T.C.
AKDENİZ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
ÖLÇME VE DEĞERLENDİRME
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

COVİD-19 PANDEMİ SÜRECİNDE MİLLİ EĞİTİM BAKANI' NIN TWİTTER MESAJLARININ METİN MADENCİLİĞİ YÖNTEMİYLE İNCELENMESİ

EMİNE İÇÖZ

ÖLÇME VE DEĞERLENDİRME
BİLİM DALI

Antalya, 2021

T.C.

AKDENİZ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EĞİTİM BİLİMLERİ ANABİLİM DALI

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME

TEZLİ YÜKSEK LİSANS PROGRAMI

COVİD-19 PANDEMİ SÜRECİNDE MİLLİ EĞİTİM BAKANI' NIN TWİTTER MESAJLARININ METİN MADENCİLİĞİ YÖNTEMİYLE İNCELENMESİ

YÜKSEK LİSANS TEZİ

EMİNE İÇÖZ

Danışman

Doç. Dr. Alper SİNAN

Antalya, 2021

DOĞRULUK BEYANI

Yüksek lisans tezi olarak sunduğum bu çalışmayı, bilimsel ahlak ve geleneklere aykırı düşecek bir yol ve yardıma başvurmadan yazdığımı, yararlandığım eserlerin kaynakçalarda gösterilenlerden ibaret olduğunu ve bu eserleri her kullanışımda alıntı yaparak yararlandığımı belirtirek bunu onurumla doğrularım. Enstitü tarafından belli bir zamana bağlı olmaksızın, tezimle ilgili yaptığım bu beyana aykırı bir durumun saptanması halinde, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara katlanacağımı bildiririm.

20 / 01 / 2021

Emine İÇÖZ

TEŞEKKÜR

Yüksek lisans tezimin başından sonuna kadar tüm süreçte sorunlarıma çözüm bulan, çalışmalarımı yönlendiren, kıymetli bilgilerini bana aktaran ve vakit ayıran değerli tez danışmanım Sayın Doç.Dr. Alper Sinan'a teşekkürlerimi sunarım. Ayrıca analizlerimde yardımlarını esirgemeyen Sayın Doç.Dr. Bilal Barış Alkan'a ve yüksek lisans eğitimim süresince bilgilerinden ve tecrübelerinden faydalandığım Akdeniz Üniversitesi Ölçme ve Değerlendirme Ana Bilim Dalı hocalarıma ayrı ayrı teşekkür ederim.

Yüksek lisans eğitimim süresince her türlü yardımda bulunan, fikir alışverişleriyle yükümü hafifleten değerli arkadaşlarım Hanife Tekeli Akdemir, Figen Göçer ve Gamze İnal'a teşekkür ederim.

Her zaman yanında olan sevgili eşim Halil İçöz'e ve pek çok zaman ihmal etmek zorunda kaldığım kıymetli çocuklarım Erva Gül ve Mehmet Emir'e sonsuz teşekkürler.

Son olarak bu zorlu süreçte hiç bir desteğini esirgemeyen sevgili annem Feride Köse ve babam Ramazan Köse'ye teşekkür ederim.

Emine İÇÖZ
Antalya,Ocak 2021

ÖZET

COVİD-19 PANDEMİ SÜRECİNDE MİLLİ EĞİTİM BAKANI' NIN TWİTTER MESAJLARININ METİN MADENCİLİĞİ YÖNTEMİYLE İNCELENMESİ

İçöz, Emine

Yüksek lisans,Eğitimde Ölçme ve Değerlendirme Bölümü

Tez Yöneticisi: Doç. Dr. Alper SİNAN

Ocak 2021, 62 sayfa

Bu araştırmanın amacı, internetin yaygın kullanılması ile birlikte artan büyük miktarlardaki verilerden anlamlı bilgi çıkarmak için günümüzün en çok talep gören konularından biri olan metin madenciliği yönteminin eğitim alanında da kullanılabileceğine ışık tutmaktır. Bu kesitsel tanımlayıcı araştırma, uzaktan eğitim sürecinde eğitime sağlıklı bir şekilde yön vermek, gelişmeleri duyurmak, her zaman ve her durumda veli ve öğrencilerin yanında olduğunu belirtmek için twitter sosyal medya hesabından paylaşımlar yapan Türkiye Milli Eğitim Bakanı Ziya SELÇUK' un 23 Mart 2020 - 17 Ekim 2020 tarihleri arasında paylaştığı mesajların içerik analizini Metin madenciliği yöntemi ile yapan nitel bir araştırmadır. Analiz aracı olarak açık kaynak kodlu bir yazılım olan R- 3. 6. 1. tercih edilmiştir. Bununla beraber elde edilen kelime sayıları ile nicel araştırma yöntemi olan X^2 analizi uygulanmıştır

Araştırmada elde edilen veriler sözlük sayma tekniğiyle analiz edilmiş, kelime bulutu tekniğiyle görselleştirlmiştir. Süreci yansıttığı düşünülen 12 kelime seçilerek sözlük sayma tekniği ile frekansları belirlenmiştir. Daha sonra kelimeler 4 ana kategoride birleştirilerek SPSS programında ki-kare homojenlik testi yapılmıştır. Toplanan mesajlar uzaktan eğitim ve telafi eğitim programının yeniden başladığı 31 Ağustos tarihi baz alınarak ikiye ayrılmış ve bu iki dönem kategorilere ve kelimelerin kullanım yüzdesine göre kıyaslanmıştır.

Anahtar Kelimeler: *Metin Madenciliği, Veri Madenciliği, Kelime Bulutu, Uzaktan Eğitim, Twitter*

.

ABSTRACT

EXAMINING THE TWITTER MESSAGES OF THE MINISTER OF NATIONAL EDUCATION WITH THE TEXT MINING METHOD DURING THE COVID-19 PANDEMIC PROCESS

İçöz , Emine

Master Thesis , Department of Assessment and Evaluation in Education

Supervisor: Asst. Prof. Dr. ALPER SİNAN

January 2021,62 pages

The aim of this research is to shed light on the use of text mining method, one of the most demanded topics of today, in the field of education in order to extract meaningful information from the large amount of data that has increased with the widespread use of the internet. This cross-sectional descriptive study is a qualitative research that analyzes the content of the messages shared between 23 March 2020 - 17 October 2020 with the Text mining method by Turkey's National Education Minister Ziya SELÇUK via his twitter social media account to direct the education wholesomely in the distance education process, to announce developments and to indicate that he is always and in all situations with the parents and the students.

The data obtained in the study were analyzed with dictionary counting technique and visualized with word cloud technique. 12 words, which were thought to reflect the process, were selected and their frequencies were determined using the dictionary counting technique. Later, the words were combined into 4 main categories and a chi-square homogeneity test was performed in the SPSS program. It has been compared according to the categories and the percentage of use of words.

***Key Words**: Text Mining, Data Mining, Word Cloud, Distance Education, Twitter*

İÇİNDEKİLER

BÖLÜM III

YÖNTEM

BÖLÜM IV

BULGULAR VE YORUMLAR

BÖLÜM V

SONUÇ VE ÖNERİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

BÖLÜM 1

GİRİŞ

Teknolojik ilerlemeler,internet kullanımındaki hızlı artış ve beraberinde getirdikleri sosyal medya araçlarını kullanmadaki artış dijital veri tabanlarının büyümesiyle sonuçlanmaktadır. Depolanan ve hızla birikmeye devam eden bu verilerle bağlantı kurma,onlardan anlamlı ve işe yarar bilgiler çıkarma konusuna ilgi artmıştır. Bu alanla ilgilenen disiplinin genel adı veri madenciliğidir (Hand,Mannila ve Smyth, 2001).

Madencilik,gizli ve ulaşılması zor değerli kaynakların çıkarılmasını ifade eder. Madenciliğin depolanan verilerle ilişkilendirilmesi,mevcut veri yığınından daha önce fark edilmemiş önemli bilgilerin bulunmasını ve derinlemesine bir analiz yapılmasını önerir.
Veri madenciliği,yapılandırılmış veriler üzerinden analizler yaparak sınıflandırma veya tahmin yoluna gider. Fakat dijital ortamda depolanan ve her geçen gün artmaya devam eden bu veriler çoğunlukla yapılandırılmamış veya yarı yapılandırılmış verilerden oluşmaktadır. Bu tür verilerle ilgilenen veri madenciliği yöntemi metin madenciliği olarak adlandırılır.

