KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

İNGİLİZCE HABER METİNLERİNDE GDT VE NOMF YÖNTEMLERİ İLE KONU MODELLEME: TÜRKİYE VE YUNANİSTAN ÖRNEĞİ

YÜKSEK LİSANS TEZİ

Sefa YAY

MART 2022
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

İNGİLİZCE HABER METİNLERİNDE GDT VE NOMF YÖNTEMLERİ İLE KONU MODELLEME: TÜRKİYE VE YUNANİSTAN ÖRNEĞİ

Sefa YAY

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"YÜKSEK LİSANS (İSTATİSTİK)"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 07/02/2022
Tezin Savunma Tarihi : 04/03/2022

Tez Danışmanı : Dr. Öğr. Üyesi Tolga BERBER

Trabzon 2022

ÖNSÖZ

"İngilizce Haber Metinlerinde GDT ve NOMF Yöntemleri ile Konu Modelleme: Türkiye ve Yunanistan Örneği" isimli bu tez çalışması, Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstatistik ve Bilgisayar Bilimleri Anabilim Dalı, Yüksek lisans Programı'nda hazırlanmıştır.

Bu tez çalışmasında bilgisi ve yardımı ile bana her türlü desteği veren, değerli görüşleri ile yol gösteren tez danışmanım Dr. Öğr. Üyesi Tolga BERBER'e, yüksek lisans eğitimi yapma fırsatı bulduğum İstatistik ve Bilgisayar Bilimleri bölümündeki hocalarıma ve tez sürecindeki vizyonumu oluşturmamdaki katkılarından ve desteğinden dolayı Dr. Öğr. Üyesi Devran YAZIR'a teşekkürlerimi sunarım.

Tüm hayatım boyunca olduğu gibi eğitim ve öğretim hayatım boyunca da maddi ve manevi her zaman beni destekleyen, her adımımda arkamda duran aileme sonsuz teşekkürlerimi sunarım.

Sefa YAY

Trabzon 2022

TEZ ETİK BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum “İngilizce Haber Metinlerinde GDT ve NOMF Yöntemleri ile Konu Modelleme: Türkiye ve Yunanistan Örneği” başlıklı bu çalışmayı baştan sona kadar danışmanım Dr. Öğr. Üyesi Tolga BERBER‘in sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuvarlarda yaptığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 04/03/2022

Sefa YAY

İÇİNDEKİLER

Sayfa No

Yüksek Lisans

ÖZET

İNGİLİZCE HABER METİNLERİNDE GDT VE NOMF İLE KONU MODELLEME: TÜRKİYE VE YUNANİSTAN ÖRNEĞİ

Sefa YAY

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik ve Bilgisayar Bilimleri Anabilim Dalı
Danışman: Dr. Öğr. Üyesi Tolga BERBER
2022, 60 Sayfa

Haber analizi, e-posta ve spam filtreleme, web sayfalarından konu çıkarımı, bloglar, film özetleri, şarkı sözleri gibi metin içeren her veri seti metin madenciliği için bir uygulama alanıdır. Bu birçok alandaki uygulamalar sayesinde büyük metin depolarından bilgi çıkarılmasına olanak sağlamaktadır. Konu modelleme ise bir belge koleksiyonunda metnin gizli anlamsal yapılarını keşfetmek için kullanılan doğal dil işleme tekniğidir. Bu tez kapsamında Türkiye ve Yunanistan'a yönelik haber metinlerini konularına göre ayırabilen otonom bir konu modellemesi gerçekleştirilmiştir. Bunun için NewsAPI haber veri sitesinden elde edilmiş olan İngilizce haber metinlerinden Gizli Dirichlet Tahsisi ve Negatif Olmayan Matris Faktorizasyonu yöntemleri kullanılmış ve bu iki yöntemin başarım karşılaştırılması yapılmıştır. Türkiye için yapılan analiz sonucundaki konular incelendiğinde dış ilişkiler ağırlıkta siyasi bir gündem olduğu görülmektedir. Yunanistan için olan analizlerde ise tek siyasi gündemin Türkiye ile aralarında yaşandığı tespit edilmiştir. Her iki algoritmanın sonuçlarında da pandeminin farklı yönlerinin çoğunluğu oluşturduğu belirlenmiştir. Böylelikle metin madenciliğinde büyük boyuttaki metin içerikli veri kaynaklarından, önceden bilinmeyen ve potansiyel olarak ihtiyaç duyulan bilginin çıkarılması sağlanmış olundu.

Anahtar Kelimeler: Konu Modelleme, Metin Madenciliği, GDT, NOMF, Veri Madenciliği

Master Thesis

SUMMARY

TOPIC MODELING WITH LDA AND NMF IN ENGLISH NEWS TEXTS: THE CASE OF TURKEY AND GREECE

Sefa YAY

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Statistics and Computer Science Graduate Program
Supervisor: Tolga BERBER
2022, 60 Pages

Every dataset containing text such as text mining, news analysis, e-mail and spam filtering, topic extraction from web pages, blogs, movie summaries, and lyrics is an application field for text mining. This enables applications to be extracted from large text stores thanks to applications in many areas. Topic modeling is a natural language processing technique used to discover hidden semantic structures of text in a document collection. Within the scope of this thesis, automatic subject modeling has been made, where we can separate the news texts for Turkey and Greece according to their subjects. For this, English news texts obtained from NewsAPI news data site were automatically analyzed using Latent Dirichlet Allocation and Non-Negative Matrix Factorization methods. Also, comparison of the two methods is provided. When the issues as a result of the analysis for Turkey are examined, it is seen that foreign relations is a predominantly political agenda. In the analyzes for Greece, it has been determined that the only political agenda is between Greece and Turkey. In the results of both algorithms, it was determined that different aspects of the pandemic constitute the majority. Thus, in text mining, previously unknown and potentially needed information has been extracted from large text-containing data sources.

Keywords: Topic Modeling, Text Mining, LDA, NMF, Data Mining

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER VE KISALTMALAR

API : Uygulama Ara Yüzü

BE : Bilgiye Erişim

DDİ : Doğal Dil İşleme

GAA : Gizli Anlamsal Analiz

GDT : Gizli Dirichlet Tahsisi

İKM : İlişkili Konu Modeli

KM : Konu Modeli

NKB : Noktasal Karşılıklı Bilgi

NNKB : Normalleştirilmiş Noktasal Karşılıklı Bilgi

NOMF : Negatif Olmayan Matris Faktorizasyonu

OGAA : Olasılıksal Gizli Anlamsal Analiz

TBF : Ters belge Frekansı

TF : Terim Frekansı

N : Tüm Belgelerin Sayısı

n : Bir Kelimenin Belgede Geçme Sayısı

w : Bir Belgedeki Bir Kelimenin Ağırlık Değeri

v : Bir Belgedeki Toplam Terim Sayısını

m : Veri Kümesindeki Örneklerin Sayısı

V : Belge-Kelime Matrisi

W : Belgelerden Keşfedilen Konular (Kümeler)

H : Katsayı matrisi

1. GENEL BİLGİLER

1.1. Giriş

Günümüz teknolojik gelişmeleri sayesinde, klasik veri analiz araçlarının birçoğunun altından kalkamayacağı miktarda veri üretilmeye başlanmıştır. Bu büyük miktardaki verilerin işlenmesi, veri temsili, belirsizlik giderme ve boyut azaltma ile ilgili yeni araştırmaların başlatılmasına sebep olmuştur [1].

Bilgi sistemleri ve teknolojinin gelişmesiyle, kamu kurum ve kuruluşları, işletmeler ve diğer kuruluşlar veritabanlarında topladıkları çeşitli türlerdeki veriler, anlamsız veri yığınları haline gelmiştirler [2,3].

Büyük ölçüde yapılandırılmamış olan ve yüzlerce yıllık bilimsel çalışmayı kapsayan milyonlarca metinsel içeriklerde otonom analiz çok önemlidir. Bu tür arşivlerin taranması, aranması ve verimli kullanımına izin verilmesi için yeni araçların geliştirilmesi önemli bir teknolojik zorluktur ve istatistiksel modelleme için yeni fırsatlar sağlamıştır [4].

Metinsel analiz 1980'lerde ortaya çıkmış olsa da teknolojik ilerlemelerle gelişme göstermiştir. Bazı teknoloji firmaları, müşterilerine İnternet kaynaklarından toplanan metinsel veriler sağlar. Bu tür veriler çoğunlukla pazarlama şirketleri tarafından kullanılmaktadır. Günümüzde e-postalar, Facebook ve Twitter'daki notlar ve bloglar güvenlik amacıyla kontrol edilmektedir. Bilimde GoPubMed gibi bilgiye dayalı arama motorları kullanılmaktadır. Kısacası bilgisayar destekli metin analizi farklı alanlarda kullanılmakta ve çeşitli teknikler verilerin analiz edilmesini mümkün kılmaktadır [5].

Otonom bir şekilde metin analizi gerçekleştirmek için geliştirilen tekniklerden biri de konu modellemedir (KM). KM, otonom bir şekilde insanların belge koleksiyonlarını anlamlandırmasını sağlamaktadır. KM'de olası konular önceden bilinmez. Amacı bir belgede hangi konuların bulunabileceğinin tespiti ve kelimeleri, her bir konu için kelime kümeleri halinde gruplamaktır. Geniş belge koleksiyonlarındaki gizli temaları ortaya çıkarmak için hem denetimsiz hem de denetimli istatistiksel makine öğrenme yöntemlerini kullanan bir metin madenciliği tekniğidir [6, 87].

1.2. Veri Bilimi

Veri bilimi, bilginin verilerden genellenebilir çıkarımlar için uygulama ve yöntemleri bir araya getiren disiplinlerarası bir bilimdir. Veri bilimi bilimsel alanlarını veri mühendisliği, bilgi bilimi, bilgisayar bilimi, istatistik, yapay zeka, makine öğrenimi, veri madenciliği ve tahmine dayalı analitik gibi düşündürücü araştırma alanları ile birleştirmektedir. Veri bilimi, istatistikten ve diğer mevcut disiplinlerden birkaç önemli yönden farklıdır. Başlangıç olarak, veri biliminin "veri" kısmı olan ham madde (metin, resim, video gibi) giderek daha heterojen ve yapılandırılmamış hale gelmektedir [15, 16].

Bilinmeyen içgörülerin çoğunun, işletmelerdeki büyük bir işlenmemiş veri havuzundan elde edildiği ve saklandığı bilgisine göre, veri yığınlarının iş dünyasının geleceği üzerinde büyük etkileri vardır. Dünya çapındaki veri bilimcilerinin, 2005 yılından bu yana 300 kat artışla 40 zettabayt'a ulaşan, farklı tiplerden çıkan yapılandırılmamış heterojen veriyi 2020 yılına kadar işlemesi gerekeceği tahmin edilmektedir. Halen eyleme geçirilebilir iş bilgisine dönüştürülmesi gereken çok sayıda veri kaynağı vardır. Veri biliminde veri tipleri o kadar heterojen ve çeşitlidir ki, tipik uygulamalarda çok boyutlu bir veri tipi için tartışılan temel yöntemler etkili olmayabilir. Bu nedenle, farklı veri türlerine ve bu farklı veri türleri bağlamında ortaya çıkan uygulamalara daha fazla önem verilmesi gerekmektedir [16, 17].

Büyük ve artan miktarda veriyle birlikte günümüzün en belirgin ifşaatlarından biri, alan bilgisi ve analizinin birbirinden ayrılamayacağıdır [18]. Provost ve Fawcett [19], "Veri bilimi, veri madenciliği algoritmalarından çok daha fazlasını içerir. Başarılı veri bilimcileri, iş sorunlarını veri perspektifinden görebilmelidir" şeklinde belirtmiştir.

1.3. Makine Öğrenmesi

Makine öğrenimi, farklı disiplinlerdeki öğrenme sistemlerinin birlikte çalışması için ayrılmış araştırma alanıdır. Bu, istatistik, bilgisayar bilimi, mühendislik, bilişsel bilim, optimizasyon teorisi ve diğer birçok bilim ve matematik disiplininden gelen fikirleri ödünç alan ve bunlara dayanan oldukça disiplinler arası bir alandır [38].

Makine öğrenimi genellikle, bir görevi bir dizi örnekten öğrenen mantıksal veya ikili işlemlere dayalı otonom hesaplama prosedürlerini kapsar [39]. Makine öğrenmesinde, veri ve kullanılan algoritma ile oluşturulan model, en yüksek performansı vermek üzere

tasarlanmaktadır. Bunun için pek çok makine öğrenmesi yöntemi geliştirilmiş olup geliştirilen yöntemlerden bazıları; k-en yakın komşu algoritması, lojistik regresyon analizi, k-ortalamalar algoritması, basit Naive Bayes sınıflandırıcı, karar ağaçları, destek vektör makinaları ve yapay sinir ağlarıdır [35].

Bu yaklaşımların bir kısmı tahmin edici bir kısmı da tanımlayıcı yöntemleri içermektedir. Bu yöntemlerde öğrenme stratejileri; denetimli, denetimsiz ve pekiştirmeli (takviyeli) olmak üzere üç gruba ayrılmaktadır.

Makine öğrenimi, insan tarafından kolayca anlaşılabilecek kadar basit sınıflandırma ifadeleri üretmeyi amaçlar. Karar sürecine içgörü sağlamak için insan akıl yürütmesini yeterince taklit etmelidirler. İstatistiksel yaklaşımlar gibi, arka plan bilgisi geliştirmede kullanılabilir, ancak tüm operasyon insan müdahalesi olmayacak şekilde varsayılır [39]. Bu yaklaşımın avantajları ne bilgi mühendislerinin ne de alan uzmanlarının müdahalesine gerek olmadığı için uzman insan gücü açısından önemli bir tasarruftur [40].

1.3.1. Denetimli Öğrenme

Ölçülen veya önceden ayarlanmış girdiler olarak gösterilebilecek bir dizi değişkenin bir veya daha fazla çıktı üzerinde bazı etkileri vardır. Amaç, çıktıların değerlerini tahmin etmek için girdileri kullanmaktır. Bu egzersize denetimli öğrenme denir [41].

Girişler ölçüm türüne göre değişir; her bir nitel ve nicel girdi değişkenine sahip olunabilir. Bunlar aynı zamanda tahmin için kullanılan yöntem türlerinde de farklılıklara yol açar: bazı yöntemler doğal olarak nicel girdiler için, bazıları doğal olarak nitel için ve bazıları her ikisi için de tanımlanmıştır. Nicel çıktılar tahmin edildiğinde regresyon ve nitel çıktılar tahmin edildiğinde sınıflandırma yöntemleri uygulanır [41].

Denetimli öğrenmede, makineye bir dizi istenen çıktı (y1, y2, …) verilir. Makinenin amacı, verilen yeni bir girdide doğru çıktıyı üretmeyi öğrenmektir. Bu çıktı bir sınıf etiketi (sınıflandırma) veya gerçek bir sayı (regresyon) olabilir [38].

Nitel değişkenler tipik olarak sayısal kodlarla temsil edilir. En kolay durum, “başarı” veya “başarısızlık”, “hayatta kaldı” veya “öldü” gibi yalnızca iki sınıf veya kategori olduğu zamandır. Bunlar genellikle 0 veya 1 olarak tek bir ikili rakam veya bit ile veya -1 ve 1 ile temsil edilir. Belirgin hale gelecek nedenlerden dolayı, bu tür sayısal kodlar bazen hedefler olarak anılır. İkiden fazla kategori olduğunda, birkaç alternatif mevcuttur. En kullanışlı ve yaygın olarak kullanılan kodlama, kukla değişkenler aracılığıyladır. Burada bir K-seviyesi

nitel değişken, bir seferde sadece biri “açık” olan K ikili değişkenleri veya bitlerinin bir vektörü ile temsil edilir. Daha kompakt kodlama şemaları mümkün olsa da kukla değişkenler faktör seviyelerinde simetriktir [41].

1.3.2. Denetimsiz Öğrenme

N boyutlu rastgele bir X vektör değişkeni varsayıldığında, genellikle veri olarak adlandırılan bu girdi, retinadaki bir görüntüye, kameradaki piksellere veya bir ses dalga biçimine karşılık gelebilir. Aynı zamanda, örneğin bir haberdeki kelimeler veya bir süpermarket alışveriş sepetindeki öğelerin listesi gibi daha az belirgin olan duyusal verilere de karşılık gelebilir [38, 41].

Denetimsiz öğrenmenin iki çok basit klasik örneği, kümeleme ve boyut azaltmadır. Denetimsiz öğrenmede amaç, her gözlem için doğru cevaplar veya hata derecesi sağlayan bir gözetmen veya öğretmenin yardımı olmadan girdilerin özniteliklerini doğrudan çıkarmaktır [38, 41]. Bununla birlikte, makinenin amacı karar verme, gelecekteki girdileri tahmin etme, girdileri başka bir makineye verimli bir şekilde iletme vb. için kullanılabilecek girdi temsillerini oluşturmaktır. Başka bir ifadeyle, saf yapılandırılmamış gürültü olarak kabul edilecek olan verilerin üstünde ve ötesinde örüntüler bulmak olarak düşünülebilir [38].

1.3.3. Pekiştirmeli Öğrenme

Pekiştirmeli öğrenme fikri, daha basit temel modellerden oluşan bir koleksiyonun güçlü yönlerini birleştirerek bir tahmin modeli oluşturmaktır. Aslında, zayıf öğrenenlerin rolüne hizmet eden temel işlevlerle, regresyon eğrileri gibi herhangi bir sözlük yöntemi, bir pekiştirmeli yöntem olarak karakterize edilebilir. Parametrik olmayan regresyon için Bayesci yöntemler, pekiştirmeli yöntemler olarak da görülebilir [41].

Pekiştirmeli (takviyeli) öğrenme yönteminde, hedef çıktıyı vermek için bir danışman yerine, elde edilen çıkışın verilen girişe karşılık iyi ya da kötü olarak değerlendiren bir kriter kullanılmaktadır [35].

Pekiştirmeli öğrenmede makine, a1, a2, … eylemleri üreterek çevresiyle etkileşime girer. Bu eylemler ortamın durumunu etkiler, bu da makinenin bazı sayısal ödüller veya

cezalar almasıyla sonuçlanır. Makinenin amacı, ömrü boyunca alacağı gelecekteki ödülleri en üst düzeye çıkaracak (veya cezaları en aza indirecek) şekilde hareket etmeyi öğrenmektir [38]. Her eylem için küçük ceza puanları ve başarılı tamamlama için büyük bir pozitif ödül atayarak, bir temsilci tarafından ideal davranışa ilişkin açık örnekler sunulmadan veya belirli bir görevin nasıl tamamlanacağı belirtilmeden, hedef odaklı ve verimli bir şekilde hareket etmek üzere eğitilebilir [42].

Pekiştirmeli öğrenme, karar teorisi (istatistik ve yönetim biliminde) ve kontrol teorisi (mühendislikte) alanlarıyla yakından ilişkilidir. Bu alanlarda çalışılan temel problemler genellikle biçimsel olarak eşdeğerdir ve genellikle problemin ve çözümün farklı yönleri vurgulansa da çözümler aynıdır [38].

1.4. Veri Madenciliği

Veri madenciliği, verilerin toplanması, temizlenmesi, işlenmesi, analiz edilmesi ve bunlardan yararlı içgörüler elde edilmesi üzerine yapılan çalışmalardır. Veri madenciliğinde gerçek uygulamalarda karşılaşılan problem alanları, uygulamalar, formülasyonlar ve veri temsilleri açısından geniş bir varyasyona sahiptir. Bu açıdan değerlendirildiğinde veri işlemenin farklı yönlerini tanımlamada kapsayıcı bir terim olmuştur [17, 20].

İstatistik, makine öğrenimi, veritabanları, sinir ağları, örüntü tanıma, ekonometri ve diğerleri gibi alanlardaki araştırmacıların hepsi “Veri madenciliği” kavramı olmadan öncesinde de aynı tür problemler üzerinde çalışmışlardır. Veri madenciliği kavramı oluşmadan öncesinde de veri madenciliği problemleri üzerine çalışan alanlar Şekil 1’de gösterilmiştir. Ancak birbirlerinin çalışmalarından tam olarak yararlanamamışlardır. Akademi dışında teknolojinin potansiyel kullanıcılarını heyecanlandıracak tek bir isim olmadan, araştırmalarını gerçek sorunlara uygulama çabaları, yaygın bir ilgi oluşturmamıştır. Bu yüzden “Veri madenciliği” kavramı tam olarak tanımlanmıştır [20].

Modern çağda, hemen hemen tüm otonom sistemler, teşhis veya analiz amacıyla bir tür veri üretir. Tahminlere göre dünyadaki veri miktarı her yirmi ayda bir ikiye katlanmaktadır. Bu, petabayt veya eksabayt mertebesine ulaşan bir veri akışına neden olmaktadır [17, 21]. Üretilen bu veriler; uzay görüntülerinin analizinden tıbbi teşhislere, toksik tehlike analizlerinden metin özeti ve duygu analizi gibi çeşitli alanlarda hızla büyüyen başarılı uygulamalar aracılığı ile incelenmektedir [22].

