İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

BORSA İSTANBUL (BIST) 100 ENDEKSİ YÖNÜNÜN EKONOMİ HABERLERİ İLE TAHMİN EDİLMESİ

YÜKSEK LİSANS TEZİ

Hakan GÜNDÜZ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

HAZİRAN 2013

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

BORSA İSTANBUL (BIST) 100 ENDEKSİ YÖNÜNÜN EKONOMİ HABERLERİ İLE TAHMİN EDİLMESİ

YÜKSEK LİSANS TEZİ

Hakan GÜNDÜZ
(504101548)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Zehra ÇATALTEPE

HAZİRAN 2013

İTÜ, Fen Bilimleri Enstitüsü'nün 504101548 numaralı Yüksek Lisans Öğrencisi **Hakan GÜNDÜZ,** ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"BORSA İSTANBUL (BIST) 100 ENDEKSİ YÖNÜNÜN EKONOMİ HABERLERİ İLE TAHMİN EDİLMESİ"** başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Doç. Dr. Zehra ÇATALTEPE** İstanbul Teknik Üniversitesi	……..
Jüri Üyeleri :	**Yrd. Doç. Dr. Ömer Sinan SARAÇ** İstanbul Teknik Üniversitesi	…….
	Yrd. Doç. Dr. Arzucan ÖZGÜR Boğaziçi Üniversitesi	…….

Teslim Tarihi : **03 Mayıs 2013**
Savunma Tarihi : **05 Haziran 2013**

Değerli anneme ve babama,

ÖNSÖZ

Yüksek lisans eğitimim süresince bana yol göstererek, benden her türlü yardımını ve desteğini esirgemeyen değerli hocam Doç. Dr. Zehra ÇATALTEPE'ye teşekkürü borç bilirim. Yaşamım boyunca maddi ve manevi desteklerini her zaman yanımda hissettiğim kıymetli anneme ve babama da şükranlarımı sunarım.

Haziran 2013 Hakan Gündüz

İÇİNDEKİLER

Sayfa

KISALTMALAR

BIST : Borsa İstanbul
df : Doküman Sıklığı
dn : Doğru negatif
DO : Doğruluk Oranı
dp : Doğru pozitif
idf : Ters Doküman Sıklığı
KAP : Kamu Aydınlatma Platformu
NB : Naive Bayes
tf : Terim sıklığı
yn : Yanlış negatif
yp : Yanlış pozitif

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

BORSA İSTANBUL 100 ENDEKSİ YÖNÜNÜN EKONOMİ HABERLERİ İLE TAHMİN EDİLMESİ

ÖZET

Yapılan tez çalışmasında internet sitelerinde yayınlanan ekonomi haberleri kullanılarak Borsa İstanbul (BIST) 100 endeksi günlük açılış fiyatının yönü tahmin edilmiştir. Çalışmada kullanılan haber metinleri Borsa İstanbul'daki şirketlerin resmi bildirimlerinin yayınladığı Kamu Aydınlatma Platformu (KAP) internet sitesinden ve iki finansal internet sitesinden elde edilmiştir. Haber metinlerine metin madenciliği teknikleri uygulanarak, her işlem gününe ait öznitelik vektörleri oluşturulmuştur. BIST 100 endeksine ait günlük açılış fiyatlarındaki anlamlı değişiklikler sınanarak değişim yönünü gösteren sınıf etiketleri oluşturulmuş ve bu etiketler öznitelik vektörlerine atanmıştır.

Öznitelik vektörlerinin boyutunun yüksek olması ve örnek sayısının az olması nedeniyle, haber metinlerinde bulunan kelimeler üzerinde öznitelik seçimi gerçekleştirilmiştir. İncelenen problemin yapısı gereği, endeks fiyatındaki artış veya azalış yönündeki değişim, endeks fiyatının sabit kalmasına oranla daha az gerçekleştiğinden, bu durum oluşturulan veri kümesinin dengesiz sınıf dağılımına sahip olmasına yol açmıştır. Öznitelik seçiminde veri kümesindeki sınıf dengesizliği ile başa çıkabilmek için, karşılıklı bilgi öznitelik seçme yöntemi ile birlikte yeniden örneklemeyi temel alan yeni bir öznitelik seçme yöntemi ortaya konmuştur. Dengeli dağılımlı karşılıklı bilgi yöntemi olarak isimlendirilen bu yöntem kullanılarak BIST 100 endeksinin anlamlı yön değişimleri %74 doğruluk ve %68,4 Makro-ortalama F-Ölçütü oranlarıyla tahmin edilmiştir. Dengesiz dağılımlı karşılıklı bilgi yöntemi; karşılıklı bilgi, bilgi kazanımı ve Ki-kare istatistiği öznitelik seçme yöntemleri ile karşılaştırılmış ve dengesiz dağılımlı bilgi kazanımının daha az sayıda öznitelik kullanarak daha başarılı sınıflandırma sonuçları elde ettiği görülmüştür.

Ayrıca kullanılan haber metinlerinin alındığı kaynaklar göz önünde bulundurularak tahmin performansı incelenmiş ve finansal internet sitelerinden alınan haber metinlerinin endeks yönüne KAP'dan alınan haber metinlerinden daha fazla etki ettiği tespit edilmiştir.

Son olarak, haber metinlerinden elde edilen özniteliklere BIST 100 endeksinin günlük geçmiş açılış, kapanış ve hacim fiyatları öznitelik olarak eklenmiş, eklenen fiyat bilgisinin tahmin performansında gelişmeye neden olmadığı görülmüştür.

Deneylerin ardından dengeli dağılımlı karşılıklı bilgi ile seçilen 50 özniteliğin BIST 100 endeksi üzerindeki etkileri incelenmiş ve endeksin artışına, azalışına ve sabit kalmasına neden olan kelimeler tespit edilmiştir.

PREDICTION OF BORSA ISTANBUL (BIST) 100 INDEX DIRECTION USING FINANCIAL NEWS ARTICLES

SUMMARY

Stock market prediction has been an attractive research topic for many years. The stock market is a complex, dynamic, and non-linear system and it is defined by data intensity, noise, non-stationary, unstructured nature, high degree of uncertainty, and hidden relationships. Many factors interact in stock market including political events, general economic conditions, and traders' expectations. But it is hard to predict the future stock price or direction because of the complex and dynamic nature of the stock market behavior due to the factors that affect it.

In order to predict the future stock price or direction, there are two analysis methods used in the literature. The first method is the technical analysis which is based on historical stock prices. By examining the daily, monthly and yearly past prices of the stock, the future stock price or stock direction is predicted. In this approach, Time Series Analysis techniques are used and market timing is very important. Strategists used this analysis approach, believe that market timing is critical and opportunities can be found through the careful averaging of historical price and volume movements and comparing them against current prices.

The second method is the fundamental analysis and this analysis includes the important numbers about the structure of the economy. When the fundamental analysis is used, features are derived from the overall economy, the particular industry's sector, or most typically, from the company itself. Features like interest rates, inflation, unemployment percentage and economic growth can all play a part in determining the price of a stock and are included in the model for prediction.

In the last 10 years, alternative analysis choices have been investigated for stock market prediction. Due to the development of information technologies, abundant information can be accessed about finance especially through the financial news on web sites. Several studies showed that there is a direct link between the stock markets and financial news. Thus, financial news which include general news articles, company releases, news of global economy affect the performance of stock market and they can be used to predict the future stock prices and their directions.

If we use the financial news articles to make the prediction, we need to extract important information that may have an effect on the stock market. News articles found on World Wide Web are unstructured and we need to convert them into a structured form in order to analyze patterns in these articles. Natural Language Processing and Data Mining methods can create feature vectors out of the news articles. Each element in a feature vector is named a term and it represents the words that occur in a news document.

In this thesis, we predicted the direction of Borsa Istanbul (BIST) 100 Index (XU 100) open price using the news articles released the day before. We used the BIST 100 Index which is the main market indicator of the ISE. Unlike other BIST (formerly called ISE) prediction studies, the presence of the words and term frequencies in the news articles were used as features. News articles were provided by Public Disclosure Platform of BIST and two financial websites “Mynet Finans” and “Bigpara”. Text mining techniques were applied on the news articles to form feature vectors for each trading day. The significant changes of BIST 100 index open price were examined to create the class labels and these labels were assigned to input vectors available for each day.

Due to the high dimensionality of inputs and small number of instances for training, feature selection on the words of the news articles was needed. By its nature, significant positive or negative changes in stock price happen much less than non-significant changes resulting in an unbalanced data set. In order to deal with the data imbalance problem, a mutual information and resampling based feature selection method was devised. The method considered the fact that classes were imbalanced for the stock market data and computed the MI after balancing the data, therefore we called this method of relevance computation Balanced Mutual Information (BMI). After feature selection process, a Naïve Bayes classifier was trained to predict the direction of BIST 100 Index. The classification performance was evaluated by means of accuracy. But, it is not necessarily a suitable metric to evaluate performance of each class in the classification process for unbalanced data. In order to assess the classifier performance on class level, we used F-Measure metric which is the combination of precision and recall metrics. F-Measure evaluates the classifier performance on the different classes separately. The overall classifier performance was evaluated with Macro-Averaged F-Measure metric.

Experimental results showed that, with balanced feature selection, the significant changes in the BIST 100 Index could be predicted, with an accuracy of 74% and a Macro Averaged F-Measure of 0.684. This balanced feature selection method were compared with three other methods, a basic Mutual Information based, Information Gain and Chi-Square feature selection and it was found out that balanced feature selection results in higher performance using a smaller number of features.

In a second experiment, the sources of the news articles were evaluated separately. It was found out that the internet news have more impact on the stock market direction than the official BIST news. The classification performance using only internet news gave better Macro Averaged F-Measure rate than the official BIST news and the combination of two types of news articles performed slightly worse than internet news using small number of features.

As a final experiment, the previous days' price information was added as features and it was found out that additional price information does not result in a reasonable improvement. These price features especially boosted the classification performance in small number of feature subsets.

The effects of the words on BIST 100 Index were also examined. Using selected words with balanced mutual information method, three different colored word clouds were formed.

These word clouds showed the effective words that caused the significant changes on BIST 100 index price. The most effective words caused the increase in BIST 100 index price were shown in the green-colored word cloud. The red-colored word cloud indicated the words that influenced the index price in decreasing direction. The blue-colored word cloud showed the words led to the index price in unchanged position. The most influential words for each word cloud were also determined.

1. GİRİŞ

Borsa tahmini son yıllarda ilgi çeken yapay öğrenme konularından bir tanesidir. Bu ilginin temel nedeni, başarılı tahmin sonuçlarının yatırımcılara daha fazla kar getireceği beklentisidir. Ancak borsanın doğasının karmaşık, gürültülü, dinamik ve doğrusal olmayan bir yapıda olması borsa endeksinin fiyatını veya hareket yönünü tahmin etmeyi zorlaştırmaktadır (Abu-Mostafa ve Atiya, 1996).

Borsa tahmininde birçok yöntem bulunmasına rağmen bu yöntemlerin çoğunda geçmiş endeks fiyatları gibi nümerik verilere dayanan zaman serileri analizi kullanılmaktadır (Marcek, 2004). Borsa endeksinin davranışını etkileyen birçok etmen olması ve bu etmenlerin birbirleriyle karmaşık ilişkisi borsa tahmininin sadece fiyat verileri ile yapılmasını engellemektedir.

Son yıllarda yapılan çalışmalarda yayınlanan ekonomi haberleri ile borsa endeksi arasında doğrudan bağlantı tespit edilmiştir (Mitchell ve Mulherin, 2001). Bu çalışmalar, borsa tahminin fiyat verileri yerine ekonomi haberleri kullanılarak da yapılabileceğini göstermiştir. Özellikle finansal internet sitelerinden alınan bu haberler beklenmedik bilgiler içermekte ve borsa endeksi üzerinde etki oluşturmaktadır. İnternet ortamındaki bu haberlerinin yoğunluğunun artması, bilgisayar bilimcileri metinlerden değerli bilgiler çıkarmaya yönlendirmiştir. Metinsel verilerin yapılandırılmamış durumda olması, bu verilerden anlamalı bilgi çıkarma için Metin Madenciliği denilen yeni bir disiplinin doğmasına sebep olmuştur. Metin madenciliği teknikleri ile çıkarılan örüntüler, yapay öğrenme algoritmalarıyla birlikte değerlendirilerek borsa endeksi tahmininde tatmin edici sonuçlar edilmiştir.

Literatürde var olan çalışmalardan yola çıkarak, tez çalışmasında internet sitelerinde yayınlanan ekonomi haberlerinin Borsa İstanbul (BIST) üzerindeki etkileri incelenecektir. Çalışmada öznitelik olarak geçmiş endeks fiyatları yerine ekonomi haberlerinde geçen kelimeler kullanılacaktır. Haber metinlerinde geçen bu kelimeler ile BIST 100 endeksi fiyatının değişim yönü tahmin edilecektir. Yön tahminiyle birlikte tahminde kullanılan kelimelerin etkinliği incelenecektir.

1.1 Tezin Amacı

Yapılan tez çalışmasının amacı, finansal internet sitelerinde yayınlanan ekonomi haber metinlerini kullanarak BIST 100 endeksinin günlük açılış fiyatının yönünü tahmin etmektir. Kullanılan ekonomi haberleri Türkiye'nin önemli 3 finans internet sitesinden alınmıştır. Haberler; şirket bildirimlerini, genel ekonomi haberlerini, araştırma raporlarını, uzman tavsiyelerini içermektedir. Haber dokümanları metin işleme teknikleri kullanılarak öznitelik vektörlerine dönüştürülecektir. İki farklı doküman gösteriminde öznitelik vektörleri oluşturulduktan sonra aynı gün içinde yayınlanmış öznitelik vektörleri birleştirilerek her gün için tek bir öznitelik vektörü elde edilecektir. BIST internet sitesinden (Url-1) alınan BIST 100 endeksinin günlük açılış fiyatları kullanılarak da her işlem günü için, BIST 100 endeksinin açılış fiyatının hangi yönde değişeceğini gösteren sınıf etiketleri oluşturulacaktır. Oluşturulan etiketler haberlerin yayınlanma tarihleri göz önünde bulundurularak öznitelik vektörlerine atanacaktır. Öznitelik vektörlerinin binlerce öznitelikten oluştuğundan, bilgi içermeyen öznitelikler Bilgi Kazanımı, Karşılıklı Bilgi ve Ki-Kare istatistiği öznitelik seçme yöntemleri kullanılarak ayrıştırılacaktır.

Öznitelik seçimi ile elde edilen öznitelikler BIST 100 endeksinin değişim yönünü tahmin eden Naive Bayes sınıflandırıcısının eğitiminde kullanılacaktır. Sınıflandırıcı eğitiminden sonra sınıflandırıcının performansı test edilecek ve Doğruluk Oranı, F-Ölçütü ve Makro-ortalama F-Ölçütü ile değerlendirilecektir.

1.2 Tezin Yapısı

Yapılan tez çalışması 5 bölümden oluşmaktadır. 1.Bölüm'de tez çalışması ile ilgili genel bilgiler ve tez çalışmasının amacı açıklandıktan sonra 2.Bölüm'de, borsa yön tahmini ve finansal haber metinleri kullanarak borsa tahmini ile ilgili çalışmalar özetlenmiştir. Bu bölümde ayrıca Borsa İstanbul ile ilgili yapılan çalışmalara da yer verilmiştir. 3.Bölüm'de haber metinlerinin borsa endeksi yönü tahmininde kullanılmasıyla ilgili genel süreçlerden ve kullanılan yöntemlerden bahsedilmiştir. Yapılan tez çalışmasının uygulama aşamaları ve deneysel sonuçları Bölüm 4'te anlatılmıştır. Son bölüm olan Bölüm 5'te ise tez çalışmasının sonuçlarından bahsedilmiştir.

2. İLGİLİ ÇALIŞMALAR

Giriş bölümünde anlatıldığı üzere, internet sitelerinde yayınlanan ekonomi haberlerini kullanarak BIST 100 Endeksi'nin günlük değişim yönünü tahmin etmek bu çalışmanın amacı olarak belirlenmiştir. Bu amaç doğrultusunda Bölüm 2'de, literatürde yer alan çalışmalardan bahsedilecektir.

2.1 Borsa Yön Tahmini

Borsa yön tahmini birçok araştırmacının ilgisini çekmiştir. Yapılan çalışmalarda, eldeki verilerden yararlı örüntüler çıkartılarak borsa endeksi fiyatının değişim yönü tahmin edilmektedir. Araştırmacılar borsa yön tahminini gerçekleştirmek için iki yaklaşımdan faydalanmışlardır. Bu yaklaşımlar Teknik Analiz ve Temel Analiz'dir (Elkan, 1999). Bir sonraki başlık altında bu iki yaklaşım detaylıca anlatılacaktır.

2.1.1 Teknik analiz

Teknik analiz yaklaşımı, hisse fiyatlarını veya değişim yönünü tahmin etmede geçmiş fiyat verilerini temel alır (Abdullah ve Ganapathy, 2000). Bu yaklaşımda geçmiş fiyat verilerinden yararlanılarak hazırlanan grafikler de önemli araçlardır. Teknik analiz, finansal zaman serileri madenciliği olarak da adlandırılan geçmiş fiyat verilerinden örüntüler çıkarma işlemidir. Teknik analiz yardımıyla, bir hissenin fiyat eğiliminin nasıl değiştiği, işlem hacminin hisse fiyatına ne yönde etkilediğini gibi bilgiler elde edilir. Bu bilgileri kullanarak borsa fiyatı ve yönü tahmini yapan çok sayıda çalışma mevcuttur (Hellmstrom ve Holmstrom,1998).

Teknik analiz sayısal zaman serilerine dayanır ve borsa tahmininde teknik göstergeleri kullanır. Teknik analizde zamanlama kritik ve önemlidir (Marcek, 2004). Analiz yapılırken geçmiş fiyat ve hacim değerlerinin ortalaması alınır ve bulunan değerler günümüzdeki fiyat ve hacim değerleriyle karşılaştırılarak borsanın veya hisse senedinin yönü tahmin edilmeye çalışılır. Tahmin işlemi yapılırken birçok zaman serileri yönteminden faydalanılabilir. En çok kullanılan iki yöntem ise

Hareketli Ortalama (İng. Moving Average) ve Otoregresyon (İng. Autoregressive)'dur.

2.1.2 Temel analiz

Temel analiz, hisseye arz ve talebi etkileyen faktörleri araştırır. Temel analizde, faktörlerle ilgili bilgiler elde edildikten sonra bu bilgiler hisse senedi fiyatına yansımadan önce harekete geçilmelidir. Bilgilerin elde edilmesiyle hisse senedine etki etmesi arasında geçen süre boşluğu hisse alım-satımı için fırsat vermektedir. Bu zaman boşluğu hisse senedinin yönünü veya fiyatını tahmin etmede kullanılmaktadır.

Temel analiz, şirketler tarafından yayımlanan ekonomik verileri temel alır ve bu verileri hisse fiyatını veya yönünü tahmin etmede kullanır. Şirketlerin periyodik olarak yayınladıkları yıllık ve 3 aylık raporlar, bilançolar, gelir bildirimleri gibi ekonomik veriler temel analiz yaklaşımında kullanılır (Gidofalvi, 2001). Ayrıca şirketler hakkında çıkan haberler de temel analizin ilgi alanındadır. Bunun nedeni ise haberlerin hisse senetlerine arzı ve talebi etkileyen faktörlerden biri olmasıdır.

Temel analiz yaklaşımında şirketlerin ekonomik verileriyle birlikte, ülkenin ekonomisiyle ilgili veriler de göz önünde bulundurulur. Enflasyon oranı, işsizlik oranı, kişi başına düşen gelir miktarı gibi ekonomik veriler hisselerin fiyatlarına etki etmektedir (Schumaker ve Chan, 2009).