Kurum ve kuruluşların pek çoğu verilerini elektronik ortamda saklamaktadır. İnternet üzerinde bloglar, sosyal medya platformları ve elektronik postalar gibi dijital kütüphaneler, veri depoları ve diğer metinsel bilgiler büyük veri yığınları oluşturmaktadır. Bu büyük veri yığınlarından önemli bilgiler elde etmeyi sağlayan uygun kalıpları ve yaklaşımları belirlemek zordur. Metin madenciliği, metinsel veri kaynaklarından ilgi çekici ve önemli bilgiler keşfetmek için kullanılan bir yöntemdir. Metin madenciliği, metinlerin işlenip anlamlı bilgiler sunabilmeleri için kullanılan bir uygulamadır. Metinlerin işlenmeye hazır hale gelmesi metin madenciliğiyle, işlenmesi ise veri madenciliği ile gerçekleştirilir (Karaca, 2012).

Günümüzde internet kullanımının yaygınlaşması ve artan teknolojiyle birlikte bireyler gündelik birçok işini bilgisayar, laptop, tablet, akıllı telefon gibi cihazlarla gazete, dergi, kitap, makale okumaktan banka işlemlerine kadar bir çok işlemi kısa sürede yapabilmektedir. Aynı zamanda artan teknoloji kullanımı insanların sosyal hayatlarını da kaçınılmaz olarak etkilemektedir. Kişilerin sosyal mecralarda paylaştıkları yazılı metinler oldukça artmıştır (Zontul ve Aydın, 2017; Kireççi, 2019).

Tüm dünya verilerini içeren "We are Social ve Hootsuite 2020" dijital raporuna göre dünyadaki nüfusun %59'u yani 4.54 milyarı internet kullanıcısı; %49'u yani 3.80 milyar sosyal medya kullanıcısı; %67'si yani 5.19 milyar insan mobil cihaz kullanıcısıdır. Bu durum da internet dunyasını metin madenciliğinin merkezine oturtmaktadır.

Bu araştırma kapsamında, sosyal medya mecralarının bu denli yoğun kullanıldığı günümüzde tüm insanlık olarak baş etmek zorunda kaldığımız koronavirüs salgını nedeniyle uzaktan eğitime sonrasında da aşamalı ve seyretilmiş modelle yüzyüze eğitime geçmek zorunda kaldığımız bu süreçte, Milli Eğitim Bakanı Ziya SELÇUK' un twitter sosyal medya hesabından paylaştığı mesajların metin madenciliği yöntemiyle analizi amaçlanmıştır.

1.1. Problem Durumu

Depolanan veri miktarı, internet kullanımının yaşantımıza girmesi ve dijital depolama aygıtlarının kapasitelerinin artrmasıyla yeni yaklaşımların ortaya çıkmasına neden olmuştur. Önceleri belgelerle ilgili her türlü işlem manuel yapılmaktayken günümüzde internet kullanımının yaşantımıza girmesiyle ve dijital depolama aygıtlarının kapasitelerini artırmasıyla bugün internette 2 milyardan fazla web sayfası bulunmaktadır ve bu bilgilere eski yöntemlerle ulaşmanın imkansız olduğu açık bir şekilde görülmektedir (Döven, 2013).

Veri madenciliği, büyük veri yığınları üzerinde analizler yapar, çıkarımlarda bulunur. Bu analiz ve çıkarımlar için yapılandırılmış verilere ihtiyaç duyar. Fakat dijital ortamlarda hızla artmakta olan veri yığınlarının çoğu yapılandırılmamış halde bulunur. Yapılandırılmamış veriler üzerinde analizler yapan veri madenciliği yöntemi metin madenciliğidir. Yapılandırılmış veriler, bulundukları yapı içerisinde organize edilebilen ve tanımlanabilen veriler için kullanılır. SQL (Structured Query Language) ve Access en yaygın kullanılan yapılandırılmış veri tabanlarıdır.Sütun ve satır düzeyinde tanımlanırsa bu verilere kayıt bazlı ulaşılabilir. Yapılandırılmamış veriler tanımlanabilir bir yapıya sahip değildir. Resim dosyaları, word, text dokümanları, web logları ve elektronik postalar en bilinen türleridir. Birçok kurumun verileri yapılandırılmamıştır. Metin madenciliği büyük yığınlardaki verilerin işlenerek gizli yapıların çıkarılması şeklinde tanımlanabilir (Dolgun ve diğ. 2009).

Metin madenciliği ile veri madenciliği yöntemleri aynı analiz yaklaşımı ve tekniklerini kullanır. Metin madenciliği ve veri madenciliği arasında etkileşimli bir iliski vardır. Metin madenciliği süreçleriyle yapısal veri elde edilir ve bu veriler, veri madenciliği modellerinde kullanılarak elde edilen sonuclar metnin yapısının incelenmesinde kullanılmaktadır (Çelikyay, 2010). Ancak veri madenciliği yapılandırılmış veri gerektirirken metin madenciliği yapılandırılmamış verilerdeki kalıpları keşfetmeyi amaçlar. Metin madenciliğinde girdi, yapılandırılmamış veya yarı yapılandırılmış olan bir metin kümesidir. Örneğin, bir metnin başlık, yazar, yayın tarihi ve kategorisi gibi birkaç yapılandırılmış bölümü olabilir. Fakat ,

yüksek bilgi değerine sahip içerik kısmı yapılandırılmamış bileşenlerden oluşur. Geleneksel veri madenciliği ile bu kısımlardan bilgi almak zordur (Dalmolen, 2010).

Weiss ve diğerlerine göre (2005), Metin madenciliğini destekleyen ana temalardan biri, metnin sayısal verilere dönüştürülmesidir, bu nedenle ilk veri farklı olsa da, bazı ara aşamalarda veriler klasik bir veri madenciliği işlemine taşınır. Yapılandırılmamış veriler yapılandırılır .

Metinlerden anlamlı bilgilerin elde edilmesi için, veri ön işleme ve özellik çıkarımı olarak adlandırılan bazı işlemlerin yapılması gerekmektedir. Bu aşamalardan sonra yapısal halde bulunmayan veriler, metin madenciliği kullanılabilecek ve bilgisayarlar tarafından işlenebilecek yapısal bir biçime dönüştürülmektedir. Bu sayede büyük veri yığınları içerisinde gizli değerli bilgiler ortaya çıkarılmış olur. Üretilen bu değerli bilgiler kullanılarak, kurum ya da kuruluşların kullanabileceği farklı sonuçlara ulaşılabilir. Temelinde matematiksel ve istatiksel yöntemler bulunan Metin madenciliği yöntemleri ; anahtar kelime elde etme, yazar tanıma, fikir madenciliği, metin sınıflama, duygu analizi, başlık elde etme gibi farklı alanlarda sıklıkla tercih edilmektedir (Kılınç ve diğ. , 2016).

Dijital ortamlarda yapılandırılmamış verilerin en çok depolandığı alanlardan biri de sosyal medya mecralarıdır. Sosyal medya, insanların paylaşım yapabildikleri, birlikte bir şeyler üretebildikleri, farklı konularda fikir alışverişi yapabildikleri, gelişmiş içerikli ve interaktif mobil teknolojilerden yararlanmaktadır. Kullanıcıların farklı kullanıcılarla çevrim içi haber, video, fotoğraf vb. paylaşım yapmalarını sağlayan sitelerin ortak adına sosyal medya denir (Özkan ve Türkmen, 2020).

Sosyal medya dendiği zaman Facebook, Twitter, Youtube gibi araçlar akla ilk gelenlerdir. Twitter, 2006 yılında kurulmuştur ve kullanıcıların ileti gönderip gönderilen iletiyi okumalarına izin veren sosyal bir ağdır (Tarhan 2012, 82). İnsanların diğer insanlardan ve onların düşüncelerinden haberdar olup iletişim kurmalarını sağlamaktadır. Ayrıca Twitter, kullanıcılara iletileri yeniden paylaşma, yorum yapma ve cevap verme imkânı sunmaktadır (İmik Tanyıldızı ve Ateş, 2018). Twitter, bedava olması ve herkese kolaylıkla ulaşılabilmesi nedeniyle günümüzde en sık tercih edilen sosyal medya araçlarından biridir.

Twitter hesaplarından paylaşılan mesajlar, resimler ve videolar yapılandırılmamış veri olarak internet ortamında depolanmaktadır. Bu büyük hacimli verilerden anlamlı çıkarımlar yapmak günümüzün merak uyandıran konularındandır.