Şekil 1. Veri madenciliği ile ilişkili problemler üzerine çalışan alanlar

Muhasebe ve envanter yöneticilerine hizmet eden işlemsel veritabanlarının aksine, analiz ve karar desteği için büyük veritabanlarının toplanmasını ifade eden veri ambarının popülaritesi, iş dünyasında veri madenciliğine duyulan ihtiyacı körüklemiştir. Toplanan veri miktarı arttıkça, verileri verimli bir şekilde analiz etmek için daha fazla araca olan ihtiyaç da artmıştır [20].

Bramer [22], veri madenciliği uygulamalarını, sınıflandırma, regresyon, ilişkilendirme ve kümeleme olarak dört ana türe ayırıyor ve hangisinin uygulanacağına karar vermeden önce etiketli ve etiketsiz veri olarak iki tür veri arasında ayrım yapılması gerektiğini belirtmiştir. Gözlemlenen bir olgunun ölçülebilir bir niteliğine öznitelik denilmektedir [86]. Özel olarak belirlenmiş bir özniteliğin görülmemiş örneklerde tahmin edilmesi için kullanılan veri kümesine etiketli veri denilmektedir. Etiketli veriyi kullanarak henüz görülmemiş örnekler için etiket olarak belirlenen öznitelik değerini tahmin etmek, veri madenciliğinde denetimli öğrenme olarak bilinir. Belirtilen etiket kategorik ise sınıflandırma, sayısal ise regresyon uygulanır [22].

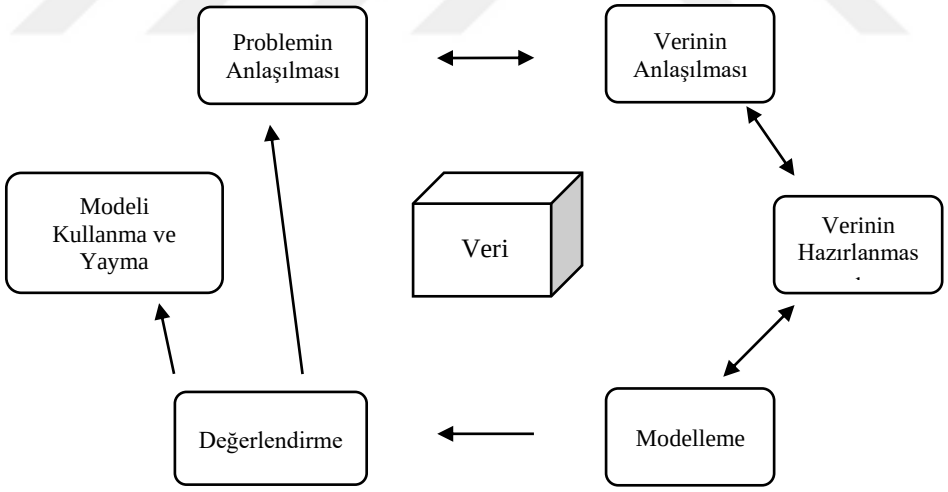
Özel olarak belirlenmiş herhangi bir özniteliği olmayan verilere etiketsiz olarak adlandırılır. Etiketsiz verilerin veri madenciliğindeki karşılığı denetimsiz öğrenme olarak bilinir. Amacı, mevcut verilerden elde edinilebilecek en fazla bilgiyi çıkarmaktır [22].

1.4.1. Veri Madenciliği Süreci

Veri madenciliği, bir problemin anlaşılması ve tanımlanması ile başlayan ve sonuçların analizi ve sonuçların avantaj elde etmek için kullanılması için bir strateji ile biten birkaç adımı içeren yinelemeli bir süreçtir. Başka bir ifadeyle, veri madenciliği süreci, problemi tanımladıktan sonra veri toplama, veri temizleme, öznitelik çıkarma ve algoritmik tasarım gibi birçok aşamayı içeren bir akıştır [17, 20].

Veri analisti, veri setini probleme göre tavsiyelerde bulunmak nasıl kullanacağına karar verir. Tavsiyeleri yerine getirmenin pek çok yolu vardır, bunlardan bazıları problemin özel tanımına bağlı olarak diğerlerinden daha etkilidir [17]. Bu nedenle, Aggarwal [17] tüm veri madenciliği sürecini, analistin becerisine dayanan bir sanat biçimi olarak tanımlamaktadır.

Bir veri madenciliği projesinin yaşam döngüsü problem tanımlama, veri anlama, veri hazırlama, modelleme, değerlendirme ve uygulama olmak üzere altı aşamaya ayrılmıştır ve bu akış Şekil 2'de gösterilmiştir [23].



Şekil 2. Veri madenciliği sürecinin akışı

1.4.1.1. Problemi Tanımlama

Veri madenciliği projesinin belki de en önemli aşaması olan problem tanımlama aşaması, proje hedeflerini iş perspektifinden anlamaya, bu bilgiyi bir veri madenciliği problem tanımına dönüştürmeye ve ardından hedeflere ulaşmak için tasarlanmış bir ön

plan geliştirmeye odaklanır. Hangi verilerin daha sonra ve nasıl analiz edilmesi gerektiğini anlamak için, veri madenciliği uygulayıcılarının çözüm buldukları işi tam olarak anlamaları hayati önem taşımaktadır. İşi anlama veya problemi tanımlama aşaması, iş hedeflerini belirleme, durumu değerlendirme, veri madenciliği hedeflerini belirleme ve proje planını üretme dahil olmak üzere birkaç temel adımı içerir [23].

1.4.1.2. Veriyi Anlama

Veri anlama, veri toplama ile başlar. Analist daha sonra verilere aşinalığı artırmaya, veri kalitesi sorunlarını belirlemeye, verilere ilişkin ilk öngörüleri keşfetmeye veya gizli bilgiler hakkında hipotezler oluşturmak için alt kümeleri tespit etmeye devam edecektir. [23].

Veri anlama aşaması, verilerin toplanması, verilerin tanımlanması, verilerin araştırılması ve veri kalitesinin doğrulanması dahil olmak üzere dört adımı içerir [23].

- Veri toplama, bir sensör ağı gibi özel donanımların, kullanıcı anketlerinin toplanması gibi el işçiliğinin veya belgeleri toplamak için Web belgesi tarama motoru gibi yazılım araçlarının kullanılmasını gerektirebilir. Bu aşama son derece uygulamaya özel ve genellikle veri madenciliği analistinin alanı dışında olsa da kritik derecede önemlidir çünkü bu aşamadaki iyi seçimler veri madenciliği sürecini önemli ölçüde etkileyebilir. Toplama aşamasından sonra, veriler genellikle bir veri tabanında veya daha genel olarak işlenmek üzere bir veri ambarında saklanır [17]. Bu aşamada analist, projenin gelecekteki tekrarlarına yardımcı olmak için karşılaşılan sorunları ve çözümlerini rapor ettiğinden emin olmalıdır [23].

- Veriyi tanımlama, bu adımda veri analisti, elde edilen verilerin “brüt” veya “yüzey” özniteliklerini inceler. Verilerin biçimi, verilerin miktarı, her veri tablosundaki kaydedilen verilerin ve alanların sayısı, alanların kimlikleri ve verilerin diğer yüzey öznitelikleri gibi konuları inceler ve sonuçları raporlar [23].

- Veri keşfetme, sorgulama, görselleştirme ve raporlama kullanılarak ele alınabilecek veri madenciliği sorularını içerir. Örneğin, bir veri analisti, belirli bir gelir grubundaki alıcıların genellikle satın aldığı ürün türlerini keşfetmek için verileri sorgulayabilir veya olası dolandırıcılık modellerini ortaya çıkarmak için bir görselleştirme analizi yapabilir. Veri analisti daha sonra ilk bulguları veya bir ilk hipotezi ve projenin geri kalanı üzerindeki potansiyel etkiyi özetleyen bir veri araştırma raporu oluşturmalıdır [23].

- Veri kalitesini doğrulama, bu noktada analist verilerin kalitesini inceler. Kontrol edilen bazı yaygın öğeler şunlardır: eksik nitelikler ve boş alanlar, tüm olası değerlerin temsil edilip edilmediği, değerlerin akla yatkınlığı, değerlerin yazılışı ve farklı değerlere sahip özniteliklerin benzer anlamlara sahip olup olmadığı. Veri analisti aynı zamanda sağduyuyla çelişen cevaplar verebilecek tüm öznitelikleri de gözden geçirmelidir [23].

1.4.1.3. Veriyi Hazırlama

Veriler toplandığında, genellikle işlemeye uygun bir biçimde olmazlar. Örneğin, veriler karmaşık günlüklerde veya serbest biçimli belgelerde kodlanabilir. Çoğu durumda, farklı veri türleri serbest biçimli bir belgede rastgele karıştırılabilir [17].

Veri hazırlama aşaması, nihai veri setini veya ilk ham verilerden modelleme araç/araçlarına, beslenecek verileri oluşturmaya yönelik tüm faaliyetleri kapsar. Görevler arasında tablo, kayıt ve öznitelik seçiminin yanı sıra modelleme araçları için verilerin dönüştürülmesi ve temizlenmesi yer alır [23].

Verileri işlemeye uygun hale getirmek için, onları çok boyutlu, zaman serileri veya yarı yapılandırılmış biçim gibi veri madenciliği algoritmalarına uygun bir biçime dönüştürmek önemlidir. Çok boyutlu biçim, verilerin farklı alanlarının öznitelik veya boyut olarak adlandırılan farklı ölçüm özelliklerine karşılık geldiği en yaygın olan biçimdir. Veri madenciliği süreci için ilgili öznitelikleri çıkarmak çok önemlidir. Öznitelik çıkarma aşaması, genellikle verilerin eksik ve hatalı kısımlarının tahmin edildiği veya düzeltildiği veri temizlemeye paralel olarak gerçekleştirilir. Çoğu durumda, veriler birden çok kaynaktan çıkarılabilir ve işlem için birleşik bir biçime entegre edilmesi gerekebilir. Bu prosedürün nihai sonucu, bir bilgisayar programı tarafından etkin bir şekilde kullanılabilen, yapılandırılmış bir veri setidir. Öznitelik çıkarma aşamasından sonra, veriler yeniden işlenmek üzere bir veri tabanında depolanabilir [17].

Süreç içerisinde yapılacak işlemler, uygulamacının bakış açısına bağlıdır. Veri kümesi üzerinde yapılan müdahaleler her algoritmada farklı neticelerle sonuçlanabilir. Yapılacak çalışmanın iyi sonuçlar üretmesi uygulamacının uygulama alanı hakkında bilgili olmasını ya da bu alan uzmanlarıyla birlikte çalışmasını gerektirir [24].

- Öznitelik çıkarımı (feature extraction), bir analist, verileri işlemek üzere çok sayıda ham belge, sistem günlükleri veya ticari işlemlerle karşı karşıya kalabilir. Bu aşama, belirli bir uygulama ile en alakalı öznitelikleri çıkarabilmesi için analiste büyük ölçüde

bağımlıdır. Örneğin, bir kredi kartı dolandırıcılık tespit uygulamasında, işlem miktarı, tekrar sıklığı ve konumu genellikle dolandırıcılığın iyi göstergeleridir. Bununla birlikte, diğer birçok özellik sahtekarlığın daha zayıf göstergeleri olabilir. Bu nedenle, doğru öznitelikleri çıkarmak genellikle eldeki belirli uygulama alanının anlaşılmasını gerektiren bir beceridir [17].

Öznitelik çıkarımı aşamasında birbiri ile ilişkili olan veya anlamsız olan değişkenlerin elenmesine dikkat edilmelidir. Amaç bilgi çıkarımı olduğundan ve birbiri ile ilişkili olan değişkenler ekstra bilgi vermediğinden, diğerine göre daha anlamlı olan değişkeni modele katmak analiz için daha faydalı olacaktır [27].

- Veri temizleme, çıkarılan verilerde hatalı veya eksik girişler olabilir. Bu nedenle, bazı kayıtların atılması, eksik girişlerin tahmin edilmesi veya tutarsızlıkların giderilmesi gerekebilir [17].

Uygulamalar için en zor görev, verileri analiz edilebilecekleri standart bir forma sokmak olabilir. Veriler standart biçimde olsa bile hatasız olduğu varsayılamaz. Gerçek dünya veri kümelerinde, ölçüm hataları, öznel yargılar ve otomatik kayıt ekipmanının hatalı çalışması veya kötüye kullanılması gibi çeşitli nedenlerle hatalı değerler kaydedilebilir. Hatalı değerler, özniteliğin olası değerleri olan ve olmayanlar olarak bölünebilir [22].

Kullanımda değişiklik gösteren gürültü terimi, bir değeri veri kümesi için geçerli olan ancak yanlış kaydedilmiş bir değeri ifade etmek şeklide tanımlanmıştır. Örneğin “69.72” sayısı yanlışlıkla “6.972” olarak girilebilir veya “kahverengi” gibi kategorik bir özellik değeri yanlışlıkla “mavi” gibi olası değerlerden biri olarak kaydedilebilir. Bu tür gürültü, gerçek dünya verileriyle ilgili kalıcı bir sorundur. “6.972” için “69.7X” veya “kahverengi” için “kkahverengi” gibi veri kümesi için geçersiz olan gürültülü değerlerde çok daha küçük bir sorun ortaya çıkar. Bunlar gürültü değil geçersiz değerler olarak kabul edilir. Geçersiz bir değer kolayca tespit edilebilir ve düzeltilebilir veya reddedilebilir [22].

- Öznitelik seçimi ve dönüşümü (feature selection and transformation), veriler çok yüksek boyutlu olduğunda, birçok veri madenciliği algoritması etkili bir şekilde çalışmaz. Ayrıca, yüksek boyutlu özniteliklerin çoğu gürültülüdür ve veri madenciliği sürecine hatalar ekleyebilir. Bu nedenle, ilgisiz öznitelikleri kaldırmak veya mevcut öznitelik kümesini analiz için daha uygun olan yeni bir veri alanına dönüştürmek için çeşitli yöntemler kullanılır. Bir başka ilgili husus, belirli bir öznitelik kümesine sahip bir veri kümesinin, aynı veya farklı türden başka bir öznitelik kümesine sahip bir veri kümesine

dönüştürülebildiği veri dönüştürmesidir. Örneğin yaş gibi bir öznitelik, analitik uygunluk için ayrı değerler oluşturmak üzere aralıklara bölünebilir [17].

1.4.1.4. Model Oluşturma

Verileri hazırlandıktan sonra, verilerden etkili analitik yöntemler tasarlamaktır. Çoğu durumda, eldeki uygulama için standart bir veri madenciliği problemini doğrudan kullanmak mümkün olmayabilir [17].

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir [16, 25].

Witten ve Frank [26], model kuruluş sürecinin denetimli ve denetimsiz öğrenimin kullanıldığı modellere göre farklılık gösterdiğini belirtmektedir. Modelleme adımları, modelleme tekniğinin seçimini, test tasarımının oluşturulmasını, modellerin oluşturulmasını ve modellerin değerlendirilmesini içerir [23].

- Test tasarımı oluşturma, bir model oluşturulduktan sonra, veri analisti modelin gücünü belirlemek için deneysel testler yaparak modelin kalitesini ve geçerliliğini test etmelidir. Sınıflandırma gibi denetimli veri madenciliği görevlerinde, veri madenciliği modelleri için kalite ölçütü olarak hata oranlarının kullanılması yaygındır. Bu nedenle, tipik olarak veri seti eğitim ve test seti olarak ikiye ayrılır. Model eğitim seti üzerinde kurulur ve kalitesi ayrı test setinde tahmin edilir. Başka bir ifadeyle, veri analisti modeli mevcut bir dizi veriye dayalı olarak geliştirir ve ayrı bir veri seti kullanarak geçerliliğini test eder. Bu yaklaşım, veri analistinin, veriyi geleceği tahmin etmek için kullanmadan önce modelin geçmişi ne kadar iyi tahmin edebileceğini ölçmesini sağlar. Modeli oluşturmadan önce test prosedürünü tasarlamak genellikle uygundur; bunun da veri hazırlığı üzerinde etkileri vardır [23].

- Model oluşturma, testten sonra, veri analisti bir veya daha fazla model oluşturmak için modelleme aracını hazırlanan veri seti üzerinde çalıştırır. Bu aşamada, veri analisti de modelleri sıralamaya çalışır. Modelleri değerlendirme kriterlerine göre değerlendirir ve iş hedefleri ile iş başarı kriterlerini dikkate alır. Çoğu veri madenciliği projesinde, veri analisti, tek bir tekniği bir kereden fazla uygular veya farklı alternatif tekniklerle veri

madenciliği sonuçları üretir. Bu ürettiği sonuçları değerlendirme kriterlerine göre karşılaştırır [23].

1.4.1.5. Modelin Değerlendirilmesi

Veri analisti tarafından oluşturulan modelin nihai dağıtımına geçmeden önce, iş hedeflerine uygun şekilde çalıştığından emin olmak için model daha kapsamlı bir şekilde değerlendirilir ve modelin yapısı gözden geçirilir. Bu aşamada bazı önemli iş konularının yeterince dikkate alınıp alınmadığını belirlemek çok önemlidir [23].

Modelin değerlendirilmesi aşamasında tespit edilecek başarısızlıkta, sonuçları iyileştirmek için süreç veri anlama aşamasına geri dönecektir [16]. Bu aşamanın sonunda, proje lideri veri madenciliği sonuçlarının nasıl kullanılacağına tam olarak karar vermelidir. Buradaki kilit adımlar, sonuçların değerlendirilmesi, sürecin gözden geçirilmesi ve sonraki adımların belirlenmesidir [23].

1.4.1.6. Modelin Kullanılması

Model oluşturma genellikle projenin sonu değildir. Edinilen bilgi, müşterinin kullanabileceği şekilde düzenlenmeli ve sunulmalıdır; bu, web sayfalarının gerçek zamanlı kişiselleştirilmesi veya pazarlama veritabanlarının tekrar tekrar puanlanması gibi, genellikle bir kuruluşun karar verme süreçlerinde “canlı” modellerin uygulanmasını içerir [23].

Gereksinimlere bağlı olarak, modelin kullanım aşaması, bir rapor oluşturmak kadar basit veya doğrudan bir uygulama olabileceği gibi kuruluş genelinde tekrarlanabilir bir veri madenciliği sürecini uygulamak kadar karmaşık olabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmini envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otonom olarak sipariş verilmesini sağlayacak veri madenciliği uygulamalarının içine gömülebilir. Kullanım adımlarını gerçekleştiren genellikle veri analisti değil müşteri olsa da müşterinin oluşturulan modellerden fiilen yararlanmak için hangi eylemlerin yapılması gerektiğini önceden anlaması önemlidir [23, 28].

Buradaki kilit adımlar plan dağıtımı, plan izleme ve bakım, nihai raporun üretilmesi ve projenin gözden geçirilmesidir [23].

- Dağıtımı planla, veri madenciliği sonuçlarını işletmeye yerleştirmek için bu görev, değerlendirme sonuçlarını alır ve dağıtım için bir strateji geliştirir [23].

- Plan izleme ve bakım, veri madenciliği sonucunun günlük işin ve çevresinin bir parçası olması gerekiyorsa, izleme ve bakım önemli konulardır. Dikkatle hazırlanmış bir bakım stratejisi, veri madenciliği sonuçlarının yanlış kullanımını önler [23].

- Nihai rapor hazırlama, projenin sonunda, proje lideri ve ekibi bir final raporu yazar. Dağıtım planına bağlı olarak, bu rapor sadece projenin ve deneyimlerinin bir özeti olabilir (henüz devam eden bir faaliyet olarak belgelenmemişlerse) veya veri madenciliği sonuçlarının nihai ve kapsamlı bir sunumu olabilir. Bu rapor, önceki tüm çıktıları içerir ve sonuçları özetler ve düzenler. Ayrıca, projenin sonunda genellikle sonuçların müşteriye sözlü olarak sunulduğu bir toplantı olacaktır [23].

- Projeyi inceleme, zamanla bütün sistemlerin özelliklerindeki değişiklikler ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin devamlı bir şekilde izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirmektedir. [28]. Veri analisti, gelecekteki projelerde kullanılmak üzere potansiyel iyileştirme alanlarının yanı sıra başarısızlıkları ve başarıları da değerlendirmelidir. Bu adım, proje sırasındaki önemli deneyimlerin bir özetini içermeli ve önemli proje katılımcıları ile yapılan görüşmeleri içerebilir. Bu belge, tuzaklar, yanıltıcı yaklaşımlar veya benzer durumlarda en uygun veri madenciliği tekniklerini seçmeye yönelik ipuçlarını içerebilir. İdeal projelerde, deneyim belgeleri, proje aşamaları ve görevleri sırasında bireysel proje üyeleri tarafından yazılan raporları da kapsar [23].

