KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

METİN MADENCİLİĞİ YÖNTEMLERİ KULLANARAK İLLERE GÖRE HABER ANALİZİ

YÜKSEK LİSANS TEZİ

Kaan TOPRAK

HAZİRAN 2018
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

METİN MADENCİLİĞİ YÖNTEMLERİ KULLANARAK İLLERE GÖRE HABER ANALİZİ

Kaan TOPRAK

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

Tezin Enstitüye Verildiği Tarih : 24/05/2018
Tezin Savunma Tarihi : 26/06/2018

Tez Danışmanı : Dr. Öğr. Üyesi Uğur ŞEVİK

Trabzon 2018

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İstatistik ve Bilgisayar Bilimleri Anabilim Dalında
Kaan TOPRAK Tarafından Hazırlanan

METİN MADENCİLİĞİ YÖNTEMLERİ KULLANARAK İLLERE GÖRE HABER ANALİZİ

başlıklı bu çalışma, Enstitü Yönetim Kurulunun 29 / 05 / 2018 gün ve 1755 sayılı kararıyla oluşturulan jüri tarafından yapılan sınavda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri

Başkan : Doç. Dr. Kamil ALAKUŞ .....

Üye : Dr. Öğr. Üyesi Uğur ŞEVİK .....

Üye : Dr. Öğr. Üyesi Tolga BERBER .....

Prof. Dr. Sadettin KORKMAZ
Enstitü Müdürü

ÖNSÖZ

Bu tez çalışması kapsamında indirilen haberler, Metin Madenciliği yöntemlerinden konu modelleme algoritması kullanılarak analiz edilmiştir. Gizli Dirichlet Tahsisi yöntemiyle işlem yapılırken Hürriyet gazetesinin kullanıcılara sunduğu https://developers.hurriyet.com.tr adresinden indirilen haberler kullanılmıştır.

Veri seti Türkçe olduğu için, Türkçe karakterlerle ilgili bazı sorunlar olduğu görülmüştür. Türkçe karakterlerinin veri setinden çok fazla olması analiz sonucunu negatif şekilde etkilemektedir. Bu sorunu önlemek için analiz aşamasına geçilmeden önce karakter dönüşümü yapılarak Türkçe karakter sorunu çözülmüş ve analiz daha doğru sonuçlar vermiştir.

Bu tezin çalışma konusu olarak "Metin Madenciliği Yöntemleri Kullanarak İllere Göre Haber Tespiti" konusu seçilmiş ve Türkiye'deki illerin haber analizleri yapılmıştır. İller hakkında 10 konu tespit edilmiştir.

Bu tez çalışmasında bilgisi ve yardımı ile bana her türlü desteği veren, değerli görüşleri ile yol gösteren tez danışmanım Dr. Öğr. Üyesi Uğur ŞEVİK'e çok teşekkür ederim.

Tezin oluşmasında ve bu noktaya kadar gelmemde katkısı olan Dr. Öğr. Üyesi Tolga BERBER'e teşekkür ederim.

Tüm hayatım boyunca maddi ve manevi her zaman beni destekleyen, her adımımda arkamda duran aileme sonsuz teşekkürlerimi sunarım.

Son olarak yüksek lisans eğitimim boyunca yanımda olan, desteğini ve yardımını bir an olsun esirgemeyen Büşra Soylu'ya teşekkür ederim.

Kaan TOPRAK

Trabzon 2018

TEZ ETİK BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum “Metin Madenciliği Yöntemleri Kullanarak İllere Göre Haber Analizi” başlıklı bu çalışmayı baştan sona kadar danışmanım Dr. Öğr. Üyesi Uğur ŞEVİK’in sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deney/analizleri ilgili laboratuvarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma süresince bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 26/062018

Kaan TOPRAK

İÇİNDEKİLER

Yüksek Lisans Tezi

ÖZET

METİN MADENCİLİĞİ YÖNTEMLERİ KULLANARAK İLLERE GÖRE HABER ANALİZİ

Kaan TOPRAK

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik ve Bilgisayar Bilimleri Anabilim Dalı
Danışman: Dr. Öğ. Üyesi Uğur ŞEVİK
2018, 44 Sayfa, 1 Ek Sayfa

Günümüzde teknolojinin gelişmesi ile birlikte insanların haber alma seçenekleri daha da artmıştır. Özellikle günlük gazete ve dergilerin yerine daha kolay erişim sağlanabilen e-gazete ve e-dergiler tercih edilmektedir. Sosyal medya kullanımının yaygınlaşması, kullanıcı yorumlarına verilen önemin değer kazanması ve e-gazete kültürünün günden güne artması ile birlikte metin içerikli verilerin analiz ihtiyacı da ortaya çıkmıştır. Ancak büyük miktardaki bu veriden ihtiyaç duyulan bilginin çıkartılması git gide zorlaşmaktadır. Bu büyük veriden anlamlı bilgiler çıkarmak için son zamanlarda araştırmacılar metin madenciliği yaklaşımları üzerinde çalışmalarını yoğunlaştırmışlardır.Veri madenciliğinde olduğu gibi metin madenciliğinde de verinin analiz edilebilmesi için bazı aşamalardan geçmesi gerekmektedir. Metin verisinin analiz edilebilmesi için verinin ön işlem adımlarından geçirilerek analize hazır hale dönüştürülmelidir. Ön işleme süreci, analiz yaklaşımlarının sonucunu doğrudan etkilediği için en önemli adımlardan biridir. Bu çalışmada, Türkiye'deki her il için yayımlanan haberlere bir konu modelleme yöntemi olan Gizli Dirichlet Tahsisi kullanılarak en yüksek frekansa sahip 10 konu belirlenmiştir. Veri seti olarak Hürriyet gazetesinin açık kaynak veri tabanında bulunan haberler kullanılmıştır.

Anahtar Kelimeler: Veri Madenciliği, Metin Madenciliği, Gizli Dirichlet Tahsisi

Master Thesis

SUMMARY

NEWS ANALYSIS BY CITIES BY USING TEXT MINING METHODS

Kaan TOPRAK

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Statistical and Computer Science Graduate Program
Supervisor: Assist. Prof. Dr. Uğur ŞEVİK
2018, 44 Pages, 1 Page Appendix

Today, with the development of technology, people have more options to receive news. Especially, E-newspapers and e-journals with easier access are preferred. in place of daily newspapers and magazines. The need for analysis of textual content has also emerged with the widespread use of social media, appreciation of user comment. the increase of e-newspaper culture from day to day. However, it is becoming increasingly difficult to extract the information needed from these large quantities. In recent years, researchers have intensified their work on text mining approaches to extract meaningful information from big data. As in data mining, it is necessary to go through some stages in order to analyze the data in text mining. The data must be passed through preprocessing stages and converted to ready for analysis. The preprocessing process is one of the most important steps because it directly affects the result of analysis approaches. In this study, 10 topics with the highest frequency were identified using the Dirichlet Allocation to published news for each provinces in Turkey. The news in the open source database of the Hürriyet newspaper were used as the data set.

Key Words: Data mining, Text mining, Latent Dirichlet Allocation

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

KISALTMALAR LİSTESİ

API	: Uygulama Ara Yüzü
DDİ	: Doğal Dil İşleme
LDA	: Gizli Dirichlet Tahsisi
TDM	: Terim Doküman Matrisi
VUM	: Vektör Uzay Modeli
WWW	: World Wide Web

1. GENEL BİLGİLER

1.1. Giriş

Bilgisayarların hayatımıza girmesi ile birlikte birçok işin kolaylaştığı, yapılacak işlemlerin daha hızlı ve güvenilir bir şekilde yapılabilmektedir. Pek çok farklı alanda bilgisayarların kullanımı, dijital ortamda verinin toplanabilmesi ve saklanabilmesinde ki teknolojik gelişmeler sayesinde, depolanan verilerin hızlı bir şekilde artmasına neden olmuştur. Bu depolanan veriler, sosyal medya kullanımının da yayılması ile kullanıcı yorumları, duygu ve düşünceleri ya da e-gazete ve e-dergilerin internet üzerinden yayımlanması metin içerikli verilerin gün geçtikçe artmasına neden olmuştur.

Metin veri madenciliği, belge bilgi madenciliği ve metin veri tabanlarındaki bilgi keşfi olarak da adlandırılan metin madenciliği büyük boyuttaki metin içerikli veri kaynaklarından, önceden bilinmeyen ve potansiyel olarak ihtiyaç duyulan bilginin çıkarılma işlemidir. Metin içeren veri setleri genellikle yapısal olmayan veriler içermektedir. Metin madenciliği yaklaşımları ise yapısal verilerle işlem yapabilmektedir. Bunun için, yapısal olmayan veriler belirli ön işleme aşamalarından geçirilerek yapısal hale getirilmesi gerekmektedir [1].

Metin madenciliği için veri setleri sadece müşteri bilgileri ya da sosyal medya ile sınırlı değildir. 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 [2].

1.2. Literatür İncelemesi

Günümüzde depolama birimlerindeki teknolojik gelişmeler sayesinde büyük boyutlu verilerin işlenip analiz edilmesinde veri madenciliği yöntemlerinin kullanımı günden güne artmaktadır. Son yıllarda ise web teknolojisinin kullanımındaki kolaylık ve mobil cihazlar sayesinde daha fazla kullanıcıya ulaşılabilmesi imkânı, metin madenciliğini popüler bir hale

getirmiştir. Yakın zamanda yapılan birçok akademik çalışma metin madenciliğinin gelişmesinde önemli rol oynamıştır.

[3] çalışmasında Türkçe twitter mesajlarında Gizli Dirichlet Tahsisine dayalı duygu analizi yapmıştır. Deneysel çalışmaların sonucunda beş farklı algoritmayı karşılaştırmış ve en başarılı sonucu %78,34 ile Naive Bayes algoritması vermiştir [3].

Gizli Dirichlet Tahsisi ile ilgili bir diğer araştırma ise metin derleminde, TF-IDF ve Gizli Dirichlet Tahsisi'nde konu olasılıkları özniteliklerinin sınıflandırma başarısı karşılaştırılmıştır [4].

Bir diğer çalışmada, SKM (Spherical K-Means) algoritmasını, belgelerin K-ortalama döngüsü içinde birden fazla kümeye atanmasına izin verilecek şekilde değiştirerek MCSKM (Multi-Cluster Spherical K-Means) algoritmasını geliştirmişlerdir. Elde edilen sonuçlar, metin kümelemesinde çok etkili olduğu bilinen BKM (Bisecting K-Means) algoritmasının sonuçlarıyla karşılaştırılmış ve daha iyi sonuçlar verdiği görülmüştür [5].

[6] çalışmasında, K-ortalama algoritması gibi çalışan K-medoid kümelemesi için yeni bir algoritma önerilmiş ve başlangıç medoidlerin seçilmesi için farklı yöntem test edilmiştir. Önerilen algoritmada, uzaklık matrisini bir kez hesaplar ve her iteratif adımda yeni medoid bulmak için kullanır. Algoritmayı değerlendirmek için, bazı gerçek ve yapay veri setleri kullanmış ve düzeltilmiş Rand indeksi açısından diğer algoritmaların sonuçlarıyla karşılaştırılmıştır. Deneysel sonuçlar, önerilen algoritmanın, medoidler etrafında bölünmeye karşı karşılaştırılabilir bir performansla hesaplamada önemli ölçüde zaman kaybı olduğunu göstermiştir [6].

Diğer çalışmada, k-ortalama algoritmasının küme merkezlerinin başlangıç yerleşimine oldukça duyarlı olduğu için çeşitli başlatma yöntemleri önerilmiş ve deneysel sonuçlar parametrik olmayan istatistiksel testler kullanılarak analiz edilmiştir. Deneysel sonuçlar, Forgy ve Macqueen gibi popüler başlatma yöntemlerinin sıklıkla kötü performans gösterdiğini ortaya koymuştur [7].

[8] çalışmasında, meme kanseri ve genler arasındaki ilişkiyi belirten aday ilişkilendirme kelimelerini bulmak için iki farklı kümeleme yaklaşımı (Basit Kümeleme ve K-ortalama) uygulamışlardır. Karşılaştırma deneyi, Basit kümelemenin K-ortalama kümelenmesinden daha üstün olduğunu göstermektedir [8].

Türkçe gazete köşe yazarlarının yazarlık öznitelikleri çıkartılarak yazar tanıma olayını gerçekleştiren diğer bir çalışmada, üç kategoriye ayrılan veri setleri Yapay Sinir Ağları (YSA) ve Destek Vektör Makineleri (DVM) ile analiz edilerek karşılaştırılmıştır.

Yapılan çalışmada, farklı kategoride yazan yazarlar ve aynı kategoride yazan yazarlar aralarında karşılaştırıldığından DVM daha iyi sonuç verdiği görülmüştür. Çalışmada, cinsiyete bağlı olan kategoride ise her iki yöntemde aynı başarıyı vermiştir [9].

[10] çalışmasında ise metin madenciliği ile ilgili kümeleme analizi ve uygulamaları yapılmıştır. Bu çalışmada K-ortalama yöntemi ve DBSCAN algoritmaları kullanılmıştır. [10].

Diğer bir çalışmada, gazetelerin web sayfalarında ki haber metinlerinden oluşan 3 farklı veri seti yapay sinir ağları ile otomatik olarak sınıflandırılmaya çalışılmıştır. Yapay sinir ağları ile elde edilen sonuçlar Destek Vektör Makinesi yöntemiyle karşılaştırılmış ve DVM yöntemi iki veri setinde daha iyi sonuç verirken bir veri setinde sonuçlar aynı çıkmıştır [11].

İris ve Vehicle veri setlerini kullanılarak yapılan çalışmada ise, ilk olarak veri setlerine boyut azaltma işlemi yapmadan K-ortalama ve Bulanık C-ortalama (BCO) algoritmaları uygulandıktan sonra Temel Bileşenler Analizi (TBA) ile boyut azaltma işlemi yapılmıştır. İris veri setine boyut azaltma işleminden sonra BCO algoritması uygulandığında elde edilen sonuçlarda Non-negative Matrix Factorization (NNMF) algoritmasının boyut azaltma işleminde daha etkili olduğu gözlenmiştir. Vehicle veri setine BCO algoritması uygulandığında algoritmanın K-ortalamaya göre daha iyi çalıştığı gözlenmiştir. Boyut azaltma işleminden sonra TBA ile elde edilen yeni veri setinde, BCO algoritması sonuçları daha doğru ve ideale yakın olduğu tespit edilmiştir. Sonuçlara bakıldığında, BCO algoritması K-ortalamaya göre kümelemede daha etkili bir algoritma olduğu görülmüştür. Boyut azaltma metotları ile veri setlerinin sahip olduğu özellikler kaybedilmeden doğru ve hızlı sonuçlar üretilebilmektedir [12].

Türkçe içerikli reklam epostalarının metin madenciliği ile otomatik olarak tespit edilen çalışmada, 800 e-postadan oluşan veri kümesi 3 farklı sınıflandırma algoritmasıyla sınıflandırılmaya çalışılmıştır. En yüksek sınıflandırma başarısına Naive Bayes ve En Yakın N-Komşuluk (K-NN) algoritmalarıyla ulaşmışlardır [13].

K-NN algoritmasıyla ilgili bir diğer çalışmada ise, R programlama dili kullanarak bilimsel makale tasnifi yapılmış ve %96,67 başarı oranı elde edilmiştir [14].

1.3. Veri Madenciliği

Veri madenciliği, bilgisayar ortamında saklanan büyük boyuttaki verilerden çeşitli yaklaşımlar yardımı ile ihtiyaç duyulan bilgiye ulaşım olarak tanımlanmaktadır. Geniş anlamda veri madenciliği, hipotez oluşturma ve doğrulama, modelleme ve yorumlama gibi işlem sonrası adımlara ek olarak, veri çıkarma, veri indirgeme ve veri temizleme gibi ön işleme görevlerini içeren bilgiye erişim yaklaşımıdır [15].

