Ankara Üniversitesi
SBF Dergisi,
Cilt 74, No. 1, 2019, s. 1 - 34

Araştırma Makalesi

BORSA İSTANBUL'DA FİNANSAL HABERLER İLE PİYASA DEĞERİ İLİŞKİSİNİN METİN MADENCİLİĞİ VE DUYGU (SENTİMENT) ANALİZİ İLE İNCELENMESİ *

Dr. Suat Atan
Tarım ve Kırsal Kalkınmayı Destekleme Kurumu
Uzman
ORCID: 0000-0003-3170-0969

Doç. Dr. Yetkin Çınar
Ankara Üniversitesi
Siyasal Bilgiler Fakültesi
ORCID: 0000-0002-4710-0346

• • •

Öz

Şirketlere ilişkin haber metinleri ile finansal değerler arasındaki ilişkilerin nesnel bir temelde analiz edilebilmesini sağlamak için, öncelikle ilişkili haberlerin ve bu haberlerde yer alan olumlu ya da olumsuz ifadelerin sayısal değerlere dönüştürülmesi gerekmektedir. Bu amaçla literatürde "metin madenciliği" ve "Duygu (Sentiment) Analizi" yaklaşım ve yöntemleri kullanılmaktadır. Bu çalışmada da, Borsa İstanbul'da işlem gören BIST30 şirketlerine ilişkin 2014 yılında farklı kaynaklarda yayınlanmış 14.108 haber metin madenciliği teknikleri ile alınarak, yıllık ve çeyreklik bazda haber sayıları tespit edilmiştir. Haber içeriklerinde yer alan ifadeler de, Türkçe diline çevrilerek oluşturulmuş bir "Duygu Sözlüğü" yardımıyla, sayısal değerlere dönüştürülmüştür. Daha sonra, bu sayısal skorlar ile aynı dönemde piyasada oluşan şirket değerleri arasındaki ilişkiler analiz edilmiştir. Ortaya çıkan temel sonuç, finansal piyasalarla yayınlanan haberler ve bunların duygu tonları ile finansal değerler arasında anlamlı ilişkilerin var olduğudur. Bu sonuç Türk finansal piyasalarının değerlendirilmesinde önemli bir araç olarak Türkçe haber kaynaklarının da kullanılabileceğini ortaya koymaktadır.

Anahtar Sözcükler: Duygu analizi, Makine öğrenmesi, Finansal piyasalar, Metin madenciliği, Veri madenciliği

Relations Between Financial News and Market Capitalizations of Companies in BIST30 Index: Text Mining and Sentiment Analysis Methods

Abstract

In order to analyze the relations between financial values and the textual news related with the companies, there is a need for a conversion of textual content to quantitative values. For this aim, "text mining" approaches and "sentiment analysis" tools are used. In this framework, this study evaluates the relations between market capitalization of companies in Borsa Istanbul (BIST30) and published financial news about them. For this aim, published 14.108 news in the year 2014, have been extracted from 313 different news content providers and have been quantified by using text mining approaches. Then, for the statements in the news, a sentiment analysis has been performed by using a sentiment dictionary which is translated into Turkish in this study. Finally, the relations between these quantities and market capitalization values of the companies have been examined. The result of the analyses is the existence of relations between financial values of and published financial news about the companies. This study, via providing a methodological framework may help and show that the sources of financial news in Turkish also can be used as a novel tool for the analysis of Turkish financial markets.

Keywords: Sentiment analysis, Machine learning, Financial markets, Text mining, Data mining

* Makale geliş tarihi: 30.01.2017
Makale kabul tarihi:27.04.2018

Borsa İstanbul'da Finansal Haberler ile Piyasa Değeri İlişkisinin Metin Madenciliği ve Duygu (Sentiment) Analizi ile İncelenmesi

Giriş

Finans kuramının klasik savlarından biri olan Etkin Piyasa Hipotezi, piyasa fiyatının piyasalardaki herkes tarafından aynı anda ulaşılarak aynı şekilde anlaşılan ve rasyonel şekilde yorumlanan tüm bilgiyi yansıttığını iddia eder. Diğer taraftan, finans yazınında modern yaklaşımlardan biri olan Davranışsal Finans ekolü, "fiyat" olgusunun piyasada yatırımcılar tarafından "algılanan" fiyatı yansıttığını kabul eder. Bu varsayıma göre söz konusu algıyı etkileyen faktörler de önemli hale gelmektedir. Son dönem yazınında bu iki yaklaşımı uyumlaştırma çabalarından ortaya çıkan Adaptif Piyasa Hipotezi'ne göre ise; piyasalarda oluşan fiyat, piyasanın bulunduğu ortam ile piyasa aktörlerinin algı ve davranışlarının izin verdiği ölçüde piyasa bilgisini yansıtır (Lo, 2005; Urquhart ve Hudson, 2013).

Karar teorisinde de son dönemde gelişen yaklaşımlara göre, insanlar kararlarında "sınırlı rasyonellik"le hareket edebilir; özellikle büyük veri kümelerinin olduğu ortamlarda veriyi depolama kısıtı ve değerlendirmede uzmanlık eksikliği gibi nedenlerle belirli "höristiklere dayalı" kararlar verebilirler. Höristik kısaca, insanların karar verirken kullandıkları "zihinsel kısa yollar" yani basit ama etkili kurallardır. İnsanlar, örneğin çok boyutlu (kriterli) bir karar problemlerinde alternatiflerin tüm özellikleri yerine bazı özelliklerine bakarak değerlendirme yapabilirler. Tüm yönlerle değerlendirme yapmaksızın basit ve hızlı bir karar vermek rasyonelliğe aykırı gibi gözükse de, yapılan deneysel çalışmalar yaygın kullanılan höristikler yardımıyla kabul edilebilir doğrulukta seçimlerin yapılabildiği göstermiştir. Benzer şekilde, insanlar bilişsel işlemlerinde rasyonellikten sapmaya yol açan bazı eğilimlere veya önyargılara (bias) da sahiptir. Bu eğilimler de satın alma kararlarından yatırımcı davranışına kadar birçok alanda görülebilmektedir (Gigerenzer ve Todd, 1999; Tversky ve Kahneman, 1974).

Finans ve karar literatürünün ele aldığı bu çerçevede, finansal haber ve raporlar ile piyasalar arasında karşılıklı etkileşimin varlığı önemli bir araştırma

konusu haline gelmiştir. Nitekim günümüzde, insanların finansal piyasalarda yatırım yaparken klasik teorilerin öne sürdüğü analizleri yapmak yerine (veya bunun yanı sıra) belirli kısa yollara başvurarak karar verebildikleri; bu höristikleri oluşturdukları süreçte dayandıkları kaynaklardan birinin de yazılı haber metinleri (ve bunların içeriğinin yansıttığı durum algısı) olduğu bilinmektedir. Yatırımcılar, farklı kaynaklardaki haber metinlerini inceleyerek şirketler hakkında bilgi sahibi olmakta, olumlu ya da olumsuz duygular ve beklentiler oluşturmakta; bu bilgilere dayalı yorum ve duygularına bağlı olarak yatırım kararlarını gözden geçirmekte ve/veya yeni yatırım kararları vermektedirler. Başka bir deyişle, yatırımcılar bu haberlerin etkisi ile varlık portföylerini revize edebilirler.

Diğer taraftan, finansal piyasalardaki gelişmeler de doğal olarak haber içeriklerini etkilemektedir. Yani karşılıklı bir etkileşimden söz edebiliriz. Bu doğrultuda, medyada çıkan haberler ile piyasadaki yatırımcıların duygu ve davranışları ve/veya piyasalar arasındaki karşılıklı ve karmaşık ilişkileri inceleyen literatür de gittikçe genişlemektedir (García, 2013; Gidofalvi ve Elkan, 2001; Kearney ve Liu, 2013; Khadjeh ve diğerleri, 2014; Mitra ve Mitra, 2011; Ryan ve Taffler, 2002; Schumaker ve diğerleri, 2012; Tetlock, 2007; Wisniewski ve Lambe, 2013).

Finansal piyasalarda yatırımcılar genellikle kendilerine en çok tanıdık gelen hisse senetlerine yatırım yapmaktadır (Liu, Sherman ve Zhang, 2014; Nieuwerburgh ve Veldkamp, 2009). Firmaların yatırımcılara "tanıdık gelme" olgusunun ardında yatan en temel etmenlerden biri de firmalar hakkında çıkan haberlerdir. Özellikle bireysel yatırımcıların yatırım kararları dikkat çekme etkisi olarak adlandırılan psikolojik bir etki altında kalmaktadır. Dikkat çekme etkisi sınırlı süre ve kaynaklara sahip yatırımcıların dikkatlerini en çok çeken şirketlere yönelmelerini sağlamaktadır. Dikkat çekme etkisinin yaratılmasının yollarından biri ise haberlerdir (Barber ve Odean, 2012). Barber'in çalışmasında yatırımcıların halka açık haberlerde bahsi geçen şirketlerin hisse senetlerine ve ani yükselişe sahip hisse senetlerine yöneldikleri ortaya konmuştur. Bu çalışmada ifade edilen haber metinlerinin özellikle bireysel yatırımcılar üzerindeki etkisi dikkate alındığında haberlerin analizinin önemi ortaya çıkmaktadır. Haberlerin yatırımcıları etkileyen bir faktör olduğu göz önüne alındığında, finansal piyasalar hakkında analizlerde bir kaynak olarak kullanılmaları mümkündür. Bunun dışında finansal piyasalar hakkında yayınlanan sayısal veriler de mevcuttur. Finans yazını sayısal ve metinsel kaynaklar arasından genellikle sayısal kaynaklara yönelmektedir (Brun ve Teigen, 1988; Shang, Brooks, ve McCloy, 2014) .

Zira finansal haberlerle piyasalar arasında ilişkinin nesnel bir tabanda analiz edilmesi, öncelikle haber gibi sayısal olmayan, metin biçiminde ve

yapılandırılmamış verilerin analiz edilebilir (sayısal) değerler dizisine dönüştürülmesini gerektirir. Bu yapılırken haber içeriklerinin öznel yönünün (olumlu veya olumsuz içeriklerinin) tespit edilerek ayrıştırılması da önemlidir. Bu süreçteki en önemli zorluk, haber kaynaklarının sayıca fazla ve bu haberlerdeki metinsel içeriğin hacim bakımından büyük olmasından kaynaklanmaktadır. Bu durum, haber metinlerinin tek tek okunması ve değerlendirilmesini çok zor, hatta bazı durumlarda imkânsız hale getirmektedir. Bu nedenle, haber metinlerinin doğru bir biçimde analiz edilebilmesi için, analiz boyunca yanılgıyı en aza indirecek ve öznelliğe yer vermeyecek otomatik sistemlere ihtiyaç duyulmuştur. Finansal haberler ile piyasalar arasındaki ilişkilerin incelenmesi amacıyla "metin madenciliği" ve "makine öğrenmesi" adı verilen tekniklerden yararlanılmaktadır. Metin madenciliği, büyük ölçekli veriler arasından bilgi çıkarma (veri madenciliği) tekniklerinin bir alt dalıdır. Makine öğrenmesi ise, bilgisayar bilimlerinde yapay zekâ ile ilintili ve günümüzde birçok teknolojinin arka planında var olan önemli teknolojilerden biridir.[1]

Yazında bu alanda yapılan birçok çalışma, genellikle finansal piyasalar ile metinsel kaynaklar arasındaki ilişkileri "duygu analizi (sentiment analysis)" çerçevesinde analiz etmektedirler. Duygu analizinin anlamı, metinde geçen özel anlama sahip kelimelerin sayılarak bu kelimelere dayalı çeşitli skorların hesaplanmasıdır. Bu analiz sadece olumlu (pozitif) veya olumsuz (negatif) duygu tonunun hesaplanması şeklinde olabileceği gibi, daha detaylı duygusal durumlar (belirsizlik, kesinlik veya panik durumu vb.) için de olabilmektedir.