1.4.2. Veri Madenciliği Teknikleri

Veri madenciliğinde kullanılan modeller, tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki ana başlık altında incelenmiştir. Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçlan bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır [29].

Tanımlayıcı modeller verilerdeki örüntü veya ilişkileri tanımlarlar. Bu modeller tahmin edici modellerin aksine analiz edilen verilerin özniteliklerini incelemek için

kullanılan modellerdir. Örnek olarak sigorta poliçesini yenilememiş müşterilerin benzer özniteliklerini belirleyecek bir kümeleme çalışması verilebilir. Kümeleme, özetleme, birliktelik kuralları, sıra örüntüleri keşfi modelleri tanımlayıcı modeller olarak nitelendirilir [30].

Veri madenciliği modellerini gördükleri işlevlere göre,

1- Sınıflama (Classification) ve Regresyon (Regression)
2- Kümeleme (Clustering)
3- Birliktelik Kuralları (Association Rules)

olmak üzere üç ana başlık altında incelenmiştir [31].

Sınıflama ve regresyon modelleri tahmin edici, kümeleme ve birliktelik kuralları modelleri tanımlayıcı modellerdir [31].

1.4.2.1. Sınıflandırma ve Regresyon

Sınıflandırma, veri madenciliği için en yaygın uygulamalardandır. Günlük yaşamda sıklıkla meydana gelen bir göreve karşılık gelebilir. Örneğin, bir hastane, tıbbi hastaları belirli bir hastalığa yakalanma riski yüksek, orta veya düşük olanlar olarak sınıflandırmak isteyebilir, bir kamuoyu araştırması şirketi, görüşülen kişileri bir dizi siyasi partiye oy verme olasılığı yüksek olan veya kararsız olanlar olarak sınıflandırmak isteyebilir veya bir öğrenci projesini, başarılı veya başarısız olarak sınıflandırmak isteyebilir [22].

Belirli bir veri kümesinden yapılan çıkarımlarla görünmeyen örnekleri bir sınıfa atayan herhangi bir algoritmaya sınıflandırıcı denilmektedir. Bir veri kümesinden birçok farklı sınıflandırıcı üretilebilir. Her bir sınıflandırıcının bir dizi görünmeyen örnekte farklı başarım göstermesi muhtemeldir. Sınıflandırıcıların başarımlarını tahmin etmek için kullanılacak en belirgin kriter, tahmine dayalı doğruluktur. Tahmine dayalı doğruluk, doğru sınıflandırdığı tahminlerin toplam veri kümesine oranı ile hesaplanmaktır. Bu genellikle en önemli kriter olarak görülür, ancak sınıfların ciddi şekilde dengesiz olduğu durumlarda, tahmin doğruluğu kendi başına bir sınıflandırıcının etkinliğinin güvenilir bir göstergesi değildir. Farklı olarak duyarlılık ve kesinlik gibi sınıflandırıcıların başarımlarını tahmin eden başka tahmin edicilerde vardır. Bazıları diğerlerinden çok daha fazla hesaplama veya bellek gerektirir. Bazıları güvenilir sonuçlar vermek için önemli sayıda eğitim örneği gerektirir. Duruma bağlı olarak veri analisti, çalışma süresini/bellek

gereksinimlerini ve/veya ihtiyaç duyulan eğitim örneklerinin sayısını azaltmak için en uygun tahmin ediciyi tercih etmelidir [22].

Regresyon analizi, bir bağımlı değişkenin başka bağımsız değişkenlere olan bağımlılığını, bağımlı değişkenin ana kütle ortalama değerini, bağımsız değişkenin yinelenen örneklerdeki bilinen ya da değişmeyen değerleri cinsinden tahmin etme ve/veya kestirme amacı ile çalışır [32]. Berry ve Linoff [33], tamamı sürekli değişkenlerden oluşan veri kümeleri için regresyonun muhtemelen en iyi yöntem olduğunu söylemişlerdir. Regresyon, doğrusal regresyon ve lojistik regresyon olmak üzere iki çeşittir. Lojistik regresyonda ikili lojistik regresyon ve çoklu lojistik regresyon olarak ikiye ayrılmaktadır.

Regresyon analizi, durum değerlerini en az hatayla sonuç değeriyle ilişkilendiren ek bir işlevi tanımlar. Regresyon modelleri, vakaları veya nesneleri tanımlayan değişkenler arasındaki etkileşimler gibi doğrusal olmayanları yansıtacak şekilde tasarlanabilir. Model elde edildikten sonra, yeni nesneler için karşılık gelen sınıfı bulmak için kullanılabilir [34].

Lojistik regresyon, sayısal ölçeklerle bir Sınıflandırma problemine geleneksel bir yaklaşımdır. Regresyon analizi ile aynı varsayımlara sahiptir, ancak sınıflar için ayrık değerlerin kullanılmasına izin verir [34]. Lojistik regresyon, diskriminant analizi ve sinir ağları, tüm niteliklerin sürekli ölçeklerle sunulduğu durumlarda en iyi şekilde kullanılırken, karar ağaçları genellikle sürekli ve kategorik verilerle aynı anda ilgilenebilir.

1.4.2.2. Kümeleme

Kümeleme, birbirine benzer ve diğer kümelere ait nesnelere benzemeyen nesneleri bir arada gruplamakla ilgilidir. Kümeleme analizlerinde nesnelerin önceden belirlenmiş bir kritere göre gruplandırması yapılmakta olup bu sebeple denetimsiz bir öğrenme yaklaşımıdır [22, 34].

Kümeleme algoritmaları, benzer öğe gruplarını bulmak için verileri inceler. Örneğin, bir sigorta şirketi müşterileri gelire, yaşa, satın alınan poliçe türlerine veya önceki hasar deneyimine göre gruplayabilir. Bir arıza teşhis uygulamasında, elektrik arızaları belirli anahtar değişkenlerin değerlerine göre gruplandırılabilir. Bir finansal uygulamada, benzer finansal performansa sahip şirket kümelerini bulunabilir. Bir pazarlama uygulamasında, benzer satın alma davranışına sahip müşteri grupları bulmaktan, tıbbi bir uygulamada, benzer semptomları olan hasta gruplamaya kadar geniş uygulama alanına sahiptir [22].

Kümeleme, küme içerisindeki gözlemlerin birbirine benzer oldukları anlamına gelirken diğer kümelerdeki gözlemlerden farklı oldukları anlamına da gelmektedir. Bunun için benzerlik ve uzaklık ölçüleri geliştirilmiştir. Benzerlik nicel olup, iki gözlem arasındaki ilişkinin kuvvetini ortaya koyarak gözlemlerin ayırt edilmesini sağlamaktadır. Bu nicel değer -1 ile 1 arasında ölçeklendirilir. Uzaklık, farklılıkları ölçer. Farklılıklarda çeşitli özelliklere dayalı olarak iki nesne arasındaki zıtlık ya da uyumsuzlukların bir ölçümüdür. Bu sayede gözlemler benzerlik veya farklılıklarına dayalı olarak kümelendirilmektedir. Kümeleme için hesaplanan uzaklık değerlerinden yararlanarak gözlemlerin kümelere atanması gerçekleştirilmektedir. En çok bilinen uzaklık ölçüleri;

1. Minkowski Uzaklığı,
2. Öklid (Euclide) Uzaklığı,
3. Pearson Uzaklığı,
4. Manhattan (City-Blok) Uzaklığı,
5. Mahalanobis Uzaklığı,
6. Hotelling T2 Uzaklığı,
7. Canberra Uzaklığı, olarak verilmektedir [84, 85].

Kümelemenin birçok yöntemi vardır. En yaygın kullanılan ikisi olan k-ortalamalar kümeleme ve hiyerarşik kümelemedir. k-ortalamalar kümeleme, özel bir kümeleme algoritmasıdır. Her nesne kesin olarak bir küme grubundan birine atanır. Hiyerarşik kümelemenin arkasındaki fikir basittir. Her nesneyle kendi kümesini oluşturarak başlar. Ardından her şeyi içeren tek bir kümeyle sonuçlanana kadar en yakın küme çiftini tekrar tekrar birleştirir. Nesneleri kümelemek için mesafeye dayalı bir kümeleme algoritması kullanmadan önce, iki nokta arasındaki mesafeyi ölçmenin bir yoluna karar vermek gerekir [22].

1.4.2.3. Birliktelik Kuralları

Birliktelik kuralları, büyük veri kümeleri arasında kurallar biçiminde birliktelik ilişkileri bulurlar. Verilen herhangi bir veri setinden türetilebilen birçok olası birliktelik kuralı vardır. Bu değişkenlerinin çoğu arasında ilişki düzeyi düşüktür veya hiç değeri yoktur. Bu nedenle birliktelik kurallarının ne kadar güvenilir olduklarını gösteren bazı ek bilgilerle belirtilmesi olağandır. Bu birliktelik kurallarını belirlemek için bazen bir eğitim seti kullanılabilir [22, 35].

Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır. Bu işlem, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını analiz eder [37].

Örneğin market sepeti uygulamasında, bir mağazada tüm müşterilerin yaptığı satın alma işlemleri (bir hafta boyunca) bilindiğinde, mağazanın gelecekte ürünlerini daha etkili bir şekilde pazarlamasına yardımcı olacak ilişkiler bulunabilir. Değişkenler arasındaki ilişkiler ile tezgahların pozisyonlarına dahi karar verilebilir [22].

1.5. Metin Madenciliği

Metin madenciliği, büyük metin koleksiyonlarındaki bilgileri keşfetme ve metin verilerindeki ilginç kalıpları ve ilişkileri otonom olarak tanımlama sürecidir. Esas olarak Web'de ve başka yerlerde sürekli artan bilgi miktarı nedeniyle son zamanlarda araştırma ve endüstri toplulukları arasında büyük ilgi uyandırmıştır. Metin madenciliği, veri madenciliği, doğal dil işleme (DDİ), makine öğrenimi ve bilgi erişimi alanlarında araştırma anlayışlarını bir araya getiren, disiplinler arası bir araştırma alanıdır. Metin madenciliğinde kaynak veriler yapılandırılmış veri tabanlarından ziyade yapılandırılmamış belge koleksiyonlarından oluşur [43].

Metin verilerini sınıflandırmak, artan basılı materyaller ve uzmanlık alanlarındaki bilgi bulmanın zorlaşması nedeniyle giderek daha önemli hale gelmektedir [22, 44]. Kitaplıklarda ve dijital ayarlarda depolanan verilerin çoğu yapılandırılmamış metinler, ses ve resim dosyalarıdır. Bu tür verilerin araştırmacılar için bilgi olarak kullanılması araştırma yöntemleri açısından çok önemlidir. Bu verilerin doğru ve sistematik bir şekilde analiz edilebilmesi için teknik ve yazılımlara ihtiyaç vardır [5].

Dijital ortamlarda dil, ses ve görsel formlarda saklanan ve işlenmeye hazır yapılandırılmamış verilerle ilgilenen Metin madenciliği, veri madenciliğinin bir alt bölümüdür. Metin madenciliği ve veri madenciliği arasında etkileşimli bir ilişki vardır. Metin madenciliğinden elde edilen yapılandırılmamış veriler, veri madenciliği modelleri kullanılarak değerlendirilir ve bulgular metinsel yapıyı analiz etmek için kullanılır [5].

Metin sınıflandırmasını diğer sınıflandırma görevlerinden ayıran önemli bir konu, çoklu sınıflandırma olasılığıdır. Sınıflandırma yaparken, birbirini dışlayan bir dizi kategori olduğunu ve her nesnenin kaçınılmaz olarak bunlardan birine ve yalnızca birine uyması gerektiği varsayımı metin sınıflandırmasında oldukça farklıdır. Tıp, İşletme, Finans, Tarih,

Biyografi, Yönetim ve Eğitim gibi N kategori olabilir. Bir belgenin bu kategorilerin birkaçına, hatta muhtemelen hepsine veya hiçbirinde yer almaması tamamen mümkündür [22].

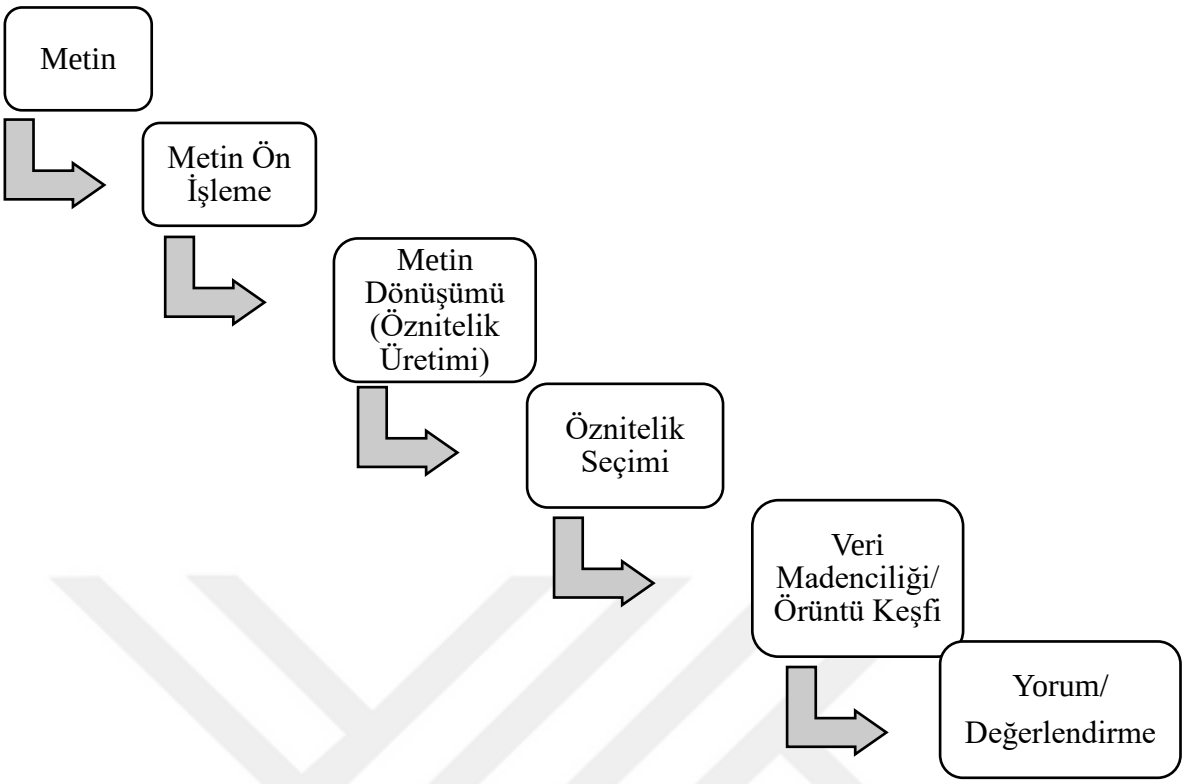
Metin belgelerinde sınıflandırmak için standart yöntemlerden birisi kullanılabilirken (Naive Bayes, En Yakın Komşu, Karar Ağaçları vb.) kümeleme yöntemleri standardın dışında ayrı bir açıklama gerektirir [22]. Metin kümeleme, doküman koleksiyonlarının birbirlerine olan benzerliklerine bağlı olarak gruplara veya kümelere ayrıştırılmasıdır. İşlem sonucunda her küme içerisinde yer alan dokümanların benzer bir konuda olmaları beklenir [45, 46]. Metin kümelemesi için oluşturulan hipotez, birbiriyle ilgili dokümanlar ilgisizlere göre birbirine daha çok benzerdir şeklindedir. Bu hipoteze dayanarak, metin kümeleme sayesinde geniş doküman koleksiyonları içerisinde dolaşım ve arama işlemlerinin iyileştirilmesi öngörülmektedir [43, 46].

1.5.1. Metin Madenciliği Adımları

Metin madenciliği adımları da veri madenciliğine benzer şekilde, metin veri seti oluşturmak ile başlar. Metin ön işleme, metin dönüşümü – öznitelik üretimi, öznitelik seçimi, veri madenciliği işlemlerinin uygulanması ve sonunda sonuçların yorum ve değerlendirme aşaması ile son bulur. Şekil 3'te veri madenciliğinin şamaları şematik olarak gösterilmiştir.

1.5.1.1. Veri Seçimi

Bilgisayarlarda kayıtlı dosyalar, e-posta kutularındaki mailler, bir şirketin her ay sonunda hazırladığı raporlar, forum sitelerinde yazılan yazılar, hastaların tahlil sonuçları, metin verisi için çeşitli alanlardan birkaçıdır. Konuya göre çeşitli platformlardaki metin içeriklerinin derlenme aşamasıdır.



Şekil 3. Metin madenciliği adımları

1.5.1.2. Metin Madenciliğinde Ön İşleme

Veri madenciliğinde olduğu gibi metin madenciliğinde de analiz edilecek verilerin belirli bir biçime sahip olması ayrıca bozuk veya gereksiz verilerden temizlenmiş olması gerekmektedir. Metin madenciliğinin en büyük problemi, yapısal olmayan veri kümesinin işlenecek olmasıdır. Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışılan metin madenciliği alanında ön işleme aşaması, veri temizlemenin yanında veriyi uygun biçime getirme işlemini de gerçekleştirmektedir [47].

Veri madenciliği için metin dokümanlarının temsil edilmesi (Bag of Words): Metin madenciliği için veri kümesi genellikle belgelerin kendisinden oluşur. Öznitelikler, sınıflandırma algoritması uygulanmadan önce içeriklerine göre belgelerden otonom olarak çıkarılır. Genelde çok fazla sayıda öznitelik vardır, bunların çoğu nadiren meydana gelir ve yüksek oranda gürültülü ve ilgisiz öznitelikler vardır [22].

Belgelerin düz metinden sabit sayıda özniteliğe sahip örneklere dönüştürülmesinin birkaç yolu vardır. Örneğin, belirtilen tümceciklerin kaç kez geçtiğini veya iki ardışık sözcüğün herhangi bir kombinasyonu sayılabilir. İki veya üç karakter kombinasyonunun

(sırasıyla bigram ve trigramlar) oluşumu da sayılabilir. Kelime torbası temsili ile bir belge, içinde en az bir kez geçen kelimelerin bir toplamı olarak kabul edilir. Kelimelerin sırası, oluştukları kombinasyonlar, paragraf yapılanması, noktalama işaretleri ve tabii ki kelimelerin anlamları göz ardı edilir [22].

Birden fazla sınıflandırmalarda, eğitim dokümanları koleksiyonu için kelime sözlüğü oluşturmanın iki yolu vardır. Birincisi yerel sözlük yaklaşımı ikincisi genel bir sözlük oluşturmaktır. Yerel sözlük yaklaşımı kullanmanın, genel bir sözlük kullanmaktan daha iyi performans sağlama eğiliminde olduğunu gösteren bazı kanıtlar vardır [22].

Yerel sözlükte her kategori için farklı bir sözlük oluşturulur. Sadece o kategoride sınıflandırılan belgelerde görünen kelimeler kullanılır. Böylece, her bir sözlüğün, N kategorinin olduğu N tane oluşturmaya ihtiyaç duyma pahasına nispeten küçük olmasını sağlar. İkinci bir yaklaşım olarak genel sözlük kullanımı, belgelerin herhangi birinde en az bir kez geçen tüm kelimeleri içeren genel bir sözlük oluşturmaktır. Bu daha sonra N kategorilerinin her birine sınıflandırma için kullanılır. Küresel bir sözlük oluşturmak, açık bir şekilde N yerel sözlük oluşturmaktan çok daha hızlı olacaktır, ancak bu, kategorilerin her birini sınıflandırmak için kullanmak üzere daha da fazla bir sunum yapma pahasına olacaktır.

Etkisiz kelimelerin çıkartılması (Stop-Words): Kelime torbası yaklaşımında birçoğu öğrenme görevi için önemli olmayan on binlerce farklı kelimenin yer alması mümkündür. Kullanılması halinde performansı önemli ölçüde düşürebilir. Öznitelik alanının boyutunu (yani sözlüğe dahil edilen sözcük grubunu) mümkün olduğunca azaltmak performans için zorunludur [22].

Gereksiz kelimelerden kurtulmanın yaygın yolu, yaygın kullanılan bu kelimelerin bir listesini oluşturup kelime torbası oluşturmadan önce listedeki kelimeleri kaldırmaktır. Evrensel olarak kullanılan kesin bir etkisiz kelime listesi yoktur. Liste dilden dile farklı olacaktır. İngilizcede ‘a’, ‘an’, ‘the’, ‘is’, ‘I’, ‘you’ ve ‘of’ gibi bazı bariz kelimeler vardır. Ayrıca ilgilenilen alana göre çok sık geçen ve başarımı düşürdüğü belirlenen kelimeler de belirlenip etkisiz kelime listesine dahil edilerek oluşturulan kelime torbasından çıkarılmaktadır.