Teknik analiz ve temel analiz yaklaşımlarının dışında finansal haberler gibi metinsel veriler kullanılarak da borsa tahmini yapılmaktadır. Şirketler hakkında yayınlanan basın bültenleri, şirket içi haberler, küresel ekonomi haberleri, ülkelerin politik durumları ve son dakika haberleri şirketlerin hisselerinin fiyatını etkileme gücüne sahiptir. Literatürde yer alan çalışmalarda, ekonomi haberlerinin borsa endeksi veya hisse senedi fiyatı üzerine etkileri ve bunların yayınlanan haberlere karşı verdiği tepkiler incelenmiştir. Borsanın yayınlanan haber metinlerine tepki verdiği ve haber metinlerinin borsa endeksinin yönünü etkilediği çalışmalar da mevcuttur. Bir sonraki bölümde haber metinlerinin borsa üzerinde etkilerini gösteren akademik çalışmalardan bahsedilecektir. Bu çalışmalar, haber metinlerinin hem borsa endeksinin fiyatının hem de endeks fiyatının hareket yönünün tahmininde kullanılmasının önünü açmıştır. Bölüm 2.3' te ise borsa endeksi yönünün tahmininin yapıldığı çalışmalar ayrıntılı biçimde incelenecektir.

2.2 Haber Metinlerinin Borsa Üzerine Etkileri

Borsa ve piyasa ile ilgili haberler metinleri genellikle politik ve ekonomik mesajlar içerirler. Günümüzde web teknolojilerinin gelişmesiyle bu haber metinlerine elektronik olarak kolaylıkla birçok kaynaktan ulaşılabilir.

Ulaşılan haber metinlerine örnek olarak küresel ve yerel ekonomi haberleri, politika haberleri, şirket raporları ve finans uzmanı kişilerin tavsiye raporları verilebilir. Haber metinleri ile borsa endeksi arasındaki ilişkinin tespit edildiği çalışmalar aşağıda bahsedilmiştir.

Mitchell ve Mulherin (1994), Amerika Birleşik Devletleri (ABD) Borsası Dow Jones tarafından yapılan günlük duyurular ile piyasa faaliyetlerinin doğrudan etkilendiğini tespit emiştir. Bu tespiti yaparken günlük piyasa dönüşlerini ve hacim miktarlarını kullanmıştır. Öne çıkan duyuruların ise hisse senedi fiyatları üzerinde mutlak değişikliğe neden olduğunu keşfetmiştir.

Berry ve Howe (1994), Reuters Haber Servisi tarafından yayınlanan haberleri kullanarak, bu haberlerin gün içerisindeki borsa faaliyetlerine nasıl etki ettiğini incelemiştir. Çalışmada, yayınlanan haberler ile işlem hacmi arasında anlamlı bir ilişki olduğu bulunmuştur.

Chan ve diğ. (2001), borsanın haber metinlerine tepki gösterdiğini kabul etmiştir ve ekonomi haberlerinin işlem gören hisse sayısı üzerinde olumlu veya olumsuz etkilere sahip olduğunu bulmuştur. Çalışmada kullanılan ekonomi ve politika haberlerinin hisselerin alım ve satım sayısına ve sıklığına ve hisse fiyatlarında değişikliğe etki ettiğini tespit etmiştir.

Yukarıda bahsedilen çalışmalar, öne çıkan ekonomi haberlerinin borsa piyasası üzerinde doğrudan etkisi olduğunu göstermektedir. Yapılan bu çalışmalar, araştırmacıların ekonomi haberlerini borsa endeksinin fiyatını veya yönünü tahmin etmede de kullanılmasının yolunu açmıştır. Metin madenciliği ve doğal dil işleme tekniklerinin kullanılarak haber metinlerinden önemli bilgilerin çıkarılması sayesinde, çıkarılan bilgilerin borsa tahmininde kullanıldığı çok sayıda çalışma bulunmaktadır.

Bir sonraki başlıkta bu çalışmalarda kullanılan yöntemlerden ve bu yöntemlerin kullanılmasıyla ortaya çıkan durumlardan bahsedilecektir.

2.3 Haber Metinleri Kullanılarak Yapılan Borsa Tahmini Çalışmaları

Wultrich ve diğ. (1998), 5 büyük borsa endeksinin (Dow Jones (ABD), Nikkei 225 (Japonya), FTSE 100 (İngiltere), HSI (Hong Kong) ve SSI (Singapur)) günlük açılış fiyatlarının hangi yönde değişeceğini tahmin eden çevrimiçi bir sistem tasarlamıştır. Tahmini yaparken Wall Street Journal'da yayınlanan elektronik haberleri kullanmıştır. Sistem, Asya ülkelerindeki borsaların açılışından önce tahmin üretmektedir. Sistem borsa endeksini etkileyeceği düşünülen ve uzman kişi tarafından seçilmiş 400'den fazla anahtar kelime öbeğini temel almıştır. Bu kelime öbekleriyle birlikte son 100 işlem gününe ait kapanış fiyatları ve haber verileri de eğitim verisi olarak belirlenmiştir. Haber metinlerinde tespit edilen anahtar kelimelerin sayılarına bakılarak çeşitli ağırlıklandırma yöntemleri denenmiş ve Kural Tabanlı, En Yakın Komşu ve Yapay Sinir Ağları gibi yapay öğrenme metotları kullanılarak sistem eğitilmiştir. Eğitilen sisteme gün içinde çıkan haberler ve o gün sonunda endekslerin kapanış fiyatları verilerek, ertesi gün endekslerin fiyatının hangi yönde değişeceği tahmin edilmiştir. Çalışmada Kural Tabanlı yapay öğrenme yöntemi diğer kullanılan yöntemlere göre daha başarılı doğruluk oranı vermiştir. Sistem rastgele tahmine göre de anlamlı bir başarı sağlamıştır. Haber metinleri kullanılarak eğitilen sisteme, geçmiş 100 güne ait fiyat verileri de eklendiğinde sistemin doğruluk oranında artış görülmüştür. Sistemin dezavantajlı yönlü ise sınıflandırıcıların eğitimi belirlenen anahtar kelimeler kullanarak yapıldığından, haber metinlerinde geçen yeni ve endeksi doğrudan etkileyen bir kelime ortaya çıktığında sistemin bunu göz ardı etmesidir.

Gidofalvi (2001), 2001 yılında haber metinlerini kullanarak kısa süre içerisinde hisse senetlerinin hareket yönünü tahmin edilebileceğini ortaya koyan bir model tasarlamıştır.12 şirketin hisse senetlerinin gün içindeki fiyat verileri ile finansal haber metinlerinin yayınlanma zamanları dikkate alınarak çeşitli deneyler yapılmıştır. Yapılan çalışmada, gün içerisinde 10'ar dakikalık zaman aralıklarında hisse senetleri fiyatlarının değişim miktarları ve NASDAQ endeksinin değişim miktarı göz önünde bulundurularak skor hesaplanması yapılmıştır. Haber metinlerinin yayınlanma zamanına göre bulunan skorlar kullanılarak hangi zaman aralığında haber metinlerinin hisse senedi fiyatında anlamlı bir değişikliğe neden olduğu bulunmaya çalışılmıştır. Bunun için tüm haber metinlerine bulunan skorlar kullanılarak aşağı, yukarı ve sabit diye 3 sınıf etiketi atanmıştır. Etiket atama işleminden sonra Naive

Bayes sınıflandırıcısı kullanılarak deneyler gerçekleştirilmiştir. Deneylerden elde edilen sonuçlarda haber metinlerinin yayınlamasından 20 dakika öncesinin ve yayınlandıktan 20 dakika sonrasının, hisse senedi fiyatları üzerinde anlamlı değişikliğe neden olduğu görülmüştür. Bu çalışmanın eksikliği ise haber metinleri üzerinde herhangi bir önişleme yapılmamış olması ve öznitelik seçim yöntemlerinin kullanılmamış olmasıdır.

2004 yılında Mittermayer, NewsCats isimli bir tahmin sistemi önermiştir. Bu tahmin sistemi şirket basın bültenlerinin yayınlanmasından hemen sonra hisse senetlerinin fiyatının hangi yönde değişeceğini tahmin etmektedir. NewsCats sistemi 3 bileşenden oluşmaktadır. Birinci bileşen, metin işleme teknikleri ile basın bildirilerinden gerekli bilgiyi çıkarmaktadır. İkinci bileşen, işlenen basın bildirilerini önceden belirlenmiş sınıflara göre sıralamaktadır. Üçüncü bileşen ise yapılan bu sınıflandırma işlemine göre hisse senetlerinin fiyatlarının hareket yönünü belirlemektedir. Çalışmada kullanılan metin işleme teknikleri, öznitelik seçimini ve doküman gösterimini içermektedir. Birinci bileşende öznitelik seçimi, terim sıklığı (İng. Term frequency- tf), ters doküman sıklığı (İng. Inverse document frequency- idf) ve terim sıklığı –ters doküman sıklığı (tf x idf) gibi ölçütleri temel alarak yapılmıştır. Doküman gösterimi ise ikili gösterim veya tf, idf ve tf x idf gibi ölçütler kullanılarak gerçekleştirilmiştir. İşlenen basın bültenleri ikinci bileşende iyi haber veya kötü haber diye iki sınıfa ayrılmış, sınıflandırılan basın bültenleri kullanılarak Destek Vektör Makineleri yöntemiyle metin sınıflandırıcısı oluşturulmuştur. Yayınlanmış yeni basın bülteni, sistemin birinci bileşeninde metin işleme yöntemine tabi tutulmuş, ikinci bileşende bu bültenin ne tür bir haber olduğuna karar verilmiş, üçüncü bileşende ise bültenin hissenin fiyatını hangi yönde etkileyeceğinin sinyali oluşturulmuştur. Sistemin başarısı kesinlik (İng. Precision) ve duyarlılık (İng. Recall) ölçütleri kullanılarak belirlenmiştir. Bunun yanında piyasa benzetimi yapılarak da sistemin başarısı ölçülmüştür. NewsCats sistemi basın bülteni yayınlandıktan sonra rastgele olarak alım satım yapan yatırımcıya göre anlamlı bir başarı sağlamıştır (Mittermayer, 2004).

2009 yılında Schumaker ve Chan yaptığı çalışmada ekonomi haberlerini 3 farklı metin temsili şeklinde işleyerek borsanın hareket yönünü tahmin etmiştir. Kullanılan metin temsilleri Kelimeler Çantası (İng. Bag of Words), İsim Öbekleri (İng. Name Phrases) ve İsim Varlıkları (İng. Name Entities)'dır. Çalışmada kullanılan yapay

öğrenme yöntemi ise Destek Vektör Makineleri'nin bir türevi olan Ardışık Asgari Eniyileme'dir. Kullanılan yapay öğrenme yöntemi Doğrusal Regresyon yöntemine göre daha başarılı sonuç vermiştir.

2.4 Borsa İstanbul İlgili Yapılan Çalışmalar

Çalışmanın bu bölümünde İstanbul Menkul Kıymetler Borsası ile ilgili geçmiş çalışmalardan bahsedilecektir. İstanbul Menkul Kıymetler Borsası yeni adıyla Borsa İstanbul (BIST) 1985 yılında kurulmuştur. Gelişen piyasalar arasında en hızlı büyümeyi gösteren İstanbul Borsası'nda 2012 Kasım ayı itibari ile 402 şirketin hisse senetleri işlem görmektedir. 2012 yılında ortalama günlük işlem hacmi 1.3 Milyar ABD dolarına yükselmiştir (Url-1). Borsa İstanbul yoğun günlük işlem hacmi ve dalgalanmalardan dolayı yatırımcıların ilgisini çekmektedir. Bunun nedeni ise yoğun dalgalanmaların yatırımcılara büyük kazançlar sağlayacağı beklentisidir. Borsa İstanbul ilgili çalışmalarda BIST 100 endeksinin tahmini yapılmaktadır. BIST 100 endeksi Borsa İstanbul'un temel göstergesidir ve seçilen 100 şirketin hisse senedi fiyatları kullanılarak hesaplanır.

BIST 100 endeksine seçilen şirketlerin sermayesi toplamı bütün borsa da bulunan şirketlerin sermayelerinin yaklaşık olarak %75'ini oluşturmaktadır (Bildik, 2001).

Literatürde, gelişmiş ülkelerin borsalarının tahminiyle ilgili birçok çalışma bulunmasına rağmen, Türkiye gibi gelişen ülkelerin borsaları ile ilgili çalışma sayısı az olmaktadır. Yapılan çalışmalar incelendiğinde, BIST 100 endeksinin günlük dönüşlerinin ve günlük dalgalanmalarının tahmin edildiği görülmüştür. Bu tahminler fiyat verileri kullanılarak yapılmıştır.

Bildirici ve Ersin (2008) tarafından yapılan çalışmada Yapay Sinir Ağları ile birlikte Oto regresyon modelleri kullanılarak BIST 100 endeksinin günlük dönüşleri tahmin edilmeye çalışılmıştır.

Boyacıoğlu ve Avcı (2010) çalışmalarında BIST 100 endeksinin günlük dönüşünü, makroekonomik fiyat verilerini ve 3 yabancı ülkenin (ABD, Almanya ve Brezilya) borsalarının günlük kapanış fiyatlarını kullanarak tahmin etmiştir. Çalışmada Uyarlamalı Ağ Tabanlı Bulanık Mantık Çıkarım Sistemi (İng. Adaptive Neuro Fuzzy Inference System-ANFIS) kullanılmıştır. ANFIS'in diğer yapay öğrenme yöntemlerine güçlü bir alternatif olduğu çalışmadan elde edilen sonuçlardan biridir.

Kara ve diğ. (2011) yaptıkları çalışmada 10 adet teknik gösterge kullanarak BIST 100 endeksinin günlük kapanış fiyatının hareket yönünü tahmin etmiştir. Tahmin sürecinde yapay öğrenme yöntemlerinden Destek Vektör Makineleri ve Yapay Sinir Ağları kullanılmıştır. BIST 100 endeksinin geçmiş 10 yıldaki günlük kapanış fiyatı verileri kullanılarak iki yöntem için de uygun parametre kümeleri belirlenmiştir. Belirlenen parametreler ile iki yöntemin eğitimi tamamlanmış ve tahmin sonuçları elde edilmiştir. İki yapay öğrenme yöntemi de endeksin hareket yönünün tahmininde anlamlı başarı sağlamıştır. Yapay Sinir Ağları yöntemi ile elde edilen ortalama başarı Destek Vektör Makineleri ile elde edilenden yüksek çıkmıştır.

Yukarıda bahsedilen çalışmaların tümünde giriş olarak borsa endeksine ait geçmiş fiyat verileri kullanılmıştır. Borsa endeksi yönü tahmininde fiyat verilerinin yerine ekonomi haberlerinin kullanıldığı çalışma ise 2013 yılında Gündüz ve Çataltepe (2013) tarafından yapılmıştır. Çalışmada finansal internet sitelerinden alınan haber dokümanlarının BIST 100 endeksini hangi yönde etkileyeceği tespit edilmiştir. Haber dokümanları metin işleme teknikleriyle öznitelik vektörlerine dönüştürülmüş ve öznitelik vektörlerinden bilgi verici öznitelikleri seçmek için öznitelik seçme yöntemleri kullanılmıştır. Seçilen öznitelikler ile Naive Bayes sınıflandırıcısı eğitilmiştir. Sınıflandırıcının tahmin performansı incelendiğinde ise TFxIDF öznitelik seçme yönteminin Karşılıklı Bilgi seçme yöntemine göre daha başarılı olduğu kanaatine varılmıştır.

Tez çalışmasının bu bölümünde borsa endeksi tahmini ile ilgili geçmişte yapılan çalışmalar incelenmiştir. Bir sonraki bölümde haber metinlerinin borsa tahmininde kullanılabilmesi için gerekli aşamalar anlatılmış ve bu aşamalarda kullanılan yöntemlere değinilmiştir.

3. BORSA TAHMİNİNDE HABER METİNLERİNİN KULLANIM SÜREÇLERİ

Ekonomi haberleri kullanılarak yapılan borsa yön tahmini genel olarak metin sınıflandırma problemi olarak düşünebilir. Metin sınıflandırma problemi, metin dokümanlarının daha önceden tanımlanış sınıflara kategorize etme işlemidir. Yapılan çalışmada gün içinde yayınlanan haber metinlerinin içeriği kullanılarak bir gün sonra BIST 100 endeksinin açılış fiyatının hangi yönde değişeceği bulunacaktır. Bunun için endeksin değişim yönünü gösteren yukarı (+1), aşağı (-1) ve sabit (0) diye üç sınıf tanımlanmıştır. Yukarı sınıfı yayınlanan haber metinlerinin borsa endeksi fiyatının artışına neden olduğunu, aşağı sınıfı haberin borsa endeksi fiyatının azalışına neden olduğunu, sabit sınıfı ise haberin fiyat üzerinde hiçbir etkisi olmadığını gösterecektir. Sınıflandırılmış haber metinleri yapay öğrenme yöntemleri ile sınıflandırıcı eğitiminde kullanılacak, eğitilen sınıflandırıcı yeni bir haber metni yayınladığında da bu metnin borsa endeksi üzerinde hangi yönde etki edeceğini tahmin edecektir.

Bu başlıkta tez çalışmasının gerçekleşme aşamaları ve bu aşamalarda kullanılan yöntemler açıklanacaktır. Bu bölüm boyunca tez çalışması metin sınıflandırma problemi şeklinde anılacaktır.

3.1 Metin Sınıflandırma Süreci

Metin sınıflandırma problemi birden fazla adımı içermektedir. Bu problemin adımları araştırmacılar tarafından değişik şekilde ifade edilmektedir.

Montanes ve diğ. (2003), yaptıkları çalışmada metin sınıflandırma sürecinin 3 aşamadan oluştuğunu belirtmiştir. Bu aşamalar doküman gösterimi (İng. Document representation), öznitelik boyutunun indirgenmesi (İng. Feature reduction) ve sınıflandırmadır (İng. Classification).

Wang Y. ve Wang X. J. (2005) ise metin sınıflandırma problemlerinin içerdiği süreçleri metin gösterimi (İng. Text representation), sınıflandırıcı eğitimi (İng. Classifier training) ve performans değerlendirmesi (İng. Performance evaluation)

olarak tanımlamıştır. Metin gösterimi aşaması, metin önişleme (İng. Text preprocessing), öznitelik boyutunun indirgenmesi ve doküman gösterimi işlemlerini içermektedir.

Yukarıda bahsedilen iki çalışmadan yola çıkarak tez çalışmasında izlenen süreçler, bu bölümün alt başlıkları altında incelenecektir. Bölümde 3.2'de metin sınıflandırmanın ilk aşaması olan metin önişleme süreci ve ön işlemede kullanılan yöntemler anlatılacaktır. Doküman gösterimi aşaması Bölüm 3.3'de, öznitelik seçimi ve öznitelik seçiminde kullanılan yöntemler Bölüm 3.4'te ele alınacaktır. Sınıflandırıcı eğitimi ve sınıflandırma işleminin performansını değerlendirmede kullanılan ölçütler ise sırasıyla Bölüm 3.5 ve Bölüm 3.6'da yer almaktadır.

3.2 Metin Ön İşleme

Metin ön işleme aşaması, metinlerin olabildiğince dile bağımlı faktörlerden arındırılmasını ve dilin yapısal olarak sınırlarının belirlemesini sağlar (Wang Y. ve Wang X. J., 2005). Bu aşama, etkisiz kelimelerin atılması (İng. Stop word removal) ve kelimelerin eklerinden arındırılarak kök halinin elde edilmesi (İng. Stemming) işlemlerini içermektedir (Lee ve Chan, 2006).

Etkisiz kelimeler (İng. Stop words), bir dilde sık kullanılan zarf, zamir ve bağlaç gibi kelimelerdir. Bu kelimeler, metin dokümanlarında çok sayıda geçtiğinden dolayı metin sınıflandırma problemlerine pozitif bir katkıda bulunamamaktadır. Metin dokümanlarından bu sözcüklerin atılması, dil bilimciler tarafından belirlenmiş etkisiz kelime listesi kullanılarak yapılabilir. Etkisiz kelimelerin atılması, sınıflandırmada kullanılacak özniteliklerin boyutunu azaltacağından öznitelik seçme yöntemi olarak da düşünülebilir (Liu ve diğ, 2003).