Duygu analizini finans alanında kullanan çalışmalar arasında güncel bir örnek olarak Bajo ve Romando (2016) şirketlerin ilk kez halka arz edilmeleri öncesinde yayınlanan haberlerin duygu analizini yapmışlardır. Bu çalışmada şirketlerin ilk halka arzları öncesinde yayınlanan teknik raporların sadece özel erişimi olan yatırımcılar tarafından analize konu edildiği, diğer yatırımcıların ise ikincil kaynak olan haber metinlerine yöneldiği ortaya konmaktadır. Bu analiz sonucunda yayınlanan haberlerin duygu tonu ile ilk kez halka arz edilen şirketlerin ilk gün sonu getirisi arasında ilişki bulunmuştur. Özellikle halka arz öncesinde birçok yatırımcının haberlere yöneldiği bilindiğine göre, haber

1 Bu araçlar sadece akademik değil, endüstriyel amaçlarla da kullanılmaktadır. Haber Analitiği (News Analytics) olarak bilinen ve haber metinlerinin sürekli olarak izlenmesi ve analizine olanak veren bilgi sistemleri sayesinde yatırımlar daha etkili olarak yapılabilmektedir. Bunun daha ileri kullanımı, önceden tanımlanmış dinamik yatırım stratejilerini insan müdahalesine gerek bırakmaksızın uygulayabilen, haber analitiği de dâhil olmak üzere birçok karmaşık modeli eş zamanlıya yakın hızda kullanarak yatırım süreçlerini otomatik olarak gerçekleştiren "Algoritmik yatırım" (Algorithmic Trading) araçlarında görülebilir. Bu araçlar, son dönemde uluslararası işlemlerde geniş kullanım olanağı bulmaktadırlar.

metinlerinin yatırımcılar üzerindeki etkisinin toplu olarak analizi incelemeye değer bir alan teşkil etmektedir. Bir diğer örnek olarak, Da, Engelberg ve Gao (2011) tarafından yapılan çalışmada ise şirketlerin Google'da aranma sıklıkları ile bu şirketlere yönelik yatırımcı ilgisi arasında ilişkiler bulmuştur. Bu ilişkiler ilk halka arzda da aynı şekilde görülmektedir. İlgili çalışmada, Google Trends hizmeti üzerinden derlenen ve arama sıklığını ifade eden arama sıklığı indeksi ile yatırımcıların ilgili araması arasındaki ilişkinin varlığı da dikkat etkisine bağlanmaktadır. Buna göre yatırımcılar arama motorlarında daha sık aradıkları şirketlere daha çok ilgi göstermektedirler. Anılan çalışmanın duygu analizi ile ilgili çalışmalarla ortak noktası dikkat etkisini yatırımcıların davranışlarının odak noktası olarak ele almasıdır.

Dünyada yukarıda sayılanlara benzer şekilde finansal haber metinlerini gerek metin madenciliği, gerekse duygu analizi kullanarak analiz eden ve son zamanlarda hızla gelişen bir uygulama literatürü olmasına karşın, Türkiye için bu kapsamda yapılmış bir çalışmaya rastlanmamıştır. Bunun temel sebeplerinden biri Türkçe haber kaynaklarının analiz edilmesini kolaylaştıracak bir yöntemsel çerçeve ya da yapının eksikliğidir. Bu çalışma da bu eksiliği bir ölçüde gidermeyi amaçlamaktadır. Bu temel düşünüşten hareketle, ayrıca Türkçe bir "Duygu Sözlüğü" oluşturma ile ilgili bir yaklaşımın da geliştirildiği çalışmamızda, konuya ilişkin Türkçe haber kaynaklarını ele alan bir analiz çerçevesi sağlanması temel hedef olmuştur. Bu doğrultuda çalışmada, Borsa İstanbul'da işlem gören (BIST30) şirketler ele alınarak; bu şirketler hakkında bir yıl (2014) süresince çeşitli kaynaklardan ulaşılan haberler ile şirket değeri arasındaki ilişkiler metin madenciliği ve duygu analizi teknikleri kullanılarak incelenmiştir.

Çalışmanın birinci bölümünde, finansal haber kaynaklarına dayanan metin madenciliği uygulamalarına ilişkin literatürde kullanılan veri kaynakları ve yöntemsel çerçeve tanıtılmaktadır. İkinci bölümde ise uygulama yöntem ve sonuçları sunulmaktadır. Bu bölümde, Borsa İstanbul'da işlem gören BIST30 şirketleri için bir yılda çıkan haberler metin madenciliği ve haberlerdeki ifadeler duygu analizi ile incelenerek elde edilen sonuçlar ile şirket değerleri arasındaki karşılıklı ilişkiler, korelasyon katsayıları hesaplanarak ortaya konulmuştur. Çalışma genel değerlendirmenin yapıldığı bölümle sonlandırılmaktadır.

1. Metin Madenciliği ve Duygu Analizinin Veri Kaynakları ve Yöntemsel Çerçevesi

Bu bölümde, çalışmada gerçekleştirilen uygulamanın temelini oluşturan analiz altyapısı tanıtılacaktır. Bu doğrultuda, finansal haber kaynaklarına dayanan metin madenciliği ve duygu analizi uygulamalarına ilişkin literatürde

kullanılan veri kaynakları ve yöntemsel çerçeve açıklanmaktadır. Bu açıklamalar iki alt bölümde ele alınacaktır. Birinci alt bölümde, finansal metin madenciliğinde kullanılan haber kaynakları sınıflandırılarak tanımlanmış; ikinci alt bölümde ise, bu konudaki analizlerin yöntemsel çerçevesi hipotetik bir örnek yardımı ile açıklanmıştır.

1.1. Metinsel Veri (Haber) Kaynakları

Yatırıma dönük analizlerde kullanılan temel metinsel veri kaynakları aşağıda kısaca açıklanmıştır. Bu analizler, genellikle belirli bir kaynaktan düzenli aralıklarla çıkan, yapı ve zaman bakımından standart haberler üzerinden yapılmaktadır.

Şirket Yayınları

Şirketlerin düzenli olarak yayınlamakta oldukları yılsonu finansal raporlar ve diğer faaliyet raporları metin madenciliğine konu olmaktadır. Genellikle uygulamalı çalışmalarda, söz konusu rapor metinlerinde geçen ifadelerle şirketin finansal verileri arasındaki açık ya da zımni ilişkiler incelenmektedir. Şirket yayınlarına dayanarak geleceğe dair finansal tahminler elde etmeye yönelik çalışmalar da mevcut olmakla birlikte, bu verilerin en fazla çeyreklik dönemler gibi düşük bir frekansta yayınlanması nedeniyle şirket kaynaklarının zaman serisi analizleri için ideal olmadığı ileri sürülmektedir (Kearney ve Liu, 2013).

Medyada çıkan haberler

Medyada çıkan haberler de metin madenciliği teknikleri için kaynak teşkil etmektedir (Khadjeh Nassirtoussi ve diğerleri, 2014). Yatırımcıları etkileyen önemli kaynaklardan biri de basılı ve (sanal) görsel medyadır. Gazeteler, televizyonlar, haber ajansları, haber derleyici siteler ve başka farklı kaynaklar sayesinde yayınlanan haberler yatırımcılara ulaşmakta ve bu haberler yatırımcıları etkilemektedir. Finansal piyasalarla ilgili haberler yatırımcıları etkilediği gibi, haber üreten kaynaklar da finansal piyasalardaki gelişmelere dayanarak haber üretmektedir. Bu bağlamda birbirini besleyen etkileşimli bir süreç bulunmaktadır.

Haber kaynaklarını metinsel veri olarak ele almak suretiyle analiz eden araştırmacılardan Davis ve diğerleri (2001) ve Tetlock (2007), Wall Street Journal ve New York Times gazetelerinde ortaya çıkan haberleri ele almışlardır. Aynı şekilde, Engelberg, García ve Sinha'nın (2008;2013;2015) çalışmaları da haber metinleri ile duygu analizi gerçekleştiren güncel araştırmalar arasında sayılabilir.

Medyada çıkan bir finansal haber metninin bir veri kaynağı olarak kullanılmasındaki temel hedef, daha çok gelecekte nelerin planlandığını ve öngörüldüğünü ortaya koymak, bu bilgileri ortaya çıkarmaktır (Kearney ve Liu, 2013). Analiz edilecek haber kaynaklarının hangi yayınlar olacağı konusunda ise literatürde farklı yaklaşımlar görülmektedir. Sadece birkaç yayın (gazete veya ajans) haberlerini veya çok geniş sayıda haber kaynağını kullanan çalışmalar bulunmaktadır. Ferguson ve arkadaşları da mümkün olan en geniş sayıda yayını analize dâhil etmenin daha doğru olduğunu ifade etmektedirler (Ferguson, Philip, Lam ve Guo, 2014).

Sosyal Medya

Sosyal medyada oluşturulan içerikler haber kaynaklarında göre çok daha geniş bir kitle tarafından özgürce oluşturulduğundan ve ortaya çıkışı bakımdan haberlere göre çok daha yüksek frekanslı olduğundan ve hızlı geliştiğinden; sosyal medya, metin ve duygu analizlerinde ağırlıklı olarak kullanılan önemli bir kaynak olmaktadır.

Bu temel avantajlarının yanı sıra bu kaynağın kullanılmasının bazı dezavantajları da vardır:

Birincisi, "gürültü" olarak tabir edilen ve piyasada ortaya çıkan gelişmelerin kasıtlı veya kasıtsız olarak ilgisiz veya yanlış bilgiler içeren metinlere dönüşmesidir. Örneğin bazı sosyal medya kampanyaları nedeniyle sosyal medyaya aşırı miktarda içerik aktarılmaktadır. Sosyal medya sitelerinin çalışma ilkesi genelde en çok konuşulan konunun en üste çıkarılması şeklinde olduğundan, bireyler bunu sağlamak adına, bilgi içersin içermesin çok fazla sayıda içeriği, bazen mükerrer biçimde sosyal medyada paylaşmaktadır. Bu durum ise metin ve duygu analizlerinde sapmalar yaratabilmekle birlikte, analizi teknik olarak da zor hale getirmektedirler. İkinci dezavantaj ise; resmi haber sitelerinde yazım hataları nadiren görülmesine rağmen sosyal medyada çok fazla sayıda yazım hatasının bulunmasıdır. Yazım hataları ise metinsel verileri analiz eden algoritmalar için sorun teşkil etmektedir. Daha da ötesinde, "emoji" olarak adlandırılan metinsel karakterlerden ikonlar oluşturma alışkanlıkları, aşırı ironik dil, kelimelerdeki bazı harflerin gereksiz biçimde uzatılması ve tekrarı gibi alışılagelmiş yazım stilleri metinsel verilerin analizini güç hale getirmektedir (Miner, IV ve Hill, 2012).

1.2. Biçimsel Altyapı ve Analiz Adımları

Bu bölümde metin madenciliği alanında kullanılan temel yöntemsel (biçimsel) çerçeve ve analiz adımları basit bir hipotetik örnek üzerinden açıklanacaktır. Bu örnek bu çalışma kapsamında gerçekleştirilen uygulamanın işleyiş şeklini ortaya koyma adına esas uygulamanın daha basitleştirilmiş halidir.

Metin madenciliği sürecindeki kavramlar da bu örnek senaryo üzerinden ifade edilmiştir.

Metinsel veri analizinin gerçekleştirilebilmesi, yani bu metinlerin makineler tarafından anlaşılıp analiz edilebilmesi için, öncelikle bu verilerin sayısal değerlere dönüştürülmesi gerektiğini belirtmiştik. Bu dönüştürme işlemi de metin madenciliği teknikleri kullanılarak ve genellikle bilgisayar programlama dilleri (Java, Python, C/C#) veya istatistiksel/matematiksel paket yazılım kodlama dilleri (RStudio, Matlab) ile yazılan kodlarla gerçekleştirilmektedir.

Bu sayısallaştırma öncelikle konu ile (örneğin finans) "ilişkili" haberlerin diğer haberlerden ayrılarak alınmasını ve sayılmasını içerir. Haber kaynaklarından alınan metinlerde örneğin hangi haberin olumlu, hangilerinin olumsuz veya nötr duygu oluşturacağını tespit ederek bir etiketlendirme yapmak ise duygu analizinin konusudur. Pozitif haberler p, negatif haberler n ve nötr haberler z olarak ifade edilirse, haberlerin bu 3 kategorik değerle etiketlenmesi ile aşağıdaki gibi bir tablo elde etmek bu analizin hedefi olacaktır.

Bu tablo **etiketlenmiş metinsel veri tablosu** olarak adlandırılır:

Tablo 1: Etiketlendirilmiş metinsel veri tablosu

Haber (n_i)	Kategori
n_1	p
n_2	p
n_3	n
n_4	z

İstenen sayıda ilişkili haber metninin insan müdahalesi olmadan, etiketlenmiş olarak elde edilmesi metin madenciliği çalışmalarında gerçekleştirilen görevlerden biridir.[2] Bu etiketlemenin nasıl gerçekleştirileceği sorununun yanı sıra, ne tür etiketlemeler yapılabileceği sorununun çözümü için geliştirilen yaklaşımlar da farklılaşabilir. Örneğin etiketlenecek metnin içeriğinin

2 Etiketleme işlemi az sayıda haber için insanlar tarafından sübjektif olarak değerlendirme yoluyla da gerçekleştirilebilir. Ancak, haber metinlerinin sayısı insanların değerlendirebileceği miktardan fazla olduğunda bu imkânsız hale gelecektir. Örneğin bu çalışma kapsamında 14.108 haber analiz edilmiştir.

hangi konuda (spor, hayat, politika, teknoloji, finans) olduğuna ya da metin içeriğinin verdiği mesajın pozitif, negatif, nötr gibi kategorilere ayrılmasına ilişkin etiketleme yapılabilir.[3] Çalışmamızda metinsel veri sayısallaştırılmasında, konu ile "ilgili / ilişkili" haberlerin ayrılması ve duygu sözlüklerine dayalı bir yaklaşım kullanılarak, metinlerin duygu içeriklerinin tespiti (olumlu-olumsuz ve nötr kategorilerine ayırma) işlemleri gerçekleştirilecektir.