Bu tür kelimelerin sıklığını ve dağılımını incelemek, bir dizi olası yazardan hangisinin bir roman veya bir oyun yazdığına karar vermek gibi üslup analizi için yararlı olurken, bir belgeyi Tıp, Finans vb. gibi konulara ayırmak için ise yararlı değildirler [22].

Kök bulma (Stemming): Kök bulma, belgelerdeki kelimelerin çoğu kez birçok morfolojik değişkene sahip olduğu gözlemine dayanmaktadır. Çoğu durumda, bu varyantlar benzer semantik yorumlara sahiptir ve bilgiye erişim (BE) uygulamaları için eşdeğer olarak ele alınabilir. Dolayısıyla kök bulma, aynı dil kökenine sahip kelimelerin tespiti ve tek bir kelime olarak indirgemek için kullanılmaktadır. Kök bulmanın kullanılması, bir kelime torbası temsilindeki kelime sayısına nispeten yönetilebilir bir sayıya düşürmenin çok etkili bir yolu olmaktadır. Bu nedenle, morfolojik varyantları kök biçimlerine indirgeyerek BE için bir dizi köklendirme veya birleştirme algoritması geliştirilmiştir [22, 48].

Kelimeleri türlerine ayırma (Lemmatizer): Kelime türlerine ayırma ile her kelime, cümle içerisindeki uygun bölümünü gösteren isim, sıfat, fiil, zamir, edat gibi bir etiketle etiketlenir. Duygu analizinde sıfatların önemli görüşlerin göstergesi olduğu birçok araştırmada bulunmuştur. Böylece kelimeler isim, sıfat, fiil, zamir olarak ayrıştırılarak özel nitelikler olarak ele alınmıştır [49]. Kelimelerin türlerine ayrılması, ayrıştırma ve makine çevirisi dahil olmak üzere birçok DDİ görevi için önemlidir [7].

1.5.1.3. Metin Dönüşümü – Öznitelik Üretimi

Vektör uzay modeli: Değerleri N boyutlu vektörler olarak yazmak, veri madenciliği dalındaki verilere bakmanın geleneksel bir yoludur [22].

Vektör uzay modelinde her nesne, vektör yapısında tanımlanmaktadır. Nesnelerin sahip olduğu farklı öznitelikler, vektör uzayının eksenlerini oluşturmakta ve belgelerin içeriğinde yer alan terimler sahip olduğu özniteliklere göre vektör uzayında belli bir koordinata sahip olmaktadır. Vektör uzaydaki koordinatları bu terimlerin ağırlık değerleri ile ilişkilidir [47,50].

Ağırlıklandırma: Kelimenin doküman içerisindeki frekansını hesaplamak o dokümanın sınıflandırılmasında ve yapısal olmayan verinin yapısallaştırılmasında önemlidir [47, 51]. Eğer bir sözcük dokümanlarda sık geçiyorsa, o doküman için belirleyici olmadığı düşünülebilir. Eğer sözcük dokümanlarda çok sık geçmiyorsa o sözcüğün o doküman için belirleyici özelliği olduğu kabul edilebilir [47].

Ağırlıkları hesaplamanın yaygın bir yolu, frekansa göre ağırlıklandırmadır. Verilen belgedeki her bir terimin kaç defa kullanıldığını sayarak ağırlıklandırma yapılır. Diğer bir yol bitsel ağırlıklandırmadır. Anahtar sözcük sözlüğünde yer alan sözcüklerin metinde yer

alıp almadığını gösteren vektörel bir gösterge oluşturulmaktadır. Belgede, 1'in terimin varlığını ve 0'ın terimin yokluğunu gösterdiği bir ikili gösterim kullanmaktır. Ne kadar karmaşıkta olsa diğer yöntemlere göre gelişmiş performansa sahip olan Terim Frekansı Ters Belge Frekansı (TF-TBF) yöntemi de ağırlıklandırma için kullanılır [22, 47].

Herhangi bir kelimenin ağırlığının TF-TBF değeri, sırasıyla sıklık terimine ve ters belge sıklığına karşılık gelen iki değerin çarpımı olarak hesaplanır [22].

TF-TBF ağırlıklandırmasında her bir dokümandaki sözcüklerin frekansı rol oynamaktadır. Böylece dokümanda daha fazla geçen TF o doküman için daha değerli olmaktadır [47].

TBF, sistemdeki tüm belgelerin sayısının (N), bir kelimenin en az bir defa geçtiği belge sayısına (n_k) oranının logaritması alınarak hesaplanmaktadır [46].

$$tbf = \log\left(\frac{N}{n_k}\right) \tag{1}$$

Bu değer, tüm belgeler içerisinde yer alan ve ayırt ediciliği fazla olan kelimeler ile belgelerin içeriğinde sıklıkla yer alıp ayırt ediciliği fazla olmayan kelimelerin belirlenmesinde kullanılmaktadır [47, 50].

Bir belgedeki bir kelimenin ağırlık değerleri (w_{ik}), bu kelime için hesaplanmış TF*TBF değerinin, belgelerindeki her bir kelime için hesaplanan TF*TBF değerlerinin karelerinin toplamının karekökünе oranı ile hesaplanmaktadır [52].

$$w_{ik} = \frac{tf_{ik} * tbf}{\sqrt{\sum_{j=1}^{v} (tf_{ij})^2 * \left(\log\left(\frac{N}{n_j}\right)\right)^2}} \tag{2}$$

Burada w_{ik}, i. D belgesindeki k. terimin ağırlık değerini, tf_{ik}, k. terimin i. D belgesinde geçme sayısını (TF), N, belge sayısını, v: bir belgedeki toplam terim sayısını ifade etmektedir [52].

1.5.2. Metin Madenciliğinde Kullanılan Teknikler

Bilgisayarlara metinlerin nasıl analiz edileceğini, anlaşılacağını ve üretileceğini öğretmek için teknolojiler DDİ ile üretilir. Bilgi çıkarma, özetleme, sınıflandırma,

kümeleme ve bilgi görselleştirme gibi teknolojiler metin madenciliği sürecinde kullanılmaktadır [53].

1.5.2.1. Bilgiye Erişim (BE)

Büyük metin veri kümeleriyle etkileşim genellikle samanlık probleminde bir iğne olarak ortaya çıkar. Okuması on yıl sürecek belgelerle yüz yüze gelen kullanıcı, aradığı şeyle eşleşen bir veya birkaç belge arar. Bunlar e-posta, bir kavram veya bir sorunun cevabını en iyi temsil eden belge olabilir. Bilgi erişim disiplini, bu sorunu sistematik hale getirmek, çözmek ve değerlendirmek üzerine inşa edilmiştir [6].

BE, kullanıcının bilgi ihtiyacıyla ilgili belgeleri bulma görevidir. En iyi bilinen bilgi erişim sistemleri örnekleri, World Wide Web'deki arama motorlarıdır. Bir web kullanıcısı bir arama motoruna bir sorgu yazabilir ve ilgili sayfaların listesini görebilir [54].

Çoğu BE sistemi Boole modelini terk edip kelime sayımlarının istatistiklerine dayanan modelleri kullanmıştır. Bir puanlama işlevi bir belgeyi ve bir sorguyu alır ve sayısal bir puan döndürür; en ilgili belgeler en yüksek puana sahiptir. BM25 (Best Matching – En iyi eşleşme) fonksiyonunda skor, sorguyu oluşturan kelimelerin her biri için skorların doğrusal ağırlıklı kombinasyonudur. Bir sorgu teriminin ağırlığını üç faktör etkiler: İlk olarak, bir sorgu teriminin belgede görünme sıklığı. Buna TF olarak da bilinir. [Kansas'ta tarım] sorgusu için, sık sık “tarım” dan söz eden belgeler daha yüksek puanlara sahip olacaktır. İkinci olarak, terimin TBS. "içinde" kelimesi hemen hemen her belgede görünür, bu nedenle yüksek belge sıklığına ve dolayısıyla düşük TBF’na sahiptir ve bu nedenle sorgu için "tarım" veya "Kansas" kadar önemli değildir. Üçüncü olarak, belgenin uzunluğu. Milyon kelimelik bir belge muhtemelen tüm sorgu kelimelerinden bahsedecektir, ancak aslında sorgu ile ilgili olmayabilir. Tüm kelimelerden bahseden kısa bir belge çok daha iyi bir adaydır [55].

BE deneylerinin çoğu, ortalama kesinlik ve ortalama hatırlama (average precision and average recall) gibi ölçülerle değerlendirilir. Bir BE tekniğinin diğerine üstünlüğüne ilişkin temel kararlar, yalnızca bu ölçümler temelinde alınır. Ortalama performans rakamlarının dikkatli bir istatistiksel analizle doğrulanması gerektiğini ve bireysel sorguların sonuçlarına yakından bakılarak ortaya çıkarılabilecek çok sayıda ek bilgi olduğu ortaya atılmıştır [48].

Kelime analizleri dilbilimsel yaklaşımları doğrudan kök çıkarma algoritmaları olarak kullandığında problem teşkil etmektedir. Bu problemlerden birincisi, sözlüğe dayalı bu yöntemlerde, sözlükte yer almayan sözcükleri doğru bir şekilde türetemezler. İkinci olarak, dilsel perspektiften uygun şekilde motive edilmiş bir kelimenin kök biçimiyle ilgili birçok karar, bilgi erişim performansı için optimal değildir. Dilsel analiz araçlarının bilgi alma uygulamaları için uyarlanması gerektiği ve belirli bir alan için potansiyel olarak daha fazla optimizasyonun önemli olabileceği ortaya atılmıştır [48].

BE sisteminin rolü

- Bir problem alanının terminolojisini, kavramlarını ve yapısını araştırmak, keşfetmek, anlamak.
- Bir bilgi ihtiyacının sınıflandırılması, rafine edilmesi ve formüle edilmesi.
- Açıklama gerektiren bilgilerle eşleşen belgelerin keşfedilmesi [56].

1.5.2.2. Bilgi Çıkarımı

Bilgi Çıkarımı, bir metni gözden geçirip belirli bir nesne sınıfının oluşumlarını ve nesneler arasındaki ilişkilerini arayarak bilgi edinme işlemidir. Tipik bir görev, cadde, şehir, eyalet ve posta kodu için veritabanı alanları olan Web sayfalarından adres örneklerini çıkarmaktır veya sıcaklık, rüzgar hızı ve yağış alanlarına sahip hava raporlarından fırtına ihtimalini çıkarsamaktır. Sınırlı bir alanda, bu yüksek doğrulukla yapılabilir. Alan daha genel hale geldikçe, daha karmaşık dil modelleri ve daha karmaşık öğrenme teknikleri gereklidir. Bilgi çıkarmanın sınırlı ihtiyaçları için, tam İngilizce modeline yaklaşan sınırlı modeller tanımlanır ve sadece eldeki görev için gereken parçalara odaklanılır [55].

Bilgi çıkarımının çözdüğü problem, belirli bir belgeye özgü temel detayları özetleme şeklindedir. Bilgi çıkarımının rol oynadığı en büyük zorluk, metnin yönetimi, iletimi, depolanması veya görüntülenmesi yerine otonom yöntemlerin geliştirilmesidir. Bilgi çıkarımı, bir tür ucuz, yönlendirilmiş doğal dil anlayışı olarak kabul edilebilir. Bilgi çıkarımı, sınırlı bir söylem alanından, her belgenin diğer belgelerde açıklananlara benzer, ancak ayrıntılarda farklılık gösteren bir veya daha fazla varlığı veya olayı açıkladığı bir dizi belgenin varlığını varsaymaktadır [58].

Tablo 1'de iki metin madenciliği tekniği olan BE ve bilgi çıkarımının altı maddede karşılaştırılması yapılmıştır.

Tablo 1. Bilgiye erişim ve bilgi çıkarımı arasındaki farklar [56].

	Bilgiye Erişim	**Bilgi Çıkarımı**
1.	Bir kullanıcının bilgi ihtiyacıyla ilgili metin belgelerini bulma görevidir.	Amaç, belgelerden veya görüntü bilgilerinden önceden belirlenmiş öznitelikleri çıkarmaktır.
2.	Belge alma.	Öznitelik alma.
3.	Belgenin içine gömülü gerçek bilgilere erişim.	Belgenin içinden saklı bilgiyi alma.
4.	Uzun belge listesi.	Tüm koleksiyon boyunca toplama.
5.	Web için en iyi BE sistemi olan Google'ın ayrıntılarını açıklar.	Çıkarılan öznitelikler genellikle bir veritabanına otonom olarak girilir.
6.	Bir BE sisteminin çıktısı, kullanıcının sorgusuyla ilgili bir belge alt kümesidir.	Daha zordur çünkü bir belge hakkında daha ayrıntılı bilgi gerektirir. Genellikle öznitelikler arasında ilişki kurulmasını gerektirir.

1.6. Konu Modelleme

KM, dilin bilgisayar tarafından açık bir şekilde anlaşılması olmaksızın kullanışlı ve yorumlanabilir bir yapı ortaya çıkarmaktadır [4].

Bu strateji, Deerwester vd. [58] tarafından Tekil Değer Ayrışımı kullanılarak bu amaç için terim-belge matrislerinin ayrıştırılmasını öneren gizli anlamsal indeksleme üzerine yapılan ilk çalışma olan Gizli Anlamsal Analiz (GAA) adı verilen doğrusal bir cebir yaklaşımı ile başlamıştır [6].

KM, hedefin hangi konuların dikkate alındığını ve bir doküman içerisinde birbiriyle nasıl ilişkili olduğunu belirleyen teknikleri içeren, makine öğrenimi içinde bir çalışma alanı olarak ortaya çıkmıştır. Konular, belgelerle ilgili olan ve genellikle birden çok terimle ilişkilendirilen odak anlamsal konular olarak tanımlanır [59].

KM algoritmaları, orijinal metinlerin kelimelerini analiz ederek bunlarda geçen temaları, bu temaların birbirine nasıl bağlandığını ve zamanla nasıl değiştiklerini analiz eden istatistiksel yöntemlerdir. KM algoritmalarında konular, belgelerin önceden herhangi bir ek açıklamasını veya etiketlenmesini gerektirmeden doğrudan orijinal metinlerin analizinden ortaya çıkar. KM, elektronik arşivleri insan açıklamasının imkansız olacağı bir ölçekte düzenlenmesini ve özetlenmesini sağlamaktadır [60].

KM, istatistik ve makine öğrenimi araştırmalarına dayanarak, hiyerarşik olasılık modelleri kullanarak belge koleksiyonlarındaki kelime kalıplarını bulmak için geliştirilen tekniklere verilen addır [61].

KM'nin temelinde, kelime kullanım kalıplarını ve benzer kalıpları paylaşan belgelerin nasıl birbirine bağlanacağını keşfetmek vardır. Dolayısıyla, KM fikri, belgelerle çalışabilen bir terimdir. Bu belgeler, bir konunun kelimelere göre olasılık dağılımı olduğu konuların karışımlarıdır. Diğer bir deyişle, KM belgeler için üretken bir model olup belgelerin oluşturulabileceği basit bir olasılık prosedürünü ortaya koyar [62].

Keşfedilen modellerin genel kullanım alanı belgeleri oluştursa da görüntüler, biyolojik veriler ve anket verilerinin analizi gibi diğer veri türlerinede kolaylıkla genelleştirilmiştir (6).

Kelimeler, belgeler ve konularla ilgilenen KM yöntemleri; GAA, Olasılıksal Gizli Anlamsal Analiz (OGAA), Gizli Dirichlet Tahsisi (GDT) ve İlişkili Konu Modeli (İKM) olarak sayılabilir [62].

DDİ alanındaki bir yöntem olan GAA, metinlerin vektör tabanlı temsilini oluşturarak anlamsal içerik oluşturur. GAA ile vektör gösterimi, ilgili kelimeleri seçmek için metinler arasındaki benzerliği hesaplar. OGAA ise GAA'de bulunan bazı dezavantajları gidermek için yayınlanan bir yaklaşımdır. OGAA, sayım verilerinin faktör analizi için istatistiksel bir gizli sınıf modeline dayalı belge indekslemeyi otonomlaştırabilen bir yöntemdir. İKM, DDİ ve makine öğreniminde kullanılan bir tür istatistiksel modeldir. İKM, belge gruplarında gösterilen konular arası ilişkileri keşfeder ve keşif için lojistik normal dağılım kullanır [62].

Zaman geçtikçe, bir belge topluluğundaki konular gelişir, konuları zamanı dikkate almadan modellemek konu keşfini karıştırır. Konuları zamanı dikkate alarak modellemeye konu evrim modellemesi denir. Konu evrim modellemesi, metin dokümanındaki önemli gizli bilgileri açığa çıkarabilir. Aynı zamanda konuların zamanın görünümüyle tanımlanmasına ve zamanla gelişimlerinin kontrol edilmesine olanak tanır [62].

1.6.1. Gizli Dirichlet Tahsisi (Latent Dirichlet Allocation)

GDT, her belgedeki kelimelerin, her konunun sabit bir kelime dağarcığı yerine çok terimli olduğu, konuların karışımından ortaya çıktığını varsayar. Konular koleksiyondaki

tüm belgeler tarafından paylaşılır, ancak konu oranları, bir Dirichlet dağıtımından rasgele çekildiği için belgeler arasında stokastik olarak değişir [4].

En basit haliyle, sürecin temel fikri, her belgenin konuların bir karışımı olarak modellenmesidir. Her konu, her bir kelimenin belirli bir konuda ne kadar olası olduğunu tanımlayan ayrı bir olasılık dağılımıdır. Bu konu olasılıkları, bir belgenin kısa bir temsilini, özetini sağlar [62].

KM yöntemleri, bir metin veritabanında inceleme yaparken hem bir belge içerisinde geçen kelimeleri, hem de farklı belgelerde kelimelerin beraber kullanıldığı diğer kelimeleri inceleyerek, her belgedeki metnin bir veya daha fazla konuya ait olabileceğini ifade eden modelleri üreten yöntemlerdir. GDT, veritabanlarındaki metinsel verileri sınıflandırmak için kullanılabilen bir yöntemdir. Her belgede GDT algoritmasının belirlediği konular üzerine olasılık dağılımını bir öznitelik vektörü olarak kullanmaktır. GDT ile belge veritabanının özetini oluşturan ve en etkin kelimeleri içeren konular ifade edilebilmektedir [63].

GDT, metinsel verilere uygulanan ve sıklıkla kullanılan, birçok konu çıkarım modelinin temelini oluşturan ve yaygın olarak kullanılan, gözlemlenmemiş gruplar tarafından açıklanmasına izin veren istatistiksel bir model ve bir KM algoritmasıdır. Her konu, belgeler arasındaki benzerlikleri ve farklılıkları anlamak için gerekli olan ağırlıklı anahtar sözcüklerin bir listesidir. Bir veri ambarında saklanan metin dosyalarından bir metin araştırılırken kullanışlı bir tekniktir. Bu model ayrıca Twitter veya e-posta analizi için de kullanışlıdır [9, 16, 64].

En yaygın olarak kullanılan GDT algoritmasının, gramer ve yazım için gözden geçirilen, düzenlenen ve kontrol edilen gazete ve akademik dergi makaleleri gibi iyi yapılandırılmış metin belgelerinin analizine uygulandığı Guo tarafından belirtilmiştir [65].

GDT tekniğinin geliştiricilerine göre, “GDT, bir koleksiyonun her bir öğesinin temeldeki bir konu kümesi üzerinden sonlu bir karışım olarak modellendiği üç seviyeli hiyerarşik bir Bayes modelidir. Her konu, sırayla, temeldeki bir konu olasılıkları kümesi üzerinde sonsuz bir karışım olarak modellenmiştir. Metin modelleme bağlamında, konu olasılıkları bir belgenin açık bir temsilini sağlamaktadır” [66].

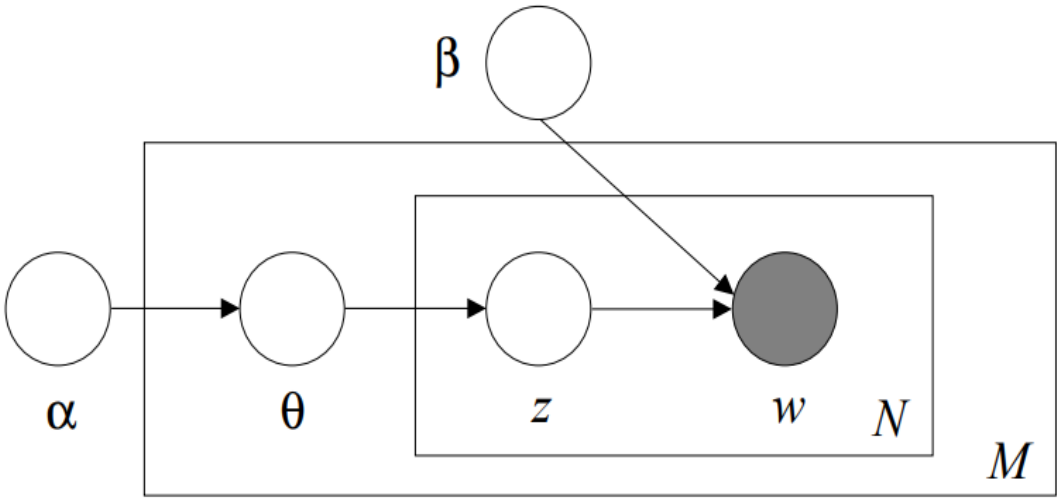
GDT üretken bir modeldir; bilimsel bir makale üretmenin karmaşık sürecini az sayıda basit olasılık adımına indirger ve böylece tüm olası belgeler üzerinde bir olasılık dağılımını belirtir. Bu üretken model, bir dizi konudan oluşan gizli bir yapıyı varsayar; her

belgeden konular üzerinden bir dağılım seçilerek ve daha sonra bu dağılım kullanılarak seçilen bir konudan rastgele olarak her kelime üretilerek üretilir [67].

