarak belirlenmiştir. Ayrıca sınıflandırmada etkisi olmayan kelimeler elenmiştir. Özellik değerleri olarak özelliğin haber içerisinde geçme sayısı yerine özelliğin haberde geçip geçmediği bilgisi kullanılmıştır. Çünkü daha önce yapılan çalışmalara göre finansal makalelerin sınıflandırmasında sözcüklerin geçip geçmediği bilgisi geçme sayısına göre daha önemli bir bilgidir. Her bir makale özellik olan kelimeleri barındırıp barındırmadığı bilgisini tutan bir vektör olarak gösterilmiştir. Özellik sayısının indirgenmesi amacı ile bilgi kazancı yöntemi kullanılmış ve en yüksek bilgi kazancı sağlayan 100 kelime seçilmiştir.

Koppel vd. (2004) çalışmasında doğrusal SVM yöntemini kullanarak 10-fold çapraz doğrulama sonucunda %70.3 başarı oranı elde etmiştir. 2000-2002 yılları arasında yayınlanan haber makalelerini eğitim seti olarak, 2003 yılında yayınlanan haber makalelerini ise test seti olarak kullandığında %65.9 başarı oranı elde etmiştir. Naive Bayes ve karar ağaçları gibi diğer sınıflandırma yöntemleri kullanıldığında da benzer sonuçlar elde edilmiştir. SVM yönteminin diğer kerneller ile uygulanması da sonuç üzerinde fazla bir değişikliğe sebep olmamıştır.

Kroha vd. (2004) çalışmasında haber makalelerinin iyi veya kötü olmasının uzun vadeli piyasa hareketlerine etkisini araştırmıştır. Bu çalışmada metin madenciliği, sınıflandırma ve bilgi erişimi yöntemleri arasındaki ilişki incelenmiştir. Çalışmada benzer kelimelerden oluşan fakat farklı anlamlar taşıyan kelime gruplarından örnekler verilmiştir. Ayrıca okumadan olumlu ve olumsuz olduğu hakkında fikir edinmenin çok zor olduğu cümle örneklerine yer verilmiştir. Çalışma Alman Borsası ile ilgili geçmişte yayınlanan haberler üzerinden test edilmiştir. Çalışma sonucunda bilgi erişimi yönteminin haber makalelerinin iyi veya kötü olarak ayrıştırılmasında yeteri kadar etkili olamadığı ortaya çıkarılmıştır. Bu tür problemlerde sınıflandırma yöntemlerinin daha etkili olduğu tespit edilmiştir.

2004 yılında Mittermayer (2004) NewsCATS (News Categorization and Trading System) isminde bir tahminleme sistemi gerçekleştirmiştir. Bu sistem finansal haber makalelerinin yayınlanmasından hemen sonra hisse senetleri fiyatları üzerinde tahminler üretmektedir. NewsCATS sistemi üç temel bileşenden oluşmaktadır. İlk bileşen metin işleme teknikleri ile haber makalelerinden uygulamanın ihtiyaç duyduğu bilgileri getirmektedir. İkinci bileşen haber makalelerini önceden tanımlanmış kategorilere ayırmaktadır. Son bileşen ise kategorize edilmiş haber makaleleri üzerinden yatırım stratejileri oluşturmaktadır. Şekil 1.6'da

Mittermayer sisteminin mimarisine yer verilmiştir.



Şekil 1.6 Mittermayer sisteminin mimarisi (Mittermayer, 2004)

NewsCATS çeşitli kategorize etme kurallarını öğrenerek kategorizasyon modülünün haber makalelerini önceden tanımlanmış kategorilere ayırmasına olanak sağlamaktadır. Her bir kategori hisse senetlerinin fiyatları üzerinde oluşan etkiler ile ilişkilendirilmiştir. Bu etkiler artış veya azalış olabilmektedir. Kategorizasyon modülünden elde edilen sonuçlara göre yatırım modülü yatırım aksiyonları tavsiyeleri oluşturmaktadır.

Haber makalelerinin otomatik olarak ön işlemesi aşamasında özellik çıkarımı, özellik seçimi ve doküman gösterimi işlemleri gerçekleştirilmektedir. Özellik seçiminde tf (Term Frequency), idf (Inverse Document Frequency) ve tf*idf (Term Frequency-Inverse Document Frequency) frekans ölçümlemeleri kullanılmıştır. Ön işleme modülünün çıktıları kategorizasyon modülüne geçirilerek haber makalelerinin iyi veya kötü olarak kategorize edilmesi sağlanır. Kategorizasyon işlemi SVM metin sınıflayıcı ile gerçekleştirilmiştir.

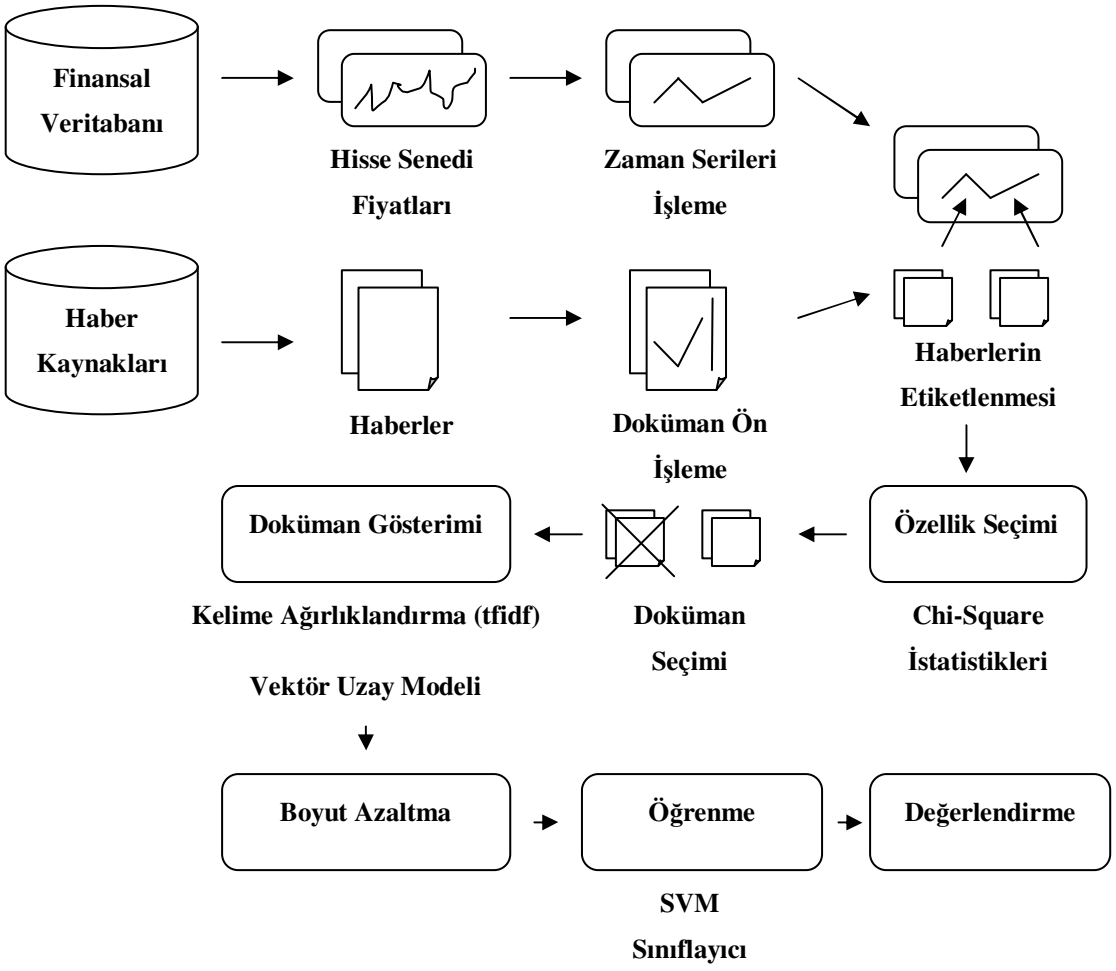
NewsCATS sistemi 2002 yılına ait haber makaleleri ve hisse senetleri fiyatları üzerinden test edilmiştir. Testler sonucunda NewsCATS sisteminin oluşturduğu yatırım stratejilerinin

rasgele karar verilen yatırımlardan daha iyi sonuç verdiği tespit edilmiştir.

Fung vd. (2005) finansal haber makalelerinin analiz edilerek hisse senetleri fiyatlarındaki değişimlerinin tahmin edilmesini sağlayan bir çatı oluşturulması üzerine bir çalışma gerçekleştirmiştir. Bu çalışmada Hong Kong borsasında işlem gören hisse senetleri fiyatları ve bu hisse senetleri ile ilgili yayınlanmış 350.000'den fazla doküman kullanılmıştır. Doküman seçimi amacı ile ki-kare istatistikleri tekniği kullanılmıştır. Seçilen dokümanların gösterimi tf*idf ölçümlemesi üzerinden yapılmıştır. Sınıflandırma işleminde ise SVM yönteminden faydalanılmıştır. Çalışma sonucunda hisse senetleri fiyatları ve finansal haber makaleleri arasında güçlü bir ilişki olduğu tespit edilmiştir.

Falinouss (2007) çalışmasında metin formatında yer alan haber verilerinin sayısal verilere nazaran daha zengin bir bilgi kaynağı olduğunu öngörerek haber makalelerinden hisse senedi fiyatlarının tahmin edilmesi üzerine bir model geliştirmiştir. Bu çalışmada temelde çeşitli veri ve metin madenciliği teknikleri kullanılarak ikili bir sınıflandırma yapılmaya çalışılmıştır.

Bu modeli hayata geçirmek için İran Khodro şirketine ait iki yıllık hisse senedi fiyatları ve bu şirket ile ilgili yayınlanan haber makaleleri kullanılmıştır. Yeni bir istatistiksel segmentasyon algoritması geliştirilerek zaman serileri üzerindeki trendler tanımlanmıştır. Haber makaleleri ön işlemeye tabi tutulup içinde bulunduğu segmentlenmiş trende göre etiketlenmiştir. Ki-kare yöntemine göre seçilen dokümanların tf*idf ölçümlemesi üzerinden vektörel gösterimi sağlanmıştır. Dokümanların sınıflandırmasında SVM yöntemi kullanılmıştır. Falinouss sisteminin mimarisi Şekil 1.7'de gösterilmiştir.



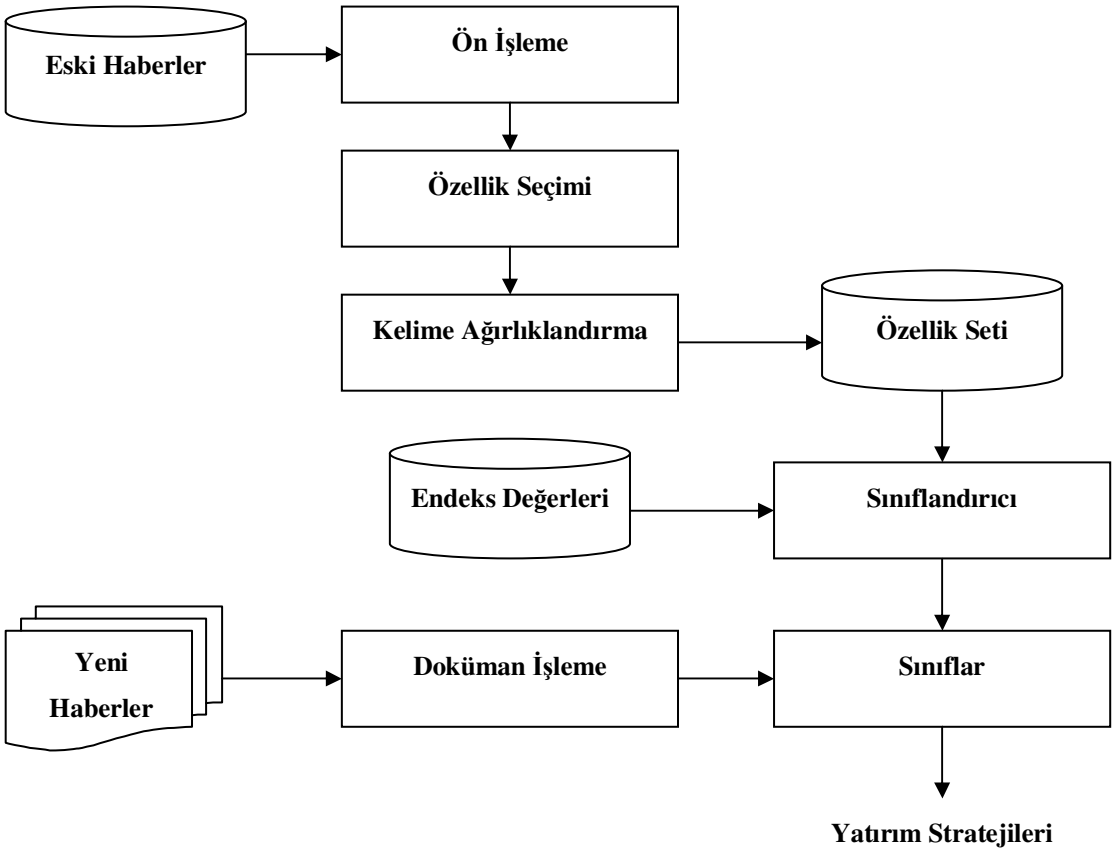
Şekil 1.7 Falinouss sisteminin mimarisi (Falinouss, 2007)

Sistemin performansını değerlendirmek amacı ile çeşitli testler gerçekleştirilmiştir. Testler sonucunda tahmin modelinin rastgele gerçekleştirilen tahminlere göre çok daha başarılı sonuçlar ürettiği tespit edilmiştir.

Wu (2007) metin madenciliği tekniklerinden faydalanarak Taiwan borsasındaki hisse senedi fiyatlarının hareketlerinin tahmin edilmesi üzerine bir çalışma gerçekleştirmiştir. Çevrimiçi finansal haber makalelerini sınıflandıran bir sistem geliştirilmiştir. Sınıflandırma sonuçlarına göre yatırım stratejileri oluşturulmuş ve sistemin performansı Taiwan Borsası Ağırlıklı Hisse Senedi Endeksi üzerinden değerlendirilmiştir.

Şekil 1.8'de Wu sisteminin mimarisi gösterilmiştir. Buna göre sistem üç temel bileşenden oluşmaktadır. İlk bileşen yeni yayınlanan haberlerin otomatik olarak ön işlemesini gerçekleştirmektedir. İkinci bileşen yeni yayınlanan bu haberleri sınıflandırmaktadır. Üçüncü

bileşen ise yatırım stratejilerini oluşturmaktadır.



Şekil 1.8 Wu sisteminin mimarisi (Wu, 2007)

Performans sonuçları değerlendirildiğinde sistemin öngördüğü yatırım stratejileri ile ayda ortalama %5.4 getiri sağlanabileceği tespit edilmiştir. Bu bakımdan sistem istatistiksel olarak %5 getiri sağlayan mevduat sertifikasından daha yüksek bir getiri sağlamaktadır. Bu bilgilerden yola çıkılarak sistemin öngördüğü yatırım stratejilerinin kısa dönem yatırımlar için kayda değer bir öneme sahip olduğu ispatlanmıştır.

Bu bölümde haber makalelerinden hisse senedi fiyatlarının tahmin edilmesi üzerine daha önce gerçekleştirilen çalışmalardan bahsedilmiştir. Bu çalışmalarda çeşitli kabullenmeler üzerinden farklı modeller kullanılmakla birlikte çoğu çalışmanın mimarisinde ortak kısımlar olduğu gözlemlenmektedir. Genel olarak daha önce yayınlanmış haber makaleleri eğitim amaçlı kullanılarak yeni haber makalelerinden hisse senedi fiyatlarına yönelik tahminler geliştirilmesi prensipleri izlenmiştir. Bu aşamaların gerçekleştirilmesinde farklı makine öğrenmesi tekniklerinden faydalanılmıştır.

Çalışmalarda araştırmacıların ortak bir veritabanı kullanmak yerine kendi oluşturdukları veritabanlarını kullandıkları gözlemlenmiştir. Bu nedenle çalışmaların sonucunda elde edilen başarı oranlarının çok farklı bir aralıkta dağıldığı tespit edilmiştir. Fakat çalışmaların çoğunda rasgele yatırım stratejilerden daha başarılı sonuçlar elde edildiği görülmüştür.

2. HİSSE SENETLERİ VE FİNANSAL HABER MAKALELERİ

Bu bölümde çalışmanın anlaşılmasına yardımcı olmak amacı ile hisse senetleri ve finansal haber makaleleri hakkında bilgiler verilmiştir.

2.1 Hisse Senetleri

Hisse senetleri çok ortaklı şirketlerin şirket ortaklığına belirli bir katılımı temsil eden yasal olarak düzenlenmiş kıymetli evraklardır. Hisse senetleri devlet veya özel kuruluşlar tarafından kurulan borsalarda işlem görürler. Hisse senetlerini alıp satmak isteyen yatırımcılar borsalar aracılığı ile bu işlemlerini gerçekleştirebilmektedirler.

Hisse senetlerinin alım satım işlemleri hisse senetlerinin fiyatları üzerinden gerçekleştirilmektedir. Borsada işlem gören her hisse senedinin belirli bir anda bir fiyatı bulunmaktadır. Hisse senedini almak isteyen yatırımcılar bu fiyat üzerinden alım yaparken, hisse senedini satmak isteyen yatırımcılar da bu fiyat üzerinden ellerindeki hisse senetlerini satabilmektedir.

Borsada hisse senetleri gün içerisinde birinci ve ikinci seans olmak üzere iki oturum süresince yatırımcılar tarafından işlem görebilmektedir. Hisse senetlerinin her iki seans sonunda da kapanış fiyatları belirlenmektedir. İkinci seansın kapanış fiyatı ilgili hisse senedinin ilgili gündeki fiyatı olarak kabul edilmektedir. Bir hisse senedinin belirli bir gündeki fiyatının bir önceki güne ait fiyatına göre artış göstermesi hisse senedi değerlenmesi anlamına gelmektedir. Hisse senedini bir önceki gün fiyatından alan bir yatırımcı bir gün sonra hisse senedinin değerlenmesi ile kazançlı duruma geçmektedir. Hisse senedinin fiyatının bir önceki güne göre azalış göstermesi hisse senedinin değer kaybetmesi anlamına gelmektedir. İlgili hisse senedine yatırım yapan yatırımcı da ilgili günde zararlı duruma geçmiş bulunmaktadır. Yatırımcılar hisse senetlerini belirli bir fiyattan alıp ilerleyen günlerde hisse senedinin fiyatında artış olması ile yatırımlarında getiri elde etmeyi amaçlamaktadır.

Hisse senetleri fiyatları yatırımcıların arz talep dengelerine göre şekillenmektedir. Yatırımcılar olumlu gördükleri şirket faaliyetlerine ilgili şirketin hisse senedine talepte bulunarak tepki vermektedirler. Örneğin bir şirketin satış gelirlerini bir önceki döneme göre artırması, piyasaya yeni bir ürün sunması veya başka bir şirketi satın alması yatırımcılar tarafından olumlu karşılanan gelişmelerdir. Bu gelişmelerden sonra şirket hisselerine talep artar ve hisse senedi fiyatında artış meydana gelir. Öte yandan yatırımcılar şirket adına olumsuz bulduğu gelişmelerde ise ilgili şirketin hissesine yatırım yapmamayı tercih ederler.

Yatırım yapılmaması hisse senedine talebin azalması anlamına gelmektedir ve bu da hisse senedinin fiyatında düşüşe neden olur.

Yatırımcıların talepleri sadece şirkete ait haberler ile şekillenmez. Ulusal veya global, ekonomik veya siyasal gelişmeler de yatırımcıların taleplerini belirleyen unsurlardır. Örneğin ülkede veya dünya üzerinde yaşanabilecek bir kriz kaçınılmaz olarak yatırımcıların hisse senetlerine olan ilgisini azaltacaktır.

2.2 Finansal Haber Makaleleri

Hisse senetlerini borsada işleme açan şirketlerin yasal olarak bazı yükümlülükleri bulunmaktadır. Bu şirketler hisse senetleri yatırımcılarının bilmesi gereken faaliyetleri ve faaliyetleri sonucu oluşan finansal pozisyonlarını kamuya bildirmek durumundadırlar. Borsada işlem gören şirketlerin yaptığı bu tür bildirimler yatırımcılara yön vermek amacı ile hazırlanan finansal yayınlarda yer almaktadır. Finansal makalelerde ayrıca şirketlere ait finansal sonuçların değerlendirmesine de yer verilmektedir.

Finansal haber makalelerinde genellikle bir hisse senedine yönelik yatırımcılara yön verecek haber veya yorumlara yer verilmektedir. Çoğunlukla bir makale sadece bir şirkete ait hisse senedine yönelik olarak yayınlanmaktadır. Bazı durumlarda makaleler aynı sektörde birbirine rakip konumdaki birden fazla şirketin hisse senedine yönelik içerikte olabilmektedir.

Haberlendirme amaçlı makalelerin içeriklerinde genellikle bir şirketin dönemsel finansal sonuçları, bir şirketin başka bir şirketi satın alması veya şirket birleştirmeleri gibi haberlere yer verilmektedir. Finansal sonuçların yayınlandığı makalelerde şirketin karlılık veya satış gelirlerine ait niceliksel bilgiler bazı durumlarda geçmiş dönemler ile karşılaştırılarak verilmektedir. Bu tip makaleler bu çalışmanın amacı olan makalelerin hisse senedi fiyatına olumlu veya olumsuz anlamdaki etkisinin tahmin edilmesi işleminde en uygun makalelerdir. Çünkü bu finansal sonuçların yayınlandığı makalelerde "A şirketinin satışları arttı", "B şirketinin karı azaldı" gibi sınıflandırmada çok etkili metinler yer alabilmektedir. Şirket satın alma veya şirket birleşmeleri haberleri genellikle yatırımcılar tarfından olumlu algılanıp, ilgili şirket veya şirketlerin hisse senedi fiyatlarında artışa sebep olmaktadır. Bu nedenle bu içerikte habere sahip makaleler kendine has belirli anahtar kelimeleri barındırmaları ve olumlu içeriklerinden ötürü sınıflamada etkili olmaktadır.

