MEHMET AKİF ERSOY ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ DERGİSİ
Mehmet Akif Ersoy University Journal of Social Sciences Institute

e-ISSN: 1309-1387
Sayı/Issue: 31
Yıl/Year: 2020
ss./pp.: 220-239

METİN MADENCİLİĞİ: İMKÂNLAR, YÖNTEMLER VE KISITLAR

TEXT MINING: POSSIBILITIES, METHODS AND LIMITATIONS

Suat ATAN[1]

1. Dr., Tarım ve Kırsal Kalkınmayı Destekleme Kurumu, suat.atan@tkdk.gov.tr, https://orcid.org/0000-0003-3170-0969

Makale Türü Derleme
Article Type Review Article

Başvuru Tarihi/Appliation Date
30.10.2018

Yayına Kabul Tarihi/Acceptance Date
10.03.2020

DOI
10.20875/makusobed.476524

Öz

Dünyada ortalama olarak her gün 2,5 milyar GB verinin üretildiği hesaplanmaktadır. Bu miktarın yaklaşık olarak %80'inin ise metin formunda olduğu tahmin edilmektedir. İnsanların bilgiyi tablolar halinde değil düz yazı formunda, doğal dille kaydetmeleri ve doğal dilin esnekliği nedeniyle bu durum var olmaya devam edecektir. Devasa miktardaki bu metinlerin insanlar tarafından okunarak değerlendirilmesi mümkün değildir. Bu verilerden anlamlı sonuçlar üretmek metin madenciliğinin konusudur. Metin madenciliği sayesinde, metinlerin programlanan algoritmalar yardımıyla özetlenmesi, sınıflandırılması, etiketlenmesi ve seçilmesi mümkündür. Metin Madenciliği bu özellikleri ile tüm organizasyonlar için çok çeşitli fırsatlar sunmaktadır. Türkçe literatürde metin madenciliği alanının uygulamalarından ziyade metin madenciliğinin arka planını ve olanaklarını inceleyen çalışmalara ihtiyaç bulunmaktadır. Bu çalışma da bu boşluğu doldurmayı ve araştırmacıların metin madenciliği olanaklarını incelemelerine yardımcı olmayı hedeflemektedir.

Anahtar Kelimeler: *Veri Madenciliği, Metin Madenciliği, Veri Yönetimi*

Abstract

It is estimated that on average 2.5 billion GB of data is generated every day in the world. More than 80% of this data is in textual form. This will continue to exist because of the tendency of people toward keeping information in the natural language due to its flexibility. It is not possible for people to read and interpret the huge amount of text written in almost every field. In order to manage this information flux and derive results from it, a research field called text mining has emerged. By text mining, processes such as summarization, classification, clustering, labeling and similarity detection can be done with the help of machines. Due to the fact that text mining is a young research field, there are few studies on text mining in Turkish literature. The purpose of this study is to fill the gap and help researchers to assess text mining and research opportunities.

Keywords: *Data Mining, Text Mining, Data Management*

EXTENDED SUMMARY

Research Questions

All of the studies encountered in the Turkish literature on text mining used to analyze related data in various fields. There is a need for a review of text mining methods and approaches. This study aims to fill this gap.

Literature

There are various definitions related to Text Mining: In one of these, text mining is considered as the process of obtaining interesting information or insight from unstructured text (Chen, 2011, p. 5). In another definition, regarding the quality of the information to be obtained, text mining has often been defined as the process of obtaining high-quality information from texts (Mucherino, Papajorgji, & Pardalos, 2009, p. 17). Some studies exploring the relations between text-only topics with the help of text mining as text mining (Babu et al. 2014, p. 12). Hansen and Johnson, on the other hand, have defined as the process of revealing text mining, hidden patterns and patterns in the text (2011, p. 10). Miner et al. interpreted text mining as the process of analyzing people's communication (2012, p. 1009). Indeed, the materials used in text mining are the result of correspondence of people, social media inputs or speeches, as opposed to the records accumulated in databases.

Text mining is defined as an interdisciplinary field (Gupta & Lehal, 2009, p. 60). There are many theoretical and applied studies in the literature on text mining other than Turkish. The methods used by a large number of studies published in the literature other than Turkish are discussed in the areas of usage in this article. Various analyses based on textual data have also been carried out in Turkish literature. One of the first studies on text mining in the Turkish literature was carried out by İlhan et al. In this study, a question answering system was established using text mining methods (2008). Dolgun et al. compared the decision tree algorithms and the methods based on text mining (2009). Montenegro and Takçı, on the other hand, developed an algorithm that can automatically classify similar news (2010). Kılınç et al. have developed a system that supports text mining and machine learning for classifying scientific articles (2016). Çalış and others have developed an application that can detect the content of e-mails with 96.5% accuracy by using the text mining methods and machine learning algorithms, kNN (k-Nearest Neighborhood), and NaiveBayes algorithms (2013). These algorithms are used in many areas, including text mining. Kaşıkçı and Gökçen have developed a Java-based application that can understand whether this site is an e-commerce site by looking at the content of a website by using the kNN and Naive Bayes algorithms (2013). All of these studies are about practical applications rather than a review of approaches and applications of text mining.

Methodology

After examining the definitions related to text mining, the studies in the literature are summarized. Afterward, the usage areas of text mining were classified under the title of possibilities, through numerous studies in the literature written in a foreign language. Finally, limitations in the field of text mining have been studied. In the study, ready-to-use libraries and codes are also discussed.

Conclusion

Most of the common tools used in text mining studies are free and open source. This has led to the fact that these tools have a widespread user base as well as a large amount of explanatory documentation on the use of these resources on the Internet. This creates important opportunities for organizations and researchers at all levels. While analyzing a large number of headlines about a topic by using a text mining method, this can help to get an idea about its overall picture. Financial, political organizations, or brand owners can collect the information from the public by using social media analysis instead of asking public opinion with a questionnaire. Scholars can perform bibliometric analyses instead of reading the whole corpus.

1. GİRİŞ

İnternetin yaygınlaşması, insanlar tarafından üretilen bilgilerin de katlanarak artmasına neden olmaktadır. Veri depolama çözümleri üreten Seagate adlı şirketin raporuna göre, tüm dünyada 2018 yılında 33 zetabayt (1 ZB = 1 katrilyon GB, takriben 1,7 trilyon adet CD kadar yer kaplar) veri üretilmiştir. Bu miktarın 2025'te 175 zetabayt civarında olacağı tahmin edilmektedir. İçerikler arttıkça içeriklerle ilgili etkileşimler de artmaktadır. Haber sitelerinde, yayınlanan haberlere yapılan yorumlar buna örnek olarak gösterilebilir. Aynı şekilde sosyal medya artık bilgi akışının tek yönlü değil çok yönlü olarak gerçekleşmesini sağlamaktadır. İnternette gittikçe artan bu bilgilerin çoğu metin formundadır. Bu verilerin bir anket veya özel olarak doldurulan bir form şeklinde yapılandırılmamış olması yanında miktarının devasa boyutlarda olması bu verilerin yönetimini ve analizini zorlaştırmaktadır. Diğer taraftan her türlü veri için depolama ve işlenme maliyetleri giderek ucuzlamakta ve organizasyonlar verileri kullanarak etkinliklerini arttırmak istemektedirler (Davenport, 2014, s. 22). Söz gelimi Google Fotoğraflar ve Yandex Disk gibi uygulamalar kullanıcılara sonsuz fotoğraf yedekleme alanı hizmeti sunmaktadır. Peki, bu kadar bilgi ne için muhafaza edilmektedir? Müşteri tatmininin ötesinde büyük verinin (big data) onları saklama hizmeti veren işletmelere de analitik faydası bulunmaktadır. Gerek hacim, gerek tür bakımından çok miktardaki bu veriler çok değerli öngörülere vesile olabilir. Bu nedenle iş dünyası ve literatür verilere odaklanmış ve adeta veriyi "madenciliği" yapılacak değerde bir meta olarak kabul etmiştir. Bu sürecin sonunda veri bilimi ortaya çıkmıştır. Veri madenciliği ya da daha geniş anlamda veri bilimi derlenen bu verilerden organizasyonların veya kişilerin etkinliklerini arttırabilmelerine olanak sağlayacak yollar barındırmaktadır. Veri madenciliğinin bir alt dalı olan "metin madenciliği" ise bir veri tipi olan metinlerin kendine has özellikleri ve zorluklarını özel olarak inceleyen ve metinlerin topluca yorumlanabilmesine olanak veren bir alan olarak ortaya çıkmıştır.

Organizasyonlar veri varlıklarını kullanarak geleneksel istatistiksel metotların ötesinde çeşitli yöntemlerle veri kaynakları üzerinde inceleme yapabilirler (Cady, 2017). Madencilik sıfatı ile nitelendirilen bu süreç veri tabanlarında tablolar halinde tutulan yapılandırılmış veriler üzerinden gerçekleştirilen analizlere dayalıdır (Atan, 2016b). Yapılandırılmış (structured) ifadesi bir veri setinin tablolar halinde düzenli olarak tutulduğu bilgisine işaret eder. "Yapılandırılmış" ifadesi çoğunlukla aritmetik işlemlere tabi olabilecek (toplama, alt toplama, ortalama alma, standart sapma vb. hesaplamalar) sayısal verilere gönderme yapmaktadır. Yapılandırılmış veri tipinde tüm veriler tablo yapısı içerisinde satır ve sütun (kolon) formunda bulunmakta, her bir hücrede bulunan bilgi en küçük düzeyde bir sayı, kısa ve kategorik bir metinsel ifade veya doğru/yanlış şeklinde bir 'boolean' değeri şeklinde kayıt altındadır. Elbette uzun metinlerin girildiği alanlar da mevcut olabilir. Ancak bu uzun metinler dışındaki nesneler SQL komutları yardımı ile bir takım analizlere tabi olabilmektedir. Metin bazında ise SQL komutları sadece temel düzeyde sorgulama yapmaya yardımcı olabilir.