Yakın geçmişte web, bilimsel açıdan ilginç bloglar, haber makaleleri ve literatür gibi çok sayıda elektronik belge koleksiyonu veri madenciliği topluluğundaki araştırmacılara yeni zorluklar getirmiştir. Özellikle bu belge koleksiyonlarını görselleştirmek, analiz etmek ve özetlemek için giderek artan otonom tekniklere ihtiyaç vardır. Yakın geçmişte, gizli konu modellemesi, büyük belge koleksiyonlarında konu keşfi için tamamen denetlenmemiş bir teknik olarak çok popüler hale gelmiştir [62].

Geçici metin madenciliği, yazar-konu analizi, denetimli konu modelleri, gizli dirichlet ortak kümelenmesi ve GDT tabanlı biyoinformatik dahil olmak üzere onlarca GDT tabanlı model vardır [62].

Bu model göz önüne alındığında, keyfi belgeler için konu dağılımları çıkarılabilir. Belgeleri konu kompozisyonları açısından açıklama fikri, bilgi yönetimi uygulamalarında geniş bir uygulama görmüştür. Örneğin, bir "elmalı turta" sorgusunda, GDT "turta" nın varlığından "elma" nın anlamının meyve kavramına bilgisayar kavramından daha yakın olduğu sonucuna varabilir. Bir GDT modelini öğrenerek elde edilen bu anlam bilgisini kullanarak, "meyve" anlamına gelen belgeler etkili bir şekilde tanımlanabilir ve "elmalı turta" sorgusunu yanıtlamak için iade edilebilir [68].



Şekil 4. GDT'nin olasılıksal model gösterimi

Şekil 4’te gösterildiği gibi, GDT temsilinin üç seviyesi vardır. α ve β parametreleri bir derlem oluşturma sürecinde bir kez örneklendiği varsayılan derlem düzeyindeki

parametrelerdir. θ_d değişkenleri, belge başına bir kez örneklenen belge düzeyindeki değişkenlerdir. Son olarak, z_{dn} ve w_{dn} değişkenleri kelime seviyesi değişkenleridir ve her belgedeki her kelime için bir kez örneklenmiştir [66].

GDT modelinin temel fikri, doküman içindeki belgelerin gizli konular üzerinde rastgele bir karışım olarak temsil edilmesi ve her bir konunun belgelerdeki kelimeler üzerinde bir dağılım ile karakterize edilmesidir [69].

GDT, bir D topluluğunda her N_i uzunluğundaki M belge için aşağıdaki üretken süreci varsayar:

1. $\theta_i \sim Dir(\alpha)$ seçin, burada $i \in \{1, \dots, M\}$ ve $Dir(\alpha)$, tipik olarak kesikli olan (α <1) simetrik bir α parametresine sahip bir Dirichlet dağılımıdır.

2. $\varphi_k \sim Dir(\beta)$ seçin, burada $k \in \{1, \dots, K\}$ ve β tipik kesikli dağılımdır.

3. i, j kelime pozisyonlarının her biri için, $i \in \{1, \dots, M\}$ ve $j \in \{1, \dots, N_i\}$

(a) Bir konu $z_{i,j} \sim Multinom(\theta_i)$ seçin.

(b) bir kelime $w_{i,j} \sim Multinom(\varphi_{z_{i,j}})$ seçin.

Bu temel modelde, birkaç basitleştirici varsayım yapılmıştır. İlk olarak, Dirichlet dağılımının k boyutunun (ve dolayısıyla konu değişkeni z'nin boyutsallığı) bilindiği ve sabit olduğu varsayılır. İkincisi, kelime olasılıkları, k*V matrisi ile parametrelendirilen β, tahmin edilecek sabit bir miktar olarak ele alınır. Burada β $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$ şeklinde hesaplanır.

Son olarak, Poisson varsayımı takip eden hiçbir şey için kritik değildir ve gerektiğinde daha gerçekçi belge uzunluğu dağılımları kullanılabilir. Ayrıca, N'nin diğer tüm veri üreten değişkenlerden (θ ve z) bağımsız olduğuna dikkat edilir. Bu nedenle, yardımcı bir değişkendir ve genellikle sonraki gelişimdeki rastgeleliği göz ardı edilmiştir.

Bir k-boyutlu Dirichlet rasgele değişkeni θ, (k −1) -simplex'te değerler alabilir (k-vektör θ, $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$ ise (k − 1) -simplex'te bulunur) ve bu simpleks üzerinde aşağıdaki olasılık yoğunluğuna sahiptir:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1} \quad (3)$$

burada α parametresi, $\alpha_i > 0$ bileşenlerine sahip bir k-vektörü ve burada Γ (x), gamma fonksiyonudur. Dirichlet, simpleks üzerinde uygun bir dağılımdır- üstel ailede,

sonlu boyutlu yeterli istatistiklere sahiptir ve çok terimli dağılıma eşleniktir. Bu özellikler, GDT için çıkarım ve parametre tahmin algoritmalarının geliştirilmesini kolaylaştıracaktır.

A ve β parametreleri göz önüne alındığında, bir konu karışımının θ'sı, bir dizi N tane konunun z ve bir dizi N tane kelimenin w olan ortak dağılımı:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad (4)$$

burada $p(z_n|\theta)$ benzersiz i için $z_n^i = 1$ olacak şekilde basitçe θ_i'dir. θ üzerinden entegre ve z üzerinden özetleme ile bir belgenin marjinal dağılımı elde edilir:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z\_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \quad (5)$$

Son olarak, tek belgelerin marjinal olasılıklarının ürününü alarak, bir doküman olasılığı elde edilir:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d \quad (6)$$

GDT'ni basit bir Dirichlet-çok terimli kümeleme modelinden ayırt etmek gerekir. Klasik bir kümeleme modeli, bir doküman için bir Dirichlet'in bir kez örneklendiği, dokümandaki her belge için bir kez çok terimli bir kümeleme değişkeninin seçildiği ve küme değişkenine bağlı olarak belge için bir kelime kümesinin seçildiği iki seviyeli bir modeli içerir. Birçok kümeleme modelinde olduğu gibi, böyle bir model bir belgeyi tek bir konuyla ilişkilendirilecek şekilde kısıtlar. Öte yandan GDT, üç seviye içerir ve özellikle konu düğümü belge içinde tekrar tekrar örneklenir. Bu model altında, belgeler birden çok konu ile ilişkilendirilebilir [66].

1.6.2. Negatif Olmayan Matris Faktorizasyonu (Non-negative Matrix Factorization)

Negatif Olmayan Matris Faktorizasyon (NOMF) kavramı 1994 yılında Paatero ve Tapper [70] tarafından tanıtılmıştır. Bu çalışma pozitif matris çarpanlara ayırma üzerine olsa da NOMF'nun atası olarak kabul görmüştür.

Negatif olmayan sınırlama, tüm fiziksel varlıklar negatif olmadığı için fiziksel bir korelasyon sağlar. Yani, çeşitli görüntülere veya varlıklara ek olarak bir görüntü gibi fiziksel bir nesneyi ayrıştırmak için NOMF, mantıksal bir yöntem sağlar [71].

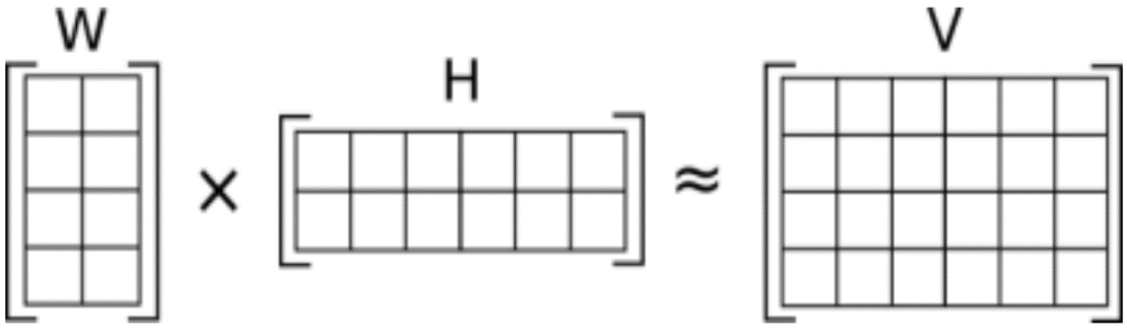
Çoğunlukla analiz edilecek veriler negatif değildir ve fiziksel gerçeklerle çelişmekten kaçınmak için düşük sıralı verilerin ayrıca negatif olmayan değerlerden oluşması gerekir. Klasik araçlar, negatif olmamayı korumayı garanti edemez. Belirli bir negatif olmayan veri matrisine yaklaşmak için azaltılmış sıralı negatif olmayan faktörler bulma yaklaşımı bu nedenle doğal bir seçim haline gelmektedir [1].

Azaltılmış bir sistem modeli, orijinal sistem seviyesine yakın bir doğruluk sağlayabilir. Gürültü giderme, model azaltma, fizibilite yeniden yapılandırma gibi çeşitli yaklaşımlardaki ortak nokta, orijinal verileri alt uzay yaklaşımı yoluyla elde edilen daha düşük boyutlu bir temsil ile değiştirmektir. Faktör analizi ve temel bileşen analizi, değişkenlerin sayısını azaltma ve değişkenler arasındaki yapıları tespit etme amacına ulaşmak için kullanılan birçok klasik yöntemden ikisidir [1].

NOMF, temel bileşen analizi ve vektör niceleme gibi bütünsel temsilleri öğrenen yöntemlerin tersidir. Yani negatif olmayan kısıtlamaları kullanması ile diğer yöntemlerden ayrılır. Bu kısıtlamalar, parçalara dayalı bir temsile yol açar. Çünkü bunlar, eksiltici değil, yalnızca toplamsal kombinasyonlara izin verir [72]

Temel olumlu yönlerden biri, matris elemanlarının ayrı ayrı işlenmesine izin veren girdi verileri olarak bilinen deneysel belirsizliklerin kullanılmasıdır. Bununla birlikte, nokta nokta ölçeklendirme, Tekil Değer Ayrıştırmasına dayalı geleneksel bir faktör analizi ile yeniden üretilemeyen ölçekli bir veri matrisiyle sonuçlanır [73].

Negatif olmayan bir veri matrisi V verildiğinde, NOMF, negatif olmayan faktörler W ve H'den yaklaşık bir çarpanlara ayırma ile $V \approx WH$ olarak bulur [74]. Şekil 5'te temsili olarak gösterilmiştir.



Şekil 5. NOMF'nin grafik model gösterimi

NOMF, çok değişkenli verilerin istatistiksel analizine uygulanabilir. Bir dizi çok değişkenli n boyutlu veri vektörü verildiğinde, vektörler bir n x m matrisinin V sütunlarına yerleştirilir. Burada m, veri kümesindeki örneklerin sayısıdır. Bu matris daha sonra belli bir yaklaşıklık ile bir n x r boyutlu W matrisi ve bir r x m boyutlu H matrisi olarak çarpanlara ayrılır. Genellikle r, n veya m'den küçük olacak şekilde seçilir, böylece W ve H, orijinal matris V'den daha küçük olur. Bu, orijinal veri matrisinin sıkıştırılmış bir versiyonu oalrak sonuçlanır [75].

W ve H'nin bulunabileceği birçok yol olsa da popüler olan Lee ve Seung'un [75] çarpımsal güncelleme kuralına göre olan algoritma aşağıda verilmiştir.

1. W ve H'yi negatif olmayan olarak tanımla.

2. Ardından, yinelemenin bir indeksi olarak "n" ile aşağıdaki (7). ve (8). formülleri hesaplayarak W ve H'deki değerleri durağan olana kadar güncelle.

$$W_{[i,k]}^{n+1} \leftarrow W_{[i,k]}^{n} \frac{(V(H^{n+1})^T)_{[i,k]}}{(W^n H^{n+1}(H^{n+1})^T)_{[i,k]}} \tag{7}$$

$$H_{[k,j]}^{n+1} \leftarrow H_{[k,j]}^{n} \frac{((W^n)^T V)_{[k,j]}}{((W^n)^T W^n H^n)_{[k,j]}} \tag{8}$$

Lee ve Seung'un [75] teorisi W ve H'nin güncellemeleri mesafenin durağan bir noktasındaysa öklid mesafesi değişmeyeceği şeklindedir.

1.6.3. Modelin Konu Tutarlılık Ölçütleri

KM, olasılık modelleri tarafından etkin bir şekilde kullanılan fiili standart yöntem haline geldiği çeşitli belge koleksiyonlarında gizli anlamsal yapının keşfi için önemli bir araçtır [66]. Tespit edilen konular genellikle karşılık gelen en üst sıradaki en yüksek dereceli N terim kullanılarak tanımlanır. Olasılıklı konu modellerinde, model uygunluğunu değerlendirmek için, Şaşkınlık veya Beklemede olma olabilirliği gibi bir dizi nicel ölçütler kullanılmıştır [77].

Konu tutarlılığı ölçümleri, konuyla ilgili yüksek puan alan kelimeler arasındaki anlamsal benzerlik derecesini ölçerek tek bir konuyu puanlar. Bu ölçümler, anlamsal olarak yorumlanabilen konular ile istatistiksel çıkarımın yapay öğeleri olan konular arasında ayrım yapmaya yardımcı olur [78].

Literatürde tutarlılık ölçüleri arasında C_V, C_P, C_{UCI}, C_{UMASS}, C_{NPMI} ve C_A vardır. Kısaca tanımları:

C_V: kayan bir pencereye, en üstteki kelimelerin tek set segmentasyonuna ve Normalleştirilmiş Noktasal Karşılıklı Bilgi (NNKB) ve kosinüs benzerliğini kullanan dolaylı bir doğrulama ölçüsüne dayanır.

C_P: kayan bir pencereye, en üstteki kelimelerin bir önceki segmentasyonuna ve Fitelson'ın tutarlılığının onay ölçüsüne dayanmaktadır.

C_{UCI}: kayan bir pencereye ve verilen en iyi kelimelerin tüm kelime çiftlerinin Noktasal Karşılıklı Bilgisine (NKB) dayanmaktadır.

C_{UMASS}: belge birlikte bulunma sayılarına, bir önceki segmentasyona ve doğrulama ölçüsü olarak logaritmik koşullu olasılığa dayanır.

C_{NPMI}: NNKB kullanan, C_{UCI} tutarlılığının geliştirilmiş bir versiyonudur.

C_A: bir bağlam penceresine, en üstteki kelimelerin ikili karşılaştırmasına ve NNKB ve kosinüs benzerliğini kullanan dolaylı bir doğrulama ölçüsüne dayanır [79].

1.6.4. Tutarlılık Analizi

C_V, tutarlılık ölçülerinin konfigürasyon uzayının sistematik olarak incelenmesiyle bulunan bir kombinasyondur. Literatürde şimdiye kadar göz ardı edilmiş bir yöntemdir [80].

C_V dört bölümden oluşur [81]:

(i) verilerin kelime çiftlerine bölünmesi,

(ii) kelime veya kelime çifti olasılıklarının hesaplanması,

(iii) bir kelime kümesinin başka bir kelime kümesini ne kadar güçlü desteklediğini ölçen bir doğrulama ölçüsünün hesaplanması,

(iv) bireysel doğrulama önlemlerinin genel bir tutarlılık puanına toplanması.

Her $S_i = (W', W^*)$ için, W^*'nin W''yı ne kadar güçlü desteklediğini hesaplayan ve W'deki tüm kelimelerle ilgili olarak W^* ve W''nin benzerliğine dayanan bir doğrulama ölçüsü φ hesaplanır. Bu benzerliği hesaplamak için, W^* ve W', W içindeki tüm kelimelerin anlamsal desteğini yakalamanın bir yolu olarak bağlam vektörleri olarak temsil edilir [82]. Bu $\vec{v}(W')$ ve $\vec{v}(W^*)$ vektörleri, formül (9)'da örneklendiği gibi, W'deki tüm kelimelerle eşleştirilerek oluşturulur. Tek tek w_i ve w_j kelimeleri arasındaki uyum, formül (11)'de gösterildiği gibi NNKB yoluyla hesaplanır. NNKB, NKB'nin aksine, insan konu sıralama verileriyle daha yüksek bir korelasyon gösterir [83].

Sıfırın logaritmasını hesaba katmak için ε kullanılır ve daha yüksek NNKB değerlerine daha fazla ağırlık vermek için γ kullanılır. Bir S_i çiftinin doğrulama ölçüsü ϕ, formül (12)'de ifade edildiği gibi $\vec{v}(W') \in \vec{u}$ ve $\vec{v}(W^*) \in \vec{w}$ ile S_i içindeki tüm bağlam vektörlerinin $\phi S_i(\vec{u}, \vec{w})$'nin kosinüs vektör benzerliğini hesaplayarak elde edilir.

$$\vec{v}(W') = \left\{\sum_{w_i \in W'} NNKB\left(w_i, w_j\right)^{\gamma}\right\}_{j=1,\ldots,|W|} \tag{9}$$

$$NKB\left(w_i, w_j\right) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{10}$$

$$v_{ij} = NNKB\left(w_i, w_j\right)^{\gamma} = \left(\frac{NKB\left(w_i, w_j\right)}{-\log(P(w_i, w_j)+\epsilon}\right)^{\gamma} \tag{11}$$

$$\emptyset S_i(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{||\vec{u}||_2 \cdot ||\vec{w}||_2} \tag{12}$$

Nihai tutarlılık puanı, tüm ∅ doğrulama ölçütlerinin aritmetik ortalamasıdır [81].

1.7. Literatürde Yapılmış Çalışmalar

İnternetin gelişmesiyle yazılı bilgilerde artmıştır. Bilgi üretim fabrikasına dönüşen internet aracılığıyla bilgi çıkarımı için üretilen veri madenciliği ve metin madenciliği teknikleri de gelişmektedir. Ayrıca çalışma alanları da bir hayli artmaktadır. Literatür incelendiğinde siyaset, ticaret, spor, ulaşım, doğal kaynaklar gibi birbirinden çok farklı alanlarda, çok farklı dillerde metin madenciliğine yönelik çalışmaların olduğu görülmüştür.

Chowdhury ve Zhu [7], makalelerinde çevrimiçi haber kaynaklarından gelen bilgileri kullanarak bir ön sınıflandırma önermişler, ulaşım altyapısının uzun vadeli planlanmasına ve karar verilmesine yardımcı olabilecek bütünleşmiş bir ontolojinin geliştirilmesini başlatmayı amaçlamışlardır. Bu alanda, çevrimiçi haber kaynakları, sosyal medya, teknik belgeler ve çok daha fazlasının büyük miktarda veri üretiminde kaynak olan bu alanların ulaşım sektöründe kullanımı için mevcut ulaşım altyapı planlaması uygulamalarını potansiyel olarak iyileştirebilecek çok büyük bir kapsam teşkil edecek çalışmada bulunmuşlardır. Ulaştırma altyapısı planlama sınıflandırması, devlet kurumları tarafından yayınlanan 20 ulaştırma planlama belgesindeki metinsel bilgilere dayalı KM teknikleri olan GDT ve NOMF algoritmaları kullanılarak ulaşım alanında açık ve yapılandırılmış bilgi sağlayabildikleri için yararlı bir çalışma olmuştur. Bu çalışmada KM analizine dahil edilen belgelerin hem sayısı hem de kapsamı sınırlı olmuştur. Gelecekteki çalışmalar için yazarlar, KM'yi uygulamak ve ulaşım altyapısı planlamasıyla ilgili ortaya çıkan kavramları belirlemek için çevrimiçi sosyal medya verileri de dahil olmak üzere daha fazla kaynaktan daha fazla veri toplayarak sınıflandırma yakalanan yeni kavram ve bilgilere dayalı olarak daha da geliştirilecektir.