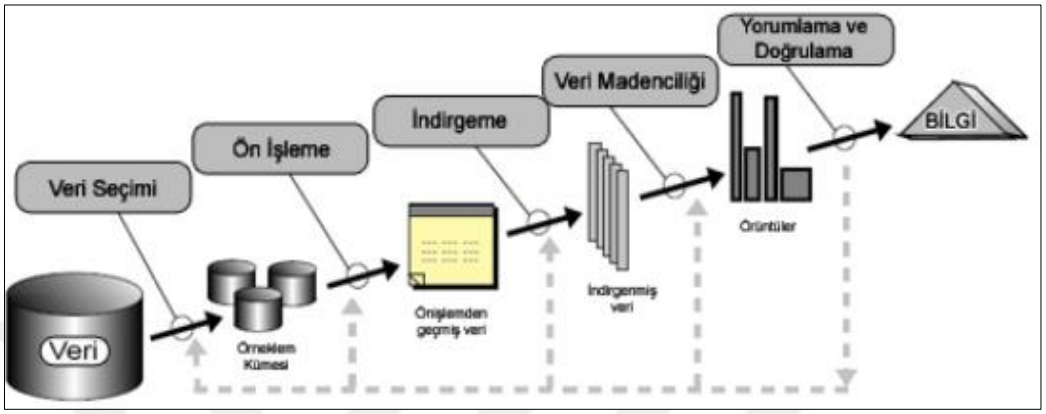
Veri madenciliği kapsamında, sınıflandırma ve tahmin, kümelenme, aykırı değer saptama, ilişkilendirme kuralları, dizi analizi, zaman serileri analizi, metin madenciliği, sosyal ağ analizi ve duygu analizi gibi özel analiz teknikleri bulunmaktadır. İş dünyasında ise iş anlayışı, veri anlama, veri hazırlama, modelleme, değerlendirme ve uygulama olmak üzere veri madenciliği süreci altı ana başlıkta toplanabilmektedir [16].

Veri madenciliği içerisinde, istatistik, yapay zekâ, makine öğrenmesi, doğal dil işleme ve veri tabanı gibi çeşitli disiplinleri barındırmaktadır. Veri madenciliğinde kullanılan bazı yaklaşımları, Sınıflandırma (Classification), Kümeleme (Clustering), Zaman Serisi Analizi (Time Series Analysis), Metin Madenciliği (Text Mining), Duygu Analizi (Sentiment Analysis) ve Sosyal Ağ Analizi'dir (Social Network Analysis) şeklinde sıralanabilir.

Veri madenciliği yaklaşımlarını, iş dünyası ve akademi başta olmak üzere birçok şirket, kurum ve kuruluşlar ihtiyaç duyulan bilgiye erişim için kullanmaktadır. Ayrıca finans ve sağlık gibi devlet uygulamalarında hayatı kolaylaştıran birçok sistem geliştirilmiştir. Örneğin, ülkemizde, emniyet birimleri için suç istatistiklerine dair online raporlama, hangi profildeki insanların ne tür suçlara meyilli olduklarını belirleme gibi yeni geliştirilen karar destek sistemleri kullanılmaktadır. Benzer şekilde, sağlık sektöründe belirli yaş aralıklarındaki hastaların nelerden şikâyet ettiğini hızlı ve doğru şekilde belirleyebilmek için veri madenciliği yöntemleri kullanılmaktadır. Bunun yanı sıra, bürokrasi nedeniyle hantallaşan bazı devlet hizmetleri e-devlet uygulaması üzerinden vatandaşlara sunulmuştur ve hangi devlet kurumlarının sayfalarının hangi amaçla kullanıldığı tespit edilerek bu hizmetlerin kalitesi arttırılabilmektedir.

1.3.1. Veri Madenciliği Aşamaları

Veri madenciliği süreci genel olarak beş temel aşamadan meydana gelmektedir.



Şekil 1. Veri Madenciliği Aşamaları [17]

Veri Seçimi: Bu aşama, veri madenciliğinde en çok zaman alan kısımlarından biridir. Bundan sonraki bütün aşamalar seçilen bu veri seti üzerinden yapılacağı için verinin seçilmeden önce iyi bir şekilde analiz edilmesi ve bilgi edinmek istenen problemle ilişkilendirilmesi şarttır.

Ön İşleme: Ön işleme aşamasına verinin gürültülü olduğu durumlarda ihtiyaç duyulur. Gürültülü veri, istenilen amaca uygun olmasına rağmen bilgi elde etmeyi zorlaştıran verilerdir. Bu aşamada sonraki aşamaları için veri uygun hale getirilir. Verinin içinde eksik ya da kayıp varsa, eksik veri, veri kümesinden atılırken kayıp verinin yerine verinin ortalama değeri yazılarak bir çözüm üretilebilir. Ön işleme aşamasının başarısı analiz yaklaşımlarının performanslarını direk olarak etkilemektedir. Bu nedenle, ön işlemenin etkin bir şekilde yapılması çok önemlidir. Ön işleme aşaması ne kadar iyi yapılırsa analiz sonucu da o kadar doğru ve güvenilir çıkmaktadır [18].

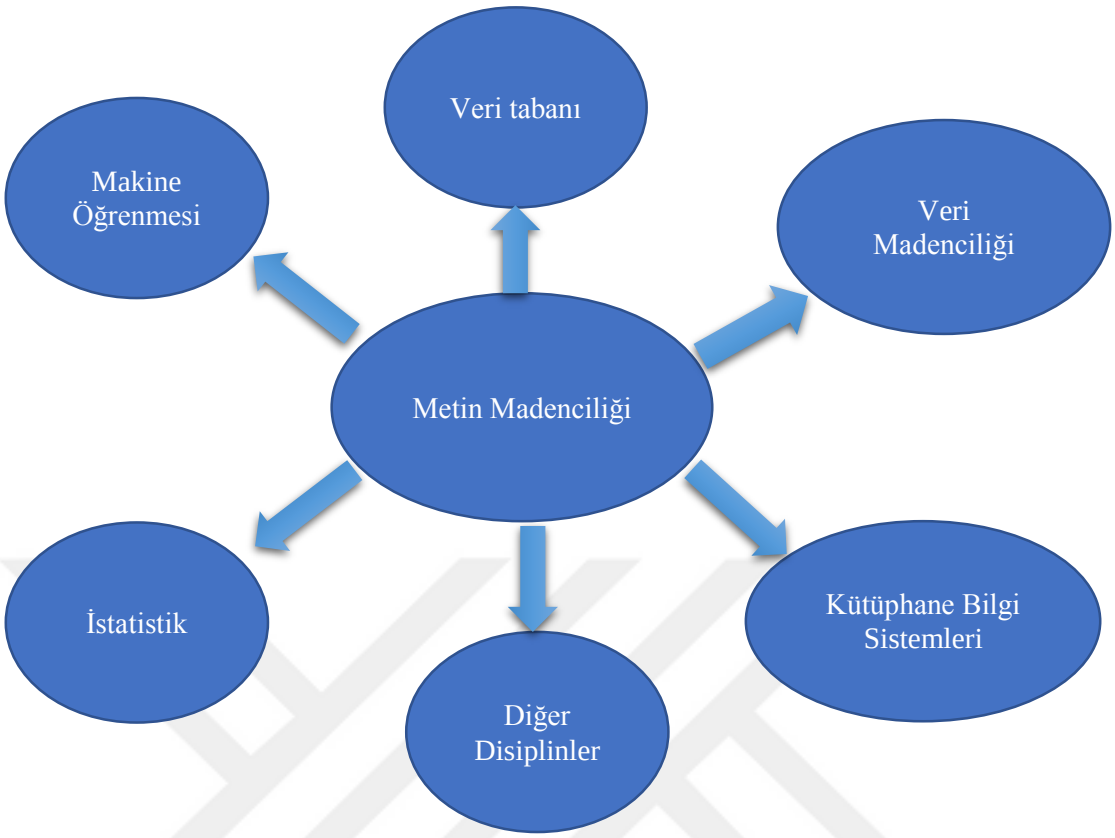
İndirgeme: Veri madenciliği çalışmalarında kullanılacak veriler her ne kadar ön işleme aşamasından geçseler de sonraki aşamalarda kullanılmak için tam anlamıyla uygun olmayabilir. Bu nedenle, sonuca etkisinin az olduğu düşünülen verilerin ya da değişkenlerin veri setinden atılması analiz sonucunun daha güvenilir olmasına ve performans artışına yardım etmektedir. Böylelikle veri boyutu azaltılarak hem daha etkin sonuçlar elde edilebilir, hem de veri boyutu küçüldüğü için hesaplama maliyeti azalır.

Veri madenciliği: Ön işleme ve veri indirgeme aşamalarından geçerek analiz yapmaya hazır hale getirilen veri, veri madenciliği aşamasında veriye uygun yaklaşımlarla analiz edilir. Veri madenciliği yöntemlerini uygulamak için birçok yaklaşım mevcuttur. Her yaklaşımın kendisine ait karakteristik özellikleri ve uygulama alanları vardır.

Yorumlama ve Değerlendirme: Analize uygun hale getirilen veri, en etkin veri madenciliği yöntemleri uygulandıktan sonra çıkan sonuçlar yorumlanır ve elde edilen sonuçlar yapılan diğer çalışmalarla karşılaştırılıp doğrulanabilir [17], [19].

1.3.2. Veri Madenciliği Yaklaşımları

Veri madenciliği, büyük veri kaynaklarından daha önce bilinmeyen verinin keşfi için istatistik, yapay zekâ, makine öğrenmesi ve sinir ağları gibi farklı yaklaşımları kullanan bir süreçtir. Veri madenciliği ve istatistik arasında büyük bir örtüşme vardır. Aslında veri madenciliğinde kullanılan tekniklerin çoğu çok değişkenli istatistiksel yaklaşımlardan oluşmaktadır. Fakat, veri madenciliği teknikleri geleneksel istatistik teknikleri ile aynı değildir. Geleneksel istatistiksel yöntemler, genel olarak bir modelin doğruluğunu onaylamak için çok sayıda kullanıcı etkileşimi gerektirir. Ayrıca, istatistiksel yöntemler genellikle çok büyük veri kümelerini iyi ölçemez ve hipotezleri test etmeye veya daha büyük bir popülasyonun daha küçük temsili örneklerine dayanan korelasyonları bulmaya dayanır. Veri madenciliği ise büyük veri kümeleri için daha uygun olmaktadır [20].



Şekil 2. Veri Madenciliğinin Diğer Disiplinlerle İlişkisi [21]

Veri madenciliğinin yaklaşımları genel olarak üç bölümden oluşmaktadır.

İstatistiksel Yaklaşımlar: Veri madenciliği ile istatistiğin ortak özelliklerinden biri verinin ihtiyaç duyulan bilgiye dönüştürülmesi veya veriden çıkartılmasıdır. Her iki yaklaşımda verinin anlamlarını belirlemek için kullanılır. Veri madenciliği için Regresyon Analizi, Korelasyon Analizi ve Bayes Ağı gibi birçok istatistiksel yaklaşım kullanılabilir[21].

Makine Öğrenmesi Yaklaşımı: Makine öğrenmesi, veriden öğrenilen bilgi ile yine o verinin analiz edilmesini sağlayan yapay zekanın alt dalı olarak tanımlanabilir. Makine öğrenimi algoritmaları, veri madenciliği için neredeyse bir ön koşuldur, ancak bunun tersi doğru değildir. Diğer bir ifadeyle, makine öğrenimi veri madenciliği içermeyen görevlere de uygulanabilir ancak veri madenciliği yöntemlerini kullanılıyorsa, makine öğrenimi yaklaşımları da kullanılmaktadır. Makine öğrenmesi yöntemleri denetimli (supervised) ve denetimsiz (unsupervised) olmak üzere ikiye ayrılmaktadır. Denetimli makine öğrenmesi veriyi test ve eğitim veri kümesi olarak ikiye bölmektedir. Eğitim kümesinden öğrenilen bilgi, test kümesine uygulanır ve analiz sonucunun güvenli olup olmadığına karar verilir.

Denetimli makine öğrenmesine sınıflandırma ve regresyon tabanlı yaklaşımlar örnek olarak verilebilir. Denetimsiz makine öğrenmesinde ise algoritmalar girdiler arasındaki benzerlikleri tespit etmeye çalışarak, ortak özellikleri olan girdileri birlikte gruplara ayırır. Denetimsiz makine öğrenmesine örnek olarak kümeleme ve birliktelik analizi yaklaşımları örnek olarak verilebilir. Makine öğrenmesi veri toplama, veri keşfetme ve hazırlama, veri içinden eğitim verisi oluşturma, modelin performansını değerlendirme ve model performansını geliştirme şeklinde beş adımda gerçekleştirilir. Bu adımlardan sonra makine öğrenmesi tamamlanır. Eğer model performansı tatmin edici derecede ise model hedeflenen problem için uygulanabilir [22], [23].

Diğer Yaklaşımlar: Veri madenciliği için veri tabanı ve sinir ağları gibi başka yaklaşımlarda bulunmaktadır.

1.3.3. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemlerinde çeşitli algoritmalar mevcuttur. Bu algoritmalar üç ana başlıkta incelenebilir. Bu başlıklar sınıflandırma, kümeleme ve birliktelik analizidir.

1.3.3.1. Sınıflandırma

Sınıflandırma, sınıf değerleri önceden bilinen veri kümesinden elde edilen model ile, sınıfı bilinmeyen yeni verinin sınıfının tahmin edilmesi işlemine denir. Sınıflandırmada, eğitim kümesi olarak isimlendirilen ve sınıf değeri önceden bilinen bir veri kümesine ihtiyaç vardır. Temel olarak iki aşamada gerçekleşir. İlk aşamada, veri kaynağı eğitim kümesi ve test kümesi olarak ikiye bölünür. Sınıflandırmanın başarısı için veri kümesi ile eğitim kümesi oranı iyi seçilmelidir. Aksi halde overfitting ve underfitting olarak adlandırılan bazı sorunlarla karşılaşılır. Overfitting, eğitim kümesinin çok fazla seçilmesinden kaynaklanan bir ezber problemidir. Algoritmalar, eğitim verisinin modele etkisi çok az olan detaylarını ya da gürültüsünü ezberleyebilmektedir. Bu ezber problemi ise modeli olumsuz yönde etkileyerek modelin güvenilirliğini azaltabilir. Underfitting ise, eğitim kümesinin çok az seçilmesinden kaynaklanan bir problemdir. Underfitting, hem eğitim verilerini modellerken hem de test veri kümesine genelleme yaparken düşük başarı oranına neden olur. Bu gibi sorunlarla karşılaşmamak için eğitim kümesi ile veri

kümesinin oranı çok iyi seçilmelidir. İkinci aşamada ise eğitim kümesindeki verilerle üretilen model, test veri kümesindeki verilerin sınıflarını bulmaya çalışır. Sınıflandırma yaklaşımlarından bazıları Destek Vektör Makineleri, Naive Bayes, Karar Ağaçları, K-En Yakın Komşu şeklinde örnek olarak verilebilir [13], [24].

1.3.3.2. Kümeleme

Kümeleme analizi, birey veya nesneleri benzerliklerine göre kümelere veya gruplara ayırmak için kullanılan çok değişkenli istatistik analiz tekniğidir. Kümeleme, veri kümelerinde bulunabilecek gizli kalıpların belirlenmesinde kullanılmaktadır. Eğitim veri kümelerini kullanarak önsel bilgi ile analiz yapan sınıflandırma ve regresyon yaklaşımlarından farklı olarak kümeleme analizi, sınıf etiketlerine başvurmadan gözlem değerlerini gruplamaya çalışır. Kümeleme analizi veriler arasındaki benzerliklere göre yapılır. Küme analizi, grup içerisindeki verilerin yüksek benzerliğe sahip olması, gruplar arası benzerliğinde çok düşük olmasını amaçlamaktadır. Diğer bir deyişle, küme içi mesafeleri minimum, kümeler arası mesafeleri maksimum yapılmaya çalışılır. Kümeleme analizi hiyerarşik ve hiyerarşik olmayan kümeleme olarak iki başlık altında toplanabilir [25].

1.3.3.2.1. Hiyerarşik Yöntemler

Aşamalı kümeleme olarak da adlandırılan bu yöntem, veri setindeki gözlem değerlerinin birbirlerine olan uzaklık değerlerini kullanarak, veri setindeki bu değerlerin hiyerarşik ayrıştırmasını sağlar. Kümelenme iki farklı şekilde gerçekleştirir. İlk yöntem Gruplayıcı (*Agglometrive)* yöntem olup, süreç başlangıcında her değer ayrı bir küme olarak ele alınır. Daha sonraki adımlarda en yakın iki gözlem yeni bir kümede birleşir. Böylece her adımda küme sayısı azalır. Ayırıcı (Divistive) yöntemde ise, başlangıç da bütün gözlemler bir küme olarak ele alınır ve benzer olmayan gözlemler yeni kümelere ayrılarak küme sayısı azaltılıp, her bir değer ayrı bir küme olana kadar devam eder [26].

Tek bağlantı yöntemi: En yakın mesafe temeline dayanmaktadır. Uzaklık matrisini kullanarak birbirine en yakın gözlemleri birleştirmeyi hedeflemektedir. İlk olarak en yakın

iki gözlem bir küme oluşturur. Daha sonra en yakın diğer gözlemler bu kümeye eklenir ya da yeni bir küme oluşturulur.

Tam bağlantı yöntemi: Tek bağlantı yönteminden farkı en uzak mesafe esasına dayanmaktadır.

Ortalama bağlantı yöntemi: Bir kümenin ortasına düşen gözlemi esas alarak işlemleri yapar. Aşırı değerlerden fazla etkilenmemektedir.