Yorum içerikli makalelerde genellikle finansal konularda uzman bir yazarın makalenin ilgili olduğu hisse senedine yönelik kişisel görüşleri ve okuyuculara tavsiyeleri yer almaktadır. Bu

tip makalelerin hisse senedinin fiyatına olumlu veya olumsuz etkili olarak sınıflandırılması insanlar tarafından okunup değerlendirildiğinde kolay olmakla birlikte doğal dil işleme teknikleri kullanıldığında oldukça zor olmaktadır. Çünkü doğal dil işleme teknikleri sınıflandırmada kullanmak üzere anahtar olacak bazı kelimelere ihtiyaç duymaktadır. Bu tip makalelerde ise metin içerisinde çok farklı standart olmayan kelimeler geçebilmektedir. Ayrıca kullanılan kelimelerin olumlu veya olumsuz sınıflandırmada etkisi daha düşük olmaktadır.

Sonuç olarak finansal haber makaleleri şirketler hakkında oldukça geniş bilgiler içermektedir. Hem niceliksel hem de niteliksel olarak oldukça zengin olan finansal haber makalelerinden faydalanılarak karar vermeye yardımcı anahtar bilgilerin üretilmesi mümkün olmaktadır.

2.3 Etkin Piyasa Hipotezi ve Tesadüfi Yürüyüş Modeli

Etkin piyasa hipotezinin ilk klasik tanımı Fama (1970) tarafından yapılmıştır. Bu tanımlamaya göre menkul kıymet fiyatları daima ulaşılır tam bilgileri yansıttığı durumda ancak etkin bir piyasanın varlığından söz etmek mümkün olmaktadır. Etkin piyasa hipotezinin temel dayanağı tesadüfi yürüyüş modelidir. Tesadüfi yürüyüş modeli ise etkin bir sermaye piyasasında her türlü bilgi piyasaya yansımış ve yatırımcılar tarafından değerlendirilmiş ise herhangi bir andaki hisse senedinin fiyatı, hisse değerinin gerçek değerine eşit olacaktır. Etkin piyasa hipotezi temel ve teknik yaklaşımların geçersiz olduğunu ileri süren bir yaklaşımdır. Yani piyasadaki menkul kıymet fiyatlarının şekillenmesi ancak bilgi ve bu bilginin tüm piyasa katılımcıları tarafından eş zamanlı ve tamamen paylaşımı ile mümkün olmaktadır. Günümüzde sermaye piyasaları ve reel piyasalar daha fazla birbirini etkiler hale gelmiştir. Bu nedenle sermaye piyasalarının etkinliği sermaye piyasalarına konu olan menkul kıymetlerin gerçek değerlerini yansıtması ile mümkün olabilmektedir.

Etkin piyasa hipotezinin varsayımları aşağıdaki şekilde listelenebilir:

- Piyasada çok sayıda alıcı ve satıcı vardır. Hiçbir alıcı veya satıcı piyasayı tek başına etkileyecek güçte değildir.
- Menkul kıymetle ilgili bilgiler düşük bir maliyetle ve en kısa sürede yatırımcılara sağlanmaktadır.
- Etkin bir piyasada işlem giderleri oldukça düşüktür.
- Piyasaların kurumsal yapısı çok gelişmiştir. Yani düzenleyici mevzuatlar ile piyasaların istikrarlı çalışması sağlanmaktadır.

- Vergi ile ilgili düzenlemeler yoktur.
- Tüm finansal varlıklar tamamen bölünebilir.

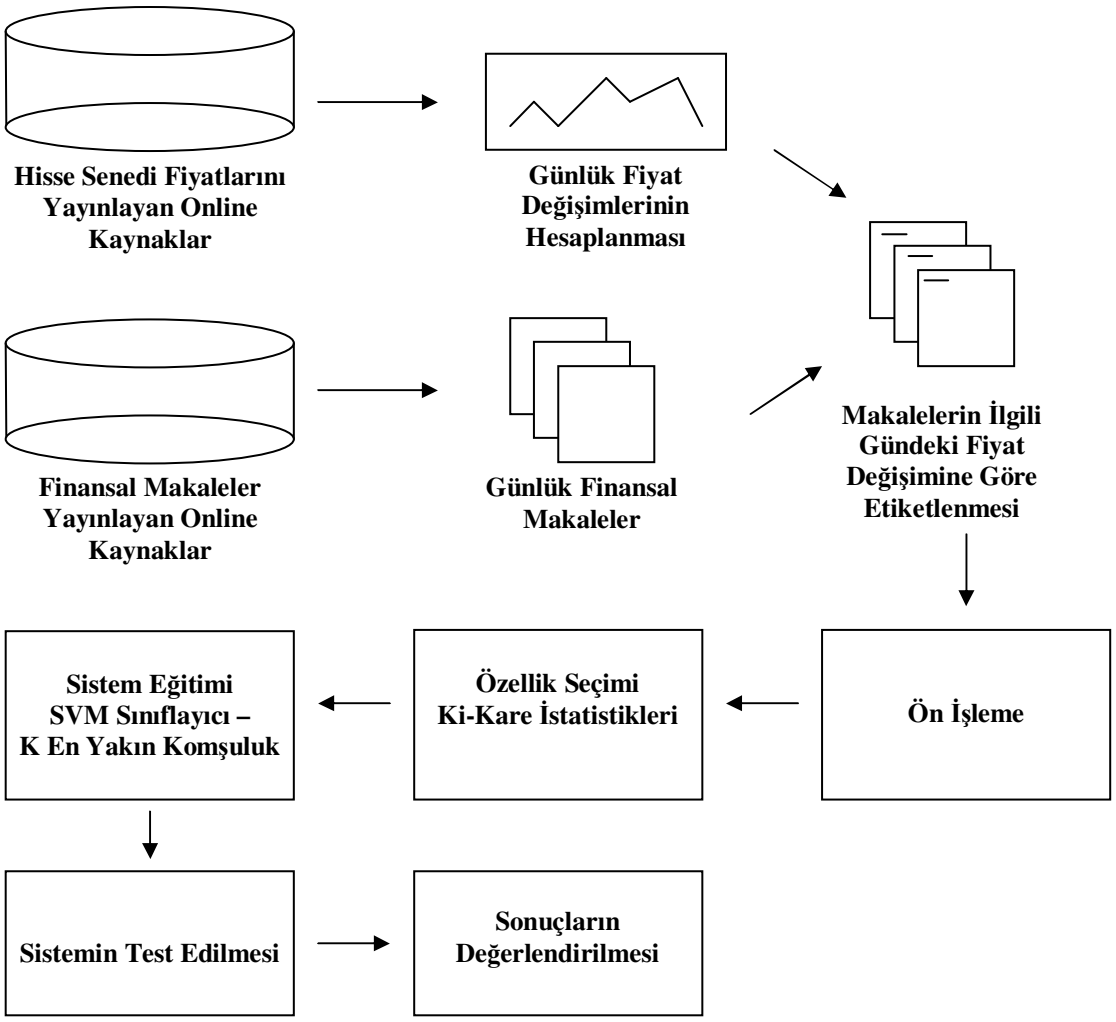
Piyasanın etkin olabilmesi için şartlar; bilgi ve veriler üzerinde tekelleşme olmaması, piyasadaki işlem giderlerinin rekabetçi bir biçimde oluşmasıdır. Bu hipoteze göre 3 çeşit form taşıyan bir piyasa oluşmaktadır. Bunlar zayıf, yarı etkin ve tam etkin formlu piyasalardır. Zayıf form piyasalarda sadece geçmiş verilere ve bilgilere dayanarak yatırımcı yatırım politikasını belirler ve getiri elde eder. Yarı etkin form piyasalarda ise yatırımcıya geçmiş bilgilerin verilerin yanında yatırım yapılacak elemanın iç faktörlerinden yani şirket içinde çalışan, şirketin ne yapacağını bilen biri tarafından bilgi verilmesi gerekmektedir. Bu şart sağlandığı zaman yatırımcı yarı etkin piyasalarda rahatça yatırım yapıp, getiri sağlayabilir. Üçüncü ve son form ise tam etkin piyasa formlarıdır. Bu tarz piyasalarda geçmiş veriler ve şirket içi bilgi dışında diğer tüm farklı bilgilerin de yatırımcıya ulaşması gerekmektedir. Böylece yatırımcı yatırım profilini çizer.

3. SİSTEM TASARIMI

Bu bölümde hisse senetleri fiyatlarındaki değişimin tahmin edilmesi sisteminin tasarımı ve kullanılan teknikler detaylı bir biçimde anlatılmıştır. Bölüm 3.1'de sistemin genel mimarisinden ve süreçler arası akıştan bahsedilmiştir. İlerleyen bölümlerde ise sistem mimarisini oluşturan aşamalar detaylı olarak anlatılmıştır.

3.1 Genel Mimari

Hedeflenen sistemin mimarisi diğer metin madenciliği uygulamalarında kullanılan mimari ile paralellik göstermekle birlikte çalışmanın finansal boyutundan ötürü bazı ek aşamalara sahiptir. Şekil 3.1'de sistemin genel mimarisine yer verilmiştir.



Şekil 3.1 Sistemin genel mimarisi

İlk aşamada belirli bir şirkete ait hisse senedi seçilmektedir. Seçilen hisse senedi için uygun

bir periyottaki tüm fiyat bilgileri ve ilgili hisse senedine ait yayınlanmış makaleler elde edilmektedir. Bir sonraki aşamada hisse senedinin her gün için bir önceki güne göre değişim bilgisi artış veya azalış olarak belirlenmektedir. Bu artış veya azalış bilgisi kullanılarak ilgili günde yayınlanan makaleler olumlu veya olumsuz olarak etiketlenerek sistemin eğitimi ve doğrulama testi gerçekleştirilmektedir. Daha sonra eğitim makaleleri işlenerek sınıflandırılmada kullanılacak özellikler belirlenecektir. Belirlenen özelliklerden sınıflandırmaya etkisi en fazla olan belirli sayıdaki özellik seçilir. Seçilen özellikler ve eğitim makaleleri kullanılarak sistemin eğitim aşaması gerçekleştirilir. Eğitim sonucunda elde edilen model kullanılarak test makaleleri sınıflandırılır ve sistemin başarısı değerlendirilir.

3.2 Hisse Senedi Fiyatları ve Finansal Makalelerin Elde Edilmesi

Hedef sistem herhangi bir dile özel olarak geliştirilmemiştir. Fakat eğitim ve test verileri İngilizce dilinde daha kolay elde edilebileceği ve çalışma uluslararası anlamda diğer çalışmalar ile daha kolay kıyaslanabileceği için sistemin değerlendirilmesi İngilizce dokümanlar kullanılarak yapılmıştır. Ayrıca İngilizce metinlerin cümle ve kelimelere ayrılması işlemini gerçekleştiren açık kaynak kodlu kütüphaneler bulunmaktadır. Bu çalışmada da akademik kullanıma izin verilen bu tür kütüphanelerden faydalanılması amaçlanmıştır.

Çalışma sonucunda ortaya çıkarılacak başarı oranlarının benzer çalışmalarla kıyaslanabilmesi açısından literatürde bu amaçla kullanılan genel erişime açık bir veritabanı elde edilmeye çalışılmıştır. Bu şekilde aynı veritabanını kullanan farklı çalışmaların başarı oranları karşılaştırılabilecek ve benimsenen yöntemin problem üzerine uygunluğu konusunda yorum yapılabilecektir. Bu kapsamda benzer çalışmalarda kullanılan veritabanlarının nasıl elde edildiği konusunda araştırmalar yapılmıştır. Diğer taraftan akademik çalışmalarda kullanılmak üzere veritabanı sağlayan KDD (Knowledge Discrovery in Databases) gibi bazı organizasyonların yayınları incelenmiştir. Bu çalışmalar sonunda proje amacına tam olarak uygun genel kullanıma açık ortak bir veritabanı olmadığı, benzer çalışmalarda veritabanlarının çalışmayı gerçekleştirenler tarafından oluşturulduğu tespit edilmiştir.

3.2.1 Veritabanının Oluşturulmasında Online Haber Kaynaklarının Kullanımı

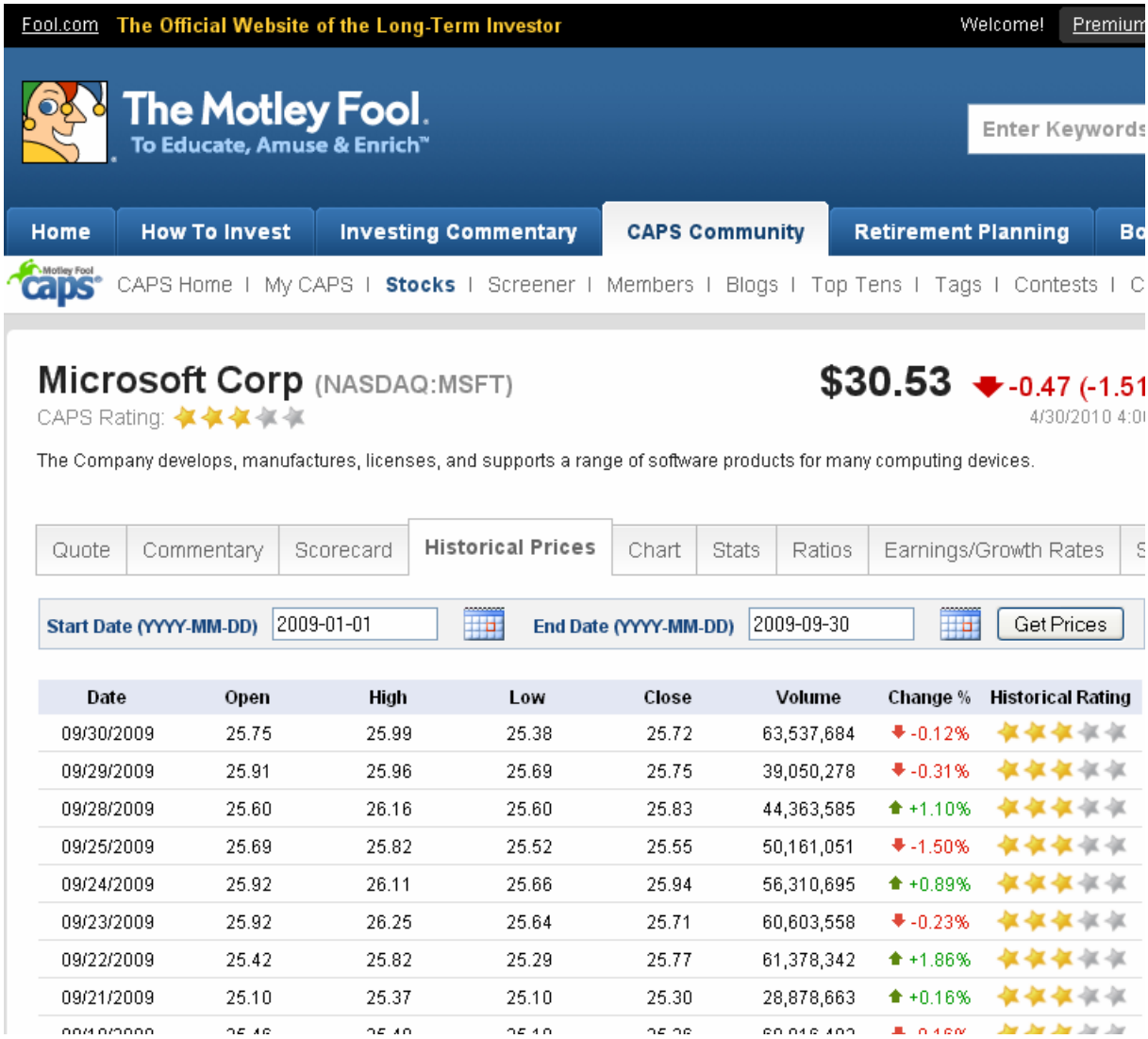
Veritabanının internet üzerinden yayın yapan haber sitelerindeki makalelerden elde edilmesine karar verilmiştir. Bu amaçla http://www.fool.com sitesinde yayınlanan ve Amerika Birleşik Devletleri borsalarında işlem gören şirketlere ait yatırım amaçlı haber makaleleri

kullanılabilmektedir. Ayrıca aynı siteden hisse senetlerinin geçmişe dönük fiyatları da elde edilebilmektedir.

Veritabanının Microsoft firmasına ait hisse senetleri üzerinden hazırlanmasına karar verilmiştir. Microsoft firmasının seçilme nedeni teknoloji şirketlerinin ekonomik kriz, petrol fiyatlarındaki artış gibi global etkenlerden daha az etkilenmesidir. Hisse senedi fiyatının global etkenlerden daha az, şirkete ait haberlere ise daha duyarlı olması hedef sistemin başarı oranının ortaya çıkarılması açısından önemlidir.

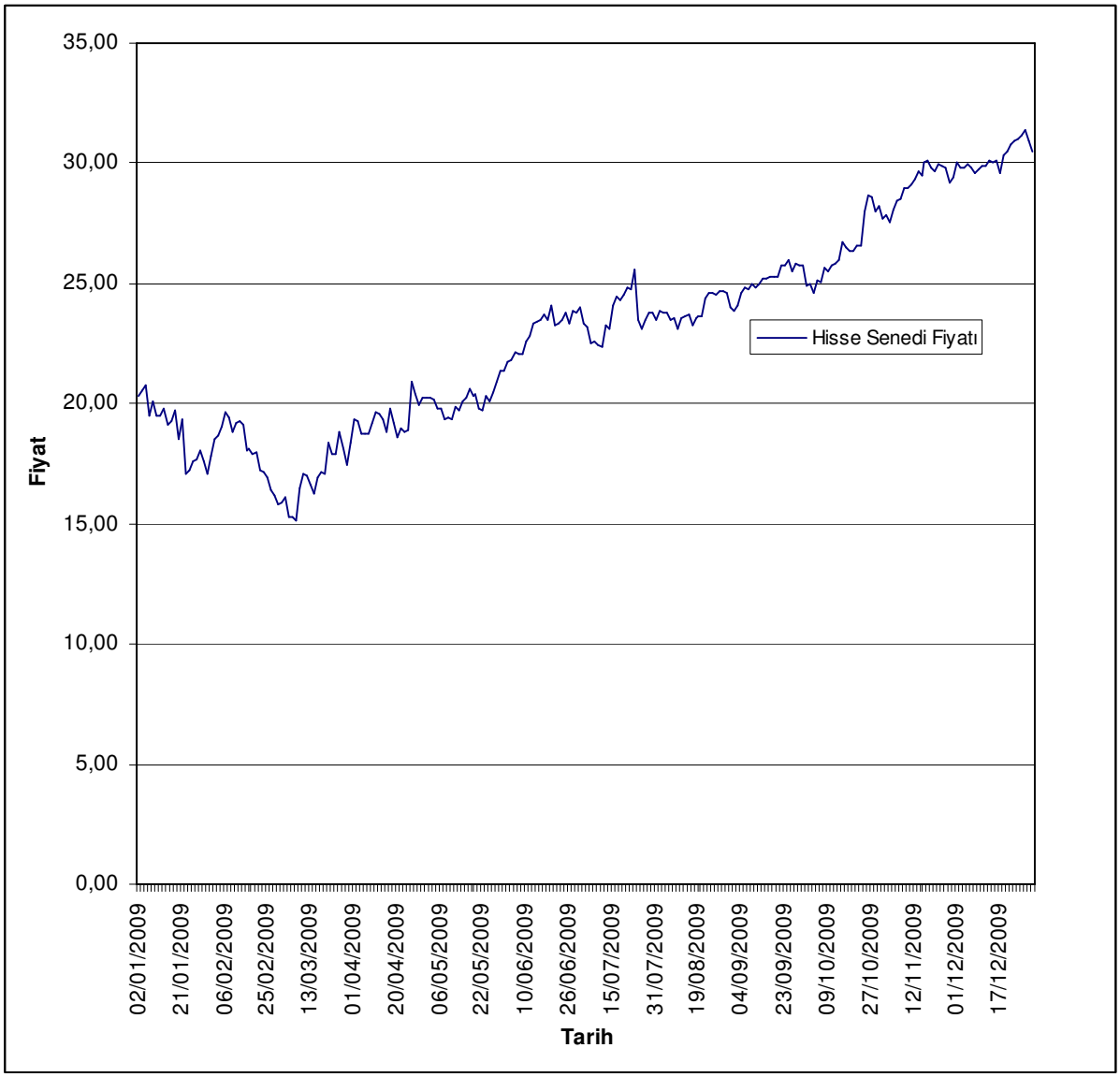
Çalışmanın son bir yıla ait haber makaleleri ve hisse senedi fiyatları üzerinden yapılmasına karar verilmiştir. Son bir yılda yayınlanan haber makalelerinin sistemin başarısını yansıtabilmesi açısından yeterli uzunlukta olduğu düşünülmüştür. Bu sürenin daha uzun tutulmasından belirlenen sürenin bir kısmında global bir ekonomik gelişmenin hisse senedi fiyatlarında etkisinin olması diğer kısımda ise bu global gelişmenin etkisini kaybetmesi gibi olasılıkları artırabileceğinden ötürü kaçınılmıştır. Örneğin son iki yıla ait makaleler ve hisse senedi fiyatları ile çalışılsa idi bu iki yılın ilk yılında global kriz olmadığından ötürü hisse senetleri fiyatlarında global krizin etkisindeki ikinci yıla nazaran genel olarak daha fazla artış olduğu gözlenecekti. Eğitim makaleleri ilgili gündeki fiyat artışına göre etiketlendiği için eğitim makalelerinin içeriklerinde bulunmayan başka bir parametre de kullanılarak etiketlenmiş olacaktı.

http://www.fool.com sitesinde çeşitli borsalarda işlem gören hisse senetlerine ilişkin güncel hisse senetleri fiyatları, ilgili şirkete ait en son haberler ve hisseye yatırım yapmak isteyen yatırımcılara öneriler gibi bilgiler bulunmaktadır. Ayrıca hisse senetlerinin tarihsel fiyatları ve ilgili şirkete ait haber arşivi de yer almaktadır. http://www.fool.com sitesinde hisse senetleri hakkında bilgileri bulunan borsalar arasında New York borsası ve NASDAQ borsası yer almaktadır. Bu çalışmanın başarısının test edilmesi amacı ile seçilen Microsoft şirketine ait MSFT kodlu hisse senedi de NASDAQ borsasında işlem görmektedir. http://www.fool.com sitesinde belirli bir tarih aralığı verilerek ilgili tarih aralığında belirli bir hisse senedinin günlük açılış, kapanış fiyatı, bir önceki güne göre yüzde değişimi gibi bilgiler elde edilebilmektedir. Şekil 3.2'de http://www.fool.com/ sitesinin MSFT hissesi için tarihsel fiyatlar sayfası ekran görüntüsüne yer verilmiştir.