Tüm dünyada ve ülkemizde yaşamakta olduğumuz Covid-19 salgını nedeniyle yüz yüze eğitime ara vermek zorunda kaldığımız 23 Mart 2020 tarihinden bu yana, süreci sağlıklı bir şekilde yürütebilmek ve veli-öğrenci ve öğretmenlerle sağlıklı iletişim kurabilmek ve gelişmelerden haberdar etmek için Milli Eğitim Bakanımız Ziya SELÇUK da Twitter

hesabını çok sık kullanmaktadır. Bu araştırma kapsamında uzaktan eğitim sürecinin başladığı tarihten bu yana Milli Eğitim Bakanı Ziya SELÇUK' un paylaştığı mesajlar üzerinde metin madenciliği yöntemi kullanılarak bilgi çıkarımı yapılmak istenmektedir.

1.2. Problem Cümlesi

Koronavirüs salgını nedeniyle 23 Mart 2020 tarihinde yüz yüze eğitime ara verilerek uzaktan eğitime geçilmiştir. 19 Haziran 2020'de yaz tatiline giren okullar, pandemi nedeniyle eksik kalan konuların telafisi amacıyla 31 Ağustos 2020'de açılmış ve 3 haftalık telafi programı uzaktan eğitim yoluyla sürdürülmüştür. 21 Eylül 2020 tarihinde aşamalı ve seyreltilmiş modelle eğitim başlamıştır. 23 Mart 2020 tarihinden 17 Ekim 2020 tarihleri arasında Milli Eğitim Bakanı Ziya SELÇUK' un Twitter hesabından paylaştığı mesajlar toplanarak oluşturulan metinlerden bilgi çıkarımı yapmak amaçlanmıştır. Bu amaçla süreçte en çok üzerinde durulan konulardan olduğu düşülen "YKS, LGS, telafi, veli, eğitim, EBA, salgın, oyun, sınav, teknoloji, lise, ilkokul" kelimeleri seçilmiştir. Bu amaçla aşağıdaki alt problemlere cevap aranmıştır:

1) Milli Eğitim Bakanı Ziya SELÇUK uzaktan eğitimin başladığı tarih olan 23 Mart 2020' den 17 Ekim 2020 tarihine kadar seçilen anahtar kelimeleri kaç kez kullanmıştır?
2)31 Ağustos 2020 yeni bir dönem olarak kabul edilirse anahtar kelimelerin kullanımları iki dönem arasında anlamlı bir şekilde farklılaşmakta mıdır? Dönemler arasındaki bu farklılaşma nasıl yorumlanbilir?

1.3. Araştırmanın Önemi

Teknoloji ve internetin hayatın ayrılmaz bir parçası olduğu gnümüzde haberleşmeden alışverişe pek çok konuda her işi internetten ve dijital ortamdan yürütülmektedir. E-postalar, dökümanlar, sosyal paylaşımlar vb. dijital ortamlardaki pek çok bilgi sürekli bir artış eğiliminde olan bir veri yığını oluşturmaktadır.

Hayatın vazgeçilmezi olan internet kullanımı eğitim alanına da yansımış gerek konu anlatımı sayfaları, dijital dökümanlar, e-kitaplar gerekse sosyal iletişim araçlarıyla eğitimi kapsayan her alanda kendini göstermiş ve incelenmeyi-analiz edilmeyi bekleyen veri yığınları oluşturmuş ve oluşturmaya devam etmektedir. Oluşan bu büyük miktarlardaki veri yığınlarını inceleyerek anlamlı bilgiler çıkarmak günümüzün en çok talep gören konularından biri haline gelmiştir. Verilerden anlamlı bilgilerin elde edilmesi için çeşitli veri madenciliği teknikleri kullanılmaktadır. Veri madenciliği, yapısal veriler üzerinde işlem yapar. Fakat internet

ortamında ve dijital ortamlarda bulunan veriler her zaman yapılandırılmış halde bulunmaz. Yarı yapılandırılmış veya yapılandırılmamış veriler üzerinde çalışan veri madenciliği türü ise Metin madenciliği olarak adlandırılır.

Metin madenciliğinde kullanılan makine öğrenmesi algoritmaları yapay zekaya dayanmaktadır. Milli Eğitim Bakanı Ziya SELÇUK' un da yapay zeka kullanımına ve metin madenciliğine verdiği önem 2023 Eğitim Vizyonunda dikkat çekmektedir. Eğitim vizyonunda büyük yer tutan 'Veriye Dayalı Yönetimle' ülke çapında eğitimin baştan sona verimli bir şekilde yönlendirilmesi amacıyla hem geçmiş kararlara yönelik nesnel çıkarımlar hem de geleceğe yönelik gerçekçi planlar yapılabileceği bunun için de farklı ve büyük yığılardaki verilerin işlenerek ilişkilendirilmesi, sürekli değişen şartlara göre yapılandırılması ve sebep sonuç ilişkisi açısından anlamlandırılmasının gerekliliği vurgulanmaktadır.

Bu gelişmeler göz önüne alındığında metin madenciliği uygulamalarının kısa bir zaman sonra eğitim ve öğretim alanında çok sık kullanılacağını göstermektedir. Bu araştırmanın önemi de bu konuya ışık tutup öncülük etmektir. Çünkü literatür taraması yapıldığı zaman görülmüştür ki metin madenciliği araştırmaları ülkemizde genellikle istatistik, mühendislik, işletme ve ekonometri alanlarında yapılmaktadır.

Dünya olarak pandemi nedeniyle zorlu bir süreçten geçtiğimiz bu dönemde ülkemizde eğitim uzaktan yürütülmektedir. Bu durumda internet ve teknoloji kullanımına verilen önemi arttırmıştır. Milli Eğitim Bakanı Ziya SELÇUK bu zorlu süreçte eğitimi sağlıklı bir şekilde yönlendirmek, veli ve öğrencileri desteklemek ve bilgilendirmek için sosyal medya iletişim araçlarını sıkça tercih etmektedir. Bu araştırmada, Bakan Ziya SELÇUK' un twitter hesabından yaptığı paylaşımlar toplanarak metin madenciliği yöntemiyle analiz edilmiş ve süreç değerlendirilmeye çalışılmıştır. Buradaki amaç metin madenciliği yöntemlerinin eğitim alanında da oldukça kullanışlı olduğunu göstermektir.

1.4. Sayıltılar

Bu araştırmada Milli Eğitim Bakanı Ziya SELÇUK' un twitter sosyal medya hesabından 23 Mart tarihinden 17 Ekim tarihine kadar paylaştığı mesajlar analiz edilmiş ve bu bu dönemin Covid- 19 pandemi sürecini yansıttığı varsayılmıştır.

1.5. Sınırlılıklar

Bu araştırmada mesaj toplamaya uzaktan eğitim sürecinin başladığı tarih olan 23 Mart' tan başlanmıştır. Fakat çalışılmakta olan araştırmanın analiz sürecine geçileceği için 17 Ekim'de sonlandırılmıştır. Toplamda yaklaşık 7 aylık bir süreç analize alınmıştır.

1.6. Tanımlar

Veri madenciliği: Veritabanınında net ve faydalı sonuçlar elde etmek amacıyla ilk başta bilinmeyen düzenleri veya ilişkileri keşfetmek için büyük miktarlarda verinin seçilmesi, araştırılması ve modellenmesi sürecidir.

Veri ambarı: Birbiriyle ilişkili verilerin sorgulandığı ve analizlerinin yapılabildiği bir veri deposudur.

Gürültülü veri: Büyük veri tabanlarında, veri girişi sırasında yapılan insan hataları ya da girilen değerin yanlış ölçülmesinden kaynaklı değerlere denir (Şen, 2008).

Boş değerler: Kendisi de dahil olmak üzere hiç bir değere eşit olmayan değerlere boş değer denir.

Artık veri: Problemde istenilen sonucu elde etmek için kullanılan örneklem kümesindeki gereksiz verilerdir.

Dinamik veri: İçeriğinin sürekli olarak değişen verilerdir.

Sosyal medya: Web 2.0 teknolojisini kullanan ve kullanıcıların kendi içeriklerini oluşturmasına ve paylaşmasına olanak sağlayan bir dizi internet tabanlı araçlardır.

Metin madenciliği: Metinsel veri kaynaklarından ilgi çekici ve önemli bilgiler keşfetmek için kullanılan bir süreçtir.