Merkezi yöntem: Bir kümeyi oluşturan gözlemlerin ortalaması esasına dayanarak çalışır. Eğer kümede tek gözlem var ise o gözlemin değeri merkez olarak kabul edilir.

Varyans yöntemi: Yönteme her birinin içinde tek bir gözlem bulunan n küme ile başlanır. İlk aşamada her gözlem bir küme olduğu için hata kareler toplamı sıfırdır. Her aşamada, iki alt küme bir sonraki seviyeyi oluşturmak için birleştirilir [27], [28].

1.3.3.2.2. Hiyerarşik Olmayan Yöntemler

Bazı durumlarda küme sayısı önceden bilinmektedir ve araştırmacı bu küme sayısına göre çözümler üretmek durumundadır. Hiyerarşik olmayan kümeleme yöntemleri, n adet birimden oluşan veri setini başlangıçta belirlenen $k<n$ olmak üzere k adet kümeye ayırmak için kullanılır [27].

K-ortalama yöntemi: En çok bilinen kümeleme algoritmasıdır. K ortalama tekniğinde küme sayısı önceden bilinmektedir ve her iterasyonda oluşan kümeler için değişkenlerin ortalamaları alınır. K-ortalamalar yönteminde kümelerde herhangi bir veri çakışması olmaz. Bir kümenin her elamanı ait olduğu kümeye her zaman daha yakındır. Başlangıç da her biri rasgele tek gözlemden oluşan k tane kümeye atanarak işlem başlatılır. Bu k küme başlangıç küme merkezlerini oluşturmaktadır. İkinci adımda, her gözlem en yakın veya benzer küme merkezine atanır. Üçüncü adımda, her yeni oluşan küme değişkenlerinin ortalamasına göre güncellenir. İki ve üçüncü adımlar küme atamalarında herhangi bir değişiklik olmayana kadar devam ettirilir [10], [29].

1.3.4. Vektör Benzerlik ve Uzaklık Ölçüleri

Öklid Uzaklığı: Öklid uzaklığı çok boyutlu uzaydaki nesnelerin birbirine geometrik uzaklığıdır. İki gözlem arasındaki Öklid uzaklığı

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{in}-x_{jn})^2} \quad (1.1)$$

formülü ile hesaplanır. İki vektör arasındaki Öklid uzaklığı vektörlerin her elemanının farklarının karelerinin toplamının karekökü alınarak hesaplanmaktadır.

Kosinüs Benzerliği: Kosinüs benzerliği, iki vektör arasındaki açının kosinüs değeri hesaplanarak vektörlerin benzerliği bulunur. Kosinüs benzerliğinin en önemli özelliği vektörler arasındaki açı ile işlem yapıldığı için vektör uzunluğundan etkilenmemesidir. Vektörler arasındaki açı küçüldükçe vektörlerin benzerlikleri artmaktadır. Çünkü aradaki açı 0'a yaklaştıkça kosinüs değeri 1'e yaklaşmaktadır ve vektörlerin benzerliği artacaktır. Kosinüs benzerliği,

$$\cos(\theta)=\frac{d*d^*}{|d|\left|d^*\right|}=\frac{\sum_{i=1}^{n} d_i d_i^*}{\sqrt{\sum_{i=1}^{n}(d_i)^2}\sqrt{\sum_{i=1}^{n}(d_i^*)^2}} \quad (1.2)$$

formülü ile hesaplanmaktadır. Burada d ve d^* birbirinden farklı iki belgeyi temsil eden çok boyutlu vektörlerdir. Kosinüs benzerliği bu iki vektörün iç çarpımlarının vektörlerin uzunluğuna bölünmesi ile elde edilmektedir.

Pearson Uzaklık Ölçüsü: Pearson uzaklık ölçüsü de kosinüs benzerliği gibi vektörler arası açıdan yararlanarak vektörlerin benzerliğini hesaplamaktadır. Kosinüs benzerliğinden farkı iki vektörün iç çarpımı yapılmadan önce her birinin ayrı ayrı ortalama değerleri hesaplanır ve her ortalama değer ait olduğu vektörün tüm elemanlarından çıkarılır.

$$d(i,j)=\sqrt{\frac{(x_{i1}-x_{j1})}{S_1^2}+\frac{(x_{i2}-x_{j2})}{S_2^2}+\cdots+\frac{(x_{in}-x_{jn})}{S_n^2}} \quad (1.3)$$

Denklemdeki i ve j, farklı iki belgeyi temsil eden çok boyutlu vektörlerdir. Burada S_n, uzaklığın hesaplandığı değişkenlere ait varyans değeridir. Bununla birlikte, farklı gruplar hakkında önceden bilgi sahibi olunmadığı için, uzaklık hesaplanmasında S değerinin kullanılması doğru olmayabilir. Bu nedenle, Pearson Uzaklık Ölçüsü yerine genellikle Öklid uzaklık ölçüsü tercih edilir.

Manhattan City Block Uzaklık Ölçüsü: Bu uzaklık ölçüsü iki vektörün toplamı ile elde edilir. Manhattan city block uzaklık ölçüsü

$$d(i,j)=(|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\ldots+|x_{in}-x_{jn}|) \quad (1.4)$$

formülü ile hesaplanır.

Minkovski Uzaklık Ölçüsü: P sayıda değişken alınarak aralarındaki uzaklığın hesaplanması için kullanılır. Minkovski uzaklık ölçüsü

$$d(i,j)=\left[\sum_{k=1}^{p}\left(|x_{ik}-x_{jk}|\right)^{m}\right]^{\frac{1}{m}} \quad (1.5)$$

formülü ile hesaplanmaktadır. Minkovski uzaklık ölçüsünde m=2 yazılırsa Öklid uzaklığı, m=1 için Manhattan city block uzaklık ölçüsü elde edilir [30], [31],[32].

1.4. Metin Madenciliği

Veri madenciliği son yıllarda bilgisayar teknolojilerinin büyük gelişme göstermesinden dolayı son zamanlarda çok hızlı bir gelişme göstermiştir ve bu durum çeşitli veri türlerine uygulanan alt veri madenciliği alanı oluşturmaktadır. Bu büyük miktarda metin içeren web ve sosyal ağ analizlerinin yapılabilmesini sağlamaktadır [9].

Metin madenciliği, doğal dille yazılmış büyük metinlerden ilgili bilgileri titizlikle çıkaran ve ilginç ilişkiler, sözdizimsel korelasyon veya semantik ilişki arayışında olan bir veri madenciliği yöntemidir. Metin madenciliği ile ilişkili algoritmalar, metin kümeleme, metin sınıflandırma, doğal dil işleme (DDİ) ve web madenciliğini içerir.

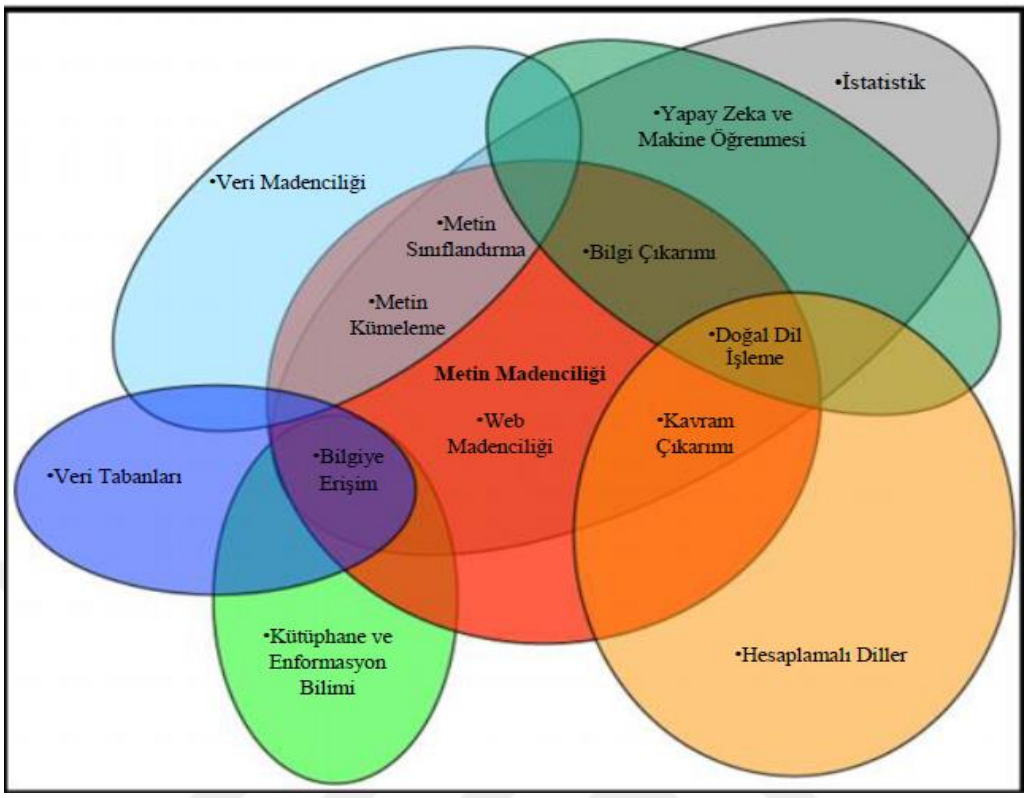
Metin madenciliği, analiz edilmek üzere yapılandırılmamış metin verileri oluşturarak veri tabanlarındaki mevcut verileri arttırmak için kullanılabilecek yeni bir teknolojidir. Bu, klasik veri madenciliğinin bir uzantısından metinlere, "dünyanın kendisi hakkında yeni gerçekleri ve eğilimleri keşfetmek için büyük çevrimiçi metin koleksiyonlarının kullanılması" gibi daha karmaşık formülasyonlara kadar çeşitli farklılıklara imkân sağlamaktadır. Metin madenciliği, veri madenciliği, doğal dil işleme, istatistik ve

bilgisayar bilimi gibi çeşitli yaklaşımları içerisinde bulunduran disiplinler arası bir araştırma alanıdır [33].

Metin madenciliği her türlü metin içerikli belgeleri veri kaynağı olarak kabul eder. Metin madenciliğinin amacı, yapılandırılmamış (metinsel) bilgileri işlemek, metinden anlamlı sayısal indeksler çıkarmak ve böylece metindeki çeşitli veri madenciliği (istatistik ve makine öğrenimi) algoritmalarına ait bilgilerin erişilebilir olmasını sağlamaktır. Yararlı veya ilginç kalıplar, modeller, eğilimler veya yapılandırılmamış metinden kurallar oluşturma süreci olan metin madenciliği, veri madenciliği tekniklerinin, metinden gelen bilginin otomatik olarak çıkartılması uygulanmasını tanımlamak için kullanılır [34]. Şekil 3'te Metin madenciliğinin diğer bilim dalları ile ilişkisi gösterilmiştir.

Metin madenciliği süreci şu şekilde açıklanmaktadır:

- Metin madenciliği, raporlar, mektuplar ve benzerleri olan metin corpusunun hazırlanması ile başlar.
- İkinci adım, metin corpusunu temel alan yarı yapılandırılmış bir metin veri tabanı oluşturmaktır.
- Üçüncü adım, terim frekansının dahil edildiği bir terim-döküman matrisi oluşturmaktır.
- Son aşama ise, metin analizi, anlamsal analiz, bilgi alma ve bilgi özetleme gibi daha ileri analizi kapsamaktadır [35].



Şekil 3. Metin Madenciliği ve İlişkili Olduğu Alanlar [36].

1.4.1. Veri ve Metin Madenciliği

Veri madenciliği, veride kalıp aramak gibi basit bir şekilde tanımlanabildiği gibi, metin madenciliği metinde kalıp aramakla ilgilidir. Fakat bu iki yaklaşım arasındaki yüzeysel benzerlik gerçek farkların görülmesine engel olabilmektedir. Veri madenciliği, veriden önceden bilinmeyen ve potansiyel olarak yararlı bilginin çıkarılması olarak tanımlanabilir. Metin madenciliğinde çıkarılacak bilgiler metinlerde açık ve anlaşılır bir şekilde belirtilmiştir. Buradaki problem, bilginin otomatik işlemeye uygun bir şekilde yapılamamasıdır. Metin, bir insan aracılığına gerek olmaksızın doğrudan bilgisayar tarafından kullanılmaya uygun bir biçimde çıkarmaya çalışmaktadır. Felsefi olarak net bir farklılık olsa da bilgisayar açısından sorunlar oldukça benzerdir.

Hem veri hem de metin madenciliğinde ortak olan bir gereksinim, çıkarılan bilginin potansiyel olarak yararlı olması gerektiğidir. Bu işlenmeye hazır anlamına gelmektedir ve işlemlerin otomatik olarak yapılmasına temel oluşturabilmektedir. Metin madenciliğide bu kavram aynı veri kaynağından gelen veriler üzerinde önemli tahminlerin yapılması anlamına gelmektedir. Aynı problem üzerinde farklı veri madenciliği yöntemlerini

karşılaştırmak için istatistiksel teknikler uygulanarak başarı ve başarısızlık performansı ölçülebilmektedir. Fakat Metin madenciliğinde bunu tanımlamak daha zordur [37].

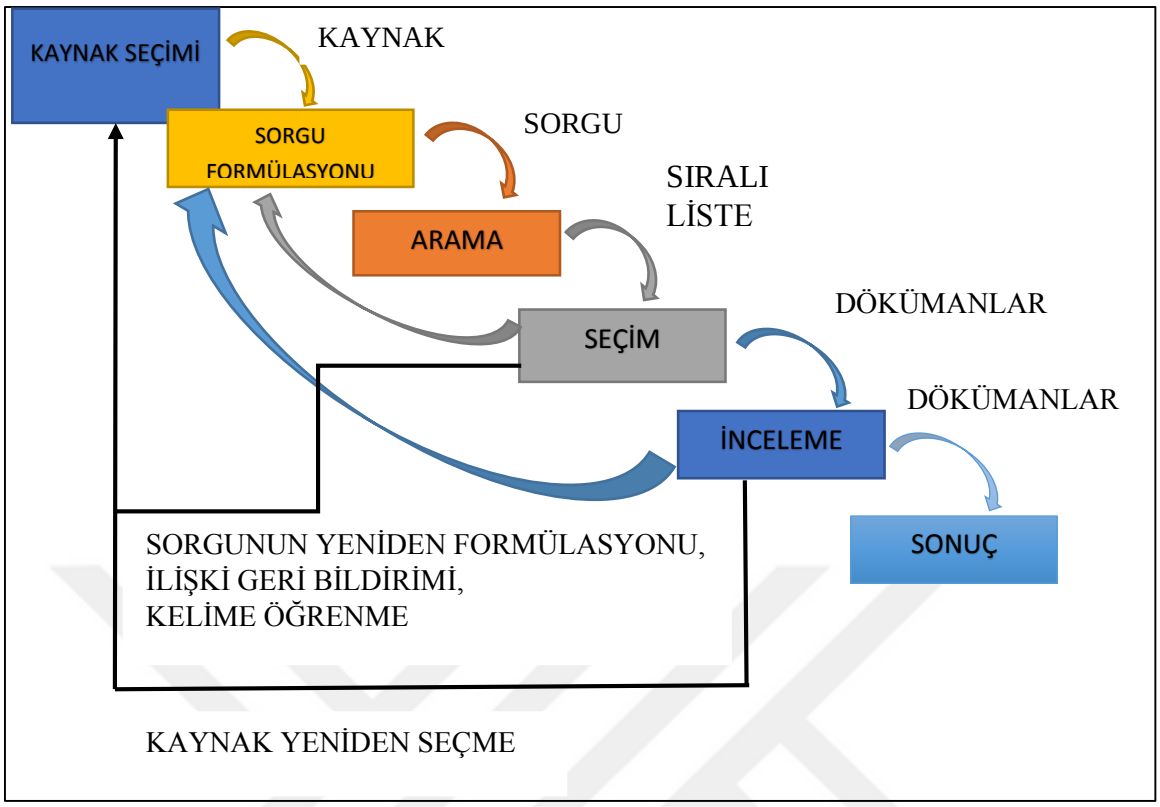
1.4.2. Metin Madenciliği Yöntemleri

Veri madenciliğinde olduğu gibi metin madenciliğinde de verilerin güvenilir bir şekilde analiz edilebilmesi için bazı işlemlerden geçirilmesi gerekmektedir. Bu hem metin madenciliği algoritmalarının uygulanması için hem de analiz sonuçlarının doğru olması için gereklidir.