Bu noktada, bu çalışmadaki metin madenciliği uygulamasının da esasını oluşturacak biçimsel altyapı ve analiz adımları aşağıdaki şekilde ifade edilebilir:

1. Adım. Veri Kütlesi (Korpus) Oluşturularak İlgili Haberlerin Ayrılması (Filtering)

2. Adım. Doküman-Terim Matrisinin (DTM) Oluşturulması ve Sadeleştirilmesi

- ✓ Kelime Köklerine İnme (Stemming)
- ✓ Temizleme (Cleaning)

3. Adım. Dökümanların Etiketlenmesi ve Duygu Analizine Hazırlık

4. Adım. Duygu Sözlüğü Oluşturma ve Analiz

Bu adımlar aşağıda açıklanmıştır.

"Korpus"un Oluşturulması: Haberlerin alınması ve İlgili Haberlerin Ayrılması (Filtering)

Korpus, bir tür doküman veri tabanıdır. Metin madenciliğinde sıkça kullanılan bir kavram olan korpus, "üzerinde metinsel analiz yapılacak ana veri kütlesi" anlamına gelir. Metin dosyalarının bir arada durduğu bir dizini temsil edebileceği gibi, bu metinlerin bir araya gelmesi ile oluşan büyük dosya veya veri tabanını da temsil edebilir. Örneğin içerisinde bir şirkete ait dosyalar, e-postalar ve resmi yazışmalar olabilir. Yine herhangi bir konuda çıkan derlenmiş haberler de korpusu teşkil edebilir.[4]

---

3 Günümüzde birçok gelişmiş web sayfasında, barındırılan mevcut içeriğin konusu bu şekilde dış müdahale olmaksızın belirlenebilmektedir. Sık kullanılan ve iyi bilinen diğer bir sınıflandırma, e-postalarımızı "gereksiz (spam)" olup olmadıkları biçiminde etiketleyen algoritmalardır.

4 Normalde tüm bilgisayar programlama dillerinde metinler "string" olarak adlandırılan nümerik olmayan formatta hafızada tutulurlar. Ancak korpus nesnesi doküman terim

Herhangi bir alanda, d_i tekil bir doküman olmak üzere, $D = d_1, d_2, \ldots d_n$ şeklinde dokümanlar mevcut olsun. Bu durumda D, "korpus" olarak adlandırılmaktadır.

Hipotetik Örnek (1. Adım):

XFinans adlı bir şirkete ait veya onunla "ilgili" haberlerin tutulduğu bir korpus mevcut olsun. Anlatım ve gösterimde basitliği sağlamak adına bu korpusta herhangi bir zaman diliminde şirket hakkında çıkmış 5 adet kısa metinli haber bulunduğu varsayılacaktır:

n_1*: XFinansın değeri ciddi şekilde arttı*

n_2*: Piyasalardaki durgunluk XFinans'ı etkilemedi*

n_3*: Küresel kriz XFinans'ı da vurdu*

n_4*: XFinans CEO'su etkinliğe katıldı*

n_5*: XFinans'ın değeri artmaya devam ediyor*

Doküman-Terim Matrisinin (DTM) Oluşturulması

Metin biçimindeki verilerin sayısal hale getirilmesi için ikinci aşamada "Doküman Terim Matrisi (DTM)" oluşturulur. İsmindeki "doküman" ifadesi ayrışık metinsel nesneyi ifade etmektedir. Bir kütüphanedeki her kitap bu tanıma göre dokümandır. Bir haber sitesindeki her bir haber ise yine doküman olarak tanımlanır. "Terim" ise bu dokümanlarda geçen her bir kelimeyi işaret etmektedir. DTM genellikle bir dokümanda belirli bir kelimenin var olduğunu veya olmadığını sırasıyla 1 ve 0 değerleri ile işaretleyen bir matristir. Aynı kelime aynı dokümanda birden fazla varsa ilgili değer 1'den büyük olacaktır. Bu matrisin sütunlarında terimler, satırlarında ise doküman adları veya tanımlayıcıları bulunmaktadır.

Hipotetik Örnek (2. Adım):

XFinans ile ilgili haberler için bu matris aşağıdaki gibi olacaktır.

---

matrisine çevrilmeye hazır, tüm metinlerin bazı özellikleri ile birlikte tek nesnede tutulduğu matris benzeri bir yapıdır.

Tablo 2: Doküman- Terim Matrisi (Hipotetik Örnek)[5]

Doküman | Terim:	XFinansın	değeri	ciddi	…	arttı	artmaya	devam	ediyor
n_1	1	1	1	…	1	0	0	0
…	…	…	…	…	…	…	…	…
n_5	1	1	0	…	0	1	1	1

DTM; n adet doküman ve m adet tekil terim varsa, n x m boyutlu bir matris olmaktadır. Yukarıdaki Tablo 2'de sunulan matriste, dikkat edilecek olursa "arttı" ve "artmaya" şeklinde geçen kelimeler, kökleri aynı olmasına karşın matriste ayrı kelimeler gibi ifade edilmişlerdir. Ancak bu kelimeler anlam ve duygu açısından aynı anlama gelmektedir ve ayrı gösterilmeleri gereksizdir. Bu durum büyük boyutlu matrislerde, büyük bir zaman ve sistemsel kaynak israfına neden olacaktır. İşte bu problemi aşmak için metin madenciliğinde yukarıda örnek olarak gösterilen doküman terim matrisi oluşturulurken kelime köklerine inilerek sadeleştirme işlemleri yapılmaktadır.

Doküman-Terim Matrisinin Sadeleştirilmesi: Kelime Köklerine İnme (Stemming)

Doküman-terim matrislerinin önemli bir özelliği de hücrelerinin ağırlıklı olarak sıfırlardan oluşmasıdır. Bu nedenle bu matrisler "seyrek matrisler" olarak nitelendirilir ve gerektiği takdirde düşük frekanslı terimler elenerek matrisin seyrekliği azaltılabilmektedir. Konumuz özelinde bu özellik kelime köklerine inme işlemi ile birlikte kullanılarak matris sadeleştirilmektedir. Bu süreçte, korpustaki dokümanlarda geçen her bir kelimenin kökleri atılarak kelimelerin temeline inilmektedir.

Bu işlemi bazı metinsel veri kütüphaneleri ve araçları, İngilizce ve bazı yaygın diller için tek fonksiyonla kolayca gerçekleştirebilmektedir. Ancak bu çalışmada kullanılan R dili kütüphanesi içerisinde Türkçe için bu özellik henüz bulunmamaktadır. Bu nedenle Türkçede kelime köklerine inmek için mevcut olan araçlar araştırılmıştır. Sonunda, açık kaynaklı olarak kullanımda olan ve

5 Buradaki DTM, normalde 5 satır ve 21 sütundan (Toplam 23 kelimeden, tekil olan 21 terimden ötürü) oluşması gerekirken, gösterimde kolaylık olması açısından sadeleştirilerek sunulmuştur.

Java dili ile geliştirilmiş Türkçe Doğal dil işleme kütüphanesi olan "Zemberek Kütüphanesi"nin kullanılmasına karar verilmiştir[6]. Ancak bu kütüphane kelime köklerine inme işlemini metin korpusu bazında tek seferde gerçekleştirememektedir; nitekim amacı metin madenciliği değildir. Bu nedenle çalışma kapsamında Zemberek Kütüphanesi kelime köklerine inme işlerini fonksiyonel olarak gerçekleştirecek şekilde uyarlanmıştır. Bu uyarlamanın esasları ve kütüphanenin analiz sürecine ne şekilde dâhil edildiği uygulama bölümünde açıklanacaktır.

Hipotetik Örnek (3. Adım):

Yukarıda gösterilen örnek doküman terim matrisinin kelime köklerine inilerek yeniden hazırlanması ile Tablo 3'te görülen matris elde edilir:

Tablo 3: Kelime Köklerine İnilmiş Terimlerle Doküman Terim Matrisi

Doküman/Terim	XFinans	değer	ciddi	....	art	devam	et
n_1	1	1	1	..	1	0	0
n_5	1	1	0	..	1	1	1

Artık kelime köklerine inilmiş olduğundan bir önceki tabloda "arttı" ve "artmaya" şeklinde ayrı ayrı gösterilen kelimeler yerine kelime / fiil kökü olan "art" sözcüğü gösterilmiştir. Böylece matris bir sütun azalmıştır. Bu durum bu basit örnekte sadece tek sütunda azalma yaratmıştır ancak bu çalışmanın uygulama bölümünde de bir benzeri olduğu gibi, binlerce haber ve yüzbinlerce terimden oluşan büyük matrislerde önemli bir kazanç sağlamaktadır.

Bu örnekte kelime köklerine inilmesi DTM'nin sadeleştirmesi için yeterli olmuştur. Ancak özellikle sosyal medya üzerinde metinsel madencilikte ön aşama işlemine eklenmesi gereken bir temizlik süreci de emoji olarak nitelendirilen, harf ve sayılardan oluşan karakterler dışındaki karakterlerin (gülücük, kızgınlık, el çırpma vs.) ifadelerinin temizliğidir. Emojilerin temizlenmesi görece kolay ise de yazım hataları çok çeşitli olduğundan bu hataların ayrıştırılarak orijinal kelimesine dönüştürülmesi görece daha fazla uğraş gerektirmektedir. Haber metinlerinin bu noktada en önemli avantajı emoji kullanımı ve yazım hatalarının neredeyse hiç olmamasıdır.

6 https://github.com/ahmetaa/zemberek-nlp

Dökümanların Etiketlenmesi ve Duygu Analizine Hazırlık

Doküman-Terim matrisi bir kez elde edildikten sonra metin madenciliği içerisinde birçok amaca yönelik olarak kullanılabilir. Bu amaçlardan en yaygın örneği, cümle içerisinde en çok tekrar eden kelimelerin bulunmasıdır. Bu basit yöntemle çok büyük metinlerin içerikleri hakkında tek bakışta kolayca bilgi edinilebilmektedir. Bir korpusta sık tekrar eden kelimeler, korpus içerisindeki dokümanların etiketlenmesi için kullanılabilmektedir. Duygu analizi de elde edilen ve etiketlenen DTM üzerinden işlemektedir.

En sık tekrar eden kelimenin bulunması doküman terim matrisinde sütun toplamlarının en büyükten en küçüğe doğru sıralanması ile basitçe elde edilebilir. En sık tekrar eden terimlerden, sıklık değerine göre kelimelerin puntosunu büyüterek bu kelimelerin bir arada gösterilmesi yöntemine ise "kelime bulutu" yöntemi denmektedir. Bu yöntem ise yine büyük metinsel verilerin içeriğini görselleştirmede ve hızlıca anlaşılabilmesini temin etmede işe yarar yöntemlerden biridir. Bir örnek olması ve yapılan analizleri görselleştirebilmek açısından Şekil 1'de Garanti Bankasına ait 2014 yılında yayınlanan haberlerin kelime bulutu gösterilmektedir. Kelime bulutunda en sık tekrar eden kelimelerin puntoları daha büyük görünmektedir.

Şekil 1. Bir Şirkete Ait Kelime Bulutu

Bu çalışmada da sık tekrar edilen kelimelerin yanı sıra terimlerin içerdiği anlamların verdiği duygu tonu, olumlu-olumsuz ve nötr olarak işaretlenmiştir. Bu işleme duygu analizi adı verilmektedir.

Duygu analizlerinde kullanılan iki temel araç grubu aşağıda kısaca açıklanmıştır. Bu araçlardan sadece biri veya bir kaçı bir arada kullanılmak suretiyle analizler gerçekleştirilmektedir. Bu çalışmada metinsel verinin analizinde kullanılan yaklaşımlardan "duygu sözlükleri" ile analiz (sözlük bazlı yaklaşım) ele alınacaktır.