Kelimelerin kök halinin elde edilmesi işlemi de öznitelik uzayının boyutunu azaltan bir yöntemdir. Bu yöntemle; aynı kök formuna sahip, yapım eki veya çekim eki almış kelimeler aldıkları eklerden arındırılarak kök formunda tek bir kelime haline getirilir.

Kelimeleri kök formuna indirgeme, Türkçe için büyük kazananımlar sağlamaktadır. Türkçe sondan eklemeli bir dildir ve kelimelerin kök halinin elde edilmesiyle, "okul", "okudu", "okuyacak", "okutman" gibi çekim eki veya yapım eki almış kelimeler, bu kelimelerin kök formu olan "oku" kelimesi altında birleşecektir.

3.3 Doküman Gösterimi

Metin dokümanlarından ön işleme yoluyla öznitelikler elde edildikten sonra, bu öznitelikler kullanılarak; dokümanlar n boyutlu öznitelik vektörleriyle ifade edilmektedir. n sayısı ön işleme sonucu elde edilen öznitelik sayısını göstermektedir. Oluşturulan öznitelik vektörlerinin yapay öğrenme yöntemleriyle sorunsuz kullanılabilmesi ve metinlerin içerdiği anlamın kaybolmaması için uygun doküman gösterimi seçimi metin sınıflandırmada önemli bir aşamadır. Seçilen doküman gösterim yöntemi farklı sınıflara ait metin dokümanlarının birbirinden kolayca ayırt edilmesini sağlayacak nitelikte olmalıdır. Metin sınıflandırma probleminde dokümanlar genellikle Kelimeler Çantası (İng. Bag of Words) gösterimi kullanılarak öznitelik vektörlerine dönüştürülür. Dönüştürülen öznitelik vektörlerinin her bir elemanı bütün doküman koleksiyonun işlenmesi sonucu elde edilen tekil kelimelerdir. Kelimeler Çantası gösteriminde kelimelerin metinde geçme sırası göz önünde bulundurulmaz (Schumaker ve Chan, 2009).

Doküman koleksiyonundaki her bir doküman için Kelimeler Çantası gösterimiyle bir öznitelik vektörü oluşturulduktan sonra bu vektörlerin her bir elemanına değer ataması yapılması gerekmektedir. Öznitelik vektörünün her bir boyutu bir kelimeyi temsil ettiğinden, eğer o kelime metin dokümanında geçiyorsa, öznitelik vektöründe kelimeye ait boyuta değer olarak 1, dokümanda geçmiyorsa değer olarak 0 atanır.

Öznitelik vektörlerine değer ataması yaparken, dokümanda geçen kelimenin o doküman için ne kadar önemli olduğunu gösteren ağırlıklandırma ölçütleri de kullanılmaktadır. Bu ölçütler terim sıklığı (İng. Term frequency) ve ters doküman sıklığıdır (İng. Inverse document frequency).

Terim sıklığı

Terim sıklığı, her bir kelimenin bir doküman için ne kadar öneme sahip olduğunu o kelimelerin dokümanda kaç defa geçtiğine bakarak hesaplayan ölçüttür.t kelimesinin d dokümanı için terim sıklığı değeri, $tf(t,\mathrm{d})$, aşağıdaki şekilde hesaplanır.

$$tf(t,\mathrm{d}) = \begin{cases} 1 + \log\#(t,d) & , \quad eğer \quad \#(t,d) > 0 \\ 0 & , \quad diğer\ durumda \end{cases} \tag{3.1}$$

$\#(t,d)$, t kelimesinin d dokümanında geçme sayısını gösterir. Terim sıklığı, doküman bazında hesaplandığından, dokümanlardaki yüksek terim sıklığına sahip

kelimeler düşük terim sıklığına sahip kelimelere göre ön plana çıkmaktadır. Ancak, bir kelime bütün dokümanlarda yüksek terim sıklığına sahip ise bu durum kelimenin ayırt edici nitelikte olmadığını gösterir (Manning ve diğ, 2008). Bundan dolayı bu tür kelimeler sınıflandırıcı eğitiminde ihmal edilmektedir.

Ters Doküman Sıklığı

Terim sıklığı bir kelimenin bir doküman içerisinde geçme sayısını göz önünde bulundururken, ters doküman sıklığı kelimenin doküman koleksiyonunda ne kadar sayıda geçtiğiyle ilgilenir. Ters doküman sıklığının temel felsefesi, doküman koleksiyonlarında nadir olarak bulunan kelimelere daha fazla ağırlık vermektir. Dokümanlarda geçen her bir kelimenin önem derecesi bu kelimeyi içeren doküman sayısıyla ters orantılı olarak hesaplanmaktadır (Manning ve diğ, 2008). t kelimesine ait ters doküman sıklığı değeri, $idf(t)$, Eşitlik 3.2'deki bağıntı kullanılarak hesaplanabilir.

$$idf(t) = \log \frac{|N|}{\#N(t)} \tag{3.2}$$

Bağıntıda $|N|$, doküman koleksiyonundaki toplam doküman sayısını; $\#N(t)$ ise t kelimesini içeren doküman sayısını göstermektedir.

3.4 Öznitelik Seçimi ve Öznitelik Seçiminde Kullanılan Yöntemler

Metin sınıflandırma işleminde dokümanlar genellikle binlerce öznitelikle ifade edilir ve bu özniteliklerin büyük bir kısmı önemli bilgi içermez. Metin ön işleme yöntemleri kullanıldığında öznitelik uzayının boyutunda azalma olmasına rağmen, ön işlemeyle elde edilen öznitelikler sınıflandırıcı eğitiminde hafıza ve işlemci yetersizliği gibi sorunlar ortaya çıkarmaktadır. Ortaya çıkan kaynak sorunlarının önüne geçmek ve sınıflandırma performansını arttırmak için 2 farklı boyut indirgeme tekniği kullanılmaktadır. Bu teknikler, öznitelik seçimi (İng. Feature selection) ve öznitelik çıkartımı (İng. Feature extraction) 'dır.

3.4.1 Öznitelik seçimi ve öznitelik çıkartımı

Öznitelik uzayının yüksek boyutlu olmasının negatif etkilerini ve veri seyrekliğini ortadan kaldırmak için öznitelik uzayı üzerinde boyut indirgenmesi yapılmalıdır.

Boyut indirgenmesi, öznitelik seçimi veya öznitelik çıkartımı teknikleri kullanılarak yapılmaktadır.

Öznitelik seçimi, mevcut özniteliklerden yeni bir öznitelik alt kümesi oluşturarak yapılır. Tüm öznitelikler arasından daha önce belirlenmiş bir ölçüte göre en yüksek değere sahip belirli sayıda öznitelik seçilir (Montanes ve diğ, 2003). Öznitelik çıkarımı ise, öznitelik kümesinden asıl öznitelikleri almak yerine bu özniteliklerin çeşitli yöntemlerle dönüştürülmüş halini alarak boyut indirgemesi yapmaktadır (Guyon ve diğ, 2006).

Öznitelik seçimi metin sınıflandırma problemlerinde sıklıkla kullanılır. Bu sınıflandırma problemlerindeki öznitelikler genellikle kelimelerden oluşur ve sayısı binlercedir. Öznitelik sayısının fazla olması sınıflandırıcı performansını negatif yönde etkilemektedir. Öznitelik seçimiyle çok sayıda öznitelikten veri kümesini en iyi şekilde temsil eden belirli sayıda öznitelik seçilecektir. Öznitelik seçimi ile gürültü olarak yer alan öznitelikler elenecek ve sınıflandırıcı performansı yükselecektir. Seçilen özniteliklerin metnin içeriğini yansıtması sınıflandırma performansı açısından önem taşır.

Metin sınıflandırma uygulamalarında öznitelik seçimi yapılırken hangi özniteliğin "iyi" olduğunun tespit edilmesi sınıflandırıcı performansı için kritiktir. Öznitelik seçiminin temel amacı öznitelik vektörlerinin boyutunun indirgenmesi olmasına rağmen; boyut indirgeme işlemiyle birlikte seçilen özniteliklerin sınıflandırıcının performansını arttırması gerekliliği de göz önünde bulundurulmalıdır. İyi bir öznitelik seçimi hem bilgi içermeyen özniteliklerin ihmal edilmesini hem de sınıflandırıcı eğitilirken işlem kaynaklarının daha verimli kullanılmasını sağlamalıdır. Ayrıca öznitelik seçimiyle sınıflandırma performansının artmasıyla birlikte sınıflandırıcının karmaşıklığı da azalmaktadır.

3.4.2 Öznitelik seçimi yaklaşımları

Tez çalışmasının bu bölümünde öznitelik seçiminin temelini oluşturan kavramlardan söz edilecektir. İlk aşamada Gözetimli (İng. Supervised) ve Gözetimsiz (İng. Unsupervised) öznitelik seçimi kavramları açılanacak ve farklarına değinilecek, sonraki aşamada sıklıkla kullanılan gözetimli öznitelik seçimi yaklaşımları açıklanacaktır. Son aşamada ise gözetimli öznitelik seçiminde kullanılan ölçütlerden bahsedilecektir.

Gözetimli ve Gözetimsiz Öznitelik Seçimi

Öznitelik seçme yöntemleri genellikle metin sınıflandırma problemlerine uygulanmakla beraber, nadiren herhangi bir sınıf etiketine sahip olmayan metin kümeleme problemlerine de uygulanabilir. Bu problemlerde kullanılan kümeleme algoritmalarının performansı, veri seyrekliği ve verinin yüksek boyutlu olmasıyla doğrudan bağlantılıdır. Kümeleme performansını artması ise öznitelik uzayının boyutunu indirgemeyle gerçekleşebilir.

Yapay öğrenmede girdilerin sınıf etiketine sahip olmasına bağlı olarak, öznitelik seçimi iki yöntemle yapılmaktadır. Bunlar, gözetimli öznitelik seçimi ve gözetimsiz öznitelik seçimidir. Metin sınıflandırma problemlerinde boyut indirgenmesi, sınıf etiketine sahip olmayan metin dokümanlarının kümelenmesi görevlerinde gözetimsiz öznitelik seçimi ile sınıf etiketi kullanılarak yapılacak sınıflandırma görevlerinde ise gözetimli öznitelik seçimi ile yapılabilir (Liu ve diğ, 2003).

Tez çalışması sınıflandırma problemini içerdiği için gözetimli öznitelik seçiminin ayrıntıları Bölüm 3.4.3 ‘te ele alınacaktır. Gözetimsiz öznitelik seçimi ile ilgili ayrıntılı bilgiye Guyon ve Elisseeff (2003) ‘in yaptığı çalışmadan ulaşılabilir.

Süzgeç Yaklaşımı ve Dürümcü Yaklaşım

Gözetimli öznitelik seçiminde süzgeç yaklaşımı (İng. Filter approach) ve dürümcü yaklaşım (İng. Wrapper approach) olmak üzere temel 2 yaklaşım bulunmaktadır. Süzgeç yaklaşımında öznitelikler sınıflandırıcıdan bağımsız olarak puanlanır ve elde edilen puanlara göre öznitelik seçimi yapılır. Dürümcü yaklaşımda ise öznitelik alt kümeleri ve sınıflandırıcı kullanılarak öznitelik seçimi yapılmaktadır (Kohavi ve John, 1998). Dürümcü yaklaşım, süzgeç yaklaşımına göre daha iyi performans göstermesine rağmen, öznitelik sayısının binlerle ifade edildiği sınıflandırma problemlerinde işlem yükü yaratmaktadır.

Süzgeç yaklaşımında öznitelik seçimi sınıflandırıcının kullandığı algoritmadan bağımsız olarak yapılır. Bu nedenle süzgeç yaklaşımı, sınıflandırıcı eğitiminden önceki ön işleme aşaması olarak değerlendirilebilir. Dürümcü yaklaşımda ise öznitelik alt kümesini seçmek için sınıflandırıcı algoritması değerlendirme fonksiyonu olarak kullanılır. Çok büyük verilerde dürümcü yaklaşımı uygulamak süzgeç yaklaşımına göre daha zordur. Bu tip veri kümelerinde öznitelik seçimi sınıflandırıcıdan bağımsız olması gerektiğinde süzgeç yaklaşımı uygulanır.

Fazla sayıda özniteliğe sahip metin sınıflandırma problemlerinde süzgeç yaklaşımı kullanılır (Yang ve Pedersen, 1997).

3.4.3 Gözetimli öznitelik seçiminde kullanılan yöntemler

Metin sınıflandırma problemlerinde gözetimli öznitelik seçimi, metin dokümanlarında geçen kelimeler için seçilen bir ölçüte göre puan hesabı yapılması ve yüksek puana sahip belirli sayıda kelimelerin seçilmesi ile gerçekleştirilir. Literatürde, öznitelik seçiminde kullanılan birçok ölçüt mevcuttur. Metin sınıflandırma probleminde ise, doküman sıklığı (İng. Document frequency), bilgi kazanımı (İng. Information gain), karşılıklı bilgi (İng. Mutual information) ve Ki-Kare istatistiği (İng. Chi-Square statistics) gibi geleneksel ölçütler kullanılır (Yang ve Pedersen, 1997).

Doküman Sıklığı

Doküman sıklığı bir kelimenin geçtiği doküman sayısını gösterir. Bu ölçütte, her bir özniteliğin (kelimenin) doküman koleksiyonunda kaç defa geçtiği hesaplanır ve doküman sıklığı önceden belirlenen eşik değerinin altında kalan öznitelikler ihmal edilir. Doküman sıklığı boyut indirgemede kullanılan basit bir yöntem olmasına karşın bilgi içeren özniteliklerin seçiminde etkili değildir.

Bilgi Kazanımı

Yapay öğrenme alanında, bilgi kazanımı bir kelimenin iyiliğini ölçmede kullanılan bir kıstastır. Bilgi kazanımı Shannon'un bilgi teorisini (İng. Shannon's information theory) temel alır.

Bilgi kazanımı kavramı düzensizlik (İng. Entropy) kavramı ile ilişkilidir ve basitçe bilgi kazanımı, düzensizliğin tersi olarak ifade edilir. Bilgi kazanımı 0 ile 1 arasında değer almaktadır ve sınıflandırma süreci, kullanılan öznitelik ile yapıldığında kaç bitlik bilgi kazanılabileceğini gösterir. Örneğin; sınıflandırma sürecinde bir öznitelik her sınıf için farklı bir değer alıyorsa, bu durumda bu öznitelik için düzensizlik değeri 0 ve bilgi kazanımı değeri 1 çıkacaktır. Bunun anlamı kullanılan öznitelik ile sınıflar arasında birebir bağlantı kurulabildiğidir. Buna karşılık elimizdeki öznitelik, sınıflardan ne kadar bağımsızsa bilgi kazanımı da o kadar düşük çıkacaktır (Cover ve Thomas, 1991). Bu yöntemle öznitelik seçimi yapılırken her bir öznitelik için bilgi kazanımı değeri hesaplanır.

Hesaplanan kazanım değerleri önceden belirlenen eşik değerinin altında kalıyorsa, bu öznitelikler sınıflandırma sürecinde ihmal edilir.

Karşılıklı Bilgi

Karşılıklı bilgi, iki bağımsız değişken arasındaki ortak bağlılığın miktarıdır. Yapay öğrenme uygulamalarında karşılıklı bilgi öznitelik vektörünün her bir boyutu ile bu öznitelik vektörüne atanan sınıf etiketleri arasında bağlılığı ölçer. Karşılıklı bilginin zayıf yönü ise özniteliklerin marjinal olasılıklarından güçlü bir şekilde etkilenmesidir. Karşılıklı bilgi yöntemiyle her bir özniteliğin sınıf etiketlerine karşılıklı bilgi tabanlı benzerlik oranı olağanlık çizelgesi (İng. Contingency table) ile hesaplanır ve yüksek benzerlik oranına sahip özniteliklerden başlanarak belirli sayıda öznitelik seçimi yapılır (Manning ve diğ, 2008).

Ki-Kare (χ2) İstatistiği

Ki-Kare (χ2) bağımsızlık testi, iki ya da daha fazla sınıfa sahip iki nitel değişken arasında bağımsızlık olup olmadığını inceler. χ2 testi ile bu değişkenler arasında ilişki olup olmadığı olağanlık çizelgesi (İng. Contingency table) yardımı ile bulunur. Bazı zamanlarda ilişkinin olup olmadığının tespiti yeterli değildir. Bunun yanı sıra aradaki ilişkinin derecesini (gücünü) de bilmek gereklidir. Bu da olağanlık çizelgeleri kullanılarak bulunan olağanlık katsayısı ile mümkündür. Bu katsayı, değişkenler arasındaki ilişkinin düzeyini ölçemeye yarar. Olağanlık katsayısı ilişkinin olmadığı durumlarda "0", çok yüksek ilişki olduğunda ise "1"e çok yakın bir değer çıkar (A.Ö.F Ders Notu, 2010).

Metin sınıflandırma problemlerinde her sınıf için her bir öznitelik ile o sınıf arasındaki Ki-Kare istatistiği hesaplanır. Karşılıklı bilgi ile Ki-Kare istatistiği yöntemi arasındaki temel fark Ki-Kare istatistiğinin hesaplanması sonucu elde edilen değerlerin normalize edilmiş halde olmasıdır. Bu da aynı sınıfta var olan özniteliklerin karşılaştırılmasında kolaylık sağlamaktadır.

Metin sınıflandırma problemlerin de kullanılan bu öznitelik seçme yöntemlerine ait matematiksel formüller Çizelge 3.1'te yer almaktadır. Formüllerde t_k metin dokümanında geçen k. kelimeyi, c_i ise dokümanın i. sınıfa ait olduğunu belirtmektedir. $\overline{t_k}$ ise k. kelimenin yer almadığı dokümanları göstermektedir. $\overline{c_i}$ ise i. sınıf haricindeki diğer sınıflara ait dokümanları ifade etmektedir. N ise koleksiyondaki doküman sayısını göstermektedir. Formüllerdeki $P(\ )$ ifadesi ise

olasılık anlamındadır. Örneğin $P(t_k, c_i)$, k. kelimenin i. sınıfa ait dokümanlarda geçme olasılığını göstermektedir.

Çizelge 3.1: Metin sınıflandırma problemlerinde kullanılan öznitelik seçme yöntemleri (Yang ve Pedersen, 1997).

Yöntem	Gösterimi	Formül
Doküman Sıklığı	$\#(t_k, c_i)$	$P(t_k, c_i)$
Bilgi Kazanımı	$IG(t_k, c_i)$	$P(t_k, c_i) \log \frac{P(t_k, c_i)}{P(t_k)P(c_i)}$ $+P(\overline{t_k}, c_i) \log \frac{P(\overline{t_k}, c_i)}{P(\overline{t_k})P(c_i)}$
Karşılıklı Bilgi	$MI(t_k, C)$	$\sum_{t_k \in T} \sum_{c_i \in C} P(t_k, c_i) \log \frac{P(t_k, c_i)}{P(t_k)P(c_i)}$
Ki-Kare İstatistiği	$\chi 2(t_k, c_i)$	$\frac{\mathrm{N}[P(t_k, c_i), P(\overline{t_k}, \overline{c_i}) - P(t_k, \overline{c_i})P(\overline{t_k}, c_i)]^2}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})}$

Çizelge 3.1'de gösterilen bu 4 yöntemin performansları Yang ve Pedersen (1997) tarafından yapılan metin sınıflandırma çalışmasında karşılaştırılmıştır. Yang ve Pedersen yaptıkları istatistiksel metin sınıflandırma çalışmasında Ki-kare istatistiği, bilgi kazanımı, karşılıklı bilgi ve doküman sıklığı öznitelik seçim yöntemlerinin yaygın kullanılan kelimelere önem verme, dokümanların sınıflarından faydalanma ve dokümanda kelimelerin var olup olmaması gibi 3 kıstasa bakarak performansı karşılatırmıştır. Karşılaştırma yaparken En Yakın K Komşu ve Doğrusal En Küçük Kare Uydurma sınıflandırıcılarını kullanılmıştır. Çalışmada, bilgi kazanımı ve Ki-kare istatistiği yöntemlerinin doküman sınıflandırmada ve öznitelikleri indirgemede birbirine yakın performans elde ettikleri anlatılmıştır. Bu iki yöntemin karşılıklı bilgi yöntemine üstünlük sağladığı bunun nedenin ise karşılıklı bilgi yönteminin nadir geçen kelimeleri ön plana çıkarmaya çalışmaya meyilli olduğu söylenmiştir.