1.4.2.1. Bilgiye Erişim (Information Retrieval)

Bilgi erişimi, metin madenciliğinin ilk adımı olup, yapılandırılmamış dokümanlardan yapılandırılmış bilgiyi çıkarmaya çalışılır. Belgeler genellikle doğası gereği yapılandırılmamış ve çok miktarda metinsel veri içermektedir. Diğer bir deyişle, çeşitli veri kaynaklarından gerekli bilgiyi elde etmeye yarayan yöntemdir. Yapılandırılmamış veri bilgisayar için temiz, kolay veya anlaşılır olmayan verilerdir. Bilgi erişimi, genellikle bilgisayarlarda saklanan büyük koleksiyonların içinden bir bilgi ihtiyacını karşılayan ve genellikle metinlerden oluşan, yapısal olmayan bir veriyi bulmakta kullanılmaktadır. Bilgi erişimi aynı zamanda, doküman koleksiyonlarına kullanıcıların göz atmasını, filtreleme veya bir dizi kurtarılmış dökümanın işlenmesini de kapsamaktadır.

Geçtiğimiz on yılda bilgi erişim sürecinin birçok farklı uygulamaları geliştirilmiştir. Bunlardan en yaygın olarak kullanılanlarından biri, World Wide Web (WWW)'de bilgi aramaktır. Google, Bing ve Yahoo gibi pek çok arama motoru, birçok gelişmiş yöntemler kullanarak bu işlemi kolaylaştırmaktadır. Ayrıca online kütüphanelerin çoğu, kullanıcılarının bilgiye erişim tekniklerine dayanarak arama yapmasını sağlamaktadır. Genel olarak, bir bilgi alma sisteminin verimliliği, bir kullanıcının sorgusunu bir corpus içindeki en alakalı belgelerle eşleştirme becerisinde yatmaktadır. Bir bilgiye erişim sisteminin verimliliğini artırmak için belgelerin orijinal içeriğine göre organize edilmesi gerekmektedir [38]



Şekil 4. Bilgiye Erişim Döngüsü [39]

Bilgi erişimi için kullanılan iki ölçüt vardır.

- Doğruluk (Recall): Doğruluk, bir bilgi sisteminin sorgu ile ilgili olarak bulduğu dokümanların içinde gerçekten sorgu ile ilgili olan doküman sayısının veri tabanında bulunan ilgili doküman sayısına oranını ile ifade edilmektedir.
- Duyarlılık (Precision): Duyarlık bir bilgi sisteminin sorgu ile ilgili olarak bulduğu dokümanların içinde kullanıcının istediği dokümanların sayısının arama sonucu bulunan tüm dokümanların sayısına oranıdır.

Tablo 1.Karışıklık Matrisi

		TAHMİN	
		Doğru (Elde Edilen)	Yanlış (Elde Edilemeyen)
GERÇEK	Doğru (Uygun)	Doğru Positif (DP)	Yanlış Negatif (YN)
	Yanlış (Uygun Değil)	Yanlış Pozitif (YP)	Doğru Negatif (DN)

$$\text{Doğruluk}=\frac{DP}{DP+YN} \tag{1.6}$$

$$\text{Duyarlılık}=\frac{DP}{DP+YP} \tag{1.7}$$

formülleri ile hesaplanmaktadır [40].

1.4.2.2. Bilgi Çıkarımı (Information Extraction)

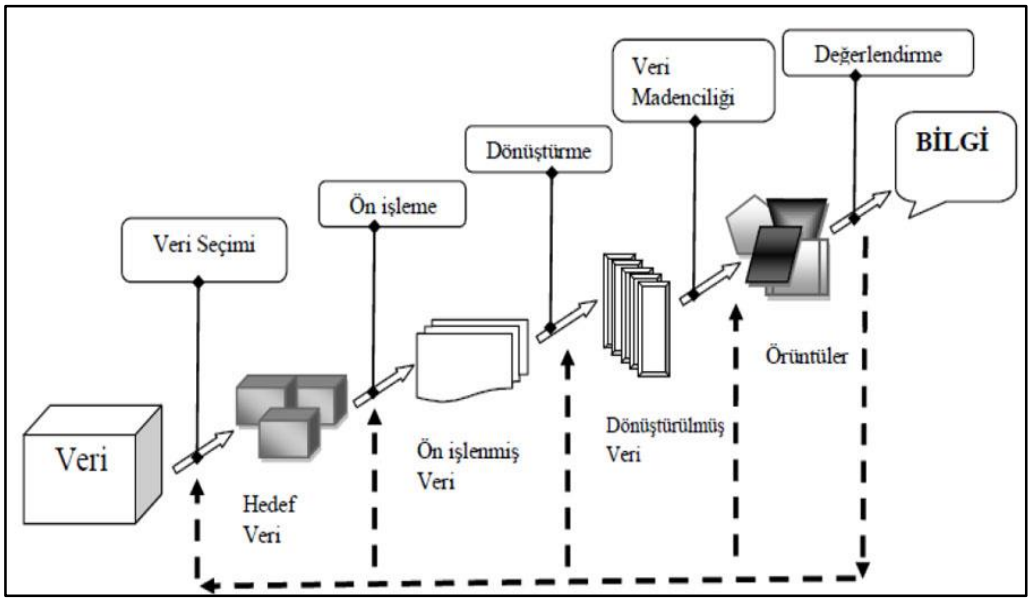
Araştırmacılar günümüzde, belgelerden yüksek performans ve otomatik bilgi çıkarımı yapmak için neredeyse tüm yapay zekâ yöntemleri ve makine öğrenimi algoritmalarını kullanmaya çalışmaktadır. Tüm tekniklerin altında, en temel teknikler söz dizim kuralları ve doğal dil işleme teknikleridir. İlk teknikle, kelime seviyesindeki bazı sözdizimsel kurallar ve kalıplar (normal ifadeler, belirteç tabanlı kurallar vb. gibi) metinden ayrıntılı bilgi elde etmek için kullanılır. Bilgi çıkarımında Doğal Dil İşleme'nin kullanılmasının temel fikri, cümle düzeyinde dil bilgisel yapıların incelenmesi ve daha sonra cümle içinde bazı yararlı bilgiler için dil bilgisel kuralların oluşturulmasıdır. Bilgi çıkarımı, bilgisayarın metin içindeki anahtar ifadeleri ve ilişkileri tanımlayarak yapılandırılmamış metni analiz etmesinin ilk adımıdır. Bilgi çıkarma işlemi, metinlere gömülü yapılandırılmamış bilgi çıkarma bilgisini, yapılandırılmış veriye dönüştürür. Bilgi çıkarma görevi, tokenizasyon, adlandırılmış varlıkların tanımlanması, cümle bölütlemesi ve konuşma bölümü atamasını içerir.

Bilgi çıkarma görevlerinin çoğunda ilk adım, bir metinde belirtilen uygun adları veya adlandırılmış varlıkları bulmaktır. Adlandırılmış varlık tanıma görevi, metin içinde adlandırılmış bir varlığın her bir sözünü bulmak ve türünü etiketlemektir. Bilgi çıkarımında ikinci adım ise ilişki çıkarımıdır. İlişkinin çıkarılması görevi, çoğu zaman eş, çocuk, iş, işçi bağı, üyelik gibi ikili ilişkilerde metin oluşumları arasındaki ilişkilerin anlamsallığını bulmak ve sınıflandırmaktır.

Son olarak, birçok metin verisi tekrarlanan kalıplaşmış durumları tanımlamaktadır. Şablonların doldurma görevi ise, dokümanlarda bu tür durumları bulmak ve şablon yuvalarını uygun malzeme ile doldurmaktır. Bu yuva dolgu maddeleri, doğrudan metinden çıkarılan metin bölümlerinden veya ek işlemlerle metin öğelerinden çıkarılan zamanlar, miktarlar veya ontoloji varlıkları gibi kavramlardan oluşabilir [41],[42].

1.4.3. Metin Madenciliği Adımları

Metin madenciliği de tıpkı veri madenciliği sürecinde olduğu gibi daha doğru analizlerin yapılabilmesi için bazı işlemlerin yapılması gerekmektedir. Bu işlemler genel olarak beş adımda yapılmaktadır.



Şekil 5. Metin Madenciliği Adımları [43]

1.4.3.1. Veri Seçimi

Veri seçimi adımı, dokümanlar, kullanıcı yorumları, web sayfaları, kitap, gazete, dergiler veya XML dosyaları gibi metin içeren verilerden metin madenciliği yapılacak olan verinin seçilme aşamasıdır.

1.4.3.2. Metin Ön işleme

Veri madenciliğinde olduğu gibi analiz yapılmadan önce analizin daha doğru ve güvenilir olabilmesi için veride bazı işlemlerin yapılması gerekmektedir. Metin madenciliği için yapılacak bu işlemler metin ön işleme aşamasında gerçekleştirilmektedir. Bu ön işleme işlemlerinden bazıları; *html* etiketleri gibi veride bulunan istenmeyen noktalama işaretlerini kaldırma, boşlukları kaldırma, kelimeleri küçük-büyük harfe çevirme, kelimeyi ek ve köklerini ayırma, yazım kurallarına uygun olup olmadığını tespit etme, stopwords ismi verilen çok kullanılan ve genel anlamı olan kelimelerin temizleme ve yapısal olmayan verinin analize uygun hale getirilmesi gibi işlemlerdir. Genel olarak metin ön işleme aşamasında belgeleri oluşturan kelimelerle ilgili işlemler yapılmaktadır [44].

- Stopwords: Durma kelimeleri, her dokümanda sıkça bulunan, tek başına herhangi bir anlam içermeyen ve dokümanlar için herhangi bir ayırt edici özelliği olmayan kelime listelerine denilmektedir. Stopwords kelimelerini temizlemek hem analiz süresini kısaltırken hem de daha doğru ve güvenilir sonuçlar alınmasına neden olmaktadır.
- Kelime Çantası (Bag of words): Stopwords kelimeleri çıkarıldıktan sonra gruplandırılacak tüm dokümanlarda bulunan tüm kelimelerin kullanım sıklıkları hesaplanarak, kelime çantası olarak düşünülebilecek bir havuzda toplanmasıdır. Kelime çantası oluşturulmasının amacı ise kelime ağırlıklandırma işleminin yapılmasını sağlamaktadır. Kelime ağırlıklandırma, bir kelimenin hangi alan ile ilgili olduğu ve o alanla ilgili olacak diğer dokümanlar da bulunma ihtimalinin yüksek olacağı anlamına gelmektedir [40].

1.4.3.3. Metin Dönüşümü

Metin dönüşümü aşamasında ise ön işlemden geçmiş ve hemen hemen analize uygun hale getirilmiş verinin kelimelerine ek-kök ayrımı, boyut azaltma ve verilerin kategorize edilme işlemleri yapılmaktadır.

1.4.3.4. Veri Madenciliği

İlk olarak yapısal olmayan veriler toplanır ve bu veriler üzerinde belirli ön işlemler yapılarak, veri yapısal hale dönüştürülür. Eldeki veri, yapısal olduktan sonra veri madenciliği algoritmalarından veriye uygun olan ya da yapılmak istenen yönteme göre ilgili algoritma seçilip veriye uygulanarak analiz edilir.

1.4.3.5. Yorum ve Değerlendirme

Ön işleme aşamasından geçerek yapısal hale getirilen veri üzerinde klasik veri madenciliği algoritmalarının uygulanması sonucu ortaya çıkan sonuçları anlaşılabilir bir dille raporlanma aşamasıdır.

1.4.5. Vektör Uzay Modeli

Vektör uzayı modeli (VUM), dokümanları oluşturan terimleri ağırlıklandırarak dokümanların her birini bir terim vektörü olarak ele alan modeldir. Dokümanlarda bulunan her terim vektör uzayında bir boyutu ifade etmektedir. Dokümanlar vektör uzayına taşınırken ilk olarak durma kelimeleri silinmelidir. İkinci adım ise, aynı köklü kelimelerin belirlenmesi ve tek kökte vektör uzayında yer almalarıdır. Örneğin, *seviyorum* ve *seviyorsun* kelimeleri normalde ayrı ayrı bir vektör olarak belirtilirken sadece *sevmek* kelimesinin vektör uzayına alınması analiz sonucunda daha doğru sonuçlar elde edilmesine yardımcı olacaktır. Vektörler arasında ki açı ne kadar az ise vektörlerin benzerliği de o kadar yüksek olacaktır [45].

1.4.6. Ağırlıklandırma

Sözcüklerin dokümanlar üzerinde etkisine ağırlıklandırma denir. Sözcükler, dokümanlarda bulunup bulunmama veya tekrar sayısına göre değerler almaktadırlar. Sözcüklerin aldıkları bu değerler sayısal olarak ifade edilir. Sözcük ağırlıklandırma yöntemlerinden bazıları terim frekansı, ters belge frekansı ve terim frekansı-ters belge frekansıdır.

- Terim frekansı (TF): Bir belgede bir terimin hangi sıklıkla geçtiğini gösteren en basit ağırlıklandırma yöntemidir. Terim frekansı $tf_{t,d}$, t terimi ve d dokümanı temsil etmektedir. $tf_{t,d}$ değeri ne kadar yüksekse t terimi d dokümanını o kadar iyi temsil etmektedir.

Terim frekansı bir sorgudaki alaka düzeyini değerlendirirken tüm terimler eşit derecede önemli kabul edilir. Aslında belirli terimlerin alaka düzeyini belirlemede çok az ya da hiç ayrımcı güçleri yoktur. Örneğin metin madenciliği ile ilgili olan bir doküman koleksiyonunda metin kelimesi neredeyse her dokümanda geçeceği için ayırt edici bir özelliği olmayacaktır. Bu sorunu gidermek için ise log normalizasyonu kullanılmaktadır. Log normalizasyonu $\log(1 + tf_{t,d})$ formülü ile hesaplanmaktadır.

- Ters belge sıklığı (IDF): Ters belge sıklığı bir kelimenin doküman için ne kadar belirleyici olduğunu belirten bir ağırlıklandırma yöntemidir. Bir sözcük sadece bir dokümanda geçiyor ise IDF değeri yüksek olacaktır. Fakat, bir kelime her dokümanda geçiyor ise veya hiçbir dokümanda geçmiyorsa IDF değeri 0 olacaktır. IDF $idf_{w_i} = \log(\frac{N}{df_t})$ formülü ile hesaplanmaktadır. Formülde N toplam doküman sayısı iken df_t ise sorgunun geçtiği doküman sayısıdır.
- Terim sıklığı-Ters belge sıklığı (TF-IDF): Metin madenciliğinde en sık kullanılan ağırlıklandırma yöntemlerinden biri TF-IDF tir. TF-IDF ağırlıklandırma yönteminde bir terimin hem kendi bulunduğu dokümandaki terim frekansı hem de diğer dokümanlarda bulunması önemlidir. TF-IDF ağırlığı $w_{t,d} = (1 + \log(tf_{t,d}) . \log(\frac{N}{df_t})$ formülü ile hesaplanmaktadır.

TF-IDF değeri bazı durumlarda artmaktadır. Bunlardan bazıları,

- Bir belge içinde tekrar sayısı ile artar. (TF)

- Bir doküman koleksiyonunda herhangi bir terimin nadir geçmesi ile artar(IDF) [40],[46], [47], [48].

1.4.7. Terim Doküman Matrisi

Birçok metinsel ifade içeren yapılandırılmamış verinin belirli işlemlerden geçerek yapısal hale dönüştürülmesinin ardından, verinin sayısal hale dönüştürülmesi için gereken ilk adım Terim-Doküman Matrisi (TDM) oluşturmaktır. TDM oluşturulmasının amacı, terimlerin dokümanlar için anlamını belirlemektir. Yapısal olmayan veriler kelimelerden oluşurken TDM sıfır ve birlerden oluşmaktadır. TDM de 0 değerleri ilgili terimin dokümanda geçmediğini gösterirken 1 değerleri ilgili terimin o dokümanda bulunduğunu göstermektedir. TDM m adet doküman ve n adet kelime olmak üzere $n \times m$ boyutlu bir matristir. TDM'de satırlar terimleri temsil ederken sütunlar dökümanları temsil etmektedir [49].

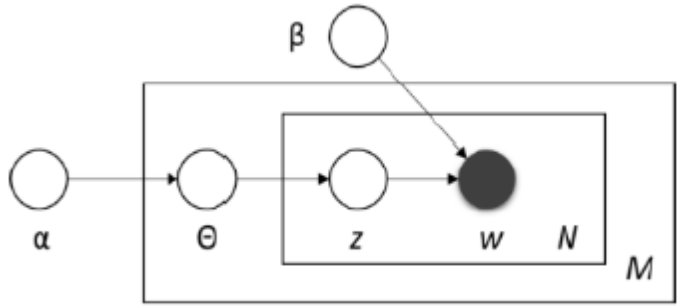
1.4.8. Gizli Dirichlet Tahsisi (GTD)

Konu modelleme yöntemleri bir metin doküman koleksiyonunda, hem doküman içerisinde geçen kelimeleri, hem de diğer dokümanlarda geçen kelimelerin birlikte kullanıldığı diğer kelimeleri inceleyerek veri koleksiyonundaki her dokümanı bir veya daha fazla konudan oluşacak şekilde sentezleyen bir model üretmektedir.