Duygu Sözlükleri ile Analiz

Haber korpusu içerisinde yer alan haberlerin duygu tonlarını insan müdahalesi olmaksızın belirleyebilmek ve haberleri buna göre etiketlemek için kullanılabilecek yaklaşımlardan biri, bu çalışmada da kullanılan "sözlük bazlı (leksikal)" yaklaşımdır. Duygu sözlükleri ile analiz, çeşitli duygu durumlarına ilişkin en basit haliyle "olumlu", "olumsuz" ve "ne olumlu ne olumsuz"verilen kategoriler için özel olarak hazırlanmış sözlüklerdeki kelimelerin korpusta geçme sıklığına göre haberin "pozitif", "negatif" veya "nötr" olarak nitelendirilmesi / etiketlenmesi metodudur.

Duygu sözlüğü, insanların yazdıkları metinlerde herhangi bir duygu ile ilgili olduğu büyük olasılıkla bilinen tüm kelimelerin derlenmesi ile oluşturulan ve her duygu için sözcük listelerinden oluşan setlerdir (Agarwal ve Mittal, 2015, s. 64). Duygu sözlükleri, metin madenciliğinde bilgisayar yardımı ile herhangi bir metnin duygu durumunu bu duygu sözlüklerini referans alarak ortaya çıkarmak için dayanak noktası teşkil ederler. Daha açık olarak; *olumlu* ve *olumsuz* duygu kategorileri için özel olarak hazırlanmış iki ayrı sözlükteki kelimelerin haberlerde geçme sıklığına göre haber pozitif veya negatif olarak nitelendirilir. Eğer her iki sözlükteki pozitif veya negatif kelimeler bir haber içinde geçmiyorsa bu haber nötr olarak değerlendirilecektir. Örneğin; "başarı, güven, azim, kararlılık, kar" gibi kelimeler olumlu duygu sözlüğünde yer alırlar ve bu kelimelerin geçtiği metinler olumlu kategorisinde etiketlenirler. Aynı şekilde "iflas, zarar, istifa" gibi kelimeler ise olumsuzluk ifade ettiklerinden negatif duygu sözlüğünde yer alacak ve bu kelimeleri içeren metinler negatif etiketlenecektir.

Metinsel verileri analiz etmek için duygu analizi dışında kullanılan yaygın yöntemlerden biri de makine öğrenme algoritmalarının kullanımıdır. Makine öğrenme algoritmaları ise duygu sözlükleri olmaksızın, metinsel verilerin daha önce hazırlanmış eğitim verileri (training data) ile eğitilmeleri suretiyle, metinlerin içerik yapıları üzerinden duygu durumlarının analiz edilmesini sağlayan araçlardır. Makine öğrenme algoritmalarının kullanımındaki temel zorluklardan biri algoritmaları eğitmek için kullanılacak verilerin insan eliyle teker teker tasnif edilmesi zorunluluğudur (Kearney ve Liu, 2013).

Bu çalışmada da kullanılan sözlük bazlı yaklaşımda dökümanların içerdikleri duygu tonu bakımından etiketlenerek analiz edilmesi işlemi, ele aldığımız örnek üzerinden aşağıda açıklanmaktadır:

Hipotetik Örnek (4. Adım):

XFinans örneği için, pozitif terimler sözlüğü Lp={'değer','art'}, negatif terimler sözlüğü ise Ln={'durgun','kriz'} kelimelerinden oluşsun. Doğaldır ki gerçek duygu sözlükleri çok daha fazla sayıda kelime içermektedir.

L_p pozitif terimler sözlüğünde bulunan kelimeler n_1 ve n_5 haberleri içerisinde 2 kez geçmektedir. Bu durumda basitçe bu iki haber pozitif olarak etiketlenecektir ki, bu etiketleme doğru yapılmış olmaktadır. Ancak L_n sözlüğünde geçen "durgun" kelimesi n_2 haberinde geçtiğinden bu haber negatif olarak etiketlenecektir. Oysa bu haberde ifade *"durgunluk XFinansı etkilemedi"* şeklinde ifade edilmişti. Buna göre haber negatif olmadığı halde negatif olarak etiketlenmiş olmaktadır. Ancak n_3 haberinde ise "kriz" kelimesinin varlığından ötürü haber isabetli bir şekilde negatif olarak etiketlenmiş olmaktadır. n_4 içerisinde her iki sözlüğe tekabül eden hiç bir kelime bulunmadığından sözlük bazlı etiketleme içerisinde bu haber nötr (z) olarak değerlendirilecektir. Bu açıklamalar aşağıda özet bir tablo olarak sunulmuştur:

Tablo 4: Sözlük Bazlı Etiketleme Matrisi ve Sonuçların Doğruluğu

Haber	Metin	İnsan yorumu	Sözlük bazlı makine yorumu	İsabetlilik
n1	XFinansın değeri ciddi şekilde arttı	p	p	doğru
n2	Piyasalardaki durgunluk XFinansı etkilemedi	p	n	yanlış

Etiketleme işleminde isabetlilik düzeyinin artırılması için duygu sözlükleri hazırlanırken veya seçilirken birçok faktöre dikkat etmek gerekmektedir. Bu faktörler ve çalışma kapsamında bulunan çözüm yolları aşağıdaki alt bölümde açıklanmaktadır.

Duygu Sözlükleri Literatürü ve Türkçe Duygu Sözlüğü

Sözlük bazlı yaklaşımda, haberlerin makine tarafından araştırılan değerlere göre kategorik olarak etiketlenmesi işleminde; etiketler, her bir kategoriyi temsil ettiği düşünülen sözlükler aracılığı ile verilmektedir. Bu nedenle duygu sözlükleri çok önemlidir. Diğer taraftan, her dilde kullanılabilen ve birçok amaca aynı anda hizmet edebilen bir sözlük oluşturmak da, uygulamalarda etiketleme işleminin hatasız yapılması da mümkün değildir. Bu nedenle ilgili sözlüklerin hazırlanması dilbilim ve bilgisayar bilimlerini yakından ilgilendiren ve her duruma özel çalışma gerektiren bir konudur. Nitekim analiz edilen dilde ele alınan duygularla ilgili kelimelerin tamamının, dilbilimsel yapıları da göz önüne alınarak ele alınması gerekir. Diğer bir anlatımla, iyi bir sözlüğün oluşturulması bu konudaki zorlukların ve özel durumların farkında olmakla mümkün olabilmektedir.

Bu zorluklar öncelikle doğal dilin karmaşık yapısından kaynaklanmaktadır. Biçimsel dil veya bilgisayar dillerinde nesnelerin kategorileri çok net bir şekilde tanımlanmış iken (örneğin s önermesinin değillemesi $!s$ şeklinde gösterilmektedir); doğal dil için bu o kadar kolay olmamaktadır. Çünkü her dilde sözcüklerin anlamları ekler, kelimeler veya başka kelimelerle değişebilmektedir. Örneğin, Türkçede "iyi" kelimesinin aksi "iyi değil" ve "kötü" şeklinde iki farklı kelime ile ifade edilebilmektedir. Burada ilk yapıda anlam ikincil bir sözcükle, diğer yapıda ise başka bir sözcük kullanımı ile değişmektedir.

Ayrıca iyi bir duygu sözlüğü, sadece ele alınan duygu durumuna dair kelimeleri içermekle kalmamalı, bu sözlüğe dayalı özel analizlerde de durumu gerçeğe yakın şekilde yansıtabilmelidir. Zira bu sözlüklerin duygu analizinde kullanımı ile ilgili olarak en önemli problemlerden biri, ilgili sözlüklerin hazırlanması sürecinde dilbilimsel olarak genel olarak *günlük dil* içerisinde kullanılan olumlu/olumsuz veya diğer duygu ifadelerini içeriyor olmasıdır. Bu durumda, farklı kapsamda ve özel olarak yapılacak analizlerde, örneğin ekonomi alanında farklı algılanan kelimeler analizde sapmalara neden olabilecektir. Bu problem literatürde bilindiğinden, birçok araştırmacının birden fazla sözlük kullanarak ayrı ayrı analizler yaptıkları, aynı zamanda farklı sözlüklerin özelliklerini inceledikleri görülmektedir (Kearney ve Liu, 2013; Khadjeh Nassirtoussi ve diğerleri, 2014).

En basit duygu sözlüğünde sadece *olumlu* ve *olumsuz* duygu durumları ile ilgili sözcük listeleri vardır. Ancak, duygu durumları basitçe olumlu ve olumsuzdan ibaret de olmayabilir. Örneğin, bunların dışında *güven, güvensizlik, öfke, tatmin, iğrenme* gibi başka duygu durumlarının da kategorik değer olarak etiketlenmesi gerekmekte ve farklı duygu durumları için de sözlükler hazırlanabilmektedir.[7]

Duygu sözlükleri ilgili kelimelerin ilişkilendirildiği duygu durumunun derecesini sayısal olarak gösterecek değerler de içerebilir. Bu tür bir sözlükte ise "kötü" kelimesi ile "berbat" kelimesi negatif duygu sözlüğünde yer alacak olmasına rağmen ikinci sözcük birincisinden daha yüksek bir (negatif) sayısal skora sahip olacaktır.

Dilbilimciler tarafından hazırlanan ve literatürde en yaygın kullanılan sözlükler (Li, Xie, Chen, Wang ve Deng, 2014) tarafından aşağıdaki gibi listelenmiştir.

---

7 Özellikle, otel ve filmler gibi insanların çoğu tarafından sıkça kullanılan ve internette de bolca yorumlanan ürün veya hizmetler için daha kapsamlı duygu analizleri için farklı duygu türlerine dönük sözlükler hazırlanmaktadır.

Tablo 5: Duygu Sözlüğü Oluşturma Yaklaşımları

Çalışma	Sözlük Oluşturma Yaklaşımı
Hatzivassiloglou and McKeown	Yarı-otomatik
Wiebe	Yarı-otomatik
Kim and Hovy	Yarı-otomatik
Nasukawa and Yi	Yarı-otomatik
Godbole Srinivasaiah and Skiena	Yarı-otomatik
Cambria et al.	Yarı-otomatik
Harvard IV-4	Elle
Loughran and McDonald	Elle

Burada gösterilen sözlüklerden sadece son ikisi elle oluşturulmuştur. Diğerleri makine öğrenmesi yardımı ile yarı otomatik olarak oluşturulmuşlardır.[8] Bu da söz konusu sözlüklerin hazırlanmasının dilbilim ve bilgisayar bilimleri ile yakın ilgisini açıklamaktadır.

Geliştirilen bir duygu sözlüğünün performansı çeşitli şekillerde test edilmelidir. Bu performans testleri, insanlar tarafından etiketlenmiş büyük metinsel veri setlerinin geliştirilen sözlüğe dayalı olarak etiketlendirdikten sonra geliştirilen sözlüğün isabetliliğini değerlendirmek suretiyle gerçekleştirilmektedir. Uygulamalı karşılaştırma yapan çalışmalar da mevcuttur. Örneğin Li ve diğerleri (2014) bu farklı sözlüklere göre Hong-Kong borsasında faaliyet gösteren şirketlere yönelik duygu analizi gerçekleştirilmişler; aynı zamanda farklı sözlükler kullanmanın analize ne yönde etki ettiğini incelemişlerdir.

Duygu sözlükleri genellikle İngilizcedir. Türkçe duygu sözlüğü geliştirilmesi ile ilgili az sayıda deneme bulunmasına karşın geliştirilen sözlüklerin hiç biri açık kaynaklı olarak yayınlanmamaktadır. Bu nedenle bu sözlükler kullanılamamıştır. Karşılaşılan çalışmalardan biri aynı sorundan hareketle, İngilizce duygu sözlüklerden birinin çevrilmesi ve değerlendirilmesine dairdir. Bu çalışmada ortaya konan sözlük ise, açık kaynaklı değildir (Uçan, 2014). Bu nedenle bu sözlüğe ulaşılamamıştır. Bir dilde duygu sözlüğü bulunmadığı durumlarda, İngilizce sözlüklerinin makine çevirisi yardımı ile

8 Bu sözlükler hazırlanırken günlük dilde olumlu ve olumsuz anlamlı terimler seçilmektedir. Ancak belirli bir alana has (sinema sektörü, finans) özel duygu sözlüğü çalışmaları da bulunmaktadır.

hedef dile çevrilerek kullanması yoluna gidilmektedir (Tutar, Ünalır ve Toker, 2015).