Doğal dil işleme: Yaygın olarak NLP (*Natural Language Processing*) olarak bilinen, yapay zekâ ve dilbilimin alt kategorisi olan, Türkçe ve İngilizce gibi doğal dillerin işlenmesi ve kullanılması amacı ile araştırma yapan bilim dalıdır.

Yapılandırılmış veri: Biçimlendirilmiş bir ambarda, tipik bir veritabanı olarak düzenlenmiş verilerdir.

Yapılandırılmamış veriler: Elektronik tablo sayfaları, veritabanı tabloları veya diğer doğrusal veya sıralı veri kümeleri gibi öğelerden daha az düzenli formu olan verilerdir.

Bilgi çıkarımı: Varlıklar, varlıklar arasındaki ilişkiler ve varlıkları yapılandırılmamış kaynaklardan tanımlayan öznitelikler gibi yapılandırılmış bilgilerin otomatik olarak çıkarılmasını ifade eder.

Metin özetleme: Belirli bir metnin kullanıcı için yararlı bilgiler sağlayan kısaltılmış bir örneğini otomatik olarak oluşturma işlemidir.

Sınıflandırma: Metinlerin içeriklerine göre önceden tanımlanmış konulara göre ayrılmasıdır.

Kümeleme: Bir araştırmada çalışılan öğeleri benzerliklerine göre belirli gruplar içinde toplayarak sınıflandırmaya, birimlerin ortak özelliklerini ortaya çıkarmaya ve bu sınıflar hakkında genel tanımlamalar yapmaktır.

Korpus: Mesaj, web sayfası, blog, haber, makale vb gibi yapılandırılmamış belgelerin analiz için belli bir amaçla toplandığı alandır.

Token: Parçalara ayrılmış bir metinde her bir parçaya verilen isimdir.

Gövdeleme: Çekimli fiilleri yalın hallerine dönüştüren bir algoritmadır.

Kökenine Döndürme (Stemming) : Metinde geçen aynı kökten türemiş kelimeleri basit hallerine döndürerek tek bir kelimeye indirgeme işlemidir.

Normalleştirme: Metnin içerisinde geçen kelimelerin küçük harflere dönüştürülerek farklılıklarının ortadan kaldırılması işlemidir.

Durak kelimeleri: Edat ve bağlaç gibi bir metnin içeriği hakkında nadiren bilgi veren kelimelerdir.

Döküman terim matrisi (DTM): Bir metin korpusunu temsil etmek için kullanılan en yaygın biçimlerden biridir. DTM, satırların belgeleri, sütunların terimleri ve hücrelerin her bir terimin her belgede ne sıklıkta gerçekleştiğini gösterdiği bir matristir.

Denetimli makine öğrenmesi: Bir algoritmanın bir veri üzerinden yapıları (gizli kalmış yapıları) öğrendiği tüm sınıflandırma tekniklerini içeren yaklaşımdır. Genellikle, bu algoritmalar, nasıl kodlama yapması gerektiğine dair yeterli örnekler verilirse , metinleri nasıl kodlayacağını öğrenebilir.

Denetimsiz makine öğrenmesi: Bir algoritmanın metindeki belirli kalıpları tanımlayarak bir model ortaya çıkardığı yaklaşımdır.

BÖLÜM II

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

Araştırmanın kuramsal çerçevesi ve konuyla ilgili yapılmış araştırmalar bu bölümde yer almaktadır. Literatürde yapılan araştırmalar ve sonuçlarına değinilmiştir. Kuramsal çerçeve sosyal medya, veri madenciliği ve metin madenciliği olarak ayrılmış ve her başlıkta konuyla ilgili bilgilere yer verilmiştir.

2.1. Sosyal Medya

2.1.1. Sosyal Medya Tanımı

Hızla değişen ve gelişen dünyada bilgisayarın icadı ve internetin bulunmasıyla birlikte giren sosyal medya, insanlık tarihi için büyük bir devrim niteliğindedir. İnternetin hayatımızın önemli bir parçası haline gelmesiyle ve Tim O'Reilly' nin ilk kez bahsettiği Web 2.0 terimi ile gelişmeye başlayan sosyal medya insan hayatının her alanında kullanılmaya başlanmıştır (Erestin, 2019).

Kişiler arkadaşlıklarını sosyal ağlar üzerinden başlatmakta ve devam ettirmektedir. Bu anlamda insanların diğer bireylerle fotoğraf, video, yazılı metin paylaştıkları platform sosyal medya olarak ifade edilmektedir (Kireççi, 2019). Bu gelişme ile birlikte iletişim biçimlerinde yaşanan bu değişiklikler sonucunda bilgi edinme ve öğrenme süreci, katılım, etkileşim, işbirliği, sosyalleşme gibi alanlarda farklılaşmanın sonucunda dönüşüm yaşanmıştır. Bireyler bu değişimle birlikte seçici ve katılımcı olarak karşılıklı etkileşime girme ve iletişimi etkileme fırsatı bulmuşlardır. Bu sayede insanlar özellikle bir uzmanlık alanı hakkında bilgi edinmek için başka öğrenme yöntemlerine göre hem katılımcı hem de paylaşımcı rolünde olmaktadırlar (Çevik, 2019).

Sosyal medya tanımına uygun olan bir uygulama veya web sitesi aşağıdaki özelliklere sahip olmadır:

a) Bağımsız kullanıcılara sahip olmalı,

b)Bağımsız kullanıcı kaynaklı içeriğe sahip olmalı,

c) Kullanıcıları arasında etkileşimin olmalı,

d) Zaman ve mekanda sınırla olmamalıdır (Erkul, 2009).

Literatür incelendiğinde sosyal medyanın tek ve kabul görmüş tek bir tanımının olmadığı görülmekte; farklı kaynaklarda farklı tanımlara rastlanmaktadır. Sosyal medya

kullanıcıların kendi içeriklerini oluşturmasına ve paylaşmasına olanak sağlayan, web 2.0 teknolojisini kullanan bir dizi internet tabanlı araçlardır (Uzun vd., 2016).

Web 2. 0 platformunda bulunan ve internet kullanıcılarına tüm dünya ile etkileşime girme; iletişim kurma; fikirler, içerikler, düşünceler, deneyimler, perspektif ve bilgi paylaşma imkanı sağlayan internet temelli uygulamalardır (Chan ve Denizci Guillet, 2011: 347). Bir diğer tanıma göre ise sosyal medya; kullanıcılara katkı sağlama, geliştirme, puanlama, işbirliği yapma, internet içeriğini çok sayıda kullanıcıya ulaştırma ve internet uygulamalarını kişiselleştirme fırsatı veren akıllı internet servisleridir (Hvass ve Munar, 2012).

2.2. Veri Madenciliği

2.2. 1. Veri Madenciliği Nedir?

Dijital veri toplama ve depolama teknolojisindeki ilerlemeler, veritabanlarının büyümesiyle sonuçlandı. Bu büyüme, en sıradan verilerden (süpermarket işlem verileri, kredi kartı kullanım kayıtları, telefon görüşmesi ayrıntıları ve hükümet istatistikleri gibi) daha dikkat çekici (astronomik organların görüntüleri, moleküler veritabanları ve tıbbi kayıtlar gibi) verilere kadar gerçekleşti. Bu verilerle bağlantı kurma, onlardan değerli bilgiler çıkarma konusuna ilgi artmıştır (Hand, Mannila ve Smyth, 2001).

Veri madenciliği, karmaşık veri kümelerinde gizli kalmış ve değerli kaynakların ayıklanmasını ifade eder. Bilimsel araştırmalar açısından veri madenciliği, temel olarak bilgi işlem, pazarlama ve istatistik gibi diğer alanlarda yürütülen çalışmalardan geliştirilen nispeten yeni bir alandır. Veri madenciliğinde kullanılan yöntemlerin çoğu, makine öğrenmesi ve istatistiğe dayanmaktadır.

Makine öğrenmesi bilgisayar bilimi ve yapay zeka ile bağlantılıdır. Verilerde genel geçer doğrulara çevrilebilecek ilişkiler bulmakla ilgilidir. Makine öğrenmesinin amacı, veri bilimcilerin gözlemlenen verilerden yola çıkarak gözlemlenmemiş yeni yapılara genelleme yapmalarını sağlayan bir model oluşturma sürecidir.