Şekil 3.2 http://www.fool.com/ sitesi MSFT hissesi için tarihsel fiyatlar sayfası

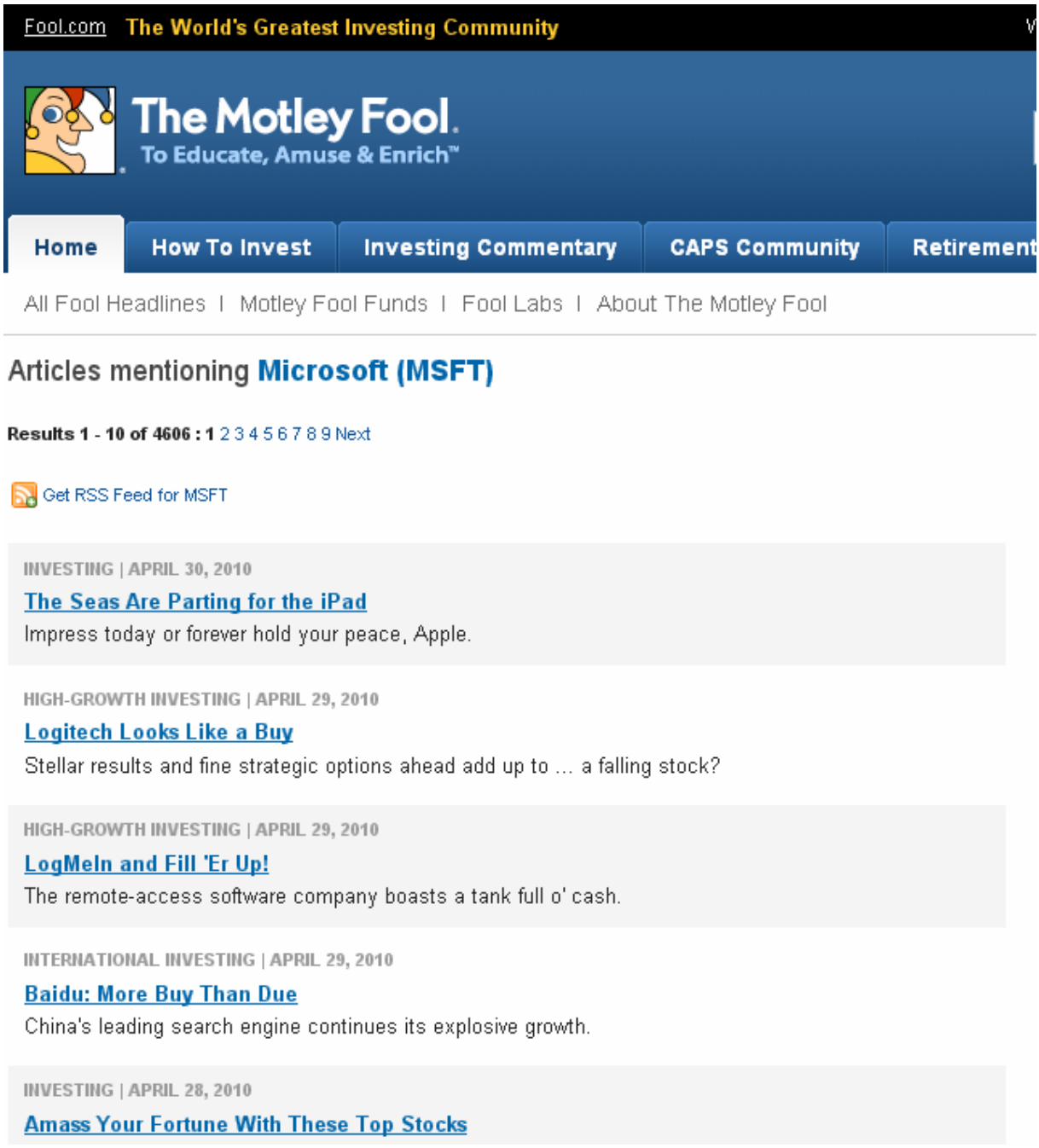
Şekil 3.2'de ekran görüntüsü verilen sayfaya bağlanılarak son bir yıldaki MSFT hisse senedinin günlük kapanış fiyatı ve bir önceki güne göre yüzde değişimi bilgileri elde edilmiştir. Elde edilen fiyat bilgileri Şekil 3.3'de grafiksel olarak gösterilmiştir. Her gün için fiyatlardaki değişim bilgisi ilgili günde yayınlanan haber makalesinin pozitif veya negatif olarak etiketlenmesinde kullanılmıştır. Örneğin ilgili günde hisse senedi fiyatı artış gösterdi veya aynı kaldı ise ilgili günde yayınlanan haber makaleleri pozitif olarak etiketlenmiştir. Eğer fiyat azalış gösterdi ise ilgili güne ait makaleler negatif olarak etiketlenmiştir. Şekildeki grafik incelendiğinde MSFT hisse senedinin 2009 yılı içerisinde çoğunlukla artış trendi içerisinde olduğu gözlenebilmektedir. Bu artış trendi 2009 yılı içerisinde yayınlanan haber makalelerinde olumlu olarak sınıflanan makalelerin çoğunlukta olacağı anlamına gelmektedir.



Şekil 3.3 MSFT hissesinin 2009 yılına ait fiyat bilgileri grafiği

http://www.fool.com sitesinde belirli bir hisse senedi seçilip ilgili hisse senedinin şirketine ait haber arşivine ulaşılabilmektedir. Bu haber arşivinde yayınlanan her haber makalesi için yayınlanma tarihi bilgisi de bulunmaktadır. Haber arşivinin ana sayfasında en güncel haberden geriye doğru her habere ilişkin sayfaya yönlendirme sağlayan bağlantılar yer almaktadır. Şekil 3.4'te http://www.fool.com/ sitesinin MSFT hissesi için haber arşivi ana sayfası ekran görüntüsüne yer verilmiştir. Ekran görüntüsünde görülen haber başlıkları ilgili habere yönlendirme sağlayan bağlantılardır. Geçmişte yayınlanan tüm haberlere ait bağlantılar tek sayfada kullanıcıya sunulamayacağı için ekranda sayfalama yoluna gidildiği görülmektedir. Haber başlıklarının bulunduğu bölümün hemen üstünde sayfalar arasında gezinmeyi sağlayan ve sayfa numaraları ile belirtilen bağlantılar görülmektedir. Bu bağlantılar

kullanılarak geriye dönük istenilen tarihe kadar yayınlanmış haber bağlantılarına ulaşılabilmektedir.



Şekil 3.4 http://www.fool.com/ sitesi MSFT hissesi için haber arşivi ana sayfası

Sistemin eğitim ve test amaçlı veri setini oluşturan haber makalelerinin ayrı metin dosyaları halinde kaydedilmesi gerekmektedir. Bu amaçla bir uygulama geliştirilip haber arşivinde son bir yıl içerisinde yayınlanan makaleler ve yayınlanma tarihleri otomatik olarak elde edilmiştir. Uygulama Java programlama dili ile geliştirilmiş ve Java internet kütüphanesinden

faydalanılmıştır.

Java internet kütüphanesinde URL adresi verilen internet sayfalarının HTML kaynak kodlarını okuyabilen yordamlar yer almaktadır. HTML kaynak kodları bir internet sayfasının tarayıcıda gösterilmesini sağlayan, sayfa içeriği ile birlikte içeriğin nasıl görüntüleneceği bilgisini barından HTML taglerini de içeren kodlardır. HTML kaynak kodlarından sayfa içeriğini elde edebilmek için HTML taglerinin çıkarılması gerekmektedir.

Geliştirilen uygulama haber arşivinin ana sayfasını okuyarak haber sayfalarına ilişkin URL adres bilgilerini almaktadır. Daha sonra her bir haberin URL adresi kullanılarak haber sayfalarına bağlanılmış ve haber sayfasının HTML içeriği elde edilmiştir. Haber sayfalarında haberin temel içeriğine ek olarak sitede yönlendirmeyi sağlayan değişik bağlantılar veya reklam amaçlı metin, resim ve bağlantılar yer almaktadır. Şekil 3.5'de 28 Ağustos 2009 tarihinde yayınlanan Microsoft firması ile ilgili bir haberin http://www.fool.com/ sitesindeki ekran görüntüsüne yer verilmiştir.

Şekil 3.5 http://www.fool.com/ sitesi MSFT hissesi için bir haber sayfası

Ekran görüntüsünde görüldüğü üzere haber metni sayfanın sol alt bölümünde yer almaktadır. Sayfadaki diğer bölümlerin haber içeriği ile bir ilgisi bulunmamaktadır. Uygulama tarafından haber sayfalarının HTML içeriğinden bu kısımlar çıkarılmıştır. Sayfadaki haber metni bölümü incelendiğinde haber metni bölümünde de bazı kelimelerin diğer sayfalara bağlantı olarak kullanıldığı görülmektedir. Bu kelimeler altı çizgili ve mavi renkli font ile takip edilebilmektedir. Haber içeriğine ilişkin HTML kısımdan HTML tagleri ayıklanarak haber

metni net olarak elde edilmiştir. Şekil 3.5'de internet sayfası görüntüsü verilen haber makalesinin ayıklama işlemlerinden geçirildikten sonra elde edilen salt içerik hali Şekil 3.6'da görülmektedir.

```
Microsoft(Nasdaq: MSFT) is a master at playing "follow the leader." This
time, the Redmond Rumbler is dropping the price of its fanciest Xbox 360
video game system by $100, only days after Sony   (NYSE: SNE) did the
same to its PlayStation 3 system. Both the PS3 Slim and Xbox 360 Elite
will cost you about $300 and come equipped with 120 GB of hard-drive
storage.
Of course, Microsoft would argue that the price drop was more of an open
secret, and that the near-simultaneous timing just a coincidence. Aaron
Greenberg, a director of product management for Microsoft's gaming
division, tells Ars Technica that the price drop had been planned for
months, as part of a product life cycle Microsoft plans "years in
advance." Whatever the case, consumers will see the high-end versions of
both Sony's and Microsoft's consoles taking 25% cuts to their price
tags, while adding more storage and new features.
That could be enough to spark a powerful upgrade cycle, which might
spell trouble for Nintendo (OTCBB: NTDOY.PK). The Nintendo Wii was never
the most powerful video game unit available, and it lacks even the
option to add a hard drive without hacking the system. But its $250
price point and unique, family-friendly games have kept the Wii
outselling all comers for years. Those days may be over for Nintendo,
unless the company trims prices on the Wii as well. The competition is
getting a wee bit too close.
I believe that these console price drops should kick-start the stalled
engines of the whole sector -- just in time for the critical holiday
shopping season. However, this two-pronged catalyst may not yet be
priced into the value of the game producers. So this could be the
perfect time to pick up a few shares in Take-Two Interactive   (Nasdaq:
TTWO), Activision Blizzard   (Nasdaq: ATVI), THQ   (Nasdaq: THQI), or
Electronic Arts   (Nasdaq: ERTS) before Mr. Market reacts to
skyrocketing game system sales. Where there's system sales, the games
will follow.
```

Şekil 3.6 Örnek bir haber makalesi içeriği

Uygulama http://www.fool.com/ sitesine bağlanıp yukarıda bahsedilen tüm bu adımları takip ederek son bir yılda MSFT hisse senedi ile ilgili yayınlanan makalelerini okumaktadır. Haber makalelerinin salt içerikleri metin dosyası olarak bir dizine kaydedilmiştir. Uygulama aynı zamanda makalelerin ait olduğu güne ait hisse senedi fiyatı değişimini kullanarak metin dosyasının ilk satırında makalenin pozitif veya negatif olduğu bilgisini de eklemektedir. Metin dosyasının ikinci satırında makalenin ait olduğu tarih bilgisi bulunmaktadır. Üçüncü ve sonraki satırlarda makalenin içeriği yer almaktadır.

3.2.2 Haber Makalesi Dokümanlarının Birleştirilmesi

Son bir yıl içerisinde yayınlanan 949 adet haber makalesi veri setine eklenmiştir. Haber makaleleri nispeten kısa olduğu için özellik olarak belirli bir kelime çiftinin bir makalede geçme ihtimali düşük olmaktadır. Bu nedenle aynı gün içerisinde yayınlanan tüm haberler tek bir doküman altında toplanarak kelime çiftlerinin dokümanlarda geçme ihtimali artırılmaya çalışılmıştır. Aynı günde yayınlanan tüm haber makalelerinin sınıfı aynı olacağı için böyle bir düzenlemenin gerçekleştirilmesinde bir sakınca bulunmamaktadır. Bu kapsamda son bir yıldaki 182 iş gününde yayınlanan 949 haber 182 farklı dokümanda birleştirilmiştir.

3.3 Finansal Makalelerin Etiketlenmesi

Belirli bir şirketin belirli bir döneme ait hisse senedi fiyatları ve şirketle ilgili finansal makaleler elde edildikten sonra bu finansal makalelerin etiketlenmesi gerekmektedir. Finansal makalelerin dahil olabileceği etiketler olumlu veya olumsuz olabilmektedir. Olumlu makale ilgili şirketin hisse senedinin fiyatının artmasında etkisi olan makale anlamına gelmektedir. Olumsuz makale ise makalenin ait olduğu şirketin hisse senedi fiyatının düşmesinde etkisi olduğu düşünülen makale sınıfıdır.

Bir finansal makalenin olumlu veya olumsuz olarak etiketlenmesinde makalenin yayınlandığı gündeki hisse senedinin fiyat değişimi kullanılmaktadır. Bu amaçla hisse senetlerinin günlük fiyatlarından günlük fiyat değişimleri hesaplanır. Bir hisse senedinin günlük fiyat değişimi belirli bir gündeki hisse senedi fiyatının bir önceki gündeki hisse senedi fiyatından farkının bir önceki gün fiyatına bölünmesi ile elde edilir. Hisse senetleri sadece iş günlerinde işlem gördüğü için bir önceki gün hafta sonu veya tatile denk geliyorsa bir önceki gün fiyatı olarak bir önceki iş günü fiyatı alınır. P_t t. güne ait fiyat, P_{t-1} ise (t-1). güne ait fiyat olarak kabul edilirse ΔP fiyat değişimi (3.1)'deki bağıntıya göre hesaplanır.

$$\Delta P = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{3.1}$$

ΔP'nin sıfırdan büyük olması ilgili günde hisse senedi fiyatında bir artış olduğu anlamına gelmektedir. ΔP'nin sıfırdan küçük olması ise ilgili günde hisse senedi fiyatında bir düşüş olduğu anlamına gelmektedir. ΔP'nin sıfıra eşit olması da olumlu olarak değerlendirilir. Makale yayınlandıktan sonra eğer hisse senedi fiyatında bir artış olmuş ise ilgili makale olumlu olarak etiketlenir. Eğer makale yayınlandıktan sonraki gün hisse senedi fiyatı düşmüş ise ilgili makale olumsuz olarak etiketlenir. Olumlu etki gösteren makaleler 1, olumsuz etki

gösteren makaleler ise -1 olarak temsil edilmektedir. Bu durumda ΔP ve makale sınıfı arasındaki ilişki (3.2) bağıntısı ile ifade edilebilir.

$$\text{Makale Sınıfı} \begin{cases} \Delta P >= 0 \longrightarrow 1 \\ \Delta P < 0 \longrightarrow -1 \end{cases} \tag{3.2}$$

182 dokümandan 104 tanesi hisse senedi fiyatının arttığı günü temsil ederek olumlu olarak etiketlenmiştir. 78 doküman ise hisse senedi fiyatının azaldığı günü temsil ederek olumsuz olarak etiketlenmiştir.

3.4 Ön İşlemler

Ön işleme bir veri madenciliği uygulamasında veri setinin ön işlemeden geçirilerek özelliklerin elde edilmesine hazır hale getirilmesidir. Ön işlemler kapsamında veri temizliği, veri bütünleştirme, veri dönüştürme ve veri azaltımı gibi işlemler gerçekleştirilir. Veri temizliği veri setindeki eksik verilerin tamamlanması, gürültülü ve tutarsız verilerin çıkarılması işlemidir. Veri bütünleştirme farklı veri kaynaklarından gelen verilerin birleştirilmesidir. Veri dönüştürme işlemi ham verinin veri madenciliğin işleyebileceği formlara dönüştürülmesidir. Veri azaltımı işleminde ise veri içerisindeki bilgi kaybı en aza indirgenerek verinin azaltılması gerçekleştirilir.

Sınıflandırmaya dayalı veri madenciliği çalışmalarında veri setindeki ham verilerin ön işlemeye tabi tutularak özellik verilerine dönüştürülmesi gerekmektedir. Sınıflandırma çalışmasının konusuna ve kullanılan veriye göre belirlenen özellikler farklı olabilmektedir. Örneğin bir resim formatındaki verinin veri madenciliği uygulamasında resimde kullanılan renkler özellik olarak belirlenebilir. Çünkü resimde kullanılan renk, resim bilgilerinin değerlendirilmesinde önemli bir bilgi olabilir. Ses formatındaki verilerin sınıflandırmasına yönelik bir uygulamada seslerin frekansları önemli bir bilgi olabilmektedir. Bir metin madenciliği uygulamasında da metinlerin sınıflandırılması için sınıflandırmada kullanılacak özelliklerin belirlenmesi gerekmektedir.

Metin formatındaki verilerin sınıflandırılmasında kullanılan en klasik veri dönüştürme tekniği metinde bulunan kelimelerin metinde geçme frekanslarının özellik olarak belirlenmesidir. Fakat bu veri dönüştürme yönteminin başarısı uygulamanın konusuna bağlıdır. Örneğin bu

yöntem metin yazarının belirlenmesi veya bir e-postanın reklam amaçlı olup olmadığının belirlenmesi uygulamalarında oldukça başarılı olabilmektedir. Fakat finansal makalelerin sınıflandırılmasında ne derece başarılı olduğu tartışmalıdır. Bir makalede isim türünden bir kelimenin geçmesi bilgisi tek başına ilgili makalenin hisse senedinin fiyatına olumlu veya olumsuz etkisi olduğu konusunda geçerli bir bilgi vermemektedir. Örneğin bir makalede "satışlar" kelimesinin geçmesi ilgili makalenin hisse senedi fiyatı üzerinde olumlu veya olumsuz etki yarattığı konusunda net bir fikir vermemektedir. Çünkü "satışlar" kelimesi "A şirketinin satışları arttı" gibi bir cümlede olumlu etkiye sahipken "A şirketinin satışları azaldı" gibi bir cümlede olumsuz etkiye sahip olabilmektedir. Aynı şekilde fiil türünden bir kelimenin makalede geçmesi bilgisi tek başına ilgili makalenin hisse senedinin fiyatına olumlu veya olumsuz etkisi olduğu konusunda net bir fikir vermemektedir. Örneğin "azaldı" gibi bir fiil türünden kelimenin bir makalede geçmesi de makalenin olumlu veya olumsuz olduğu konusunda fikir verememektedir. "B şirketinin karı geçen yıla göre %10 azaldı" cümlesinde olumsuz bir etki verirken "B şirketinin zararı geçen yıla göre %10 azaldı" cümlesinde olumlu bir etki vermektedir.

Bu çalışmada aynı cümlede geçen isim ve fiil çiftleri özellik olarak belirlenmiştir. Çünkü aynı cümlede geçen biri isim diğeri fiil olan kelime çifti makalenin olumlu veya olumsuz olduğu konusunda net bir fikir verebilmektedir. Yukarıda örnek olarak verilen cümlelere tekrar göz atarsak "A şirketinin satışları arttı" cümlesinde "satışları" ve "arttı" kelime çifti makalenin olumlu bir etkiye sahip olduğu bilgisini verebilmektedir. Diğer örnek "B şirketinin karı geçen yıla göre %10 azaldı" cümlesinde ise "karı" ve "azaldı" kelime çifti makalenin olumsuz etkiye sahip olduğu fikrini çıkarmaktadır.

Çalışma İngilizce dilindeki makaleler üzerinden yapılmıştır. İngilizce dilinde kelime sonunda eklerin nispeten daha az olması benimsenen yöntemin bu dilde daha başarılı sonuçlar verebileceğininin öngörülmesini sağlamıştır. Çünkü kurgulanan sistem yazılışı farklı olan her kelimeyi farklı olarak algılamaktadır. Örneğin Türkçe' de "artmak" fiili makaleler içerisinde "arttı", "artmıştır", "artacak", "artması bekleniyor" şeklinde farklı ekler alarak geçebilir. Fakat İngilizce'de "increase" kelimesi genellikle "increased" ya da sadece "increase" olarak kullanılmaktadır.

"X Company sales increased 20 percent from last year" cümlesi için Çizelge 3.1' de belirtilen kelime ikilileri özellik olarak çıkartılır.

Çizelge 3.1 Örnek bir cümle için özellik olarak seçilen kelime ikilileri

1. Kelime (İsim)	**2. Kelime (Fiil)**
X	increased
Company	increased
sales	increased
20	increased
percent	increased
year	increased

Örnek cümle için geliştirilen sistem 6 farklı kelime ikilisini özellik olarak çıkartır. Örnek cümlede fiil türünde sadece "increased" kelimesi bulunduğu için tüm özelliklerin fiil olan ikinci kelimesi "increased" olarak belirlenmiştir. "X", "Company", "sales", "20", "percent" ve "year" kelimeleri ise isim türünde olduğu için özelliklerin isim olan ilk kelimesinde yer almaktadırlar.

Hedef sistemin yazılımının gerçekleştirilmesi aşamasında literatürde akademik çalışmalarda kullanılmak üzere sunulan yazılım kütüphanelerinden mümkün mertebe faydalanılmıştır. Bu şekilde gerçeklemeye ayrılan efor sistemin tasarımının zenginleştirilmesine ve aynı çalışmada daha fazla yöntemin kullanılabilmesine olanak sağlamıştır. Open NLP kütüphanesi de haber makalelerinin içeriğinde yer alan cümle ve kelimelerin işlenmesinde kullanılmıştır.

OpenNLP [1] doğal dil işleme üzerinde gerçekleştirilen açık kaynak kodlu projelerin toplandığı bir organizasyondur. OpenNLP üzerinden Java tabanlı doğal dil işleme kütüphanelerine ulaşılabilir. Bu kütüphaneler kelime tespiti, tokenlara ayırma, isimlendirilmiş bileşen tespiti, pos-tagging gibi fonksiyonları gerçekleştirmede kullanılabilir.