Rogati ve Yang (2002) tarafından yapılan çalışmada ise bilgi kazanımı, Ki-Kare istatistiği ve doküman sıklığı öznitelik seçme yöntemlerinin performansları 4 farklı sınıflandırıcı kullanılarak karşılaştırılmıştır. Bu sınıflandırıcılar Naive Bayes, Roccio

Stili, En Yakın K Komşu ve Destek Vektör Makineleridir. Elde edilen sonuçlar Ki-Kare istatistiğinin bilgi kazanımı veya doküman sıklığı yöntemleriyle birlikte kullanıldığında en iyi sonucu verdiği ve Destek Vektör Makinelerinin öznitelik seçimine en az duyarlı sınıflandırıcı olduğudur.

3.5 Sınıflandırıcı Öğrenmesi

Sınıflandırıcı öğrenmesi, öznitelik seçimi aşamasından sonra gelmektedir. Boyutu indirgenmiş öznitelik vektörlerinden oluşan veri kümesinden sınıflandırıcı eğitimi için eğitim örnekleri alınmaktadır. Sınıflandırıcılar, alınan eğitim örneklerini yapay öğrenme algoritmalarıyla kullanarak problemdeki her bir sınıf için gerekli ilişkisel bilgiyi öğrenir.

Metin sınıflandırma çalışmalarında yapay öğrenme algoritmaları kullanılmadan önce sınıflandırma işlemleri kural tabanlı sistemler kullanılarak yapılmıştır. Uzman kişiler tarafından tanımlanan kurallar kullanılarak metin dokümanları sınıflara ayrılmış, bu durum yoğun iş yükü, maliyet ve zaman kaybı getirirmiştir.90'lı yıllarda yapay öğrenme felsefesinin ortaya çıkmasıyla kural tabanlı öğrenme popülerliğini yitirmiştir. Tümevarımsal bir süreç olan yapay öğrenme ile önceden sınıfları belli olan metin dokümanlarından metinleri otomatik kategorize eden sınıflandırıcılar eğitilmiştir. Yapay öğrenme algoritmalarının sağladığı avantajlar ise sınıflandırma başarısının uzman kişiler tarafından yapılana yakın olması, uzman iş gücünden tasarruf sağlaması ve sınıflandırma işlemi yapılırken alanında uzman kişilere ihtiyacın en aza indirilmesi olarak söylenebilir.

Metin sınıflarıma çalışmalarında birçok yapay öğrenme algoritması kullanılmaktadır. Bu algoritmalar En yakın K komşu, Yapay Sinir Ağları, Destek Vektör Makineleri, Karar Ağaçları, Naive Bayes Sınıflandırıcısıdır (Yang ve Liu, 1999).

Yapılan tez çalışmasında sınıflandırıcı olarak Naive Bayes sınıflandırıcısı seçilmiştir. Bir sonraki başlıkta Naive Bayes sınıflandırıcısıyla ilgili detaylı bilgi verilecektir.

3.5.1 Naive Bayes (NB) sınıflandırıcısı

Naive Bayes sınıflandırıcısı metin sınıflandırma problemlerinde başarılı performans göstermektedir (McCallum ve Nigam, 1998). NB Sınıflandırıcısı Bayes Kuralını temel almaktadır. Bu sınıflandırıcı bir örneğin bir sınıfa ait olma olasılığını

hesaplamak için özniteliklere ait şartlı olasılıkları kullanarak olasılık modeli bulur. Naive Bayes modelinde bulunmak istenen olasılık değeri $p(c|w_1, \dots, w_n)$'dir. Bu olasılık değeri sonsal (İng. Posterior) olasılık olarak adlandırılmaktadır. Metin sınıflandırma problemi açısından düşünüldüğünde; bu olasılık gösteriminde her bir w, öznitelik vektöründe yer alan öznitelikleri yani metin dokümanında geçen kelimeleri gösterir. c ise metin dokümanının hangi sınıfa ait olduğunu belirtir. Naive Bayes sınıflandırıcısında veri kümesindeki örnekler kullanılarak her sınıf için sonsal olasılık değeri hesaplanır. Hangi sınıfın sonsal olasılık değeri yüksek ise bu örneklere o sınıfın ataması yapılır. Ancak yüksek boyuta sahip veri kümelerinde sonsal olasılık değerini direkt olarak hesaplamak zordur. Bu durumda hesaplama işlemini kolaylaştırmak için Bayes olasılık kuralı kullanılır (Alpaydın, 2010).

$$p(c|w_1, \dots, w_n) = \frac{p(c)p(w_1, \dots, w_n|c)}{p(w_1, \dots, w_n)} \quad \textbf{(3.3)}$$

Eşitlik 3.3'te $p(c)$ önsel (İng. Prior) olasılık değerini göstermektedir. Bu olasılık, eğitim kümesindeki sınıf dağılımına bakılarak hesaplanır. $p(w_1, \dots, w_n)$ olasılığı ise kanıt (İng.Evidence) olasılığı olarak bilinmektedir. Bu olasılık değerinin sınıflar üzerinde bağımlılığı olmadığı için sonsal olasılık değeri hesaplanırken kanıt olasılığı ihmal edilir. $p(w_1, \dots, w_n|c)$ olasılık değeri ise olabilirlik (İng. Likelihood) olasılığıdır ve bu olasılık değerinin hesaplanması önsel olasılığa göre daha zordur. Bunun nedeni ise olabilirlik olasılığındaki şartlı olasılığın bütün özniteliklere bağımlı olmasıdır. Bu olasılık değerinin hesaplanmasını kolaylaştırmak için Naive Bayes sınıflandırıcısı özniteliklerin şartlı bağımsızlığını göz önünde bulundurur. Bu, Naive Bayes sınıflandırıcısında her özniteliğin diğer tüm özniteliklerden şartlı bağımsız olduğu anlamındadır. Olabilirlik olasılığını hesaplarken Naive varsayımının nasıl kullanılacağı Eşitlik 3.4'te gösterilmiştir.

$$p(w_1, \dots, w_n|c) \sim p(c) \prod_{i=1}^{n} p(w_i|c) \quad \textbf{(3.4)}$$

Bu varsayımın etkileri Zhang (2004) tarafından incelenmiştir. Zhang, öznitelikler arasındaki bağımlılıklara bakarak Naive Bayes sınıflandırıcısının performansını incelemiştir. Bu çalışmada öznitelikler arasında güçlü bir bağımlılık olması durumunda bile NB sınıflandırıcısının tatmin edici bir performans ortaya koymuştur.

3.6 Sınıflandırıcı Performansını Değerlendirme Ölçütleri

Sınıflandırıcı eğitiminin tamamlanmasından sonraki aşama ise test verilerinin kullanılarak sınıflandırıcının performansının değerlendirilmesidir. Sınıflandırıcının etkinliğinin ölçülmesi için geleneksel bilgi getirimi yöntemlerinde kullanılan ölçütlerden faydalanılır. Bu ölçütler doğruluk oranı (İng. Accuracy), kesinlik (İng. Precision), duyarlılık (İng. Recall) ve F-Ölçütüdür. Kullanılan bu ölçütlerden doğruluk oranı hariç diğer ölçütler sınıflandırıcı performansını sınıf bazında değerlendirmektedir. Sınıflandırıcının genel performansını değerlendirmek için bu ölçütleri temel alarak oluşturulan makro-ortalama yöntemi bulunmaktadır. Makro-ortalama performansı hesaplanırken ilk önce her sınıf için performans ölçümü yapılmakta daha sonra sınıf performansı ölçümlerinin ortalaması alınmaktadır. Makro-ortalama her sınıfa eşit ağırlık vermektedir (Özgür ve diğ, 2005).

İki sınıf içeren sınıflandırma problemlerinde, örnekler pozitif veya negatif olmak üzere iki sınıfa ayrılmaktadır. Sınıflandırıcıların iki sınıfa ait bu örnekleri nasıl sınıflandırdıkları hata matrisi (İng. Confusion matrix) denilen çizelge yapısıyla gösterilir (Yang, 1997). Hata matrisi Çizelge 3.2' de gösterilmektedir. Hata matrisi 4 kategoriye sahiptir. Doğru pozitif (dp) (İng. True positive), pozitif sınıf etiketine sahip olan örneklerin pozitif olarak doğru sınıflandırıldığını gösterir. Yanlış pozitif (yp) (İng. False positive), negatif sınıf etiketine sahip örneklerin pozitif olarak sınıflandırıldığını anlatır. Doğru negatif (dn) (İng. True negative) negatif sınıfındaki örneklerin negatif olarak sınıflandırılmasıdır ve bu istenen bir durumdur. Son olarak yanlış negatif (yn) (İng. False negative) ise pozitif sınıfında olan örneklerin negatif sınıfındaymış gibi sınıflandırılmasını anlatır. Hata matrisindeki bu 4 kategori kullanılarak kesinlik, duyarlılık, doğruluk oranı ve F-Ölçütü gibi performans ölçütleri hesaplanabilmektedir (Davis ve Goadrich, 2006).

Yapay öğrenme uygulamalarında performans değerlendirme ölçütü olarak genellikle doğruluk oranı kullanılmaktadır. Doğruluk oranı (DO), doğru olarak sınıflandırılan örnek sayısının tüm örnek sayısına bölünmesiyle bulunur. İyi bir sınıflandırıcı yüksek doğruluk oranına sahip olmalıdır. Doğruluk oranının hata matrisi değerleri kullanılarak hesaplanması Eşitlik 3.5' te gösterilmiştir.

$$\mathrm{DO}=\frac{dp+dn}{dp+dn+yp+yn} \tag{3.5}$$

Çizelge 3.2: İki sınıf için oluşturulan hata matrisi.

Gerçek Sınıf/Tahmin Edilen	**Pozitif**	**Negatif**
Pozitif	dp	yn
Negatif	yp	dn

Eğer sınıflandırma sürecinde kullanılan veri kümesi dengesiz sınıf dağılımına sahip ise bu durumda doğruluk oranının değerlendirme ölçütü olarak kullanılması yanıltıcı olabilmektedir. Örneğin; pozitif sınıftaki örnek sayısının negatif sınıftaki örnek sayısına oranla çok az olduğu durumda, bütün test örneklerini negatif olarak sınıflandırdığımızda sınıflandırma işleminin doğruluk oranı yüksek çıkacaktır. Bu nedenle, veri kümesinde bir sınıfa ait örnek sayısı diğer sınıflara ait örnek sayısına baskın olduğunda sadece doğruluk oranının değerlendirme ölçütü olarak kullanılması uygun değildir. Sınıflandırıcının her sınıfı ayırt etme başarısının ölçülebilmesi için doğruluk oranının yanında kesinlik, duyarlılık ve ikisinin bileşimi olan F-Ölçütü ölçütlerinin kullanılması gerekir. Kesinlik ve duyarlılık ölçütleri pozitif olarak sınıflandırmış örnekleri temel alır. Kesinlik pozitif olarak sınıflandırılmış olan örneklerden kaç tanesinin gerçekten pozitif olduğunu, duyarlılık ise toplam pozitif örnek sayısından kaç tanesinin doğru olarak tanındığını ölçer.

Diğer bir performans değerlendirme kıstası da F-Ölçütü'dür. F-Ölçütü, kesinlik ve duyarlılık ölçütlerinin bileşiminden oluşur. F-Ölçütü, kesinlik ve duyarlılık değerlerinin ağırlıklı harmonik ortalamasının alınması sonucu hesaplanır. Yüksek F-Ölçütü değeri iyi bir sınıflandırma performansı anlamına gelmektedir. Kesinlik, duyarlılık ve F-Ölçütünün hata matrisi elemanları ile bulunması gösteren formüller Çizelge 3.3 'de gösterilmiştir.

İki sınıf içeren sınıflandırma problemlerinde kesinlik, duyarlılık ve F-Ölçütü başarı değerlendirmesini pozitif örnekler üzerinden yapar. Çoklu sınıf içeren yapay öğrenme problemlerinde ise her bir sınıf ayrı olarak ele alınır ve bu sınıfa ait örnekler pozitif olarak geriye kalan tüm örnekler negatif olarak ayrılır. Böylece her bir sınıf için bu ölçütler cinsinden performans değeri bulunur. Sınıflandırıcının genel performansı ise her sınıf performans değerlerinin Makro-ortalaması alınarak bulunur.

Çizelge 3.3: Kesinlik, duyarlılık ve F-Ölçütünün formülleri.

Ölçüt	Formül
Kesinlik(K)	$\frac{dp}{dp + yp}$
Duyarlılık(D)	$\frac{dp}{dp + yn}$
F-Ölçütü	$\frac{2*K*D}{K+D}$

Dengesiz dağılımlı sınıflandırma problemlerinin başarımının değerlendirilmesinde Eğri Altında Kalan Alan (İng. Area Under Curve-AUC)'da kullanılan diğer bir ölçüttür (Chawla ve diğ, 2004).

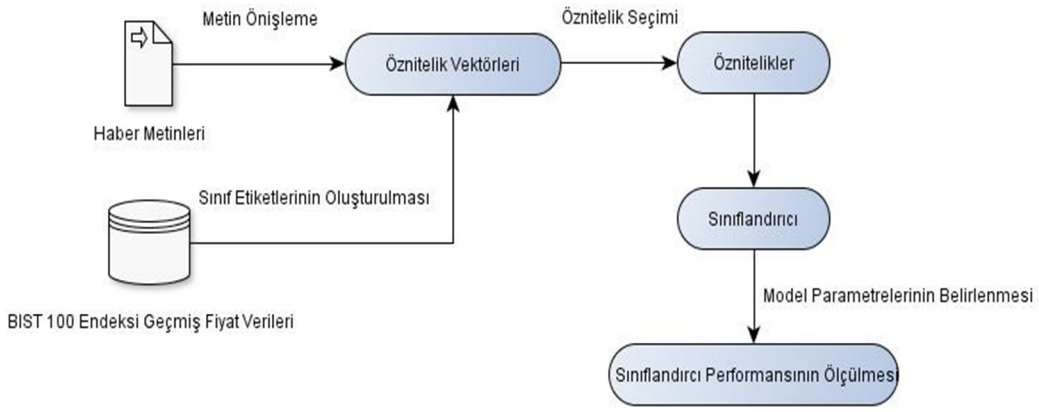
3.7 Bölüm Özeti

Tez çalışmasının bu bölümünde metin sınıflandırma problemlerinde var olan süreçlerden bahsedilmiştir. Bu sürecin ilk aşaması metin dokümanlarının ön işlenmesidir. Ön işleme aşamasından sonra metin dokümanları, doküman gösterimleri kullanılarak öznitelik vektörlerine dönüştürülür. Sonraki aşamada ise öznitelik seçme yöntemleri ile bilgi içermeyen öznitelikler ihmal edilir ve öznitelik vektörlerinin boyutu indirgenir. En son aşamada ise boyut indirgemesi yapılan öznitelik vektörleri metin sınıflandırıcısına aktarılır ve sınıflandırıcının başarısı değerlendirilir.

Bir sonraki bölümde ise tez çalışmasının araştırma sürecinden söz edilecektir.

4. HABER METİNLERİ İLE BIST 100 ENDEKSİ YÖN TAHMİNİ UYGULMASI

Bu bölümde tez çalışmasının uygulama sürecinden bahsedilecektir. Çalışmanın genel süreci ve sıralı adımları Şekil 4.1'de gösterilmektedir. Çalışmanın ilk adımını veri toplama aşaması oluşturmaktadır. Bu aşamada finansal web sitelerinden ekonomi ile ilgili haber metinleri, Borsa İstanbul'un resmi web sitesinden ise BIST 100 endeksine ait geçmiş fiyat verileri alınacaktır. Sonraki adımda, toplanan haber dokümanları ön işlemeye tabi tutulacak ve bu dokümanlar öznitelik vektörlerine dönüştürülecektir. Geçmiş fiyat verileri de haber dokümanlarının endeks fiyatı üzerindeki etkisini gösteren sınıf etiketlerinin oluşturulmasında kullanılacaktır. Öznitelik vektörlerinin oluşturulmasından sonra öznitelik seçme yöntemleri kullanılarak bilgi içeren öznitelikler bulunacak ve seçilen öznitelikler kullanılarak öznitelik uzayından yeni öznitelik alt kümeleri oluşturulacaktır. Son aşamada ise oluşturulmuş öznitelik alt kümelerinden alınan eğitim örnekleriyle sınıflandırıcı eğitimi gerçekleştirilecek ve test örnekleriyle de sınıflandırma performansı değerlendirilecektir.



Şekil 4.1: Tez çalışmasının genel süreci.

Şekil 4.1'de gösterilen işlemleri yazılımsal olarak gerçekleştirebilmek için Python, JAVA ve MATLAB programlama dillerinden faydalanılmıştır.

Python dili kullanılarak yazılan betikler ile web sitelerinden haber metinleri ve fiyat verileri toplanmıştır. Alınan haber metinlerinin işlenmesi ve bu metinlerden tekil kelimelerin elde edilmesi JAVA programlama dilinde yazılmış paketler kullanılarak yapılmıştır. MATLAB programlama dili ise elde edilen tekil kelimelerin kullanılarak haber dokümanlarından öznitelik vektörlerinin oluşturulması aşamasında kullanılmıştır. Öznitelik seçimi, sınıflandırıcı eğitimi ve sınıflandırıcı performansının ölçülmesi ise WEKA Veri Madenciliği Yazılımı ile yapılmıştır (Hall ve diğ, 2009).

Bölümün devamında araştırma sürecinin aşamaları ayrıntılarıyla ele alınacaktır. Bölüm 4.1'de Veri Toplama aşamasından bahsedilecektir. Bölüm 4.2'de Veri Kümesinin Oluşturulması anlatılacaktır. Öznitelik Seçimi Bölüm 4.3'te, Sınıflandırıcının Eğitimi Bölüm 4.4'te verilecektir. Bölüm 4.5'te ise Yapılan Sınıflandırma Deneyleri ve Alınan Sonuçlar gösterilecektir.