Veri madenciliği terimi, ilk olarak Usama Fayaad tarafından 1995 yılında Montreal' de düzenlenen ve hala bu konuyla ilgili ana konferanslardan biri olarak kabul edilen Birinci Uluslararası Bilgi Keşfi ve Veri Madenciliği Konferansı' nda ortaya konmuştur. Daha önce bilinmeyen bilgileri, gözle görülür bir düzen veya önemli bir ilişkiye sahip görünmeyen kitlesel gözlemlenmiş verilerden tahmin etmek amacıyla çeşitli aşamalara bölünmüş bir dizi entegre analitik teknikten bahsedilmiştir.

Veri madenciliğin tarihsel gelişimi şu şekilde sıralanablir:

- 1950' ler, ilk bilgisayarlar (sayım için) ,

- 1960' lar, veri tabanı ve verilerin depolanması,
- 1970' ler, ilişkisel veri tabanı yönetim sistemleri,basit kurallara dayanan uzman sistemler ve makine öğrenimi,
- 1980' ler, büyük miktarda veri içeren veri tabanları,SQL sorgu dili,
- 1990' lar, veri tabanlarında bilgi keşfi çalışma grubu ve sonuç bildirgesi,veri madencliği için ilk yazılım,
- 2000' ler, tüm alanlar için veri madenciliği uygulamaları.

Veri madenciliğinin tanımı, veritabanınında net ve faydalı sonuçlar elde etmek amacıyla ilk başta bilinmeyen düzenleri veya ilişkileri keşfetmek için büyük miktarlarda verinin seçilmesi, araştırılması ve modellenmesi sürecidir (Gıudıcı, 2003).

Farklı kaynaklarda geçen veri madenciliği tanımları şu şekildedir:

- Veri madenciliği, çok büyük miktarlarda bilginin depolandığı veri tabanlarından, amaç doğrultusunda, gelecekle ilgili tahminler yapma imkanı sunan, anlamlı olan veriye ulaşma ve o veriyi kullanma şeklinde tanımlanabilir (Savaş, Topaloğlu ve Yılmaz,2012).
- Veri madenciliği, daha önce beklenmeyen bilgileri büyük veritabanlarından çıkarmak ve sonuçları karar vermeye uygulamak için çok aşamalı bir süreçtir (Ayre, 2006).
- Veri madenciliği, bir bilgisayar programı kullanarak geleceğe dair tahminlerde bulunmak için çok yararlı olacak çok sayıda veri arasındaki ilişkileri bulmamızı sağlayan bilgiye ulaşma işidir (Doğan ve Türkoğlu, 2007).
- Veri madenciliği, veri tabanında beklenmeyen ilikilerin bulunmasına yardım eden, yapay zeka ve veri görselleştirme alanındaki çalışmaların zorunlu hale geldiğini ve bu açılardan geleneksel istatistiksel analizlerden farklılık gösteren, Sınıflama, tahmin, bölümlendirme/kümeleme, tanımlama/ özetleme olmak üzere dört boyutta incelenen bir analiz yöntemidir (Oğuzlar, 2005).

Veri madenciliği, depolanan verilerde gizli kalmış çok sayıda veri arasındaki ilişkiyi irdeleyerek ilişkilerini tespit etme; istatistiksel ve matematiksel tekniklerle gereksiz verilerin elenerek anlamlı yeni korelasyonları, yapıları ve eğilimleriortaya çıkarma; gizli kalmış, değerli, anlamlı ve yararlı olan veriyi bulma; ileriye yönelik tahminlerde bulunmaya imkan sağlayan veri analiz tekniğidir. Temel amacı pekçok bilgi içinden gizli, potansiyel ve değerli bilgileri bulmaktır.

Bu işlemlerin uygulama sahası hayli büyüktür. Bahsedilen bu sahalar veri tabanı sistemleri, Veri Görselliği, Yapay Sinir Ağları, İstatistik, Yapay Öğrenme, vb. gibi disiplinleri içermektedir. Bu durum Şekil 4' te özetlenmektedir.

Şekil 2.1. *Veri Madenciliği ve İlgili Alanlar* (Savaş, Topaloğlu ve Yılmaz, 2012)

2.2.2. Veri Madenciliğinin Kullanım Alanları

Veri madenciliği günümüzde bir çok alanda etkili bir şekilde kullanılarak en çok uygulanan disiplinlerden biri haline gelmiştir. Kolay uygulanabilir olması ve etkili sonuçlar vermesi her geçen gün daha da yaygın bir kullanım alanı bulmasını sağlamaktadır. Literatür taramasıyla elde edilen veri madenciliği ile gerçekleştirilmiş uygulamaları ve kullanım alanları başlıca eğitim, bankacılık ve borsa, ticaret, mühendislik, spor, tıp, telekominikasyon, sismoloji, güvenlik, makine ve biyoloji olarak özetleyebiliriz (Savaş, Topaloğlu ve Yılmaz, 2012; Murathan ve Devecioğlu, 2018).

2.2.3. Veri Madenciliği Süreçleri

Veri madenciliği süreci problemin tanımlanmasından sonuçların değerlendirilmesi kadar olan işlemlerin bütünüdür. Bir aşamanın sonucu diğer aşamanın girdisi olduğu için tüm süreç birbiriyle bağımlıdır (Bölükbaş, 2013). Veriden anlamlı yapılar çıkarma sürecine literatürde, veri madenciliği, bilgi çıkarımı, bilgi keşfi, veri arkeolojisi ve veri yapı işleme gibi isimler verilmektedir.

Veri madenciliğinin beş ana sürece ayrıldığı görülmektedir. Bu süreçler:

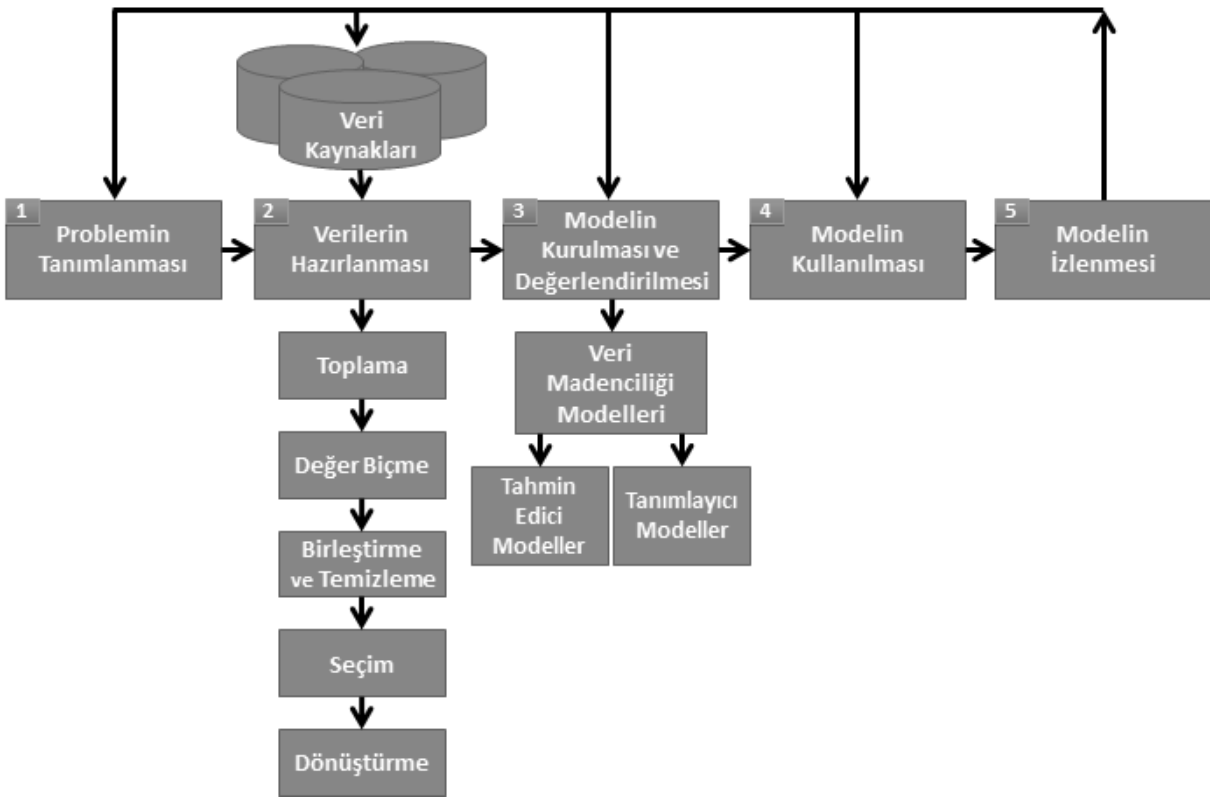
1. Problemin Tanımlanması,
2. Verilerin Hazırlanması,

3. Modelin Kurulması ve Değerlendirilmesi,

4. Modelin Kullanılması,

5. Modelin İzlenmesi,

olarak tanımlanmaktadır (Albayrak, 2008).