GTD (Latent Dirichlet Allocation) olasılıksal bir konu modelleme (topic modelling) yöntemidir. Başka bir ifade ile GDT, metin belgelerinin Bayesien olasılıklı bir modelidir. Konu modellenmesi, büyük ölçekli metinsel bilgi içeren koleksiyonları anlama ve özetleme gibi yöntemler sunmaktadır. Özet olarak, konu modellemesi koleksiyondaki bilgileri en iyi şekilde temsil eden bir belge topluluğundan bir grup konu bulmak için kullanılan bir yöntem olarak tanımlanabilir. GDT'nin amacı, gözlemlenen veriye göre metin dokümanlarının temel ve gizli konularını bulmaktır [4], [50].

GTD yönteminde metin dokümanları, konuların birleşimi olarak temsil edilmektedir. Yani, metin dokümanları konular üzerinde bir olasılık dağılımı olarak temsil edilirken, konular ise sözcüklerin üzerinde bir olasılık dağılımı olarak temsil edilmektedir. Bir dokümanın konusu doğal olarak o dokümanda geçen kelimelere bağlı olacaktır.

Dokümanlarda geçen kelimeler gizli olarak dokümanın ana konu ya da konularını temsil edecektir. Bu kelimeler ilişkili oldukları konuların sözcük grubunu oluşturacaktır. Örneğin, siyasetle ilgili bir konuda *meclis* ve *parti* gibi kelimeler olurken, ekonomi ile ilgili olan konularda *euro* ve *dolar* gibi kelimeler bulunacaktır. Metin koleksiyonlarından oluşan veri seti ne kadar büyük olursa hangi kelimelerin hangi konu grubuna dahil olduğunu belirlemek o kadar zorlaşacaktır. GTD ise bu işlem için geliştirilmiş bir metin araştırma aracıdır.



Şekil 6. Gizli Dirichlet Tahsisi [51]

Şekil 6'da GTD süreci grafiksel olarak gösterilmiştir. Burada rastgele değişkenler düğümler ile gösterilmektedir. Şekil 6'da α, Dirichlet parametresini, Θ belge seviyesi konu değişkenlerini, *z* her bir kelime için atanan konuları, *w* gözlemlenen kelimeleri ve *β* ise konuları temsil etmektedir [52], [53],[3].

2. YAPILAN ÇALIŞMALAR

Metin madenciliği günümüzde birçok çalışma yapılarak ülkemizde de popüler hale gelmiştir. Bu tez çalışmasındaki amaç, Türkiye'deki il bazında gerçekleşen haberlerde hangi olayların sıklıkla yaşandığını tespit etmek ve söz konusu iller hakkında bilgi sahibi olmaktır.

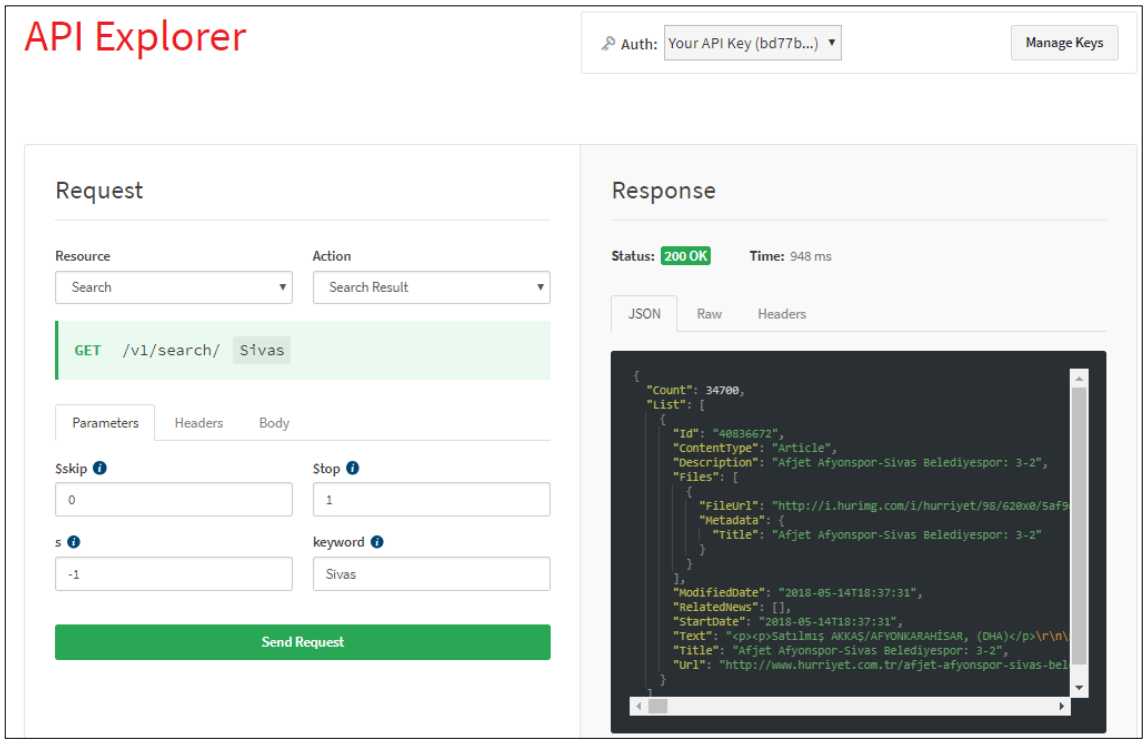
Bu çalışmayla birlikte il yöneticilerinin cinayet, taciz, hırsızlık, terör gibi konuların sıklıkla yaşandığı illerde alacakları önlemler artırılabilecek ve bu tarz olayları engellemede katkı sağlayacaktır. Buna ek olarak, Türkiye'deki hangi illerin güvenilir, hangi illerin sadece spor ya da benzeri haberlerle ön plana çıktığı tespit edilebilecektir.

Dünya genelinde birçok gazete veri setlerini açık kaynak olarak kullanıcının hizmetine sunarken Türkiye'de sadece Hürriyet gazetesi bu imkânı sağlamaktadır. Dünya çapında haberlerine ulaşabileceğiniz gazetelerden en bilinen ve geniş olanı New York Times gazetesidir. New York Times kullanıcılara 1980 yılından itibaren yayımladığı her haberi bulma ve kullanma imkânını açık kaynak olarak sağlamaktadır. Yine aynı şekilde, *https://newsapi.org* adresinde içlerinde CNN, BBC, Daily Mail, Reuters ve Four Four Two, Fortune gibi dünyaca ünlü gazete ve dergilerin bulunduğu 30,000'den fazla haber kaynağı ve blog haberlerini kullanıcılara sunmaktadır [54], [55].

Bu tez çalışmasında, Türkiye'de sadece Hürriyet gazetesinde bulunan ve kullanıcıya açık kaynak olarak sunduğu, *http://www.hurriyet.com.tr* ve Hürriyet gazetesinde yayımlanan haberlerden istenilen konu ya da yazar hakkındaki olanları indirilmesini mümkün kılan *https://developers.hurriyet.com.tr* adresinden il bazında haberler indirilmiştir.

API (Application Programming Interface) yani uygulama programlama arayüzü, yazılım programlarının birbirleriyle iletişim kurmak için takip edebilecekleri belirli kurallar kümesidir. Farklı yazılım programları arasında bir arayüz görevi görür ve kullanıcı arayüzünün insanlar ve bilgisayarlar arasındaki etkileşimi kolaylaştırmasına benzer şekilde etkileşimlerini kolaylaştırır [56].

Şekil 7'de görülen Hürriyet gazetesinin kullanıcıya sunduğu ve halen deneme (Beta) sürümünde olan https://developers.hurriyet.com.tr arayüzünden veri elde etmek için API gerekmektedir. API anahtarı (API-key) elde etmek için *https://developers.hurriyet.com.tr* sitesine kayıt olmak yeterli olmaktadır. Bu sistem arayüzü deneme sürümünde olduğu için istenilen haber sayısı elde edememe, boş haber metni gibi bazı problemler yaşanabilmektedir. Yine de bu arayüz, kullanıcıya saniyede maksimum 5 ve saatte maksimum 500 istek yapma imkânı sunmaktadır.



Şekil 7. https://developers.hurriyet.com.tr arayüzü [57]

2.1. Veri Elde Etme

Yapılan çalışmaların bu aşaması, hangi verilerin kullanılacağını ve kullanılacak verinin elde etme işlemlerinin yapıldığı aşamasıdır. Bu tez çalışmasında, R programlama dili ve Hürriyet API'si kullanılarak illere ait *www.hurriyet.com* adresinden illere ait belli sayıdaki son haberler indirilmiştir. İndirilen haberler üzerinde herhangi bir işlem yapılmamıştır.

R ile gerçekleştirilen sistemde, Hürriyet API'si ile veri elde etmek için birçok paket kullanılmıştır. Bu R kütüphanelerinde bazıları Jsonlite, RCurl, XML ve TM paketleridir.

Veri elde ederken *https://developers.hurriyet.com.tr* adresinden kaynaklanan bazı sorunlar meydana gelmiştir. Bu sorunlardan bazıları istenilen sayıda haber indirilememesi ve indirilen haberlerden bazılarının boş olmasıdır. Her il için toplam haber sayısı, indirilen haber sayısı ve boş haber sayısı Tablo 1 de verilmiştir.

Tablo 2. Haber Sayıları

İller	**İndirilen**	**Toplam**	**Dolu**	**İller**	**İndirilen**	**Toplam**	**Dolu**
ADANA	5,000	84775	4932	KONYA	3,000	60618	2970
ADIYAMAN	1,000	18034	997	KÜTAHYA	1,000	18045	995
AFYON	1,000	14430	998	MALATYA	1,500	30272	1488
AĞRI	1,000	26918	998	MANİSA	3,000	46547	2983
AMASYA	1,000	9204	997	KAHRAMANMARAŞ	1,500	15703	1495
ANKARA	10,000	370012	9952	MARDİN	1,000	25937	993
ANTALYA	7,000	136441	6943	MUĞLA	4,000	43130	3970
ARTVİN	1,000	8508	998	MUŞ	2,000	16722	1987
AYDIN	5,000	110647	49778	NEVŞEHİR	1,000	11269	996
BALIKESİR	3,000	32648	2976	NİĞDE	1,000	12831	996
BİLECİK	0,500	10866	496	ORDU	2,000	43196	1992
BİNGÖL	2,000	17801	1982	RİZE	500	22105	497
BİTLİS	1,000	14217	993	SAKARYA	2,500	30949	2483
BOLU	2,000	23317	1982	SAMSUN	5,000	48145	4981
BURDUR	1,000	10825	1000	SİİRT	1,000	16495	984
BURSA	5,000	87958	4947	SİNOP	1,000	8299	989
ÇANAKKALE	4,000	42969	3952	SİVAS	4,500	33986	4472
ÇANKIRI	1,000	8197	998	TEKİRDAĞ	1,000	20796	999
ÇORUM	1,500	13425	1489	TOKAT	2,000	26543	1994
DENİZLİ	4,000	55422	3972	TRABZON	3,000	58452	2957
DİYARBAKIR	5,000	65344	3966	TUNCELİ	1,000	13697	981
EDİRNE	2,000	28726	1990	ŞANLIURFA	3,000	33683	2988
ELAZIĞ	1,500	19588	1481	UŞAK	1,000	16651	995
ERZİNCAN	1,500	16310	1491	VAN	3,000	57350	2910
ERZURUM	3,000	39666	2980	YOZGAT	1,000	13464	997
ESKİŞEHİR	3,000	44607	2986	ZONGULDAK	2,000	21600	1985
GAZİANTEP	4,000	59043	3971	AKSARAY	0,500	12450	495
GİRESUN	1,000	13099	990	BAYBURT	1,000	6026	986
GÜMÜŞHANE	1,000	8528	989	KARAMAN	1,000	17912	997
HAKKARİ	2,000	26392	1985	KIRIKKALE	1,000	11976	998
HATAY	3,000	28236	2956	BATMAN	1,000	17824	991
ISPARTA	2,000	16672	1981	ŞIRNAK	2,000	20968	1963
İÇEL	4,000	58560	3965	BARTIN	1,000	7770	991
İSTANBUL	9,000	595849	8907	ARDAHAN	1,500	11594	1488
İZMİR	9,000	225459	8951	IĞDIR	1,000	9237	994
KARS	1,000	17283	986	YALOVA	1,000	16848	993
KASTAMONU	1,000	13734	994	KARABÜK	1,000	13421	992
KAYSERİ	4,000	61771	3978	KİLİS	0,500	10509	497
KIRKLARELİ	0,700	12345	697	OSMANİYE	1,000	12089	997
KIRŞEHİR	1,000	7304	997	DÜZCE	1,000	14619	992
KOCAELİ	3,000	38171	2992				

2.2. Veri Ön İşleme

Veri ön işleme aşaması veri madenciliğinin olduğu gibi metin madenciliğinin de en önemli aşamasıdır. Bu aşamada yapılan her işlem analiz sonucunu doğrudan etkilemektedir. Veri ön işleme aşaması ne kadar doğru ve iyi yapılırsa analiz sonucu da o kadar doğru ve güvenilir olmaktadır.

Veri elde etme aşamasının sonucunu oluşturulan veri seti üzerinde bazı ön işlemler yapılmıştır. Bu ön işleme işlemleri *html* etiketlerini, boşlukları, noktalama işaretlerini ve stopwords adı verilen durma sözcüklerini silmek, büyük harfleri küçük harfe dönüştürme şeklinde sıralanabilir. Veri setinde çok fazla bulunan Türkçe karakterlere eğer karakter dönüşümü yapılmazsa işletim sistemindeki karakter kodlama problemi nedeniyle analiz sonucu negatif yönde etkilenebilir.

Veri ön işleme işlemleri için R programlama dilinde birçok paket kullanılmıştır. TM (Text Mining) paketi *html* taglerini, noktalama işaretlerini, boşlukları ve sayıları silmek, büyük harfleri küçük harfe dönüştürmek gibi işlemleri yaparken *gsubfn* paketi karakter dönüşümü yapmaktadır. Yapılan karakter dönüşümüne "*ç*" harfini "*c*" harfine, "*ü*" harfini "*u*" harfine dönüştürmek örnek olarak verilebilir. Diğer bir ön işleme işlemi ise stopwordsleri doküman içeriğinden ayıklamaktır. Bu tez çalışmasında kullanılan veri seti Türkçe metinlerden oluştuğu için Türkçe stopwords listesi gerekmektedir. R programında Türkçe stopwords kütüphanesi istenilen seviyede olmadığı için Türkçe kelimelerden oluşan bir stopwords listesi bu çalışma kapsamında oluşturulmuştur. Oluşturulan bu listedeki kelimeler her il için indirilen veri setinden ayrı ayrı çıkarılmıştır.

2.3. Sayısallaştırma

Yapısal olmayan metinsel veri, ön işleme aşamasından geçirilerek yapısallaştırılmaya hazır hale getirilmiştir. Veri üzerinde analiz yapılabilmesi için metin verilerinin sayısal hale dönüştürülmesi gerekmektedir. Bu dönüşüm iki evrede gerçekleştirilmiştir. İlk evrede, veri setini oluşturan her bir kelimeye terim frekansı ağırlıklandırılması yapılmıştır. Terim frekansı ağırlıklandırma işleminin yapılmasının nedeni bu çalışmada kullanılan ve konu modelleme algoritması olan GDT'nin terim frekansı ile işlem yapmasıdır. İkinci evre ise doküman terim matrisinin oluşturulmasıdır. Doküman terim matrisi, metin verisini 0 ve

1’lerden oluşan bir matriste gösterilmesidir. Hem terim frekansı ağırlıklandırma yöntemi hem de döküman terim matrisinin oluşturulması için R dilinde

DocumentTermMatrix(t, control = list(weighting=weightTf)) komutu kullanılmıştır. Bu matrisin elde edilmesiyle birlikte veri analize uygun hale getirilmiştir.

2.4. Analiz

Bu aşamada metin madenciliğinin bir yaklaşımı olan konu modelleme (Topic Modeling) ve kelime bulutu (Word Cloud) işlemleri yapılmaktadır. Sayısallaştırma aşamasında elde edilen doküman terim matrisi kullanılarak GDT işlemi yapılmıştır.

GDT’nin kullanılması için R programında *topicmodel* paketi kurulması gerekmektedir. Konu modelleme işleminin gerçekleştirilmesi için R, versiyon 3.5.0 programlama dili kullanılmıştır. Yapısal hale gelen veriye GDT yapılarak her il için 10 adet konu (topic) belirlenmiştir. Bu işlem,

```
lda <- LDA (dtm, 10)
```

komutu ile gerçekleştirilmiştir. Bu kod parçasındaki “*dtm*” komutu doküman terim matrisini temsil etmektedir.