Yu, Lu ve Muñoz-Justicia [8], İspanyol haber medyasının sosyal medya platformlarındaki halk sağlığı krizlerini nasıl kapsadığının anlaşılmasına yönelik yaptığı çalışmalarında COVID-19 salgını hakkındaki haberleri analiz etmiş ve karşılaştırmış. Twitter'dan El País ve El Mundo'nun haber verilerinden yararlanılmış. Çalışmalarında her gazetenin Twitter hesabı için sekiz haber çerçevesi tanımlanmış. Ayrıca, tüm pandemik gelişme süreci üç döneme ayrılmış- kriz öncesi dönem, kilitlenme dönemi ve iyileşme dönemi. Sosyal medya aracı Twiter'dan toplanılan verilerin temizliği için ilk olarak retweetler kaldırılmış. Ardından hariçteki web site adresleri, hashtagler, emojiler, rakamlar ve etkisi az olan kelimeler (stop-words) kaldırılmış. GDT ile bir model oluşturulup analiz gerçekleştirilmiş.

Panasyuk, Yu ve Mehrotra [9] yaptıkları çalışmada, Twitter'ın kullanıcı toplulukları arasındaki çatışmayı belirlemek için nasıl kullanılabileceğini ele almışlar. Makalelerinde, Amerika'daki kongre üyelerini endişelendiren önemli haber medyası konularının otonom bir şekilde nasıl çıkarılabileceği gösterilmiştir. Amaçları sadece popüler konuları değil, cumhuriyetçiler ve demokratlar arasında yüksek düzeyde anlaşmazlık bulunan konuları araştırmak olmuş. Konuları, bir topluluk için olumsuz, diğeri için olumlu olacak şekilde karşıt görüşlere göre sıralanmış. GDT ile model kurulmuş. Başarılı bir sonuç elde ettikleri çalışmalarında gelecekteki çalışmaları için, sonuçların nasıl tahmin edileceğini araştırmak ve uzun vadeli olarak, henüz tespit edilmemiş topluluklar için bu hedefleri ve anlaşmazlık seviyelerinin olası şiddet ve huzursuzluk için nasıl bir uyarı olarak kullanılabileceğini tekrarlamak olarak ifade edilmiş.

Petrol piyasasındaki risklerin oluşumu ve kapsamı hakkında yapılan çalışmada Zhao ve arkadaşları [10] GDT modeline dayalı ağ haberlerinden risk faktörlerini çıkarmak için yöntem önermişlerdir. Web tarayıcılarını kullanarak 18.000 haber ile analiz gerçekleştirilmiş. 28 risk faktörü tespit edilmiş ve risk faktörleri beş kategoriye ayrılmış. Bunları gerçek vakalarla karşılaştırarak, risk faktörlerinin temelde makul ve etkili olduğu sonucuna varılmış. Sonucunda çeşitli risk faktörlerinin jeopolitik, savaş çatışmaları, çevre koruma, OPEC politikaları ve piyasa arz ve talebi etrafında merkezlendiğini göstermiştir. Özellikle, Latin Amerika ekonomik yaptırımları ve Suriye savaşı ile ilgili olarak, siyasi çatışmalar, ekonomik yaptırımlar ve petrol üreten bölgeleri içeren savaşlar en önemli faktör olarak görülmüş. Petrol piyasası risk faktörleri yapısını değerlendirmek ve petrol piyasası tedarik zincirinin belirli bir bölümünün analizi bu faktörler arasındaki ilişkiler ile analiz edileceği öne sürülmüş.

Gao ve arkadaşları [11] haber ve tweetlerden temsilci ve tamamlayıcı bilgileri birlikte keşfederek trend olan konuları özetlemek için KM'sine dayalı yeni bir denetimsiz yaklaşım önermişler. Önerdikleri yöntemde, tamamlayıcı cümle-tweet çiftlerinin tanımlanmasını güçlendirerek konuyu zenginleştiren içeriği yakalar. Bir çiftin tamamlayıcılığını değerlendirmek için, çok-belgeli özetleme literatüründe iki boyutlu bir konu yönü modelini ve bir çapraz koleksiyon yaklaşımını birleştirerek KM biçimciliğinden yararlanmışlar. Önerdikleri yöntemi LexRank, KL-divergence, Cosine and Language Modeling ve LexRank+Complementarity (LexComp) ile karşılaştırdıklarında daha iyi ROUGE puanları elde etmişler.

Hidayatullah ve arkadaşları [12] Endonezya'daki futbol haberleriyle ilgili tweetlerin konusunu belirlemek için KM'si uygulamışlardır. Çalışmada kullanılan veriler, futbol hakkında her zaman güncelleme yapan ve daha önce seçtikleri birkaç resmi Endonezya Twitter hesabından alınmış. Konu başlıklarını tespit edecek yöntem olarak GDT kullanılmış. Yapılan analize göre, maç öncesi analiz, canlı maç güncellemesi, futbol kulübü başarıları vb. gibi birçok konu elde edilmiş. Genel olarak, futbol haber sağlayıcısının Twitter hesabı tarafından yayınlanan başlıklar da Endonezya, İngiltere, İspanya, İtalya ve Almanya gibi bazı ülkelerdeki futbol rekabeti hakkında da bilgi verdiği tespit edilmiş.

Toprak [13], Türkçe olarak analiz gerçekleştirmiştir. Türkiye'deki her il için yayımlanan haberlere bir KM yöntemi olan GDT'yi kullanılarak her il için en yüksek frekansa sahip ayrı ayrı 10 konu belirlenmiş. Türkçe yapılan bir çalışma olmasından Türkçe karakterlerinin veri setinde çok fazla olması analiz sonucunu negatif şekilde etkileyeceğinden analiz aşamasına geçilmeden önce karakter dönüşümü yapılarak Türkçe karakter sorunu çözülmüş ve analizin daha doğru sonuçlar elde etmesi sağlanmıştır.

Aslan [14] tarafından yapılan bu çalışmada, mikroblog sitelerindeki haber mesajlarını tanımlamak için bir yaklaşım önermiştir. Bu yaklaşım mesajlar arasındaki benzerlikleri yakalamak ve "Haber sağlayıcılar mikroblogları nasıl kullanıyor? Bireysel kullanıcılar haber yayıyor mu? Yayıyorsa nasıl yayıyorlar? Zamana ait ve niceliksel özellikler nelerdir?" gibi sorulara cevaplarıyla onların zamansal ve nicel özelliklerini analiz edip üzerinde çalışmak için uygulanmış. Tezin yazıldığı dönemde en popüler mikroblog sistemi olarak Twitter kullanılmış. Bir kısım, farklı küresel olaylara ait ve bireysel mikroblog kullanıcı mesajları incelenmiş. Deneyler iki kategoride gerçekleştirilmiş. İlk olarak seçilen 60 kullanıcının tweetleri, tweetlerin N-V-N modeliyle eşleşip eşleşmediğini kontrol etmek için ayrı ayrı analiz edilmiş. Haber tweet oranlarına göre yapılan sıralamada haber içerik üreticisi olan kullanıcılar listenin başında yer alması mikrobloglardaki haber tweetlerinin çıkarılması için N-V-N modelini kullanmanın faydalı olacağı düşündürmüş. İkinci adım olarak küresel olayların analizine geçilmiş ve çok sayıda tweet analiz edilmiş. Buldukları sonuçlar: Futbol maçları gibi spor etkinliklerinde insanlar maç günü diğer günlere göre daha fazla tweet atma eğiliminde oldukları ancak maç gününde haber tweet oranları en düşük seviyede olduğu ortaya çıkmış. Ayrıca haber tweetlerinde isim veya isim sayısı incelendiğinde bazı kelimelerin diğerine göre çok daha fazla kullanıldığı görülmüştür.

2. YAPILAN ÇALIŞMALAR

Bu tez kapsamında İngilizce olarak toplanılan metinsel verilere metin madenciliği işlemlerini uygulayarak KM'si ile kümeleme analizi gerçekleştirilmiştir. Analiz aşamasında GDT ve NOMF yöntemleri kullanılmış ve iki yöntemin karşılaştırılması sağlanmıştır.

NewsAPI'nin ücretsiz sürümünde haber istemcisini sağlayan kod ile bir aylık zaman aralığındaki haberlere ulaşılabilmektedir. Ayrıca her aramada en fazla 100 habere erişilebilmektedir. Bu durum haber toplama aşaması için büyük bir kısıt olarak karşımıza çıkmaktadır. Bu problemi aşabilmek için hazırlanan yazılım günlük olarak çalıştırılarak veri çekilmiştir. NewsAPI kaynağından toplanılan her haberin; yazar, içerik, açıklama, yayınlanma tarihi, kaynak, başlık, internet bağlantısı ve resim bilgilerine erişilebilmektedir.

Türkiye'ye ve Yunanistan'a yönelik haber metinlerini GDT ve NOMF yöntemlerini kullanarak otonom olarak analiz edebilecek KM'si yapmak hedeflenmiştir. 13 Mart 2020-13 Eylül 2020 tarih aralığındaki Türkiye ve Yunanistan ile ilişkili haberleri derleyip çıkan sonuçlar ile iki ülke arasındaki konuları görme olanağı sağlanmıştır. Bununla birlikte iki yöntemin karşılaştırılması da sağlanmıştır. Böylelikle metin veri tabanlarındaki bilgi keşfi olarak da adlandırılan metin madenciliğinde büyük boyuttaki metin içerikli veri kaynaklarından, önceden bilinmeyen ve potansiyel olarak ihtiyaç duyulan bilginin çıkarılmasında kolaylık sağlayabilecek bir uygulama oluşturulmuştur.

2.1. Veri Toplama

Bu çalışma için NewsAPI haber veri kaynağından faydalanılmıştır. NewsAPI, şu anda web' de yayınlanan başlıklar ve makaleler için JSON meta verilerini döndüren basit ve kullanımı kolay bir API'dir. NewsAPI, dünya çapında 75.000'den fazla büyük ve küçük haber kaynağından ve blogdan milyonlarca makaleye erişim imkanı sağlamaktadır [72]. Tablo 2'de NewsAPI haber kaynağından elde edilen içeriklerin isimleri, veri tipleri ve içeriklere dair örnekler verilmiştir.

Tablo 2. NewsAPI'nin bir haberden sağladığı bilgilere örnek

İçerik İsimleri	Veri Tipi	İçerik
Author	str	ABC News
Content	str	Now Playing: Fisherman hooks great white shark off Cape Cod beach…
Description	str	Beachgoers had a close encounter with a shark swimming in shallow waters near the shore in Benidorm, Spain, before it returned to deeper waters, according to police.
PublishedAt	str	2021-08-15T20:21:06Z
Source	dict	{“id”: abc-news, “name”: “ABC News”}
Title	str	WATCH: Shark swims close to shore in Spain
Url	str	https://abcnews.go.com/International/video/shark-swims-close-shore-spain-79469504
UrlToImage	str	https://s.abcnews.com/images/International/210815_abc_social_shark_hpMain_16x9_992.jpg

2.2. Veri Ön İşleme

Bu çalışmada, ön işleme olarak kelimelerden noktalama işaretlerinin kaldırılmış, metin karakterlerinin tek tip olunması için harflerin hepsi küçük harfler dönüştürülmüştür. Bağlaç, zamir gibi belgelerde çokça geçen kelimelerin hesaplamalarda bir anlam ifade etmeyecek, analizi kötü etkileyecek kelimeler olmasından dolayı belgelerden çıkartılmıştır. Kök bulma ve kelime eklerinden arındırma işlemlerinin uygulanması ile kelimelerdeki çekim ekleri kaldırılmış ve İngilizce ’deki farklı zamanları ifade eden kelimeler ilk hallerine dönüştürülmüştür.

Bu tez çalışmasında kullanılan veri seti İngilizce metinlerden oluştuğu için İngilizce etkisiz kelimeler (stop-words) listesi gerekmektedir. Gereken bu liste Python programındaki kütüphanelerce karşılanmış ve Tablo 3’te etkisiz kelimelerin listesine yer verilmiştir.

Tablo 3. İngilizce ’deki etkisiz kelimelerin (stop-words) listesi

i	it	have	between	why	just	haven
me	its	has	into	how	don	haven’t
my	it’s	had	through	all	don’t	isn
myself	itself	having	during	any	should	isn’t
we	they	do	before	both	should’ve	ma
our	them	does	after	each	now	mightn
ours	their	did	above	few	d	mightn’t
ourselves	theirs	doing	below	more	ll	mustn
you	themselves	a	to	most	m	mustn’t
you’re	what	an	from	other	o	needn
you’ve	which	the	up	some	re	needn’t
you’ll	who	and	down	such	ve	shan
you’d	whom	but	in	no	y	shan’t
your	this	if	out	nor	ain	shouldn
yours	that	or	on	not	aren	shouldn’t
yourself	that’ll	because	off	only	aren’t	wasn
yourselves	these	as	over	own	couldn	wasn’t
he	those	until	under	same	couldn’t	weren
him	am	while	again	so	didn	weren’t
his	is	of	further	than	didn’t	won
himself	are	at	then	too	doesn	won’t
she	was	by	once	very	doesn’t	wouldn
she’s	were	for	here	s	hadn	wouldn’t
her	be	with	there	t	hadn’t	
hers	been	about	when	can	hasn	
herself	being	against	where	will	hasn’t	

Sosyal medyada toplanılan metin verileri yazım kurallarına uygun olmayabilir. İçerisinde fazlaca eksik harfler ve emojiler barındırabilmektedir. Sosyal medyanın aksine tez kapsamında gazete haberlerinden toplanmış olan verilerin kullanılmış olması nedeniyle ön işleme aşamasında veri temizliği sosyal medya verilerine göre daha kolay aşılmıştır.

Tablo 4 ve Tablo 5’te Türkiye ve Yunanistan için toplanılan ham haber verilerine yer verilmiştir.

Tablo 4. Türkiye için toplanılan ham verilere örnek

Sıra	Veri Tipi	Haber Metni
1	str	TURKEY CANNOT HOST ANY MORE REFUGEES, EITHER FROM SYRIA OR BEYOND, FOREIGN MINISTER MEVLÜT ÇAVUOLU S
2	str	A POLICE OFFICER TAKES A PICTURE OF A MIGRANT WHO ARRIVED IN A CLOSED CAMP OF KLEIDI NEAR PROMAHONAS, NORTHERN GREECE, ON MARCH 21, 2020. (AFP PHOTO)
3	str	FM ÇAVUSOGLU SAYS TURKEY CANNOT HOST ANY MORE REFUGEES TURKEY CANNOT HOST ANY MORE REFUGEES, EITHER FROM SYRIA OR BEYOND, FOREIGN MINISTER MEVLÜT CAVUSOGLU SAID ON MARCH 23, IN AN OPINION PIECE IN THE FINANCIAL TIMES. ANKARA
4	str	TURKEY TO MINIMIZE TROOP MOVEMENTS IN SYRIA CITING VIRUS CONCERN <P>DOCTORS HAVE BEEN SENT TO SYRIA TO CONDUCT MEDICAL TRAINING ON COVID-19</P>
5	str	WITH JOHNSON AILING, UK FACES A LEADERSHIP QUANDARY THE BRITISH GOVERNMENT HURTLED INTO UNCHARTED TERRITORY TUESDAY, WITH ITS FOREIGN SECRETARY, DOMINIC RAAB, TAKING UP THE DAY-TO-DAY DUTIES OF PRIME MINISTER BORIS JOHNSON, WHO WAS BEING TREATED IN AN INTENSIVE CARE UNIT

Tablo 5. Yunanistan için toplanılan ham verilere örnek

Sıra	Veri Tipi	Haber Metni
1	Str	THE DISPUTE HAS REIGNITED THE LONG-RUNNING RIVALRY BETWE… ISTANBUL (AFP)
2	Str	HOSPITALS IN FRANCE REPORTED A 10% INCREASE IN THE NUMBER OF DEATHS
3	Str	WHAT IF THE AXIS HAD CUT OFF THE SOVIET LIFELINE AT MURMANSK? WARFARE HISTORY NETWORK
4	Str	AN EARTHQUAKE WITH A PRELIMINARY MAGNITUDE OF 5.1 STRUCK OFF THE COAST OF SOUTHERN GREECE MONDAY MORNING, GREECE'S INSTITUTE OF GEODY... ATHENS, GREECE
5	Str	THE LATEST: BUSINESSES TO START REOPENING IN TENNESSEE THE LATEST ON THE CORONAVIRUS PANDEMIC. THE NEW CORONAVIRUS CAUSES MILD OR MODERATE SYMPTOMS FOR MOST PEOPLE. FOR SOME, ESPECIALLY OLDER ADULTS

2.3. Sayısallaştırma

Verinin ön işleme adımları uygulandıktan sonra yapısal hale dönüşen veri bu aşamada terimlere ayrılmıştır. Bu aşamada veri üzerinde analizler yapabilmek için metin verisinin sayısal ifadelere çevrilmesi gerekmektedir. Her bir haber için terimler elde edilerek iki aşamada genel bir terimler sözlüğü oluşturulmuştur. İlk aşamada, çalışmada KM algoritması olarak kullanılmış olan GDT'nin terim frekansı ile işlem yapmasından dolayı veri setini oluşturan her bir kelimeye terim frekansı ağırlıklandırması yapılmıştır. İkinci aşama ise doküman terim matrisinin oluşturulmasıdır. Her bir terim için terim frekansı ve doküman terim matrisi elde edilerek metin verilerinin analize hazır hâle gelmesi sağlanmıştır.

Terim frekansı ve doküman terim matrisini elde etmek için kullanılacak öznitelik seçimi yapmak başlı başına bir iş yüküdür. Bunun için bir "Boru Hattı" ve "Izgara Araması" oluşturulmuştur.

Analizin sağlıklı bir şekilde gerçekleştirilebilmesi için, öznitelik seçiminin ayarlanması gerekmektedir. Analiz için bütün kelimelerin incelenmesi hem iş yükü oluşturması hem de analizi kötü etkileyecek olmasından bu aşamada "öznitelik seçim" yöntemi kullanılmıştır. Bunun için öznitelik seçim optimizasyonu için bir konfigürasyon oluşturulmuştur. Bu yöntem ile gerekli gereksiz kelimeleri belirleyip gerekli kelimelerle yola devam edilmiştir.

Izgara Araması yardımıyla belirlenilen kelime frekansının başarım oranları tespit edilebilmektedir. Terim frekansı ve terim doküman matrisi elde edilerek analiz için veri hazırlanmıştır. Türkiye haber verisi için kullanılacak sayısallaştırılmış kelime frekansının 0.909'luk başarı üreten öznitelikleri kullanarak veri madenciliği aşamasına gidilmiştir. Yunanistan haber verisi için ise 0.913'lük başarı gösteren öznitelikler uygulanmıştır. Türkiye haber verileri için (30511, 2000) bir matrisle işlem yapılmış. Yunanistan haber verileri için (28646, 2000) bir matrisle işlem yapılmıştır.

2.4. Veri Madenciliği Yöntemlerinin Uygulanması

Bu aşamada veri madenciliği yöntemlerinin (sınıflandırma, regresyon, kümeleme ve birliktelik kuralları) uygulması gerçekleştirilmektedir. Metin madenciliği olarak tezde kullanılacak KM yöntemlerinin karşılaştırılması bu aşamada gerçekleştirilmiştir.

Sayısallaştırma aşamasında elde edilmiş olan terim frekansı ve doküman terim matrislerini kullanarak metin kümeleme işlemi gerçekleştirilmiştir. KM yapabilmek için GDT ve NOMF algoritmaları uygulanmıştır. Python programında GDT ve NOMF uygulayabilmek için Scikit-learn'ün kütüphanelerinden "LatentDirichletAllocation" ve "NMF" kütüphaneleri kullanılmış algoritmaların performansları karşılaştırılmıştır.

Her iki yöntem ve her iki ülkenin veri setleri için konu tutarlılığı analizine göre konu sayıları belirlenmiştir. Her konuya ait 15 kelime getirecek şekilde analiz gerçekleştirilmiştir.

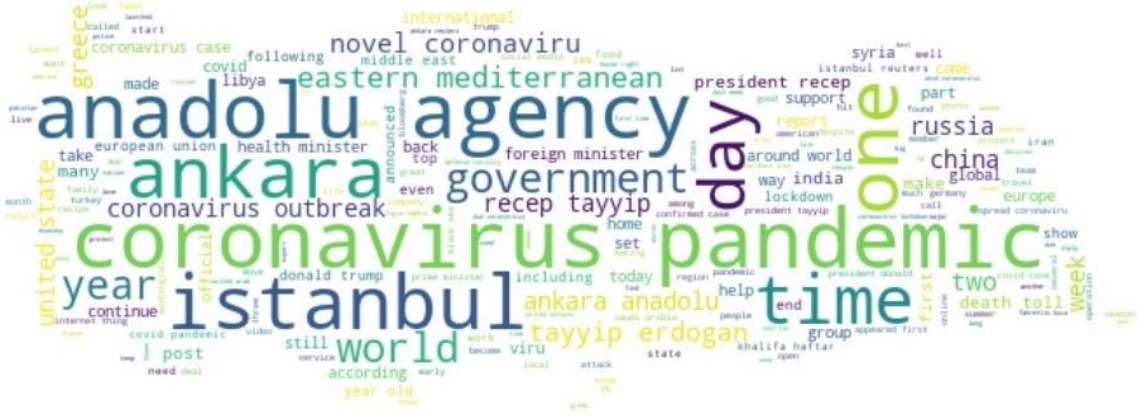
3. BULGULAR

Bu bölümde belirlenilen iki konu hakkında yapılmış İngilizce haberler NewsAPI servisi aracılığı ile derlenmiştir. Bu derlenmiş metin verileri işlenerek GDT ve NOMF algoritmaları ile KM'si gerçekleştirilmiş ve iki algoritmanın karşılaştırılması sağlanmıştır. Konu tutarlılıklarına göre konu sayıları belirlenmiş ve her konuda 10 kelime yer almıştır.