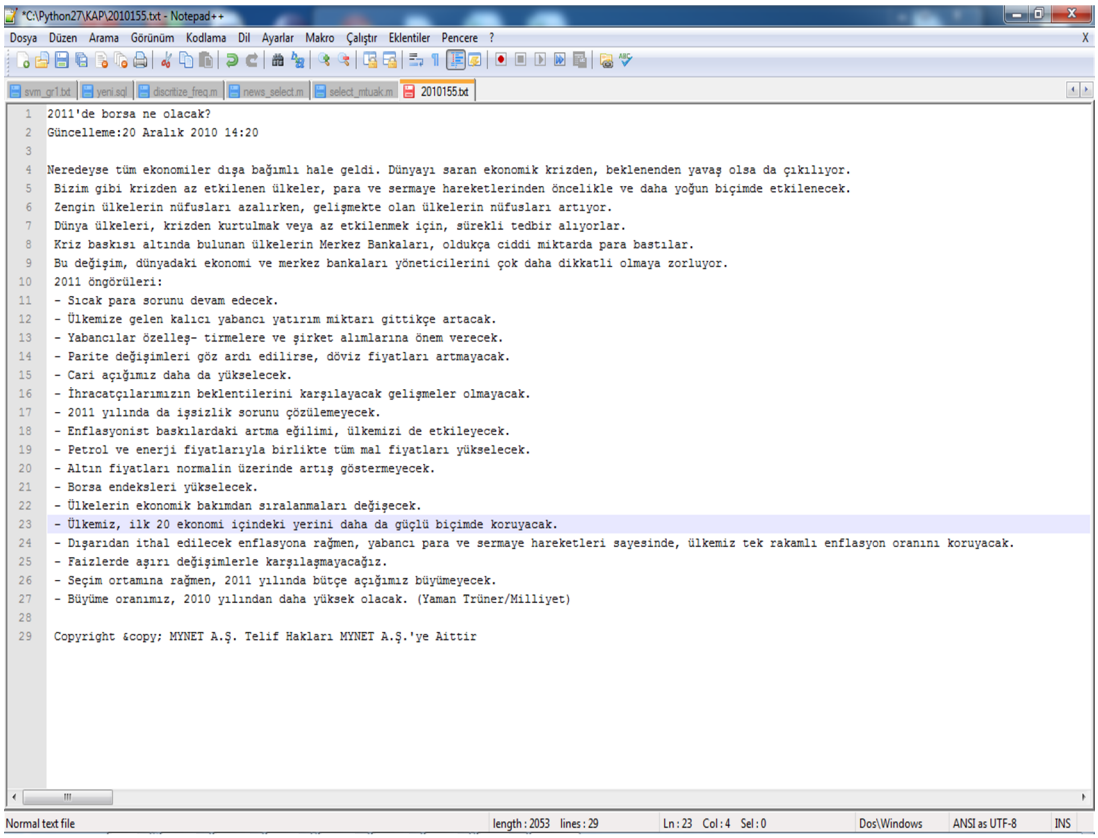
4.1 Veri Toplanması

Tez çalışmasının amacı finansal haber dokümanlarının BIST 100 endeksi açılış fiyatı üzerindeki etkilerini araştırmak olduğundan veri toplama aşamasında iki çeşit veriye ihtiyaç duyulmaktadır. Bunlar finansal haber metinleri ve BIST 100 endeksine ait fiyat verileridir.

Çalışmada finansal haber metinlerinin toplandığı iki tür haber kaynağı bulunmaktadır. Bu kaynaklardan birisi Borsa İstanbul'daki şirketlerin resmi bildirimlerinin yayınlandığı Kamu Aydınlatma Platformu (KAP)'nun internet sitesidir. KAP, sermaye piyasası ve Borsa mevzuatı uyarınca kamuya açıklanması gerekli bildirimlerin elektronik imzalı olarak iletildiği ve kamuya duyurulduğu elektronik sistemdir. İşletimi Borsa İstanbul-Kotasyon Müdürlüğü-KAP İşletim Birimi tarafından yürütülmekte olan sistem, 7/24 esasına göre çalışmaktadır. Uygulama, tüm Türkiye'ye yayılmış 550'yi aşkın şirketi ve 2500'ü aşkın kullanıcıyı kapsamaktadır. Sistem, tüm kesimlerin Borsa İstanbul şirketleri ile ilgili doğru, anlaşılır, tam bilgiye, internet üzerinden eş anlı ve düşük maliyetle erişebilmelerine imkân tanıyacak şekilde tasarlanmıştır. Ayrıca, geçmişe dönük bilgilere de kolay ve düşük maliyetle erişim imkânı sağlayan elektronik bir arşiv niteliğindedir (Url-2). KAP'tan alınan haber dokümanları, şirketlerin özel durum bildirimlerini, 3 aylık ve yıllık kar bildirimlerini, bilançolarını, yeni bir iş ilişkisine başlama veya sipariş alma haberlerini içermektedir.

İkinci haber kaynağı olarak da Türkiye'nin iki önemli finans internet sitesi Bigpara (Url-3) ve Mynet Finans (Url-4) seçilmiştir. Bu internet sitelerinden alınan haber dokümanları ise genel ekonomi haberlerini, politika haberlerini, küresel ekonomi ile ilgili gelişmeleri, şirketlerin basın bültenlerini, şirketler ile ilgili basında çıkan haberleri, ekonomi yazarlarının araştırma raporlarını içermektedir.

Bahsedilen iki haber kaynağı taranarak 2010 ve 2011 yıllarında yayınlanmış 111587 adet haber dokümanı toplanmıştır. Haber metinlerinde içerik olarak haberin yayınlanma tarihi ve saati, haberin başlığı ve haberin ayrıntılarının açıklandığı paragraflar yer almaktadır. Haber dokümanlarına ait örnek ekran çıktısı Şekil 4.2' de görülmektedir.



Şekil 4.2: Haber dokümanına ait ekran çıktısı.

Haber dokümanı koleksiyonunun oluşturulmasından sonra, haber metinlerinin BIST 100 Endeksi'nin açılış fiyatına hangi yönde etki ettiğini gösterecek sınıf etiketlerinin oluşturulması için BIST 100 Endeksinin 2010 ve 2011 yıllarındaki günlük açılış fiyatı, kapanış fiyatı, işlem hacim verileri Borsa İstanbul internet sitesinden elde edilmiştir.

Haber dokümanlarının ve fiyat verilerinin toplanmasından sonra veri kümesi oluşturma aşamasına geçilmiştir.

4.2 Veri Kümesinin Oluşturulması

Toplanan haber dokümanları yapılandırmamış düzendedir ve bu dokümanların sınıflandırma sürecinde kullanılabilmesi için ön işlemeye tabi tutulması gerekmektedir. Ön işlemenin ilk aşaması Türkçe'deki etkisiz kelimelerin haber dokümanlarından çıkarılmasıdır. Bu işlemi yapabilmek için Türkçe'de yer alan etkisiz kelimelerin listesi Fatih Üniversitesi Doğal Dil İşleme Grubu'nun internet sitesinden alınmış ve bu liste kullanılarak tüm haber dokümanlarından etkisiz kelimeler silinmiştir. Etkisiz kelimelerin ayıklanmasından sonra dokümanlardaki noktalama işaretleri ve sayısal ifadeler de temizlenmiştir.

Metin ön işlemesinin ikinci aşaması ise haber dokümanlarında geçen kelimelerin kök halinin elde edilmesidir. Bu işlem JAVA programlama dilinde yazılmış olan Türkçe Dil İşleme Paketi ZEMBEREK kullanılarak yapılmıştır (Akın ve Akın, 2007). Türkçe'nin sondan eklemeli dil yapısına sahip olmasının avantajıyla kökleri aynı olan sözcükler tespit edilmiş ve bu kelimeler tek bir kök altında birleştirilmiştir. Kelimelerin kök halinin bulunmasıyla kelime uzayının boyutu indirgenmiştir. Bu işlemin tamamlanmasıyla doküman koleksiyonundan 13800 adet tekil kelime elde edilmiştir

Metin ön işlemenin son aşamasında ise doküman sıklığı temel alınarak boyut indirgemesi yapılmıştır. Bir önceki aşamada bulunan tekil kelimelerin doküman sıklıkları bulunmuş ve sıklık değerleri büyükten küçüğe doğru sıralanmıştır. Doküman sıklığı 100'den az olan tekil kelimeler silinmiştir. Bu işlemin sonucunda 13800 olan tekil kelime sayısı 2888'e indirilmiştir.

Ön işleme aşamasından elde edilen 2888 tekil kelime ile her bir haber dokümanı öznitelik vektörü olarak temsil edilmiştir. Öznitelik vektörleri Kelimeler Çantası gösterimdedir ve bu vektörlerin her bir boyutu bir tekil kelimeyi temsil eder. Kelimeler çantası gösteriminde kelimelerin dokümanda geçme sırası önemli değildir. d. haber dokümanına ait vektör temsili Eşitlik 4.1'de gösterilmektedir.

$$\mathrm{x(d)} = [w_1, w_2, w_3, w_4, \dots, w_n] \qquad \textbf{(4.1)}$$

$\mathrm{x(d)}$, n boyutlu öznitelik vektörü olarak adlandırılır. w'ler ise her bir tekil kelimeyi ifade etmektedir. Eğer bir tekil kelime haber dokümanında geçiyorsa, öznitelik vektöründe bu kelimeye ait boyutun değerine 1; aksi durumda ise 0 atanmaktadır.

İkili değerler kullanılarak tüm haber dokümanlarının öznitelik vektörleri oluşturulmuştur. Haber dokümanları için öznitelik vektörleri oluşturulduktan sonra, haberlerin yayınlanma tarihleri göz önünde bulundurularak, aynı gün içerisinde yayınlanan haber dokümanlarının öznitelik vektörleri birleştirilmiş ve o günü temsil eden tek bir öznitelik vektörüne dönüştürülmüştür. 2010 ve 2011 yılları için resmi tatiller ve hafta sonu günleri ihmal edilerek Borsa İstanbul'un işleme açık olduğu 503 işgününe ait 503 adet öznitelik vektörü oluşturulmuştur.

Kelimeler Çantası gösterimiyle ikili değerli öznitelik vektörleri oluşturulduktan sonra, alternatif olarak farklı ağırlıklandırma yönteminin kullanıldığı ikinci çeşit öznitelik vektörleri oluşturulmuştur. Bu vektörler de öznitelik değerleri olarak ikili değer yerine kelimelerin ham terim sıklıkları kullanılmıştır.

Haber metinleri kullanılarak öznitelik vektörlerinin oluşturulmasından sonra BIST 100 endeksinin günlük açılış fiyatı, kapanış fiyatı ve günlük hacim miktarı verileri kullanılarak da öznitelikler oluşturulmuştur. Belirlenen 9 fiyat özniteliği Çizelge 4.1'de gösterilmiştir.

Çizelge 4.1: Fiyat verisi kullanılarak belirlenen öznitelikler.

Öznitelik /Zaman	$\boldsymbol{t-1}$	$\boldsymbol{t-4}$	$\boldsymbol{t-19}$
Açılış	Op	w_Op	m_Op
Kapanış	Cl	w_Cl	m_Cl
Hacim	Vl	w_Vl	m_Vl

Çizelge 4.1'de gösterilen özniteliklerin t . gün için $\mathrm{gp}(t)$ ile temsil edilen 9 boyutlu öznitelik vektörü aşağıdaki gibi olacaktır:

$$\mathrm{gp}(t) = [Op\,, Cl\,, Vl\,, w\_Op\,, w\_Cl\,, w\_Vl\,, m\_Op\,, m\_Cl\,, m\_Vl] \qquad \textbf{(4.2)}$$

$\mathrm{gp}(t)$ vektörünün değerleri hesaplarken, $Op,\ Cl,\ Vl$ özniteliklerinin t.gündeki değeri için $t-1$. işlem günündeki BIST 100 endeksinin açılış fiyatı, kapanış fiyatı ve hacim miktarı aynen alınmıştır. $w\_Op,\ w\_Cl,\ w\_Vl$ özniteliklerinin t. gündeki değerleri, kendinden önceki 4 işlem gününün $(t-4)$ değerlerinin ortalaması alınarak hesaplanmıştır. $m\_Op,\ m\_Cl,\ m\_Vl$ özniteliklerinin t. gündeki değeri ise kendinden önceki 19 işlem gününün değerlerinin ortalaması $(t-19)$ alınarak bulunmuştur.

2010 ve 2011 yılları arasındaki her gün için bulunan fiyat öznitelik vektörlerinin değerleri daha sonra ayrıklaştırılmıştır. Ayrıklaştırma işleminde tüm fiyat öznitelik vektörlerinin her bir boyutunun ortalaması, μf, bulunmuştur. t.gün için f özniteliğin ayrık değeri olan $df(t)$ aşağıdaki gibi atanmıştır:

$$\mathrm{df}(t) = \begin{cases} +1\ , & \text{Eğer } \mathrm{f}(t) > \mu f \\ 0\ , & \text{Eğer } \mu f \leq \mathrm{f}(t) \end{cases} \qquad \textbf{(4.3)}$$

Bu işlem sonucunda fiyat verileri kullanılarak her işlem günü için ayrıklaştırılmış fiyat öznitelik vektörleri elde edilmiştir.

Öznitelik vektörlerinin oluşturulmasından sonra vektörlere BIST 100 endeksinin hareket yönünü gösteren sınıf etiketlerinin ataması yapılmıştır. Haber metinlerinin yayınlandıktan sonraki günde BIST 100 endeksinin yönünü tahmin etmek için, 2010 ve 2011 yıllarındaki her işlem gününde BIST 100 endeksinin açılış fiyatının değişim oranı incelenmiştir. Eşitlik 4.4, t.gün için açılış fiyatı değişim oranının, $p(t)$'in, nasıl hesaplanacağını göstermektedir.

$$p(t) = \frac{[s(t) - s(t-1)]}{s(t-1)} \qquad \textbf{(4.4)}$$

Eşitlik 4.4'de, $s(t)$ t.gündeki endeks açılış fiyatını, $s(t-1)$ ise $t-1$. gündeki endeks açılış fiyatını göstermektedir. İki yıllık süre zarfında, günlük açılış fiyatındaki değişim oranları göz önünde bulundurularak, bu değerlerin standart sapması (σ) bulunmuştur. σ değeri anlamlı fiyat değişimine karar vermede kullanılmıştır.$p(t)$ ve σ değerleri kullanılarak sınıf etiketlerinin oluşturulması Eşitlik 4.5 'te verilmiştir.

$$r(t) = \begin{cases} +1\ , & p(t) \geq +\sigma \\ 0\ , & -\sigma < p(t) < +\sigma \\ -1\ , & p(t) \leq -\sigma \end{cases} \qquad \textbf{(4.5)}$$

Eşitlik 4.5'e göre, eğer $p(t)$ değeri σ değerinden büyük ise bu t. günde endeks fiyatında anlamlı bir artış olduğunu göstermektedir ve o güne etiket olarak "+1" atanmıştır. Eğer $p(t)$ değeri σ değerinden küçük ise o güne etiket olarak anlamlı azalışı gösteren "-1" etiketi tanımlanmıştır. "0" etiketi ise endeks fiyatında anlamlı bir değişme olmadığını göstermektedir. Sınıf etiketlerinin oluşturulmasındaki bu ölçüt diğer borsa tahmini çalışmalarında da kullanılmıştır (Gidofalvi, 2001).

2010 ve 2011 yıllarındaki her işlem günü sınıf etiketleri belirlendikten sonra, bu etiketler haber dokümanlarının yayınlanma tarihleri göz önünde bulundurularak öznitelik vektörlerine atanmış ve veri kümeleri oluşturulmuştur.

4.3 Öznitelik Seçimi ve Sınıflandırma Sonuçları

Bir önceki başlıkta anlatıldığı üzere haber dokümanları kullanılarak Kelimeler Çantası gösteriminde ikili değere ve kelimelerin terim sıklığı değerlerine sahip iki çeşit veri kümesi oluşturulmuştur. İki veri kümesi de 2888 öznitelik içermektedir. Öznitelik uzayından bilgi içeren ve ilişkili öznitelikleri seçebilmek için ilk önce gözetimli öznitelik seçme yöntemlerinden biri olan Karşılıklı Bilgi kullanılmıştır. Bu yöntem ikili öznitelik değeri içeren veri kümesine uygulanmıştır. Aynı veri kümesi üzerinde Bilgi Kazanımı ve Ki-Kare istatistiği öznitelik seçme yöntemleri kullanılarak sınıflandırma performansı elde edilmiştir. Yapılan bu deneylerin ardından kullanılan öznitelik seçme yöntemleri terim sıklığı değerlerine sahip veri kümesi üzerinde uygulanmıştır. Elde edilen deneysel sonuçlar sonraki alt başlıklar da gösterilmiştir.

4.3.1 İkili değere sahip veri kümesi üzerinde yapılan deneyler

Karşılıklı bilgiye göre öznitelik seçimi

Karşılıklı bilgi yöntemi kullanılarak öznitelik vektörlerinin her bir boyutu ile bu vektörlere atanan sınıf etiketleri arasındaki ilişkiye bakılmıştır. Her bir öznitelik için karşılıklı bilgi puanı hesabı yapılmış ve en yüksek puana sahip ilk 1000 öznitelik seçilmiştir. Öznitelikler belirlendikten sonra veri kümesi, Naive Bayes sınıflandırıcısını eğitmek ve sınıflandırma performansını test etmek için iki kısma ayrılmıştır. Ayrım yapılırken borsa işlem günleri dikkate alınmıştır. 2 yıllık süre içerisinde ilk 18 aya ait öznitelik vektörleri eğitim kümesini, son 6 aya ait öznitelik vektörleri de test kümesini oluşturmuştur. Eğitim kümesi ile 12 farklı öznitelik alt

kümesi oluşturulmuştur. Bu alt kümelerin öznitelik sayıları sırasıyla 10, 20, 50, 100, 200, 300, 500, 600, 700, 800, 900 ve 1000'dir. Aynı alt kümeler test kümesi içinde seçilmiştir. İkili değere sahip öznitelik alt kümeleri için sınıflandırma sonuçları Çizelge 4.2'de görülmektedir. Sınıflandırma performansı, kesinlik ve duyarlılık ölçütlerinin birleşimi olan F-Ölçütü, Makro-ortalama F-Ölçütü ve doğruluk oranı kullanarak hesaplanmıştır.

Çizelge 4.2: Karşılıklı bilgi yöntemiyle seçilen öznitelikler ile sınıflandırma sonuçları.

Karşılıklı Bilgi					
Öznitelik Sayısı	**F-Ölçütü**			**Makro-Ortalama F-ölçütü**	**Doğruluk Oranı**
	-1	**0**	**+1**		
10	0,33	0,77	0,00	0,37	0,63
20	0,54	0,77	0,20	0,50	0,65
50	0,54	0,74	0,50	0,60	0,66
100	0,51	0,72	0,48	0,57	0,62
200	0,51	0,73	0,69	0,64	0,66
300	0,50	0,72	**0,71**	0,64	0,65
500	0,51	0,75	0,69	0,65	0,67
600	0,52	0,75	0,56	0,61	0,66
700	0,53	0,76	0,63	0,64	0,68
800	**0,56**	**0,80**	0,65	**0,67**	**0,72**
900	0,54	0,79	0,67	0,66	0,70
1000	0,52	0,78	0,53	0,61	0,68

Sonuçlar incelendiğinde, sadece 20 öznitelik kullanılarak %65'lik doğruluk oranı elde edildiği görülmektedir. Sınıflandırıcının performansına sınıf seviyesinde bakıldığında ise "0" etiketli sınıfa ait F-Ölçütü oranı %77 olarak gerçekleşmiştir. Ancak bu 50 öznitelikle "-1" ve "+1" etiketine sahip sınıfların tahmin performansı düşük kalmaktadır.

Sınıflandırma da kullanılan öznitelik sayısı 800'e çıkarıldığında ise ,"+1" sınıfı için F-Ölçütü oranı %56'ya, "0" sınıf için ise %80'e çıkmıştır. "-1" sınıfının F-Ölçütü oranı da %65'e yükselmiştir.

Çizelge 4.2'teki sonuçlar, Naive Bayes sınıflandırıcısının az sayıda öznitelik kullanarak "-1" ve "+1" etiketli örnekleri yüksek performansla tahmin edemediğini yansıtmaktadır. Bu durumun temel nedeni ise veri kümesindeki sınıf dağılımının dengesiz olmasıdır. Çizelge 4.3'te gösterilen veri kümesinin sınıf dağılımında, "0" sınıfındaki örneklerin sayısının baskın olduğu görülmektedir. Karşılıklı bilgi yöntemi özniteliklere ait puan hesabını yaparken, sınıf dağılımlarını da göz önünde bulundurduğundan, "0" sınıfındaki örnekler öznitelik sürecini doğrudan etkilemektedir. Bu durumu ortadan kaldırmak için öznitelik seçimine "-1" ve "+1" sınıfındaki örneklerin de adil bir şekilde katılacağı yeni bir öznitelik seçme yöntemi geliştirilmiştir.

Çizelge 4.3: Veri kümesi sınıf dağılımı.