3. BULGULAR

Bu aşamada her il için indirilen haberlere konu modelleme yönteminin bir algoritması olan Gizli Dirichlet Tahsisi (GDT) uygulanarak her il için 10 adet konu belirlenmiştir. Her il için özel durma kelime listesi bulunmaktadır. Bunun nedeni, örneğin Trabzon ili için çıkartılan konu başlıklarında *Trabzon* kelimesi doğal olarak yoğun bir frekansa sahiptir. Ancak ilin kendi isminin anlamsal olarak (topic bazlı) etkisi yoktur. Bu nedenle, bu tip kelimeler il bazında haber içeriğinden (corpus) çıkarılmıştır. GDT ile elde edilen analiz sonuçları Tablo 2'de verilmiştir.

GDT sonucu, 81 il için elde edilen 10 konuya bakıldığında 69 tanesinde "üniversite" kelimesi 10 konudan biri olmuştur ve Türkiye'deki illerde en çok konuşulan kelimelerden biridir. Terim frekansı yüksek olan bir diğer kelime ise "şehit" kelimesidir. Bu kelime ise 81 ilin 44 tanesinde 10 konudan biri olmuştur. Özellikle son zamanlarda ülkemizin terörle mücadele çalışmaları yoğun bir şekilde sürdüğünden dolayı böyle bir konu başlığının çıkması manidardır. En yüksek üçüncü frekansa sahip kelime olarak "Gözaltına" kelimesi sonuçlar arasında göze çarpmaktadır. Yine bu kelimenin yüksek frekansa sahip olmasında özellikle 15 Temmuz olayları sonrasında birçok terör örgütüne düzenlenen operasyonlar etkili olmuştur. Bu operasyonlarla ilgili haber yoğunluğunun fazla olduğu ve hatta il bazında nerelerde daha fazla gözaltı yapıldı, sonuçlara bakıldığında görülebilmektedir. Ayrıca, *gözaltı* kelimesi ile suç oranın orantılı olarak hangi illerde yüksek olduğu sonucuna da varılabilmektedir.

Yüksek frekansa sahip diğer bir kelime ise 81 ilin 38 tanesinde bulunan "spor" dur. Buna ek olarak "futbol, basketbol, maç, lig, puan" gibi sporla alakalı kelimelerin yanı sıra takım isimleri ve sponsorların analiz sonucu elde edilmesi, Türkiye'de iller arasında fark olmaksızın sporla ilgili konuların konuşulduğu anlaşılmaktadır.

"Konut" kelimesi genel olarak az nüfuslu illerden oluşan 14 ilde bulunmaktadır. Bu sonuç yardımıyla Artvin, Hakkâri, Kırşehir, Aksaray gibi illerde normalden daha fazla sayıda konut inşası yapıldığı hakkında fikir elde etmek mümkündür.

İl bazında analiz sonuçlarına bakıldığı zaman diğer bir dikkat çekici kelime ise "Temmuz" dur. Temmuz kelimesi 15 Temmuz hain darbe girişimine atıfta bulunmaktadır. Her il için indirilen haberler üzerinde GDT yapılmadan önce bazı ön işleme işlemleri yapılmıştır. Bu ön işleme işlemlerinden bir tanesi de veri setlerinden sayıların silinmesidir. Bu işlem başarıyla yapıldığı için analiz sonucunda 15 Temmuz olarak değil sadece "temmuz" kelime bulunmaktadır.

"Terör" kelimesi ise 19 farklı ilde bulunmaktadır. Terör kelimesinin doğal olarak doğu illerinde bulunması beklenmektedir. Fakat analiz sonucuna bakıldığında Afyon ve Karaman illeri için bu kelimenin bulunduğu görülmektedir. Bu iki ilin coğrafi konumu da düşünüldüğünde terör kelimesinin daha çok FETÖ terör örgütü ile ilgili olduğu düşünülmektedir.

"Turizm" kelimesinin bulunduğu illerden bazıları Antalya ve Nevşehir dir.

GDT sonucu her il için bulunan kelimelerin güvenilir olduğunu gösteren bazı sonuçlar görülmektedir. Artvin için "hes", Manisa için "macun", Kırşehir için "Neşet", Rize ve Trabzon için "çay" kelimesi ve Türkiye'nin ikinci nükleer santralinin kurulacağı Sinop için elde edilen sonuçlarda "nükleer" kelimesinin bulunması buna örnek olarak verilebilir. Konya ili için bulunan kelimelere bakıldığında "Mevlâna" ve "Selçuklu" kelimelerinin olması bir diğer örnektir. "Proje" kelimesinin bulunduğu 4 farklı il olmasına rağmen son zamanlarda 1923 proje üreten Mersin ilinde bu kelimenin çıkması bu tez çalışmasının da yapılan çalışmanın güvenilirlik düzeyini göstermektedir. Tablo 2'de Gizli Dirichlet Tahsisi ile elde edilen sonuçlar verilmiştir.

Tablo 3. Gizli Dirichlet Tahsisi ile Elde Edilen Sonuçlar

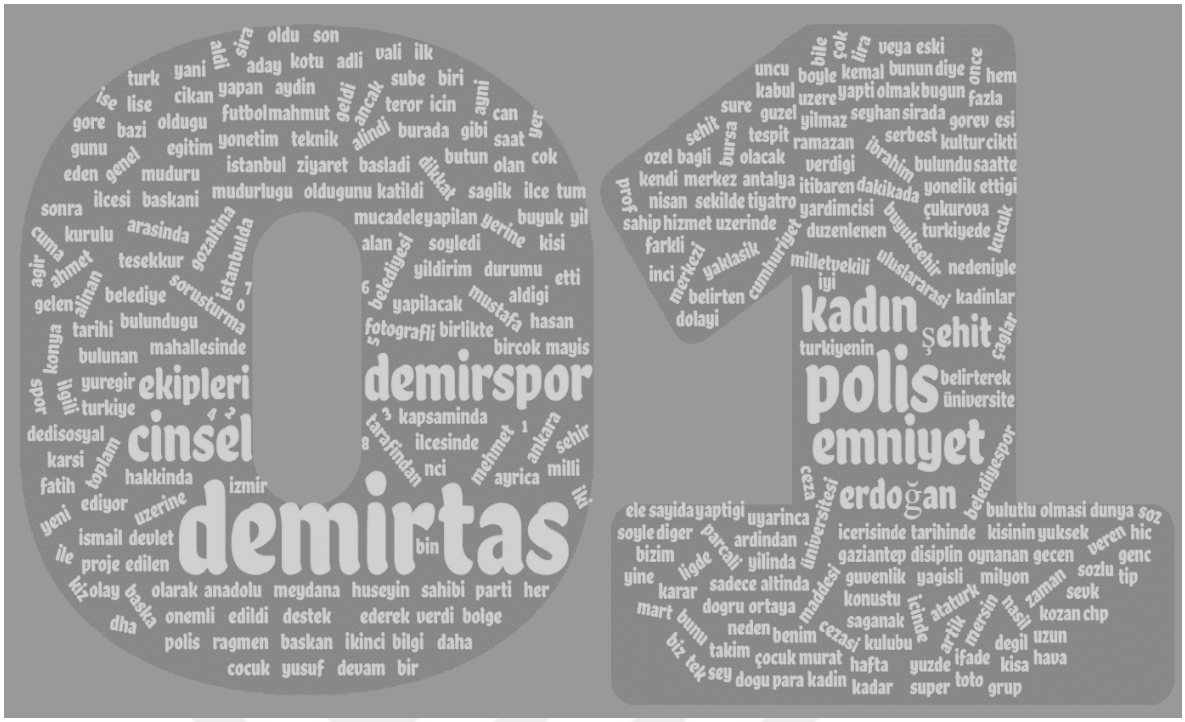
İller	Konu1	Konu2	Konu3	Konu4	Konu5	Konu6	Konu7	Konu8	Konu9	Konu10
Adana	Tiyatro	Cuma	Çukurova	Türkiye	Polis	Spor	Sınav	Kadın	Üniversite	Demirtaş
Adıyaman	Gölbaşı	Üniversite	Sağlık	Trafik	Gözaltına	Şehit	Çelikhan	Futbol	Hastane	Spor
Afyon	Sonuçlar	Üniversite	Uyuşturucu	Basketbol	Terör	Belediye spor	Yıldırım	Şeker	Adli	Kurban
Ağrı	Âdem	Tedavi	Üniversite	Mayıs	Burun	Sosyal	Eğitim	Doğum	Terör	Şehit
Amasya	Üniversite	Kurban	İstanbul	Olay	Elma	Gözaltına	Şehit	Mülakat	Milli	Mayıs
Ankara	Türkiye	Cumhurbaşkanı	Mahkeme	Terör	Tiyatro	Spor	Eğitim	Gözaltına	İzmir	Cuma
Antalya	Gözaltına	Olay	Sağanak	Sağlık	Üniversite	Belediye	Turizm	Ceza	Sınav	Spor
Artvin	Üniversite	Erdoğan	Havalimanı	Doğu	Rol	Yusufeli	Konut	HES	Cankurtaran	Boğa
Aydın	Üniversite	Sosyal	Doğan	İzmir	İnci	Hayırlı	Kuşadası	Sanık	Hapis	Parti
Balıkesir	Banvit	Milletvekili	Parti	Konut	Bandırma	Ayvalık	Gözaltına	Spor	Şehit	Burhaniye
Bilecik	Lira	Bozüyük	Eğitim	Yaralı	Üniversite	TÜSİAD	Eğitim	Doğu	Kurban	Yönetim
Bingöl	Tiyatro	Terör	CHP	Şehit	Üniversite	Lig	Eğitim	Doğu	Parti	Yaralandı
Bitlis	Puan	İstanbul	Van	Eğitim	Üniversite	Doğu	Kurban	Terör	Ustaoğlu	Şehit
Bolu	Eğitim	Üniversite	Şehit	Olay	Kurban	Atatürk	Mülakat	Gözaltına	Hapis	Tabiat
Burdur	Mesut	Yılmaz	Ekipleri	Şeker	Türk	Antalya	Üniversite	Mağara	Spor	Üniversite
Bursa	Nilüfer	Türkiye	Gözaltına	Belediye	Üniversite	Olay	Teşekkür	Tiyatro	Bursaspor	Konut
Çanakkale	Gökçeada	Belediye spor	Türk	Yıldırım	Beşiktaş	Eğitim	Bulutlu	Erdoğan	Şehit	CHP
Çankırı	Temmuz	Bayram	Eğitim	Kurban	Milletvekili	Spor	Şehit	Cumhuriyet	Spor	Kuvvetli
Çorum	Anadolu	Polis	Şehit	Trafik	Üniversite	Gözaltına	Olay	Milletvekili	Türkiye	Erdoğan
Denizli	Kar	Fenerbahçe	Denizlispor	Trafik	Olay	Gözaltına	Üniversite	Spor	Eğitim	Hakimliği
Diyarbakır	Türkiye	Mehmet	Parti	Adet	Eğitim	İstanbul	Kürt	Terör	Ceza	Şehit
Edirne	Üniversite	Elektrik	Eğitim	Kurban	İstanbul	Yunan	Keşan	Milletvekili	Jandarma	Parti
Elazığ	Şehit	Elazığ spor	Terör	Türkiye	Malatya	Üniversite	Atatürk	Kurban	Gözaltına	Olay
Erzincan	Eğitim	Jandarma	Üniversite	Yıldırım	Belediye spor	Zeynep	Gözaltına	Hapis	Doğan	Şehit
Erzurum	Terör	Üniversite	Proje	Teknik	Şehit	Ovit	Eğitim	Şeker	Kurban	Sınav
Eskişehir	Gözaltına	Milli	Termik	Basketbol	Lira	Odun pazarı	Şehit	Parti	Yerli	Üniversite
Gaziantep	Ramazan	Hastane	Spor	İhracat	Terör	Belediye	Üniversite	Gazişehir	Özel	Gözaltına

Tablo 3'ün devamı

Giresun	Giresun spor	Üniversite	Atatürk	Trafik	Fındık	Turizm	Maç	Şehit	Ertuğrul	Trabzon
Gümüşhane	Kurban	İstanbul	Türkiye	Doğan	Üniversite	Gümüşhane spor	Polis	Temmuz	Mülakat	İşçi
Hakkâri	İnce	Kar	Üniversite	Şehit	Emniyet	Eğitim	Terör	Konut	Kadın	Parti
Hatay	Türk	Spor	Eğitim	Üniversite	Gözaltına	Erdoğan	Zeynep	Yardım	Şehit	Zeytin
Isparta	Gözaltına	Üniversite	İhracat	Şehit	Olay	Lisesi	Gül	Burdur	Mahkeme	Taşeron
İçel	Kadın	Belediye spor	Havalimanı	Eğitim	Üniversite	Polis	Spor	Proje	Gözaltına	Türkiye
İstanbul	Kripto	Dolar	İsrail	Cumhurbaşkanı	Eğitim	Gözaltına	Üniversite	Ramazan	Yönetim	Sanık
İzmir	Lisesi	Belediye	Konut	Yönetim	Teknik	Seçim	CHP	Üniversite	Çeşme	Gözaltına
Kars	Eğitim	Yaralı	Belediye	Türkiye	Bakan	Harun	Kar	Yusuf	Sarıkamış	Tarım
Kastamonu	Şeker	Kastamonu spor	Doğan	Çataloğlu	Kurban	Müze	Devlet	Üniversite	Kiralık	Puan
Kayseri	Hastane	Üniversite	Parti	Kocasinan	Şehit	Gözaltına	Kayseri spor	Olay	Tutuklu	Ticaret
Kırklareli	Şehit	Üniversite	Vahit	Edirne	Şeker	Lüleburgaz	Atatürk	Çiftçi	Kar	Trafik
Kırşehir	Şehit	Temmuz	Neşet	Üniversite	Hapis	Emniyet	Konut	Şeker	Proje	Tarım
Kocaeli	Darıca	Türkiye	Belediye spor	Üniversite	Eğitim	Gözaltına	Kitap	Ceza	Sosyal	Gebze
Konya	Ali	Adli	Hapis	Polis	Üniversite	Mevlâna	Gözaltına	Selçuklu	Spor	Şehit
Kütahya	Gediz	Simav	Üniversite	Polis	Kurban	İstanbul	Eğitim	Hastane	Kar	Şehit
Malatya	Üniversite	Tüfenkçi	Gözaltına	Ordu	Spor	Kuru	Beşiktaş	Belediye	Battalgazi	Olay
Manisa	Parti	Trafik	Zeytin	Eğitim	Polis	İzmir	Üzüm	Mesir	Lig	Teknik
Kahramanmaraş	Taşeron	Olay	Ömer	Üniversite	Yurdakul	Eğitim	Gözaltına	Parti	İstifa	Hastane
Mardin	Türkiye	Eğitim	Terör	Parti	Kurban	Şehit	HDP	Kız	İstanbul	Toz
Muğla	Parti	Zeytin	Deprem	Mahkeme	Dalaman	Gözaltına	Konut	Bodrum	Olay	Marmaris
Muş	Terör	Eğitim	Kar	Trafik	Kurban	Temmuz	Karşılıksız	Varto	Erdoğan	Üniversite
Nevşehir	Kurban	Üniversite	Şehit	Spor	Parti	Turizm	Kapadokya	Patates	Lira	Vali
Niğde	Koro	Üniversite	Ömer	Belediye spor	Milletvekili	Yıldırım	Doğu	Ali	Patates	Şehit
Ordu	Fındık	Üniversite	Eğitim	Trabzon	İstanbul	Erdoğan	Lisesi	Parti	Hapis	Şehit
Rize	Puan	Üniversite	Beyaz	İnce	Kitap	Çay	Eğitim	ÇAYKUR	Şeker	Lisesi
Sakarya	Şehit	Gözaltına	Fındık	Üniversite	Evet	Deprem	Basketbol	Elektrik	Mustafa	Spor