Öte yandan, e-postalar, resmi yazışmalar, raporlar, bilgi notları, haberler, tweetler, özgeçmişler gibi birçok içerik türü de veri tabanlarında saklanabilmekle birlikte içerdikleri bilgi tablo halinde değil doğal dil ile yazılmış metin formundadır. Doğal dille yazılmış ifadeler ise ifadenin türüne göre değişen miktarlarda farklı bilgileri ihtiva eder. Bağlaçlar, ünlemler ve bazı ara biçimdeki sözcükler ise hiçbir bilgi içermeyebilir. Bir metnin içeriğini alarak tablolara benzer şekilde analiz yapmak olası değildir (Gupta & Lehal, 2009). Örneğin bir özgeçmiş metni içinde kişinin geçmiş tecrübeleri ve bunun süreleri mevcut olmakla birlikte bu metinden toplam kaç yıl tecrübenin var olduğunun hesaplanması eğer özgeçmişler standart bir formatta değilse bilgisayarlarca yapılamaz. Bunun yerine bu hesaplama ilgili dokümanların insanlar tarafından okunması ile gerçekleştirilir. Metinsel verilere bu nedenle sayısal verilerdeki "yapılandırılmış" ifadesinin aksine "yapılandırılmamış" veri adı verilmektedir. Yapılandırılmamış veri kategorisi tıpkı metinler gibi işlenmeden analitik amaçlarla kullanılması mümkün olmayan ses, video ve fotoğraf gibi verileri de kapsar. Dünyada bulunan ve bilgi ihtiva eden basılı veya dijital kayıtlarda veri tabanlarında tutulan yapılandırılmış verilerden çok yapılandırılmamış verilerin bulunduğu tahmin edilmektedir ve yapılandırılmamış verilerin tüm verilerin %80'ine tekabül ettiği ifade edilmektedir (Gupta & Lehal, 2009). Örneğin, Türkiye'de

kullanılan EBYS (Elektronik Belge Yönetim Sistemi ya da BelgeNet) içerisinde yazılan resmi yazıların ihtiva ettiği bilgiler de yapılandırılmamış veridir. Metin madenciliği yardımıyla EBYS'de en çok yazışma yapılan konular ortaya çıkarılabilir.

Yapılandırılmamış veri ile yapılandırılmış verinin işleyiş şekilleri birbirinden çok farklıdır. Metinlerin ilişkisel veri tabanlarında bir tabloya ait sütunlarda muhafaza edilmeleri onları ilişkisel hale sokmamaktadır. Örnek olarak bir bilgi dizisinin yapılandırılmış ve yapılandırılmamış biçimi Tablo 1 ve Tablo 2 içerisinde örnekle ifade edilmiştir.

Örnek olarak, bir kan bankasına ait veri tabanında donörlerin geçmişte sigara, alkol ve uyuşturucu kullanıp kullanmadıkları bilgisi bu durumda iki biçimde kaydedilebilir. Tablo 2 ilişkisel veri tabanlarında tutulan kayıt şeklidir. Her bir bilgi için bir sütun açılmakta ve bu bilgi varsa "1", yoksa "0" şeklinde kayıt yapılmaktadır. Eğer bu sütunlarda istenen bilgi temin edilememişse ilgili alan boş bırakılmakta ya da "NA (Not Applicable)" olarak ifade edilen kayıt girilmektedir. Tablo 1 biçimde ise veriler yine veri tabanında tutulmakla birlikte iki sütunda tutulmakta ve donörlere ait bilgiler düz metin formunda kaydedilmektedir. Düz metin formunda olan bu kayıtlar bu halleri nedeniyle "yapılandırılmamış" olarak kabul edilmektedir. Yapılandırılmamış veriler, yapılandırılmış forma dönüşümü için özel bir çaba gereken herhangi bir veri türü olabilir.

Tablo 1. Yapılandırılmamış Veri (Yazım hataları bilinçli olarak gösterilmiştir)

Donör	**Hikâye**
A	Donür uyuşturucu, alkol ve sigara kullanmıyor
B	Eroin ve sigara kullanmış
C	Esrar kullanıyor, cigarayı bırakmış

Tablo 2. Yapılandırılmış Veri

Donör	**Sigara**	**Alkol**	**Uyuşturucu**
A	1	0	0
B	1	1	NA
C	1	1	0

Sesler, fotoğraflar ve video kayıtları da tıpkı metinler gibi yapılandırılmamış veri sınıfına girmektedir. Tüm yapılandırılmamış veri türlerinde olduğu gibi metinsel verilerde de en büyük problem Tablo 2 içerisinde görüleceği üzere metinsel verilerin doğal dilin esnek kurallarına tabi olmaları nedeniyle sahip oldukları durumdur. Örneğin 2. tablodaki 1. satırda A donörü ile ilgili tüm bilgiler metinsel olarak yer almaktadır. Aynı tablodaki 2. satırda ise alkol kullanımı ile ilgili bilgi yer almamaktadır. Bu bilginin var olmadığı için mi yoksa donör gerçekten alkol kullanmadığı için mi mevcut olup olmadığı belli değildir. Diğer bir problem ise, örnekte görüldüğü üzere uyuşturucu kelimesi yerine "eroin" kelimesi kullanılmasıdır. Bu durumda geleneksel veri tabanlarında SQL sorgularındaki LIKE komutu ile benzer ifadeler aranmak istediğinde veri analisti uyuşturucu ile ilgili tüm olası kelimeleri tahmin etmek zorunda kalacaktır. Tablo 2'de 3. kayıtta da benzer bir durum mevcuttur, bu kez "esrar" kelimesi kullanılmıştır. Yine, sigara kelimesi ise yazım hatası ile "cigara" yazılmıştır. Bu durumda da yazım hatalarını algılayabilecek standart SQL komutları bulunmamaktadır. Bu durum yapılandırılmamış verilerin standart zorluklarından birini teşkil etmektedir. Diğer yandan Tablo 2'de ise her donöre ait her kategorik konu spesifik olarak tanımlanmıştır. Yazım hataları, farklı kelimelerle tanımlamalar gibi durumlar söz konusu değildir. Buna göre Tablo 1 yapılandırılmamış veriyi Tablo 2 ise yapılandırılmış veriye örnektir. Ancak doğası gereği, dünya üzerinde benzer birçok kayıt metinsel formda bulunmak durumundadır.

Söz konusu yapılandırılmamış veri tiplerinden biri olan metinsel veri tiplerini analiz etmek için kullanılan metin madenciliği ile ilgili Türkçe literatürde karşılaşılan çalışmaların (Dolgun, Özdemir, & Oğuz, 2009; İlhan, Duru, Karagöz, & Sağır, 2008; Kaşıkçı & Gökçen, 2013; Kılınç vd., 2016) tamamı metin madenciliği yöntemlerinin çeşitli alanlarda uygulamaları ile ilgili olup, metin madenciliğinin yöntemleri, kullanım alanları ve özellikleri hakkında Türkçe çalışma ile karşılaşılmamıştır. Bu çalışma bu boşluğu doldurmayı amaçlamaktadır. Bu kapsamda, literatür bölümünde, metin madenciliği ile ilgili tanımlar irdelendikten sonra literatürde yapılmış çalışmalar özetlenmiştir. Daha sonra yabancı dilde yazılmış literatürdeki çok sayıda çalışma üzerinden metin madenciliğinin kullanım alanları imkânlar başlığında sınıflandırılarak ele alınmıştır. Kullanılan araçlar bölümünde araştırmacıların metin madenciliği çalışmaları yaparken kullanabilecekleri ücretsiz ve açık kaynak kodlu araçlar tanıtılmış, bu araçlar kullanılarak yapılacak analizlerdeki adımlar analiz süreci bölümünde ayrıntıları ile ele alınmıştır. Son olarak metin madenciliği alanındaki kısıtlar incelenmiştir.

2. LİTERATÜR

Metin Madenciliği ile ilgili muhtelif tanımlar bulunmaktadır: Bunlardan birinde metin madenciliği, yapılandırılmamış metinden ilgi çekici bilgi veya iç görü elde etme süreci olarak değerlendirilmektedir (Chen, 2011, s. 5). Başka bir tanımda, elde edilecek bilginin kalitesine atıfla, metin madenciliği genellikle metinlerden yüksek kaliteli bilgi elde etme süreci olarak tanımlanmıştır (Mucherino, Papajorgji, & Pardalos, 2009, s. 17). Metin madenciliği yardımı ile salt metin içinde geçen konular arasındaki ilişkileri keşfetmeyi de metin madenciliği olarak kabul edenler de bulunmaktadır (Babu vd. 2014, s. 12). Hansen ve Johnson ise metin madenciliğini, metin içindeki gizli kalıp ve örüntüleri ortaya çıkarma süreci olarak tanımlanmıştır (2011, s. 10). Miner ve diğerleri metin madenciliğini insanların iletişimini analiz etme süreci olarak yorumlamıştır (2012, s. 1009). Gerçekten de metin madenciliğinde kullanılan materyaller, veri tabanlarında biriktirilen kayıtların aksine insanların yazışması, sosyal medya girdileri veya konuşmalarının metinleştirilmesi sonucu oluşmaktadır. Bu değerlendirmeler ışığında metin madenciliği şöyle tanımlanabilir: Metin madenciliği, veri kaynağı olarak sadece insanlar tarafından ve doğal dilde serbest biçimde yazılmış metinleri ele alan ve bu metinlerden daha önce bilinmeyen önemli çıkarımların yapılabilmesine olanak veren metotlar ve araçlar bütünüdür. Tüm bu tanımlar metin madenciliğinin bilgisayarlar yardımı ile yapıldığı varsayımına dayalıdır nitekim metin madenciliğindeki metotların bilgisayar yardımı olmaksızın uygulanması mümkün değildir. Milyonlarca dokümanda geçen ve geçmeyen kavramların sayılması, tasnifi ve sıralanması ancak bilgisayarlar yardımı ile mümkündür. Diğer yandan metinlerin manuel olarak tasnifi ve analizi ise metin madenciliği olarak değil "içerik analizi" olarak tanımlanabilir. Her halükarda iki yöntem de bütünden anlam çıkarılması amacına hizmet etmektedir.