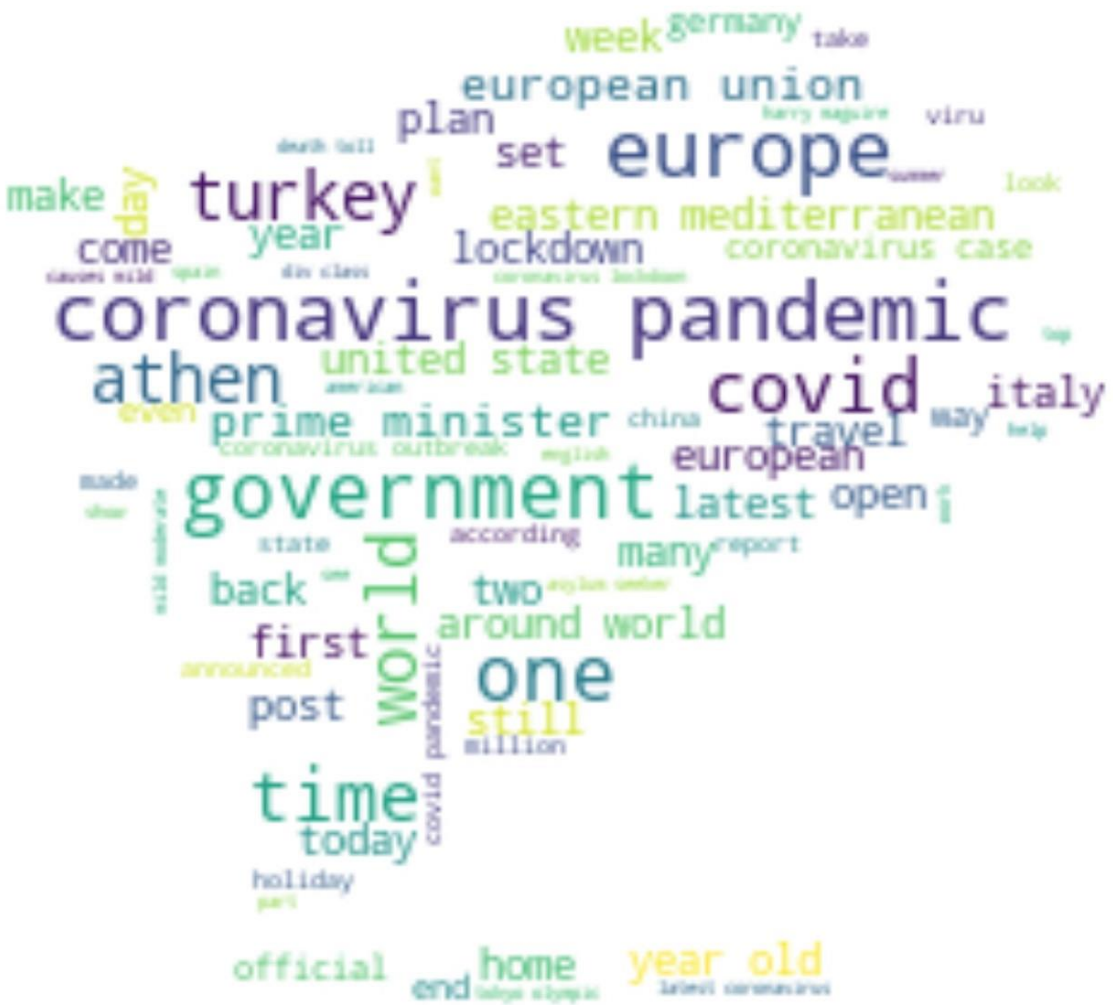
Belirlenilen iki konu Türkiye ve Yunanistan olmuştur. NewsAPI servisinden kısıtlar "Turkey" ve "Greece" şeklinde oluşturulmuştur. Sistemin izin verdiği ölçüde olabildiğince çok veriyi alabilmesi denenmiştir. Tablo 6'da gösterildiği gibi Türkiye için yaklaşık 32243 haber çekilebilmiştir. Bunlardan da bazı içerik eksikliklerinden ve aynı haberlerin birden fazla kullanılmaması için temizlenilen haberlerden sonra Türkiye için 27956 haber verisi ile çalışma gerçekleştirilmiştir. Yunanistan için ise yaklaşık 30056 haber çekilirken yine düzenlemelerden sonra 26081 haber verisi ile analiz gerçekleştirilmiştir. Toplamda ham 62299 İngilizce haber metnine ulaşılmış işlenmiş 54037 veri ile analiz gerçekleştirilmiştir. Şekil 6 ve Şekil 7'de Türkiye ve Yunanistan hakkında yapılan haber verileri için kelime bulutu olarak ayrı ayrı gösterilmiştir.

Tablo 6. Toplanılan ham haber sayısı ve toplam işlenmiş veri sayısı

Türkiye		**Yunanistan**		**Toplam**	
Türkiye için toplanılan ham veri sayısı	Temizlenen haberlerden sonra işlenmiş veri sayısı	Yunanistan için toplanılan ham veri sayısı	Temizlenen haberlerden sonra işlenmiş veri sayısı	Toplam Ham Veri	Toplam İşlenmiş Veri
32243	27956	30056	26081	62299	54037



Şekil 6. Türkiye için toplanılan haber metinlerinden oluşan kelime bulutu



Şekil 7. Yunanistan için toplanılan haber metinlerinden oluşan kelime bulutu

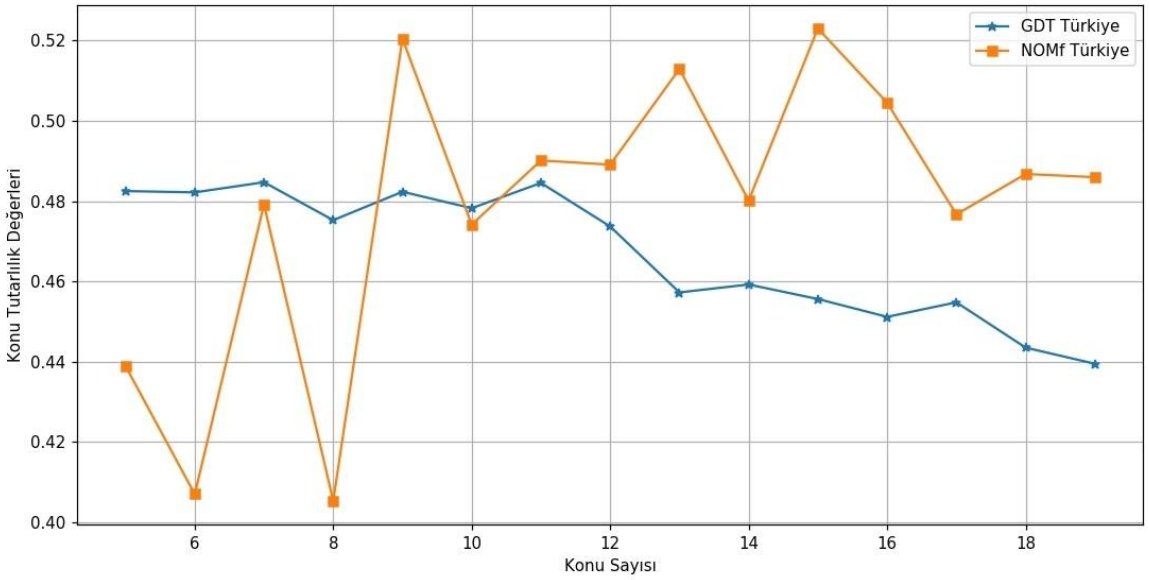
NewsAPI kaynağından toplanılan her haberde; yazar, içerik, açıklama, yayınlanma tarihi, kaynak, başlık, internet bağlantısı ve resim bilgilerine ulaşılmış olsa da içerik, açıklama ve başlık bölümlerinin birleştirilmesi ile analiz gerçekleştirilmiştir.

TF'nı oluşturabilmek için fonksiyonun ihtiyaç duyduğu parametreler optimize edilmiştir. Bunun için Python'ın Izgara Arama fonksiyonu kullanılmıştır. Izgara Arama fonksiyonunu kullanabilmek için Türkiye ve Yunanistan verilerinin duygu durumları etiket olarak kullanılmıştır. En iyi parametrelerin seçimi esnasında Türkiye ve Yunanistan verileri için ayrı ayrı 2800 kombinasyon denenmiştir. Izgara Arama sonucunda %91'lık performans gösteren en iyi parametreler seçilmiş ve Tablo 7'de kullanılan parametrelere yer verilmiştir. En iyi TF-TBF'nı oluşturabilecek parametrelerin belirlenmesi ile TF-TBF elde edilmiştir. Elde edilen TF-TBF ile GDT ve NOMF algoritmaları çalıştırılmış ve her iki ülkenin metin verilerine uygulanmıştır.

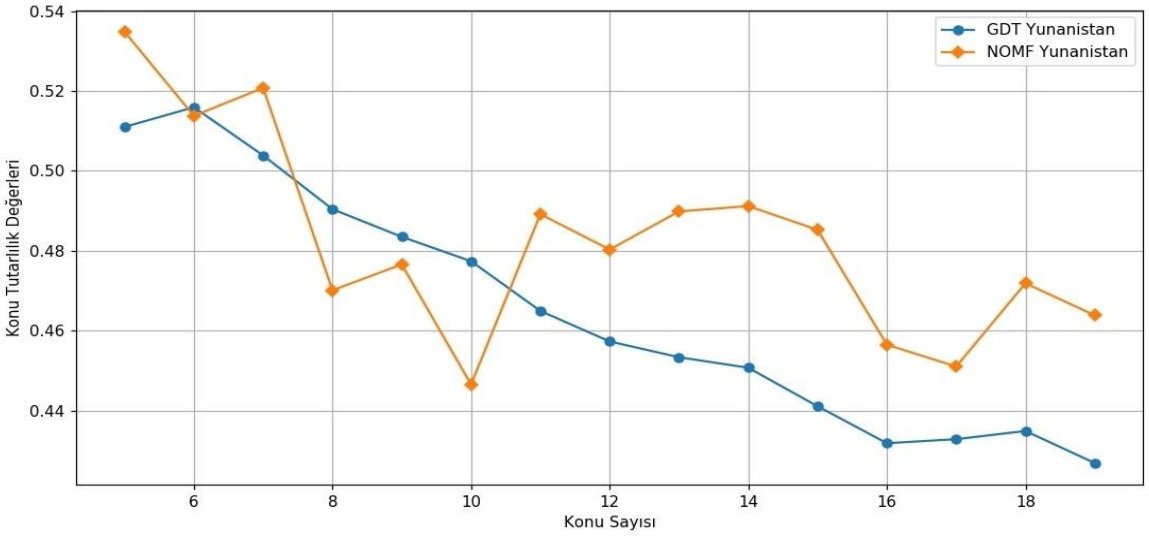
Tablo 7. Izgara Arama sonucunda TF için seçilen parametreler ve değerleri

En İyi Skor	0,91
Tfvect__max_df	0,5
tfvect__max_features	2000
tfvect__ngram_range	(1, 1)
tfvect__norm	'l2'
tfvect__use_idf	True

Konu sayısının seçiminde konu tutarlılık ölçütlerinden C_V yöntemi kullanılmıştır. Her iki veride ve her iki yöntem için ayrı ayrı çalıştırılmıştır. 5 ile 20 konu arasında en iyi sonucu veren konu sayısı tespit edilmiştir. Şekil 8'de Türkiye verisi için GDT yöntemiyle en iyi sonucu 11 konu belirlendiğinde elde etmiştir. NOMF yöntemiyle 15 konu oluşturulduğunda en iyi sonucu elde etmiştir. Şekil 9'da Yunanistan verisi için GDT yöntemiyle 6 konu belirlendiğinde en iyi sonucu ürettiği anlaşılmaktadır. Yunanistan verisi için NOMF yöntemine göre ise 5 konu belirlendiğinde en iyi sonucu üretmiş olduğu görülmüştür.



Şekil 8. Türkiye verisi için konu sayılarına göre Konu Tutarlılık değerleri



Şekil 9. Yunanistan verisi için konu sayılarına göre Konu Tutarlılık değerleri

Türkiye ve Yunanistan verileri için yapılacak konu çıkarımında kaçar konu olması gerektiğine konu tutarlılığı analizi ile ölçülmektedir. Konu tutarlılığını ölçmek için C_V yöntemi seçilmiştir. Konu sayısınca gösterdikleri performansların toplamına göre hangi yöntemin daha iyi performans gösterdiği böylece tespit edilmiştir. Bu yöntem ile GDT ve NOMF algoritmalarının konu sayısına göre oluşan performansları karşılaştırmalı olarak Tablo 8'de yer almıştır. Her bir konu sayısının başarımlarının ortalaması alındığında NOMF yönteminin GDT yöntemine göre daha yüksek bir ortalamaya ulaştığı

gözlemlenmiştir. Türkiye ve Yunanistan'ın 13 Mart 2020-13 Eylül 2020 tarih aralığındaki haber verileri için NOMF algoritması daha iyi sonuç üretmiş olduğu tespit edilmiştir.

Tablo 8. GDT ve NOMF algoritmalarının Konu sayılarına göre skorlarının listesi

Konu Sayısı	**Türkiye**		**Yunanistan**	
	GDT	**NOMF**	**GDT**	**NOMF**
5	0.482498	0.438888	0.511025	0.534840
6	0.482179	0.406939	0.515944	0.513787
7	0.484706	0.479156	0.503873	0.520809
8	0.475249	0.405269	0.490420	0.470122
9	0.482331	0.520496	0.483502	0.476586
10	0.478228	0.474135	0.477398	0.446532
11	0.484529	0.490138	0.464942	0.489178
12	0.473674	0.489088	0.457299	0.480318
13	0.457231	0.512951	0.453352	0.489856
14	0.459207	0.480136	0.450713	0.491204
15	0.455582	0.523038	0.441072	0.485245
16	0.451141	0.504539	0.431820	0.456483
17	0.454763	0.476689	0.432829	0.451015
18	0.443473	0.486750	0.434880	0.471863
19	0.439447	0.485960	0.426852	0.463850
Ortalama	0.466949	**0.478278**	0.465061	**0.482779**
En İyi Sonuç	**0.484706**	**0.523038**	**0.515944**	**0.534840**

Tablo 9'da Türkiye hakkında yapılmış haberlerin GDT sonuçları gösterilmiştir. Tablo 10'da ise Türkiye hakkında yapılmış haberlerin NOMF yöntemi sonuçlarına yer verilmiştir. Tablo 9'daki sonuçlarda Suriye, Libya, Yunanistan gibi ülkeler ve Covid-19, ve Türkiye'nin Akdeniz'deki sorun ve atılımları GDT'nin oluşturduğu konu başlıkları arasında yer almıştır. Konu 2'de Doğu Akdeniz'deki tatbikatın Yunanistan ile oluşturduğu gerilim belirlenmiştir. Konu 3'te Türkiye, Suriye ve Libya üçgeni belirlemiş. Konu 4'te Yunanistan'ın Suriye'den gelen göç dalgası ile ilgili vakalarını içeren konuları içermiş olduğu belirlenmiştir. Konu 5 pandemiyi ölüm vakaları üzerinden ayrıştırmıştır. Konu 7 ise küresel pazarda internet ve teknolojinin yer aldığı bir konu başlığını oluşturmaktadır.

Tablo 9. Türkiye verisi için GDT sonuçları

Konu 1	photo team week time food like home year image recipe
Konu 2	mediterranean greec eastern post tension john drill disput relat explor
Konu 3	coronavirus presid libya ankara erdogan govern istanbul syria state minist
Konu 4	greec ministri migrant syria refuge publish kill iraq defens northern
Konu 5	coronavirus case covid death health number updat peopl trade confirm
Konu 6	liverpool class deal ancient best promot entri point bournemouth disast
Konu 7	market stori report internet research year tech global thing miss

Tablo 10. Türkiye verisi için NOMF sonuçları

Konu 1	coronavirus pandem covid world test global help virus china million
Konu 2	libya militari forc russia govern unit state nation call nato
Konu 3	presid erdogan tayyip trump recep donald speak black leader phone
Konu 4	photo imag appear getti alami express post file pakistan tribün
Konu 5	greec mediterranean eastern european tension migrant cyprus island athen union
Konu 6	case death number confirm toll rise covid total health report
Konu 7	minist foreign trade ministri statement interior turkey health meet visit
Konu 8	syria russia defens iraq northern terrorist oper idlib ministri group
Konu 9	market publish today read stori time thing report global busi
Konu 10	agenc ankara anadolu updat coronavirus novel medic measur nation outbreak
Konu 11	bank year compani announc research product central billion deal accord
Konu 12	peopl latest kill caus accord polic suspect arrest die terror
Konu 13	year home court famili go know saudi come liverpool khashoggi
Konu 14	week post team secur middl east meal leagu point game
Konu 15	istanbul citi travel lockdown restrict flight manchest reopen europ provinc

Tablo 10'daki NOMF sonuçlarına göre dış ilişkilerde Rusya, Yunanistan, Libya, Irak ile yaşanılan terör, göçmenlik ve Akdeniz konuları yer alırken Covid-19 virüsü ve yol açtığı etkilerde farklı konular halinde yer aldığı gözlemlenmektedir. Konu 1 covid ve yapılan milyonlarca testi konu edinmiş olduğu anlaşılmaktadır. Konu 2 Libya, Rusya ve Nato kelimelerinin yanında askeri kuvvetin olduğu durumları barındırmıştır. Konu 3 Türkiye başkanı Recep Tayyip Erdoğan ve Amerikan Birleşik Devletleri başkanı Donald Trump arasındaki telefon görüşmesini yansıtmaktadır. Konu 5 sıcak gündem olan Doğu Akdenizi içermiş. Taraflar arasında Kıbrıs, Avrupa Birliği ve Yunanistan'ı da dahil etmiştir. Konu 6 Tablo 9'daki GDT sonuçlarından beşinci konunun benzeri olarak

pandemiyi ölüm vakalarına bağlı olarak derlemiştir. Konu 7'de Türkiye'nin Dış İşleri, Sağlık ve Ticaret Bakanlıklarının toplantıları yer almıştır. Konu 8 dış ilişkilerimize dair bir konu olarak belirlenmiş. Suriye ve Kuzey Irak'taki terörist gruplarını içermektedir. Konu 9'da küresel Pazar ve iş dünyasını ilgilendiren bir konuyu oluşturmuştur. Anadolu ajansının salgının etkilerini güncellemeleri konu 10'u oluşturmuş. Konu 11 Merkez Bankasının yaptığı milyarlık anlaşmayı kapsamaktadır. Konu 13 gazeteci Cemal Kaşıkçı'nın cinayetini içermektedir. Konu 15 Avrupa'ya uçuşların yeniden başlamasını konu olarak derlemiştir.

Tablo 11'de Yunanistan hakkında yapılmış haberlerin GDT sonuçları gösterilmiştir. Tablo 12'de ise Yunanistan hakkında yapılmış haberlerin NOMF yöntemi sonuçlarına yer verilmiştir.

Tablo 11. Yunanistan verisi için GDT sonuçları

Konu 1	europ turkey itali year mediterranean croatia nation albania eastern union
Konu 2	coronavirus cyprus minist franc europ covid prime publish test migrant
Konu 3	select europ histori hour today world nato assassin creed year
Konu 4	travel tourist coronavirus europ olymp holiday summer reopen lockdown open
Konu 5	post world english sport coronavirus event game olymp presid major
Konu 6	coronavirus death case peopl class caus latest entri pandem health

Tablo 12.Yunanistan verisi için NOMF sonuçları

Konu 1	turkey mediterranean eastern cyprus turkish presid tension ankara erdogan militari
Konu 2	select camp migrant island refuge lesbo asylum moria govern quarantin
Konu 3	case death coronavirus class entri report total covid confirm olymp
Konu 4	europ itali travel reopen spain border tourist citi manchest open
Konu 5	caus peopl latest symptom coronavirus mild moder exist older adult

Tablo 11 ve Tablo 12 Yunanistan için toplanılan haber verisinden elde edilen sonuçları göstermektedir. Türkiye verisine nazaran her iki yöntemde konu sayıları olarak daha azdır. Tablo 11'de GDT yönteminin sonuçları incelendiğinde; Konu 1'de Avrupa birliği ve Türkiye'yi de içine alan Akdeniz ülkelerini barındıran bir konu olduğu gözlemlenmektedir. Konu 2'de Avrupa ve pandemi ele alınmıştır. Yapılan testlerin yayınlanmasını içermektedir. Konu 4 pandemideki seyahat kısıtlamalarının yeniden

açılması konunun odağını oluşturmaktadır. Konu 5'te 2020 yılındaki olimpiyat oyunlarını barındırdığı gözlemlenmektedir. Konu 6'da vaka ve ölüm ifadelerinden pandeminin etkilerini içeren bir konu olduğu anlaşılmaktadır.

