METİN MADENCİLİĞİ

Metin Madenciliği, işletme dokümanları, müşteri yorumları, web sayfaları ve XML dosyalarını içeren, yapısal olmayan verilerden, önceden bilinmeyen, potansiyel olarak kullanışlı bilgiyi keşfetme sürecidir. Elde edilen bilgiyle, analiz edilecek olan metin kaynaklarında açık olarak görülmeyen ilişkiler hipotezler veya eğilimler olduğu anlaşılır (MECCA, RAUNICH, & PAPPALARDO, 2007; WITTEN, 2003)

Metin Madenciliği, Metin Veri Madenciliği (Text Data Mining) ve Metin Veri tabanlarından Bilgi Keşfi (Knowledge Discovery from Textual Databases) olarak da adlandırılır (DELEN & CROSSLAND, 2008).

Metin Madenciliği, işletme arşivinde veya internet üzerindeki belgelerde bu belgeye benzer belgelerin olup olmadığı elle bir sınıflandırma gerekmeden benzerliği hesaplayabilmektir. Bu genelde otomatik olarak çıkarılan anahtar kelimelerin tekrarı sayesinde yapılır (ALPAYDIN, 2000).

Metin madenciliği, veri madenciliğinin bir parçası olarak düşünülmesine rağmen alışılagelen veri madenciliğinden farklıdır. Ana farklılık, metin madenciliğinde örüntülerin olay tabanlı veri tabanlarından daha çok, *doğal dil metinlerinden* çıkartılmasıdır (DELEN & CROSSLAND, 2008).

Metin madenciliğinin yararları, metinsel verilerin büyük bir çoğunluğunun işletme işlemlerinden elde edildiği alanlarda açıkça görülür. Örneğin, müşteri serbest formundaki şikayet ve memnuniyet metinlerinden gelen anlamlı bilgiler, ürün geliştirme, hata izleme garanti süresi gibi konularda işletmeye girdi oluşturur (DELEN & CROSSLAND, 2008).

Metin Madenciliğinin ne yaptığına bakıldığında en temel seviyede yapısal olmayan metin belgelerini sayısallaştırıp daha sonra veri madenciliği araç ve tekniklerini kullanarak onlardan anlamlı örüntüler çıkarttığı görülür. Başka bir deyişle metin madenciliği, en genel haliyle doğal dilde yazılmış metinler içinden,

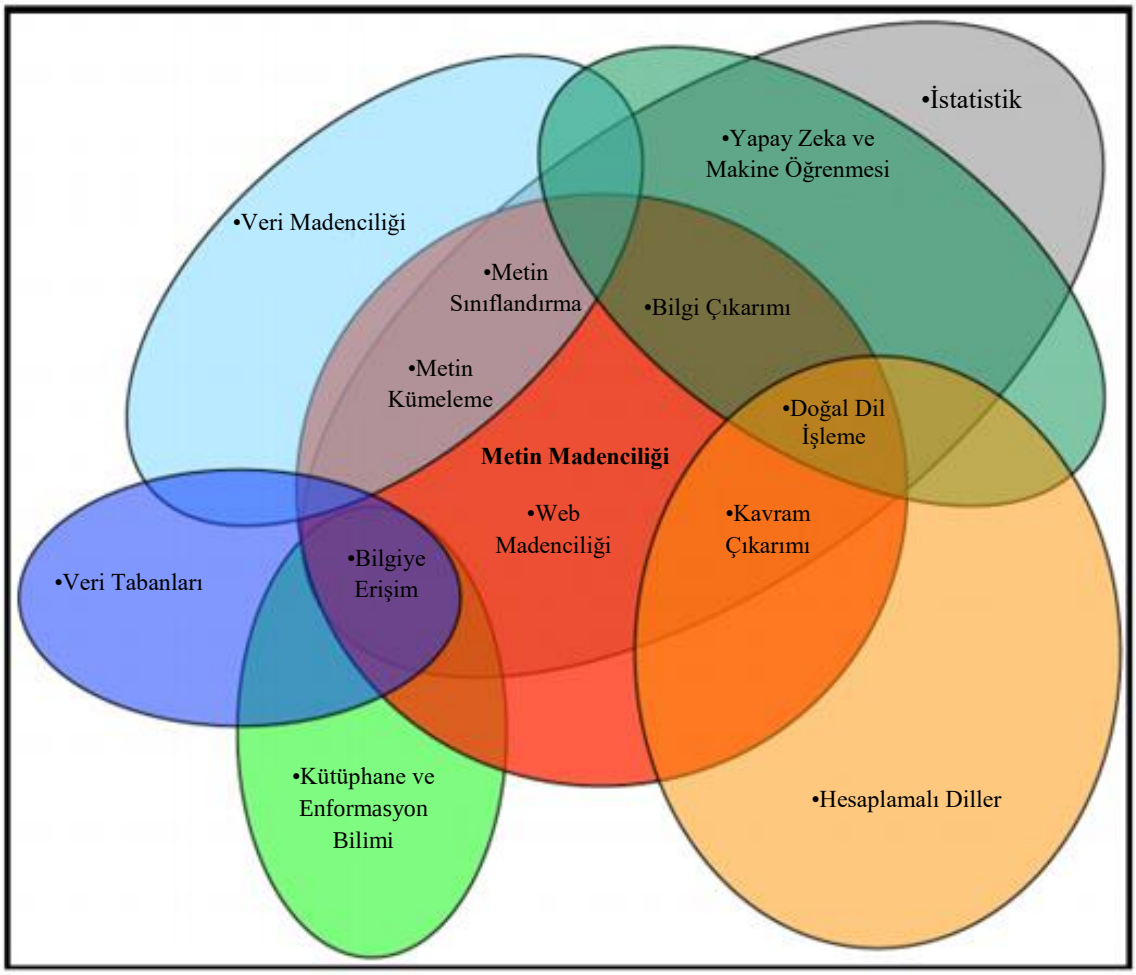
- aynı konudaki belgeleri bulur,
- birbiriyle ilişkili belgeleri bulur,
- ve bulunan belgeleri sıralar.

Daha ileri düzeyde bakıldığında Metin Madenciliği teknikleri belgeyi özetlemek, bilgi çıkarmak amacıyla da kullanılır.

Metin madenciliğinin yararları, metinsel verilerin büyük bir çoğunluğunun işletme işlemlerinden elde edildiği alanlarda açıkça görülür. Örneğin, müşteri serbest formundaki şikayet ve memnuniyet metinlerinden gelen anlamlı bilgiler, ürün geliştirme, hata izleme garanti süresi gibi konularda işletmeye girdi oluşturur (DELEN & CROSSLAND, 2008; MOHAMMAD, 2007).