Metin madenciliği veri madenciliğinin alt dalı olmakla birlikte kullandığı araçlar ve girdiler oldukça farklıdır. Veri madenciliğinde girdiler genellikle veri tabanları veya dosyalardaki tablo şeklindeki veriler iken metin madenciliğinin girdisini tablo formunda olmayan dosyalar, HTML web sayfaları veya PDF ya da Word dokümanları oluşturmaktadır. Metin madenciliğinin herhangi bir veri tabanında yapılan standart bir arama davranışından farkı şudur: Sözgelimi Ctrl+F veya SQL komutları ile yapılacak normal bir arama işleminde, aranacak olan kavram önceden belirlidir. Bu durumda kavramın kaynak içinde mevcudiyeti sorgulanmaktadır, metin madenciliğinde bir arama durumundan ziyade metnin içeriğinde veya genelinde önceden bilinmeyen birtakım öğelerin ortaya çıkarılması sağlanmaktadır. Başka bir deyimle metin madenciliği, bir metin içerisinde analiz öncesinde bilinmeyen kavramların bulunması sürecidir.

Metin madenciliği disiplinler arası bir alan olarak kabul edilmektedir (Gupta & Lehal, 2009, s. 60). Metin madenciliği ile ilgili Türkçe dışındaki literatürde birçok teorik ve uygulamalı çalışma mevcuttur. Türkçe dışındaki literatürde yayınlanmış çok fazla sayıdaki çalışmanın genel olarak kullandığı metotlar bu makaledeki kullanım alanları bölümünde ele alınmıştır. Türkçe literatürde de metinsel verilere dayalı çeşitli analizler gerçekleştirilmiştir. Türkçe literatürde metin madenciliği ile ilgili karşılaşılan az sayıdaki çalışmalar arasında ilk çalışmalardan biri İlhan ve diğerleri tarafından gerçekleştirilmiş ve bu çalışmada metin

madenciliği yöntemleri kullanılarak soru cevaplama sistemi kurulmuştur (2008). Dolgun ve diğerleri ise karar ağacı algoritmaları ile metin madenciliğine dayalı yöntemleri karşılaştırmışlardır (2009). Karadağ ve Takçı ise benzer haberleri otomatik olarak sınıflandırabilen bir algoritma geliştirmişlerdir (2010). Kılınç ve diğerleri ise bilimsel makaleleri tasnif için metin madenciliği ve makine öğrenmesi destekli bir sistem geliştirmişlerdir (2016). Çalış ve diğerleri metin madenciliği yöntemleri ve makine öğrenme algoritmaları olan kNN (k-Nearest Neighborhood) ve NaiveBayes algoritmalarını kullanarak e-postaların içeriğinin reklam olup olmadıklarını %96.5 doğrulukta tespit edebilen bir uygulama geliştirmişlerdir (2013). Bu algoritmalar metin madenciliği de dâhil birçok alanda kullanılmaktadır. Kaşıkçı ve Gökçen de kNN ve Naive Bayes algoritmalarını kullanarak bir web sitesinin içeriğine bakarak bu sitenin e-ticaret sitesi olup olmadığını anlayabilen Java tabanlı bir uygulama geliştirmiştir (2013).

Metin madenciliği sadece akademik çalışmalar için değil kurum, kuruluş, şirket gibi her türden organizasyonlar için de pratik bazı çözümler ortaya koymaktadır. Metin madenciliği, yayıncılık ve medya, telekomünikasyon, enerji, bilgi teknolojileri, bankalar, sigorta şirketleri, finansal şirketler, siyasi kurumlar, hukuk, sağlık ve eczacılık gibi alanlarda etkin bir şekilde kullanılabilmektedir (Bolasco, Canzonetti, Capo, Ratta-Rinaldi, & Singh, 2005, s. 323). Somut bir örnek olarak insan kaynakları yönetiminde adayların özgeçmişlerinin analizinde de metin madenciliği kullanılabilmektedir çünkü özgeçmişler metin formundadır. Bu özgeçmişler içinden metin madenciliği kullanılarak örneğin en sık karşılaşılan yeteneklerin çıkarılması gibi analizler yapılabilmektedir (Gupta & Lehal, 2009). Aynı şekilde metin formunda olan ve herhangi bir ürün veya hizmet hakkında yazılan her türlü haber, sosyal medya mesajı veya raporlar üzerinde de metin madenciliği analizi yapılarak müşterilerin beklentileri ile şirketlerin imajı takip edilebilmektedir (Atan, 2016a). Otomotiv sektöründe araç tamir servisleri sürekli olarak metinsel formda arıza raporları hazırlamaktadırlar. Bu raporlarda geçen yedek parça adları ve arıza türleri anahtar terimleri (örneğin yağ sızıntısı) ile bu türlerin mevsimsel olarak hangi dönemlerde ortaya çıktığı bilgisi kullanılarak arızalar ve mevsimselliği üzerine önemli içgörüler elde edilmek suretiyle maliyet avantajının yaratılması mümkündür (Derick, 2010). Metin madenciliği bilişim sistemlerinin arka planında da önemli fonksiyonlar icra etmektedir. Arama motorları, spam ve reklam epostalarını filtreleme, ürün önerme algoritmaları, sahtekârlık tespiti ve sosyal medya analiz uygulamaları bu alanlara örnek teşkil etmektedir (Talib, Kashif, Ayesha, & Fatima, 2016). Metin madenciliği sohbet programlarında siber suç unsurlarının araştırılması gibi amaçlarla da kullanılmaktadır (Kaushik, 2013). Finans alanında ise algoritmik yatırım araçları metin madenciliğini kullanarak borsada şirketler hakkında çıkan haberlere göre eş zamanlı analizler yapma olanağı sunmaktadır (Atan, 2016a).

3. İMKÂNLAR

Bu bölümde metin madenciliği kullanılarak gerçekleştirilebilecek araştırma imkânları ele alınmıştır. Metin madenciliği araçları tek başına birçok sonucun üretilmesine yardımcı olabildiği gibi yapay zekâ veya başka bilişim sistemi yaklaşımları ile farklı sonuçları elde etmeye yönelik imkânlar sağlamaktadır.

3.1. Bilgi Ekstrasyonu

Metin madenciliğindeki en temel imkânlardan biri bilgi ekstrasyonudur (Gupta & Lehal, 2009, s. 61). Yapılandırılmamış veri içerisinde geçen özel adlar, yer adları, kişi adları, zaman, olay gibi özel ifadeleri kalıp eşleştirme ve anahtar terim süzme vb. yöntemlerle ayıklama bu alanın konusudur. Bilgi ekstrasyonu ile metin halinde verilen örneğin haber metinleri içinde geçen kişi adları ayıklanabilmekte, bir haber sitesinde en çok ele alınan politikacıların, yer isimlerinin hatta metinlerde en sık kullanılan fiillerin ortaya çıkarılması mümkün olabilmektedir. Bilgi ekstrasyonu metin yığınlarından sonuç üretmekten ziyade belirli konuların ayrıştırılması sürecidir ve ayrıca metinlerin teker teker tasnifi yerine bu metinlerin ilgili oldukları konu ve kategorileri ortaya çıkarmaya yardımcı olmaktadır. Bu işlev sadece belirli kişiler ve organizasyonlarla ilgili haberleri takip eden araştırmacıların haber yığınları içinde standart olarak kelime arayarak incelemeleri yerine birebir ilgili haberleri bulmalarına yardımcı olmaktadır. Bir algoritmanın herhangi bir metin içerisinden yer adları, özel işimler veya tarih, olay ve etkinlik gibi bilgileri çıkarması yığın halindeki metinler

tasavvur edildiğinde çok önemli faydalar sağlayabilmektedir. Örneğin metinler içerisinden etkinliklerin ad, yer ve zamanlarını ayrıştırabilme algoritmaları yardımı ile yayınlanan haberleri sürekli tarayarak etkinlikleri süzen modeller geliştirilmesi mümkündür. Normal şartlarda bu tür bilgiler elle de aranabilir ancak bu arama davranışında nelerin aranacağı önceden belirlidir. Belirlenmemiş bir terim o metinler içerisinde aranmayacaktır. Ancak bilgi ekstrasyonu algoritmaları yardımı ile bu durumda önceden belirlenmiş bir arama terimlerine ihtiyaç bulunmamaktadır.

Aşağıdaki örnekte Python kodları yardımı ile bir haber sitesinden derlenen bir haber bilgi ekstrasyonu amacı ile kullanılan kütüphanelerden olan Spacy Kütüphanesi (http://www.spacy.io) yardımı ile analiz edilerek içerisindeki özel adlar (kişi, yer vb.) ortaya çıkarılmıştır. Elde edilen sonuçlar ilgili kütüphane tarafından etiketlenmiştir: Bu kütüphane bilgi ekstrasyonu amacıyla her türlü analiz için ücretsiz olarak kullanılabilmektedir. Bu kütüphane birçok dili desteklemektedir. Ancak henüz Türkçe desteklenen diller arasında değildir. Bu kütüphane metin madenciliği ile yapay zekâ çalışmaları için de metinler içinde hızlı şekilde özel ifadelerin aranmasına yardım ederek bilgi ekstrasyonu için araştırmacılara çeşitli olanaklar sunmaktadır.

Tablo 3. Spacy Kütüphanesi Örneğinden Bilgi Ekstrasyonu

```
# Bu kodlara şu çalışma alanında erişim sağlanabilir ve test edilebilir: http://j.mp/spacy-ornek
import spacy
nlp = spacy.load("en core web sm")
haber_metni = """
Turkey has not promised the United States to not install or use the Russian S
400 systems, the Foreign Minister Mevlüt Çavuşoğlu said on Nov. 26.
“We have no commitment to anyone that we will not install or use the S-
400. We purchased it because we needed an air defense system. Is an air defense system taken to keep
 it in the box?” Çavuşoğlu told reporters at the parliament.
"""
doc = nlp(haber metni)
print("Ad öbekleri:", [chunk.text for chunk in doc.noun_chunks])
print("Fiiler:", [token.lemma_ for token in doc if token.pos_ == "VERB"])
for entity in doc.ents:
    print(entity.text, entity.label_)
```

Bu kodlar çalıştırıldığında içerikteki özel adlar aşağıdaki gibi ortaya çıkmaktadır[1]:

```
Ad öbekleri: ['\nTurkey', 'the United States', 'the Russian S-400 systems','Çavuşoğlu']
Fiiler: ['have', 'promise', 'install', 'use', 'say', 'have', 'will', 'install', 'use', 'purchase',
'need', 'be', 'take', 'keep', 'tell']
Turkey GPE (Jeopolitik Varlık)
the United States GPE
Russian NORP
Mevlüt Çavuşoğlu PERSON
Nov. 26 DATE
```

Görüldüğü üzere Spacy Kütüphanesi yardımı ile metin içindeki fiiller, insan isimler ve tarihleri otomatik olarak ortaya çıkarılmıştır. Bu özelliğin elle yapılması tüm metinlerin baştan sona okunarak elle

[1] Bu kodlara şu çalışma alanında erişim sağlanabilir ve test edilebilir: http://j.mp/spacy-ornek

işaretlenmesi gibi uzun süren bir işlemi gerektirirken, algoritma bu işlemi milisaniyeler içerisinde gerçekleştirmektedir. Öte yandan dilsel yapı içerisinde aslında Türkçe olan "Çavuşoğlu" ifadesi de algoritma tarafından algılanabilmekte ve işaretlenmektedir.