OpenNLP kütüphanesinin sağladığı cümlelere ayırma fonksiyonu kullanılarak her bir makale cümlelere ayrılmıştır. Elde edilen cümlelerdeki kelimeler ayrıştırılarak her kelimenin türü elde edilmiştir. OpenNLP kütüphanesi maximum entropy yöntemini kullanarak cümlede geçen her bir kelimenin türünü belirleyebilmektedir. Aynı cümlede geçen ve türleri belirlenen kelimelerden biri isim diğeri fiil olacak şekilde tüm kelime ikilileri çıkarılarak genel bir özellik listesinde saklanmıştır. Bu işlem yapılırken bitiş kelimelerini içeren ikililer hariç tutulmuştur. Örneğin "is", "are" gibi fiillerin geçtiği tüm ikililer veya "it", "they" gibi isim türünden kelimelerin geçtiği tüm ikililer özellik setine dahil edilmemiştir. Tüm dokümanlardaki her cümle için aynı işlemler tekrar edilerek veritabanındaki tüm kelime ikilileri elde edilmiştir.

3.5 Özellik Seçimi

"X Company sales increased 20 percent from last year" örnek cümlesi için Çizelge 3.1'de belirtilen kelime ikilileri özellikleri incelendiğinde bu özelliklerden hepsinin sınıflandırmada etkisi olmadığı görülmektedir. Örneğin "X" ve "increased" kelime ikilisinin bir makalede aynı cümle içinde geçmesi makalenin hisse senedi fiyatına olumlu veya olumsuz bir etki yaratacağı konusunda net bir fikir verememektedir. Bu iki kelimenin beraber geçtiği bir cümlede arttığı ifade edilen nesne firmanın kârı da olabilir. Bu durumda olumlu bir makale olarak sınıflandırılabilir. Fakat arttığı ifade edilen nesne firmanın zararı da olabilir. Bu durumda ise olumsuz bir makale olarak sınıflandırılmalıdır.

Uygun kelime ikilileri özellik olarak çıkarıldıktan sonra bu özelliklerden sınıflandırmada en çok etkisi olan ve uygun sayıda özellik içeren bir alt kümenin seçilmesi gerekmektedir. Çünkü çıkarılan her özelliğin sınıflandırmaya etkisi olmayabilir hatta sınıflandırmanın başarısını azaltabilir.

Özellik seçimi işlemi için literatürde kullanılan bazı yöntemler incelenmiştir. Bu yöntemlerden bazılarının çalışma prensibi aşağıda anlatılmıştır.

3.5.1 Özellik Seçimi Yöntemleri

Bu bölümde şimdiye kadar metin sınıflandırma çalışmalarında kullanılan farklı özellik seçimi yöntemlerinden bahsedilmiştir. Ayrıca özellik seçim yöntemlerinin performanslarını karşılaştırılan çalışmalarda elde edilen sonuçlara yer verilmiştir.

3.5.1.1 Kelimenin Bulunduğu Doküman Sayısı

Kelimenin bulunduğu doküman sayısı yönteminde bir kelimenin kaç farklı dokümanda geçtiği dikkate alınır. Eğitim setindeki dokümanlar kullanılarak her bir kelimenin kaç farklı dokümanda geçtiği tespit edilir. Doküman sıklığı belirli bir eşik seviyesinden düşük olan kelimeler çıkarılır. Apte vd. (1994) çalışmasında kelimelerin bulunduğu doküman sayısı metodunu kullanmıştır. Bu yöntem çalışmalarda birincil özellik seçimi yöntemi olarak kullanılmaktan ziyade genellikle etkinliği artırma amaçlı kullanılmaktadır. (3.3) eşitliğine göre i özelliğinin c sınıfındaki doküman sayısı i kelimesini içeren ve c sınıfına dahil olan doküman adedidir.

$$DS(i,c) = P(i,c) \tag{3.3}$$

3.5.1.2 Bilgi Kazanımı

Bilgi kazanımı makine öğrenmesi alanında sıklıkla bir özelliğin sınıflandırmadaki kullanışlılığını sınamak amacı ile kullanılır. Bu yöntemde bir özelliğin bulunması veya bulunmamasının bilinmesinin sınıflandırmaya kattığı bilginin bit cinsinden değeri hesaplanmaktadır. Her bir özelliğin bilgi kazanımı hesaplanır ve belirli bir eşik değerinin altındaki özellikler elenir. Lewis ve Ringuette (1994) bilgi kazanımı yöntemini Naive Bayes ve karar ağaçları yöntemleri ile ikili sınıflandırma yaptığı çalışmasında kullanmıştır. i sınıfının c dokümanındaki bilgi kazanım değeri (3.4) eşitliğine göre hesaplanır.

$$BK(i,c) = P(i,c) \cdot \log \frac{P(i,c)}{P(c) \cdot P(i)} + P(\bar{i},c) \cdot \log \frac{P(\bar{i},c)}{P(i) \cdot P(\bar{c})} \tag{3.4}$$

3.5.1.3 Ortak Bilgi

Ortak bilgi istatistiksel dil modelleme ve kelime ilişkilendirmede sıklıkla kullanılan bir ölçümleme yöntemidir. Wiener vd. (1995) ortak bilgi yöntemini ki-kare istatistik yöntemi ile beraber özellik seçimi amacı ile kullanmıştır. Ortak bilgi yönteminin zayıf özelliği kelimelerin aykırı olasılıklarından güçlü bir şekilde etkilenmesidir. Örneğin aynı koşullu olasılığa sahip kelimelerden daha az rastlanan kelimeler daha yüksek skorlara sahip olmaktadır. Bu nedenle farklı sıklıkta kullanılan kelimelerin skorlarının karşılaştırılması pek mümkün olmamaktadır. (3.5) eşitliğinde i özelliğinin c dokümanındaki ortak bilgi değerinin nasıl hesaplandığı belirtilmektedir.

$$OB(i,c) = \log \frac{P(i,c)}{P(i) \cdot P(c)} \tag{3.5}$$

3.5.1.4 Ki-kare İstatistiği (Chi Square Statistic)

Ki-kare istatistikleri yönteminde bir sınıf ve bir özellik arasındaki bağımlılık ölçülür. Her sınıf için eğitim setindeki tüm özellikler ile bu sınıf arasındaki ki-kare istatistikleri hesaplanır. Ki-kare istatistiği ve ortak bilgi yöntemi arasındaki temel fark ki-kare istatistiği yönteminde normalize edilmiş skorların kullanılmasıdır. Bu nedenle ki-kare istatistiği yönteminde elde edilen skorlar karşılaştırılabilir olmaktadır. Fakat çok düşük sıklıkta tekrar eden özellikler için bu normalizasyon işe yaramamaktadır. Bu nedenle ki-kare istatistiği yöntemi düşük sıklıktaki özellikler için uygun değildir. (Dunning, 1993)

Bu yöntemde i özelliğinin c sınıfındaki ki-kare istatistik değeri (3.6) eşitliğine göre

hesaplanır.

$$X^2(i,c) = \frac{N \cdot \left[P(i,c) \cdot P(\bar{i},\bar{c}) - P(i,\bar{c}) \cdot P(\bar{i},c)\right]^2}{P(i) \cdot P(\bar{i}) \cdot P(c) \cdot P(\bar{c})} \tag{3.6}$$

3.5.1.5 Terim Gücü

Terim gücü yöntemi ilk olarak Wilbur ve Sirotkin (1992) tarafından ortaya atılmış ve metin erişimi çalışmalarında sözlük azaltımında kullanılmak amacı ile geliştirilmiştir. Daha sonra Yang (1995) tarafından metin sınıflandırma çalışmalarında gürültülü veriyi azaltmak amacı ile kullanılmıştır. Yang ve Wilbur (1996) ayrıca terim gücü yöntemini doğrusal regresyon ve en yakın komşuluk sınıflandırma çalışmalarında değişken sayısının azaltımında kullanmışlardır. Bu yöntemde bir terimin yakın ilişkili dokümanlarda gözükme olasılığı temel alınarak terimin önemi tahmini olarak hesaplanmaktadır. Yakın ilişkili dokümanlar bulunması eğitim setindeki dokümanlar kullanılarak birbirine benzerliği belirli bir eşik değerinden yüksek olan doküman çiftleri elde edilmesi ile gerçekleştirilir. i özelliğinin terim gücü değeri (3.7) eşitliği ile hesaplanır.

$$TG(i) = P_r\left(i \in y \middle| t \in x\right) \tag{3.7}$$

3.5.1.6 Eşitsizlik Oranı

Eşitsizlik oranı ilk olarak Rijsbergen (1979) tarafından ilgililik geribildirimi için terim seçmek amacı ile geliştirilmiştir. Bu yöntemin ardındaki temel prensip, özelliklerin alakalı dokümanlardaki dağılımının alakasız dokümanlar üzerindeki dağılımından farklı olduğudur. Bu yöntem Mladenic (1998) tarafından metin sınıflandırmada terim seçimi amacı ile kullanılmıştır. i özelliğinin c dokümanına ait eşitsizlik oranı değeri (3.8) eşitliğindeki gibi hesaplanabilmektedir.

$$EO(i,c) = \frac{P(i,c) \cdot \left[1 - P(i,\bar{c})\right]}{\left[1 - P(i,c)\right] \cdot P(i,\bar{c})} \tag{3.8}$$

3.5.1.7 Korelasyon Katsayısı

Korelasyon katsayısı yöntemi Ng vd. (1997) ve Sebastiani (2002) tarafından tanımlanmıştır. Bu yöntem ki-kare istatistiği yönteminin farklı bir uyarlamasıdır. (3.9) eşitliğinde i özelliğinin c dokümanındaki korelasyon katsayısı değerinin nasıl hesaplanacağı belirtilmiştir.

$$KK(i,c) = \frac{\sqrt{N} \cdot \left[P(i,c) \cdot P(\bar{i},\bar{c}) - P(i,\bar{c}) \cdot P(\bar{i},c)\right]}{\sqrt{P(i) \cdot P(\bar{i}) \cdot P(c) \cdot P(\bar{c})}} \quad (3.9)$$

3.5.1.8 GSS Katsayısı

GSS Katsayısı yöntemi Galavotti vd. (2000) tarafından tanımlanmış ki-kare istatiği yönteminin farklı bir uyarlamasıdır. i özelliğinin c dokümanına ait GSS katsayısı değeri (3.10) eşitliğindeki gibi hesaplanabilmektedir.

$$GSS(i,c) = P(i,c) \cdot P(\bar{i},\bar{c}) - P(i,\bar{c}) \cdot P(\bar{i},c) \quad (3.10)$$

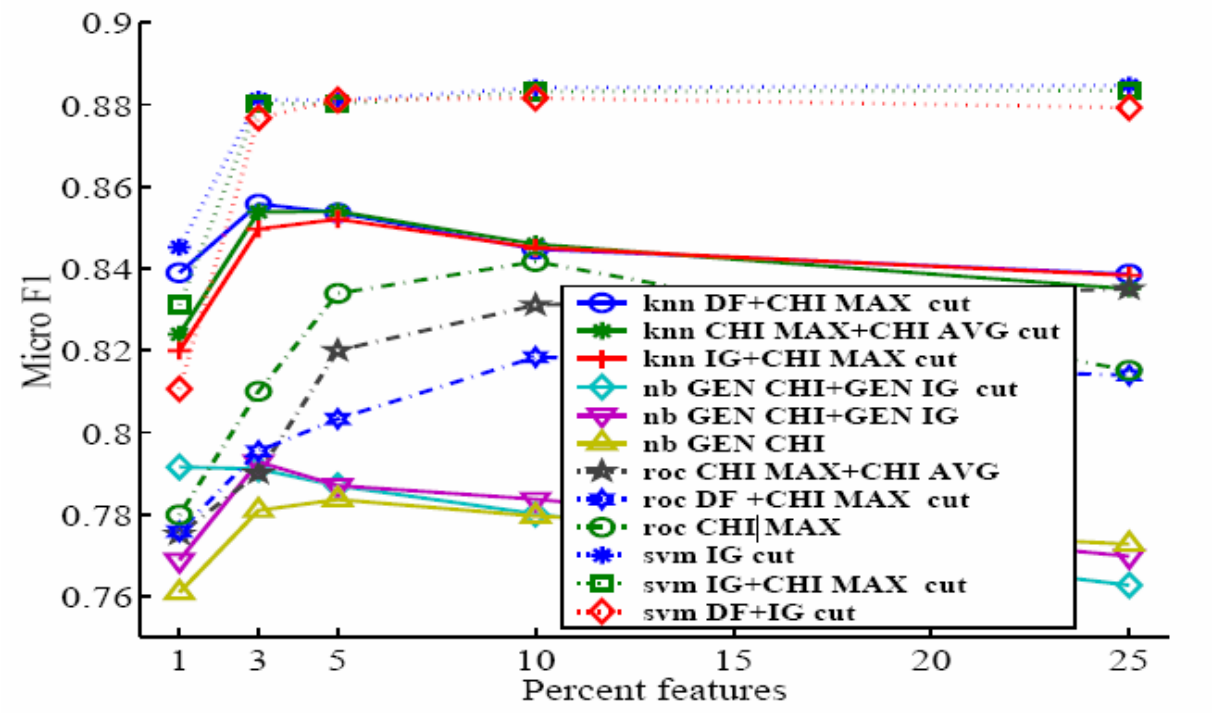
3.5.1.9 Özellik Seçimi Yöntemlerinin Karşılaştırılması

Yang ve Pedersen (1997) metin sınıflandırma çalışmalarında kullanılan özellik seçimi yöntemlerini karşılaştıran bir çalışma gerçekleştirmiştir. Beş farklı metin sınıflandırma yöntemi üç değişik kriter doğrultusunda karşılaştırılmıştır. Bu kriterler genel kelimelere eğilim, sınıf değerlerini dikkate alma ve özelliğin yokluğu durumunu dikkate alma şeklindedir. Değerlendirmeler Reuters-22173 ve Ohsumed veri setleri kullanılarak gerçekleştirilimiştir. Sınıflandırma işlemlerinde k-en yakın komşuluk ve doğrusal en küçük kare uydurma yöntemleri kullanılmıştır. Çalışmanın sonuçları Çizelge 3.2'de gösterilmiştir. Bu sonuçlara göre bilgi kazanımı ve ki-kare istatistikleri yöntemlerinin diğer yöntemlere göre daha başarılı olduğu gözlenmiştir. Bu iki yöntemden ki-kare istatistiklerinin hesaplamaların daha kolay olmasından ötürü uygulamada daha çok tercih edildiği düşünülmektedir. Doküman sayısı yöntemi sınıf değerlerini dikkate alma ve özelliğin yokluğu durumunu dikkate alma kriterlerini karşılamasa da iyi bir sınıflandırma performansı göstermiştir. Ortak bilgi yönteminin ise belirgin olarak diğer yöntemlere göre zayıf olduğu gözlenmiştir.

Çizelge 3.2 Özellik seçimi yöntemleri karşılaştırma sonuçları (Yang ve Pedersen, 1997)

Özellik Seçimi Yöntemi	**Doküman Sayısı**	**Bilgi Kazanımı**	**Ki-kare İstatistikleri**	**Ortak Bilgi**	**Terim Gücü**
Genel kelimelere eğilim	Evet	Evet	Evet	Hayır	Evet
Sınıf değerlerini dikkate alma	Hayır	Evet	Evet	Evet	Hayır
Özelliğin yokluğunu dikkate alma	Hayır	Evet	Evet	Hayır	Hayır
Sınıflandırma Performansı	Çok iyi	Çok iyi	Çok iyi	Zayıf	İyi

Rogati ve Yang (2002) bilgi kazanımı, doküman sayısı ve ki-kare istatistikleri özellik seçim yöntemlerini karşılaştıran bir çalışma gerçekleştirmiştir. Bu çalışmada Reuters-21578 ve RCV1 veri setleri kullanılmıştır. Naive Bayes, Rocchio algoritması, k-en yakın komşuluk ve destek vektör makinesi yöntemleri sınıflandırma işlemlerinde kullanılmıştır. Her sınıflandırma yöntemi her özellik seçimi yöntemi ile tek tek ve birkaç özellik seçim yönteminin birlikte kullanılması ile test edilmiştir. Her sınıflandırma yöntemi için elde edilen en iyi üç sonuç Şekil 3.7'de grafiksel olarak gösterilmiştir. Grafikte yer alan eğrilerin açıklamalarında kullanılan "+" işareti birden fazla özellik seçim yönteminin birlikte kullanıldığı anlamına gelmektedir. "cut" ibaresi ise çok az sayıda dokümanda geçen kelimelerin elendiğini belirtmektedir. Grafik incellendiğinde kullanılan sınıflandırma yöntemine bakılmaksızın ki-kare istatistik yönteminin tek başına veya diğer özellik seçimi yöntemleri ile birlikte kullanımının en iyi sonuçlar arasında yer aldığı gözlemlenmektedir. Diğer bir tespit ise sınıflandırma yöntemi olarak destek vektör makinesi yönteminin kullanılan özellik seçimi yöntemine en az duyarlı olan yöntem olduğudur.



Şekil 3.7 Özellik seçimi yöntemleri karşılaştırma sonuçları (Rogati ve Yang, 2002)

Seo vd. (2004) ontoloji öğrenme çalışması kapsamında konsept kelimelerin belirlenmesi amacı ile dört farklı özellik seçimi yöntemini kullanmış ve elde ettiği sonuçları karşılaştırmıştır. Bu yöntemler ortak bilgi, ki-kare istatistikleri, Markov Blanket ve bilgi kazanımı yöntemleridir. Çalışma sonucunda ontolojiye has kelimeleri belirlemede bilgi

kazanımı yönteminin en kötü performansı sergilediği görülmüştür. Markov Blanket yöntemi bilgi kazanımı yöntemine göre daha iyi bir performans göstermiştir. Ortak bilgi ve ki-kare istatistikleri yöntemleri diğer yöntemlere göre çok daha iyi sonuçlar üretmiştir. Fakat bu iki yöntem arasında herhangi birinin diğerine net bir üstünlük kuramadığı gözlenmiştir. Her iki yöntem de ontolojiye has anlamlar içeren kelimeleri başarı ile belirleyebilmektedir.

3.5.2 Özellik Seçimi İşleminin Gerçekleştirilmesi

Önceki bölümlerde anlatılan özellik seçim yöntemlerinin litetürdeki uygulamaları incelendiğinde herhangi bir yöntemin diğerine göre performans açısından net bir üstünlük göstermediği tespit edilmiştir. (Falinouss, 2007) Her yöntemin kendine has güçlü ve zayıf yönleri bulunmaktadır. Özellik seçim yöntemlerinin uygulamadaki başarıları daha çok kullanılan veritabanına ve birlikte kullanıldığı sınıflandırma yöntemine bağlı olmaktadır. Metin sınıflandırma çalışmalarında bilgi kazanımı, ki-kare istatistikleri, korelasyon katsayısı ve eşitsizlik oranı yöntemlerinin daha etkili olarak öne çıktığı görülmüştür. Diğer bir deyişle klasik özellik seçim yöntemleri metin sınıflandırma problemleri için en iyi olma özelliklerini devam ettirmektedir. Bu çalışmada özellik azaltılırken sınıflandırmada faydalı bilgi kaybının en aza indirgenmesi ve zamana bağlı özellikleri daha iyi ele almasından ötürü ki-kare istatistikleri yönteminin kullanılmasına karar verilmiştir.

Her özellik için ilgili özelliğin tüm dokümanlardaki ki-kare değerleri hesaplanmış ve bu değerler toplanarak özellik bazında toplam ağırlıklar elde edilmiştir. Ki-kare istatistik değerleri hesaplanırken (3.6) eşitliği kullanılmıştır. Toplam ağırlığı en yüksek olan belirli sayıdaki özelik hedef özellik seti olarak seçilmiştir. Bu işlemler Java programlama dili kullanılarak geliştirilen bir uygulama ile gerçekleştirilmiştir.

Özellik seçimindeki hedef özellik setinde bulunacak özellik sayısı bir sistem parametresi olarak belirlenmiştir. Bu şekilde farklı özellik sayılarının sistemin başarısını nasıl etkilediği analiz edilebilecektir. Hedef özellik setinde az özellik bulunması makaleleri sınıflandırmada ayırıcı olan bazı özelliklerin ihmal edilmesi açısından başarıyı azaltan bir faktör olabilmektedir. Aynı şekilde gerektiğinden fazla özellik kullanmak örneğin makale sınıflandırmada etkili olmayan özelliklerin kullanılması aynı şekilde yanlış sınıflandırma sonuçlarına sebebiyet verebilecektir. Bu nedenle ideal özellik sayısı farklı özellik sayılarında sistemin denenmesi ile elde edilebilecektir.

3.6 Sınıflandırma İşlemi

Özellik seçimi aşamasında sınıflandırma işleminde kullanılacak en etkili özellikler belirlendikten sonra sistemin eğitimi gerçekleştirilmiştir. Sistemin eğitim aşamasında otomatik bir öğrenme algoritması eğitim verilerinden gerekli bilgileri çıkararak her kategori için bir sınıflayıcı oluşturur. (Ng vd., 1997)

80'li yılların sonuna kadar metin sınıflandırma problemlerinde en çok kullanılan yaklaşım bilgi mühendisliğiydi. Bu yaklaşımda sınıflandırma kuralları ilgili konuda uzman olan kişilerin görüşüne başvurularak elle hazırlanmaktaydı. 90'lı yıllarda bu yaklaşım yerini makine öğrenmesi tekniklerine bıraktı. Makine öğrenmesi tekniklerinde sınıfı belirli olan dokümanlar otomatik olarak analiz edilerek bir sınıflayıcı oluşturulmaktadır. Daha sonra bu sınıflayıcı sınıfı bilinmeyen örneğin sınıfını tahmin etmektedir. Makine öğrenmesi teknikleri bazı avantajları da beraberinde getirmiştir. Bu teknikler kullanıldığında elde edilen doğruluk oranları ilgili konuda uzmanların yaptığı sınıflandırmalarla karşılaştırılabilir başarıdadır. Aynı zamanda çok fazla insan eforu gerektirmemektedir.