Yapısal olmayan metinleri otomatik işlemenin kullanıldığı diğer alan, elektronik iletişim ve e-maillerdir. Metin madenciliği yalnızca sınıflandırmaya ve junk mailleri filtrelemeye yardım etmez aynı zamanda e-maillere otomatik olarak cevap vermekte de kullanılır. Metin madenciliği yargı, sağlık ve diğer endüstrilerde geleneksel olarak zengin belgeler ve sözleşmelerle elde edilen verilere de ulaşmayı sağlar (DELEN & CROSSLAND, 2008) (MOHAMMAD, 2007).

Miner vd.(2012)'e göre Metin Madenciliğinin ilişkili olduğu disiplinler ve yöntemler Şekil 1.'de gösterilmiştir (MINER, DELEN, ELDER, FAST, HILL, & NISBET, 2012)



Şekil 1. Metin Madenciliği ve İlişkili Olduğu Alanlar

Başka bir açıdan bakıldığında giderek artan belge yığınlarının faydalı bilgiye dönüştürülmesini sağlamak için geliştirilen Metin Madenciliği çalışmaları, Bilgiye Erişim (Information Retrieval) ve Bilgi Çıkarımı (Information Extraction) olmak üzere iki alanda incelenmektedir.

Zohar'a (2002) göre Metin Madenciliği metotları,

- Bilgiye Erişim (Information Retrieval),
- Bilgi Çıkarımı (Information Extraction),
- Web Madenciliği (Web Mining),
- Kümeleme (Clustering),

olmak üzere dört grupta toplanmaktadır (ZOHAR, 2002). Buna göre Tablo 1.'de bu metotların girdi ve çıktıları özetlendiği gibidir.

Tablo 1. Metin Madenciliği Metotlarının Girdi ve Çıktıları (ZOHAR, 2002)

Bilgiye Erişim	Bilgi Çıkarımı	Web madenciliği	Kümeleme
Girdi: Metin Belgesi Kaynağı, Kullanıcı sorgusu (metin tabanlı) *Çıktı:* Sorgu ile ilişkili olan sıralanmış belgeler kümesi	Girdi: Metinsel belgeler kaynağı İyi tanımlanmış sınırlandırılmış sorgu Çıktı: İlişkili bilgi cümleleri İlişkili bilginin çıkarımı ve ilişkili olmayan bilginin yok sayılması Önceden belirlenmiş formatta çıktı ve ilgili bilgi linki.	Webteki özel bilginin çıkarımı ve metinsel belgelerin erişimi ve indekslenmesi	Benzer metin belgelerinin toplanması

Bilgiye Erişim (Information Retrieval)

Bilgiye Erişim kavramı ilk kez Calvin Mooers tarafından 1948 yılında "Application of Random Codes to the Gathering of Statistical Information" başlığını taşıyan yüksek lisans tezinde *Information Retrieval* terimi altında kullanılmıştır. Vickery, Mooers'in kavrama İngilizce olarak getirdiği ilk tanımı şu şekilde aktarır. Bilginin bir depodan özelliklerine göre konusal olarak aranarak erişilmesidir (TÜRKEEŞ, 2007).

Bilgiye Erişim (IR), metin madenciliğinde ilk adımdır. IR'nin amacı kullanıcıların bilgi ihtiyaçlarını karşılayacak olan belgeleri bulmasına yardımcı olmaktır.

IR, birçok konu alanına sahipliği nedeniyle geniş bir alana yayılmaktadır ve kullanıcıların belirli konulardaki belgeleri bulabilmesi gibi büyük bir topluluktan oluşan metni sunması için modeller geliştirmiştir. Problem, kullanıcı şu an ne ile ilgilenmekte ve belirli bir konu kümesi hakkında belgeler nasıl sunulmalı ve tanımlanması gibidir (SEZER, 2006).

Bilgiye Erişim, bilgi ihtiyacını karşılayan yapılandırılmamış materyalleri (genellikle dokümanlar) geniş bir koleksiyonun içerisinden bulmaktır. Eskiden bilgiye erişim sadece bazı meslek grupları tarafından özel amaçlar için kullanılmaktaydı. Fakat değişen günümüz dünyasında, milyonlarca insan mail ve web aramaları için kullanmaktadır. Böylelikle IR geleneksel veri tabanı arama yöntemlerinin önüne geçmeye başlamıştır.

Bilgiye Erişim sistemlerinde kullanılan standart iki ölçü vardır (CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

Recall (Doğruluk): Araştırmacı tarama yaptığı konularda bütün kaynaklara erişmek istemektedir. Bilgi sistemlerinde araştırmacının bu isteğinin karşılanma derecesi Recall ile ifade edilir. Recall, bir bilgi sisteminin sorgu ile ilgili olarak bulduğu yayınların içindeki gerçekten sorgu ile ilgili olan yayınların sayısının veritabanında bulunan ilgili yayınların

sayısına oranını gösterir (PİLAVCILAR, 2007; SEZER, 2006; CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

$$Recall = \frac{Veritabanı\ içinde\ dönen\ ilgili\ belge\ sayısı}{ilgili\ toplam\ doküman}$$

Precision (Duyarlık):Araştırmacı istediği bilgileri çok fazla zaman harcamadan bulmak istemektedir. Zaman söz konusu olunca ilk akla gelen bilgi sisteminin tarama hızıdır. Ancak hızlı bir tarama sistemi araştırmacının amacı açısından yeterli değildir. Araştırmacının bilgi sisteminin kendisine sorgu ile ilgili olarak gösterdiği yayınlarda gerçekten ilgili olanları seçmesi gerekmektedir. Araştırmacının zamanının büyük bir kısmı da bu evrede harcanmaktadır. Araştırılan yayınları bulma süresini doğrudan etkileyen ve tarama sonuç listesinin iyiliğini gösteren bu özellik ise Precision olarak adlandırılır. Precision bir bilgi sisteminin sorgu ile ilgili olarak bulduğu yayınların içindeki kullanıcının istediği yayınların sayısının bulunan yayınların sayısına oranıdır (PİLAVCILAR, 2007; SEZER, 2006; CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

$$Precision = \frac{Veritabanı\ içinde\ dönen\ ilgili\ belge\ sayısı}{geri\ dönen\ doküman}$$

Recall ve *Precision* ölçümlerinin her ikisini birden arttırmak bilgilerin tasnif edilmesi ile olur. Bu konudaki robotların *Recall* ve *Precision* oranları düşüktür. Kütüphanelerin ise yüksektir.