3.2. Konu Tanıma ve Sınıflandırma

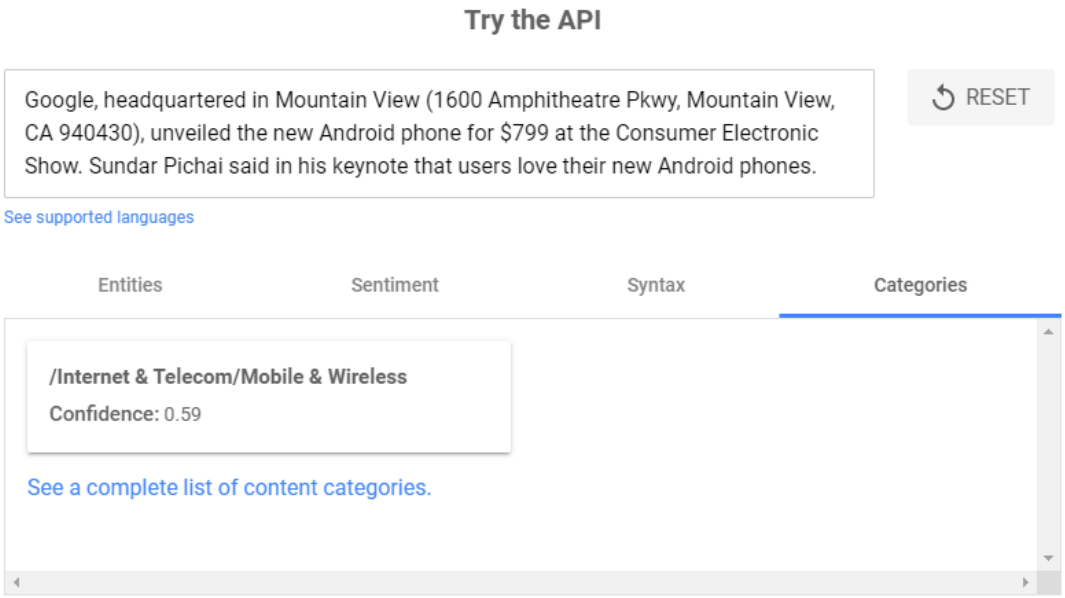
Çeşitli algoritmalar yardımı ile mevcut metinlerin hangi konu hakkında olduğuna dair çıkarım yapılması işlevi de metin madenciliği yöntemleri tarafından sağlanabilmektedir. Konu modelleme algoritmaları makine öğrenmesi yardımıyla herhangi bir etiketlenmiş veri seti üzerinden etiketleri bilinmeyen haberlerin otomatik olarak etiketlenmesini sağlayabilir. Konu tanıma algoritmaları daha önce konuları belirli olan kaynakların makineye öğretilerek makinenin konuları belirlenmemiş metinlerin konularını tahmin etmesine olanak verir. Bir haber sitesinde daha önce ekonomi, sağlık, politika gibi başlıklarda yazılmış haberlerin etiketleri mevcut olduğundan bu haber metinleri makine öğrenme algoritmasına öğretilerek algoritmanın herhangi bir başka haber metnini alıp hangi kategoriye gireceğini tahmin etmesi istenebilir. Bu durumda kullanılan algoritma 'gözlemli' algoritma (supervised algorithm) olup dışarıdan eğitim sürecini içermektedir. Diğer taraftan 'gözlemsiz' (unsupervised) algoritmalar yardımı ile bir metin seti içinde olası konu ayrımları saptanabilir (Silge & Robinson, 2017). Bu durumda saptanan konu başlıkları yerine metin setindeki sınıflar sayısal olarak numaralandırılır. Daha sonra bu numaraların hangi konular olabileceğini geliştirici eşleşme yardımı ile bulmaktadır. Haber kategorisi örneğinde yeterli sayıda sağlık, politika ve ekonomi haberi mevcut olduğunda algoritma 1,2 ve 3 şeklinde kategorileri isimsiz olarak tespit ederek her haberin hangi kategoride olduğunu tespit edebilmektedir. LDA (Latent Drillecht Algorithm) en bilinen gözlemsiz metin sınıflandırma algoritmalarından biridir.

Konu sınıflandırma algoritmalarından Google Natural Language API adlı hizmet bir kütüphane olarak sunulmayıp API (Application Programming Interface) olarak sağlanmaktadır. Bu hizmet aynın zamanda bilgi ekstrasyonu ve sentiment analizi desteği de vermektedir. Şekil 1'de girilen bir metnin kategorisi ilgili hizmet tarafından %59 eminlikte "İnternet ve Telekom" konusu olarak tespit edilmiştir.

3.3. Metinler Arası Benzerlik Ve İntihal Tespiti

İki farklı metinin tamamen aynı olup olmadığını veya birinin diğerinden türetilmiş olup olmadığının tespiti de metin madenciliğinin ilgi alanına girmektedir. İki metnin benzerliğinin tespitinde birebir kullanılan kelimelerin mukayesesi ile gerçekleştirilebildiği gibi (Baker, 1995), metinlerin stilleri ve atıf benzerlikleri gibi ölçütlere göre de metinler arası benzerlik incelenebilmektedir. Metinler arası benzerlik algoritmaları akademik intihal tespiti yazılımlarında kullanılmaktadır. Bu programlar girilen bir metnin daha önce taranmış devasa bir metin setindeki herhangi bir metnin herhangi bir parçası ile benzeşimini tespit ederek tahlil edilecek metinde intihal edilme olasılığı bulunabilecek her parçayı algılayabilen başarılı algoritmalardır. iThenticate gibi ticari uygulamalar yanında online olarak metinler arası benzerlik tespiti yapan servisler de bulunmaktadır (Şekil 2). Bu servislerin tamamı metinler arası benzerlik algoritmalarını kullanmaktadır. Metinler arası benzerlik algoritmaları yalnızca intihal tespitinde değil, girilen bir metnin yanlış yazılması halinde dâhil esas kastedilenin ne olabileceğinin tahmini için de kullanılmaktadır. Arama motorlarına girilen yanlış yazılmış bir ifadenin arama motoru tarafından doğru anlaşılması bu algoritmalar yardımı ile gerçekleştirilmektedir (Şekil 3).

Şekil 1. Google Natural Language API Demo



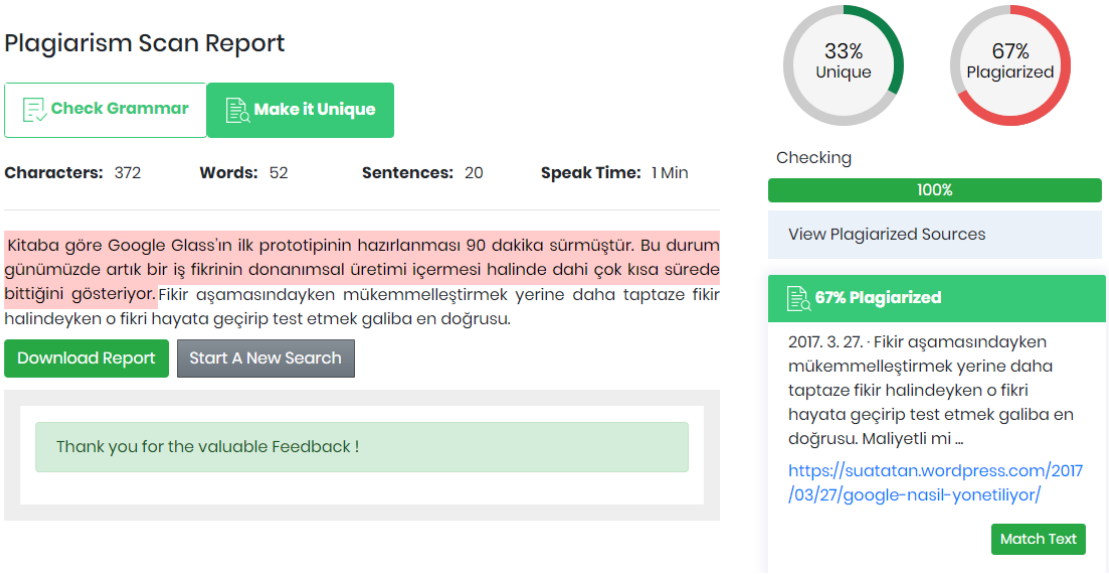
3.4. Konu Özetleme

Uzun yazılmış metinlerin özetlenmesi de metin madenciliğinin inceleme alanlarından biridir. Bu işlev yardımıyla çok uzun bir metni özetleme algoritmaları yardımı ile bir veya bir kaç paragrafa indirerek konu hakkında fikir edinmek mümkündür (Shaalan, Hassanien, & Tolba, 2017, s. 436). Konu özetleme algoritmaları çeşitli kelimelerin ağırlıklarını ve metin içindeki sıklıklarını kullanarak önemli olduğunu tahmin ettiği cümleleri bir araya getirmektedir. Bu algoritmaların kullandığı diğer bir yöntem ise cümlenin metin içinde geçtiği yere göre tahmin yürütmektir. Örneğin metnin son kısımlarının metnin tamamı hakkında özet bilgi içerme ihtimali daha fazla olmaktadır. Ya da "özellikle", "kesinlikle", "önemli bir nokta da" gibi ifadelerden sonra gelen cümleler metnin geri kalanına göre daha kritik bilgiler içerebildiğinden algoritmalar bu pozisyonlarda bulunan ifadeleri de ayrıştırarak özet içerisine koyabilmektedir. Özetleme işlevi de yine çok fazla metnin bizzat insanlar tarafından okunması gerektiğinde hepsinin okunması yerine özetlerinin okunarak, odaklanılacak esas metnin bulunmasına olanak sağlamaktadır. Konu özetleme ile ilgili olarak R dili ile birlikte kullanılan TextRank adlı kütüphane (TextRank Algorithm, t.y.) bulunmaktadır. Bu kütüphane ne kadar fazla olursa olsun bir metni özetleyebilmekte ayrıca en önemli bölümlerini ortaya çıkarabilmektedir. Ayrıca online çalışan örnekler de bulunmaktadır (ReSoomer Text Summarization Tool, t.y.) . Online çalışan bu uygulamalar bilgisayara kurulum yapılmasını gerektirmemektedir. Bilgisayarlara kurulum yapılmaksızın analiz özellikle sınırlı işlemci kaynaklarına sahip bilgisayarlarda analizlerin daha hızlı yapılmasına olanak sağlayabilir.