Şekil 2.2. *Veri Madenciliği Süreçleri* (Özbay, 2015)

2.2.3.1. Problemin Tanımlanması

Veri madenciliğinin hedefleri süreç başlamadan önce iyi anlaşılmalıdır. Veri madenciliği uygulamalarında problemin tanımlanması ilk ve en önemli aşamalarından biridir. Problem ve hedeflerin net olarak ifade edilmesi analizin doğru olarak yapılandırılması için ön koşuldur. Problemin doğru tanımlanması başarıyı arttıracak ve problemin çözümü hızlı ve doğru bir şekilde gerçekleşecektir. Problemin yanlış tanımlanması zaman ve maliyet kaybına neden olacağı için böyle bir durumda ortaya çıkacak maliyet ve doğru tahminlerde elde edilecek faydalara ilişkin tahminler de bu aşamada yer almalıdır. Bu adımda yapılan tanımlamalar ve çalışmanın ne şekilde yapılacağının belirleniyor olması bu aşamayı veri madenciliği uygulamalarında en zor adım kılmaktadır. Bu yüzden amaç ve problemler şüphe ve belirsizlik içermemelidir (Ayre, 2006; Guidici, 2003; Aydın, 2007; Onat, 2008; Özbay, 2015).

2.2.3.2. Verilerin Hazırlanması

Veri madenciliği kullanılırken oluşturulacak modelin veri kaynaklarının neler olduğunun tespiti ve modelde kullanılmak için uygun hale getirilmesi bu aşamada oluşturulur (Özbay, 2015). Veri madenciliğinin en önemli asamalarından bir tanesi olan verinin hazırlanması (veri ambarının olusturulması) asaması veri bilimcinin toplam zaman ve enerjisinin %50-70' ini harcamasına neden olmaktadır (Albayrak, 2008). Çünkü model kurulma basamağında karşılaşılan problemler, bu kısıma sıklıkla geri dönülmesine ve verilerin yeniden gözden geçirilip düzenlenmesine sebep olmaktadır (Savaş, Topaloğlu ve Yılmaz, 2012; Boyacı, 2017). Verilerin hazırlanması süreci "toplama", "değer biçme", "birleştirme ve temizleme", "seçim" ve "dönüştürme" olmak üzere 5 aşamadan oluşmaktadır (Ayık, Özdemir ve Yavuz, 2007).

a)Toplama (Collcetion)

Bu aşamada tanımlanan problem için gerekli olan verilerin toplanacağı veri kaynaklarının neler olduğu ve bu verilerin hangi kaynaklardan toplanacağı belirlenir. Verilerin hangi kaynaklardan toplandığı önemlidir; çünkü, verilerin güvenilirliği doğru sonuçlara ulaşılıp ulaşılamayacağını etkilemektedir. Ayrıca yetersiz veri ya da gereğinden fazla veri de veri madenciliği sürecini etkiler ve zaman kaybına sebep olabilir (Şengür, 2013; Özbay, 2015).

b)Değer Biçme (Assesment)

Veri uyumsuzlukları analiz edilecek verilerin farklı kaynaklardan toplanmasından kaynaklanmaktadır. En sık karşılaşılan uyumsuzluklar farklı ölçü birimleri, farklı zamanlarda gerçekleşmiş olma, kodlama farklılıkları (örneğin bir veride cinsiyet e/k, diğer veride 0/1 olarak kodlanması)dır. Bunların yanı sıra analiz edilecek verilerin hangi koşullar altında, nasıl, nerede ve toplandığı da önemlidir. Bu yüzden, iyi sonuç alınabilecek modeller sadece iyi verilerin üzerine kurulabileceğinden toplanan verilerin ne derece uyumlu oldukları bu aşamada değerlendirilmelidir (Karataş, 2019; Karabatak ,2008; Ayık,Özdemir ve Yavuz, 2007).

c)Birleştirme ve Temizleme (Consolidation and Cleaning)

Bu aşamada, bir önceki aşamada uyum sorunu giderilen farklı kaynaklardan elde edilen veriler tek veri tabanında birleştirilir. Birleştirilen veri tabanındaki hatalı veya eksik veriler belirlenerek temizlenir. Bu aşamada yapılacak hatalar sürecin başına dönülmesine sebep olabileceği için titiz davranılmalı ve sadece çıkarılması gereken veriler çıkarılmalıdır.

Basit yöntemlerle ve gelişigüzel yapılan sorun giderme işlemleri, ileride daha büyük sorunlara yol açabilir (Şengür, 2013; Özbay, 2015; Karataş, 2019; Karabatak, 2008; Onat, 2008).

d) Seçim (Selection)

Bu adımda, kurulacak olan modele göre veri seçilir. Mesela,öngörü yapan(tahmin edici)bir model için, bu aşama bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamına gelir. Sıra numarası, kimlik numarası vb bir anlam taşımayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da sebep oluşturabilecek değişkenlerin modele girmemesi şarttır. Bazı veri madenciliği algoritmaları konuyla alakasız bu tip değişkenleri otomatik olarak yok etse de, uygulamada bu işlemin kullanılan yazılıma bırakılmaması daha mantıklıdır (Karabatak, 2008).

Çok büyük veri tabanı kullanılması durumunda tesadüfiliği bozmayacak bir örnekleme yapılması uygundur. Tüm veri tabanının kullanılarak bir kaç model denenmesinden ziyade tesadüfi olarak örneklenmiş bir veri tabanında bir çok modelin denenmesi ve bunlardan en güvenilir ve güçlü modelin seçilmesi daha akılcı olacaktır.

e) Dönüştürme (Transformation)

Bu adımda, modelde kullanılacak veriler bazı kodlamalar kullanılarak tanımlanır veya gösterim şekli değiştirilir. Veri tabanında özet veya bağlantılı olan veriler daha anlamlı bir yapıya dönüştürülür (Özbay, 2015; Karataş, 2019). Veri dönüşümünde verilerin veri madenciliği için uygun formlara dönüştürülmesi düzeltme, bir araya getirme, genelleme, normalleştirme ve özellik oluşturma işlemleriyle gerçekleştirilir (Taşdemir, 2012).

2.2.3.3. Modelin Kurulması ve Değerlendirilmesi

Veri madenciliğinde belirlenen probleme uygun ve bizi sonuca götürecek en iyi modele ancak çok sayıda modelin kurulup denenmesiyle ulaşılabilir. Bu yüzden modelin kurulma aşaması en iyi modele ulaşılıncaya kadar yinelenir (Özbay, 2015).

Denetimli ve denetimsiz öğrenmenin kullanıldığı modellere göre model kurma aşaması değişmektedir. Denetimli öğrenme, örnekten öğrenme olarak da isimlendirilir. İlgili sınıflar önceden belirlenen bir kritere göre bir denetçi tarafından ayrılır ve her sınıf için farklı örnekler verilir. Burada amaç örneklerden yola çıkarak her bir sınıfa ilişkin niteliklerin keşfedilmesi ve bu niteliklerin kural cümleleri ile oluşturulmasıdır. Öğrenme süreci bittiğinde, yeni örneklere tanımlanan kural cümleleri uygulanır ve kurulan model tarafından bu yeni örneklerin hangi sınıfa ait olduğu tanımlanır. Denetimsiz öğrenmede ise, örneklerin gözlenmesi ve bahsedilen örneklerin özellikleri arasındaki benzerliklerden yola çıkarak

sınıfların belirlenmesi hedeflenir ki bu süreç kümeleme analizinde olduğu gibidir. Denetimli öğrenme modelinde seçilen algoritmaya uygun veriler hazırlanır.İlk aşamada verinin bir bölümü model öğrenimi, diğer bölümüyse model geçerliliğinin test edilmesi için ayrılır. Model öğrenimi, öğrenim kümesi kullanılarak sağlandıktan sonra, test kümesi ile modelin doğruluk derecesi saptanır (Karabatak, 2008).

Kurulan modelin değerlendirilmesinde kullanılan bir diğer ölçü de modelin önerdiği uygulamadan elde edilecek kazancın bu uygulamanın gerçekleştirilmesinde katlanılacak maliyete bölünmesi ile ulaşılacak yatırımın geri dönüş oranıdır. Kurulan modelin doğruluk derecesi ne kadar yüksek olsa da gerçek dünyayı tam olarak modellediğini garanti etmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca sebepler, modelin kurum aşamasında kabul edilen varsayımların ve modelde kullanılan verilerin doğru olmamasıdır.