Bilgiye Erişim sistemlerinde ağırlık(w) verme önemli bir rol oynar ve birçok farklı ağırlık verme modeli geliştirilmiştir. En yaygın olarak kullanılan model, yerel(local) ve genel(global) ağırlık verme şemalarının bir arada kullanılmasıdır. Yerel ağırlık vermede terim frekansı (Term Frequency, TF), genel ağırlık vermede ise ters doküman frekansı (Inverse Document Frequency,IDF) kullanılır (PİLAVCILAR, 2007; SEZER, 2006; CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

Terim Frekansı (TF), bir doküman içerisinde bir terimin tekrar sıklığıdır. Ters Doküman Frekansı (IDF) bir terimin bütün doküman koleksiyonu *(D)* içindeki önemidir.

Bu modele göre, terimin önemi, belge içerisinde o terimin geçme sayısıyla doğru orantılıyken; bütün belge havuzu içerisinde o terimin geçme sıklığıyla ters orantılıdır. *D* belgesinde, *i* teriminin ağırlığı şu şekilde hesaplanır

$$w_i = tf_i \times \log\frac{D}{df_i}$$

Frekansı düşük olan terimler için IDF skoru yüksek, frekansı yüksek olan terimler için IDF skoru düşüktür.

TF-IDF değeri, az miktarda doküman içerisinde terim yüksek miktarda geçiyor ise yüksek değer alır. Eğer terim her dokümanda geçiyorsa TF-IDF değeri en düşük değerini alır (PİLAVCILAR, 2007; SEZER, 2006; CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

Bilgi Çıkarımı (Information Extraction)

Bilgi çıkarımı en basit şekliyle geniş ölçekli bilgilerden özet çıkarılması olarak adlandırılabilir. Başka bir ifadeyle büyük veri yığınları içerisinden özet bilgiler elde edilmesidir. Anahtar kelimeler veya örnek dokümanlar gibi kullanıcı girişleriyle bağlantılı olan bilgi ya da dokümanların bulunması bilgi çıkarımı örnekleridir. Bu çalışmalar sonucunda web sayfalarından bilgiler karşılaştırılarak bulunabilir, geniş ölçekli metinlerden özet bilgiler çıkarılabilir, sorgulara karşılık gelen ifadeler bulunabilir (SODERLAND, 1999).

Bilgi Çıkarımı, yöntemleri metin içindeki unsurları varlıkları otomatik olarak çıkarır ve bunlar arasındaki ilişkileri ortaya koyar. Metin içindeki cümleler ve paragraflar içerdikleri önermelerle varlıklara ait bilgiler taşır. Bilgi Çıkarımı teknikleri bu önermelere bağlı olarak belgeyi oluşturan varlıkları ve bu varlıklar arasındaki ilişkileri çıkarırlar (DAŞ, 2008; KAISER & MIKSCH, 2005).

Bilgi çıkarım işleminin en zor adımlarından birisi de veriyi işlerken belirli bir yapıya oturtmaktır. Örneğin internet üzerinde yayınlanan verilerin herhangi bir standart yapısı bulunmamakta, veriler dağınık halde istenildiği gibi yayınlanmaktadır.

Bilgiye Erişim yöntemlerine nazaran daha etkin sonuçlar elde edilmesini sağlayan bilgi çıkarma tekniklerinin avantajı belge içindeki içeriğin anlamını ön plana çıkaran terimlerin ve terimler arası ilişkilerin bulunmasında yatar. Ancak bazen belgelerin incelenmesindeki amaç daha önceden fark edilmemiş gerçeklerin ve ilişkilerin ortaya çıkarılmasıdır. Bu aşamada devreye bilgi keşfi teknikleri girer. Bilgi keşfi için kullanılan yöntemler metnin içeriklerini derler, birbiri ile entegre eder ve başka kaynaklardan elde edilen sonuçlarla birleştirilerek üst seviye bir anlam ve ilişki kümesi oluşturmaya çalışır. Özellikle konuya bağlı olarak terimler ve terimler arası ilişkilerin üzerine de çıkılır ve konuya özel yapılar ve fonksiyonlara bağlı bir ilişki kümesi oluşturulur. Bu amaçla geliştirilen sistemlerin sadece belgeleri değil veri tabanlarındaki verileri de kullanması gerekir (KUSHMERICK, 1997).

Bilgi çıkarım işlemi, temelde anahtar kelime ve/veya benzerlik tabanlı çıkarımlara dayanmaktadır (KUSHMERICK, 1997). Anahtar kelime tabanlı bilgi çıkarımında, herhangi bir doküman ya da metinden bilgi çıkarılırken anahtar kelimelerden oluşan bir küme oluşturulur.

Benzerlik tabanlı çıkarım sistemleri ortak anahtar kelimeler kümesini temel alarak, benzer dokümanları bulmaktadır. Bu tür bir çıkarımın çıktısı, kelimelere yakınlığı ve birbirleriyle ilişki derecelerini temel almaktadır. Günümüzde internet ve bilgi teknolojilerinin hızla gelişmesi ve insanların hayatında önemli bir yer tutması sebebiyle, bu ortamlardan bilgi çıkarımı önem kazanmıştır. Herhangi bir ürünün satış sitelerinden aranması ve karşılaştırmalı olarak değerlendirilmesinden, elektronik posta içeriklerinin yorumlanmasına kadar çeşitli uygulamalar internetten bilgi çıkarımı işlemine örnek olarak düşünülebilir.

Bilgi Çıkarım sistemi sonuçlarının değerlendirilmesinde bilgi erişim sistemlerinde de olduğu gibi duyarlık ve doğruluk ölçütleri kullanılmaktadır. Fakat burada belgeler yerine, yapılan tahminler ölçüm değişkenleri olarak kullanılmaktadır. Duyarlık, sistemin doğru yaptığı tahminlerin tüm tahminlere bölümü ile hesaplanmaktadır. Doğruluk ise sistemin yaptığı doğru tahminlerin metinde bulunan bütün varlıkların sayısına bölünmesi ile elde edilmektedir (OFLAZER, 2002; GÜVEN, 2007)

Bilgiye Erişim ve Bilgi Çıkarımının Karşılaştırılması

Bilgi Çıkarımı, bilgi parçalarını çıkarmak için doğal dil işlemeyi temel alan bir teknolojidir. Bu süreç girdi olarak metinleri ele alır ve çıktı olarak belirli bir formatta açık şekilde ifade edilebilecek veriler üretir. Bu veri kullanıcıların görüntü elde etmesi için doğrudan kullanılabilir veya daha sonra analiz etmek için veri tabanında veya elektronik tablolarda saklanabilir veya Google gibi internet arama motorlarında olduğu gibi Bilgiye Erişimi uygulamalarında dizinleme amaçlarını yerine getirmek için kullanılabilir (TURMO, AGENO, & CATALA, 2006).