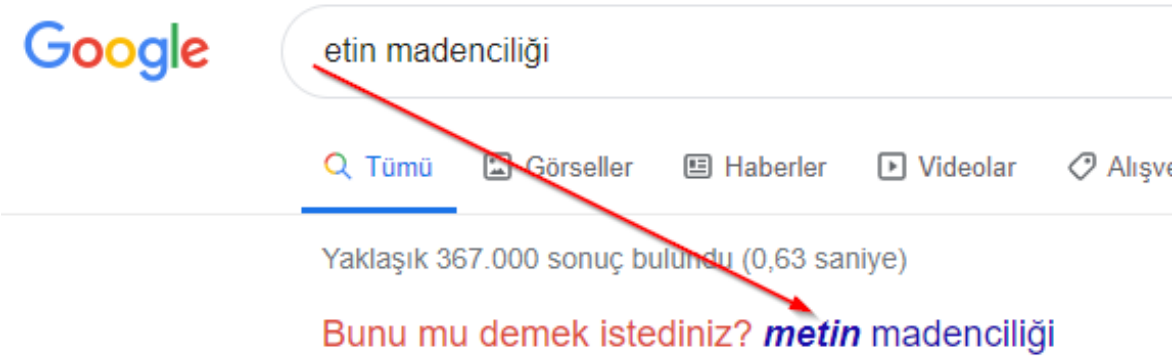
3.5. Kavram Bağlanımı Tespiti

Bir metin setinde farklı metinler içinde sıkça ifade edilen kavramların tespiti kavram bağlanımı (concept linkage) olarak ele alınmaktadır (Gupta & Lehal, 2009, s. 64). Kavram bağlanımı birden fazla doküman arasında dışarıdan fark edilebilen mevcut kavramsal ilişkiler yanında dışarıdan fark edilmeyen olası ilişkileri de tespit edebilmektedir. Örneğin biyomedikal alanında bir hastalık üzerine yazılmış metinlerle, başka bir kaynakta bu hastalığın tedavisi ile ilgili metinlerin arasındaki olası ilişkileri kavram bağlanımı ile tespit etmek mümkündür (Patel & Soni, 2012).

Şekil 2. https://plagiarismdetector.net/ adlı web sayfasında bir analiz örneği



Şekil 1. Google tarafından doğru anlaşılan metin ("Did you mean" özelliği). Arama terimi yanlış yazıldığı halde algoritma doğru ifadeyi tahmin etmektedir



3.6. Sentiment Analizi

Duygu analizi (Sentiment Analysis), metinlerin yansıttığı görüşleri ve duyguları tanımlamaya ve analiz etmeye yarayan bir metottur (Chen, 2011). Duygu analizi metotları metin setindeki her bir metnin ayrı ayrı okumaksızın bilgisayarlar tarafından değerlendirilmek suretiyle analiz edilmesi yoluyla gerçekleşir. Duygu analizi metodu genellikle duygu sözlüklerine dayalı analiz üzerinden gerçekleştirilir. Sözlüklere dayalı analiz önceden tanımlanmış ve her duygu için ayrı ayrı kelimeler barındıran duygu sözlüklerinde geçen kelimelerle her bir metindeki kelimelerden eşleşenlerin sayılarak en sık karşılaşılan duygu ifadesinin ortaya çıkarılmasına dayalıdır. Örneğin bir metin içinde "iflas", "ekonomi", "zarar" ve "konkordato" gibi ifadeler geçiyor olsun, olumsuz ifadeler sözlüğünde geçen kelimeler ise "iflas" ve "zarar" kelimeleri ise bu metinin olumsuzluk bakımından duygu skoru 2 olarak kabul edilecektir. Metinler kelimelerden teşkil edildiğinden bir metindeki ifadeler ne kadar yansız olursa olsun (olumluluk veya olumsuzluk barındırmayan) metnin herhangi bir yerinde olumsuzluk söz konusu ise bu kelimelere yansımakta ve analiz algoritması tarafından algılanabilmektedir (Atan, 2016a).

4. KULLANILAN ARAÇLAR

Metin madenciliğinin de gerçekleştirilebildiği ücretli ve ücretsiz çeşitli veri madenciliği araçları ve programlama dilleri mevcuttur. Aşağıda bu araçlardan açık kaynak kodlu ve ücretsiz olanlarından en yaygınları tanıtılmaktadır. Bu araçlar ücretli alternatifleri kadar yaygın ve güçlü özelliklere sahiptirler.

4.1. R/Studio

R, çeşitli analitik problemleri çözmek için hesaplama istatistikleri ve veri bilimi topluluğu tarafından yaygın olarak kullanılan açık kaynaklı bir istatistik programlama dilidir (Fontama, Barga, & Tok, 2015, s. 81). R dili ile SPSS ve Minitab gibi programlarla gerçekleştirilebilen her türlü analiz gerçekleştirilebildiği gibi ilgili kütüphaneler kullanılarak makine öğrenmesi, veri madenciliği ve metin madenciliği için de kullanılmaktadır. R dili "ggplot" ve seaborn gibi başarılı görselleştirme kütüphaneleri yardımıyla raporlama imkânı sunmaktadır. R dilinin kişisel bilgisayarlarda kullanılabilmesi için RStudio adı verilen bir IDE (Entegre geliştirme ortamı) bulunmaktadır. RStudio R dili ile birlikte kullanılabilecek RMarkdown adlı işaretleme dili desteği de sunarak verilerle sürekli etkileşim halinde kalabilen bir kelime işlemci ortamı sunar. Böylece herhangi bir akademik veya veriye dayalı başka bir raporun yazıldığı ortam ile analizlerin sonuçlarının görselleştirildiği ortam aynı dokümanda yer almaktadır. RStudio Shiny adı verilen modülü ile analiz sonuçlarının web ortamında da kolayca yayınlanabilmesine olanak vermektedir. R dili yalnızca fen bilimleri değil sosyal bilimler alanında da en yaygın kullanımı olan istatistiksel programlama dilidir (Atan & Emekci, 2018).

4.2. Python

Python da tıpkı R gibi bir programlama dilidir. Python dili de veri madenciliği ve metin madenciliği desteğinin yanında, makine öğrenme algoritmalarının da hazırlanabildiği açık kaynaklı bir dildir. Python dilinin R diline göre güçlü yanı ise şudur: R dili sadece veri bilim amacıyla kullanılmakta olup R dili ile Shiny modülü dışında gerçek anlamda web sayfaları, masaüstü uygulamaları veya mobil uygulamalar geliştirilmemektedir. Python ise veri bilim amacıyla kullanılabilmesinin yanında web, masaüstü ve mobil ortamlarda program geliştirilebilmesine olanak vermektedir. Bu durum Python'un veri bilimi ile ilgili hazırlanacak modellerin ürüne dönüştürülebilmesi noktasında avantaj sağlamaktadır. Python için çok çeşitli entegre geliştirme ortamları (IDE) mevcut olmakla birlikte veri bilim amacıyla kullanılan IDE Jupyter adlı web tabanlı IDE olup Rstudio'ya göre daha az özelliğe sahiptir. Rstudio benzeri bir uygulama olarak Yhat adlı bir IDE de mevcuttur. Ayrıca Google Colab ve Microsoft Azure Notebook online araçları, Python ile veri analizi çalışmalarının bulut üzerinden gerçekleştirilmesine olanak vermektedir.

4.3. Weka

Weka, Yeni Zellanda'da bulunan University of Waikato tarafından geliştirilmiş ücretsiz ve açık kaynaklı bir veri madenciliği aracı olup, metin madenciliği için de modüller içermektedir. Weka bir programlama dili değil programlamaya en az ihtiyaç bırakacak şekilde sürükle ve bırak olarak adlandırılan görsel araç seti ile analiz yapılabilmesine olanak veren bir araçtır.

4.4. Orange

Orange de tıpkı Weka gibi görsel olarak veri madenciliği çalışmaları için geliştirilmiş ve Slovenya'da bulunan University of Ljubljana menşeili bir uygulamadır. Orange de metin madenciliğinin yapılabilmesine olanak vermektedir. Orange içerisinden görsel programlama yanında Python ile kod yazma olanağı da bulunmaktadır.

4.5. Google Sentiment API

Google Sentiment API, duygu analizi için geliştirilmiş temel kullanım düzeyinde ücretsiz ancak belirli bir seviyedeki kullanımdan sonra ücretli olan bir API'dir (Application Programming Interface: Uygulama Programlama Arayüzü). Bu yönü ile arayüzü olan bir uygulama olmayıp daha ziyade uygulama geliştiricilerin yazdıkları programlar içerisinden çağırabildikleri servislerdir. Normalde duygu (sentiment) analizi için herhangi bir yazılım dilinde (Python veya R) gibi programlama yapmak, bu programlara duygu sözlükleri ile işlenecek metinleri tanıtmak gerekmektedir. Google Sentiment API ise bu işlemlere gerek bırakmaksızın programcıların hazırladıkları uygulamalarda verilen bir metnin olumluluk veya olumsuzluğunun ölçülebilmesine olanak verir. Ayrıca bu API'nin uyum olanakları sayesinde geliştirilecek

herhangi bir mobil, web veya masaüstü uygulama içerisinden direkt olarak çağrılabilme imkânı bulunmaktadır.