Son yıllarda birçok istatistiksel makine öğrenmesi tekniği ortaya atılmış ve birçok araştırmacı kendi metin sınıflandırma problemlerinde bu teknikleri başarı ile uygulamıştır. Bu tekniklerden bazıları şu şekildedir: Hiyerarşik sınıflandırma (Larkey, 1998; Jenkins vd., 1999), çok değişkenli regresyon modelleri (Fuhr vd., 1991; Yang ve Chute, 1994; Schutze vd., 1995), k-en yakın komşuluk sınıflandırma (Creecy vd., 1992; Yang, 1994, 1999; Kwon ve Lee, 2000), Bayes sınıflayıcı (Tzeras ve Hartman, 1993; Lewis ve Ringuette, 1994; Lam vd., 1997; McCallum ve Nigam, 1998), karar ağaçları (Lewis ve Ringuette, 1994), yapay sinir ağları tabanlı sınıflandırma (Chen vd., 1994; Yang, 1994; Wiener vd., 1995; Ng vd., 1997; Ruiz ve Sririvasan, 1999), Rocchio algoritması (Buckley vd., 1994; Joachims, 1997), sembolik kural öğrenimi (Apte vd., 1994; Cohen ve Singer, 1996). Ayrıca lojistik regresyon ve destek vektör makinesi (Cortes ve Vapnik, 1995; Joachims, 1998, 2002; Kwok, 1998; Dumais vd., 1998; Hearst, 1998; Dumais ve Chen, 2000; Siolas ve d'Alche-Buc, 2000) gibi doğrusal sınıflandırma teknikleri de metin sınıflandırma yöntemlerinde kullanılmaya başlamıştır. Bu iki yöntem üstündüzlem bulunması konusunda benzerlik barındırmaktadır. Fakat destek vektör makinesi gösterdiği üstün performans ile en ileri sınıflandırma yöntemi olarak yerini korumaktadır. (Zhang ve Oles, 2001)

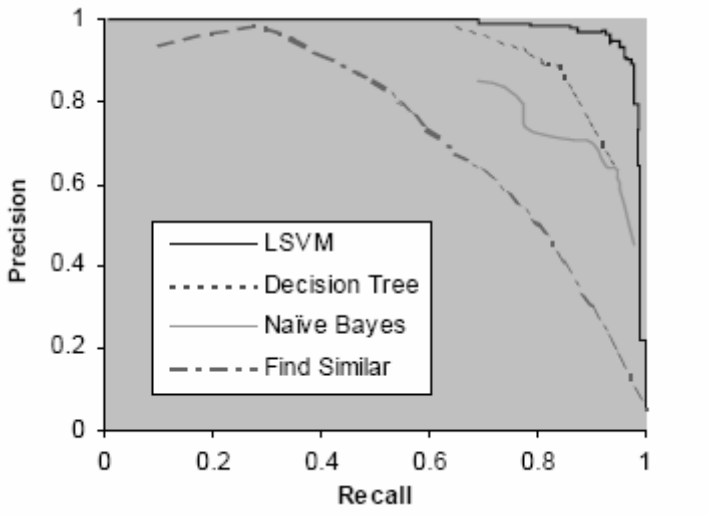
Finansal makalelerin seçilen özellikler ile eğitim ve doğrulama testlerinde destek vektör makinesi ve k-en yakın komşuluk yöntemleri kullanılmıştır.

3.6.1 Sınıflandırma Yöntemlerinin Karşılaştırması

Yukarıda da belirtildiği gibi metin sınıflandırma problemlerinde bir çok farklı yöntem kullanılmaktadır. Bazı çalışmalarda bu farklı yöntemlerin etkinliği karşılaştırılmıştır. Bu bölümde bu karşılaştırma sonuçlarına yer verilmiştir.

Brucher vd. (2002) sonuçların ağırlıklı olarak test veri setine bağlı olduğunu belirtmiştir. Sınıflandırma algoritmaları için belirli parametreler tanımlanmıştır ve bu parametreler çalışmaların performansını etkileyebilmektedir. Kullanılan veritabanına göre parametrelerin en etkili sonuçlar elde edilecek şekilde ayarlanması gerekmektedir. Tüm koşullar ve kısıtlamalar göze alındığında en iyi sonuçların SVM yöntemi ile elde edildiği tespit edilmiştir.

Dumais vd. (1998) çalışmasında beş farklı yöntemi karşılaştırmıştır. Bu yöntemler benzeri bul, karar ağaçları, Naive Bayes, Bayes Net ve destek vektör makineleri yöntemleridir. Çalışmada Reuters-21578 veritabanı kullanılmıştır. Çalışma sonucunda doğrusal destek vektör makinesi en etkili yöntem olarak bulunmuştur. Bu yöntemin oldukça doğru sonuçlar üretmesi ile birlikte eğitim ve test işleminin hızlı ve kolaylıkla yapılabildiği tespit edilmiştir. Şekil 3.8'de bu karşılaştırmanın sonuçlarına yer verilmiştir.



Şekil 3.8 Metin sınıflayıcı yöntemlerinin karşılaştırması (Dumais vd., 1998)

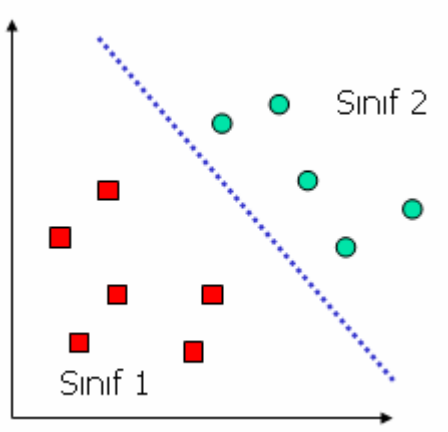
Joachims (1998) çalışmasında destek vektör makinesi yöntemini diğer geleneksel metin sınıflandırma yöntemleri ile karşılaştırmıştır. Bu yöntemler Naive Bayes, Rocchio algoritması, k-en yakın komşuluk ve karar ağaçları yöntemleridir. Veritabanı olarak Reuters-21578 ve Ohsumed veritabanları birlikte kullanılmıştır. Denemeler sonucunda destek vektör makinesi yönteminin diğer yöntemleri performans anlamında geride bıraktığı tespit edilmiştir.

Ayrıca destek vektör makinesi yönteminin radyal tabanlı kernel kullanıldığında polinom kernele göre daha iyi sonuçlar ürettiği gözlenmiştir.

3.6.2 Destek Vektör Makinesi (SVM)

Destek vektör makinesi yöntemi (Cortes ve Vapnik, 1995; Vapnik, 1998) sınıflandırmada en etkili yöntemlerden birisidir. Özellikle metin sınıflandırmasında oldukça başarılı olduğu çeşitli çalışmalarda ispatlanmıştır (Joachims, 1998; Dumais vd., 1998). Destek vektör makinesi iki sınıflı tanıma problemlerinde sıkça kullanılan bir makine öğrenmesi tekniğidir. El yazısı tanıma, yüz tanıma ve metin sınıflandırması gibi büyük ölçekli girdileri olan problemlerde çok iyi performans sağladığı çeşitli çalışmalar ile kanıtlanmıştır (Dumais vd., 1998; Joachims, 1998; Yang ve Liu 1999; Fukomoto ve Suzuki 2001). Destek vektör makinesi yöntemi yapısal risk azaltılması prensibini uygulayan bir istatistiksel öğrenme teorisine dayanır. Ayrıca hem regresyon hem de sınıflandırma görevlerini yerine getirebilmektedir.

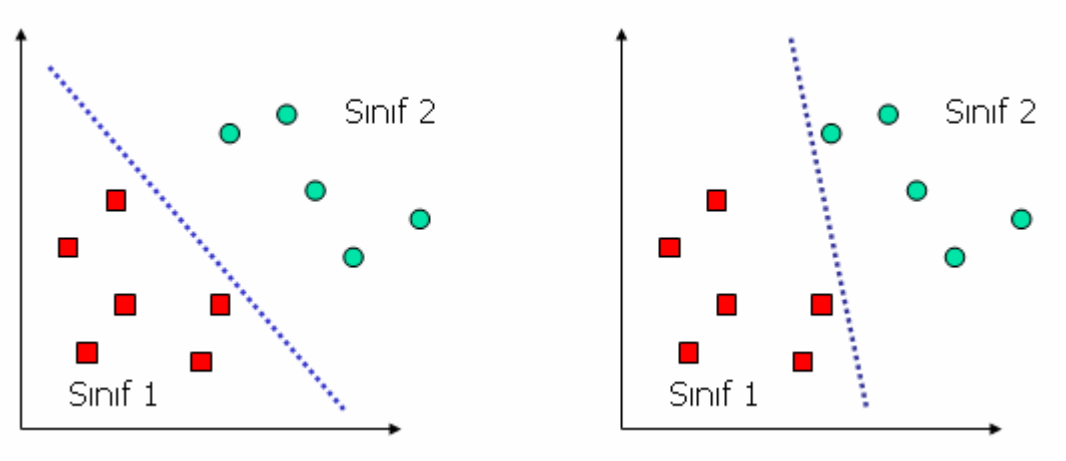
Destek vektör makinesi yöntemi karar sınırlarını belirleyen karar düzlemleri prensibine dayanır. Bir karar düzlemi farklı sınıflara ait örnekleri ayıran düzlemdir. Şekil 3.9'da bir karar sınırı ve sınıflara ayırdığı örnekler görülmektedir. Şekle göre her bir örnek Sınıf 1 veya Sınıf 2 sınıflarından birine ait olmaktadır. Doğru ise karar sınırını belirlemektedir. Buna göre doğrunun solundaki örnekler Sınıf 1 sınıfına, sağındaki örnekler ise Sınıf 2 sınıfına ait olmaktadır.



Şekil 3.9 Örnek bir karar sınırı

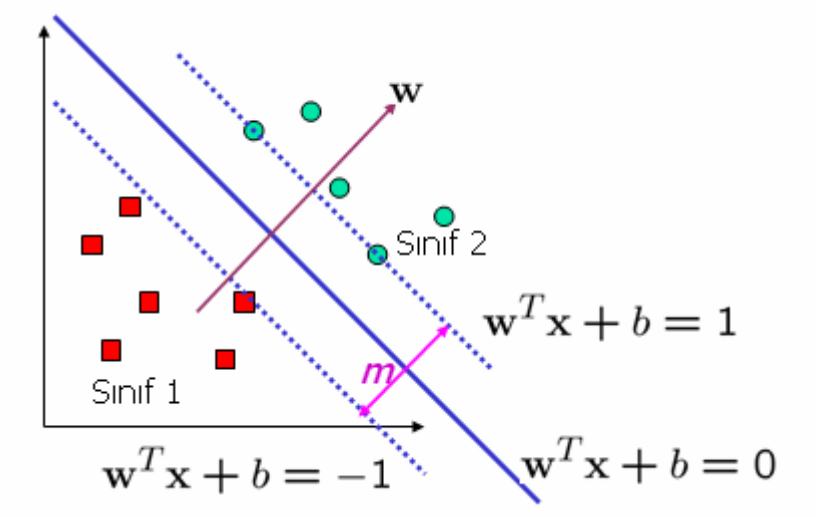
Şekil 3.8'deki sınıfları ayırabilecek çok sayıda karar sınırı çizilebilmektedir. Fakat bu karar sınırlarının hepsi sınıfları ideal bir şekilde ayıramamaktadır. Şekil 3.10'da ideal olmayan bazı

karar sınırları örnekleri gösterilmektedir.



Şekil 3.10 İdeal olmayan bazı örnek karar sınırları

İdeal bir karar sınırı ayırdığı iki sınıfın örneklerine de mümkün mertebe uzak olmalıdır. Şekil 3.11'de ideal bir karar sınırı ve ayırdığı sınıflar görülmektedir. Şekilde görülen kesiksiz kalın doğru üstündüzlemi ifade etmektedir. Üstündüzleme paralel olan kesikli doğrular ise her bir sınıfın sınırlarını ifade etmektedir. Sınırlar üzerindeki örnekler destek vektörler, sınır üzerinde olmayan örnekler normal vektörlerdir. İdeal bir karar sınırı elde edebilmek için şekilde *m* ile gösterilen sınırlar arası marjın maximize edilmesi gerekmektedir. Marjın değeri (3.11) eşitliği ile hesaplanmaktadır.



Şekil 3.11 İdeal bir karar sınırı

$$m = \frac{2}{\|w\|} \tag{3.11}$$

Karar sınırlarını ifade eden üstündüzlem (3.12) eşitliği ile ifade edilmektedir. Eşitliğe göre w^T normal vektör, b ise ofset değeridir.

$$w^T \cdot x + b = 0 \tag{3.12}$$

Maximum marjın oluşmasını sağlayan sınır doğruları (3.13) ve (3.14) eşitlikleri ile gösterilebilmektedir. Bu sınır doğruları arasındaki uzaklık 2/||w|| değerine eşittir. m marj değerinin maximize edilmesi için ||w|| değerinin minimize edilmesi gerekmektedir.

$$w^T \cdot x + b = 1 \tag{3.13}$$

$$w^T \cdot x + b = -1 \tag{3.14}$$

$\{x_1, x_2, \ldots, x_n\}$ değerleri veri setindeki örnekler ve $y_i \in \{1, -1\}$sınıflar olmak üzere eğitim setindeki her örneğin (3.15) eşitliğini sağlayarak doğru sınıflandırılması gerekmektedir.

$$y_i\left(w^T \cdot x + b\right) \geq 1 \tag{3.15}$$

Bu problemi çözmek için problem Lagrangian formülasyonuna çevrilir. Bu şekilde kısıtlar Lagrange çarpanları cinsinden ifade edilerek daha kolay ele alınabilir bir forma dönüştürülür. Problemin Lagrangian formu (3.16) eşitliği ile ifade edilebilmektedir.

$$L = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left(y_i\left(w^T \cdot x_i + b\right) - 1\right) \qquad \alpha_i \geq 0 \tag{3.16}$$

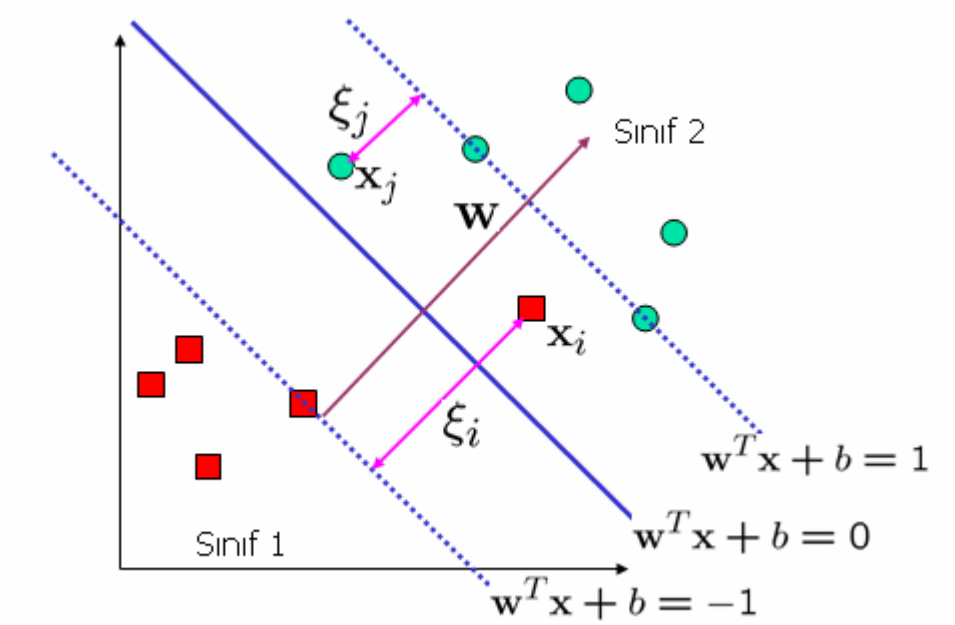
Yukarıdaki eşitlikteki L değeri minimize edilmelidir. Fakat problemin çözümü bu hali ile oldukça zaman alıcı bir işlemdir. Problem (3.17) eşitliğinde belirtilen değişiklikler gerçekleştirilerek bir ikili probleme dönüştürülür.

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,\, j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \qquad \alpha_i \geq 0\,,\ \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.17}$$

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

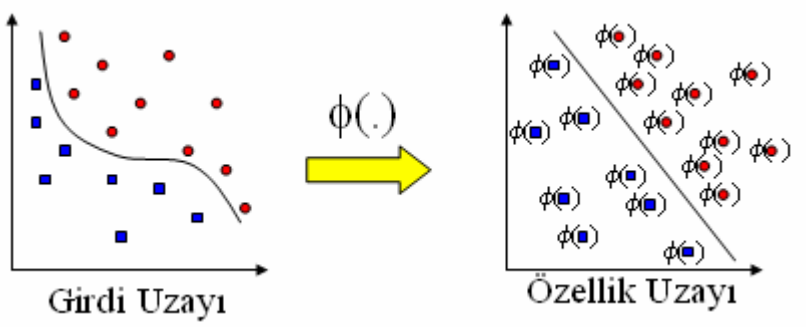
Bu çözümde çoğu örnek için α_i değeri 0'dır. α_i değeri 0'dan farklı olan x_i örnekleri destek vektörleri olarak ifade edilmektedir. Karar sınırları sadece destek vektörleri tarafından belirlenmektedir.

Yukarıdaki bölümler doğrusal sınıflama örnekleri üzerinden anlatılmıştır. Doğrusal sınıflamalarda sınıflar bir doğru ile birbirinden tamamen ayrılabilmektedir. Fakat çoğu sınıflandırma probleminde sınıfların ayrılması bu kadar kolay olmamaktadır. En iyi karar sınırlarını belirlemek karmaşık yapıların oluşturulmasını gerektirebilmektedir. Şekil 3.12'deki örnekte görüldüğü gibi Sınıf 1 ve Sınıf 2 örnekleri doğru şeklindeki bir karar sınırı ile ayrılamaya çalışıldığında ξ olarak belirtilen hatalar oluşmaktadır.



Şekil 3.12 Doğrusal ayrılamayan bir sınıflandırma örneği

Şekil 3.13'te destek vektör makinesi yönteminin altındaki temel fikir anlatılmaktadır. Şeklin sol tarafındaki girdi uzayında örneklerin orijinal durumları yer almaktadır. Şeklin sağ tarafında ise örneklerin kernel adı verilen çeşitli matematiksel fonksiyonlar ile yeniden gösterimi yer almaktadır. Özellik uzayı verilen bu yeni gösterimde artık örnekler sınıflara doğrusal bir çizgi ile ayrılabilmektedir. Bu sayede girdi uzayındaki örnekleri karmaşık bir eğri ile sınıfları ayırmak yerine özellik uzayındaki gösterimleri ayıran doğruyu tanımlamak sınıflandırma için yeterli olmaktadır.



Şekil 3.13 Girdi uzayı ve özellik uzayı

Kernel fonksiyonları doğrusal ayrılamayan sınıfların ayrılması için kullanılır. Girdi nesnelerinin birbirleri ile benzerliğinin ölçümlenmesi olarak düşünülebilir. Fakat tüm benzerlik ölçüleri kernel fonksiyonu olarak kullanılamaz. Kernel fonksiyonlarının Mercer koşullarını sağlaması gerekmektedir. En sık kullanılan kernel fonksiyonları polinom, radyal tabanlı ve sigmoid fonksiyonlarıdır. (3.18), (3.19) ve (3.20) eşitliklerinde bu fonkiyonların matematiksel gösterimlerine yer verilmiştir.

Polinom kernel fonksiyonu:

$$K(x, y) = (x^T y + 1)^d \tag{3.18}$$

Radyal tabanlı kernel fonksiyonu:

$$K(x, y) = \exp\left(-\|x - y\|^2 / (2\sigma^2)\right) \tag{3.19}$$

Sigmoid kernel fonksiyonu:

$$K(x, y) = \tanh\left(\kappa x^T y + \theta\right) \tag{3.20}$$

Destek vektör makinesi yönteminin çok çeşitli sınıflandırma problemlerinde çok iyi genelleme başarısı gösterdiği deneysel çalışmalar ile kanıtlanmıştır. Ayrıca yöntemin metin sınıflandırmadaki başarısının nedenlerini teorik olarak açıklayan bazı kaynaklar da bulunmaktadır. (Joachims, 1998) Aşağıda destek vektör makinesi yönteminin metin sınıflandırmadaki başarısını açıklayan bazı argümanlar maddeler halinde belirtilmiştir.

- Metin sınıflandırma problemlerinde genellikle kelimeler özellik olarak kullanıldığı için çok fazla sayıda özellik bulunmaktadır. Destek vektör makineleri aşırı uyum prensibini kullandığı için büyük özellik uzaylarını ele almada çok başarılıdır.
- Her bir dokümanı temsil eden özellik vektöründe sadece birkaç özellik için özellik

değerinin sıfırdan farklı olduğu durumlar için destek vektör makinesi yöntemi çok uygundur.

- Çoğu metin sınıflandırma problemi doğrusal olarak ayrılabilmektedir. Destek vektör makinesi yönteminde de amaç sınıfları doğrusal ayıran karar sınırını bulmaktır.

Destek vektör makinesi yönteminin metin sınıflandırma problemlerine uygunluğu hem deneysel hem de teorik olarak ispatlanmıştır. Yöntemin diğer geleneksel sınıflandırma yöntemlerine göre bir diğer avantajı ise sağlamlığıdır. Destek vektör makinesi tüm deneylerde iyi bir performans göstermektedir, öyle ki diğer geleneksel sınıflandırma yöntemlerinde istisnai olsa da gözlenebilen büyük başarısızlıklar destek vektör makinesi yönteminde görülmemektedir. Ayrıca destek vektör makinesi yöntemi herhangi bir parametre iyileştirme işlemine gerek duyurmadan en ideal parametre değerlerini kendisi ayarlayabilmektedir. Tüm bu avantajlar destek vektör makinesi yöntemini metin sınıflandırma problemleri için kolay kullanılabilen ve etkili bir yöntem haline getirmiştir. Temelde bir metin sınıflandırması olan bu çalışmada da destek vektör makinesi yönteminin temel sınıflandırma yöntemi olarak kullanılmasına karar verilmiştir.