Bilgi Çıkarımı, Bilgiye Erişimden oldukça farklıdır;

- Bilgiye Erişim sistemi uygun metinleri bulur ve bunları kullanıcıya sunar.
- Bilgi Çıkarımı uygulaması metinleri analiz eder ve sadece kullanıcıların ilgilendikleri metinlerden özel bilgi elde ederler (TURMO, AGENO, & CATALA, 2006).

Örneğin, tarım ürünlerini ilgilendiren ticari grup yapılarından bilgi bekleyen bir Bilgiye Erişim sistemi kullanıcısı uygun kelime listesini girecektir ve karşılığında olası eşleşmeleri içeren belge kümesine (örneğin gazete makaleleri) ulaşacaktır. Daha sonra kullanıcı belgeleri okuyacaktır ve bilgilerin içerisinde bir ayıklama işlemi gerçekleştirecektir. İşlem uygulandıktan sonra elektronik tablo halinde bilgi girişi yapılabilir ve bunlardan rapor veya sunu çizelgeleri oluşturulabilir. Bunun tersine bir Bilgi Çıkarımı sistemi uygun şirket ve grup adlarını doğrudan ilgilendiren değerleri otomatik olarak elektronik tablo halinde sunacaktır (TURMO, AGENO, & CATALA, 2006).

Metin Madenciliğinin Diğer Uygulamaları

1) ***Konu izleme:*** Kullanıcı profillerini kullanarak ve kullanıcı görüşlerinden oluşturulan belgelere bağlı olarak kullanıcı için ilginç olabilecek diğer belgelerin tahmin edilmesidir (DELEN & CROSSLAND, 2008). Sosyal ağlarda kullanıcı profillerine göre farklı kullanıcılara farklı reklamların gösterilmesi bu konuya örnek olarak verilebilir.

2) ***Özetleme (Summarization):*** Okuyucuya zaman kazandırmak amacıyla belgenin aslını bozmadan metnin özetlenmesi olarak tanımlanabilir. Başka bir deyişle otomatik metin özetleme bir bilgisayar programı aracılığıyla istenilen metinlerin özetinin çıkarılmasıdır. Belge özetlemenin amacı bir belgenin amacını anlatan kısa bir özetinin otomatik olarak oluşturulmasıdır. Etkin bir özetleme sistemi kullanıcıların arama sonucu olarak elde ettikleri belgelerin özetlerine bakarak tüm belgeyi inceleme zorunluluğu olmadan doğru belgeye ulaşıp ulaşamadıklarını belirleyebilmeleridir (DELEN & CROSSLAND, 2008; TÜLEK, 2007).

Bu konudaki ilk çalışma 1959 yılında Luhn adlı bir bilim adamı tarafından yapılmıştır. Luhn kelimelerin kullanım frekansından yararlanmıştır (TÜLEK, 2007).
Metin Özetleme çeşitleri iki grupta toplanabilir.

- Cümle Seçerek Özetleme (Extract Summarization)
- Yorumlayarak Özetleme (Abstract Summarization)

Cümle seçerek özetlemede (Extract Summarization) özetlenecek metin önemli cümleler, istatistiksel metotlarla, sezgisel çıkarımlarla veya bunların ikisinin kombinasyonuyla seçilerek bu cümlelerden oluşan bir özetleme yapılır. Bu özetlemede özeti oluşturan cümleler içeriği akıllıca değişik şekilde anlatan cümleler değil, yazı içinden seçilmiş olan önemli cümlelerdir (TÜLEK, 2007).

Yorumlayarak özetlemede (Abstract Summarization) özetleme, özetlenecek metnin akıllıca yorumlanması ile yapılır. Bu özetlemede orijinal metindeki ifadeler akıllı bir şekilde kısaltılarak tekrar yazılmaya çalışılır (TÜLEK, 2007).

3) ***Sınıflandırma:*** İçinde önceden tanımlanmış konu kategorilerinin yer alacağı şekilde bir belgenin ana temalarının tanımlanmasıdır. İçerik bazlı belge yönetimi işi belgelere ulaşımda esnekliği amaçlamaktadır. Metin sınıflama çalışması bu amaca ulaşmak için kullanılan bir adımdır ve konuşma dili ile yazılmış metinleri anlamlarına göre daha önceden belirlenmiş sınıflara ayırmaya çalışır. Günümüzde metin sınıflama kontrollü bir kelime haznesine bağlı olarak belgeleri indeksleme, belgeleri filtreleme, otomatik olarak metadata oluşturma web sayfalarını otomatik olarak hiyerarşik düzenlemeye tabi tutma gibi pratik olarak uygulanan pek çok alanda görmek mümkündür (DELEN & CROSSLAND, 2008; TÜLEK, 2007).

4) ***Kümeleme:*** Metin madenciliğindeki önemli noktalardan biri de kümeleme metotlarıdır. Kümeleme önceden belirlenmiş bir kategoriler kümesine sahip olmaksızın birbirine benzer belgelerin gruplandırılmasıdır. Karar ağaçları, makine öğrenmesi, istatistik gibi çeşitli teknikler bu nokta için kullanılmaktadır. Bunların içinden en önemlileri, karar ağaçları, yapay sinir ağları bulanık mantık, yaklaşımlı kümeler ve içerik öğrenmedir. Benzer belgelerin aranması da metin madenciliği uygulamasıdır ve benzer olarak ön işleme ve sınıflandırma kümeleme aşamalarını içerir (AMASYALI, 2008).

Başka bir deyişle kümeleme verilerin kendi aralarındaki benzerlikleri göz önüne alınarak gruplandırılması işlemidir.

Herhangi iki doküman arasındaki benzerlik, dokümanların Vektör Uzay Modeli ile vektör haline getirilmesinden sonra **Kosinüs Benzerliği** ile hesaplanır.

$d_1 \; ve \; d_2$ iki doküman vektörü olduğunda, kosinüs benzerliği şu şekilde hesaplanır.

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\|.\|d_2\|}$$

$d_1 \bullet d_2$: iki dokümanın vektör çarpımı
$\|d_i\|$:di dokümanının (vektörünün) uzunluğu (normu) (G. ÖĞÜDÜCÜ, 2011)
Kümeleme işlemlerinde değerlendirme için iki ölçüt vardır. Entropi ve F-Ölçütü (F-Measure).