Tablo 3'ün devamı

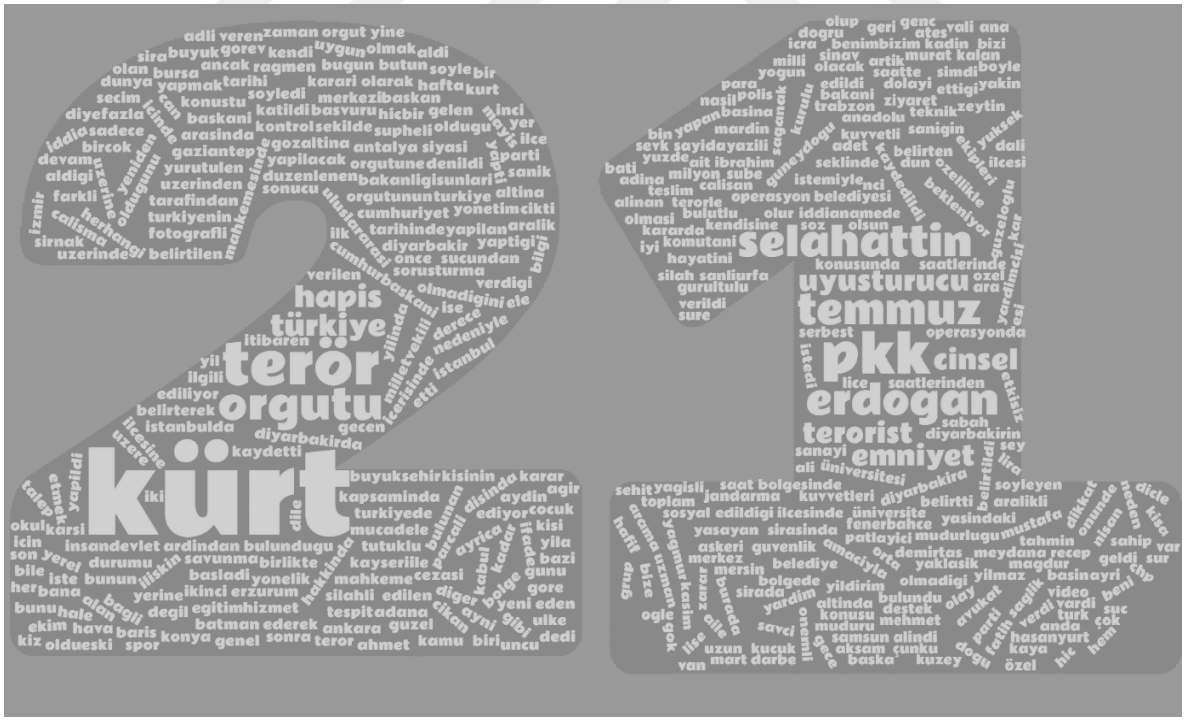
Samsun	Cumhurbaşkanı	Tekkeköy	Olay	Gözaltına	Turizm	Lira	Üniversite	Şehit	Öğretmen	Atakum
Siirt	Temmuz	Vali	Tahmin	Üniversite	Kurban	Şehit	Terör	Akıncı	Gözaltına	Komutan
Sinop	Olay	Nükleer	Eğitim	Üniversite	İstanbul	İşçi	Mağara	Samsun	Temmuz	Türkiye
Sivas	Yıldırım	Belediye spor	Kar	Cumhuriyet	Eğitim	Teknik	Beşiktaş	Gül	FETÖ	Olay
Tekirdağ	Kırklareli	Süleyman paşa	Şeker	Üniversite	Şehit	Termik	Malkara	Gözaltına	Trafik	Belediye
Tokat	Şeker	Volkan	Polis	Trafik	Eğitim	Üniversite	Kudüs	Sanık	Parti	Kadın
Trabzon	Çay	Beşiktaş	Polis	Trabzonspor	Parti	Uçak	Çalımbay	Ceza	Konut	Hamsi
Tunceli	Cenaze	Gözaltına	Terör	Şehit	Vali	Parti	Kurban	Lira	Mülakat	Üniversite
Şanlıurfa	Yardım	Sınav	Ceylanpınar	Belediye	Olay	Belediye spor	Fakıbaba	Şehit	Eğitim	Gözaltına
Uşak	İzmir	Parti	Üniversite	Olay	Güneş	Muratbey	Gözaltına	Hastane	Kurban	Konut
Van	Şehit	Belediye	Üniversite	Eğitim	Doğu	Elektrik	Belediye spor	Film	Fenerbahçe	Futbol
Yozgat	Spor	Şehit	Eğitim	Üniversite	Kurban	Lisesi	Şeker	Erdoğan	Ekipleri	Sıddık
Zonguldak	Termik	Belediye spor	Lira	Kiralık	Gözaltına	Havalimanı	Kadın	Maden	Kömür spor	Şehit
Aksaray	Oyun	Gözaltına	Belediye S.	Sanayi	Üniversite	İstanbul	Doğan	Konut	Polis	Hastane
Bayburt	Kurban	Lira	Lig	Yol	Grup	Üniversite	Konut	Emniyet	Ağbal	Erdoğan
Karaman	Anadolu	Milli	Üniversite	Terör	Zeynep	Gözaltına	Demir	Arslan	Şehit	Olay
Kırıkkale	Profesör	Yaralandı	Temmuz	Proje	Üniversite	Emniyet	Şehit	Spor	Cumhuriyet	Tiyatro
Batman	Doğu	Gözaltına	Üniversite	Terör	Şırnak	Şehit	Konut	Ana	Akşener	Yıldırım
Şırnak	Şehit	Fakıbaba	Eğitim	Terör	Turan	Üniversite	Gözaltına	Diyanet	Güvenlik	Bakan
Bartın	Termik	Üniversite	Doğum	Temmuz	Artış	Doğu	Ayhan	Yol	İstanbul	Eğitim
Ardahan	Konut	Eğitim	Üniversite	Tarım	Milli	Kar	rgan	Profesör	Karaca	Güvenlik
Iğdır	Vali	Üniversite	Eğitim	Temmuz	Cumhuriyet	Cumali	Kalp	Gözaltına	Ağrı	Doğu
Yalova	Milletvekili	Salman	Türkiye	Trafik	Konut	Önder	Üniversite	İnce	Eğitim	CHP
Karabük	Üniversite	Kardemir	Safranbolu	Futbol	Parti	Bülent	Gözaltına	Şehit	Gol	Puan
Kilis	Terör	Türk	Üniversite	Afrin	Şehit	Yardım	Gök	Zeytin	Zeynep	Trafik
Osmaniye	Ebu	Trafik	Coşkun	İbrahim	Karşılıksız	Şehit	Jandarma	Üniversite	Atatürk	Terör
Düzce	Atatürk	Profesör	Gözaltına	Milli	Olay	Basketbol	Üniversite	Deprem	Eğitim	Turizm



Şekil 8. Adana için Kelime Bulutu

Şekil 9. Ankara için Kelime Bulutu

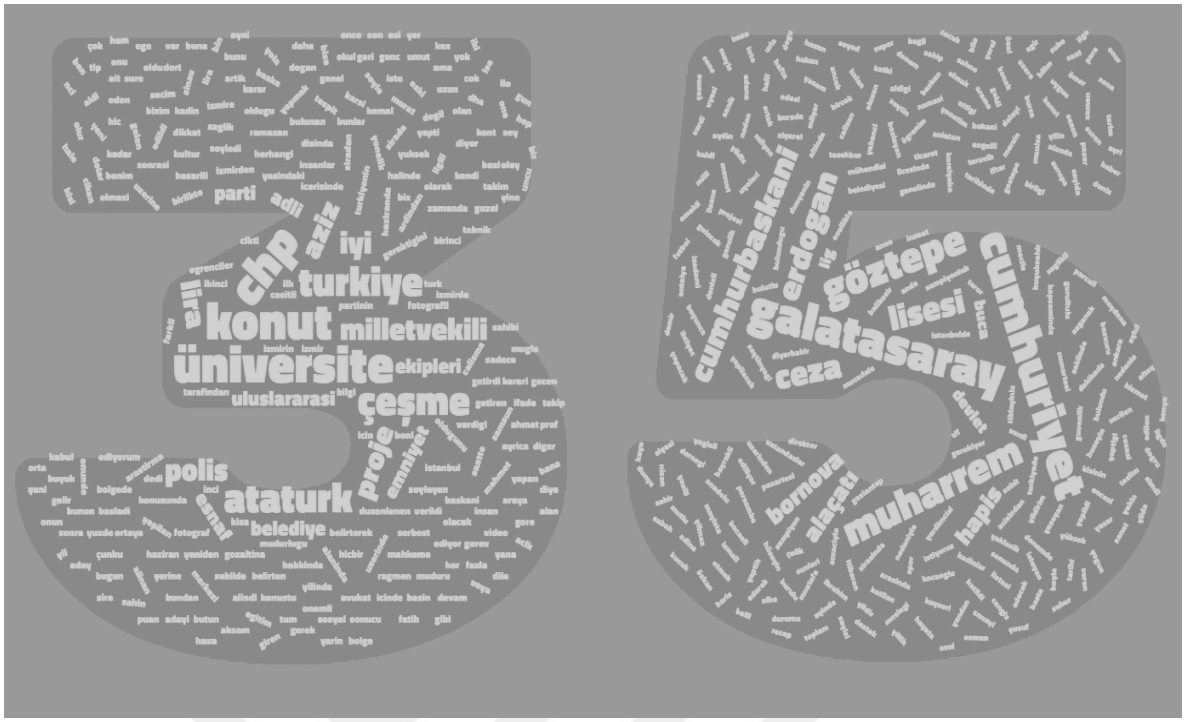
Şekil 10. Antalya için Kelime Bulutu



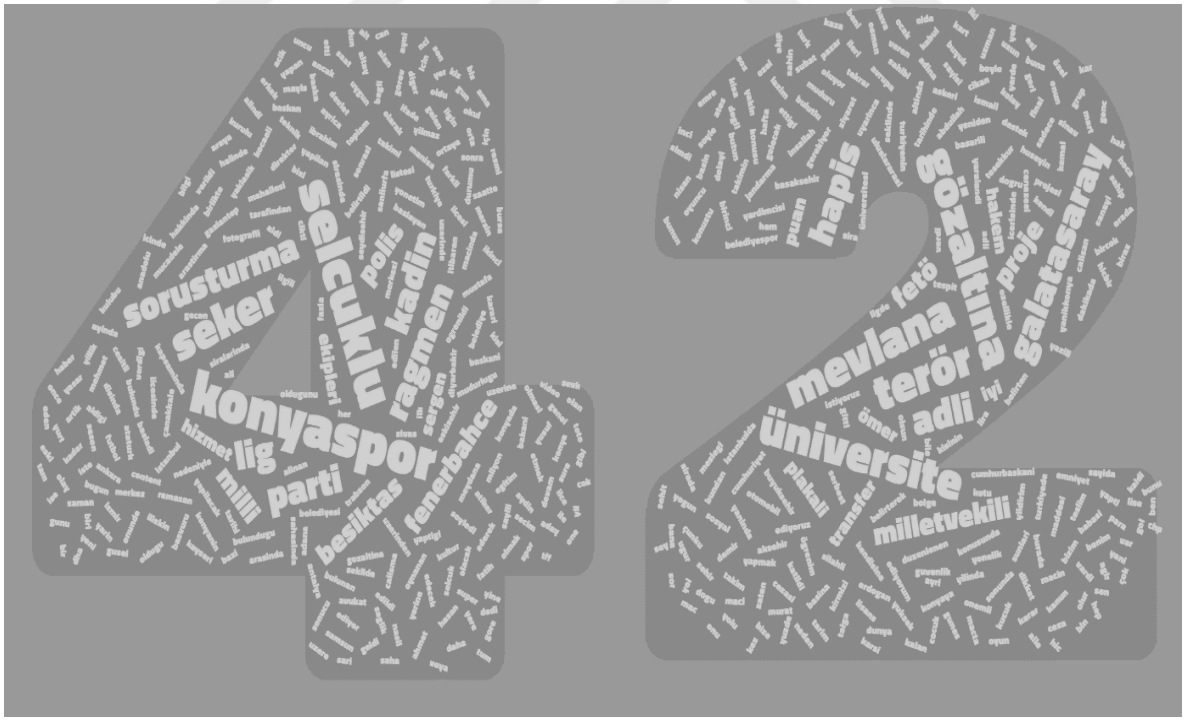
Şekil 11. Diyarbakır için Kelime Bulutu

Şekil 12. Mersin için Kelime Bulutu

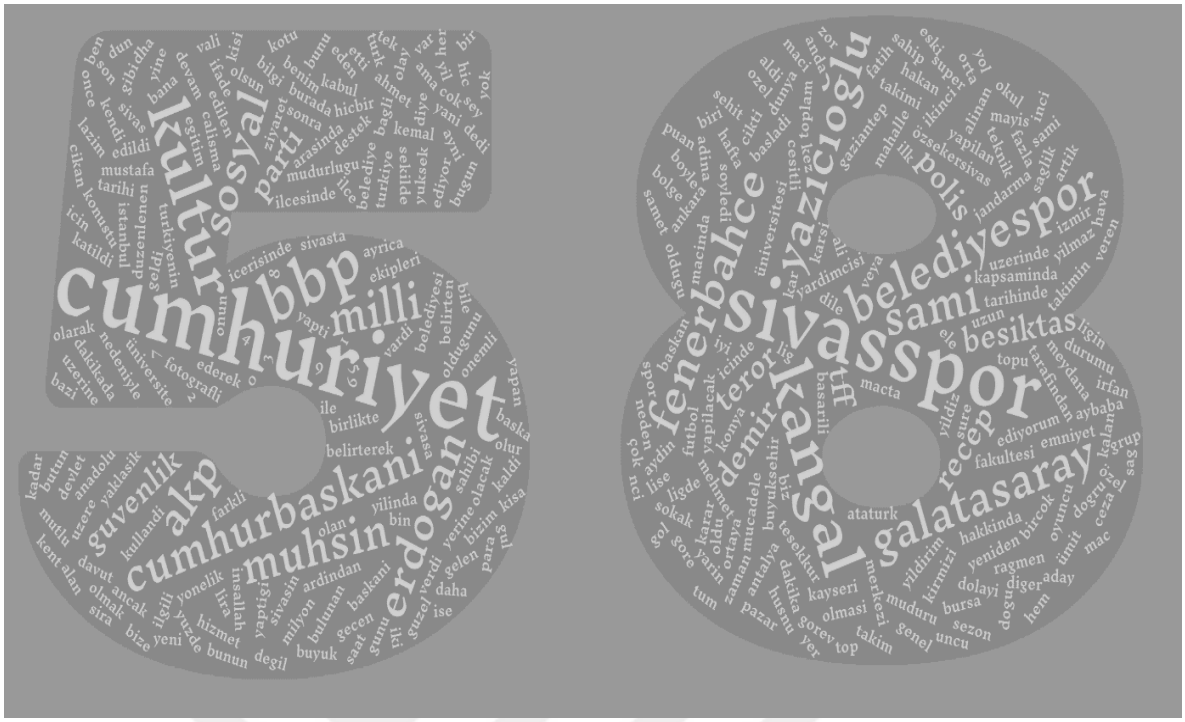
Şekil 13. İstanbul için Kelime Bulutu



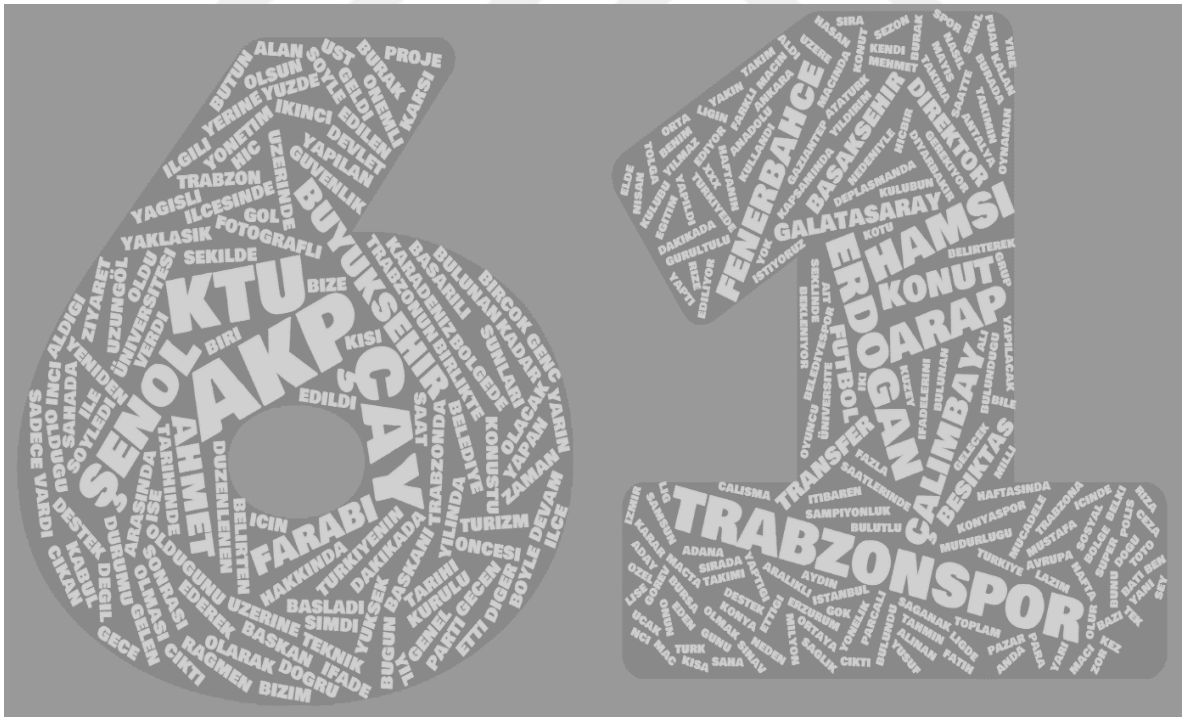
Şekil 14. İzmir için Kelime Bulutu



Şekil 15. Konya için Kelime Bulutu



Şekil 16. Sivas için Kelime Bulutu



Şekil 17. Trabzon için Kelime Bulutu

4. SONUÇLAR

Türkiye'de metin madenciliği geliştiricileri için henüz deneme (beta) aşamasında olan tek veri kaynağı Hürriyet gazetesinin haber arşividir. Bu arşive erişebilmek için bir API Key yardımıyla elde edilen haberler, R programlama dili kullanılarak analiz edilmiştir. R programının metin madenciliği ile ilgili çok sayıda kütüphanesi mevcuttur ve açık kaynak olduğundan dolayı sürekli kullanıcılar tarafından geliştirilmeye devam etmektedir. Bu kütüphaneler yardımıyla her şehir için indirilen haberler belirli ön işleme aşamalarından geçirilmiştir. Ön işlem yapılan veri setine Gizli Dirichlet Tahsisi yaklaşımı yardımıyla olasılıksal dağılımlar hesaplanarak yüksek frekansa sahip konu başlıkları çıkartılmıştır. Bu çalışmada her il için yüksek frekansa göre ilk 10 adet konu tespit edilmiştir.