5. ANALİZ SÜRECİ

Metin analizi çalışması, metinlerin web sayfaları gibi kaynaklardan derlenmesi (web scraping) daha sonra bu metinlerin ön işlenmesi ile başlar. R dili ile çalışan RCrawler ve Rvest gibi kütüphaneler bu işlevi başarı ile yerine gerçekleştirmektedir. Çeşitli analiz araçları yardımıyla metinler üzerinde istenen çıkarımları yapmaya yarayacak işlemler gerçekleştirilir ve son olarak elde edilen bulgular raporlanır. Bu noktada metin derleme metin madenciliği için verinin tedarik edildiği bir araç olarak özel bir inceleme sahasıdır. Bu nedenle metin madenciliği süreçlerinden biri olan ön işlem (pre-processing) süreci ve müteakip süreçler ifade edilecektir.

5.1. Ön İşlem

Metin madenciliği analizi yaparak gerçekçi sonuçlar elde edebilmek için analiz öncesinde bir takım veri temizliği ve ön işlemler yapılmalıdır (Atan, 2016a). Bu işlemlerin sonucunda elde edilen bazı Terim Doküman Matrisi (TDM) ve Doküman Terim Matrisi (DTM) gibi ara sonuçlar da vardır ancak genellikle bu matrisler devasa boyutlarda olduğundan analizde bilgisayar hafızasında tutulmakta ancak tamamı raporlanmamaktadır. Bunun yerine bu matrisler küçültülerek çeşitli başka analizler gerçekleştirilmektedir. Ön işlem aşamasında gerçekleştirilen işlemler aşağıdaki gibidir:

5.2. Anlamsız Sözcüklerin Temizliği

Doğal dilde cümlenin bütünlüğü içerisinde her sözcüğün bir anlamı veya işlevi vardır (O'Keeffe & McCarthy, 2010). Ancak metin madenciliğinde işlevi olan ancak tek başına anlamı olmayan edatlar, bağlaçlar, ünlemler ve benzeri sözcükler (stopwords) temizlenmelidir. Genellikle metin madenciliği araç ve yazılım kütüphaneleri çeşitli diller için temizlenecek kelime listelerini hazır olarak vermektedir.

5.3. İlgisiz kelimelerin Temizliği

Anlamları mevcut olmakla birlikte analize katkısı olmayan özel olarak belirlenmiş kelimelerin de temizlenmesi gerekebilir. Örneğin ekonomi haberleri üzerinde gerçekleştirilen bir analizde en fazla ekonomi kelimesi geçecektir ve bu durumun özel bir anlamı yoktur. Bu nedenle bu tür kelimeler analizlerden ihtiyaca bağlı olarak temizlenmektedir.

5.4. Sayıların Temizliği

Metin analizlerinde haberlerde geçen tarihler, yıllar, parasal miktarlar ve oranları barındıran birçok ifade ile karşılaşmak mümkündür. Ancak bu sayıların her birinin tam olarak neye tekabül ettiği bilinmediğinden sayılar da özel başka bir amaç olmadıkça temizlenmemelidir.

5.5. Noktalama İşaretleri ve İlgisiz Karakterlerin Temizliği

Noktalama işaretleri ve ilgisiz karakterler (emojileri teşkil eden karakterler, fazladan boşluklar vs.) her ne kadar metin analizine zarar vermeseler de analiz yapılırken sistem kaynaklarını boşa tüketebilirler. Bu nedenle bu ifadeler de temizlenmelidir.

5.6. Kelime Köklerine İnme

Bir sözcük, doğal dilde aynı anlama da gelse bazen dilbilgisi kurallarından ötürü çeşitli formlarda görülebilmektedir. Örneğin Türkçe için "başarmak", "başardı", "başarıyor," başarmış" gibi tüm kelimeler "başarmak" fiilinden türemiştir. İngilizce için ise gitmek fiili zamanına göre "go, went ve gone" gibi bir sözcüğün tümden değişebildiği formlar alabilmektedir. Bu durum metin analizindeki bilinen problemlerden biridir (Feldman & Sanger, 2006) nitekim aynı anlama gelen sözcüklerin farklıymış gibi sayılarak analiz edilmesi analiz sonuçlarında karışıklığa yol açarak doğru olmayan çıkarımlara neden olabilir. Söz gelimi,

"başarı" kelimesinden türeyen farklı biçimlerdeki kelimeler sıklık listelerinde görülmeyebilir ancak bunların farklı formlarının ayrı ayrı toplanıp bir araya getirildiği durumda en sık kelime "başarmak" fiili olabilir. Bu durumla mücadele etmek için metin setindeki her sözcüğün teker teker kelime köklerine inilmektedir. Bu işlemin elle yapılması olanaksızdır bu nedenle kelime köklerine inme (stemming) adı verilen algoritmalar kullanılmaktadır. Bu algoritmalara kendisine girdi olarak verilen metinlerin her bir kelimesinin kelime köküne inerek çıktı verir. Ancak kelime köklerine inme algoritmaları her dil için ayrı olmaktadır. Bu alanda, R dilinde SnowballC adlı kütüphane Türkçe'nin de içinde yer aldığı bilinen algoritmalardandır.

5.7. Raporlama ve Görselleştirme

Metin madenciliği yardımı ile varılan sonuçları daha anlaşılır kılmak için birtakım raporlama ve görselleştirme yöntemleri bulunmaktadır. Bu yöntemler analiz edilen metnin hacmi ne olursa olsun bu metinlere tek bakışta kanaat ortaya çıkarılabilmesine yardımcı olmaktadır. Kullanılan raporlama ve görselleştirme araçları aşağıdaki gibidir:

5.7.1. Kelime Bulutu

Kelime bulutu (wordcloud) metin seti içerisinde en sık geçen kelimelerin bir alanda rastgele dağıtılması suretiyle metin seti hakkında fikir edinmeye yardımcı olan bir diyagramdır. Kelime bulutundaki her bir kelimenin ayrıca sıklığına göre daha büyük punto ile yazılması suretiyle kelimelerin nispeten hangisinin daha sık görüldüğü de anlaşılmaktadır. Kelime bulutuna en sık geçen kelime listesindeki ilk kaç kelimenin gireceğini araştırmacı belirler. Kelime bulutunda doğal olarak bağlaç ve edatlar ya da ilgili dildeki diğer tanımlayıcı ifadeler (ve veya gibi, için benzeri terimler) sıkça yer alacaktır. Bu kelimeler kelime bulutu oluşturulmadan önce ayıklanmalıdır. Doğal dilde bazen bir kavram tek kelime ile ifade edilmez. Bu durum dilden dile değişkenlik gösterir, bir dilde iki kelime ile ifade edilen bir olgunun başka bir dilde tek bir kelime ile ifade edilmesi de olasıdır. Bu durum kelime bulutlarına da yansımaktadır. Ayrıca kelime bulutu ve diğer sıklık gösterge araçlarında kelimelerin metin seti içinde homojen olmamaları da olasıdır. Homojen olmama durumu "Kelimelerin Heterojen Dağılımı Durumu" adlı başlıkta ele alınmıştır. Ayrıca iki ya da daha fazla kelimeden oluşan farklı kelime öbeklerinde geçen ortak kelimeler (Merkez Bankası, Ziraat Bankası ifadelerindeki "Bankası") ifadesi de zaman zaman kelime bulutlarında en çok görülen kelimeler arasında yer alabilmektedir. Farklı kelime ikililerinde geçen "bankası" teriminin hangi kelime ikililerinde yer aldığı N-gram'lar sayesinde ortaya çıkarılabilmektedir. Bu durumla mücadele etmek için ise N-gramlar kullanılabilir. Hazırlanan interaktif bir kelime bulutu örneği Şekil 4'te sunulmuştur. Bu uygulama bazı gazetelerden anlık olarak tüm haberleri alarak o anda gündemde olan konuları ortaya çıkarmaktadır. Bulutta görülen bir kelimenin (gözaltı, son, dakika, kriz, fragman gibi) puntosu büyüdükçe o kelimenin haberlerde sık geçtiği anlaşılmaktadır. Bu uygulama R ve Shiny Kütüphanesi ile örnek olarak geliştirilmiştir.

Şekil 4. Kelime Bulutu Örneği (Kaynak: https://suatatan.shinyapps.io/gundem_analizi/) İlgili uygulama bazı haber sitelerinden anlık olarak haberleri alarak gündemi yansıtmaktadır.

5.7.2. Kelime Sıklık Dağılımı Histogramı

Kelime sıklık histogramları da kelime bulutunda olduğu gibi en sık geçen kelimeler listesini kullanmaktadır. Ayrıca her bir kelimenin kaç kez tekrar ettiğini belirten sayılar grafikte açıkça gösterilmektedir. Kelime sıklık grafikleri kelime bulutuna göre sıklık sıralamasının açıkça görülebilmesini sağlar. Ayrıca sık geçen kelimelerin birbirine göre durumunun da görülebilmesine olanak verir. Şekil 5'te Türk Ceza Kanunu'nun metni analiz edilerek en sık görülen kelime kökleri analiz edilmiştir. Bu analizde en sık görülen kelime hazırlanan algoritma ile "kişi" kelimesinin kökü olarak görülen "kiş" ifadesidir.

5.7.3. N-Gramlar

Metinler içinde metinde geçen diğer kelimelere göre daha sık geçen tekil kelimeler olabileceği gibi kelime ikilileri, üçlüleri ya da daha fazla kelimeden oluşan öbeklerin tekrar etmesi mümkündür. Örneğin 'Merkez' ve 'Bankası' kelimeleri bir metin setinde çok sık geçebilir ancak 'Bankası' kelimesi ' Ziraat Bankası' gibi bir ifadeden de, 'Merkez Bankası' gibi bir ifadeden de kaynaklanıyor olabilir. İşte bu durumlarda n-gram analizi ile 'Ziraat Bankası' ve 'Merkez Bankası' ifadelerinin kaç kez görüldüğü ortaya çıkacaktır. Burada n değeri kelime öbeğinin kaç kelimeden oluştuğunu ifade eder (Pattnaik, Rautaray, Das, & Nayak, 2018). Ayrıca n=2 olan n-gramlara bigram n=3 olan n-gramlara trigram denilmektedir. Bu makaledeki metinler esas alınarak hazırlanan n-gram örneği Tablo 3'te sunulmuştur. Türk Ceza Kanunu girdi olarak alınarak bigram (n =2) ve 5gram (n=5) düzeyinde kelime öbekleri çıkarılmıştır. Bu öbekler incelendiğinde ceza kanunundaki ceza tipleri kolayca ortaya çıkmaktadır. 96 sayfalık olan bu metin okunmaksızın n-gram analizi ile genel çerçeve hakkında fikir elde edilebilmektedir.