Entropi: Rassal bir değişkenin belirsizlik ölçütü olarak bilinen Entropi, bir süreç için tüm örnekler tarafından içerilen enformasyonun beklenen değeridir. Entropi sadece nesnelerin baskın sınıfta olup olmadığını ölçmekle kalmaz, kümelerdeki sınıfların dağılımlarını da dikkate alır. Entropi değerinin 0 olması kümenin tamamen tek bir sınıftan oluştuğunu gösterirken, 1'e yakın bir entropi değeri kümenin bütün sınıfların tekdüze (uniform) dağılıma göre oluşmuş bir karışımını içerdiğini gösterir.

$$E = -\sum_{i=1}^{n} p_i * \log_2(p_i)$$

p_i: i $mesajının$ üretilme olasılığı

F-Ölçütü: F-Ölçütü (F-Measure), yaygın olarak kullanılan diğer bir dışsal kalite ölçüm yöntemidir. Kümeleme Doğruluğu (clustering accuracy) olarak da bilinir. F-Ölçütü için Precision ve Recall değerlerinin hesaplanması gerekmektedir.

$$F(i,j) = \frac{2 * Recall(i,j) * Precision(i,j)}{Recall(i,j) + Precision(i,j)}$$

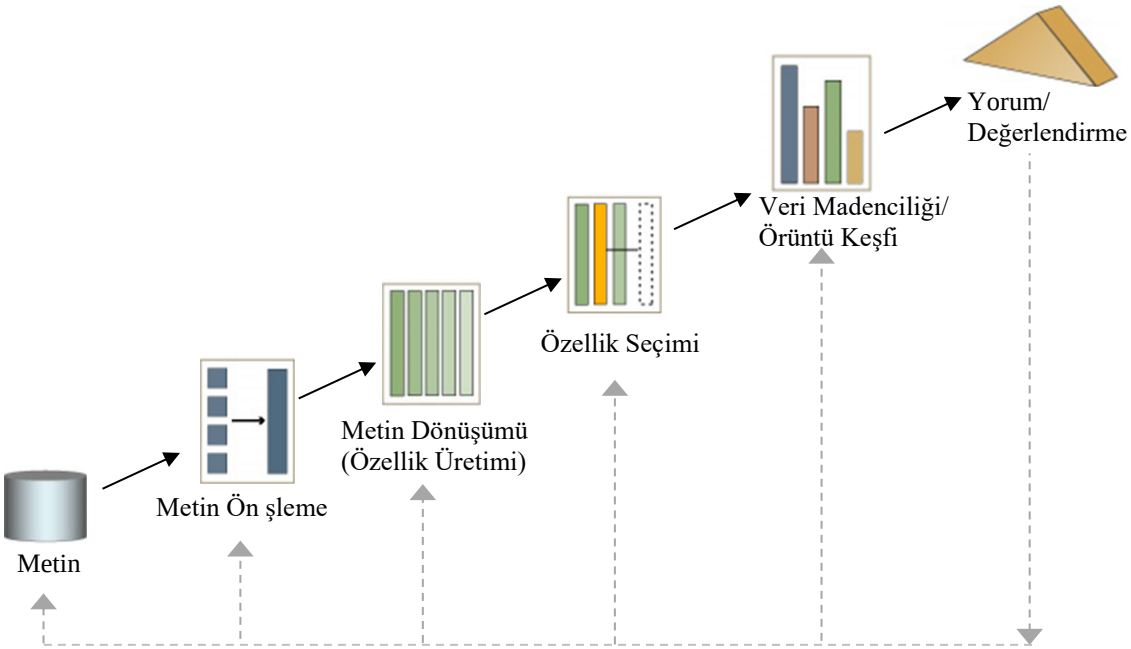
5) ***Kavram bağlama-ekleme:*** Yayın olarak paylaşılan kavramları tanımlayarak ve böyle yaparak da geleneksel arama metotlarını kullanarak aradıklarını bulamayan kullanıcılara yardım ederek ilişkili belgelerin bağlanmasıdır (DELEN & CROSSLAND, 2008).

6) ***Soru cevaplama:*** Bilginin yönlendirdiği örüntü eşleştirmeleriyle verilen bir soru için en iyi cevabın bulunmasıdır (DELEN & CROSSLAND, 2008).

Soru cevaplama sistemleri bilgiye ulaşma ihtiyacı ile ortaya çıkmış olan ve genelde bilgisayar destekli yapılardır. Soru-cevap benzerliklerini karşılaştırarak ya da varolan kaynaklar üzerinde yapay zeka gibi insan türevi teknikler uygulanarak sorulara yeni cevaplar üretmeye çalışan sistemler geliştirilmiştir (AMASYALI, 2008).

Metin Madenciliğinin Adımları

Büyük miktardaki metinsel verilerden potansiyel olarak yararlı ve önceden bilinmeyen belirli bir önemi olan bilginin çıkarılması olarak nitelendirilen Metin Madenciliği şekilde görüldüğü gibi temelde altı adımdan oluşmaktadır. Metin Madenciliği işlemleri, Veri Madenciliğine benzer olarak Şekil 2.'deki gibi özetlenebilir.



Şekil 2. Metin Madenciliğinin Adımları (ZOHAR, 2002)

Bu bilgilerden sonra Metin Madenciliği işlemleri ve içerdikleri yöntemler Tablo 2.'de görüldüğü gibi özetlenebilir. Bunlar hakkında detaylı bilgili izleyen kısımda yer almaktadır.

Tablo 2. Metin Madenciliği İşlemleri (ZOHAR, 2002)

Metin	Metin Ön İşleme	Metin Dönüşümü	Özellik Seçimi	Veri Madenciliği/ Bilgi Keşfi	Yorum/ Değerlendirme
	Söz dizimsel/ Semantik analiz Sözcük türü etiketleme Kelime anlamı belirginleştirme Ayrıştırma(parsing)	Kelime torbası, Kelimeler Kök bulma, Durdurma kelimeleri	Basit hesaplama İstatistik (boyut azaltma, ilişkisiz özellikler)	Sınıflandırma (danışmanlı) Kümeleme (Danışmansız)	Analiz Sonuçları

Metin koleksiyonu oluşturma

İlgilenilen konularda bilgiye erişim sistemleri kullanılarak metin koleksiyonu oluşturma sürecidir. Bu süreç, günümüzde genel olarak internet üzerinden, özellikle Google vb. arama motorları kullanılarak gerçekleştirilmektedir. Çevrim içi veri tabanlarının yanı sıra veri tabanlarında ya da kişisel bilgisayarlarda bulunan metin türü veriler ile oluşturulan koleksiyonlar da metin madenciliğinde kullanılmaktadır (PİLAVCILAR, 2007; OĞUZ, 2009).