Dolayısıyla bu çalışmada yukarıda bahsedilen sözlüklere benzer ve açık olan sunulan Hu ve Liu'ya ait sözlük kullanılmıştır (Hu ve Liu, 2004). Haber metinlerinin duygu durumlarını etiketleyebilmek için, bu sözlük Türkçeye Google Translate API makine çevirisi ile çevrilmiş daha sonra sözlükte geçen kelimeler teker teker incelenerek yanlış çevrildiği görülen ifadeler düzeltilmiştir. Ayrıca bazı ilgisiz sözcükler, yazım hatalarıyla yazılmış (yazım hatasıyla ifade edilen kelimeleri de yakalamak için) ve argo sözcükler, haber metinlerinde karşılaşılmayacağı gerekçesi ile sözlüklerden temizlenmiştir. Örneğin 'melun' ve 'pejmürde' sözcükleri sözlüğün çevirisinin orijinalinde mevcuttur ancak analize esas sözlükten çıkarılmıştır. Genel amaçlı sözlüklerin değerlendirilmeksizin kullanımı durumunda finansal haber metinlerinde farklı sonuçlara yol açabileceği bilinmektedir. Loughran ve McDonalds'ın yaptıkları çalışmada da genel amaçlı duygu sözlüğünün finansal metinlerin sınıflandırılması için kullanımı halinde analiz sonuçlarını olumsuz etkilediği ortaya konmuştur (2011). Bu yazarlar araştırmacıları duygu sözlüklerinin alan dışında kullanımı durumunda dikkatli olmaları yönünde uyarmaktadırlar. Çalışma kapsamında gerçekleştirilen analizlerde haber metinlerinin duygu skorlarının hesaplanması işlemi de elde edilen bu elenmiş duygu sözlüğü yardımı ile gerçekleştirilmiştir.

Haberlerin duygu durumlarının kategorik değerlerle etiketlenmesi sürecinde ise sadece olumlu ve olumsuz anlamdaki kelimelerin varlığına bakılarak karar verilmiştir. Ancak doğal olarak, aynı metinde hem olumlu hem olumsuz kelimeler olabilmektedir.

Bu noktada da bu terimlerin ilgili metin için nasıl bir ölçüye dönüştürüleceği önemli hale gelmektedir. Bir arama teriminin, bir doküman ile ilgili olup olmadığının incelenmesi bilgisayar bilimleri dâhilindeki IR (Information Retrieval-Bilgi Erişimi) alt alanı konusudur ve çeşitli yaklaşımlara göre belirlenen metrik değerlere göre gerçekleştirilmektedir. Genel olarak bir anahtar terimin metin içerisinde geçme sıklığı (frekansı) ile o metin ile ilişkisini tanımlayan DF: Document Frequency, TF.IDF (Term Frequency-Inverted Document Frequency), IG: (Information Gain) gibi metrikler bulunmaktadır (Khadjeh Nassirtoussi ve diğerleri, 2014). Bu metrikler özellikleri itibariyle (Taşcı ve Güngör, 2013) tarafından mukayese edilmiştir. Bu metriklerin kullanılması ile ilgili en çok yapılan varsayım, bir doküman içerisinde bir ifadenin sık tekrar etmesinin, o ifadenin metinle ilgili olması ihtimalini arttırdığına dair kabuldür.

Duygu analizinde ise bu problemin daha karmaşık bir hali ile karşılaşılmaktadır. Bu konuda da farklı yaklaşımlar bulunmaktadır. Bu yaklaşımlar metindeki ifadelerin doğal frekansına dayalı metriklerden ziyade

pozitif ve negatif kelimelerin bir arada olduğu bir metnin nasıl değerlendirileceğine dair farklı görüşlerden oluşmakta ise de literatürde yazarların bu amaçla kullanıldığı metrikleri mukayese eden ve değerlendiren bir çalışma ile karşılaşılmamıştır. Farklı çalışmalarda açık veya zımni olarak ifade edilen bu yaklaşımlar metinlerin analiz şekillerine göre değişkenlik göstermektedir. Bu nedenle kelime frekansları kullanılmıştır.

Burada esas alınacak doküman düzeyi de analiz sürecini ve sonuçları etkileyebilecek diğer bir faktördür. Metinsel verilerin sayısal değerlere dönüştürülmesi sürecinde pozitif ve negatif duygu etiketleme işlemi genellikle doküman düzeyinde (her bir doküman için ayrı ayrı) gerçekleştirilmektedir. Ancak analiz birimi olarak doküman dışında daha da ayrıntıya gidilerek başka birimlerin seçilmesi de mümkündür. Analiz birimi olarak seçilecek metin parçasının boyutuna göre analiz türleri (Feldman, 2013) tarafından doküman, cümle düzeyinde, özellik bazlı ve mukayeseli analiz olarak sıralanmaktadır. Bunlardan, doküman düzeyinde analiz en yalın ve yaygın kullanılan yaklaşım olduğundan, çalışmamızın esasını oluşturmuştur. Öte yandan çalışma kapsamında gerçekleştirilen analizde kullanılan haber metinlerinin tamamı, haber boyunca aynı konudan bahsetmediğinden ve haberin söz konusu şirketle ilgili ifadelerin geçtiği parçası alındığından cümle düzeyinde analize yakın bir analizin gerçekleştirildiği söylenebilir.

Çalışmanın sonraki bölümünde, elde edilen Türkçe duygu sözlüğü yardımıyla gerçekleştirilen uygulama sürecine ve sonuçlarına yer verilmektedir.

2. Uygulama ve Sonuçlar

2.1. Uygulama Aşamaları ve Analiz Araçları

Uygulama aşağıdaki akış çerçevesinde gerçekleştirilmiştir. Bu bölümde söz konusu akış adımları, verinin elde edilmesi de dâhil olmakla birlikte, Şekil 2'de sunulmuştur.

Analiz aşamalarını açıklamaya geçmeden bu çalışma kapsamında kullanılan ve geliştirilen analiz araçlarına da kısaca değinmekte fayda bulunmaktadır:

Haber metinleri ile ilgili analizlerde, haberlerin adreslerini bulmak ve listelemek için araştırmacılar "web gezgini (web crawler)[9]" adı verilen

9 Web Crawler ifadesindeki 'crawler' sözcüğünün sözlük anlamı gezgin değildir. Palet, örümcek gibi anlamlara gelmektedir. Web tarayıcısı şeklinde çeviri mevcut ise

programlar geliştirmek veya mevcut bir web gezginini analizlerine uyarlayarak kullanmak zorundadırlar. Web gezgini, herhangi bir konuda gereken web sayfalarını bir şekilde bulan yazılım algoritmasıdır (Edwards, McCurley ve Tomlin, 2001). Arama motorları da arama sonuçlarında gösterdikleri web sayfalarını özel olarak geliştirdikleri bu uygulama yardımı ile bulmaktadırlar.

Web gezgini geliştirirken kullanılan en genel yaklaşım sayfaların içerisindeki linklere giderek bu sayfaları kaydetmek, sonra da gidilen sayfalarda aynı işlemi tekrarlayarak ilgili sayfaları bulmaktır (Mitra ve Mitra, 2011). Nitekim webdeki haber kaynakları tek yerde bulunmamaktadır. Bazı servisler ise (Google News, Bing Haberler vb.) bu amaçla çeşitli konulardaki haberleri derleyerek sunmaktadırlar. Bu çalışmada da ilgili haber toplama servislerinden Google News kullanılmıştır. Bu site artık sadece günlük kullanıma uygun biçimde haberleri sunmakta, ancak toplu liste şeklinde analize uygun şekilde veri sunmamaktadır. Bu nedenle çalışma kapsamında Google News içerisindeki haberlerin çekilebilmesi için yine Python dili ile özel olarak geliştirilen bir web gezgini kullanılmıştır.

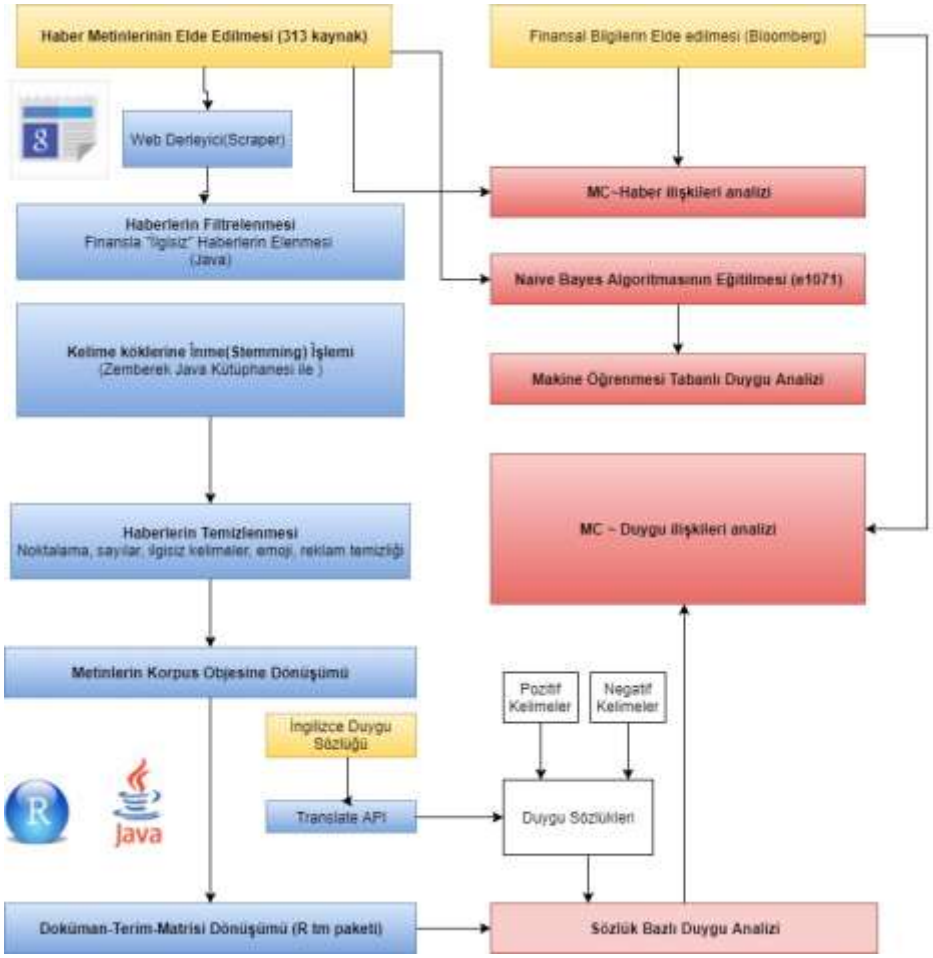
Ayrıca çalışma kapsamında web gezgini ile birlikte bütünleşmiş şekilde çalıştırılan bir "web derleyici (web scraper)[10]" de bu çalışmanın yazarları tarafından geliştirilmiştir. Web derleyici, web gezgini tarafından bulunan veya daha önce bulunarak listelenmiş web adreslerinde yer alan içeriği bilgisayara indirmeye yarayan yazılımlardır (Mitra ve Mitra, 2011). Çalışma kapsamında geliştirilen web derleyici de, web gezgininin bulduğu haber sayfalarını eş zamanlı olarak indirmekte ve metnini ayrıştırarak kaydetmektedir. Herhangi bir haberin herhangi bir web sayfası tarafından yayınlanması durumunda bu haberin hangi sayfada yer aldığı web gezgini uygulaması ile Google News veri tabanından alınmakta, yeri belirlenmiş olan haberin içeriği ise web derleyici tarafından indirilmektedir. Bu işlem yayınlanmış tüm haberler için bu iki uygulama tarafından gerçekleştirilmektedir. Ayrıca web ortamında yayınlanan metinsel verileri okumakta, daha sonra sadece haberin gövdesini almakta ve ilgili haber sitesindeki reklam ve menü adları, sayfa içi reklamlar gibi ilgisiz içerikleri

---

de bu 'browser' kavramı ile karıştırılmaktadır. Bu nedenle web crawler'in web sayfalarını arayarak bulan işlevinden ötürü Türkçesi için 'web gezgini' kavramı tercih edilmiştir.

10 Scrape fiili İngilizce'de 'kazıma' anlamına gelmektedir. Bu çeviri de Türkçe'de somut bir kavrama tekabül etmediğinden direct olarak çevrilmemiştir. Web scraper programının işlevinin adresleri bilinen sayfaların içeriğini derleme olduğundan 'web derleyici' çevirisi tercih edilmiştir. Buradaki derleme işlevi programlama dillerinin makine diline çevrilmesi anlamına gelen ve Türkçe'ye derleme olarak çevrilen 'compiling' kavramı ile karıştırılmamalıdır.

ayırabilmektedir. İlgili şirket kodlarını aldıktan sonra bu şirketlerle ilgili tüm haberleri işleyebilen bu uygulama da Python ve Java dilleri ile geliştirilmiştir.