Tablo 3. N-Gram Örnekleri (Kaynak olarak Türk Ceza Kanununu Analiz Edilmiştir.)

Bigram	**Tekrar**
yıla kadar	371
kadar hapis	366
adlî para	125
müebbet hapis	71

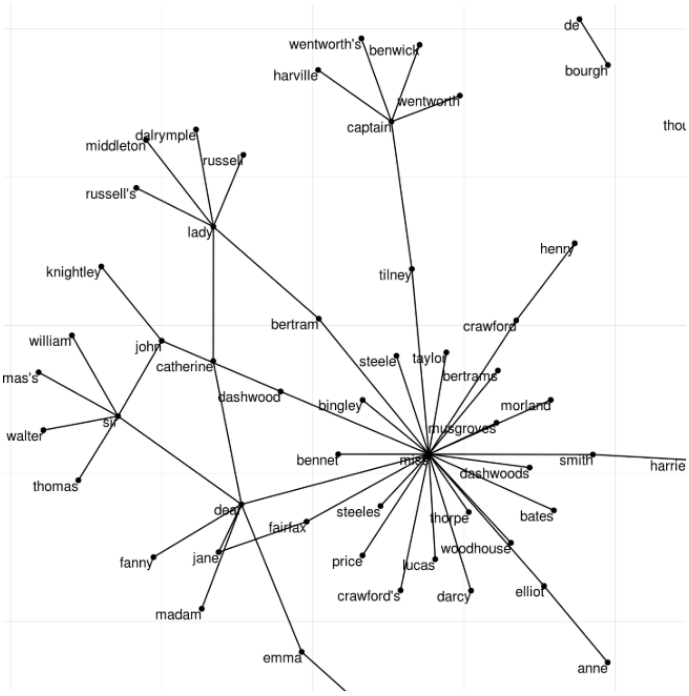
5-gram	**Tekrar**
yıla kadar hapis cezası ile	130
yıldan üç yıla kadar hapis	50
yıldan üç yıla kadar hapis	50
verilecek ceza yarı oranında artırılır	26

5.7.4. Kelime Eş Görülme Korelasyonları

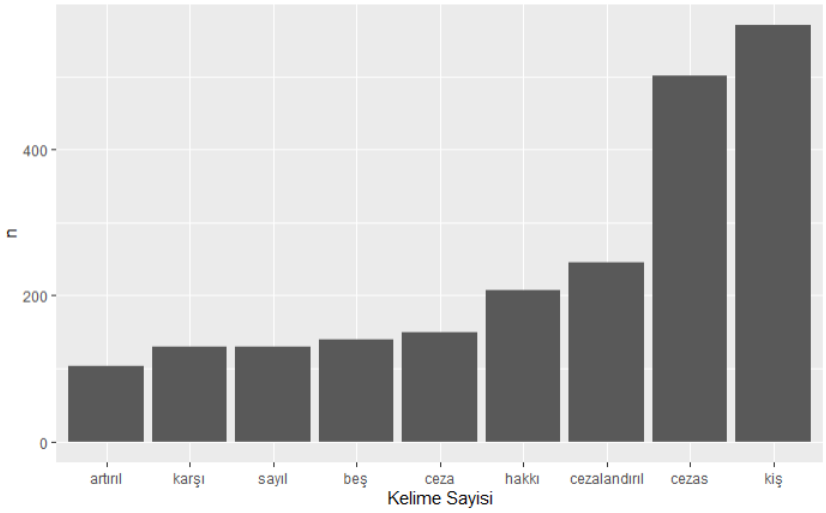
Kelime öbeklerinde geçen her kelimenin yan yana olduğu bilinmektedir. Ancak zaman zaman bir kelime öbeği teşkil etmeyen iki kelime düzenli olarak aynı dokümanlarda görülebilir. Bu durum da anlatılan konularla ilintilidir. Örneğin 'enflasyon' ifadesinin her geçtiği haberde 'ekonomi' kelimesinin geçmesi olasıdır. Bu iki kelime yan yana olmasa bile aynı dokümanda büyük olasılıkla yer alacaktır. Bu tür ilişkileri ortaya çıkarmak için de kelimelerin eş görülme korelasyonları incelenmektedir. Eş görülme korelasyonu ile yardımı ile kelime sıklık diyagramlarında geçen bir kelimenin genellikle başka hangi kelimelerle birlikte yer aldığı ortaya çıkarılarak metin setinin gündemi hakkında fikir edinmek mümkündür.

Farklı n-gramlarda geçen bazı kelime öbeklerindeki ortak kelimeler ile peşi sıra gelen tüm kelimelerden bir ağ diyagramları oluşturulabilir. Ağ bilimi düğüm adı verilen varlıklar ile bunlar arasındaki bağlantıların teşkil ettiği ağ yapısını bütünsel olarak analiz eden bir alandır (Scott & Carrington, 2011). Buna göre kelimelerin her biri düğüm ve art arda gelme durumu da iki kelime arasındaki ilgiyi gösteren bir özellik olmak üzere her metin seti bir ağ diyagramı teşkil edecektir. Ağ diyagramları yardımıyla metinlerde kavramlar arası ilişkiler görsel olarak ortaya çıkarılabilir. Şekil 5'te Jane Austen adlı meşhur romancının eserlerinin ağ diyagramları görülmektedir. Bu diyagramda görüleceği üzere ağda en çok bağlantısı olan düğümler romanın kahramanı ve etrafındaki diğer figürlerin adları olmaktadır.

Şekil 5. Ağ Diyagramı Örneği (Silge ve Robinson, 2018**)**



Şekil 6. Kelime Sıklık Histogramı Örneği (Girdi olarak Türk Ceza Kanununu Analiz Edilmiştir)



5.7.5. TF.IDF Ölçütü: Kelimelerin Heterojen Dağılımı Durumu

Sıklık ölçütleri ile ilgili yukarıda ele alınan tüm ölçütler metin setinin geneli ile ilgilidir. Bir metin seti içinde tek bir yazıda aşırı derecede tekrar etmiş bir ifadenin tüm metin seti içindeki en yüksek frekanslı kelime olma olasılığı bulunmaktadır. Bu durumda örneğin ekonomi ile ilgili yazıların yer aldığı bir metin setinde sadece bir yerde enflasyon kelimesi görüldüğü halde metin setinin tamamı üzerinde analizde bu kelime en sık görülen kelimeler listesinde birinci olarak ortaya çıkabilir. Ancak bu kelimenin dokümanlar arasında bulunmamaktadır. Bir kelime aynı şekilde diğer metinlerde nadiren de görülebilir. Bu olasılık metin madenciliği analizinin sağlığı hakkında şüphe oluşturabilir ancak yukarıda anılan risk dilin doğası göz önüne alındığında görece düşük kabul edilebilir. Nitekim bir dokümanda doğal dil söz konusu ise herhangi bir kelimenin tüm metin setini etkileyebilecek kadar çok tekrar etmesi olası değildir. Bunun için aynı kelimenin örneğin yüz kez alt alta ya da dağınık olarak dokümanda verilmesi gerekir.

Ancak bu ihtimale karşı bir kelimenin tüm dokümanlar içindeki heterojen dağılımı yerine homojen bir şekilde sık dağılımını ölçen IDF (Inverse document Frequency) adlı bir ölçtü bulunmaktadır. Öte yandan hem kelime sıklığını (TF: Term Frequency) hem de IDF değerini aynı anda göz önüne alarak bir kelimenin ne kadar "önemli" olduğunu hesaplayan ve TF ile IDF değerinin çarpımı ile hesaplanan TF.IDF adlı bir ölçüt daha bulunmaktadır. Bu ölçüt arama motorları tarafından da kullanılmaktadır. Söz gelimi "ve, veya" gibi sözcükler, hem her dokümanda olduğundan hem de her dokümanda sıkça görüldüğünden TF.IDF değerleri düşük çıkarken, o doküman setinde sık görülen ancak anlamlı olan anahtar terimler daha yüksek TF.IDF değerine sahip olmaktadır.

6. KISITLAR

Metin madenciliği birtakım kısıtlara sahiptir ve gerçekleştirilecek araştırmaların bu kısıtlar göz önüne alınarak planlanması gerekmektedir. Örneğin elde bulunan metinler içerisindeki konulara bağlı olarak tek başına yorumlanabilir net sonuçlar üretmeyebilir. Bu nedenle analizlerin genellikle bir uzman görüşü ile birlikte değerlendirmesini gerektirecektir.

Ancak metin madenciliği yapılan konuda uzmanların daha fazla araştırması hipotezler türetebilmesi için kayda değer olanaklar yaratmaktadır (Kaushik, 2013). Doğal dillerin esnek kuralları ve istisnaları ile argo ve deyim kullanımı ve yazım hataları da analizlerin önemli zorluklarından bir kısmını teşkil etmektedir. Bilgisayarlar metinler de dâhil olmak üzere çok fazla miktardaki veriyi analiz edebilmesine rağmen anılan bu dilsel öğeleri insanların günlük yaşamda hızla çözdüğü kadar kolay çözememektedir. Metin madenciliği ile ilgili bir diğer kısıt ise dilsel bağımlılıktır. Metinlerin içeriğini direkt olarak görselleştirmeye yarayan kelime bulutu gibi araçlar dışında kalan duygu analizi, özetleme gibi metnin içeriğini anlamaya yönelik algoritmalar

hangi dil için tasarlanmış ise o dilde hizmet verebilir. Bir dil için tasarlanmış algoritma başka bir dilde tam olarak çalışmamaktadır nitekim her dilin farklı söz dizim, yazım ve ifade kuralları bulunmaktadır. Bu nedenle örneğin İngilizce için geliştirilmiş duygu analizi algoritması Türkçe için geçerli olmayacaktır.