2.2.3.4. Modelin Kullanılması

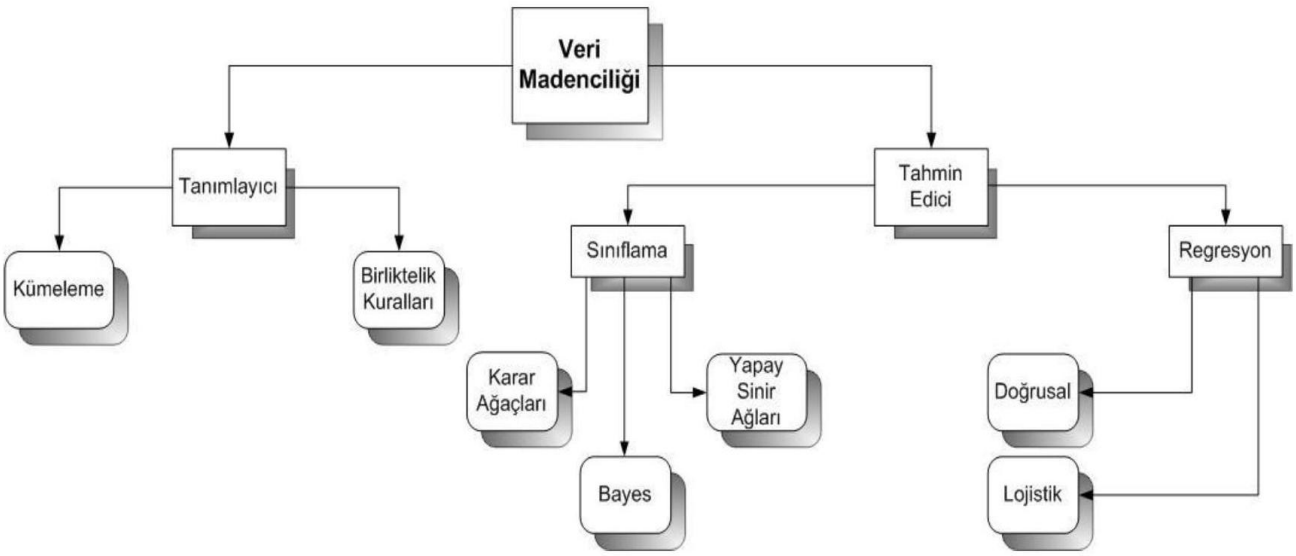
Belirlenen probleme uygun bir şekilde oluşturulan ve geçerliliği test edilerek onaylanan bir model doğrudan problemin çözümünde kullanılabileceği gibi başka problemlerin alt uygulaması olarak da tercih edilebilir.

2.2.3.5. Modelin İzlenmesi

Zamanla ortaya çıkan tüm sistem özelliklerinde ve ürettilen verilerdeki meydana gelen farklılıklar, kurulan modelin devamlı takip edilmesini ve gerektiği zaman güncellenmesini gerektirecektir. Gözlenen ve tahmin edilen ve değişkenler arasındaki farklılığı gösteren grafikler, model sonuçlarının izlenmesinde sık tercih edilen yöntemlerdendir.

2.2.4. Veri Madenciliği Modelleri

Veri madenciliğinde tercih edilen modeller iki başlık altında incelenebilir: tahmin edici(predictive) ve tanımlayıcı(descriptive).Tahmin edici modeller, sonuçları bilinen verilerden yola çıkılarak bir model geliştirilen ve bu modelden faydalanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edildiği modellerdir. Tanımlayıcı modeller ise karar vermeye rehberlik etmek için kullanılabilecek mevcut verilerdeki yapıların tanımlanması şeklindeki modellerdir (Özekeş, 2003).



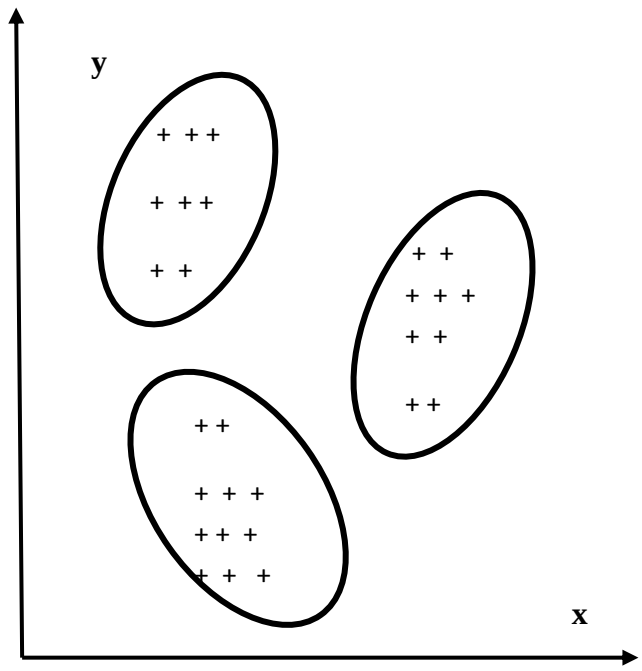
Şekil 2.3. *Veri Madenciliği Modelleri* (Aksoy, 2014)

Veri madenciliği modellerini işlevlerine göre 3 başlık altında incelemek mümkündür:

i) Kümeleme (Clustering)

Bölümleme olarak da bilinen kümeleme analizi, verileri alt kümelere ayıran bir yöntemdir. Her bir kümedeki elamanlar birbirlerine çok benzemekte, özellikleri farklı olanlar ise farklı kümelerde yer almaktadır. Başlangıçta veri tabanında bulunan kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı bir uzman tarafından belirtilebileceği gibi bilgisayar programlarından da faydalanılabilinir (Onat, 2008; Şengür, 2013).

Üç kümeye ayrılmış örnek bir veri seti şu şekilde gösterilmektedir:



Şekil 2.4. *Kümeleme Yöntemi*

ii) Birliktelik Kuralı (Association Rule) ve Ardışık Zamanlı Yapı (Sequential Pattern)

Birliktelik Kuralı, büyük veri yığınlarındaki dikkat çekici ilişkileri veya bağlantıları bulmak için kullanılırlar. Veri kümesi içinde yer alan ve sık görülen durumları tespit ederek birliktelik ilişkilerini bulur. Her geçen gün eldeki verinin büyümesi veri tabanlarındaki birliktelik ilişkilerinin tespit edilmesi ihtiyacını doğurmuştur. En yaygın kullanım alanı market sepet analizidir(Şen, 2008; Han&Kamber, 2001; Özekeş, 2003). Örneğin, "Düşük yağlı peynir ve yağsız süt alan müşteriler % 85 olasılıkla diyet süt alırlar. " ifadesi birliktelik kuralına ait bir ifadedir.

Ardışık zamanlı yapı ise birbiriyle ilişkili fakat birbirini takip eden zamanlarda gerçekleşen ilişkilerin tanımlanmasında kullanılır. "Şemsiye alan müşterilerin %10'u bir ay içerisinde yağmurluk almaktadır. " ifadesi ardışık zamanlı yapıya örnektir (Şen,2008).

iii) Sınıflama (Classification) ve Regresyon (Regression)

Veri madenciliği teknikleri arasında en sık kullanılan modellerden olan Sınıflama ve Regresyon, eldeki mevcut verilerden yola çıkılarak modelden bir kestirim için kullanılır. İki model birbirinden ayıran temel fark ise kestirilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değer olmasıdır. Sınıflandırma (classification) modeli; yeni verinin incelenmesi ve daha önceden tanımlanan ve özellikleri açık bir şekilde belirlenen bir sınıfa dahil

edilmesidir. Regresyon modeli ise, iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanarak konuyla ilgili tahminde bulunma yöntemidir. Tüm verilerin her zaman sınıflandırma, kategorize etme ve derecelendirme yöntemlerine ihtiyacı olduğu için sınıflandırma veri hazırlama aracı olarak veri madenciliğinin temelini oluşturur(Aksoy, 2015; Onat, 2008; (Murathan ve Devecioğlu, 2018).

Sınıflama teknikleri eğitim alanında öğrenci davranışlarını, Bir konuya ilgilerini ve sınav sonuçlarını kestirmek için kullanılabilir (Kumar ve Vİjayalakshmi).