Şekil 2. Uygulama Aşamaları ve Araçları

Kelime köklerine inilmesi işleminde, Java dilinde yazılmış olan Zemberek kütüphanesinden faydalanılmıştır. Nitekim yukarıda açıklandığı gibi kelime köklerine inilmiş biçimde doküman terim matrisine ulaşmak için korpustaki bazen milyonlarca kelime teker teker köklerine indirgenmelidir. Bu amaçla İngilizce için bir çok algoritma bulunmakta iken Türkçe için sadece Zemberek Java Kütüphanesi bulunmaktadır. Zemberek adlı araçtan bu amaçla

yararlanabilmek için yine Java diliyle yazılmış bir yardımcı kod kullanılmış; bu kütüphane ve yazılım ile algoritmik olarak tüm kelimelerin köküne inilmiştir. Bu amaçla kullanılan başka kütüphaneler olmasına rağmen, açık kaynaklı Türkçe kelime işlemci ofis programları (OpenOffice, LibreOffice) arkasındaki desteğinden ötürü Zemberek tercih edilmiştir.

Metin madenciliği ve temel istatistiksel analizlerle bazı görselleştirme işlemlerinde, istatistiksel analiz, veri madenciliği, matematiksel modelleme ve ağırlıklı olarak bilimsel amaçlara dönük işlemleri gerçekleştirmek için ortaya çıkarılmış olan *R*© adlı programlama dili ve bu dilin daha verimli kullanılabilmesi için geliştirilmiş *RStudio* adlı bütünleşmiş geliştirme ortamı kullanılmıştır.

2.2. Veri

Uygulamada kullanılan veriler ve nasıl elde edildikleri aşağıda açıklanmıştır.

Haber Metinleri

Ele aldığımız analiz döneminde (2014 yılı) Borsa İstanbul 30 (BIST30) endeksi içerisinde yer alan 30 firmaya ait haberleri elde edebilmek için; her bir şirketin, şirket ünvanı yer almaksızın resmi adı (Garanti Bankası, Koç Holding gibi) geliştirilen web derleyicisi yardımı ile otomatik olarak Google News haber servisinde arattırılmıştır. Bunun için şirket adları, web scraper kodları içerisinde tanımlanmış ve bu programın her bir terimi Google News'te arayarak ilgili haberleri hafızasına alması sağlanmıştır. Bu sayede toplam 14.108 adet haber, yayınlanma tarihi, başlığı, yayınlandığı haber sitesi, URL adresi ve bu haberden Google News'in kırptığı arama terimin geçtiği metin bloğu (intro) ile birlikte (XML formatında) kaydedilmiştir.

Bu noktada "ilgili" haberleri bulmak için arama işlemi sırasında hangi arama teriminin ve metodolojisinin kullanılacağına karar verilmesi önemlidir. Zira özünde "ilgili olma" kavramı bir ölçüde öznellik barındırmaktadır.

Nitekim arama motoruna yazılan farklı arama terimleri ya da ifadeler, farklı haberlerin ve haber sayılarının ortaya çıkmasına neden olmaktadır. Örneğin "Yapı Kredi" terimi, analiz tarihinde Google News servisinde aratıldığında 86.900 adet haber listelenirken, firmanın resmi adı olan "Yapı ve Kredi Bankası" şeklinde arama yapıldığında 4.550 adet haber listelenmektedir.

Aynı şekilde yazılan arama terimine göre firma ile doğrudan "ilgili olmayan" haberlerin listelenme olasılığı da göz önünde tutulmalıdır. Söz gelimi, bir kullanıcı "Pegasus" ifadesini uçak bileti almak için arama terimi olarak

yazabilir, diğer bir kullanıcı ise Yunan Mitolojisi araştırması yapıyor olabilir. Dolayısıyla "Pegasus Hava Yolları" ifadesi, halk arasındaki adı ile "Pegasus" şeklinde aratıldığında Türkiye dışından birçok şirket haberleri yanında finans konusu dışındaki haberler de listelenmektedir.

Arama motorlarında istenen dile göre arama yapılabilmekle birlikte çeşitli nedenlerden ötürü, farklı dillerde yayınlanan haberler yine de bir dile özel olarak listelemesi beklenen haberler arasına dâhil olabilmektedir.

Bu nedenlerle "ilgili" haber metinlerini bulmak amacıyla şirketin resmi adı dışındaki arama terimlerinin neler olabileceğine dair çalışma kapsamında değerlendirilen yaklaşımlar şunlardır:

Şirketin alternatif adları: Şirketlerle ilgili daha fazla habere ulaşabilmek adına kullanılabilecek bu yaklaşım, bazı şirketlerin halk arasında veya medyada resmi adı dışında kullanılan adlarını da arama terimi olarak kullanmaktadır. Bu adlar genellikle bilinen resmi adların kısaltması veya farklılaştırılması şeklinde ortaya çıkmaktadır. Örneğin "Yapı ve Kredi Bankası" adlı firma "Yapı-Kredi" şeklinde kısaltılmaktadır. Yine "Türk Telekomünikasyon A.Ş.", "Türk Telekom" şeklinde de ifade edilerek aranabilmektedir.

Şirketle ilgili ürün adları: Şirketin ürettiği, geliştirdiği veya pazarladığı bir ürüne adına dair kelimelerin de arama terimi olarak seçilmesidir. Söz gelimi "Apple Inc" şirket adı yerine "Iphone" ya da "Macintosh" ifadesinin aranması buna örnektir. Bazı durumlarda şirket adı ile ürün adı aynıdır. "Tofaş Otomotiv A.Ş." ifadesi kısaltılarak "Tofaş" denmektedir ki bu durumda aynı firmanın ürettiği bir otomobil modeli de kastedilmiş olmaktadır. Bu durum da ilgili şirketle ilgili haberlerin derlenmesinde sorun çıkarabilmektedir. Çalışma kapsamında, örneğin Tofaş model otomobillerin karıştığı kaza haberlerinin de Google News sonuçlarında çıktığı görülmüştür. Bu yaklaşım ancak ürün adlarının benzersiz sözcükler olduğu durumlarda kullanıldığı takdirde faydalı sonuçlar sağlayacaktır.

Şirketin borsa kodu: Sadece ilgili haberleri bulmak amacıyla şirketlerin borsada kotasyon kodlarını kullanma (DOHOL, KCHOL gibi) yolu da denenmiştir. Bu yol yukarıda anılan problemlerle karşılaşılmaması adına güvenli bir yol olarak ele alınmıştır. Ancak bazı şirket kodlarının doğal dilde tekabül ettiği başka yaygın anlamlardan ötürü yine ilgisiz haberlerin listelenmesine neden olduğu görülmüştür. Örneğin Şişecam A.Ş kodu olan "SISE" ifadesi bunlardan biridir. Bu arama terimi yazıldığında ilgili şirket dışında "şişe" kelimesi ile ilgili haberler de listelenebilmektedir. Aynı şekilde aynı şirket kodunun farklı dillerdeki anlamları da sorun olmuştur. Örneğin "SAHOL" ifadesi Malay dilinde, "BIMAS" ifadesi Endonezya dilinde farklı anlama geldiğinden Türkçe haberler arasında bu dillerden haberler çıkmaktadır.

Sonuç olarak, hem genellikle finansal haberler içerisinde yaygın kullanımından ötürü, hem de yukarıda anılan yaklaşımlardan daha ilgili sonuçlar verdiği görüldüğünden arama terimi olarak şirketlerin *resmi adlarının kullanımı tercih edilmiştir.*

Kuşkusuz, bu "ilgi" ilişkisinin tek referans noktası, milyonlarca farklı insan tarafından kaleme alınan yorum yapılan metinsel veriler olduğundan, ilgili dokümanı tespit etmek için kurgulanmış her algoritmanın yanılması olasıdır. Profesyonel arama motorları bu durumu aşmak için çeşitli yaklaşımlar geliştirseler de ilgisiz sayfalar varlığını sürdürmeye devam edebilmektedir.

Finansal Veriler

Analiz kapsamında BIST30'da faaliyet gösteren her bir şirkete ait 2013 sonu ve 2014 içindeki her çeyrek sonuna ait şirket değeri ve haftalık kapanış fiyatları veri olarak derlenmiştir. Veriler Thomson Reuters ve Bloomberg veri terminallerinden indirilmiştir.

Duygu Sözlüğü (İngilizce)

Elde edilen üçüncü veri grubu Türkçe'ye çevrilecek olan bir İngilizce duygu sözlüğüdür. Daha önce belirtildiği gibi bu sözlük açık olan sunulan Hu ve Liu'ya ait sözlüktür (Hu ve Liu, 2004). Bu sözlük Türkçeye Google Translate API makine çevirisi ile çevrilmiş daha sonra sözlükte geçen kelimeler teker teker incelenerek yanlış çevrilen ifadeler düzeltilmiştir.

2.3. Metin Madenciliği ve Duygu Analizi

Şirketlerle "İlgili" Haber Sayılarının Ayrılması (Filtering)

Elde edilen metinsel verilerin analizlere hazır hale getirilebilmesi amacıyla, literatürde ve endüstride genel kabul görmüş bir yaklaşımla, analizlerden önce bir ön-işleme sürecinden geçirilmesi gerekmektedir. Bu süreç analizlerin doğruluğunu artıracaktır.

Bu çalışma kapsamında yapılan ön işleme analizlerinde; spor kulüpleri ile sponsorluk anlaşmaları yapan şirketlerin ilgili bu kulüplerin adları içerisinde kullanılması nedeniyle ve Google News'in kategorik arama (finans, ekonomi, sağlık vs.) gibi bir filtreleme özelliği bulunmadığından spor haberlerinin de arama sonuçlarına karıştığı tespit edilmiştir. Analizde yanılmalara neden olabilecek bu durum özellikle dikkate alınarak haberlerin elenebilmesi amacıyla metin düzeyinde bir filtreleme (ilgisiz haberleri eleme) algoritması yazılmıştır. Bu algoritma iki basit ilkeye göre web derleyicinin indirdiği haberleri "ilgili"

veya "ilgisiz" diye ayırmakta, ilgisiz haberi ise silmektedir. Bu süreç aşağıdaki gibi ifade edilebilir:

EĞER: Haber metninin herhangi bir yerinde spor ifadesi varsa VEYA Haber başlığında geçen bir ifade, haberin intro kısmında bulunmuyorsa: HABER İLGİSİZ [11]
AKSİ TAKDİRDE: HABER İLGİLİ

Bu temizlik işleminden sonra web scaper'in ulaştığı 14.108 adet haberden, 5.674 adedi ilgisiz olarak değerlendirilerek elenmiş ve 8.434 adet haber kalmıştır.

Aşağıdaki iki haber ilgili ve ilgisiz haberlere örnek olarak verilebilir:

İlgisiz	İlgili
Kardemir *Karabükspor Teknik Direktörü Tolunay Kafkas, bazı futbolcular ve teknik heyet, Aylin Özel Eğitim ve Rehabilitasyon merkezinde eğitim gören çocukları ziyaret etti.*	*İş Yatırım,* ***Kardemir*** *hisse raporunda düşen hammadde fiyatlarının şirket büyümesine olumlu katkı yaptığı belirtilirken, yıllık büyüme beklentisinin de yüzde 23 olduğu açıklandı.*

Ayrıca elde edilen haber metinleri, doküman-terim matrisinde herhangi analitik bir değeri olmayacağı kesin olarak bilinen ilgisiz bağlaçlardan ("ve", "veya" gibi; bu ifadeler stopwords olarak adlandırılmaktadır) arındırılmıştır.

Korpus, toplam 8435 adet haber başlığından ve bu başlıklarla ilgili haber bloklarına ait 220.497 adet sözcükten oluşmaktadır.

pn: Filtrelenmemiş şirket haberleri (alınan tüm haberler) ve *rn*: Filtrelenmiş şirket haberleri (spor haberleri ve Google News'te listelenen bazı ilgisiz haberler hariç) olmak üzere; her şirket için yayınlanan haber sayılarının dökümü aşağıdaki tabloda sunulmaktadır.

11 "Haber başlığında geçen bir ifade, haberin intro kısmında bulunmuyorsa" filtresi ilgisiz haberlerin genelinde (spor haberleri de dâhil) bu durumun mevcut olmasından ötürü algoritmaya eklenmiştir.