Destek vektör makinesi yöntemi uygulanırken SVM Light [2] kütüphanesinden faydalanılmıştır. SVM Light kütüphanesi destek vektör makinesi yönteminin uygulanması amacı ile C programlama dili ile hazırlanmış açık kaynak kodlu ve akademik çalışmalarda ücretsiz olarak kullanılabilen bir yazılımdır. SVM Light Vapnik' in örüntü tanıma, regresyon problemleri ve bir sıralama fonksiyonunun öğrenilmesi gibi amaçlar için kullanılan destek vektör makinesi yönteminin bir gerçekleştirimidir (Vapnik, 1995). SVM Light kütüphanesinde kullanılan optimizasyon algoritması (Joachims, 1999) ve (Joachims, 2002)'de ifade edilmiştir. Algoritmanın hafıza gereksinimleri ölçeklendirilebilirdir ve binlerce destek vektör makinesine ihtiyaç duyulan problemlerde bile kullanılabilmektedir. SVM Light yazılımı ayrıca performans genelleştirmesini elverişli olarak atayabilmeyi sağlayan yöntemleri de içermektedir. Hata oranı ve duyarlık - geri çağırma için iki elverişli tahmin yöntemi barındırmaktadır. XiAlpha tahminleri (Joachims, 2000, 2002) neredeyse hesaplama gerektirmeden gerçekleştirilebilmektedir fakat bu tahminler geçmiş önyargılıdırlar. Önyargısız tahminler birini dışarıda bırak testlerini desteklemektedir.

SVM Light kütüphanesini kullanabilmek amacı ile doküman bazında özelliklerin geçme sayıları SVM Light uygulamasının kullanımına uygun formata çevrilmiştir. Bu formatın deseni aşağıdaki gibidir.

<Sınıf etiketi> <özellik no1>:<özellik değeri1> <özellik no2>:<özellik değeri2>…

Bu formata uygun bir örnek aşağıdaki şekilde ifade edilebilmektedir. Örnekte 1 sınıfına ait verinin 4 farklı özellik için içerdiği değerler verilmiştir.

1 1:3 2:5 3:1 4:4

SVM Light uygulaması sınıflandırmanın optimizasyonu amacı ile çeşitli parametreler ile çağrılabilmektedir. Bu parametrelerden birisi de kullanılacak kernel fonksiyonunu belirleyen parametredir. SVM Light dördü önceden tanımlı ve bir tanesi kullanıcı tanımlı olmak üzere beş farklı kernel fonksiyonu seçeneği sunmaktadır. Kernel fonksiyonu parametresi aşağıdaki değerleri alabilmektedir.

- 0: Doğrusal kernel fonksiyonu.
- 1: Polinom kernel fonksiyonu.
- 2: Radyal tabanlı kernel fonksiyonu.
- 3: Sigmoid kernel fonksiyonu.
- 4: Kullanıcı tanımlı kernel fonksiyonu.

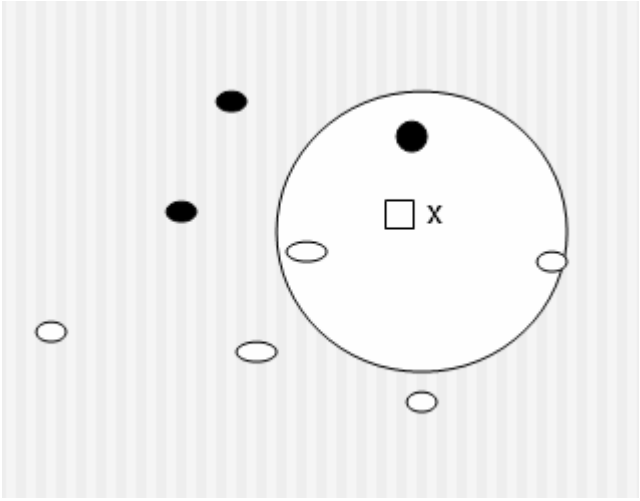
Uygun formata dönüştürülen eğitim verileri ile SVM Light öğrenme uygulaması kullanılarak model dokümanları oluşturulmuştur. Bu kapsamda kernel fonksiyonu parametresi kullanılarak doğrusal, polinom ve radyal tabanlı kernel fonksiyonları için farklı modeller oluşturulmuştur.

Test amaçlı kullanılacak dokümanlar okunarak özellik seçimi aşamasında elde edilen özelliklerin dokümanlarda kaç kere tekrar ettiği bilgisi elde edilmiştir. Bu bilgiler SVM Light uygulamasının işleyeceği formata çevrilmiştir. Eğitim aşamasında farklı kernel fonksiyonları için elde edilen model dokümanları kullanılarak uygun formata çevrilen test dokümanlarının sınıfları tahmin edilmiştir.

3.6.3 k-En Yakın Komşuluk

k-en yakın komşuluk en temel ve basit sınıflandırma yöntemlerinden birisidir. 1951 yılında Fix ve Hodges desen sınıflandırma amacı ile parametrik olmayan bir teknik geliştirdiler, daha sonraları bu teknik k-en yakın komşuluk kuralı olarak anılmaya başladı (Fix ve Hodges, 1951). Bu yöntemin temel aldığı prensip aynı sınıfa ait örneklerin özellik uzayında birbirine yakın oluşudur. Bu doğrultuda bilinmeyen bir sınıfa ait bir örnek için eğitim setindeki tüm örneklere olan uzaklıklar hesaplanır ve en yakın k adet örneğin çoğunluğunun hangi sınıf olduğuna bakılarak bilinmeyen örneğin sınıfı belirlenir. Örneğin Şekil 3.14'te 3-en yakın

komşuluk sınıflandırma örneğine yer verilmiştir. Şekilde bilinmeyen *x* örneğine en yakın üç örnek daire içine alınmıştır. Benzer üç örnekten çoğunluğu beyaz sınıfına ait olduğu için *x* örneği 3-en yakın komşuluk yöntemine göre beyaz olarak sınıflandırılmalıdır.



Şekil 3.14 3-en yakın komşuluk sınıflandırma örneği

Verinin doğasına göre uzaklık hesaplaması için farklı ölçümleme teknikleri kullanılabilmektedir. Örneğin, Öklid uzaklığı tipik olarak devamlı değişkenler için kullanılmaktadır. $\{a_1(x), a_2(x), .. , a_n(x)\}$ kümesi bir *x* örneğinin özellikleri olarak düşünülürse x_i ve x_j örnekleri arasındaki öklid uzaklığı (3.21) eşitliğinde belirtildiği gibi hesaplanmaktadır.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2} \tag{3.21}$$

k-en yakın komşuluk yöntemi sadeliğinden dolayı genellikle sınıflandırma problemlerinde diğer yöntemlerin sonuçları ile karşılaştırmak amacı ile kullanılmaktadır. Bu çalışmada da k-en yakın komşuluk yöntemi elde edilen sonuçların destek vektör makinesi yöntemi sonuçları ile karşılaştırılması amacı ile kullanılmıştır. k-en yakın komşuluk yöntemi ile sınıflandırma işlemini gerçekleştiren uygulama Java programlama dili kullanılarak geliştirilmiştir.

3.7 Doğrulama İşlemi

Tahmin etme sisteminin başarısının ölçülebilmesi amacı ile makalelerin eğitim ve test veri setleri olarak gruplandırılması gerekmektedir. Bu grupların belirlenmesi doğrulama testleri esnasında kullanılan k-fold çapraz doğrulama yöntemi ile yapılmıştır. k-fold çapraz doğrulama yönteminde tüm veri setinden k adet örnek test verisi olarak seçilir. Geriye kalan

diğer örnekler ise eğitim verisi olarak belirlenir. Daha sonra ilk aşamada seçilen k adet örnekten sonra gelen k adet örnek test verisi olarak belirlenir. Diğer örnekler ise eğitim verisi olarak seçilir. Bu şekilde farklı test aşamaları ile veri setindeki her bir örneğin bir kere test verisine dahil olması sağlanır.

Bu çalışmada, sistemin eğitim ve test veri setleri 10-fold çapraz doğrulama yöntemi kullanılarak iteratif olarak değişecek şekilde belirlenmiştir. İlk iterasyonda veri setinin ilk %10'luk kesimi test verisi olarak belirlenmiştir. Kalan %90'lık kısım da eğitim verisi olarak belirlenmiştir. İkinci iterasyonda ikinci %10'luk kesim test verisi olarak, kalan %90'lık kısım eğitim verisi olarak belirlenmiştir. Bu şekilde 10 iterasyon ile veri setindeki her dokümanın bir kere test verisine dahil olması sağlanmıştır.

3.7.1 Performans Değerlendirme Ölçüleri

Sınıflandırma performansı klasik bilgi kazanımı ölçüleri olan kesinlik ve geri çağırma oranlarının doküman sınıflandırmaya uyarlanması ile ölçülmektedir. (Sebastiani, 1999) Bir ikili sınıflandırma yönteminde sınıflayıcı, örnekleri pozitif veya negatif olarak tahmin etmektedir. Sınıflayıcının yaptığı tahminler hata matrisi veya olasılık tablosu adı verilen bir yapı ile gösterilebilmektedir. Hata matrisinde dört kategori yer almaktadır: Doğru pozitif (DP) sayısı doğru tahmin edilen pozitif örnek sayısıdır. Doğru negatif (DN) sayısı doğru tahmin edilen negatif örnek sayısıdır. Yanlış pozitif (YP) sayısı pozitif olarak tahmin edilen fakat gerçek sınıfı negatif olan örnek sayısıdır. Yanlış negatif (YN) sayısı ise negatif olarak tahmin edilen fakat gerçek sınıfı pozitif olan örnek sayısıdır. Bu kategoriler için değerler hesaplanarak elde edilen hata matrisinden kesinlik ve geri çağırma oranları hesaplanabilmektedir. (Davis ve Goadrich, 2006) Hata matrisi ya da diğer adı ile olasılık tablosunun yapısı Çizelge 3.3'de verilmiştir.

Çizelge 3.3 Olasılık tablosunun yapısı (Sebastiani, 1997)

		Tahmin Edilen Sınıf	
		Pozitif	**Negatif**
Gerçek Sınıf	**Pozitif**	DP	YN
	Negatif	YP	DN

Sınıflandırma sistemlerinin değerlendirilmesinde izlenen genel yaklaşımlardan birisi de doğruluk oranı ve doğruluk oranının tamamlayıcısı olan hata oranının hesaplanmasıdır.

Doğruluk oranı (3.22) eşitliğine, hata oranı ise (3.23) eşitliğine göre hesaplanır.

$$Doğruluk\ Oranı = \frac{DP + DN}{DP + YP + DN + YN} \tag{3.22}$$

$$Hata\ Oranı = \frac{YP + YN}{DP + YP + DN + YN} \tag{3.23}$$

Kesinlik oranı doğru tahmin edilen pozitif örnek sayısının toplam pozitif olarak tahmin edilen örnek sayısına bölümü ile elde edilir. Geri çağırma oranı ise doğru tahmin edilen örnek sayısının toplam pozitif örnek sayısına bölümü olarak hesaplanır. (3.24) ve (3.25) eşitliklerinde kesinlik oranı ve geri çağırma oranlarının hesaplanma yöntemleri verilmiştir.

$$Kesinlik\ Oranı = \frac{DP}{DP + YP} \tag{3.24}$$

$$Geri\ Çağırma\ Oranı = \frac{DP}{DP + YN} \tag{3.25}$$

Kesinlik oranı ve geri çağırma oranlarının ikisini de dikkate alan bir ölçü ise F-ölçeğidir. *KO* kesinlik oranı ve *GCO* geri çağırma oranı olmak üzere F-ölçeği (3.26) eşitliğine göre hesaplanır.

$$F\text{-}ölçeği = \frac{2 \cdot KO \cdot GCO}{KO + GCO} \tag{3.26}$$

4. UYGULAMA

Bu bölümde sistemin çalıştırılmasına ilişkin detaylar, doğrulama sonuçları ve sonuçların değerlendirmelerine yer verilmiştir. Sınıflandırma yöntemleri olarak kullanılan destek vektör makinesi ve k-en yakın komşuluk yöntemleri ile elde edilen sonuçlarının karşılaştırmaları yapılmıştır.

4.1 Özelllik Seçimi ve Doküman Gösterimi Sonuçları

Sınıflandırma işleminin yapılabilmesi için veri setindeki makalelerin özellik değerlerinin çıkarılması gerekmektedir. Önceki bölümlerde bahsedildiği üzere finansal makaleleri sınıflandırmak için kullanılacak özellikler bir isim ve bir fiilden oluşan kelime çiftinin makaledeki cümlelerde geçme sayısıdır. Burada özelliği simgeleyen isim ve fiil cinsinden kelimenin aynı cümlede geçmesi gerekmektedir. Bu kelimelerin cümlede yan yana geçmesi ya da birinin diğerinden önce geçmesi gibi bir şart aranmamaktadır.

İsim ve fiil kelime ikilileri özellik olarak belirlendikten sonra bu özelliklerden sınıflandırmada etkisi en fazla olanların seçilmesi gerekmektedir. Bu amaçla ki-kare özellik seçimi yöntemi kullanılmıştır. Özellik seçimi işlemi sonucunda kaç adet özelliğin seçileceği bir sistem parametresi olarak belirlenmiştir. Bu kapsamda değişik özellik sayıları ile sistemin performansı ölçümlenmiştir. Kullanılan sınıflandırma yöntemine göre en iyi performansın elde edildiği özellik sayısı farklılık göstermektedir. Özellik sayısı parametresinin düşük tutulduğu durumlarda seçilen özelliklerin sınıflandırmada yetersiz kaldığı gözlenmiştir. Özellik sayısı parametresi çok büyük olarak belirlendiğinde bazı özelliklerin sınıflandırma başarısını artırmak yerine yanlış sınıflandırmaya sebep olarak başarıyı azalttığı gözlenmiştir. Bu bakımdan ideal özellik sayısı parametresinin belirlenmesi sistemin gerçek performansının ortaya çıkarılması açısından önem arzetmektedir. Özellik sayısı parametresinin 250 adet olarak belirlendiğinde durumda seçilen bazı özellikler Çizelge 4.1'de listelenmiştir.

Çizelge 4.1 Özellik olarak belirlenen bazı kelime ikilileri

1. Kelime (İsim)	**2. Kelime (Fiil)**	**1. Kelime (İsim)**	**2. Kelime (Fiil)**
years	start	yahoo	struggling
buy	said	service	see
days	doing	stock	beat
people	believe	companies	nyse
year	acquired	had	share
month	hit	stocks	seeing
company	offers	gates	got
packard	buy	shares	bought
game	pick	fact	make
instance	have	company	turns
point	ask	cause	have
share	sell	profit	turn
service	believe	investor	based
gains	make	today	called
industry	buy	sell	had
shareholders	have	diversification	make
right	take	time	build
tomorrow	start	day	lost
players	make	period	held
company	own	service	join
months	hit	performance	check
right	have	year	started
netflix	have	point	get
income	held	call	believe
google	consider	cash	expect
yahoo	going	shareholders	started
cloud	have	shares	being
declines	have	stores	want
companies	found	game	play
ceo	including	everyone	seems
point	does	email	enter
market	gotten	articles	make
market	sharing	time	looking
ways	doing	market	built
interest	know	time	enjoyed
article	take	markets	fall
time	used	companies	turns
chief	operating	holdings	see
investors	expect	service	want
buys	clicks	holdings	buying
nothing	buy	world	looking
games	make	revenue	rose
technology	leading	quality	hidden
companies	changing	inside	pick
company	build	hand	sell

Çizelge 4.1 incelendiğinde özellik olarak belirlenen bazı kelime çiftlerinin makalenin olumlu veya olumsuz olduğu konusunda ilk bakışta hemen bir fikir verdiği kolaylıkla gözlenebilmektedir. Örneğin "gains" ve "make" ikilisi veya "revenue" ve "rose" ikilisi makalenin içeriğinin olumlu olduğu fikrini vermektedir. "markets" ve "fall" veya "declines" ve "have" ikililisi ise makalenin içeriğinin olumsuz olduğu fikrini vermektedir.

Bir özellik olarak fiil ve isimden oluşan kelime çiftinin makaledeki herhangi bir cümlede geçip geçmemesi sınıflandırma için önemli bir bilgi olmaktadır. Fakat aynı kelime çiftinin makaledeki birden fazla cümlede geçmesi makalenin ilgili özellikte ağırlığının daha fazla olduğu anlamına gelmektedir. Bu nedenle bir özellik için ilgili özelliği taşıyan kelime çiftinin bir makalede aynı cümle içerisinde beraber geçme sayısı özellik değeri olarak belirlenmektedir. Bu şekilde tüm dokümanlar her bir üyesinde belirli bir kelime çiftinin dokümanda geçme frekansını taşıyan vektörler olarak gösterilmiştir.

4.2 Sınıflandırma Sonuçları

Bu bölümde sistemin değişik parametreler ışığında genel performans değerlendirmesi ve doküman bazında sınıflandırma sonuçları ayrı ayrı ele alınmıştır.

4.2.1 Sistemin Genel Performansının Değerlendirilmesi

Sistemin doğrulama testleri destek vektör makinesi ve k-en yakın komşuluk yöntemleri ayrı ayrı kullanılarak gerçekleştirilmiştir. Bu kapsamda her iki yöntem kullanılırken de farklı özellik sayısı parametreleri ile sonuçlar elde edilmiştir. Ayrıca isim-fiil kelime çiftlerinden oluşan özellik kullanımına ek olarak tek tek kelimelerin özellik olarak kullanıldığında nasıl sonuçlar elde edileceği de hesaplanmıştır.

4.2.1.1 Destek Vektör Makinesi Sınıflandırma Sonuçları

Destek vektör makinesi ve isim-fiil kelime çifti özellik yöntemlerinin farklı özellik sayısı parametreleri ile doğrulaması sonucunda en iyi değerler 250 özellik ile elde edilmiştir. Ayrıca tüm farklı özellik sayısı parametreleri için radyal tabanlı kernel fonksiyonu diğer kernel fonksiyonlarına göre daha başarılı sonuçlar üretmiştir. Çizelge 4.2'de 250 özellik ve radyal tabanlı kernel fonksiyonu ile elde edilen sonuçlara ait hata matrisine yer verilmiştir.

Çizelge 4.2 Destek vektör makinesi ve isim-fiil kelime çifti özellik yöntemleri ile elde edilen en iyi sonuçların hata matrisi

		Tahminlenen Sınıf		
		Artış	Azalış	
Gerçek Sınıf	Artış	DP=90	YN=14	104
	Azalış	YP=57	DN=21	78
		147	35	Toplam=182

Çizelge 4.2'de yer alan sonuçlar incelendiğinde sistemin artış olarak tahmin ettiği ve gerçek sınıfı artış olan doküman sayısının 90 olduğu görülmektedir. Azalış olarak tahmin edilen fakat gerçek sınıfı artış olan doküman sayısı ise 14'tür. 57 adet doküman ise gerçekte azalış sınıfına dahil olduğu halde artış olarak sınıflama tahmini yapılmıştır. 21 azalış sınıfına ait doküman ise doğru olarak tahmin edilmiştir. Bu sonuçlara göre sistemin doğruluk oranı, doğru pozitif oranı (geri çağırma oranı), doğru negatif oranı ve kesinlik oranı Çizelge 4.3'de verilmiştir.

Çizelge 4.3 Destek vektör makinesi ve isim-fiil kelime çifti özellik yöntemleri ile elde edilen performans ölçüleri

Performans Ölçüsü	**Değer**
Doğruluk Oranı	%61
Doğru Pozitif Oranı	%87
Doğru Negatif Oranı	%27
Kesinlik Oranı	%61

Tek tek kelimeler özellik olarak kullanıldığında polinom kernel fonksiyonun diğer kernel fonksiyonlarına göre daha başarılı olduğu görülmüştür. Ayrıca bu yöntemde en iyi sonuçlar özellik sayısı parametresi 5000 olarak belirlendiğinde elde edilmiştir. İsim-fiil kelime çiftleri özellik olarak kullanıldığında en iyi sonuçlar 250 özellik sayısı ile elde edilmişti. Bu bakımdan isim-fiilden oluşan özelliklerin tek kelimeden oluşan özelliklere göre özellik başına daha çok sınıflandırma bilgisi taşıdığı sonucu çıkarılabilir. Çizelge 4.4'de özellik sayısı 5000 ve polinom kernel kullanımı ile elde edilen sonuçların hata matrisine yer verilmiştir.

Çizelge 4.4 Destek vektör makinesi ve tek tek kelimelerin özellik olarak kullanımı ile elde edilen en iyi sonuçların hata matrisi

		Tahminlenen Sınıf		
		Artış	Azalış	
Gerçek Sınıf	Artış	DP=92	YN=12	104
	Azalış	YP=63	DN=15	78
		155	27	Toplam=182

Çizelge 4.4'te yer alan sonuçlardan yola çıkılarak tek tek kelimelerin özellik olarak kullanımı yönteminin başarı oranı %59 olarak hesaplanmıştır. Ayrıca doğru pozitif oranı (geri çağırma oranı), doğru negatif oranı ve kesinlik oranı Çizelge 4.5'te verilmiştir.