Genel olarak değerlendirildiğinde, Türkiye tek bir konu içerisinde yer alırken pandemi üç farklı yönden başlıklar oluşturmuştur. Bunların yanında olimpiyatlarda konular arasında yer almıştır.

Tablo 12'de Yunanistan için NOMF yöntemi uygulandığında çıkan sonuçlara yer verilmiştir. Konu 1 Yunanistan'ın Türkiye ile aralarındaki Kıbrıs ve Doğu Akdeniz gerilimini içermektedir. Konu 2 göçmen kampları ve sığınakları olarak belirlenmiş. Konu 3 Pandeminin yol açtığı ölümler olarak tespit edilmiştir. Konu 4 Avrupa ve seyahat engelinin yeniden açılması olarak yer almıştır. Son olarak konu 5 ise pandemi ile ilgili yetişkin insanlara olan bulaşı hakkında olduğu belirlenmiştir.

Yunanistan ile ilgili yapılmış olan haberlerin sonuçları genel olarak değerlendirildiğinde her iki sonuçta da pandemi farklı yönleri ile çoğunluğu oluşturmaktadır. NOMF yönteminde belirgin bir şekilde Türkiye ile yaşadıkları siyasi kriz yer almıştır.

4. SONUÇLAR VE ÖNERİLER

Bu tez çalışması kapsamında indirilen haberler, Metin Madenciliği yöntemlerinden KM algoritması kullanılarak analiz edilmiştir. Bu haberlere erişebilmek için NewsAPI sisteminin kullanıcılara sunduğu https://newsapi.org adresinden indirilen haberler kullanılmıştır. NewAPI yardımıyla elde edilen haberler, Python programlama dili kullanılarak analiz edilmiştir. Python programının metin madenciliğinin kolayca uygulanabilmesi için çok sayıda kütüphanesi mevcuttur ve açık kaynak olduğundan dolayı sürekli kullanıcılar tarafından geliştirilmeye devam etmektedir. Bu kütüphaneler yardımıyla Türkiye ve Yunanistan ülkeleri için indirilen 13 Mart 2020-13 Eylül 2020 tarih aralığındaki haberler belirli ön işleme aşamalarından geçirilmiştir. Ön işlem uygulanan veri setinden GDT ve NOMF algoritmaları yardımıyla konu başlıkları çıkartılmıştır. Bu çalışmada konu sayılarını belirlemek için konu tutarlılığı analizi gerçekleştirilmiştir. Sonuç itibariyle Türkiye verisi için GDT algoritması 7 konu, NOMF algoritması için ise 15 konu en iyi sonucu üretmiştir. Yunanistan verisi için ise GDT algoritması 6 konu, NOMF algoritması 5 konu en iyi sonucu çıkartmıştır. Tespit edilen konularla iki komşu ülkenin gündemi belirlenmiş ve iki KM algoritmasının karşılaştırılması yapılmıştır. Böylece önceden bilinmeyen ve potansiyel olarak ihtiyaç duyulan bilginin çıkarılmasında kolaylık sağlayabilecek bir yöntem sağlanmış olundu.

Türkiye için yapılan analiz sonucundaki konular incelendiğinde dış ilişkiler ağırlıkta siyasi bir gündem olduğu görülmektedir. Yunanistan için olan analizlerde ise tek siyasi gündemin Türkiye ile aralarında yaşandığı tespit edilmiştir.

Türkiye için oluşturulan konu başlıklarının sayısının Yunanistan için oluşturulan konu sayısından fazla olması Türkiye'nin verilerin toplandığı 2020'nin Mart ve Eylül ayları arasında daha yoğun bir gündeme sahip olduğunun bir göstergesi olmuştur.

Türkiye'nin gündemleri arasında Libya, Kıbrıs, Doğu Akdeniz, Suriye savaşı ve beraberindeki göç probleminin yanı sıra pandeminin oluşturduğu seyahat ve karantina gibi kısıtlamalarda yer almıştır.

Yunanistan ise pandemi ve olimpiyatları barındırmıştır. Dış siyaset gündeminde ise yalnızca Türkiye yer almıştır. Türkiye için yapılan analiz sonuçlarında da yerini alan Doğu Akdeniz, Libya ve göçmen krizi Yunanistan'ın 2020 yılının Mart ve Eylül aylarındaki gündemini oluşturmaktadır.

Bu çalışmada görüldü ki, haber kaynağından daha fazla veri elde edilebilmesi ve kaynağın sağladığı haber içeriklerinin daha çok olması halinde sonuçların daha başarılı olacağı düşünülmektedir.

Sonraki çalışmalarda Türkçe olarak haber verileri derlenip yerel medya haberleri için bir KM çalışması yapılarak Türkçe analizlerin geliştirilmesine ve gündemin objektif olarak belirlenmesine katkıda bulunulabilir. Farklı bir çalışmada toplanılan haberlerin yayıncı kuruluşlarına bakarak duygu analizi tespiti yapılabilir. Böylece yayınların nesnelliğine dair fikir edinilebilir.

5. KAYNAKLAR

1. Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. ve Plemmons, R. J., Algorithms and Applications for Approximate Nonnegative Matrix Factorization, Computational Statistics & Data Analysis (CSDA), 52,1 (2007) 155 – 173.

2. Dolgun, M. Ö., Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları, Yüksek Lisans Tezi, Hacattepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2006.

3. Han, J., Kamber, M. ve Pei, J., Data Mining: Concepts and Techniques, 3. Baskı, Morgan Kaufman Publishers, San Francisco, 2012.

4. Blei, D. M. ve Lafferty, J. D., A Correlated Topic Model of Science, Annals of Applied Statistics, 1,1 (2007) 17–35.

5. Ergün, M., Using The Techniques of Data Mining and Text Mining in Educational Research, Electronic Journal of Education Sciences, 6,12 (2017) 180-189.

6. Boyd-Graber, J., Hu, Y. ve Mimno, D., Applications of Topic Models. Now Publishers Inc., 2017.

7. Chowdhury, S., ve Zhu, J., Towards The Ontology Development for Smart Transportation Infrastructure Planning via Topic Modeling, 36. International Symposium on Automation and Robotics in Construction (ISARC), 2019, Canada, 507-514.

8. Yu, J., Lu, Y. ve Muñoz-Justicia, J., Analyzing Spanish News Frames on Twitter during COVID-19—A Network Study of El País and El Mundo, International Journal of Environmental Research and Public Health, 17,15 (2020) 5414-5425.

9. Panasyuk, A., Yu, E. S. ve Mehrotra, K. G., Controversial Topic Discovery on Members of Congress with Twitter, Procedia Computer Science, 36, 2014, 160–167.

10. Zhao, L. T., Guo, S. Q. ve Wang, Y., Oil Market Risk Factor Identification Based on Text Mining Technology, 10. International Conference on Applied Energy, Ağustos 2018, Hong Kong, 3589–3595.

11. Gao, W., Li, P. ve Darwish, K., Joint Topic Modeling for Event Summarization Across News and Social Media Streams, 21. International Conference on Information and Knowledge Management (CIKM), Ekim 2012, Singapore, 1173-1182.

12. Hidayatullah, A. F., Pembrani, E. C., Kurniawan, W., Akbar, G. ve Pranata, R., Twitter Topic Modeling on Football News, 3. International Conference on Computer and Communication Systems, 2018, Japonya, 467-471.

13. Toprak, K., Metin Madenciliği Yöntemleri Kullanarak İllere Göre Haber Analizi, Yüksek Lisans Tezi, KTÜ, Fen Bilimleri Enstitüsü, Trabzon, 2018.

14. Aslan, O., An Analysis of News on Microblogging Systems, Master's Thesis, Yeditepe University, Institute for Graduate Studies in Science and Engineering, İstanbul, 2006.

15. Dhar V., Data Science and Prediction, Communications of the ACM, 56,12 (2013) 64-73.

16. Vermeulen, A. F., Practical Data Science: A Guide to Building the Technology Stack for Turning Data Lakes into Business Assets, Apress, Berkeley, CA, 2018.

17. Aggarwal, C. C., Data Mining: The Textbook, Springer, New York, 2015, 435-440.

18. Waller, M. A., ve Fawcett, S. E., Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management, Journal of Business Logistics, 34,2 (2013) 77–84.

19. Provost, F., ve Fawcett, T., Data Science and Its Relationship to Big Data and Data-Driven Decision Making, Big Data, 1,1 (2013) 51–59.

20. John, G. H., Enhancements to The Data Mining Process, Stanford University, 1997.

21. Frawley, W. J., Pietetsky-Shapiro, G. ve Matheus, C. J., Knowledge Discovery in Databases: An Overview, AI Magazine, 13,3 (1992) 57-70.

22. Bramer, M., Principles of Data Mining, 3. Baskı, Springer, London, 2016.

23. Shearer, C., The CRISP-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing, 5,4 (2000) 13-22

24. Coşkun, C. ve Baykal, A., Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması, Akademik Bilişim, 2011, 1-8.

25. Zaki, M., Introduction to Data Mining, Springer Verlag, 2003.

26. Witten I. H., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, 2. Baskı, Morgan Kaufmann Publishers, San Fransisco, 2000.

27. Çelik, M., Veri Madenciliğinde Kullanılan Sınıflandırma ve Bir Uygulama, Yüksek Lisans Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul, 2009.

28. Ayık, Y. Z., Özdemir, A. ve Yavuz, U., Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10,2 (2010) 441-454.

29. Han, J. ve Cercone, N., Dviz: A System for Visualizing Data Mining, 3. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Nisan 1999 Beijing, China, 390-399

30. Aydın, S. ve Özkul, A. E., Veri Madenciliği ve Anadolu Üniversitesi Açıköğretim Sisteminde Bir Uygulama. Eğitim ve Öğretim Araştırmaları Dergisi, 4,3 (2015) 36-44.

31. Akpınar, H., Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İşletme Fakültesi Dergisi, 29,1 (2000) 1-22.

32. Gujarati, D. N. ve Porter, D. C., Temel Ekonometri, Çeviri: Şenesen, Ü., Literatür Yayınları, İstanbul, 2001.

33. Berry, M. J. A. ve Linoff, G. S., Data Mining Tecniques For Marketing, Sales, and Customer Relationship Management, 2. Baskı, Wiley Publishing, New York, 2004.

34. Moshkovich, H. M, Mechitov, A. I. ve Olson, D. L., Rule Induction In Data Mining: Effect of Ordinal Scales, Expert Systems with Applications, 22 (2002), 303-311.

35. Atalay, M. ve Çelik, E., Büyük Veri Analizinde Yapay Zekâ ve Makine Öğrenmesi Uygulamaları, Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9,22 (2017) 155-172.

36. Han J. ve Fu, Y., Mining Multiple-Level Association Rules in Large Databases, IEEE Transactions on Knowledge and Data Engineering, 11,5 (1999) 798-805.

37. Özekes, S., Veri Madenciliği Modelleri ve Uygulama Alanları, İstanbul Ticaret Üniversitesi Dergisi, 3 (2003) 65-82.

38. Ghahramani, Z., Unsupervised Learning, Summer School on Machine Learning, Springer, 2004, Berlin, Heidelberg, 72-112.

39. Michie, D., Spiegelhalter, D. J. ve Taylor, C. C., Machine Learning, Neural and Statistical Classification, Cambridge University, London, 1994.

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

41. Hastie, T., Tibshirani, R. ve Friedman, J., The Elemenets of Statistical Learning Data Mining, Inference, and Prediction, 2. Baskı, Springer, New York, 2008.

42. Sutton, R. S. ve Barto, A. G., Reinforcement Learning, Cambridge, MA: MIT Press, 1998.

43. Feldman, R. ve Sanger, J., The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data, Cambridge University Press, 34,1 2007.

44. Liu, B., Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, 2. Baskı, Springer, 2011.

45. Li, Y., High Performance Text Document Clustering, Doktora Tezi, Wright State University, Ohio, Amerika Birleşik Devletleri, 2007.

46. Aliguliyev, R. M., Clustering of Document Collection–A Weighting Approach, Expert Systems with Applications, 36,4 (2009) 7904-7916.

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

48. Hull, D. A., Stemming Algorithms: A Case Study for Detailed Evaluation, Journal of The American Society For Information Science, 47,1 (1996) 70-84.

49. Aggarwal, C. C. ve Zhaı, C. X., Mining Text Data, Springer, 2012.

50. Karamollaoğlu, H., Doğru, İ. A. ve Dörteller, M., Makine Öğrenmesi Yöntemleri ile İstenmeyen E-postaların Tespiti, Innovations in Intelligent Systems and Applications Conference (ASYU), 2018, Adana, 1-5.

51. Seçkin, K., Metin Madenciliğinde Kullanılan Yöntemlerin Karşılaştırılması: Siyasi Parti Liderlerinin Grup Genel Toplantı Konuşmaları ile Bir Uygulama. Yüksek Lisans Tezi, Sakarya Üniversitesi, Sosyal Bilimler Enstitüsü, Sakarya, 2011.

52. Salton, G., Allan, J. ve Singhal, A., Automatic Text Decomposition and Structuring, Information Processing and Management, 32,2 (1996) 127-138.

53. Gaikwad, S. V., Chaugule, A. ve Patil, P., Text Mining Methods and Techniques, International Journal of Computer Applications, 85,17 (2014) 42-45.

54. Chang, C., Kayed, M., Girgis, M. R. ve Shaalan, K., A Survey of Web Information Extraction Systems, IEEE Transactions on Knowledge and Data Engineering (TKDE), 18,10 (2006) 1411-1428

55. Russell, S. J. ve Norvig, P., Artificial Intelligence A Modern Approach, 3. Baskı, Pearson Education, New Jersey, 2010, 867-868.

56. Ahmad, P. H. ve Dang, S., A Comparative Study on Text Mining Techniques, International Journal of Science and Research (IJSR), 3,12 (2014) 2222-2226.

57. Freitag, D., Machine Learning for Information Extraction in Informal Domains, Machine Learning, 39,2 (2000) 169–202.

58. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. ve Harshman, R., Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science, 41,6 (1990) 391–407.

59. Altaweel, M., Bone, C. ve Abrams, J., Documents as Data: A Content Analysis and Topic Modeling Approach for Analyzing Responses to Ecological Disturbances, Ecological Informatics, 51 (2019) 82-95.

60. Blei, D. M., Probabilistic Topic Models, Communications of the ACM, 55,4 (2012) 77-84.

61. Blei, D. M. ve Lafferty, J. D., Dynamic Topic Models, Proceedings of the 23. International Conference on Machine Learning (ICML), Temmuz, 2006, Pittsburgh, 113-120.

62. Alghamdi, R. ve Alfalqi, K., A Survey of Topic Modeling in Text Mining, International Journal of Advanced Computer Science and Applications (IJACSA), 6,1 (2015) 147-153.

63. Öztürk, S., Sankur, B., Güngör, T., Yılmaz, M. B., Köroğlu, B., Ağin, O., İşbilen, M., Ulaş, Ç. ve Ahat, M., Türkçe Etiketli Metin Derlemi, IEEE 22. Signal Processing and Communications Applications Conference (SIU), Nisan 2014, Trabzon, 1395-1398.

64. Li, W. ve McCallum, A., Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations, 23. International Conference on Machine Learning, Pittsburgh, 2006, 577-584.

65. Guo, L., Vargo, C. J., Pan, Z., Ding, W. ve Ishwar, P., Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling, Journalism and Mass Communication Quarterly, 93,2 (2016) 332-359.

66. Blei, D. M., Ng, A. Y. ve Jordan, M. I., Latent Dirichlet Allocation, Journal of Machine Learning Research (JMLR) 3 (2003) 993-1022.

67. Griffiths, T. L. ve Steyvers, M., Finding Scientific Topics, Proceedings of the National Academy of Sciences (PNAS), Nisan 2004, 5228-5235.

68. Wang, Y., Bai, H., Stanton, M., Chen, W. Y. ve Chang, E. Y., PLDA: Parallel Latent Dirichlet Allocation for Large-Scale Applications, 5. International Conference Algorithmic Aspects in Information and Management (AAIM), Haziran, 2009, San Francisco, 301-314.

69. Xu, G., Zhang, Y. ve Li, L., Web Mining and Social Networking: Techniques and Applications, Springer, 2011.

70. Paatero, P. ve Tapper, U., Positive Matrix Factorization: A Non-negative Factor Model with Optimal Utilization of Error Estimates of Data Values, Environmetrics, 5,2 (1994) 111-126.

71. Kayacan, C., Watermarking Algorithm Based on Modified Non-negative Matrix Factorization, Master's Thesis, Boğaziçi University, Institute for Graduate Studies in Science and Engineering, İstanbul, 2008.

72. Lee, D. D. ve Seung, H. S., Learning the Parts of Objects by Non-negative Matrix Factorization, Nature, 401,6755 (1999) 788-791.

73. Comero, S., Capitani, L. ve Gawlik, B. M., Positive Matrix Factorisation (PMF) An Introduction to the Chemometric Evaluation of Environmental Monitoring Data Using PMF, Joint Research Centre-Institute for Environment and Sustainability, Luxembourg, 2009.

74. Hoyer, P. O., Non-negative Matrix Factorization with Sparseness Constraints, Journal of Machine Learning Research (JMLR), 5 (2004) 1457–1469.

75. Lee, D. D. ve Seung, H. S., Algorithms for Non-negative Matrix Factorization, 13. Advances in Neural Information Processing Systems (NIPS), 2001, 556-562.

76. https://newsapi.org/docs/endpoints/everything, Documentation, Everything, 12 Haziran 2021.

77. Wallach, H. M., Murray, I., Salakhutdinov, R. ve Mimno, D., Evaluation Methods for Topic Models, Proceedings of the 26. International Conference on Machine Learning (ICML), 2009, 1105-1112.

78. Stevens, K., Kegelmeyer, P., Andrzejewski, D. ve Buttler, D., Exploring Topic Coherence over Many Models and Many Topics, Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Güney Kore, 2012, 952-961.

79. Mifrah, S. ve Benlahmar, E. H., Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus, International Journal of Advanced Trends in Computer Science and Engineering, 9,4 (2020) 5756-5761.

80. Röder, M., Both, A. ve Hinneburg, A., Exploring the Space of Topic Coherence Measures, Proceedings of the 8. ACM International Conference on Web Search and Data Mining (WSDM), Şubat, 2015, Shanghai 399-408.

81. Syed, S. ve Spruit, M., Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation, International Conference on Data Science and Advanced Analytics, Ekim, 2017, Tokyo, 165-174.

82. Aletras, N. ve Stevenson, M., Evaluating Topic Coherence using Distributional Semantics, Proceedings of the 10. International Conference on Computational Semantics (IWCS), Mart, 2013, Potsdam, 13-22.

83. Bouma, G., Normalized (Pointwise) Mutual Information in Collocation Extraction, Proceedings of German Society for Computational Linguistics (GSCL), 2009, 31-40.

84. Servi, T., Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi, Doktora Tezi, Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Adana, 2009.

85. Özdamar, K., Paket Programlar ile İstatistiksel Veri Analizi II (Çok Değişkenli Analizler), 7. Baskı, Kaan Kitabevi, Eskişehir, 2010.

86. Bishop, C. M., Pattern Recognition and Machine Learning, Springer Science and Business Media, Cambridge, UK, 2006.

87. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. ve Zhao, L., Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey, Multimedia Tools and Applications, 78,11 (2019) 15169–15211.

ÖZGEÇMİŞ

Sefa YAY, Tokat Teknik Endüstri ve Meslek Lisesinde Bilişim Teknolojileri Alanı Veritabanı Bölümünü okuyarak 2012 yılında mezun oldu. Yine aynı yıl başladığı Karadeniz Teknik Üniversitesi, Fen Fakültesi İstatistik ve Bilgisayar Bilimleri Bölümündeki lisans eğitimini örgün olarak 2017 yılında tamamladı. Akabinde yüksek lisans eğitimi için Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik ve Bilgisayar Bilimleri Anabilim Dalı yüksek lisans Programına kabul edildi. Yüksek lisans eğitimi sürecine dahil olarak aldığı hazırlık eğitimi ile B1 seviye İngilizce bilgisine sahiptir.