Tablo 6: Yayınlanan Haber Sayıları

Kod	Yayınlanan Haber (*pn*)	İlgili Haber (*rn*)	Kod	Yayınlanan Haber (*pn*)	İlgili Haber (*rn*)
AKBNK	713	565	**OTKAR**	155	116
ARCLK	628	398	**PETKM**	483	222
BIMAS	750	303	**PGSUS**	100	75
DOAS	520	139	**SAHOL**	500	317
DOHOL	510	273	**SISE**	220	164
EKGYO	446	303	**TAVHL**	193	145
ENKAI	116	32	**TCELL**	576	546
EREGL	67	40	**THYAO**	414	389
FROTO	666	303	**TKFEN**	105	32
GARAN	735	546	**TOASO**	690	336
HALKB	609	464	**TTKOM**	633	492
ISCTR	740	509	**TUPRS**	705	277
KCHOL	514	429	**ULKER**	400	367
KOZAL	402	133	**VAKBN**	628	415
KRDMD	533	40	**YKBNK**	267	64
			TOPLAM	**1.4018**	**8.434**

Tablo 6'da görülen 2. Sütün yayınlanan tüm haberlerin adedini göstermektedir. Bu aşamada henüz spor haberleri ve ilgisiz diğer bazı haberler elenmemiş durumdadır. 3. Sütunda görülen sayılar ise spor haberleri ile ilgisiz haberlerin elendikten sonra kalan haber sayılarını göstermektedir. Analiz döneminde şirketler hakkında çıkan haber sayıları 32 ila 565 adet haber arasında seyretmektedir. 8 şirkete ait ilgili haber sayısı olan 300-400 en yüksek frekans aralığıdır.

Kelime Köklerine İnme ve Doküman Terim Matrisi

Önişleme süreci ile birlikte Zemberek Kütüphanesi kullanılarak haberlerde geçen tüm sözcükler için kelime köklerine inme işlemi de

gerçekleştirilmiştir. Ayrıca haber metinleri üzerinde veri temizliği işlemi de gerçekleştirilmiştir. Bunun yanında tüm sözcüklerin tamamı küçük veya tamamı büyük harfe dönüştürülmektedir. Böylece sadece kelime köklerinden oluşan metin yapısı ortaya çıkarılmaktadır. Metin bu hale getirilmeden analiz yapıldığı takdirde analizin işlem süresi gereksiz ifadelerden ötürü artmaktadır. Ayrıca kelimelerin tamamının küçük harf veya tamamının büyük harf olmadığı durumlarda duygu sözlüğü ile karşılaştırma yapılırken eşleşme problemleri nedeniyle ilgili sözcük atlanabilmektedir.

Bu şekilde elde edilen ve halen sayısal hale dönüştürülmemiş metinsel veri, R dili yardımı ile doküman-terim matrisine dönüştürülmüştür.

Buna göre BIST30'daki 30 şirkete ait doküman-terim matrisinin boyutu 8.435 x 220.497=1.859.892.195 hücreli olmaktadır. Bu matris her şirket için ayrı ayrı elde edilmiş ve filtrelenerek gerekli analizlere hazır hale getirilmiştir.

Sözlük Bazlı Analizle Duygu Skorlarının Hesaplanması

Duygu tonlarının tespiti için sözlük bazlı yaklaşım kullanılarak duygu skorları hesaplanmıştır.

Sözlük temelli yaklaşımda *İngilizce sözlüğün Türkçeleştirilmesi* sürecinde aşağıdaki iki duruma sahip kelimeler elenmiştir:

Bu durumlardan ilki, çevrilen kelimenin ilgili dilde farklı anlamlar içerebilmesi dolayısıyla aslında pozitif iken negatif veya tersi durumların oluşmasıdır. Daha sık görülen bir durum da İngilizcedeki pozitif bir kavramın çevirisinin Türkçede de aynı anlama geldiği halde sık kullanılmama durumudur. Örneğin "awesome" kelimesi "harika" anlamında çevrildiğinde yine pozitif anlamlıdır ancak Türkçede özellikle finansal haber metinleri içerisinde bunun yerine "mükemmel" kelimesi kullanılıyor olabilir. Bu durumda, ilgili dilden çeviri yapılırken benzer anlamlardaki kelimeler de insan müdahalesi ile eklenmelidir. Bu amaçla çalışma kapsamında çevrilen pozitif ve negatif sözcükler teker teker incelenerek ilgisiz olanlar veya yanılmalara neden olabilecek kelimeler düzenlenmiştir. Aynı şekilde aynı kavramı karşılayacak başka Türkçe kelimeler de eklenmiştir.

İkinci durum ise, Türkçe'nin dilbilimsel yapısından kaynaklanan özel durumlardır. Türkçenin sondan eklemeli bir dil olması nedeniyle birçok kavram ekler vasıtası ile karşıt anlamlı kavramlara dönüşmektedir. Örneğin "ümit" kelimesi pozitif bir kelime iken, eklerle türetilen "ümitsiz" kelimesi negatif anlamlıdır. Sorun, kelime köklerine inme işlemi sırasında "ümitsiz" ve "ümitli" kelimesinin kökünün "ümit" olması nedeniyle anlamsal değişim yaşanmasıdır ki bu analiz sağlığını ciddi bir şekilde etkileyecektir. Bu amaçla sözlük içerisinde

ekler yolu ile anlam değiştiren kelimeler elenerek eksiz biçimde de anlamını koruyan "müthiş", "iyi", "zengin" gibi kavramlara yer verilmiştir. Ancak yine de farklı bazı kullanımların (ironik dil kullanımı vb. durumların) algoritmayı yanıltabileceği kabul edilmektedir. Bu tür yapıların tamamı doğal dil işleme alanını ilgilendirdiğinden bu çalışmanın kapsamı dışındadır.

Sözlüklerin üretilmesinden ve haber metinlerinin analize hazır hale getirilmesinden sonra her şirket için ayrı oluşturulan haber korpusu içerisinde 2014 yılı boyunca çıkan haberlerin pozitif ve negatif duygu skorları sözlükler üzerinden hesaplanmıştır.

Buradaki skor ifadesi ilgili şirket ile ilgili çıkan haberlerden duygu sözlüğü ile eşleşmeyi ifade etmektedir. Söz gelimi pozitiflik skorunun 83 olması bu şirketle ilgili korpustaki haberlerin içerisinde 83 adet kelimenin olumlu anlamdaki kelimeler sözlüğü ile eşleştiğini ifade eder. Yine, negatiflik skorunun 19 olması ilgili şirket hakkında 19 negatif kelimeyi ifade eder. Daha sonra pozitif skorlar negatif skorlardan çıkarılarak "dengelenmiş duygu skoru" elde edilebilir. Dengelenmiş duygu skoru değeri ilgili şirket hakkındaki olumlu veya olumsuz genel kanaati yansıtmaktadır. Simgesel gösterimle, P: Pozitif duygu skoru; N: Negatif duygu skoru; Δ: Dengelenmiş duygu skoru $\Delta = P - N$ olmak üzere; $\Delta > 0$ durumunda olumlu, $\Delta < 0$ durumunda olumsuz kanaatin varlığından söz edilebilir. Örnekteki durumda ilgili şirketin dengelenmiş skoru; 83 - 19 = 64 olacaktır.

Tahmin edileceği üzere Δ skoru, yayınlanan haber sayısıyla dolaylı olarak bağlantılı bir değer olmakla birlikte, özünde duygu sözlüğündeki eşleşme adedi ile de ilgilidir. Örneğin, şirket hakkında çok fazla haber çıktığı halde bu haberlerin nötr olması durumunda duygu skorları yüksek olamayacaktır. Yine başka bir şirket için az haber çıktığı halde bu haberler pozitif duygu tonunda olabilir. Duygu skorlarının bu özelliği nedeniyle şirketlerin piyasa değeri ile ilişkisi ayrıca ele alınmalıdır.

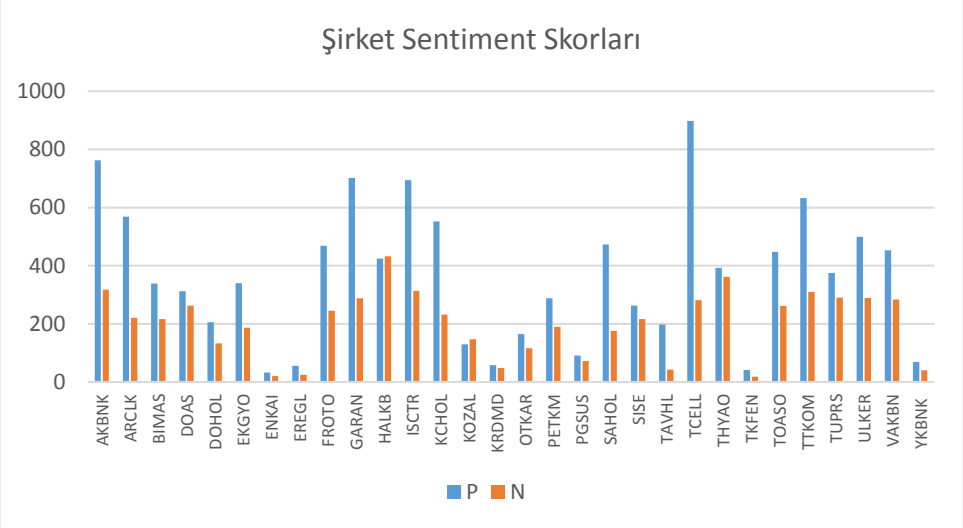
P,N ve Δ skorunun her bir şirket korpusu için hesaplanan değerleri Tablo 7'de gösterilmektedir.

Tablo 7: Şirketlerin Duygu Skorları

Şirket	*P*	*N*	Δ	Şirket	*P*	*N*	Δ
AKBNK	763	318	445	**OTKAR**	165	116	49
ARCLK	569	221	348	**PETKM**	288	190	98
BIMAS	338	216	122	**PGSUS**	91	72	19

DOAS	312	263	49	**SAHOL**	473	175	298
DOHOL	205	132	73	**SISE**	262	216	46
EKGYO	340	186	154	**TAVHL**	197	42	155
ENKAI	32	20	12	**TCELL**	898	281	617
EREGL	56	25	31	**THYAO**	392	362	30
FROTO	469	245	224	**TKFEN**	41	18	23
GARAN	702	288	414	**TOASO**	448	261	187
HALKB	424	432	-8	**TTKOM**	633	310	323
ISCTR	694	313	381	**TUPRS**	375	290	85
KCHOL	552	232	320	**ULKER**	499	289	210
KOZAL	129	147	-18	**VAKBN**	453	283	170
KRDMD	58	48	10	**YKBNK**	69	40	29

Görüldüğü üzere 2014 yılında HALKB ve KOZAL kodlu şirketler haricindeki tüm şirketler hakkında yayınlanan haberler olumludur. Tüm şirketlerin *P* ve *N* değerleri ayrıca Şekil 3'te gösterilmektedir.



Şekil 3: Şirketlere ait Pozitif (P) ve Negatif (N) Duygu Skorları

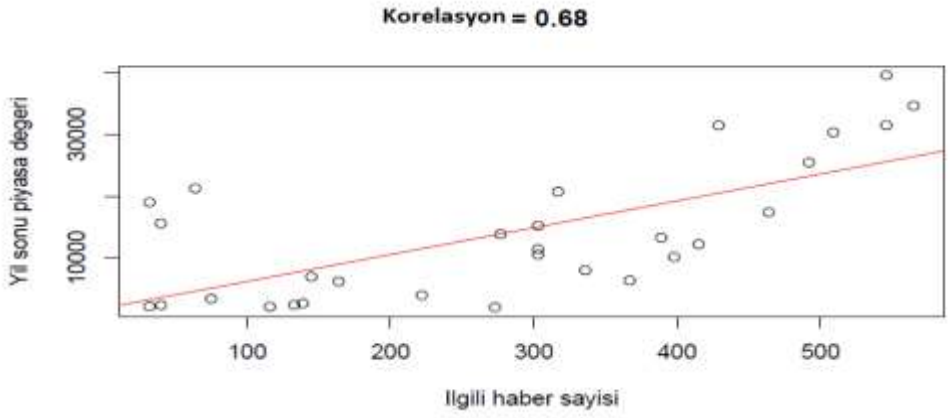
2.4. Haberler ve Şirket Değerleri Arasındaki İlişki Analizi

Uygulamanın bu aşamasında haber sayıları ile finansal değerler arasındaki ilişkilerin ortaya çıkarılması hedeflenmiştir.

Söz konusu ilişki iki yönlü olabilir: Piyasa değeri yüksek şirketlerin daha fazla kurumsal tanıtım ve imaj çalışmasından vb. ötürü daha fazla haberin yayınlanmasını sağlamaları mümkündür. Diğer yandan, bu haberlerin şirketin mevcut piyasa değerine ve performansına olumlu bir katkı yaptığı da düşünülebilir. Bu çalışma kapsamında etkinin yönünden ziyade, karşılıklı ilişkinin varlığı ve düzeyi (işareti) yayınlanan haberlerle şirketlerin piyasa değerleri arasındaki korelasyon değerleri yardımıyla incelenmiştir.