Sınıf	**Örnek Sayısı**
+1	65
0	376
-1	62

Bu yöntemde, öznitelik seçme aşamasında eğitim kümesindeki örneklerden sınıf etiketlerine göre 3 ayrı veri kümesi oluşturulmuştur ve bir veri kümesinde bulunan örnek sayısına eşit sayıda diğer iki veri kümesinden rastgele örnekler seçilmiştir. Örneğin birinci veri kümesi 62 adet "-1" etiketine sahip örnek içermektedir. Birinci veri kümesine, "0" ve "+1" etiketli örneklerden rastgele 62 adet örnek daha seçilerek veri kümesinde dengeli bir sınıf dağılımı elde edilmiştir. Aynı işlem diğer iki veri kümesi içinde yapılmıştır. 3 sınıf için de dengeli sınıf dağılımlı veri kümeleri oluşturulduktan sonra bu 3 veri kümesindeki özniteliklerin karşılıklı bilgi puanı hesaplanmıştır. Sınıf dağılımlarını dengeleme ve karşılıklı bilgi puanı hesaplama işlemi 100 defa gerçekleştirilmiş ve 100 işlem sonucunda bulunan karşılıklı bilgi puanlarının ortalaması alınmıştır. Üç veri kümesinden öznitelik seçimi sırasıyla her veri kümesindeki en yüksek karşılıklı bilgi puanına sahip özniteliğin seçilmesiyle

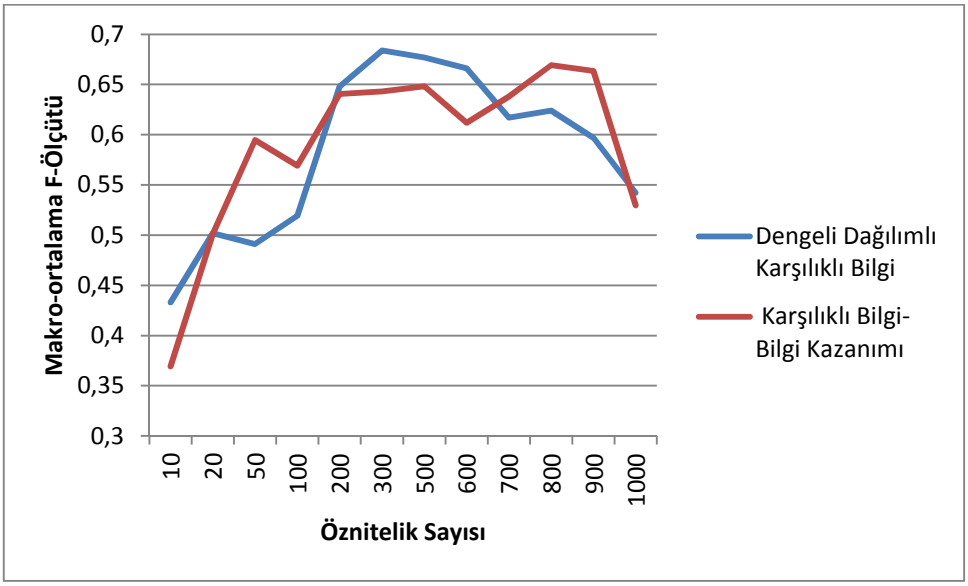
yapılmıştır. Bu seçim işlemi 1000 öznitelik seçilene kadar devam etmiştir. Bir önceki deneyde olduğu gibi seçilen 1000 öznitelik kullanılarak sınıflandırıcı eğitimi için 10, 20, 50, 100, 200, 300, 500, 600, 700, 800, 900 ve 1000 özniteliğe sahip 12 öznitelik alt kümesi oluşturulmuştur. Aynı öznitelikler test kümesi içinde seçilmiştir. Kullanılan bu öznitelik seçme yöntemiyle elde edilen sınıflandırma sonuçları Çizelge 4.4'de verilmiştir.

Çizelge 4.4: Dengeli dağılımlı karşılıklı bilgi yöntemiyle seçilen öznitelikler ile sınıflandırma sonuçları.

Karşılıklı Bilgi (Dengeli Dağılımlı)					
Öznitelik Sayısı	**F-Ölçütü**			**Makro-Ortalama F-Ölçütü**	**Doğruluk Oranı**
	-1	**0**	**+1**		
10	0,45	0,75	0,10	0,43	0,63
20	0,59	0,76	0,15	0,50	0,66
50	0,48	0,71	0,29	0,49	0,60
100	0,56	0,75	0,25	0,52	0,65
200	0,59	0,81	0,54	0,65	0,73
300	**0,62**	0,81	0,62	**0,68**	0,74
500	0,58	0,81	**0,64**	0,68	0,74
600	0,60	**0,83**	0,57	0,67	**0,75**
700	0,52	0,81	0,52	0,62	0,72
800	0,54	0,81	0,52	0,62	0,72
900	0,57	0,83	0,40	0,60	0,74
1000	0,55	0,82	0,26	0,54	0,72

Dengeli dağılımlı öznitelik seçimi ile 300 öznitelik kullanılarak %68'lik Makro-ortalama F-Ölçütü oranı elde edilmiştir. Aynı özniteliklerle "-1", "0" ve "+1" etiketli sınıflar için bulunan F-Ölçütü oranları sırasıyla %62, %81 ve % 62'dir. Öznitelik seçiminde sınıf dağılımının dikkate alınması ile doğruluk ve F-Ölçütü oranlarında önemli miktarda artış gerçekleşmiştir. Bu durum karşılıklı bilgi yöntemi ile dengeli dağılımlı karşılıklı bilgi yönteminin sınıflandırma performansları karşılaştırıldığında açıkça görülmektedir.

Şekil 4.3'teki grafikte bu yöntemlerin sınıflandırma performansları gösterilmiştir. Performans ölçütü olarak Makro-ortalama F-Ölçütü kullanılmıştır.



Şekil 4.3: Karşılıklı bilgi ve dengeli dağılımlı karşılıklı bilgi yöntemlerinin performans karşılaştırması.

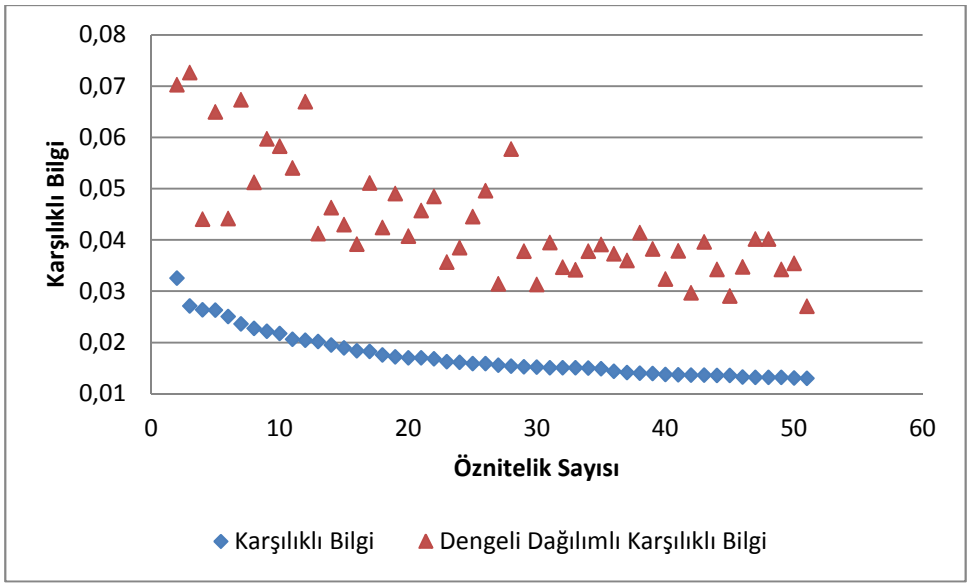
Grafik incelendiğinde dengeli dağılımlı karşılıklı bilgi yöntemi ile seçilen öznitelik alt kümeleri ile en başarılı sınıflandırma performansı elde edildiği görülmektedir. Bunun nedeni, dengeli dağılımlı karşılıklı bilgi yönteminin öznitelik seçimi sürecinde bütün sınıflara ait örneklere eşit şans vermesidir.

Sonraki aşamada ise karşılıklı bilgi yöntemiyle seçilen ilk 50 özniteliğin karşılıklı bilgi puanının, dengeli dağılımlı karşılıklı bilgi yöntemi uygulandığında nasıl değişim gösterdiği incelenmiş ve Şekil 4.4'teki grafikte karşılaştırılmıştır.

Şekil 4.4'teki grafikte görüldüğü üzere, dengeli dağılımlı karşılıklı bilgi yöntemi kullanıldığında, karşılıklı bilgi yöntemine göre bazı öznitelikler ön plana çıkmıştır. Bu özniteliklerin karşılıklı bilgi yöntemi puanları ilk aşamada düşük iken, dengeli dağılımlı karşılıklı bilgi yöntemi kullanıldığında elde edilen puanlarda artış olmuştur.

Bilgi kazanımı ve Ki-kare yöntemleri ile öznitelik seçimi

Dengeli dağılımlı karşılıklı bilgi yöntemiyle sınıflandırıcı performansı değerlendirildikten sonra, ikili değerli veri kümesi üzerine Bilgi kazanımı ve Ki-Kare istatistiği yöntemleri uygulanmıştır. İki yöntemde de veri kümesindeki her bir öznitelik için puan hesabı yapılmış ve en yüksek puana sahip 1000 öznitelik seçilmiştir.



Şekil 4.4: 50 öznitelik için karşılıklı bilgi ve dengeli dağılımlı karşılıklı bilgi puanları.

Önceki deneyde olduğu gibi bu özniteliklerle de öznitelik altkümeleri oluşturulmuştur. İki öznitelik seçme yöntemi ile elde edilen sınıflandırma performansı Çizelge 4.5'te gösterilmiştir.

Çizelge 4.5'te görüldüğü üzere, bilgi kazanımı yöntemiyle öznitelik seçimi yapıldığında, en yüksek doğruluk oranına ve Makro-ortalama F-Ölçütü başarısına 800 öznitelik kullanılarak ulaşılmaktadır. Bu yöntemle elde edilen doğruluk oranı %72'dir.

Öznitelik seçiminde Ki-kare istatistiği yöntemi kullanıldığında ise sadece 50 öznitelik ile en yüksek sınıflandırma başarısı bulunmuştur. 50 öznitelikle elde edilen doğruluk oranı %74' tür. Kullanılan 50 öznitelik ile yüksek başarı elde edildikten sonra, 100 öznitelik ile sınıflandırıcı eğitimi yapıldığında ise sınıflandırma başarısında ciddi bir düşüş olmuştur. Buna neden olarak ise öznitelik seçiminde süzgeç yaklaşımının kullanılması gösterilebilir.

Süzgeç yaklaşımı ile her bir öznitelik ile sınıf etiketleri arasındaki ilgililik (İng. Relevance) derecesi tespit edilmektedir. Bu yaklaşımda kullanılan yöntemler, özniteliklerin birbirleriyle ilinti (İng. Correlation) derecesini ölçememektedir. Süzgeç yaklaşımında yüksek ilgililik puanına sahip öznitelikler seçildiğinde, seçilen öznitelikler arasında ilinti derecesi fazla ise sınıflandırma performansında düşüş yaşanabilmektedir.

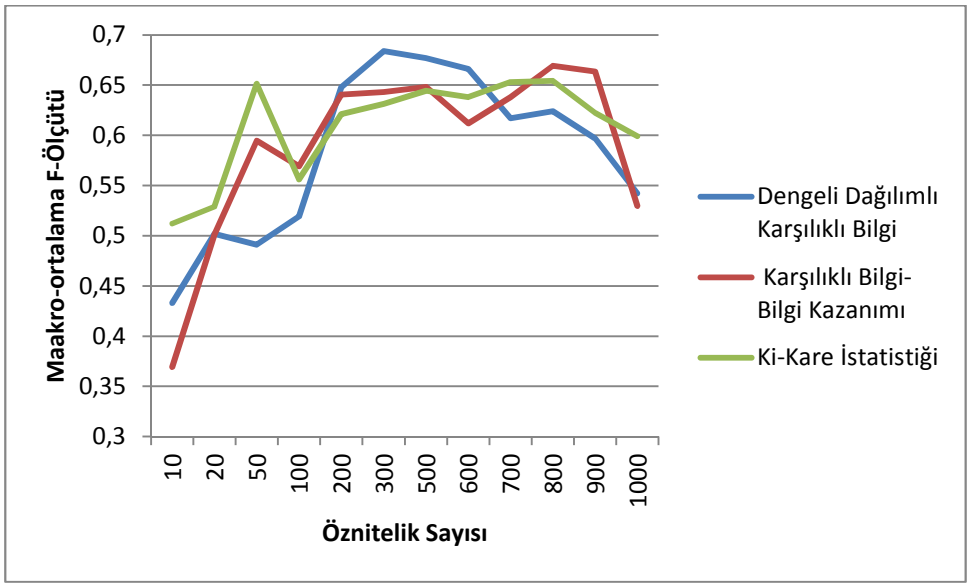
Çizelge 4.5: Ki-kare istatistiği ve bilgi kazanımı yöntemleri ile elde edilen sınıflandırma sonuçları.

	Ki-Kare İstatistiği					**Bilgi Kazanımı**				
Öznitelik Sayısı	**F-Ölçütü**			**Makro-Ortalama F-Ölçütü**	**D.O**	**F-Ölçütü**			**Makro-Ortalama F-Ölçütü**	**D.O**
	-1	**0**	**+1**			**-1**	**0**	**+1**		
10	0,48	0,82	0,24	0,51	0,70	0,33	0,77	0,00	0,37	0,63
20	0,58	0,79	0,21	0,53	0,68	0,54	0,77	0,20	0,50	0,65
50	**0,64**	**0,83**	0,48	**0,65**	**0,74**	0,54	0,74	0,50	0,60	0,66
100	0,48	0,70	0,48	0,56	0,61	0,51	0,72	0,48	0,57	0,62
200	0,52	0,74	0,60	0,62	0,66	0,51	0,73	0,69	0,64	0,66
300	0,50	0,75	0,65	0,63	0,66	0,50	0,72	**0,71**	0,64	0,65
500	0,52	0,75	**0,67**	0,64	0,67	0,51	0,75	0,69	0,65	0,67
600	0,53	0,76	0,63	0,64	0,68	0,52	0,75	0,56	0,61	0,66
700	0,54	0,78	0,65	0,65	0,70	0,53	0,76	0,63	0,64	0,68
800	0,55	0,78	0,63	0,65	0,70	**0,56**	**0,80**	0,65	**0,67**	**0,72**
900	0,52	0,80	0,55	0,62	0,70	0,54	0,79	0,67	0,66	0,70
1000	0,52	0,78	0,50	0,60	0,68	0,52	0,78	0,53	0,61	0,68

Bundan dolayı hem özniteliklerin sınıf etiketleriyle olan ilgililiğini, hem de öznitelikler arasındaki ilinti derecesini göz önünde bulunduran En az tekrar en fazla ilgililik (İng. Minimum Redundancy Maximum Relevance- mRMR) yöntemi kullanılarak öznitelik seçimi yapılmalıdır. mRMR yöntemiyle ilgili ayrıntılı bilgiye Peng ve diğ. (2005) yayınından ulaşılabilir.

İkili değerli veri kümesi üzerinde uygulanan bu iki öznitelik seçme yönteminin sınıflandırma performansları, bir önceki deneyde kullanılan dengeli dağılımlı karşılıklı bilgi ve karşılıklı bilgi yöntemleriyle karşılaştırılmıştır.

Karşılaştırmaya ait grafik Şekil 4.5'te gösterilmektedir.



Şekil 4.5: Dengeli dağılımlı karşılıklı bilgi, karşılıklı bilgi Ki-Kare istatistiği ve bilgi kazanımı yöntemlerinin performans karşılaştırması.

Grafikte, dengeli dağılımlı karşılıklı bilgi yönteminin en başarılı sınıflandırma sonucunu elde ettiği görülmektedir. Dengeli dağılımlı karşılıklı bilgi yöntemi kullanılarak seçilen 300 öznitelikle, %68,4'lük Makro-ortalama F-Ölçütü başarı oranına ulaşılmıştır. Sınıflandırmada kullanılan öznitelik sayısının 100'den az olduğu durumda ise Ki-Kare istatistiği yönteminin performansı üst düzeydedir. Öznitelik sayısının 600'den fazla olduğu sınıflandırma sürecinde bilgi kazanımı ve Ki-kare istatistiği performans açısından üst seviyedeyken, dengeli dağılımlı karşılıklı bilginin performansında ciddi düşüş yaşanmaktadır. Karşılıklı bilgi yönteminin sınıflandırma performansı ise bilgi kazanımı yöntemiyle aynıdır. Bunun nedeni, Weka Veri Madenciliği yazılımının her bir öznitelik için bilgi kazanımı değerini 4.6'da gösterilen eşitlik ile bulmasıdır:

$$IG(T,C) = H(C) - H(C|T) \quad \textbf{(4.6)}$$

Eşitlikte $H(C)$, sınıf düzensizlik değerini; $H(C|T)$, T özniteliği kullanıldığında hesaplanan koşullu düzensizlik değerini göstermektedir. Aynı eşitlik bazı kaynaklar da karşılıklı bilgi tanımı olarak da gösterilmektedir (Gray, 2013). Bundan dolayı bilgi kazanımı ve karşılıklı bilgiyle seçilen öznitelikler birbirinin aynısıdır.

4.3.2 Terim sıklığı değerlerine sahip veri kümesi üzerinde yapılan deneyler

Haber dokümanlarındaki kelimelerin terim sıklıkları kullanılarak oluşturulan veri kümesi üzerinde de ikili değerli veri kümesi üzerine uygulanan öznitelik seçme yöntemleri kullanılmıştır. Öznitelik seçme sürecinden önce her bir özniteliğe ait sütun vektörü ayrılmış ve bu vektörlerdeki öznitelik değerleri WEKA Veri Madenciliği yazılımı kullanılarak ayrıklaştırmıştır. Ayrıklaştırma işleminde sütun vektöründeki en yüksek öznitelik değeri ile en düşük öznitelik değeri bulunmuş, bulunan bu iki değer arası 10 eşit değer aralığına ayrılmıştır. Her bir değer aralığı için bir etiket tanımlanmış ve öznitelik satır vektöründeki her bir öznitelik değerinde bakılarak, bu değer hangi aralığa denk geliyorsa o değerin yerine aralığa ait etiket atanmıştır. Böylece kelimelerin terim sıklığı değerleri yerine 10 adet nominal etiket kullanılmıştır. Değerlerin ayrıklaştırılmasından sonra karşılıklı bilgi, dengeli dağılımlı karşılıklı bilgi, bilgi kazanımı ve Ki-kare istatistikleri kullanılarak bu yöntemler için en yüksek değerli 1000 öznitelik seçilmiştir. Sınıflandırıcı eğitimi için, seçilen özniteliklerle 10, 20, 50, 100, 200, 300, 500, 600, 700, 800, 900 ve 1000 özniteliğe sahip alt kümeler oluşturulmuştur. Aynı öznitelikler test alt kümelerini oluşturmada da kullanılmıştır. Karşılıklı bilgi ve dengeli dağılımlı karşılıklı bilgi yöntemleri ile edilen sınıflandırma sonuçları Çizelge 4.6'da, bilgi kazanımı ve Ki-kare istatistikleri ile bulunan sınıflandırma sonuçları ise Çizelge 4.7'de verilmiştir.

Çizelge 4.6'da gösterilen sonuçlar incelendiğinde, karşılıklı bilgi yöntemiyle %75'lik doğruluk oranı ile %68'lik Makro-ortalama F-ölçütü başarısı elde edilmiştir. Bu sınıflandırma performansında 1000 öznitelik kullanılmıştır. "-1", "0" ve "+1" sınıfları için F-Ölçütü başarıları sırasıyla %59, %83 ve %62 olarak gerçekleşmiştir. Dengeli dağılımlı karşılıklı bilgi yöntemi ile öznitelik seçim yapıldığında ise sınıflandırma performansında düşüş olmuştur. Bu yöntemle elde edilen en yüksek Makro-ortalama F-ölçütü başarısı %55 olarak gerçekleşmiştir.