Metin önişleme

Metni kelimelere ayırma, kelimelerin anlamsal değerlerini bulma (isim, sıfat, fiil, zarf, zamir vb.), kelimeleri köklerine ayırma ve gereksiz kelimeleri ayıklama, yazım kurallarına uygunluğunu tespit etmek ve var olan hataları düzeltmek gibi metin belgelerin yapıtaşı olan kelimelerle ilgili işlemleri içeren süreçtir.

Metin madenciliğinin en büyük sorunu işleyeceği veri kümesinin yapısal olmamasıdır. Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışılan metin madenciliği alanında ön işleme aşaması veri temizlemenin yanında veriyi uygun formata getirme işlemini de gerçekleştirmektedir (GÜVEN, 2007).

Belgeler için dizin oluşturmadan önce yapılacak ön işleme işlemleri şöyledir.

- Doküman doğrusallaştırma
  - ✓ *Markup & Format Removal:* Dokümanı oluşturan etiket ve özel formatların çıkarılması)
  - ✓ *Tokenization*: Metin küçük harflere çevrilmesi ve noktalama işaretlerinin çıkarılması)

Metin önişleme çalışmaları aynı zamanda doğal dil işleme çalışmaları kapsamında incelenen bir alandır. Doğal dil işlemenin belge analizi sürecindeki en önemli faydası terimlerin yani kelimelerin ayrıştırılması, eklerinden arındırılarak anlamını kaybetmeyen en kısa biçimlerine dönüştürülmesidir. Çünkü aynı anlam için kullanılan kelimeler dilbilgisi kuralları gereği farklı biçimlerde bulunabilir ve bu farklı kullanım biçimleri ortadan kaldırılmadığı takdirde farklı anlam taşıyan terimler gibi işleme alınarak, belgelerin gerçek anlamına ulaşılmasını engelleyebilirler. Doğal dil işleme çalışmaları kapsamında yürütülen girişimler dört ana grup altında toplanabilir (ÖZBİLİCİ, 2006).

a) Morfolojik (Biçimbirimsel) Analiz
b) Sözdizimsel (Sentaktik) Analiz
c) Semantik (Anlamsal) Analiz
d) Anlam kargaşasının giderilmesi

a) Morfolojik Analiz: Biçimbirim, sözcüklerin yapısıyla ile ilgilenir. Türkçe için sözcüklerin türetilmesi ve ekler çok önem taşır. Her dilde iki farklı şekilde sözcük oluşturulabilir. Bunlardan biri çekim, diğeri ise türetme yöntemidir. Çekim yoluyla sözcük oluşturulurken bir sözcüğün farklı şekilleri kullanılır. Türetme ise var olan eski sözcüklere yapım ekleri eklenmesi yoluyla yeni sözcük oluşturma yöntemidir (ÖZBİLİCİ, 2006; NABİYEV, 2010; KESGİN, 2007).

b) Sözdizimsel Analiz: Bilgisayarla doğal dil modellemelerinde anlamsal analize geçmeden önce, kelimeler yığınının geçerli bir cümle yapısı oluşturup oluşturmadığı kontrol edilmelidir. Rasgele kelimelerin yan yana gelmesiyle geçerli bir cümle meydana gelmeyecektir. Geçerli bir cümle yapısı oluşturulamadığı zaman, buradan anlam çıkarılmasını beklemek yanlış olacaktır (ÖZBİLİCİ, 2006; NABİYEV, 2010; KESGİN, 2007).

Sözdizimsel analiz, cümlenin yapısal bir tanımını oluşturabilmek için morfolojik analizin sonuçlarını kullanır. Bu işlemi yapmanın amacı, ardı ardına gelen kelime yığınlarının bu

kelimeler yığınının ifade ettiği cümle birimlerini tanımlayan bir yapıya dönüştürmektir. Cümle birimleri, kelimeler tamlamalar veya buna benzer cümle parçacıkları olabilir.

c) **Semantik Analiz:** Bir cümlenin ne demek istediğinin anlaşılması, diğer bir deyişle bir cümle ile ifade edilmek istenilen duygu veya düşüncenin ne olduğunun anlaşılması, anlamsal analiz yardımıyla yapılır.

d) **Anlam kargaşasının giderilmesi**: Anlamsal analiz yapılırken, öncelikli olarak kelimelerin tek tek veritabanından uygun nesnelerle eşleştirilme işleminin yapılması gerekir. Bu işlem, her zaman birebir eşleme olamayabilir. Diğer bir deyişle, kelimelerin ifade ettikleri anlamlar her zaman bir tane olmayabilir. Ayrık kelimelerin bir cümledeki doğru anlamını bulma işlemine "kelime anlam berraklaştırılması" denir. Bu işlem, cümle içinde geçen bir kelimenin sözlükteki anlamlarının belirlenip bunlardan uygun olanının seçilmesidir. Cümle içinde geçen her bir kelime, diğer kelimelerin doğru anlamlarının ortaya çıkarılması için önem taşımaktadır (DELİBAŞ, 2008; ÖZBİLİCİ, 2006; NABİYEV, 2010; SAY, 2003).

Metin dönüşümü

Kelimelerin düzgün bir biçimde hecelerine ve eklerine ayrılmasından sonraki işlem, kelimelerin kökünün tespit edilmesidir. İngilizce için kullanılan Porter Stemmer Yöntemi gibi bir kök bulma algoritması kullanmak hızlı olması açısından önemli olsa da Türkçe gibi sondan ekli bir dilde başarı yüzdesi istenen düzeyde değildir ve özel durumları yakalayamamaktadır.

Snowball: Kök bulmak için tasarlanmış küçük bir karakter işleme dilidir. Snowball kullanılarak birçok dil için kök bulma algoritmaları geliştirilmiştir. Türkçe için snowball kullanılarak geliştirilen kök bulma algoritmaları Evren Kapusuz tarafından yürütülmektedir (AYDIN & KILIÇASLAN, 2010; ORHAN, 2006).

Kelime Türü: Bir kelimenin kökü bulunduktan sonraki adım kelimenin türünün bulunmasıdır. Bu işleme Pos Tagging denir. Pos Tagging 2 fazdan oluşur. Birincisi eğitim(training) fazıdır. Bu fazda kelimelerin kökleri manuel olarak tanımlanmış algoritmalar kullanılarak machine learning sistemi vasıtasıyla işlenir. İkinci faz ise tagging fazıdır. Bu fazda, birinci adımda kullanılan algoritma, öğrenilen parametrelere göre yeniden işlenir ve kelimeler türlerine ayrılır.