İncelenen yıl bazında haber sayıları ile şirketlerin yılsonu piyasa değerlerinin arasındaki karşılıklı ilişkinin ölçülmeye çalışıldığı analizlere ilişkin sonuçlarsa şöyledir:

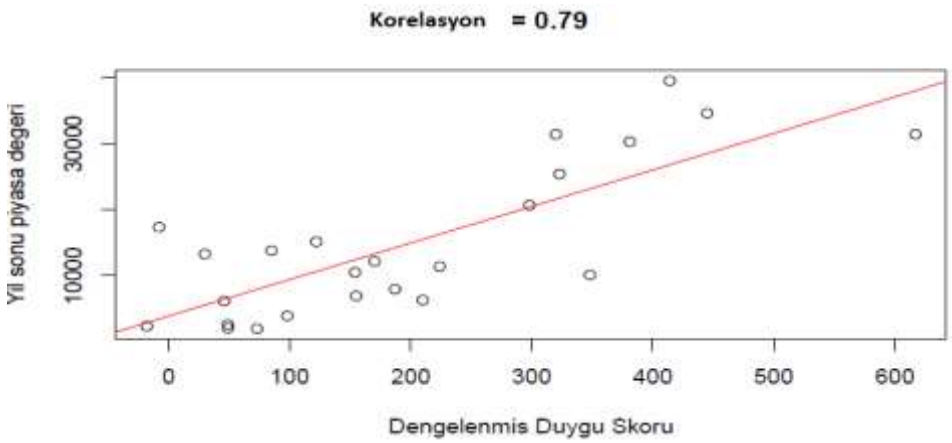
Haberler filtrelenmeksizin yayınlanan haber sayıları (*pn*) ile ilgili şirketlerin piyasa değerleri arasındaki korelasyon 0,44 iken; sponsorluk haberlerine dair spor haberleri ve benzeri ilgisiz haberler elendikten sonra, yani "filtrelenmiş" haber sayıları (*rn*) ile şirketlerin piyasa değerleri arasındaki ilişki katsayısı 0,68'e çıkmaktadır. Bu ikinci ilişki Şekil 4'te gösterilmektedir.



Şekil 4: Filtrelenmiş Haber Sayıları ile Şirketlerin Piyasa Değeri Arasındaki İlişki

Bu sonuçtan haber sayıları ile piyasa değeri arasında orta düzeyde güçlü pozitif ilişkilerin var olduğu ve bu ilişkiler haberlerin içeriğinin finans ile "ilgili"

olması ile birlikte daha iyi bir biçimde gözlemlenebildiği anlaşılmaktadır. Bu analizlerde haberlerin tonunun pozitif veya negatif olup olmamasına bakılmaksızın tüm haberler dikkate alınmıştır. Haberlerin duygu tonları ile şirket piyasa değeri arasındaki ilişkiler incelendiğinde ise bu iki parametre arasındaki korelasyon değeri Şekil 5'te görüleceği üzere 0,79'a yükselmektedir.



Şekil 5: Δ Dengelenmiş Duygu Skoru ile Şirketlerin Piyasa Değerleri Arasındaki İlişki

Sonuç olarak, şirket hakkında çıkan haberlerin duygusal tonu ile şirketin yılsonu piyasa değeri arasında, yalnızca haber sayıları ile karşılaştırıldığındaki durumdan daha yüksek düzeyde pozitif bir ilişki olduğu (0,79 > 0,68) görülmektedir.

Genel Değerlendirme

Bu çalışmada, Borsa İstanbul'da işlem gören BIST30 şirketlerine ilişkin 2014 yılında yayınlanmış haberler metin madenciliği ve duygu analizi teknikleri ile incelenmiş; bu analizler sonucunda haber sayılarına ve haber içeriklerinin duygu tonlarına (olumlu / olumsuz) ilişkin olarak elde sayısal değerler ile aynı dönemde piyasada oluşan şirket değerleri arasındaki ilişkiler analiz edilmiştir.

Bu amaçla öncelikle, analiz döneminde bu şirketlere ilişkin çeşitli kaynaklarda yayınlanan toplam 14.018 haber arasından finansal olmayanlar elenerek toplam 8434 haber başlığı, kaynağı, tarihi ve ilgili şirketin geçtiği metin ile bir veri kümesi (korpus) teşkil edilmiştir. Bu haberler literatürde önerilen ön

işlem prosesine uygun olarak işlenmiştir. Bu kapsamda öncelikle kelime köklerine inilmiş ve veri kümesi sadeleştirilerek analize uygun hale getirilmiştir. Duygu analizi için Türkçede henüz yayınlanmış duygu sözlüğü bulunmadığından İngilizce yayınlanmış duygu sözlüklerinden biri olan Minqing Hu ve Bing Liu'a ait duygu sözlüğü (Hu ve Liu, 2004) Google Translate API'si ile Türkçeye çevrilmiştir. Daha sonra sözlük üzerinde de Türkçenin dilsel özellikleri ile ilgili düzeltmeler yapılarak elde edilen doküman-terim matrisi üzerinden olumlu / olumsuz haber içeriklerine ilişkin pozitiflik / negatiflik skorları hesaplanmıştır. Nihai olarak, ele alınan şirketlere ilişkin sayısallaştırılmış metinsel verilerle, şirketlerin 2014 yılı finansal verileri arasındaki ilişkiler analiz edilerek sonuçlar yorumlanmıştır.

Ortaya çıkan temel sonuç, finansal piyasalarla yayınlanan haberler ve bunların duygu tonları ile finansal değerler arasında karşılıklı ilişkilerin var olduğudur. Örneğin BIST30 şirketleri hakkında çıkan finansla "ilgili" haberler ile şirketin piyasa değeri arasında 0,68 düzeyinde bir korelasyon değeri bulunurken; şirketler hakkında çıkan haberlerin dengelenmiş pozitif duygu skoru ile piyasa değerleri arasında daha yüksek bir ilişki bulunduğu (korelasyon: 0,79) tespit edilmiştir.

Bu bulgulardan hareketle, şirket değerlerinin klasik risk – getiri gibi faktörlerin yanı sıra karar vericilerin algı ve önyargılarından etkilenebildiğine ilişkin öncül çıkarımlara ulaşılabilir. Yani şirket fiyatları, büyüklüğü veya algılanan değer ilişkisi de çıkan haberleri ve bunların algılanma biçimini etkileyebilir. Bu çalışmada emareleri bulunan ilişkilerin karar verme literatüründeki hangi höristiklerle ve ne şekilde ilişkilendirebileceği bu çalışmanın kapsamı dışında tutulmuştur.

Çalışma temel olarak Türk finansal piyasalarının değerlendirilmesinde önemli bir araç olarak Türkçe haber kaynaklarının da kullanılabileceğini göstermeyi ve bunun için yöntemsel bir katkı sunmayı hedeflemiştir. Türkiye'deki finansal piyasalara ilişkin bu çalışmaya benzer çalışmaların çoğalmasının, yatırımcı davranışları ile ilgili önemli bilgilere ulaşılmasına yardımcı olacağı düşünülmektedir.

Kaynakça

Agarwal, Basant, Namita Mittal (2015), *Prominent Feature Extraction for Sentiment Analysis*. (Springer).

Bajo, Emanuele ve Carlo Raimondo (2017) "Media sentiment and IPO underpricing." *Journal of Corporate Finance*.

Barber, Brad M. ve Terrance Odean (2007), "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors." *The Review of Financial Studies* 21.2 785-818.

Brun, Wibecke ve Karl Halvor Teigen (1988), "Verbal probabilities: ambiguous, context-dependent, or both?" *Organizational Behavior and Human Decision Processes*, *41*(3), 390-404.

Colm Kearney ve Sha Liu. (2013), "Textual Sentiment in Finance: A Survey of Methods and Models". *International Review of Financial Analysis*, 33, 171-185.

Da, Zhi, Joseph Engelberg ve Pengjie Gao (2011), In search of attention. *The Journal of Finance*, *66*(5), 1461-1499.

Edwards, Jeny Kevin McCurley ve John Tomlin. (2001). "An Adaptive Model for Optimizing Performance of an Incremental Web Crawler", Proceedings of the 10th International Conference on World Wide Web (New York, United States of America, ACM)

Engelberg, Joseph (2008), *Costly Information Processing: Evidence from Earnings Announcements* (SSRN Scholarly Paper No: ID 1107998). Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=1107998. (03.03.2016)

Feldman, Ronen (2013). "Techniques and applications for sentiment analysis". *Communications of the ACM*, *56*(4), 82–89.

Ferguson, Nicky J, Dennis Philip, Herbert Y. T. Lam, Jie Guo (2014), "Media Content and Stock Returns: The Predictive Power of Press" (SSRN Scholarly Paper No: ID 2111352). Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2111352 (03.03.2016)

García, Diego (2013), ""Sentiment during Recession"s. *The Journal of Finance*, *68*(3), 1267–1300.

Gidofalvi, Gyozo ve Charles Elkan (2001), "Using news articles to predict stock price movement"s. *Department of Computer Science and Engineering*.

Gigerenzer, Gerd, Peter M. Todd (1999), Fast and frugal heuristics: The adaptive toolbox. *Simple heuristics that make us smart* içinde, Evolution and cognition. (ss. 3–34), (Oxford University Press).

Hu, Minqing, Bing Liu. (2004), Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* içinde (ss. 168–177).

Li, Xiaodong, Haoran Xie, Li Chen, Jianping Wang, Xiaotie Deng (2014), "News impact on stock price return via sentiment analysis". *Knowledge-Based Systems*, *69*, 14–23.

Liu, Laura Xiaolei, Ann E. Sherman, and Yong Zhang (2014), "The long-run role of the media: Evidence from initial public offerings" *Management Science*, *60*(8), 1945-1964.

Lo, Andrew W. (2005). *Reconciling Efficient Markets with Behavioral Finance: The Adaptive Markets Hypothesis* (SSRN Scholarly Paper No: ID 1702447). Rochester, NY: Social Science Research Network.

Loughran, Tim. ve Bill McDonald, (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65.

Miner, Gary, John Elder IV, Thomas Hill (2012), *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. (Academic Press)

Mitra, Gautam, Leela Mitra (Ed.) (2011), *The handbook of news analytics in finance* (Wiley).

Nassirtoussi, Arman Khadjeh, Saeed Aghabozorgi, Teh Ying Wah ve David Chek Ling Ngo. (2014), "Text mining for market prediction: A systematic review". *Expert Systems with Applications*, 41(16)

Ryan, Paul ve Richard Taffler (2002), "What Firm-Specific News Releases Drive Economically Significant Stock Returns and Trading Volumes" (SSRN Scholarly Paper No: ID 314880). Rochester, NY

Schumaker, Robert P., Yulei Zhang, Chun-Neng Huang, Hsinchun Chen (2012), Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458–464.

Shang, Zilu, Chris Brooks, and Rachel McCloy (2014), "Are investors guided by the news disclosed by companies or by journalists?" *Journal of behavioral and experimental finance*, *1*, 45-60.

Sinha, Nitish Ranjan (2015), Underreaction to News in the US Stock Market. *Quarterly Journal of Finance*, 6.02 (2016): 1650005.

Steve Skovran. (2001). System and method for influencing a position on a search result list generated by a computer network search engine. United States of America Patent No:6, 269, 361, http://www.google.com/patents/US6269361 adresinden erişildi. (03.03.2016)

Taşcı, Şerafettin, Tunga Güngör. (2013). Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications*, *40*(12), 4871–4886.

Tetlock, Paul C. (2007), Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139–1168.

Tutar, Kadir, Murat Osman Ünalır, Levent Toker (2015), Sosyal Ağlar Üzerinde Ontoloji Tabanlı Sezgi Analizi için bir Uygulama Çerçevesinin Geliştirilmesi. *Pamukkale University Journal of Engineering Sciences*, *21*(5).

Tversky, Amos. ve Daniel Kahneman, (1974). *Judgment under Uncertainty: Heuristics and Biases*. Science, 185(4157), 1124–1131.

Uçan, Alaettin (2014), *Otomatik Duygu Sözlüğü Çevrimi ve Duygu Analizinde Kullanımı*. (Yayımlanmamış yüksek lisans tezi). (Hacettepe Üniversitesi)

Urquhart, Andrew Robert Hudson (2013), Efficient or adaptive markets? Evidence from major stock markets using very long run historic data. International Review of Financial Analysis, 28, 130–142.

Van Nieuwerburgh, Stijn, and Laura Veldkamp (2009), "Information immobility and the home bias puzzle." *The Journal of Finance*, *64*(3), 1187-1215

Wisniewski, Tomasz Piotr ve Brendan Lambe. (2013), "The role of media in the credit crunch: The case of the banking sector" *Journal of Economic Behavior & Organization, Financial Sector Performance and Risk*, 85, 163–175.