Çizelge 4.7'de görüldüğü üzere Ki-Kare istatistiği öznitelik seçme yöntemiyle en iyi doğruluk oranına 900 öznitelik kullanılarak ulaşılmaktadır. Bu 900 öznitelikle "-1", "0" ve "+1" sınıfları için sırasıyla %66, %87 ve % 52 oranlarında F-Ölçütü başarısı yakalanmıştır. Ki-Kare istatistiği ile elde edilen sonuçlar, 50 öznitelik kullanılarak da sınıf seviyesinde tatmin edici performans elde edilebildiğini göstermektedir.

Çizelge 4.6: Karşılıklı bilgi ve dengeli dağılımlı karşılıklı bilgi yöntemleri ile elde edilen sınıflandırma performansları (terim sıklığı değerli).

	Karşılıklı Bilgi					**Dengeli Dağılımlı Karşılıklı Bilgi**				
Öznitelik Sayısı	**F-Ölçütü**			**Makro-Ortalama F-Ölçütü**	**D.O**	**F-Ölçütü**			**Makro-Ortalama F-Ölçütü**	**D.O**
	-1	**0**	**+1**			**-1**	**0**	**+1**		
10	0,50	0,77	0,31	0,53	0,66	0,41	0,63	0,02	0,35	0,50
20	0,47	0,71	0,35	0,51	0,59	0,38	0,44	0,00	0,27	0,37
50	0,54	0,74	0,25	0,51	0,62	0,36	0,43	0,00	0,27	0,36
100	0,47	0,65	0,38	0,50	0,55	0,38	0,45	0,14	0,32	0,38
200	0,46	0,67	0,50	0,54	0,58	0,39	0,59	0,16	0,38	0,47
300	0,47	0,71	0,46	0,55	0,60	0,42	0,63	0,35	0,46	0,52
500	0,53	0,77	0,50	0,60	0,67	0,47	0,72	**0,40**	0,53	0,61
600	0,58	0,82	0,59	0,66	0,73	0,48	0,76	0,29	0,51	0,63
700	0,56	0,82	0,59	0,66	0,73	**0,50**	**0,78**	0,36	**0,55**	**0,66**
800	0,58	0,83	0,59	0,67	0,74	0,48	0,76	0,36	0,53	0,64
900	**0,59**	**0,84**	0,56	0,66	**0,75**	0,49	0,77	0,33	0,53	0,65
1000	0,59	0,83	**0,62**	**0,68**	0,75	0,50	0,78	0,26	0,51	0,66

Sınıflandırıcı eğitiminde bilgi kazanımı yöntemiyle seçilen öznitelikler kullanıldığında ise doğruluk oranı olarak %75'e erişilebilmektedir. 1000 öznitelik kullanılarak yapılan bu sınıflandırma ile "-1" sınıfı için %59, "0" sınıfı için %83, "+1" sınıfı için ise %62 F-Ölçütü oranları elde edilmiştir. Sonuçlardan çıkarılabilecek diğer bir husus ise bilgi kazanımı öznitelik seçme yönteminin Ki-kare istatistiğine göre "+1" sınıfındaki örnekleri tahmin etmede daha başarılı olmasıdır. Terim sıklığı değerli veri kümesine ait sınıflandırma performansı ile önceki deneylerde kullanılan ikili dağılımlı veri kümesinden elde edilen sınıflandırma performansları karşılaştırılmıştır.

Karşılaştırma, her öznitelik seçme yöntemi için elde edilen en yüksek Makro-ortalama F-Ölçütü başarısı kullanılarak yapılmıştır. Karşılaştırma sonuçları Çizelge 4.8'de gösterilmektedir.

Çizelge 4.7: Ki-Kare istatistiği ve bilgi kazanımı yöntemleri ile elde edilen sınıflandırma performansları (terim sıklığı değerli).

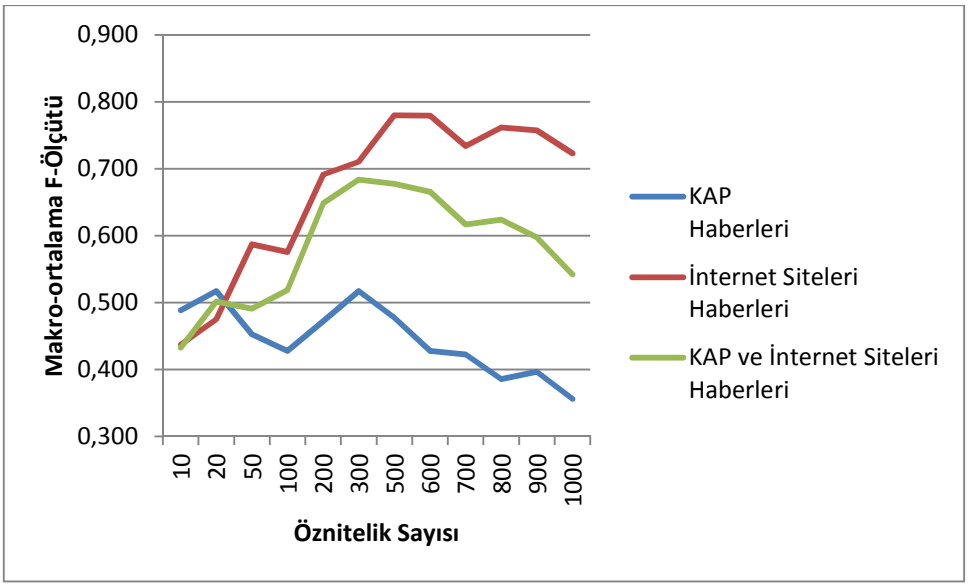
	Ki-Kare İstatistiği					**Bilgi Kazanımı**				
Öznitelik Sayısı	**F-Ölçütü**			**Makro-Ortalama F-Ölçütü**	**D.O**	**F-Ölçütü**			**Makro-Ortalama F-Ölçütü**	**D.O**
	-1	**0**	**+1**			-1	0	+1		
10	0,53	0,82	0,07	0,48	0,68	0,50	0,77	0,31	0,53	0,66
20	0,52	0,81	0,44	0,59	0,70	0,47	0,71	0,35	0,51	0,59
50	0,57	0,84	0,45	0,62	0,73	0,54	0,74	0,25	0,51	0,62
100	0,57	0,74	0,39	0,57	0,64	0,47	0,65	0,38	0,50	0,55
200	0,51	0,72	0,42	0,55	0,62	0,46	0,67	0,50	0,54	0,58
300	0,55	0,75	0,45	0,58	0,66	0,47	0,71	0,46	0,55	0,60
500	0,57	0,84	0,48	0,63	0,74	0,53	0,77	0,50	0,60	0,67
600	0,60	0,85	**0,52**	0,66	0,75	0,58	0,82	0,59	0,66	0,73
700	0,60	0,84	0,48	0,64	0,74	0,56	0,82	0,59	0,66	0,73
800	0,66	0,86	0,50	0,67	0,78	0,58	0,83	0,59	0,67	0,74
900	0,66	**0,87**	0,52	**0,68**	**0,78**	**0,59**	**0,84**	0,56	0,66	**0,75**
1000	**0,68**	0,87	0,40	0,65	0,78	0,59	0,83	**0,62**	**0,68**	0,75

Sonuçlar incelendiğinde en yüksek Makro-ortalama F-Ölçütü başarısına ikili değerli veri kümesi kullanılarak Dengeli Dağılımlı Karşılıklı Bilgi ile ulaşılmıştır. Diğer yöntemlerle de elde edilen başarılar yüksek olmasına karşın, Dengeli Dağılımlı Karşılıklı Bilgi bu başarıyı daha az öznitelik kullanarak yakalamıştır.

Çizelge 4.8: Yöntemlerle elde edilen en yüksek başarı oranları.

Veri Kümesi	Yöntem	Öznitelik Sayısı	Makro-Ortalama F-Ölçütü	Doğruluk Oranı
İkili Değerli Veri Kümesi	Karşılıklı Bilgi	800	0,67	0,72
	Dengeli Dağılımlı Karşılıklı Bilgi	**300**	**0,68**	**0,74**
	Bilgi Kazanımı	800	0,67	0,72
	Ki-Kare İstatistiği	800	0,65	0,70
Terim Sıklığı Değerli Veri Kümesi	Karşılıklı Bilgi	800	0,68	0,75
	Dengeli Dağılımlı Karşılıklı Bilgi	700	0,55	0,66
	Bilgi Kazanımı	800	0,68	0,75
	Ki-Kare İstatistiği	900	0,68	0,78

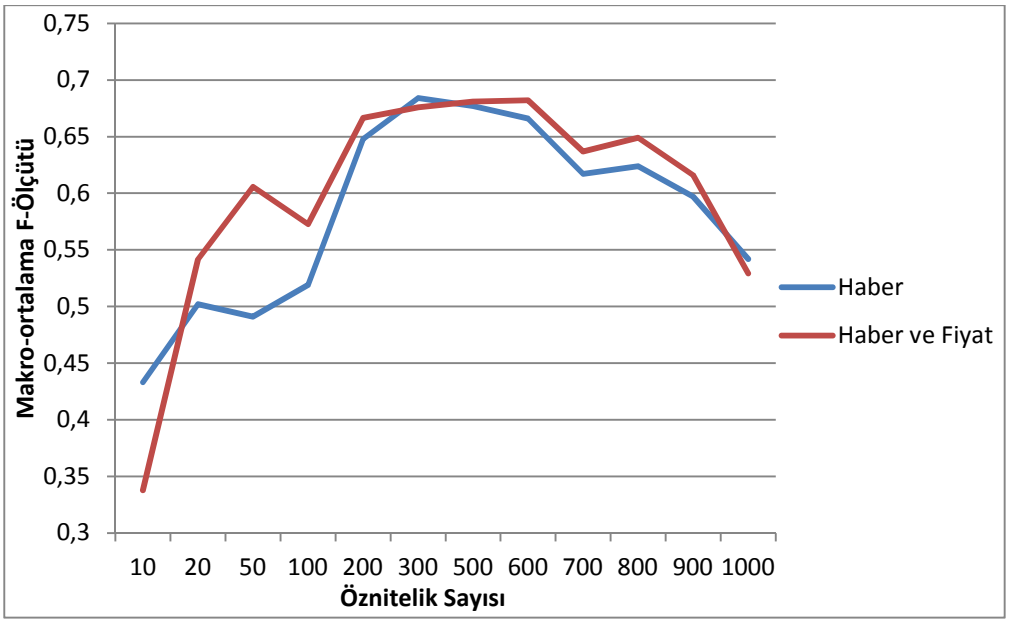
Bölüm 4.1'de anlatıldığı üzere, veri kümesi oluşturulurken iki farklı haber kaynağından alınan finansal haber metinleri kullanılmıştır. Farklı öznitelik seçme yöntemleriyle sınıflandırma performansları elde edildikten sonraki aşamada, kullanılan haber metinlerinin kaynağının sınıflandırma sürecine etkilerini incelenmiştir. Daha önce yapılan deneylerde gösterdiği yüksek performansı göz önünde bulundurularak bu deneyde de öznitelik seçme yöntemi olarak dengeli dağılımlı karşılıklı bilgi kullanılmıştır. Haber kaynaklarının türüne göre elde edilen sınıflandırma performansları Şekil 4.6'daki grafikte gösterilmiştir. Şekil 4.6'daki grafik, finansal internet sitelerinden alınan haber dokümanlarının endeks yönü tahmininde üst düzey performans sergilediğini göstermektedir. Bu dokümanlarla elde edilen Makro-ortalama F-Ölçütü başarısı %80'lere kadar çıkmaktadır.



Şekil 4.6: Haber kaynaklarının türüne göre elde edilen sınıflandırma performansları.

Kamu Aydınlatma Platformu (KAP)'ndan elde edilen haber dokümanlarıyla yakalanan tahmin performansı ise düşük olmaktadır. Buna neden olarak, problemin yapısı gösterilebilir. Çalışmada gün içinde yayınlanan haber metinlerinin sonraki günkü endeks fiyatına etkileri incelenmektedir. KAP'da yayınlanan haberler şirketlerin resmi bildirimleri olduğundan, borsa yatırımcıları gün içinde bu bildirimleri takip etmekte ve bildirimlere anlık tepki vermektedir. Yatırımcıların tepkileri sonraki günkü endeks fiyatına yansıdığından dolayı endeksin hareket yönünün tahmininde KAP'tan alınan haberler başarılı olamamaktadır. Her iki haber kaynağından alınan haber dokümanlarıyla sınıflandırıcı eğitildiğinde, sınıflandırıcının tahmin performansı sadece KAP haber dokümanlarıyla eğitilen sınıflandırıcıdan yüksek, sadece finansal internet sitelerinden alınan haber metinleriyle eğitilen sınıflandırıcıdan düşük olmaktadır.

Gerçekleştirilen son deneyde, sınıflandırma sürecinde haber metinleriyle oluşturulan öznitelik alt kümeleriyle birlikte fiyat özniteliklerinin de kullanılmasının etkileri incelenmiştir. Bu deneyde, haber metinlerinden dengeli dağılımlı karşılıklı bilgi yöntemi ile seçilen öznitelik alt kümelerine, fiyat verileri ile oluşturulmuş 9 adet fiyat özniteliği eklenmiştir. Şekil 4.7'deki grafikte sadece haber verileri ile haber dokümanları ve fiyat verilerinin kullanılması durumundaki sınıflandırma performansı gösterilmektedir. Sınıflandırıcıların performansı ise Makro-ortalama F-Ölçütü ile değerlendirilmiştir.



Şekil 4.7: Sadece haber verileri ile haber ve fiyat verilerinin kullanıldığı durumdaki sınıflandırma performansı.

Grafikte, haber öznitelikleri ile fiyat özniteliklerinin birlikte kullanılmasının az sayıda özniteliğe sahip altkümeler ile yapılan sınıflandırma sürecine olumlu katkı sağladığı görülmektedir. 50 özniteliğe sahip öznitelik alt kümesine fiyat özniteliklerinin eklenmesiyle sınıflandırma performansında yaklaşık olarak %20 oranında artış olmuştur. Bunun nedeni ise fiyat özniteliklerinin sadece fiyat değişiminin olmadığı günleri başarılı olarak tahmin etmesidir. Bu öznitelikler haber öznitelikleriyle beraber kullanıldığında, "0" sınıfındaki örneklerin tahmin başarısında artış görülmüştür ve sınıflandırıcının genel başarısı yükselmiştir. Fazla sayıda haber özniteliğinin kullanıldığı durumda ise fiyat özniteliklerinin eklenmesinin herhangi bir katkısı olmamıştır.

4.4 Seçilen Kelimelerin İncelenmesi (Kelime Bulutu)

Sınıflandırma işlemleri yapıldıktan sonra öznitelik alt kümeleri hazırlanırken kullanılan 1000 öznitelikten dengeli dağılımlı karşılıklı bilgi tabanlı benzerliği en yüksek olan 50'si alınmış ve 50 öznitelik kullanılarak Kırmızı-Yeşil-Mavi (İng. 'Red' 'Green' 'Blue'- RGB) kelime bulutu oluşturulmuştur. Kelime bulutunu oluşturmak için www.wordle.net web sitesi kullanılmıştır. Kelime bulutları oluşturulurken seçilen bu 50 kelimenin geçtiği haber dokümanlarına bakılmış ve bu kelimelerin % kaç oranında anlamlı fiyat artışına, fiyat azalışına ve fiyatın sabit kalmasına sebep olduğu bulunmuştur.

Çizelge 4.9'da her bir kelimenin sınıf bazında yüzdesel olarak etkileri gösterilmiştir. Bulunan bu oranlar sabit 100 sayısıyla çarpılarak tam sayı haline getirilmiştir. Bu oranların ağırlıklarına göre artışa sebep olan kelimeler yeşil renkli, değişime sebep olmayan kelimeler mavi renkli, azalışa neden olan kelimeler ise kırmızı renkli kelime bulutlarıyla temsil edilmiştir. Bu kelime bulutları sırasıyla Şekil 4.8'de, Şekil 4.9'da ve Şekil 4.10'da gösterilmiştir.

Çizelge 4.9: Seçilen 50 kelimenin sınıf bazında yüzdesel etkileri.

Anlaş	14	73	11	**Yerli**	15	72	12	**Eşya**	8	75	16
Kıy	16	69	13	**Biraz**	13	74	12	**Çatış**	20	66	13
Amerika	13	73	12	**İçer**	14	72	13	**Bahar**	12	67	19
Men	16	69	13	**Tbmm**	18	71	10	**Torun**	19	71	9
Gerilim	20	64	15	**Tehlike**	13	76	10	**Konya**	17	68	14
Cay	20	70	10	**Heyecan**	15	65	19	**Tetik**	17	72	10
Muhalefet	17	70	11	**Kötüle**	18	66	15	**Sarsıl**	26	67	6
Nere	13	72	13	**Transfer**	14	73	12	**Kasap**	15	80	3
Panik	20	67	12	**Müdahale**	15	71	13	**Eğer**	13	73	13
Dursun	5	84	9	**Nazım**	17	74	8	**Resim**	6	65	27
Usta	24	63	12	**Yüzlerce**	13	80	5	**Yatış**	22	58	18
Tecrübe	17	67	14	**Fayda**	13	73	13	**Boyut**	14	72	12
Böylece	13	73	13	**Bağlantı**	13	70	16	**Hale**	13	74	12
Yarala	19	59	20	**Yayıl**	14	71	13	**Düzelt**	14	71	14
Gövde	15	60	24	**Üstlen**	16	70	12	**Kese**	24	59	16
İz	13	62	23	**Um**	18	73	8	**Avans**	16	72	11
Tatmin	21	61	16	**Opsiyon**	16	70	13				

Şekil 4.8: "+1" sınıfı için seçilen 50 kelimenin endeks fiyatında artışa neden olma yüzdesine göre kelime bulutu.

Şekil 4.9: "0" sınıfı için seçilen 50 kelimenin endeks fiyatının sabit kalmasına neden olma yüzdesine göre kelime bulutu.

Şekil 4.10: "-1" sınıfı için seçilen 50 kelimenin endeks fiyatında azalışa neden olma yüzdesine göre kelime bulutu.

Şekil 4.8'de, Şekil 4.9'da ve Şekil 4.10'da gösterilen kelime bulutları incelendiğinde, fiyat artışına neden olan kelimelerin başında "resim", "heyecan" ve "iz" gelmektedir. Fiyat azalışında ise "gerilim", "panik", "sarsıl" ve "kese" gibi kelimelerin etkisi ön plana çıkmaktadır. Fiyatın sabit kalmasında ise "dursun", "müdahale" ve "yüzlerce" gibi kelimeler etkindir.

5. SONUÇLAR

Yapılan tez çalışmasında gün içinde internet sitelerinde yayınlanan finansal haberlerin Borsa İstanbul (BIST) 100 endeksinin bir sonraki günkü açılış fiyatı yönüne etkileri incelenmiştir. Çalışmada BIST ile ilgili yapılmış diğer çalışmaların aksine haber dokümanların da geçen kelimeler öznitelik olarak kullanılmıştır. Haber dokümanları metin madenciliği teknikleri kullanılarak öznitelik vektörlerine dönüştürülmüştür. Aynı gün içinde yayınlanan haber metinlerine ait öznitelik vektörleri birleştirilerek her işlem günü için bir öznitelik vektörü elde edilmiştir. Öznitelik vektörlerine BIST 100 endeksinin günlük açılış fiyatındaki anlamlı değişiklikler göz önünde bulundurularak fiyat değişim yönünü gösteren sınıf etiketleri atanmıştır.

Oluşturulan öznitelik vektörleri binlerce özniteliğe sahiptir. Öznitelik seçme yöntemleriyle bu öznitelikler arasından bilgi içeren öznitelikler seçilmiş ve öznitelik vektörlerinin boyutu indirgenmiştir. Seçilen öznitelikler Naive Bayes sınıflandırıcısı eğitilmiş ve eğitilen sınıflandırıcının tahmin performansı doğruluk oranı, F-Ölçütü ve Makro-ortalama F-Ölçütü ile değerlendirilmiştir.