Metin madenciliği devasa metinlerden sonuç üretme kapasitesine sahiptir ancak bu durum bazen bu şekildeki bir veri setinin analizi sonucunda elde edilecek çıktıların bile çok fazla miktarda olmasına neden olabilir. Söz gelimi on milyon sayfaya tekabül eden miktardaki bir metnin, on binde biri bile bin sayfalık bir metne tekabül eder. Diğer taraftan metin madenciliği metin içerisinde kavramlar arası ilişkileri tespitte çok güçlü olmakla birlikte tamamen otomatize edilmiş bir çıkarım yapma gücüne halen sahip değildir.

7. DEĞERLENDİRME VE SONUÇ

Giderek artan miktarda verinin üretildiği çağımızda bu verilerin miktar bakımından en önemli kısmını metinler oluşturmaktadır. Bu durum insanların bilgi aktarımında doğal dilleri kullanmalarından kaynaklanmaktadır. Tablolar halinde tutulan verilerin analizi için veri madenciliği alanında kullanılan yöntemler metinlerin analizi için kullanılabilir değildir. Metinlerin analiz edilebilmesi için bir takım özel araçlar ve yaklaşımlar ortaya çıkmıştır. Ortaya çıkan bu araç ve yaklaşımlar metin madenciliği olarak adlandırılan alanın doğmasına neden olmuştur. Metin madenciliği için kullanılan araçlar veri madenciliği alanında kullanılanlardan farklı olduğu gibi elde edilen bulgular da veri madenciliği bulgularına göre daha farklı yorumlanmaktadır. Bir veri madenciliği çalışması sonucunda elde edilen temel istatistiksel bilgiler ilgili veri seti hakkında direkt çıkarımlara olanak verebilir ancak metin madenciliği çalışması sonucunda elde edilen bilgiler ek yorumlamalara ihtiyaç bırakır.

Her metin setinin ve metin setinde bulunan metinlerin yazarlarının öznel özellikleri ile ilgili alanın dinamiklerinden ötürü çok ciddi değişkenlik arz ettiğinden metin madenciliği sonuçları üzerinden standart çıkarımlar yapılması kolay olmamaktadır. Yaşayan bir varlık olarak doğal dilin esnekliği, istisnaları, değişimi gibi özellikleri ile dil kullanımında deyimler, ironiler, argo kullanımı, yazım hataları ve standart olmayan kısaltmalar ve jargonlar metinler üzerinde analiz yapılmasını zorlaştıran unsurlardandır. Bu nedenle haber metinleri veya teknik raporlar gibi standart metinlerin analizleri daha kontrolsüz olan ve standart olmayan sosyal medya metinleri üzerinden gerçekleştirilen analizlerine göre daha kolay yorumlanabilir sonuçlar üretmektedir. Doğal dilin yapısından ve kullanıcılarından kaynaklanan bu zorluklar yine de metin madenciliği ile araştırmacılar ve organizasyonların günümüzdeki devasa metin yığınlarını incelemekten alıkoymamaktadır. Nitekim metinlerden süzülen çıktılar bir sonuç üretmediğinde dahi analiz edilen metin yığınları üzerinde daha önce sorulmamış yeni soruların üretilmesine ve hipoteze dönüştürülmesine olanak vermektedir. Metin madenciliği çalışmalarında bazı standart çıktılar dışındaki çıktılar araştırmacıların tercihlerine göre elde edilmektedir. Örneğin eş görülme korelasyonu bunlardan biridir. Araştırmacı metin yığınları içinde herhangi bir kelimenin en sık birlikte görüldüğü kelimeyi sorgulamak suretiyle yeni bir sonuca varabilir. Diğer taraftan metin madenciliği araçlarından örneğin metin özetleme özelliği metinlerden sonuç üretmek yerine metinlerin daha kısa forma getirilerek okunabilmesine olanak vermektedir. Bu durum ise metin madenciliğinin sadece devasa metin yığınlarından sadece sonuç üretmek için değil bu metinlerle başa çıkmak için de var olduğunu ortaya koyar. Eğer insanlar tarafından okunacaksa bile ilgisiz metinlerin baştan elenmesine ve zaman kazanılmasına hizmet eden özetleme algoritmaları yine metin madenciliği alanı içerisindedir. Spam (gereksiz reklam e-postası) filtreleme algoritmaları metin madenciliğinin bu alandaki en somut faydalarının ürüne dönüşmüş bir örneğidir. Gelişmekte olan spam algoritmaları ile artık e-posta kutularımızda daha az spam mesajla karşılaşmaktayız.

Metin madenciliği çalışmalarında kullanılan yaygın araçların çoğu ücretsiz ve açık kaynaklıdır. Bu durum bu araçların yaygın bir kullanıcı kitlesine sahip olmasına bunun yanında internette bu kaynakların kullanımı ile ilgili çok miktarda açıklayıcı dokümantasyonun ortaya çıkmasına neden olmuştur. Bu durum her düzeyde organizasyonlar ve araştırmacılar için önemli fırsatlar yaratmaktadır. Büyük ve eski bir işletme hakkında çıkan çok sayıda haberleri metin madenciliği yöntemi ile analiz ederek imajı hakkında fikir edinebilir iken yeni ve küçük bir işletme ilgili olduğu mevcut ürünlerle ilgili internette yazılan yorumları

kullanarak metin madenciliği çalışmaları ile ürününün özelliklerini geliştirmek için yeni ve değerli bilgiler ortaya çıkarabilir. Politikacılar ve politik organizasyonlar kamuoyu görüşlerini anketle sormak yerine sosyal medya analizini kullanarak kamuoyunun nabzını tutabilirler. Akademisyenler araştırmacının zaman ve imkânlarına göre sınırlı tutulan bibliyometrik çalışmalarını metin madenciliği yöntemlerini kullanarak çok daha fazla akademik materyale yöneltebilirler. Aynı şekilde metin madenciliği ile daha önce hiçbir şekilde tamamen incelenmemiş büyük miktarda haber metinleri, raporlar veya başka metinsel kaynaklar topluca analiz edilerek önemli çıkarımlar elde edilebilir.

KAYNAKÇA

Atan, S. (2016a). Metin Madenciliği ile Sentiment Analizi ve Borsa İstanbul Uygulaması, Doktora tezi, Ankara Üniversitesi.

Atan, S. (2016b). Veri, Büyük Veri ve İşletmecilik. Balıkesir Üniversitesi Sosyal Bilimler Dergisi, 19(35), 137-153.

Atan, S., & Emekci, H. (2018). İktisat ve İşletme Uygulamaları İçin R ile Veri Analizi, İstatistik, Modelleme ve Uygulama (1. bs). Seçkin Yayıncılık.

Babu, B. V., Nagar, A., Deep, K., Pant, M., Bansal, J. C., Ray, K., & Gupta, U. (2014). Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Springer.

Baker, B. S. (1995). On finding duplication and near-duplication in large software systems. 86-95. https://doi.org/10.1109/WCRE.1995.514697

Bolasco, S., Canzonetti, A., Capo, F. M., Ratta-Rinaldi, F. della, & Singh, B. K. (2005). Understanding Text Mining: A Pragmatic Approach. Içinde Knowledge Mining (ss. 31-50). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-32394-5_4

Cady, F. (2017). The Data Science Handbook. John Wiley & Sons.

Chen, H. (2011). Dark Web: Exploring and Data Mining the Dark Side of the Web. Springer Science & Business Media.

Çalış, K., Gazdağı, O., & Yıldız, O. (2013). Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti. International Journal Of Informatics Technologies, 6(1), 1–7.

Davenport, T. (2014). Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Review Press.

Derick, J. (2010). Three Real-World Applications of Text Mining to Solve Specific Business Problems by Derick Jose—BeyeNETWORK. http://www.b-eye-network.com/view/12783

Dolgun, M. Ö., Özdemir, T. G., & Oğuz, D. (2009). Veri madenciliğiâ nde yapısal olmayan verinin analizi: Metin ve web madenciliği. İstatistikçiler Dergisi: İstatistik ve Aktüerya, 2(2).

Feldman, R., & Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

Fontama, V., Barga, R., & Tok, W. H. (2015). Predictive Analytics with Microsoft Azure Machine Learning 2nd Edition. Apress.

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence, 1(1). https://doi.org/10.4304/jetwi.1.1.60-76

Hansen, C. D., & Johnson, C. R. (2011). Visualization Handbook. Elsevier.

İlhan, S., Duru, N., Karagöz, Ş., & Sağır, M. (2008). Metin Madenciliği ile Soru Cevaplama Sistemi. Elektronik ve Bilgisayar Mühendisliği Sempozyumu (ELECO), Bursa, 26–30.

Karadağ, A., & Takçı, H. (2010). Metin madenciliği ile benzer haber tespiti. Akademik Bilişim.

Kaşıkçı, T., & Gökçen, H. (2013). Metin Madenciliği İle E-Ticaret Sitelerinin Belirlenmesi. Bilişim Teknolojileri Dergisi, 7(1).

Kaushik, M. L. (2013). Text Mining—Scope and Applications. Text Mining, 55(2).

Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M., & Özçift, A. (2016). KNN algoritması ve r dili ile metin madenciliği kullanılarak bilimsel makale tasnifi.

Miner, G., IV, J. E., & Hill, T. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press.

Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). Data Mining in Agriculture. Springer Science & Business Media.

O'Keeffe, A., & McCarthy, M. (2010). The Routledge Handbook of Corpus Linguistics. Routledge.

Patel, F. N., & Soni, N. (2012). Text mining: A Brief survey.

Pattnaik, P. K., Rautaray, S. S., Das, H., & Nayak, J. (2018). Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2017. Springer.

ReSoomer Text Summarization Tool. Erişim: 3 Mart 2019, https://resoomer.com/en/

Scott, J., & Carrington, P. J. (2011). The SAGE Handbook of Social Network Analysis. SAGE.

Shaalan, K., Hassanien, A. E., & Tolba, F. (2017). Intelligent Natural Language Processing: Trends and Applications. Springer.

Silge, J., & Robinson, D. (2017). Text Mining with R: A Tidy Approach. O'Reilly Media, Inc.

Talib, R., Kashif, M., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. International Journal of Advanced Computer Science and Applications, 7(11). https://doi.org/10.14569/IJACSA.2016.071153

TextRank Algorithm for Text Summarization for R Language. Erişim: 3 Mart 2019, https://cran.r-project.org/web/packages/textrank/vignettes/textrank.html