Çizelge 4.5 Destek vektör makinesi ve tek tek kelimelerin özellik olarak kullanımı ile elde edilen performans ölçüleri

Performans Ölçüsü	**Değer**
Doğruluk Oranı	%59
Doğru Pozitif Oranı	%88
Doğru Negatif Oranı	%19
Kesinlik Oranı	%59

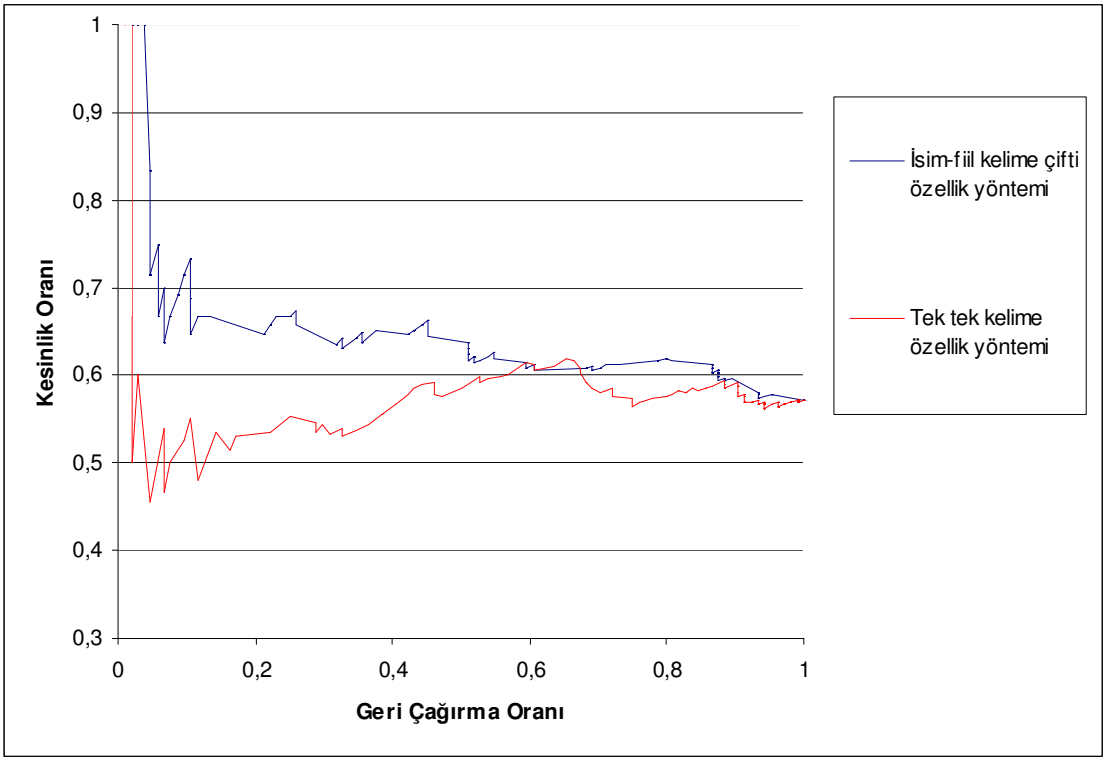
Kesinlik ve geri çağırma oranları tek tek ele alındığında bazen yanıltıcı yorumlara neden olabilmektedir. Çünkü bu iki oran arasında bir denge bulunmaktadır. Kesinlik oranı arttığında genellikle geri çağırma oranı azalmaktadır. Aynı şekilde geri çağırma oranı arttıkça kesinlik oranı çoğunlukla azalmaktadır. F-ölçeği kesinlik ve geri çağırma oranlarının uyumlu bir ortalaması olarak her iki değeri de hesaba katmaktadır. Çizelge 4.6'da destek vektör makinesi yöntemi sınıflandırma amacı ile kullanıldığında her iki özellik kullanım yöntemine ait F-ölçeği değerlerine yer verilmiştir.

Çizelge 4.6 Destek vektör makinesi yöntemi sonucunda elde edilen F-ölçeği değerleri

Özellik Kullanım Yöntemi	**F-ölçeği Değeri**
İsim-fiil özellik yöntemi	%72
Tek tek kelime özellik yöntemi	%71

Kesinlik oranı ve geri çağırma oranı arasındaki ilişkiyi analiz etmenin bir diğer yolu ise

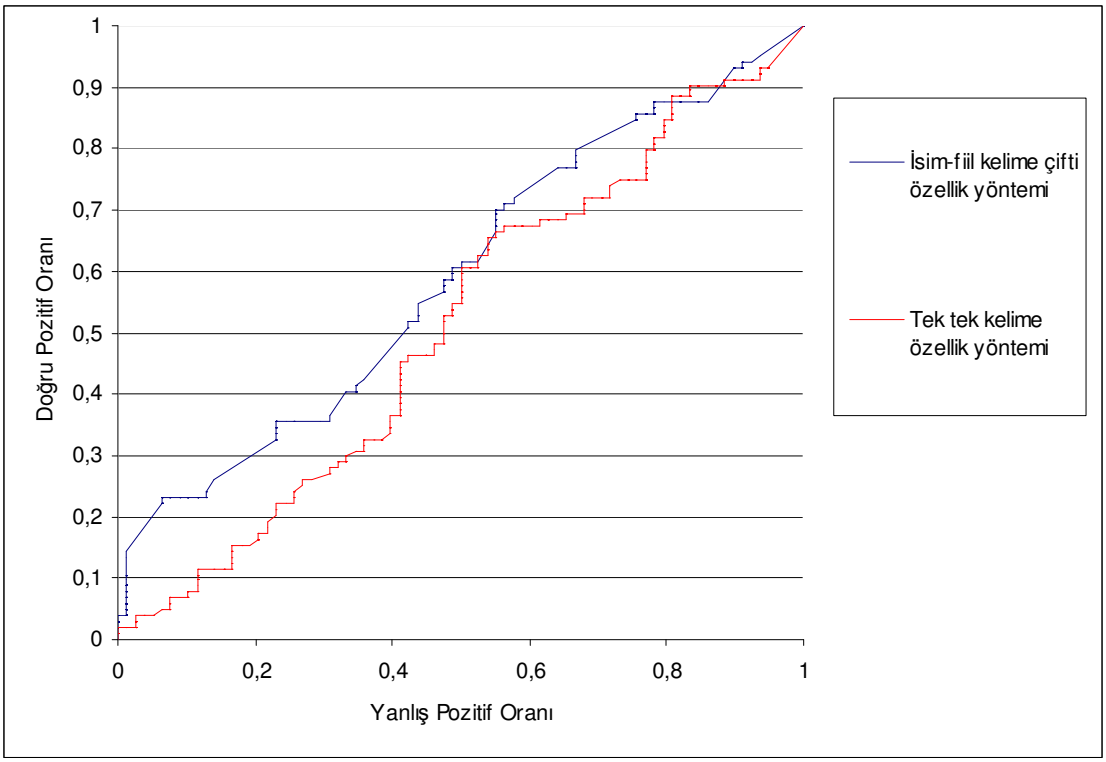
kesinlik-geri çağırma eğrisidir. Bu eğride geri çağırma oranının değişimine göre kesinlik oranının değerlerindeki değişime yer verilir. Yatay eksende geri çağırma oranı, düşey eksende ise kesinlik oranı değerleri yer almaktadır. Kesinlik-geri çağırma eğrilerinde en iyi performans hem kesinlik oranı hem de geri çağırma oranının dengeli şekilde yüksek olduğu eğrinin sağ üst köşeye en yakın olduğu noktadır. Şekil 4.1'de sınıflandırma yöntemi olarak destek vektör makinesi kullanıldığında elde edilen kesinlik-geri çağırma eğrilerine yer verilmiştir. Grafikte isim-fiil kelime çifti ve tek tek kelime özellik kullanımı teknikleri için iki farklı eğri gösterilmektedir. Mavi renkli eğri isim-fiil kelime çifti özellik yöntemi ile elde edilen sonuçları, kırmızı renkli eğri tek tek kelime özellik yöntemi ile elde edilen sonuçları göstermektedir. Mavi renkli eğrinin kırmızı renkli eğriye göre grafiğin sağ üst köşesine daha yakın olduğu gözlenmektedir. Grafikten yola çıkarak isim-fiil özellik yönteminin tek tek kelime özellik yöntemine göre daha iyi bir performans gösterdiği yorumu yapılabilmektedir.



Şekil 4.1 Kesinlik-geri çağırma eğrisi

Şekil 4.2'de her iki özellik kullanım yöntemine ait sonuçların ROC eğrisine yer verilmiştir. ROC eğrisi artış ve azalış sınıfları arasındaki eşik değerinin değişimine göre yanlış pozitif oranı ve doğru pozitif oranının ilişkisini gösteren eğridir. Grafikte yatay eksende yanlış pozitif oranı, düşey eksende ise doğru pozitif oranı değerleri yer almaktadır. ROC eğrisinde ideal durum doğru pozitif oranı artırılarak ve yanlış pozitif oranı azaltılarak en uygun

kombinasyonun elde edildiği durumdur. İdeal durum eğrinin grafiğin sol üst köşesine en yakın noktasıdır.



Şekil 4.2 ROC eğrisi

Şekil 4.2'deki grafikte mavi renkle gösterilen eğri isim-fiil kelime çifti yönteminde elde edilen sonuçları, kırmızı renkli eğri tek tek kelimelerin özellik olarak kullanıldığında elde edilen sonuçları göstermektedir. Grafik incelendiğinde aynı yanlış pozitif oranı değerinde isim-fiil kelime çifti yönteminin daha yüksek doğru pozitif oranı değerlerine sahip olduğu farkedilmektedir. Bu bakımdan isim-fiil kelime çifti yönteminin tek tek kelimelerin özellik olarak kullanımı yöntemine göre daha başarılı bir performans gösterdiği sonucuna varılmıştır.

4.2.1.2 k-En Yakın Komşuluk Yöntemi ile Sınıflandırma

k-en yakın komşuluk yöntemi kullanılarak gerçekleştirilen doğrulama testlerinde örneklerin birbirlerine olan uzaklıkları ölçülürken Öklid uzaklığı ölçümlemesi kullanılmıştır. k-en yakın komşuluk yöntemi ve isim-kelime çifti özellik yönteminin farklı özellik sayısı ve farklı k komşuluk sayısı parametreleri ile doğrulaması gerçekleştirilmiş ve en iyi sonuçlar özellik sayısı 100 ve k komşuluk sayısı 5 olarak alındığında elde edilmiştir. Bu parametreler kullanıldığında k-en yakın komşuluk ve isim-fiil özellik yöntemi için sınıflandırma sonuçları

hata matrisi Çizelge 4.7'de yer almaktadır.

Çizelge 4.7 k-en yakın komşuluk ve isim-fiil özellik yöntemi ile sınıflandırma sonuçları hata matrisi

		Tahmin Edilen Sınıf		
		Artış	Azalış	
Gerçek Sınıf	Artış	DP=96	YN=8	104
	Azalış	YP=69	DN=9	78
		165	17	Toplam=182

k-en yakın komşuluk yöntemine göre sistemin doğruluk oranı, doğru pozitif oranı (geri çağırma oranı), doğru negatif oranı ve kesinlik oranı Çizelge 4.8'de verilmiştir.

Çizelge 4.8 k-en yakın komşuluk ve isim-fiil özellik yöntemi ile sınıflandırma performans ölçüleri

Performans Ölçüsü	**Değer**
Doğruluk Oranı	%58
Doğru Pozitif Oranı	%92
Doğru Negatif Oranı	%12
Kesinlik Oranı	%58

Tek tek kelimeler özellik olarak kullanıldığında en iyi sonuçlar özellik sayısı 250 ve k komşuluk sayısı 7 olarak belirlendiğinde elde edilmiştir. Bu parametreler kullanıldığında elde edilen sonuçlar Çizelge 4.9'da verilmiştir.

Çizelge 4.9 k-en yakın komşuluk ve tek tek kelime özellik yöntemi ile sınıflandırma sonuçları hata matrisi

		Tahmin Edilen Sınıf		
		Artış	Azalış	
Gerçek Sınıf	Artış	DP=87	YN=17	104
	Azalış	YP=63	DN=15	78
		150	32	Toplam=182

k-en yakın komşuluk sınıflandırmasında tek tek kelimelerin özellik olarak kullanımı yönteminin başarı oranı %56 olarak hesaplanmıştır. Doğru pozitif oranı (geri çağırma oranı), doğru negatif oranı ve kesinlik oranı Çizelge 4.10'da verilmiştir. Bu sonuçlara göre k-en

yakın komşuluk yöntemi isim-fiil kelime çifti özellik yöntemi ile beraber kullanıldığında tek tek kelime özellik yöntemine göre daha başarılı sonuçlar elde edilmektedir.

Çizelge 4.10 k-en yakın komşuluk ve tek tek kelime özellik yöntemi ile sınıflandırma performans sonuçları

Performans Ölçüsü	**Değer**
Doğruluk Oranı	%56
Doğru Pozitif Oranı	%84
Doğru Negatif Oranı	%19
Kesinlik Oranı	%58

Çizelge 4.11'de k-en yakın komşuluk yöntemi sınıflandırma amacı ile kullanıldığında her iki özellik kullanım yöntemine ait F-ölçeği değerlerine yer verilmiştir.

Çizelge 4.11 k-en yakın komşuluk yöntemi sonucunda elde edilen F-ölçeği değerleri

Özellik Kullanım Yöntemi	**F-ölçeği Değeri**
İsim-fiil özellik yöntemi	%71
Tek tek kelime özellik yöntemi	%69

4.2.1.3 Destek Vektör Makinesi ve k-En Yakın Komşuluk Yöntemleri Sonuçlarının Karşılaştırılması

Destek vektör makinesi ve k-en yakın komşuluk yöntemleri ile gerçekleştirilen sınıflandırma işlemlerinin sonuçları incelendiğinde beklendiği üzere destek vektör makinesi yönteminin daha başarılı sonuçlar ürettiği gözlenmiştir. Özellikle k-en yakın komşuluk yönteminin özellik sayısı arttıkça başarısının düştüğü tespit edilmiştir. Bu yöntemin dezavantajlarından birisi de çok boyutlu veriler ile çalışıldığında performansının düşük olması idi. Bu çalışmada temelde bir metin sınıflandırma işlemi olduğu için boyut sayısı fazla olmaktadır. Bu nedenle k en yakın komşuluk yönteminin bu problemin karakteristiğine uygun olmadığı yorumu getirilebilir. Bunun aksine destek vektör makinesi yöntemi aşırı uyum prensibi ile çalıştığı için çok boyutlu veriler üzerinde çok başarılı sonuçlar üretebilmektedir. Bu durumu destekleyen bir diğer sonuç ise k-en yakın komşuluk yönteminde en iyi performansın 100 özellik ile edilmesine karşın destek vektör makinesi yönteminde en iyi performansın 250 özellik ile elde edilmesidir.

k-en yakın komşuluk yönteminin bir diğer özelliği ise eğitim setindeki örnek sayısının çok fazla olduğu durumlarda yüksek başarı sağlamasıdır. Bu çalışmada her bir dokümanın daha çok özellik için değer taşıyabilmesini sağlamak adına 949 haber makalesi aynı gündeki makaleler bir dokümanda birleştirilerek 182 doküman haline getirilmiştir. Bu şekilde 949 haber makalesi 182 dokümana indirgenerek destek vektör makinesi yönteminin daha başarılı olması sağlanmakla birlikte doküman sayısının azalmasından ötürü k-en yakın komşuluk yönteminin başarısının azalmasına sebebiyet verilmiştir.

4.2.2 Doküman Bazında Sınıflandırma Sonuçları

Doğrulama işlemine ait genel sonuçlardan sonra doküman bazında sonuçlar incelenerek çalışmanın genel sonucunu etkileyen detaylar araştırılmıştır. Bu kapsamda 2 Haziran 2009 tarihinde Microsoft şirketi hakkında yayınlanan haber makalelerini içeren doküman incelenmiştir. Bu tarihin seçilme nedeni yayınlanan haberlerin Microsoft şirketi hakkında belirgin olarak olumlu nitelikte olması ve sınıflandırmada önem taşıyan net cümleler içermesidir. Şekil 4.3'te bu dokümanda yer alan bazı önemli cümleler verilmiştir. Sınıflandırmada etkisinin büyük olacağı düşünülen bazı kelimeler kalın font ile gösterilmiştir. Bu tarihte Microsoft şirketi hisse senedi fiyatının haberlere uygun şekilde %3.26 oranında yükseldiği tespit edilmiştir. Çalışmanın da ilgili doküman için ürettiği tahmin sonucunun olumlu olduğu gözlenmiştir.

```
.....
...However, IT cuts are demanding tech companies to come up with
innovative solutions to keep revenue growing. Microsoft (Nasdaq:
MSFT) recently extended an engineering and sales partnership with
EMC (NYSE: EMC), the parent company of VMWare (NYSE: VMW),
despite the fact they compete head to head in the growing
virtualization field. The rationale being that the agreement can
save their clients money and keep money flowing in despite budget
cutbacks.....
.....
...Revenue was up more than 25% year-over-year, return on equity
was a robust 49%, and it was considered one of the world's most
dominant companies. Its name? Microsoft (Nasdaq:
MSFT).....
...James Early does not own any of the stocks discussed in this
article. Microsoft and Legg Mason are Motley Fool Inside Value
recommendations. The Motley Fool owns shares of Legg Mason. The
Fool has a disclosure policy with regard to such
interests.....
```

Şekil 4.3 2 Haziran 2009 tarihli dokümanda yer alan bazı cümleler

Doküman bazında sonuçlar daha fazla incelendiğinde sınıflandırma sonucunu olumsuz etkileyen bazı durumlar da tespit edilmiştir. Örneğin bazı günlerde yayınlanan haberler olumlu içeriğe sahip olmasına rağmen bazı başka faktörlerden ötürü hisse senedi fiyatında düşüş yaşanabilmektedir. Ya da tam tersi şekilde haberler olumsuz olmasına rağmen hisse senedi fiyatı artabilmektedir. Bu tür durumlarda sistem aslında haber içeriğine göre başarılı bir sınıflama yapsa bile hisse senedi fiyatındaki değişim tam tersini gösterdiği için genel sınıflandırma istatistiklerine başarısız bir sınıflama örneği olarak yansımaktadır. Test veri setinde yer aldığında yanlış sınıflamalara sebebiyet veren bu tür dokümanlar eğitim veri setinde yer aldıklarında da sistemin yanlış eğitilmesine neden olmaktadır. 24 Nisan 2009 tarihinde yayınlanan haberler için de böyle bir durum söz konusudur. Şekil 4.4'te de görülebileceği üzere bu tarihte yayınlanan haberler genellikle Microsoft şirketi hakkında olumsuz gelişmeler barındırmaktadır. Fakat bu tarihte Microsoft şirketinin hisse senedi fiyatı %10.52 nispetinde artış göstermiştir. Bu durumun sebebi diğer bir gelişmenin olumsuz haberlerin önüne geçerek yatırımcıların hisse senedine ilgisinin artması olabilmektedir. Bu önemli gelişmenin sistemin veritabanındaki haberlerde yer almaması gelişmenin Microsoft şirketi ile doğrudan ilgili olmamasından kaynaklanabilmektedir. Sistemin gerçekleştirdiği sınıflandırma çalışması sonucunda bu doküman olumsuz olarak sınıflandırılmıştır. Bu nedenle sistemin gelen performansına yanlış bir sınıflama olarak yansımıştır.

```
.....................................................................
...Microsoft is still posting losses in its online
business.....................................................
.....................................................................
...The week in tech  nor are many of the tech titans that reported
earnings this week. Microsoft  (Nasdaq: MSFT) copped to the first
year-over-year profit decline in company history last
night.....................................................
.....................................................................
...Instead, investors cheered when Microsoft's $13.65 billion in
revenue was $500 million less than the consensus
estimate.....................................................
.....................................................................
...Microsoft fits nicely into all four categories. With shares down
40% over the past year, we're now looking at a company with a
stranglehold on its key products, trading at less than 10 times
forward earnings ............................................
```

Şekil 4.4 24 Nisan 2009 tarihli dokümanda yer alan bazı cümleler

Bazı günlerde yayınlanan haberlerde yer alan gelişmelerin belirgin olarak olumlu veya olumsuz olarak sınıflandırılabilecek nitelikte olmadığı görülmüştür. Diğer bir deyişle

yayınlanan haberlerin yatırımcıların aksiyon almalarında herhangi bir etkisi bulunmamaktadır. Böyle günlerde yatırımcıların yatırımlarına yön verirken belirli bir şirkete ait olmayan genel gelişmeleri, rakip şirketler hakkında yayınlanan haberleri veya daha önceki günlerde Microsoft şirketi hakkında yayınlanan haberleri dikkate aldıkları düşünülmüştür. Bu bakımdan bu günlere ait dokümanların sistem tarafından doğru olarak sınıflandırılabilmesine olanak sağlayan bilginin veritabanında yer almadığı kararına varılmıştır.

Doküman bazında sonuçlar incelendikten sonra sistemin performansının hisse senedi fiyatlarının etkilendiği gelişmelerin çeşitliliği ve bu gelişmelerin ne kadarının sistemin veritabanında yer aldığı ile yakından ilişkili olduğu sonucuna varılmıştır.

5. SONUÇ

Bu bölümde tüm çalışma kısaca tekrar gözden geçirilmiş ve sonuç olarak önemli saptamalara değinilmiştir. Çalışmanın gerçekleştirilmesi boyunca karşılaşılan kısıtlamalar ve bu kısıtlamaların çalışmanın başarısına olan etkisinden bahsedilmiştir. Ayrıca çalışmanın devamı niteliğinde ileriye dönük yapılabilecek iyileştirmelere yer verilmiştir.

5.1 Çalışmanın Değerlendirilmesi

Hisse senetleri fiyatlarının tahmin edilmesine yönelik şimdiye dek birçok çalışma gerçekleştirilmiştir. Metin formatındaki dokümanların analiz edilerek ileriye dönük tahminler yapılması veri madenciliği alanında gittikçe önem kazanan bir konu olmaktadır. Hisse senetleri fiyatlarının tahmin edilmesinde metin formatındaki haber makalelerin kullanılması da oldukça sık karşılaşılan bir yöntemdir. Çünkü sadece sayısal bilgiden oluşan tarihsel hisse senedi fiyatları bilgileri sadece olayların sonuçları hakkında fikir vermektedir. Sayısal veri bu anlamda olayların sebepleri hakkında fikir barındırmamaktadır. Bunun aksine haber makaleleri gibi metin formatındaki veriler içerdikleri bilgi bakımından daha zengindir. Çoğu durumda olayların sebepleri ve gelişimleri hakkında bilgiler barındırmaktadır. Olayların sebeplerini barındıran metin formatındaki veriler ile birlikte olayların sonuçlarını simgeleyen sayısal bilgilerin bir arada kullanılması tahmin etme sistemlerinin başarısını artırmaktadır.

Hisse senetleri yatırımcıları yatırımlarını yönlendirirken yüksek getiri ve düşük riske sahip hisse senetlerini tercih etmektedirler. Ayrıca bir hisse senedi yatırımcısının yatırımlarından kazanç elde edebilmesi için elinde bulundurduğu hisse senedini doğru zamanda satabilmesi ve yatırım yapmayı planladığı hisse senedini doğru zamanda alması gerekmektedir. Yatırımcıların yatırım stratejilerini belirlerken en çok kullandığı bilgi kaynağı çeşitli kaynaklar tarafından yayınlanan finansal haber makaleleridir. Fakat birçok değişik haber kaynağı tarafından hisse senetleri fiyatlarını etkileyebilecek birçok haber yayınlanmaktadır. Ayrıca bu haberlerin hisse senetleri fiyatı üzerindeki etkisi çok kısa süre içerisinde gerçekleşmektedir. Yatırımcıların çok kısa süre içerisinde tüm bu haberleri okuyup analiz etmesi ve yatırım stratejilerini buna göre belirlemesine imkan bulunmamaktadır.