Borsa tahmin problemlerinde endeks fiyatındaki anlamlı artışın veya azalışın olduğu gün sayısı, endeks fiyatının sabit kaldığı gün sayısına oranla daha azdır. Aynı durum bu çalışmada da görülmektedir. Çalışmada kullanılan veri kümeleri dengesiz sınıf dağılımına sahiptir ve dengesiz dağılımlı sınıflandırma problemlerinde performans değerlendirilirken sınıflandırıcının genel tahmin performansıyla beraber sınıf seviyesinde tahmin performansı da göz önünde bulundurulmalıdır. Bu nedenle, çalışmada performans değerlendirme ölçütü olarak doğruluk oranıyla beraber Makro-ortalama F-Ölçütü kullanılmıştır. Makro–ortalama F-Ölçütü, sınıf seviyesinde bulunan F-Ölçütü oranlarının ortalaması alınarak bulunmuştur.

Öznitelik vektörlerinin boyutunun yüksek olması ve örnek sayısının az olması nedeniyle, haber metinlerinde bulunan kelimeler üzerinde öznitelik seçimi gerçekleştirilmiştir. Öznitelik seçiminde kullanılan ilk yöntem olan karşılıklı bilgi ile

anlamlı fiyat değişiminin olmadığı günler başarılı olarak tahmin edilmiştir.Karşılıklı bilgi, fiyatta anlamlı artışın veya azalışın olduğu günlerin tahmininde kötü performans göstermiştir ve sınıflandırıcının genel tahmin performansı düşmüştür.

Öznitelik seçiminde veri kümesindeki sınıf dengesizliği ile başa çıkabilmek için, karşılıklı bilgi öznitelik seçme yöntemi ile birlikte yeniden örneklemeyi temel alan yeni bir öznitelik seçme yöntemi ortaya konmuştur. Dengeli dağılımlı karşılıklı bilgi yöntemi olarak isimlendirilen bu yöntem kullanılarak BIST 100 endeksinin anlamlı yön değişimleri %74 doğruluk ve %68,4 Makro-ortalama F-Ölçütü oranlarıyla tahmin edilmiştir. Dengesiz sınıf dağılımının örnekleme işlemi ile dengeli hale getirilmesi ve öznitelik seçiminde bütün sınıf örneklerine eşit şans verilmesi sonucunda karşılıklı bilgi yönteminin performansında artış sağlanmıştır

Karşılıklı bilgi yönteminin haricinde Bilgi Kazanımı ve Ki-Kare istatistiği öznitelik seçme yöntemlerinin de sınıflandırma performansına etkileri incelenmiştir. Bilgi Kazanımı yöntemi ile öznitelik seçimi kullanıldığında %78,4 doğruluk oranıyla ile %68,1 Makro-ortalama başarısı elde edilmiştir. Ki-Kare istatistiği öznitelik seçimiyle ise %77,6'lık doğruluk oranı ve %66,7 Makro-ortalama F-Ölçütü başarısı bulunmuştur.

Dengeli dağılımlı karşılıklı bilgi, Bilgi Kazanımı ve Ki-Kare istatistiği yöntemlerinin sınıflandırma başarıları karşılaştırılmış ve en yüksek Makro-ortalama F-Ölçütü başarısına dengeli dağılımlı karşılıklı bilgi yöntemi ile ulaşıldığı görülmüştür. Dengeli dağılımlı karşılıklı bilgi bu başarıyı diğer yöntemlere göre daha az sayıda öznitelik kullanarak gerçekleştirmiştir.

Sınıflandırmada kullanılan haber dokümanlarının kaynakları göz önünde bulundurularak, bu kaynakların BIST 100 endeksi üzerindeki etkileri de incelenmiştir. Finans internet sitelerinden alınan haber dokümanlarının BIST 100 endeksine etkisi Borsa İstanbul'un resmi bildirimlerinin yapıldığı KAP internet sitesinden alınan dokümanlardan daha fazladır. Bu duruma, KAP internet sitesinde yayınlanan haber metinlerine yatırımcıların anında tepki vermesi ve bu haber metinlerindeki bilgilerin anında endeks fiyatına yansıması neden olmuştur.

Ekonomi haberleriyle birlikte BIST 100 endeksine ait geçmiş açılış, kapanış ve hacim fiyatları kullanılarak da BIST 100 endeksi yönünün tahmini yapılmıştır. Haber öznitelikleriyle fiyat özniteliklerinin birlikte kullanılmasının sınıflandırma

performansında iyileştirmeye yol açmadığı görülmüştür. Fiyat özniteliklerinin eklenmesi, sadece fiyat değişiminin olmadığı günlerin tahmin başarısı yükseltmiştir. Bu nedenle haber metinlerine fiyat verilerinin eklemesi sınıflandırıcının sınıfları ayırt etme başarısını arttırmamaktadır.

Deneylerin tamamlanmasından sonra, dengeli dağılımlı karşılıklı bilgi ile seçilen ilk 50 özniteliğin BIST 100 endeksi üzerindeki etkileri incelenmiştir. Her sınıf için farklı renklerde kelime bulutları oluşturulmuş ve BIST 100 endeksinin fiyat değişiminde ön plana çıkan kelimeler tespit edilmiştir. Fiyat artışında “resim” ve “heyecan”, fiyat azalışında “gerilim” ve “panik” ve fiyatın sabit kalmasında ise ”dursun” ve ”müdahale” gibi kelimelerin ön olana çıktığı görülmüştür.

KAYNAKLAR

Abdullah, M.H.L. ve Ganapathy, V. (2000). Neural Network Ensemble for Financial Trend Prediction. Proc. *TENCON 3*, 157-161.

Abu-Mostafa, Y. S. ve Atiya, A.F. (1996). Introduction to financial forecasting, *Applied Intelligence*, Volume 6, Issue 3, 205-213.

Akın A. A. ve Akın M. D. (2007). Zemberek, an open source NLP framework for Turkic Languages.

Alpaydın, E. (2004). Introduction to Machine Learning, *The MIT Press*.

Bildik, R. (2001). Intra-day seasonalities on stock returns: Evidence from the Turkish Stock Market. Emerging Markets Review, 2, 387–417.

Bildirici, M. ve Ersin, Ö. Ö. (2009). Improving forecasts of GARCH family models with the artificial neural networks: An application to the daily returns in Istanbul Stock Exchange, *Expert Systems with Applications 36,* 7355–7362.

Boyacioglu, M. A. ve Avci, D. (2010). An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the prediction of stock market return: The case of the Istanbul Stock Exchange, Expert Systems with Applications, Volume 37, Issue 12,7908-7912.

Chan Y., Chui, A. C. W., and Kwok, C. C. Y. (2001). The Impact of Salient Political and Economic News on the Trading Activity, *Pacific-Basin Finance Journal,* 9(3), 195-217.

Chawla, N., Japkowicz, N., ve Kolcz, A. (2004). Special Issue on Learning from Imbalanced Data Sets, volume 6(1), *ACM SIGKDD Explorations*.

Cover, T. M., ve Thomas, J. A. (1991). Elements of Information Theory*, New York: Wiley*.

Davis, J. ve Goadrich, M. (2006). The relationship between precision-recall and ROC curves, *In Proceedings of the International Conference on Machine Learning*.

Elkan, C. (1999). Notes on discovering trading strategies.

Gidófalvi, G. (2001). Using News Articles to Predict Stock Price Movements.

Gray, R. M. (2013). Entropy and Information Theory, *New York: Springer-Verlag*.

Guyon, I., ve Elisseeff, A. (2003). An introduction to variable and feature selection, Journal of Machine Learning Research, 1157–1182.

Guyon, I., Gunn, S., Nikravesh, M. ve Zadeh, L. (2006). Feature Extraction, Foundations and Applications, *Springer.*

Gündüz, H. ve Çataltepe, Z. (2013). Prediction of Istanbul Stock Exchange (ISE) direction based on news articles, *In Proceedings of the 3rd International Conference on Digital Information Processing and Communications (ICDIPC 2013)*, Dubai, UAE, 320–330.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. ve Witten, I. H. (2009). The WEKA Data Mining Software: An Update, *SIGKDD Explorations,* Volume 11, Issue 1, 10-18.

Hassan, M. R. ve Nath, B. (2005). Stock Market Forecasting Using Hidden Markov Model: a new approach, *Proc. Of 5th Int. Conf.on intelligent systems design and applications*.

Hellström, T. ve Holmström, K. (1998). Predicting the Stock Market, Technical Report Series IMATOM- 1997-07.

Kara, Y., Boyacioglu, M. A. ve Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange, *Expert systems with Applications*, 5311-5319.

Kim, K. J. (2003). Financial time series forecasting using support vector machines, *Neurocomputing,* 55(1/2), 307–319.

Kim, K. ve Han, I. (2000). Genetic Algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index, *Expert Syst. Appl. 19 (2),* 125–132.

Kohavi, R. ve John, G. H. (1998). The Wrapper Approach, in Feature Selection for Knowledge Discovery and Data Mining, *Kluwer Academic Publishers,* 33-50.

Lee, L. ve Chen, S. (2006). New Methods for TextCategorization Based on a New Feature Selection Method and a New Similarity Measure Between Documents, *Lecture Notes in Computer Science*.

Liu, T., Liu, S., Chen, Z. ve Ma, W. (2003). An Evaluation on Feature Selection for Text Clustering*, Proc. of Int'l Conference on Machine Learning.*

Manning, C. D., Raghavan P. ve Schütze, H. (2008). Introduction to Information Retrieval, *Cambridge University Press*.

Marcek, D. (2004). Stock price forecasting: Statistical, classical and fuzzy neural network approach, in MDAI, ser. *Lecture Notes in Computer Science,* V. Torra and Y. Narukawa, Eds., vol.3131. Springerpp. 41–48.

McCallum, A. ve Nigam, K. (1998). A comparison of event models for Naive Bayes text classification, *In AAAI-98 Workshop on Learning for Text Categorization*, 41-48.

Mitchell, M. L. ve Mulherin, J. H. (1994). The impact of public information on the stock market, *The Journal of Finance Vol. 49, No. 3, Papers and Proceedings Fifty-Fourth Annual Meeting of the American Finance Association*, Boston, Massachusetts, January 3-5, 923-950.

Mittermayer M. (2004). Forecasting intraday stock price trends with text mining techniques, *37th Annual Hawaii International Conference on System Sciences*, Proceedings of the, vol. 00, 64-73.

Montanes, E., Fernandez, J., Diaz, I., Combarro, E. F. ve Ranilla, J. (2003). Measures of Rule Quality for Feature Selection in Text Categorization. *Proc. 5th International Symposium on Intelligent Data Analysis* (IDA2003), Berlin, Germany, volume 2810, 589-598.

Nikfarjam, A., Emadzadeh, E. ve Muthaiyah, S. (2010). Text mining approaches for stock market prediction, *The 2nd International Conference on Computer and Automation Engineering (ICCAE),* 256-260.

Özgür, A., Özgür, L. ve Güngör, T. (2005). Text categorization with class-based and corpus-based keyword selection. *In Proceedings of the 20th International Symposium on Computer and Information Sciences*, volume 3733 of Lecture Notes in Computer Science, Springer-Verlag, 607–616.

Peng, H., Long, F. ve Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, 1226-1238.

Rogati, M. ve Yang, Y. (2002). High performing and scalable feature selection for text classification, *In Proceedings of the Eleventh International Conference on Information and Knowledge Management*, 659-661.

Schumaker, R. P. ve Chen, H. (2009). A Quantitative Stock Prediction System based on Financial News, *Information Processing and Management: an International Journal,* Volume 45 Issue 5, September, 571-583.

Wang, Y. ve Wang, X. J. 2005. A New Approach to Feature selection in Text Classification, In *Proceedings of the 4 International Conference on Machine Learning and Cybernetics IEEE*, vol.6, 3814-3819.

Wuthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J. ve Lam, W. (1998). Daily prediction of major stock indices from textual www data, in KDD, 364–368.

Yang, Y. (1997). An evaluation of statistical approaches to text categorization, *Technical Report CMU-CS-97-127*, Carnegie Mellon University.

Yang, Y. ve Liu, X. (1999). A re-examination of text categorization methods, *In Proceedings of SIGIR-99*, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, USA, 42–49.

Yang, Y. ve Pedersen, J. (1997). A comparative study on feature selection in text categorization*, In International Conference on Machine Learning.*

Zhang, H. (2004). The optimality of naive Bayes, *Proceedings of the 17th International FLAIRS conference (FLAIRS2004)*, Menlo Park, CA, AAAI Press, 562-568.

Anadolu Üniversitesi Açık Öğretim Fakültesi İstatistik Dersi Notları, 2010.

Url-1 : <*http://www.borsaistanbul.com***>**, alındığı tarih: 08.03.2013

Url-2 : <*http://www.kap.gov.tr***>**, alındığı tarih: 02.03.2013

Url-3 : <*http://bigpara.com* >, alındığı tarih: 05.3.2012

Url-4 : <*http://finans.mynet.com* >, alındığı tarih: 05.3.2012

Url-5 : < *http://www.bilgisayarkavramlari.com/2012/11/13/information-gain-bilgi-kazanimi/* >, alındığı tarih: 06.03.2013

Url-6 : <*http://nlp.stanford.edu/IR-book/html/htmledition/mutual-information1.html*>, alındığı tarih: 06.03.2013

Url-7 : <*http://web.sakarya.edu.tr/~adurmus/statistik/acikogretim/unite10.pdf* >, alındığı tarih: 06.03.2013

EKLER

EK A: Türkçe–İngilizce Terimler Sözlüğü
EK B: Uygulanan Öznitelik Seçme Yöntemleriyle Seçilen ilk 50 Kelime

EK A

Çizelge A.1: Türkçe–İngilizce terimler sözlüğü.

TÜRKÇE	İNGİLİZCE
Bilgi Kazanımı	**Information Gain**
Doğru Negatif	**True Negative**
Doğru Pozitif	**True Positive**
Doğruluk Oranı	**Accuracy**
Doküman Gösterimi	**Document Representation**
Doküman Sıklığı	**Document Frequency**
Duyarlılık	**Recall**
Dürümcü Yaklaşım	**Wrapper Approach**
Etkisiz Kelimeler	**Stop Words**
Gözetimli	**Supervised**
Gözetimsiz	**Unsupervised**
Hareketli Ortalama	**Moving Average**
İsim Öbekleri	**Name Phrases**
İsim Varlıkları	**Name Entities**
Kanıt	**Evidence**
Karşılıklı Bilgi	**Mutual Information**
Kelimeler Çantası	**Bag Of Words**
Kesinlik	**Precision**
Ki-Kare İstatistiği	**Chi-Square Statistics**
Metin Gösterimi	**Text Representation**
Metin Önişleme	**Text Preprocessing**
Olabilirlik	**Likelihood**
Olağanlık Çizelgesi	**Contingency Table**
Oto regresyon	**Autoregressive**
Önsel	**Prior**
Öznitelik Boyutunun İndirgenmesi	**Feature Reduction**
Öznitelik Çıkartımı	**Feature Extraction**
Öznitelik Seçimi	**Feature Selection**
Performans Değerlendirmesi	**Performance Evaluation**
Sınıflandırıcı Eğitimi	**Classifier Training**
Sınıflandırmadır	**Classification**
Sonsal	**Posterior**
Süzgeç Yaklaşımı	**Filter Approach**
Terim Sıklığı	**Term Frequency**
Ters Doküman Sıklığı	**Inverse Document Frequency**
Yanlış Negatif	**False Negative**
Yanlış Pozitif	**False Positive**

EK B

Çizelge B.1: Uygulanan öznitelik seçme yöntemleriyle seçilen ilk 50 kelime.

	İkili Değerli Veri Kümesi			Terim Sıklığı Değerli Veri Kümesi		
	Dengeli Dağılımlı Karşılıklı Bilgi	**Ki-Kare İstatistiği**	**Bilgi Kazanımı (Karşılıklı Bilgi)**	**Dengeli Dağılımlı Karşılıklı Bilgi**	**Ki-Kare İstatistiği**	**Bilgi Kazanımı (Karşılıklı Bilgi)**
1	tetik	sarsıl	sarsıl	tepki	tepki	tepki
2	sarsıl	resim	tetik	etkile	başla	basit
3	kasap	yarala	yarala	karşı	tansiyon	karşı
4	eğer	tetik	resim	düş	mensucat	risk
5	resim	gerilim	böylece	basit	dengele	tetik
6	anlaş	gövde	kasap	risk	günü	mensucat
7	Kıy	kıy	gerilim	seviye	sürü	kır
8	Amerika	usta	anlaş	yakın	borsa	başla
9	men	yatış	kıy	sabah	risk	hala
10	gerilim	kese	men	bıldır	piyasa	taraftar
11	Cay	kasap	eğer	fırsat	kır	dair
12	muhalefet	men	düzelt	git	basit	düş
13	nere	cay	dursun	başla	dair	tansiyon
14	panik	heyecan	gövde	tetik	tetik	etkile
15	dursun	anlaş	heyecan	yapı	politik	fırsat
16	usta	panik	cay	konu	akşam	Osman
17	tecrübe	dursun	yatış	yaşa	karşı	için
18	böylece	düzelt	nere	Avrupa	hala	sürü
19	yarala	tatmin	kese	günü	yayıl	dengele
20	gövde	nam	usta	kal	düşün	günü
21	yatış	tecrübe	panik	analiz	gergin	borsa
22	boyut	tehlike	nam	ılgı	şubat	kamuoyu
23	hale	nazım	nazım	büyük	polat	sabah
24	düzelt	tansiyon	tecrübe	dolar	pazartesi	dün
25	kese	muhalefet	muhalefet	dün	olası	yayıl
26	yerli	um	liman	için	hazine	yaşa
27	biraz	iz	Amerika	bit	hızlı	git
28	içer	kötüle	um	seyir	taraftar	dolar
29	TBMM	periyot	tür	doğrultu	test	sakin
30	tehlike	böylece	tehlike	piyasa	fırsat	şubat
31	heyecan	liman	tansiyon	ver	yarala	alternatif
32	kötüle	koy	ışık	hale	seviye	kimse
33	transfer	koş	tatmin	direnç	yaşa	plan
34	müdahale	ışık	kötüle	ülke	alternatif	herhangi
35	nazım	Mustafa	eşya	kır	inci	piyasa
36	yüzlerce	Avrasya	koş	tamamen	mısır	düşün
37	fayda	eşya	hale	al	yar	kur

Çizelge B.1 (devam): Uygulanan öznitelik seçme yöntemleriyle seçilen ilk 50 kelime.

38	bağlantı	olumsuz	bağlantı	şubat	yalanla	direksiyon
39	Yayıl	TBMM	sesle	hala	siyah	politika
40	Üstlen	teklif	iz	kaynak	direksiyon	odak
41	İz	tav	Mustafa	kaçın	yararlan	yarala
42	Tatmin	kota	TBMM	devam	odak	hazine
43	Um	sesle	Fransa	sarı	deneyim	pazartesi
44	Opsiyon	ıstıfa	Avrasya	bekle	depo	olası
45	Avans	bahar	periyot	çıkma	yatır	test
46	Eşya	bağlantı	içer	hayır	sabah	kişi
47	Çatış	tür	biraz	açı	anlam	akşam
48	Bahar	cinsi	ıstıfa	plan	dalga	yatır
49	Torun	torun	dinle	yukarı	trend	polat
50	Konya	bilanço	harç	tutar	gerilim	defa

ÖZGEÇMİŞ

Ad Soyad: Hakan GÜNDÜZ

Doğum Yeri ve Tarihi: Kırıkkale – 18.09.1986

Adres: İstanbul Teknik Üniversitesi Ayazağa Yerleşkesi Bilgisayar ve Bilişim Fakültesi Oda No:1208 Maslak-Şişli/İSTANBUL.

Lisans: Kocaeli Üniversitesi

E-posta: hakangun56@gmail.com, hakangunduz@itu.edu.tr

Yayın Listesi:

Gündüz, H. ve Çataltepe, Z. (2013). Prediction of Istanbul Stock Exchange (ISE) direction based on news articles, *In Proceedings of the 3rd International Conference on Digital Information Processing and Communications (ICDIPC 2013)*, Dubai, UAE, 320–330.