Tespit edilen konulara genel olarak bakıldığında Türkiye genelinde haberlerde çok sayıda üniversite ve spor ile ilgili haberin bulunduğu görülmektedir. Yapılan analiz sonuçları, illere ait çok simgeleşmiş bazı kişi, kurum veya değerleri içeren haberlerin sayısının oldukça fazla olduğu görülmektedir. Örneğin, Trabzon'da *hamsi*, Rize'de *çay*, Manisa'da *mesir*, Kırşehir'de *Neşet*, Konya'da *Mevlâna* eşleşmeleri örnek olarak verilebilir.

5. ÖNERİLER

Bu bölümde tez kapsamında kullanılan yöntemin başarısını ve temsil derecesini artıracak çalışmalardan bahsedilmektedir. Tez kapsamında Türkiye'deki illere ait ortalama 155,000 bin haber metni incelenmiş ve bu haber metinlerinden her il için 10'ar adet konu belirlenmiştir.

Veri madenciliği yöntemlerinin kullanımının yaygınlaşması ile birlikte birçok farklı bilgiye erişim sistemleri geliştirilmiştir. Özellikle son zamanlarda Türkçe dokümanlar içerisindeki metin verileri analiz edilerek anlamsal sonuçlar çıkartılabilmektedir. Türkçe karakterlerle yazılmış metinler üzerinden yapılan çalışmalar herhangi bir ön işleme aşamasından geçirilmeden doğru ve güvenilir sonuç vermemektedir. Bir diğer sorun ise Türkçe'de bir kelime kökünden çok fazla kelimenin türetilebilir olmasıdır. Bu durum, analiz sonucunu hem olumsuz etkilemekte hem de aynı köklü kelimelerin çıkmasına neden olmaktadır. Bu tez çalışmasının devamında Türkçe'ye uygun ek-kök ayrımı gibi morfolojik işlemlerin yapılması amaçlanmaktadır. Böylece il profilleri hakkında daha doğru bilgi çıkartılması ve il yöneticilerine illeri hakkında daha gerçekçi bilgiler verilmesine olanak sağlanacaktır.

İl profilinin tespit edilmesi ile cinayet, taciz, hırsızlık gibi konularda güvenlik güçleri tarafından alınan önlemler artırılabilir ve suç oranların düşmesi sağlanabilir Bunun yanı sıra, yerel yönetimler açısında yönetim ve hizmet faaliyetleri daha etkili bir biçimde yapılabilir.

Bu çalışmada, haber kaynağından daha fazla veri elde edilebilmesi, kaynağın geliştirici (developer) imkanlarının daha iyi olması ve analiz yaklaşımlarının Türkçe metinler için daha iyi adapte edilebilmesi sağlanırsa sonuçların daha başarılı olacağı düşünülmektedir. Özellikle çoğu il için elde edilen veri az olmasına rağmen iller hakkında beklenen bazı kelimelerin bulunması bu beklentiyi kuvvetlendirmektedir.

6. KAYNAKÇA

1 Visa A., Technology of Text Mining, International Conference on Machine Learning and Data Mining, Heidelberg 2001, 1-11.

2 Srivastava N. A. and Sahami M., Text Mining, First Edition, A Chapman & Hall Book, Boca Raton,2009.

3 Onan A., Türkçe Twitter Mesajlarında Gizli Dirichlet tahsisine Dayalı Duygu Analizi, 19. Akademik Bilişim Konferansı, Şubat 2017, Aksaray, Id:77.

4 Öztürk S., Sankur B., Göngör T. ve Yılmaz M. B., Türkçe Etiketli Metin Derlemi Turkish Labeled Text Corpus, IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı, Nisan 2014, Trabzon, Bildiriler Kitabı,1395–1398.

5 Tunali V., Bilgin T., ve Camurcu A., An Improved Clustering Algorithm for Text Mining: Multi-cluster Spherical K-means, The International Arab Joural of Information Technology, 13, 1 (2016), 12–19.

6 Park H. S. and Jun C. H., A simple and fast algorithm for K-medoids clustering, Expert Systems with Applications, 36,2 (2009), 3336–3341.

7 Celebi M. E., Kingravi H. A., and Vela P. A., A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm, Expert Systems with Applications, 40,1 (2013), 200–210.

8 Kawashima K., Text Mining and Pattern Clustering for Relation Extraction of Breast Cancer and Related Genes, 8th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, June 2017, Kanazawa, Bildiriler Kitabı, 59-63.

9 Levent V.E. ve Diri B., Türkçe Dokümanlarda Yapay Sinir Ağları ile Yazar Tanıma, 16. Akademik Bilişim Konferansı, Şubat 2014, Mersin, Bildiriler Kitabı, 735–741.

10 Bijuraj L. V, Clustering and its Applications, National Conference on New Horizons in IT, Ekim 2013, Bildiriler Kitabı, 169–172.

11 Amasyali M. ve Yildirim T., Automatic Text Categorization of News Articles, 12. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı, Ekim 2004, Kuşadası, Bildiriler Kitabı, 224–226.

12 Yıldız K., Çamurcu Y. ve Doğan B., A Comperative Analize of Principal Component Analysis, XII. Akademik Bilişim Konferansı, Şubat 2010, Muğla, 207-213.

13 Çalış K., Gazdağı O. ve Yıldız O., Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti, Bilişim Teknolojileri Dergisi, 6, 1 (2013), 1–7.

14 Kılınç D., Borandağ E., Yücalar F., Tunalı V., Şimşek M. ve Özçift A., KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi, Marmara Fen Bilimleri Dergisi, 28,3 (2016), 89–94.

15 Zaki M. J. and Meira W., Data Mining and Analysis: Fundamental Concepts and Algorithms, First Published, Cambridge University Press, New York, 2014.

16 Zhao Y., R and Data Mining: Examples and Case Studies, First Edition, Elsevier, Australia ,2011.

17 http://mgocenoglu.blogspot.com.tr/2014/06/veri-madenciligi-asamalar.html, Veri Madenciliği Aşamaları 09-May-2018.

18 Han J., Kamber M. and Pei J., Data Mining: Concepts and Techniques. Third Edition, Morgan Kaufman, Waltham, 2012.

19 Acar M. Veri Madenciliği Aşamaları https://mevlutcanvar.com.tr/veri-madenciligi-asamalari. 09-May-2018.

20 https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm #DMCON002, What is Data Mining?, 09-May-2018.

21 Savaş S., Topaloğlu N. ve Yılmaz M., Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21 (2012), 1-23.

22 Tüzüntürk S., Veri Madenciliği ve İstatistik, Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 29,1, (2010), 65–90.

23 S. Marsland, Machine Learning An Algorithmic Perspective, Second Edition, CRC, Press, Boca Raton, 2015.

24 Amanet H., Türkçe Sosyal Medya Metinlerinde Duygu Analizi, Yüksek Lisans Tezi, K.T.Ü., Fen Bilimleri Enstistüsü, Trabzon, 2017.

25 Ozdemir S., Principles of Data Science, First Published, Packt, Birmingham, 2016.

26 Jain A. K., Data clustering: 50 years beyond K-means, Pattern Recognition Letter, 31,8 (2010), 651–666.

27 Awad M. and Khanna R., Efficient Learning Machines, First Published, Apress Open, California, 2015.

28 Dinler M., Kümeleme Analizi Yöntemlerinin Hayvancilik Verilerinde Karşilaştirilmali Olarak İncelenmesi, Yüksek Lisans Tezi, Bingöl Üniversitesi, Fen Bilimleri Enstitüsü, Bingöl, 2014.

29 Gürcü Ö. Kümeleme Analizi, https://prezi.com/_qvp7mj981sx/kumeleme-analizi/. 09-May-2018.

30 Kirk M., Thoughtful Machine Learning, First Edition, O'Reilly Media, Sebastopol, 2015.

31 Seung-Seok C., Sung-Hyuk C., and Tappert C. C., A Survey of Binary Similarity and Distance Measures, Journal of Systemics Cybernetics and Informatics, 8,1 (2010), 43–48.

32 Cha S., Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions, International Journal of Mathematical Models and Methods in Applied Sciences, 1,4 (2007), 300–307.

33 Taha S. M., Metin Madencilği İle Doküman Denetleme, Yüksek Lisans Tezi, Gazi Üniversitesi, Bilişim Enstitüsü, Ankara, 2011.

34 Feinerer I., Hornik K., and Meyer D., Text Mining Infrastructure in R, Journal of Statistical Software, 25, 5 (2008), 1–54.

35 Nahm U. Y. and Mooney R. J., Text Mining with Information Extraction, Doctoral Dissertation, University of Texas, Austin, 2004.

36 Mahhabel B., Mishra P., Danneman N. and Heimann R., R: Mining Spatial, Text, Web and Social Media Data, Google Kitaplar, First Published, Packt, Birminghan, 2017, https://books.google.com.tr/books?id=HHg5DwAAQBAJ&printsec=frontcover&hl=tr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false. 11-May-2018.

37 Ergün K., Metin Madenciliği, Balıkesir Üniversitesi, 2008.

38 Tsai F. S., Text mining and visualisation of Protein-Protein Interactions, International Journal of Computational Biology and Drug Design, 4,3 (2011), 239-244.

39 Chakraborty G., Pagolu M., and Garla S., Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS, First Edition, SAS Institute Inc., North Carolina, 2013.

40 http://www.mtechprojects.org/information-retrieval-systems-projects.html., Information Retrieval Systems Projects and Research Topics – MTech Projects, 11-May-2018

41 Manning C. D., Ragahvan P. and Schutze H., An Introduction to Information Retrieval, Online Edition, Cambridge University Press, Cambridge, 2009.

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

43 L. Xiao, D. Wissmann, M. Brown, and S. Jablonski, Information Extraction from the Web: System And Techniques, Applied Intelligence, 24 (2004),195–224.

44 https://www.emaze.com/@AQLRLFLR/untitled, Veri Madenciliği:Web Sayfalarında Örüntü Keşfi, 11-May-2018.

45 Francis L.,FCAS, MAAA, Flynn M. and PhD, Text Mining Handbook, Casualty Actuarial Society E-Forum, Spring (2010), 1–61.

46 Talebi M., Facebook'da Yorum Madenciliği Kullanarak Kişilerin Cinsiyet, Yaş Ve Eğitim Düzeylerinin Tanımlanması, Yüksek Lisans Tezi, K.T.Ü., Fen Bilimleri Enstitüsü, 2013.

47 Neto J. L., Santos A. D. ,Kaestner C. A. A. and Freitas A.A., Document Clustering and Text Summarization, 4th International Conference Practical Applications of Knowledge Discovery and Data Mining, 2000, London, Bildiriler Kitabı,41-55.

48 Karaca M. F., Metin Madenciliği Yöntemi İle Haber Sitelerindeki Köşe Yazilarinin Siniflandirilmasi, Yüksek Lisans Tezi, Karabük Üniversitesi, Fen Bilimleri Enstitüsü, Karabük, 2012.

49 Balcı M., Metin İşlemede En Uzun Eşleşme Algoritmasının Karşılaştırmalı Analizi, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya, 2010.

50 Kumar A. and Paul A., Mastering Text Mining with R, First Edition, Packt, Birmingham, 2016.

51 Newman D., Asuncion A., Smyth P. and Welling M., Distributed Inference for Latent Dirichlet Allocation, 20th International Conference on Neural Information Processing Systems, Aralık 2007, Vancouver, Bildiriler Kitabı, 1081–1088.

52 Hoffman M. D., Blei D. M. and Bach F., Online Learning for Latent Dirichlet, 23rd International Conference on Neural Information Processing Systems, Eylül 2010, Vancouver, Bildiriler Kitabı, 856-864.

53 Blei D., Ng A.Y., Edu A. S. and Jordan M. I., Latent Dirichlet Allocation, Journal of Machine Learning Research, 3 (2003), 993–1022.

54 Sönmez N., Çevrimiçi Yorumlarin Metin Madenciliği ile Analizi:İstanbul'daki Alişveriş Merkezleri Üzerine Bir Çalişma, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2017.

55 https://developer.nytimes.com/.API Gallery - NYT Developers Network. 15-May-2018.

56 https://newsapi.org/sources. ,List of News Source APIs Available - News API. 15-May-2018.

57 https://www.webopedia.com/TERM/A/API.html , What is API - Application Program Interface, 15-May-2018.

7. EKLER

Ek 1. 160 Kelimeden Oluşan Durma Kelime Listesi

Acaba	Bazen	Birdenbire	Bütün	Edecek	Hangi	Madem	Sonra
Acep	Bazı	Biri	Çok	En	Hariç	Meğer	Söyledi
Açıkça	Başka	Birileri	Da	Epey	Hatta	Nasıl	Tam
Adet	Başkası	Birçok	Daha	Eski	Henüz	Neden	Tamam
Adeta	Belirterek	Birkaç	Dahil	Esnasında	Hem	Nere	Tamamen
Ait	Belki	Birşey	Daima	Etti	Hep	Neyse	Tarafından
Alanda	Ben	Biz	Dair	Eğer	Hepsi	Niye	Tek
Aldığı	Benden	Bizce	Dakika	Fakat	Herhangi	Niçin	Tüm
Altmış	Beni	Bizden	Dan	Farklı	Herkes	Olan	Üzere
Altı	Benim	Bizim	De	Gayet	Hiç	Olarak	Var
Ama	Beri	Bizzat	Dedi	Gayri	İçin	Oldukça	Ve
Amma	Beriki	Bu	Defa	Geçen	İle	Olsun	Velev
Ancak	Beş	Bulunan	Değil	Geldi	İlgili	Pekala	Verildi
Arada	Bilcümle	Buna	Demin	Gene	İse	Peki	Veya
Artık	Bile	Bundan	Derhal	Genel	İyi	Rağmen	Ya
Aslında	Bin	Bunlar	Devam	Gerek	Kadar	Saat	Yahut
Aynen	Binaen	Bunu	Değil	Geç	Kere	Sahi	Yaklaşık
Aynı	Bir	Bunun	Diye	Gibi	Keşke	Sana	Yakında
Ayrıca	Biraz	Burada	Diğer	Göre	Kişi	Sayılı	Yalnız
Az	Birazdan	Böyle	Dolayı	Halbuki	Kısaca	Sen	Yani
Bana	Birbiri	Böylece	Dolayısıyla	Halen	Lakin	Sizi	Yapacak
Bari	Birden	Böylelikle	Doğru	Hangi	Lütfen	Son	Yine

ÖZGEÇMİŞ

Kaan TOPRAK, 6 Haziran 1992 tarihinde Sivas'da doğdu. Orta öğretimini Erzincan Milli Piyango Anadolu Lisesi'nde tamamladıktan sonra 2010 yılında Karadeniz Teknik Üniversitesi İstatistik ve Bilgisayar Bilimlerinde lisans eğitimine başladı. 2015 yılında Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik ve Bilgisayar Bilimleri Anabilim Dalı'nda tezli yüksek lisans eğitimine başladı. 2017 yılında İspanya'da Oviedo Üniversitesinde 3 ay staj yaptı.

Toprak, İngilizce bilmektedir.