Stopword İşlemi: Tekrar eden ve tek başına anlam taşımayan kelimelere stopword kelimeleri denir. Bilgiye Erişimde bir stopword listesi, belgeleri bir diğerinden ayırt etme durumuna etkisi olmayan sıklıkla kullanılan kelimeleri içerir. Stopword kelimelerini azaltmak sorgu sürecinin verimini arttırır. Bir stopword listesinin yapılandırması farklı ve bazen rastgele kararları içerir. Bilgiye Erişim literatüründe, verilen özel diller için farklı uzunluklarda stopword listelerini bulmak mümkündür.

Bag of words: Bu aşamada gruplanan tüm dokümanlardaki tüm kelimelerin kullanım sıklıkları hesaplanır ve bir havuzda toplanır. Daha sonrasında ise bu kelimelerin değerleri (Word Weighting) hesaplanır. Kelime değeri, bir kelimenin belirli bir alan (sağlık, spor, politika, ...) ile ilgili bir metnin içinde bulunma sıklığı olarak açıklanabilir. Örneğin 10000 kelimelik spor kategorisindeki bir haberin içinde gol veya hakem kelimelerinin bulunma

sıklıkları, aynı kelimelerin sağlık kategorisindeki bir haber içinde bulunma sıklığına göre kat be kat fazladır (ORHAN, 2006; KARADENİZ, 2007).

Özellik seçme

Metin madenciliği uygulamalarında her zaman gürültülü ve önemsiz bilgi içeren metin koleksiyonlarıyla uğraşma ihtiyacı bulunmaktadır. İlgili verilerin saptanması üzerine odaklanan özellik seçme, büyük miktarlardaki veriler üzerinde işlem yapılırken iş yükünü azaltmada yardımcı olmaktadır. Özellik seçme aşamasında, ön işlemden geçen metinlerdeki önemli kelimeleri (varlıkları) belirleme (isimler, tamlamalar, bileşik kelimeler, kısaltmalar, sayılar, tarihler, para birimleri vb.) ve ilişkili olmayan özelliklerin çıkarılması, sadece birkaç dokümanda gözlemlenen özelliklerin çıkarılması, birçok dokumanda gözlemlenen özellikleri azaltma vb. işlemleri yapılmaktadır (CEBİROĞLU, TANTUĞ, ADALI, & ERENLER, 2003; ERYİĞİT, 2006).

Veri Madenciliği/Bilgi Keşfi

Metinsel verilerden bilgi keşfi için veri madenciliğinde geçen Sınıflandırma ve Kümeleme yöntemleri kullanılabilir. Sınıflandırma yöntemleri şu şekilde özetlenebilir.

- Entropiye Dayalı algoritmalar (ID3, C4.5)
- Sınıflandırma ve Karar Ağaçları (Twoing, Gini,)
- Bellek tabanlı sınıflandırma modelleri (En yakın komşu algoritması)
- Optimizasyon tabanlı Sınıflandırma Modelleri (Destek Vektör Makinesi)
- İstatistiksel Sınıflandırma Modelleri (Navie Bayes)

Kümeleme yöntemleri de aşağıda sıralandığı gibi özetlenebilir.

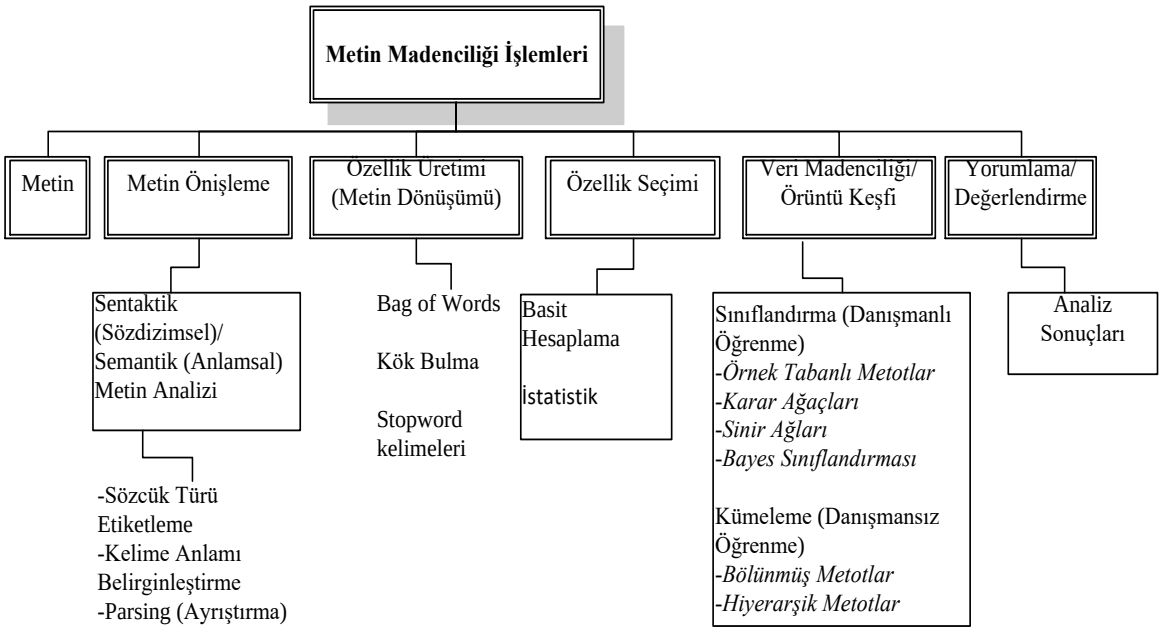
- Hiyerarşik Metotlar (En yakın komşu algoritması, En uzak komşu algoritması,)
- Hiyerarşik olmayan Metotlar (k-ortalamalar)

Veri Madenciliği için belirtilen bu sınıflandırma ve kümeleme yöntemleri ön işleme adımlarından geçirilerek metinsel verilere uygulanmaktadır (ÖZKAN, 2013).

Değerlendirme ve yorumlama

Veri madenciliği yöntemleri ile verilerin analizinden elde edilen sonuçların değerlendirilip kullanıcıya uygun ve anlaşılır bir şekilde sunulması işlemidir.

Metin Madenciliği işlemleri ve içerdikleri yaklaşımlar Şekil 3.'teki gibi özetlenebilir.