Bu çalışmada haber makalelerinin çok kısa bir sürede analiz edilerek hisse senetleri fiyatlarında kısa dönemde gerçekleşebilecek değişikliklerin tahmin edilmesi amaçlanmıştır. Bu amaçla yeni yayınlanan haberlerin hisse senedi fiyatı üzerindeki artış veya azalış etkisini belirlemek üzere bir tahmin etme modeli geliştirilmiştir. Bu şekilde çalışma temelde bir ikili

sınıflandırma problemine dönüştürülmüştür. Geliştirilen sistemin iki temel girdisi bulunmaktadır. Bu girdilerden ilki yapısal formda ve sayısal nitelikte geçmişe dönük hisse senedi fiyatları bilgileridir. Sistemin diğer girdisi ise metin formunda yer alan haber makaleleridir. Çeşitli veri madenciliği teknikleri ile hisse senetleri fiyatları ve haber makalelerinin bazı özellikleri arasında ilişkiler kurularak sistemin eğitimi gerçekleştirilmiştir.

Tahmin etme modelinin oluşturulması esnasında veri setinin elde edilmesi, verinin ön işlemeden geçirilmesi, özellik seçimi, sınıflandırma ve değerlendirme aşamaları takip edilmiştir. Aşamaların gerçekleştiriminde Java programlama dili ve SVM Light kütüphanesi kullanılmıştır. Microsoft firmasının hisse senedi fiyatları ve bu firma ile ilgili yayınlanmış haber makaleleri sistemin veri seti olarak kullanılmıştır. Ön işleme aşamasında yeni bir yaklaşım kullanılarak aynı cümle içerisinde geçen isim ve fiil kelime çiftleri özellik olarak belirlenmiştir. Özellik seçimi aşamasında ki-kare istatistikleri yöntemi kullanılarak sınıflandırmada en ayırt edici özellikler seçilmiştir. SVM ve k en yakın komşuluk sınıflandırma yöntemleri kullanılarak sistemin eğitim ve testi gerçekleştirilmiştir.

Çalışma sonucunda %61'lik bir başarı oranı elde edilmiştir. Rasgele yapılacak yatırım stratejilerinin %50'lik bir başarıya sahip olacağı düşünülürse gerçekleştirilen sistemin kayda değer bir başarı elde ettiği yorumu getirilebilmektedir. Ayrıca bu başarı oranı hisse senetleri fiyatları ve haber makaleleri arasında mutlak bir ilişki olduğunu da kanıtlamaktadır. %61 başarı oranı bir ikili sınıflandırma çalışması için düşük gibi gözükse de hisse senetleri fiyatlarının bir çok değişkenden etkilendiği düşünüldüğünde önemli bir orandır. Örneğin hisse senedine ait makalelerin olumlu olmasına rağmen hisse senedi fiyatı global krizden etkilenerek düşüş gösterebilmektedir. Bu durumda aslında sistem doğru tahmin etmesine rağmen bu doküman için sistem yanlış sınıflandırma yapmış gibi görünmektedir.

5.2 Çalışma Süresince Karşılaşılan Kısıtlar

Çalışma süresinde karşılaşılan en büyük sıkıntı ideal bir veri setinin elde edilememesidir. Bu konu ile ilgili akademik çalışmalarda kullanılmak üzere ortak bir veri tabanı yer almamaktadır. Bu nedenle araştırmacıların kendi veri tabanlarını hazırlamaları gerekmektedir. Her çalışmanın kendine özgün veri tabanı kullanması farklı çalışmaların sonuçlarının sağlıklı olarak birbiri ile karşılaştırılabilmesine engel teşkil etmektedir. Belirli bir şirkete ait haberlerin internet üzerindeki haber kaynaklarından elde edilmesindeki sıkıntı ise aynı sektörde yer alan başka şirketlerin haberlerinin de ilgili şirket kategorisi altında yayınlanmasıdır. Hisse senedi fiyatları kullanılan şirkete ait olmayan haberlerin de

sınıflandırmada kullanılması performansı olumsuz yönde etkilemektedir.

5.3 İleriye Yönelik İyileştirmeler

Mevcut durumda sistem veri seti olarak sadece bir haber kaynağından faydalanmaktadır. Veri seti hazırlanırken birden çok farklı haber kaynağından faydalanılabilir. Bu şekilde eğitim setinin farklı haber kaynakları ile zenginleştirilmesi sağlanacaktır. Daha farklı kaynaklardan ve daha çok haber makalesinin hisse senetleri fiyatlarına etki edecek daha çok bilgiyi barındırması sebebi ile sistem başarısı da artacaktır.

Çalışmanın gerçekleştirilmesinde karşılaşılan zaman kısıtından ötürü sistem çevrimdışı olarak çalışacak şekilde tasarlanmıştır. Sistem online olarak internet ile tam entegre çalışacak şekilde geliştirilebilir. Bu şekilde sistemin online haber kaynaklarını otomatik olarak tarayarak anlık yatırım tavsiyeleri gerçekleştirmesi sağlanabilir. Ayrıca tahmin etmede kullanılan yeni haberler ilgili güne ait hisse senedi fiyatı açıklandıktan sonra eğitim veri setine dahil edilerek sistemin eğitim veri setinin otomatik olarak zenginleştirilmesi sağlanabilir.

Bir şirketin hisse senedi fiyatı şirkete ait haberlerle birlikte global ekonomik ve siyasi gelişmelerden de etkilenmektedir. Bu çalışmada haber kaynaklarının şirkete ait haber kategorisinde yayınlanan haberler kullanılmıştır. Fakat bazı durumlarda global bir ekonomik gelişme şirket haberlerinin önüne geçerek hisse senedi fiyatı değişikliğindeki esas rolü üstlenmektedir. Global ekonomik ve siyasi gelişmelere ilişkin haberlerin de veri setine ayrı bir kategori altında eklenerek hisse senedi fiyatlarındaki değişikliklerin daha doğru tahmin edilebilmesi gerçekleştirilebilir.

KAYNAKLAR

Apte, C., Damerau, R. ve Weiss, S.M., (1994), "Automated Learning of Decision Rules for Text Categorization", ACM Transactions on Information Systems, 12(3):233-251.

Brucher, H., Knolmayer, G. ve Mittermayer, M.A., (2002), "Document Classification Methods for Organizing Explicit Knowledge", In Proceedings of the 3rd European Conference on Organizational Knowledge, Learning and Capabilities (OKCL), 05-06 Apr 2002, Athens/Greece.

Buckley, C., Salton, G. ve Allan, J., (1994), "The Effect of Adding Relevance Information in a Relevance Feedback Environment", In W.B. Croft and C.J. van Rijsbergen eds. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Developlment in Information Retrieval, 03-06 Jul 1994, Dublin/Ireland.

Chen, H., Hsu, P., Orwig, R., Hoopes, L. Ve Nunamaker, J.F., (1994), "Automatic Concept Classification of Text from Electronic Meetings", Communications of ACM, 37(10):56-73.

Cohen, W.W. ve Singer, Y., (1996), "Context-Sensitive Learning Methods for Text Categorization", ACM Transactions on Information Systems (TOIS), 17(2):141-173.

Cortes, C., Vapnik, V., (1995), "Support Vector Networks", Machine Learning, 20(3):273-297.

Creecy, R.H., Masand, B.M., Smith, S.J., ve Waltz, D.L., (1992), "Trading MIPS ve Memeory for Knowledge Engineering: Classifying Cencus Returns on the Connection Machine", Communication of the ACM, 35(8):48-64.

Davis, J. ve Goadrich, M., (2006), "The Relationship between Precision-Recall and ROC Curves", In W.W. Cohen and A. Moore eds. Proceedings of the 23rd International Conference on Machine Learning (ICML), 25-29 Jun 2006, Pittsburgh/Pennsylvania.

Dorre, J., Gerstl, P. ve Seiffert, R., (1999), "Text Mining: Finding Nuggets in Mountains of Textual Data", In U. Fayyad, S. Chaudhuri, and D. Madigan ed. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 15-18 Aug 1999, San Diago/CA, 398-401.

Dumais, S. ve Chen, H., (2000), "Hierarchical Classification of Web Content", In E. Yannakoudakis, N.J. Belkin, M.K. Leong, and P. Ingwersen eds. Proceedings of the 23rd Annual International SIGIR Conference on Research and Development in Information Retrieval, 24-28 Jul 2000, Athens/Greece.

Dumais, S., Platt, J., Heckermann, D. ve Sahami, M., (1998), "Inductive Learning Algorithms and Representations for Text Categorization", In G. Gardarin, J.C. French, N. Pissinou, K. Makki, and L. Bouganim eds. Proceedings of the 7th ACM International Conference on Information and Knowledge Management (CIKM) 02-07 Now 1998, Bethesda/Maryland.

Dunning, T., (1993), "Accurate Methods for the Statistics of Surprise and Coincidence", Computational Linguistics, 19(1):61-74.

Everitt, B.S., (1993), Cluster Analysis, 3rd ed., John Wiley and Sons, New York, Edvard Arnold, London, Halsted Pres, New York.

Falinouss, P., (2007), "Stock Trend Prediction Using News Articles: A Text Mining Apporach", Yüksek Lisans Tezi, Lulea University of Technology.

Fama, E.F., (1970), "Efficient Capital Markets: A Review of Theory and Empirical Work", Papers and Proceedings of the Twenty-Eighth Annual Meeting of American Finance Association, Journel of Finance, 28-30 Dec 1969, New York.

Fix, E. ve Hodges, J.L., (1951), "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties", USAF School of Aviation Medicine.

Fuhr, N., Hartmanna, S., Knorz, G., Lusting, G., Schwantner, M. ve Tzeras, K., (1991), "Air/X – a Rule-Based Multistage Indexing Systems for Large Subject Fields", In A. Lichnerowicz ed. Proceedings of the 3rd RLAO Conference, 02-05 Apr 1991, Barcelona/Spain.

Fukomoto, F. ve Suzuki, Y., (2001), "Learning Lexical Representation for Text Categorization", Proceedings of the 2nd NAACL (North American Chapter of the Association for Computational Linguistics) Workshop on Wordnet and Other Lexical Resources: Applications, Extensions and Customizations, 03-04 Jun 2001, Pittsburgh/Pennsylvania.

Fung, G.P.C., Yu, J.X. ve Lu, H., (2005), "The Predicting Power of Textual Information on Financial Markets", IEEE Intelligent Informatics Bulletin, 5(1):1-10.

Galavotti, L., Sebastiani, F., Simi, M., (2000), "Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization", In J.L. Borbinha and T. Baker eds. Research and Advanced Technology, Proceeding of the 4th European Conference on Digital Libraries, 18-20 Sep 2000, Lisbon/Portugal.

Gidofalvi, G., (2001), "Using News Articles to Predict Stock Price Movements", University of California.

Grobelnik, M., Mladenic, D. ve Milic-Frayling, N., (2000), "Text Mining as Integration of Several Reletad Research Area: Reports on KDD 2000 Workshop on Text Mining", ACM SIGKDD Explorations Newsletter, 2(2):99-102.

Hearst, M.A., Schoelkopf, B., Dumais, S., Osuna, E. ve Platt, J., (1998), "Trends and Controversies-Support Vector Machines", IEEE Intelligent Systems, 13(4):18-28.

Hearst, M.A., (2003), "What is Text Mining?", University of California.

Jeckins, C., Jackson, M., Burden, P. ve Wallis, J., (1999), "Automatic RDF Metadata Generation for Resource Discovery", Computer Networks: The International Journal of Computer and Telecommunications Networking, 31(11-16):1305-1320.

Joachims, T., (1997), "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", In D.H. Fisher ed. Proceedings of the 14th International Conference on Machine Learning (ICML), 08-12 Jul 1997, Nashville, Tennessee.

Joachims, T., (1998), "Text Categorizations with Support Vector Machines: Learning with Many Relevant Features", In C. N\'edellec and C. Rouveirol eds. Proceedings of the 10th European Conference on Machine Learning (ECML), Application of Machine Learning and Data Mining in Finance, 21-24 Apr 1998, Cemnitz/Germany.

Joachims, T., (1999), Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press, Cambridge.

Joachims, T., (2000), "Estimating the Generalization Performance of a SVM Efficiently", In Proceedings of the International Conference on Machine Learning, 2000, San Fransico/CA, 431-438.

Joachims, T., (2002), Learning to Classify Text Using Support Vector Machines, Kluwer, Boston.

Karanikas, H. ve Theodoulidis, B., (2002), "Knowledge Discovery in Text and Text Mining Software", UMIST University.

Khare, R., Pathak, N., Gupta, S.K. ve Sohi, S., (2004), "Stock Broker P – Sentiment Extraction for the Stock Market", In A. Zanasi, N.F.F. Ebecken, and C.A. Brebbia eds. Data Mining V, Proceedings of the Fifth International Conference on Data Mining, Text Mining and Their Business Applications, 15-17 Sep 2004, Malaga/Spain.

Koppel, M. ve Shtrimbers I., (2004), "Good News or Bad News? Let the Market Decide", In AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 86-88.

Kroeze, J.H., Matthee, M.C. ve Bothma, T.J., (2003), "Differentiating Data and Text Mining Technology", In J. Eloff, A. Engelbrecht, P. Kotze, and M. Eloff, eds. Proceedings of the ACM 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT) on Enablement Through Technology, 17-19 Sep 2003 Johannesburg.

Kroha, P. ve Baeza-Yates, R., (2004), "Classification of Stock Exchange News", Universidad de Chile.

Kwok, J.T., (1998), "Automated Text Categorization Using Support Vector Machine", In S. Usui, and T. Omori eds. Proceedings of the 5th International Conference on Neural Information Processing (ICONIP), 21-23 Oct 1998, Kitakyushu/Japan.

Kwon, O.W. ve Lee, J.H., (2000), "Web Page Classification Based on K-Nearest Neighbor Approach", In K.F. Wong, D.L. Lee, and J.H. Lee eds. Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages (IRAL), 30 Sep-1 Oct 2000, Hong Kong/China.

Larkey, L.S., (1998), "Some Issues in the Automatic Classificaiton of U.S. Patents", In Workshop on Learning for Text Categorization, at the 15th National Conference on Artificial Intelligence, 26-30 Jul 1998, Madison/WI.

Lam W., Low, K.F. ve Ho, C.Y., (1997), "Using a Bayesian Network Induction Approach for Text Categorization", In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI), 23-29 Aug 1997, San Francisco.

Lavrenko, V., Lawrie, D., Ogilvie, P. ve Schmill, M., (1999), "Electronic Analyst of Stock Behavior", Information Mining Seminar, University of Massachusetts.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. ve Allan J., (2000), "Language Models for Financial News Recommendation", In Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM), 06-11 Nov 2000, McLean/Virginia.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. ve Allan J., (2000), "Mining of Concurrent Text and Time Series", In Proceedings of the Workshop on Text Mining at the Sixth International Conference on Knowledge Discovery and Data Mining, 20-23 Aug 2000, Boston/MA.

Lewis, D.D. ve Ringuette, M., (1994), "A Comparison of Two Learning Algorithms for Text

Categorization" , In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR), 11-13 Apr 1994, Las Vegas/Nevada.

McCallum, A.K. ve Nigam, K., (1998), "A Comparison of Event Models for Naive Bayes Text Classification", In ICML/AAAI Workshop on Learning for Text Categorization at the 15th International Conference on Machine Learning, 26-27 Jul 1998 , Wisconsin.

Mittermayer, M.A., (2004), "Forecasting Intraday Stock Price Trends with Text Mining Techniques", In Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS), 05-08 Jan 2004, Big Island/Hawaii.

Mladenic, D., (1998), "Feature Subset Selection in Text-Learning", In C. Nedellec and C. Rouveirol eds. Proceedings of the 10th European Conference on Machine Learning (ECML) 21-23 Apr 1998 Chemnitz/Germany.

Ng, H.T., Goh, W.B. ve Low, K.L., (1997), "Feature Selection, Perception Learning, and a Usability Case Study for Text Categorization", In N.J. Belkin, A.D. Narasimhalu, P. Willet, and W. Hersh eds. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 27-31 Jul 1997, Philadelphia/Pennsylvania.

Ponte, J.M. ve Croft, W.B., (1998), "A Language Modeling Approach to Information Retrieval", In Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval, 24-28 Aug 1998, Melbourne/Australia.

Raghavan, P., (2002), "Structure in Text: Extraction and Exploitation", In S. Amer-Yahia and L. Gravano eds. Proceedings of the 7th International Workshop on the Web and Databases (WebDB): Collocated with ACM SIGMOD/PODS 2004, 17-18 Jun 2004, Maison de la Chimie/Paris/France.

Rogati, M. ve Yang, Y., (2002), "High-Performing Feature Selection for Text Classification", In C. Nicholas, D. Grossman, K. Kalpakis, S. Qureshi, H. van Dissel, and L. Seligman eds. Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM), 04-09 Nov 2002, McLean/Virginia.

Ruiz, M.E., ve Srinivasan, P., (1999), "Hierarchical Neural Networks for Text Categorization", In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 15-19 Aug 1999, Berkeley/California.

Schutze, H., Hull, D.A. ve Pederson, J.O., (1995), "Toward Optimal Feature Selection", In E.A. Fox, P. Ingwersen, and R. Fidel eds. Proceedings of the 18th Annual International ASM SIGIR Conference on Research and Development in Information Retrieval, 09-13 Jul 1995, Seatle/Washington.

Sebastiani, F., (1999), "A Tutorial on Automated Text Categorization", In A. Amandi and A. Zunino eds. Proceedings of the 1st Argentinean Symposium on Artifical Intelligence (ASAI), 08-09 Sep 1999, Buenos Aires/Argentina.

Sebastiani, F., (2002), "Machine Learning in Automated Text Categorization", ACM Computing Surveys (CSUR), 34(1):1-47.

Seo, Y.W., Ankolekar, A. ve Sycara, K., (2004), "Feature Selection for Extracting Semantically Rich Words", Technical Report CMU-RI-TR-04-18, Robotic Institute, Carnegie Mellon University.

Siolas, G. ve d'Alche-Buc, F., (2000), "Support Vector Machines Based on Semantic Kernel for Text Categorization", In S.I. Amari, C.L. Giles, M. Gori, and V. Piuri eds. Proceedings of IEEE-INNS-ENNS International Joint Conference on Neural Networks, Neural Computing: New Challenges and Perspectives for the New Millennium 24-27 Jun 2000, Como/Italy.

Tan, A. H., (1999), "Text Mining: The State of Art and Challenges", In PAKDD Workshop on Knowledge Discovery from Advanced Databases (KDAD'99) in Conjuction with Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'9), 26-28 Apr 1999, Beijing/China, 71-76.

Tzeras, K. ve Hartman, S., (1993), "Automatic Indexing Based on Bayesian Inference Networks", In R. Korfhage, E.M. Rasmussen, and P. Willett eds. Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval, 27 Jun-01 Jul, 1993, Pennsylvania.

Van Rijsbergen, C.J., (1979). Information Retrieval, 2nd ed. Butterworths, London.

Vapnik, V.N., (1995), The Nature of Statistical Learning Theory, Springer, New York.

Vapnik, V.N., (1998), Statistical Learning Theory, Wiley-Inter Science, New York.

Wallis, F., Jin, H., Sista, S. ve Schwartz, R., (1999), "Topic Detection in Broadcast News", In Proceedings of the DARPA Broadcast News Workshop 28 Feb-3 Mar 1999, Herndon/Virginia.

Wen, Y., (2001), "Text Mining Using HMM and PPM", Yüksek Lisans Tezi, University of Waikato.

Wiener, E., Pedersen, J.O. ve Weigend, A.S., (1995), "A Neural Network Approach to Topic Spotting", In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR), 24-26 Apr 1995, Las Vegas/Nevada.

Wilbur, J.W. ve Sirotkin, K., (1992), "The Automatic Identification of Stop Words", Journal of Information Science, 18(1):45-55

Wu Y.-C., (2007), "Predicting the Trend of Taiwan Weighted Stock Index with Text Mining Techniques", Yüksek Lisans Tezi, National Central University.

Wuthrich, B., Cho, V., Leung, S., Permunetillek, D., Zhang J. ve Lam, W., (1998), "Daily Stock Market Forecast from Textual Web Data", IEEE International Conference on Systems, Man, and Cybernetics, 11-14 Oct 1998, San Diego/CA.

Yang, Y., (1994), "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval", In W.B. Croft and C.J. van Rijsbergen eds. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 03-06 Jul 1994, Dublin/Ireland.

Yang, Y., (1995), "Noise Reduction in a Statistical Approach to Text Categorization", In E.A. Fox, P. Ingwersen, and R. Fidel eds. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 09-13 Jul 1995, Seattle/Washington.

Yang. Y., (1999), An Evaluation of Statistical Approaches to Text Categorization, Information Retrieval, 1(1-2) :69-90.

Yang, Y. ve Chute, C.G., (1994), "An Example-based Mapping Method for Text

Categorization and Retrieval", ACM Transactions on Information Systems (TOIS): Special Issue on Text Categorization, 12(3):252-277.

Yang, Y. ve Liu, X., (1999), "A Re-Examination of Text Categorization Methods", In Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieveal, 15-19 Aug 1999, Berkley/California.

Yang, Y. ve Pedersen, J.O., (1997), "A Comparative Study on Feature Selection in Text Categorization", In D.H. Fisher ed. Proceedings of the 14th International Conference on Machine Learning (ICML), 08-12 Jul 1997, Nashville/Tennessee

Yang, Y. ve Wilbur, J., (1996), "Using Corpus Statistics to Remove Redundant Words in Text Categorization", Journal of the American Society for Information Science, 47(5):357-369.

Zhang, T. ve Oles, F., (2001), "Text Categorization Based on Regularized Linear Classification Methods", Information Retrieval, 4(1):5-31.

İNTERNET KAYNAKLARI

[1] http://opennlp.sourceforge.net/

[2] http://svmlight.joachims.org/

ÖZGEÇMİŞ

Doğum tarihi	22.12.1981	
Doğum yeri	Nevşehir	
Lise	1994-1997	Kayseri Aydınlıkevler Lisesi
Lisans	1998-2003	Ege Üniversitesi Mühendislik Fak. Bilgisayar Mühendisliği Bölümü
Yüksek Lisans	2007-2010	Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Müh. ABD, Bilgisayar Müh. Programı

Çalıştığı kurum(lar)

2005-2008	KoçSistem Bilgi ve İletişim Teknolojileri
2008-Devam ediyor	Akbank T.A.Ş