Şekil 3. Metin Madenciliği Adımları ve İçerdikleri Yaklaşımlar (ZOHAR, 2002)

KAYNAKLAR

AKAT, Ö., TAŞKIN, Ç., & ÖZDEMİR, A. (2006). Uluslararası Alışveriş Merkezi Tüketicilerinin Satın Alma Davranışı: Bursa İlinde Bir Uygulama. *Uludağ Üniversitesi Sosyal Bilimler Dergisi, 2006*(2).

ALPAYDIN, E. (2000). *Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri.* Bilişim 2000 Eğitim Semineri.

AMASYALI, M. F. (2008). *Yeni Makine Öğrenmesi Metotları ve İlaç Tasarımına Uygulamaları.* İstanbul: Yıldız Teknik Üniversitesi FBE, Doktora Tezi.

AYDIN, Ö., & KILIÇASLAN, Y. (2010). Tümevarımlı Mantık Programlama ile Türkçe için Kelime Anlamı Belirginleştirme Uygulaması. *Akademik Bilişim.* Muğla.

CAN, F., KOÇBERBER, S., BALÇIK, E., KAYNAK, C., ÖÇALAN, H., & VURSAVAŞ, O. (2008). Information Retrieval on Turkish Texts. *Journal of the American Society for Information Science and Technology*, 407-421.

CEBİROĞLU, G., TANTUĞ, A., ADALI, E., & ERENLER, Y. (2003). Sentetik Türkçe Sözcük Kökleri Üretimi. Çanakkale: International XII. Turkish Symposium on Artificial Intelligence and Neural Networks - TAINN.

ÇİÇEKLİ, İ. (2010). *Otomatik Özetleme ve Anahtar Kelime Bulma.* Ankara: TÜBİTAK.

DAŞ, R. (2008). *Web Kullanıcı Erişim Kütüklerinden Bilgi Çıkarımı.* Elazığ: Fırat Üniversitesi FBE, Doktora Tezi.

DELEN, D., & CROSSLAND, M. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 1707-1720.

DELİBAŞ, A. (2008). *Doğal Dil İşleme ile Türkçe Yazım Hatalarının Denetlenmesi.* İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.

ERYİĞİT, G. (2006). *Türkçe'nin Bağlılık Ayrıştırması.* İstanbul: İstanbul Teknik Üniversitesi FBE, Doktora Tezi.

G. ÖĞÜDÜCÜ, Ş. (2011). *İTÜ Veri Madenciliği Ders Notları.* İstanbul.

GÜVEN, A. (2007). *Türkçe Belgelerin Anlam Tabanlı Yöntemlerle Madenciliği.* İstanbul: Yıldız Teknik Üniversitesi FBE, Doktora Tezi.

KAISER, K., & MIKSCH, S. (2005). *Information Extraction A Survey.* Vienna, Avusturya: Vienna University of Technology Institue of Software Technology & Interactive Systems.

KARADENİZ, İ. (2007). *Türkçe İçin Biçimbirimsel Belirsizlik Giderici.* İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.

KESGİN, F. (2007). *Türkçe Metinler için Konu Belirleme Sistemi.* İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.

KUSHMERICK, N. (1997). *Wrapper Introduction for Information Extraction.* University of Washington, Ph.D.

MECCA, G., RAUNICH, S., & PAPPALARDO, A. (2007). A new algorithm for clustering search results search results. *Data & Knowledge Engineering*, 504-522.

MINER, G., DELEN, D., ELDER, J., FAST, A., HILL, T., & NISBET, R. (2012). *Practical Text Mining and Statistical analysis for Non-Structured Text Data Applications.* Waltham, USA: Elsevier.

MOHAMMAD, M. (2007). *Text Mining: A Burgeoning Quality Improvement Tool.* Ankara: Msc. Thesis, METU.

NABİYEV, V. (2010). *Yapay Zeka: İnsan-Bilgisayar Etkileşimi.* Ankara: Seçkin Yayıncılık.

OFLAZER, K. (2002). *Türkçe İçin Bir Sonlu Durumlu "Hafif" Doğal Dil Çözümleyicisi ve Bilgi Çıkarımı Uygulamasının Gerçekleştirilmesi.* TÜBİTAK PROJESİ, PROJE NO:199E027.

OĞUZ, B. (2009). *Metin Madenciliği Teknikleri Kullanılarak Kulak Burun Boğaz Hasta Bilgi Formlarının Analizi,*. Antalya: Akdeniz Üniversitesi Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi.

ORHAN, Z. (2006). *Türkçe Metinlerdeki anlam Belirsizliği Olan Sözcüklerin Bilgisayar Algoritmaları ile Anlam Belirginleştirilmesi.* İstanbul: İstanbul Üniversitesi FBE, Doktora Tezi.

ÖZBİLİCİ, A. (2006). *Türkçe Doğal Dili Anlamada İlişkisel Ayrık Bilgiler Modeli ve Uygulaması.* Sakarya: Sakarya Üniversitesi FBE, Yüksek Lisans Tezi.

ÖZKAN, Y. (2013). *Veri Madenciliği Yöntemleri.* İstanbul: Papatya Yayıncılık.

PİLAVCILAR, İ. (2007). *Metin Madenciliği İle Metin Sınıflandırma.* İstanbul: Yıldız Teknik Üniversitesi FBE, Yüksek Lisans Tezi.

SARAÇOĞLU, R., TÜTÜNCÜ, K., & ALLAHVERDİ, N. (2008). A new approach on search for similar documents with multiple categories using fuzzy clustering. *Expert Systems with Applications*, 2545-2554.

SAY, B. (2003). *Türkçe İçin Biçimbirimsel ve Sözdizimsel Olarak İşaretlenmiş Ağaç Yapılı Bir Derlem Oluşturma.* TÜBİTAK EEEAG Projesi.

SEZER, E. (2006). *Web Sayfaları İçin Anlamsal Erişim Sistemi.* Ankara: Hacettepe Üniversitesi FBE, Doktora Tezi.

SODERLAND, S. (1999). Learning Information Extraction Rules for Semi-structured and Free Text. *Machine Learning*, 233-272.

TURMO, J., AGENO, A., & CATALA, N. (2006). Adaptive Information Extraction. *ACM Computing Surveys*.

TÜLEK, M. (2007). *Türkçe İçin Metin Özetleme.* İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.

TÜRKEEŞ, M. (2007). *Bilgi Erişiminde Tamlama Temelli Dizinleme.* İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.

WITTEN, I. (2003). Text Mining. Computer Science, University of Waikato.

ZOHAR, E. (2002). Introduction to Text Mining. *Supercomputing*. Automated Learning Group National Center for Supercomputing Applications, University of Illinois.