Sınıflama ve regresyon modellerinde tercih edilen teknikler :

- Karar Ağaçları (Decision Trees)
- Yapay Sinir Ağları (Artificial Neural Networks)
- Genetik Algoritmalar (Genetic ALgorithms)
- K-En Yakın Komşu (K-Nearest Neighbor)
- Bellek Temelli Nedenleme
- Lojistik Regresyon
- Bayes Sınıflandırıcıları
- Diskriminant analizi (Discriminant Analysis)
- .Naïve-Bayes
- . Kaba kümeler
- Bellek Temelli Nedenleme (Memory Based Reasoning)

şeklinde sıralanabilir(Şengür, 2013; Aksoy, 2014; Albayrak, 2008).

2.2.5. Veri Madenciliği Uygulamalarında Karşılaşılan Problemler

Veri madenciliği girdi olarak kullanılan ham veriyi veri tabanlarından aldığı için veri tabanlarındaki verinin dinamik, eksiksiz ve net veri içermemesi sorun yaşanmasına sebep olur. Ayrıca küçük verilerde hızlı ve doğru çalışan bir sistem, çok büyük veri tabanlarında doğru çalışmayabilir (Aydoğan, 2003; Şen, 2008; Albayrak, 2008).

Veri madenciliği uygulamalarında karşılaşılan sorunlar:

a) Veritabanı Boyutu

Karşılaşılan en büyük sorunlardan biri çok büyük veri boyutlarıdır. Küçük test verileri için geliştirilmiş bir algortima, çok büyük test verileri ile kullanılabilmesi çok dikkat gerektirir. Bu yüzden, veri madenciliği yöntemleri için iki olasılık söz konusudur. Birincisi,

sezgisel bir yaklaşımla arama uzayını taramak olup ikincisi ise, test verilerini en aza indirmektir (Sever ve Oğuz, 2002).

b) Gürültülü Veri

Büyük veri tabanlarında, veri girişi sırasında yapılan insan hataları ya da girilen değerin yanlış ölçülmesinden kaynaklı bir çok hatalı değer bulunabilir. Bu tür sistem dışı hatalara gürültü adı verilir (Şen, 2008).

c) Boş Değerler

Kendisi de dahil olmak üzere hiç bir değere eşit olmayan değerlere boş değer denir. Boş değere sahip bir nitelik bilinmeyen ve uygulanamaz bir değere sahiptir. Boş değerli bir nitelik veri kümesinde yer alıyorsa ya ihmal edilmeli ya da en yakın değer atanmalıdır (Altıntop, 2006).

d) Eksik Veri

Eldeki mevcut veri,kurum ihtiyaçları gözönünde bulundurularak düzenlenip toplandığı için gerçek hayatı yeterince yansıtmayabilir. Eksik veri, yapılacak istatistiksel analizlerde önemli problemler yaratabilir. Çünkü analizler ve onların yapılmasını sağlayan ilgili paket programlar, tüm verilerin var olduğu durumlar için geliştirilmiştir. Eksik veriler olduğunda yapılması gerekenler,

- Eksik veri içeren kayıt veya kayıtlar çıkarılabilir.
- Değişkenin ortalaması kullanılabilir.
- Var olan verilere dayalı olarak en uygun değer kullanılabilir.

şeklinde sıralanabilir (Altıntop, 2006; Albayrak, 2008; Savaş, Topaloğlu ve Yılmaz, 2012).

e) Artık Veri

Problemde istenilen sonuca ulaşmak için kullanılan örneklem kümesindeki gereksiz verilerdir. Artık verileri yok etmek için oluşturulmuş algoritmalar, özellik seçimi olarak isimlendirilir. Özellik seçimi arama alanını küçültür ve sınıflama işleminin kalitesini artırır (Taşdemir, 2012).

f) Farklı Tipteki Verileri Ele Alma

Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik ya da kategorik veri tipleri değil, aynı zamanda tamsayı, kesirli sayı, çoklu ortam verisi, coğrafi bilgi içeren veri gibi farklı türdeki veriler üzerinde işlem yapılmasını gerektirir.Veri tipi

çeşitliliğinin fazla olması bir veri madenciliği algoritmasının tüm veri tiplerini ele alabilmesini olanaksızlaştırmaktadır. Bu yüzden veri tipine özgü veri madenciliği algoritmaları geliştirilmektedir (Sever, Oğuz, 2002).

g) Dinamik veri

Kurumsal çevrimiçi veritabanları dinamikliği yani içeriğinin sürekli olarak değişiyor olması bilgi keşfi metotları için önemli sakıncalar doğurmaktadır. Sadece okuma yapan ve uzun süre çalışan bilgi keşfi metodu mevcut veritabanı ile birlikte çalıştırıldığında mevcut uygulamanın da performansı ciddi ölçüde düşecektir. Bir başka sakıncası ise veritabanında bulunan verilerin kalıcı olduğu varsayılıp, çevrimdışı veri üzerinde bilgi keşif metodu çalıştırıldığında, değişen verinin elde edilen yapılara yansıması gerekir. Bu işlem, bilgi keşfi metodunun ürettiği yapıları zaman içinde değişen veriye göre sadece ilgili yapıları güncelleme yeteneğine sahip olmasını gerektirir (Altıntop,2006).

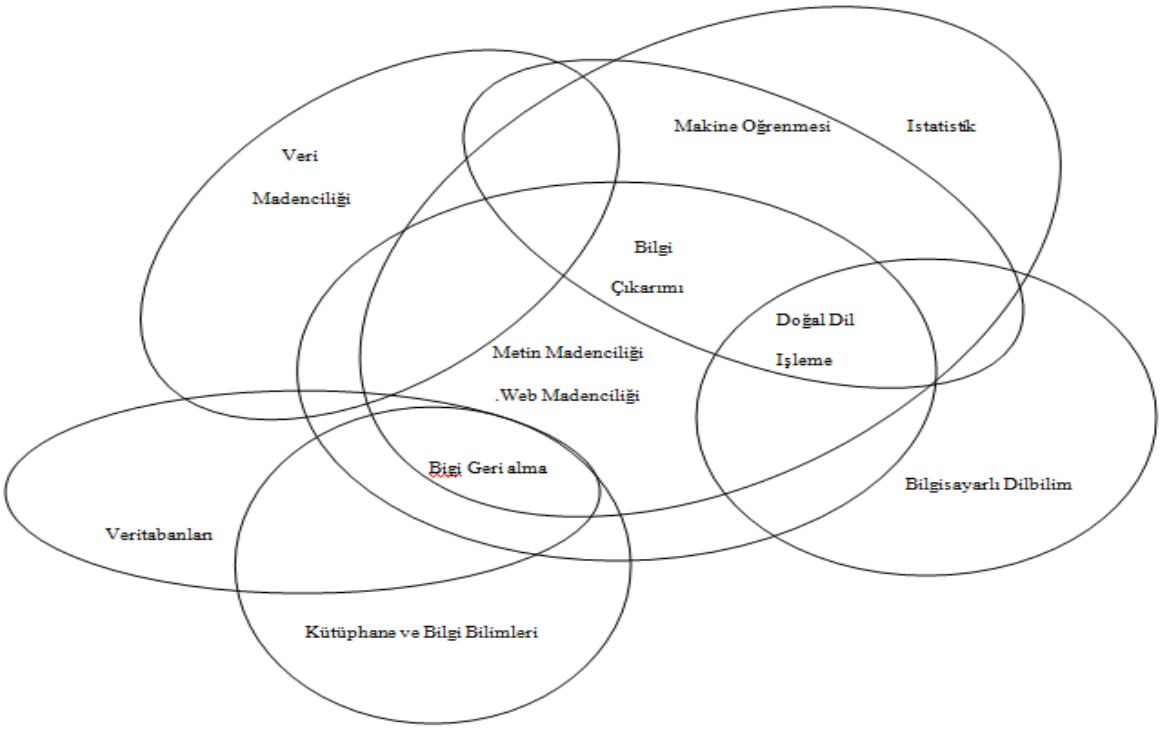
h) Belirsizlik

Yanlıslıkların siddeti ve verideki gürültünün derecesi ile ilgilidir. Veri tahmini bir kesif sisteminde önemli bir husustur (Albayrak,2008).

2.3. Metin Madenciliği

İnternet kullanımının yaşantımıza dahil olmasıyla ve dijital depolama aygıtlarının kapasitelerinin hızlı bir şekilde artmasıyla depolanan veri hacimleri çok büyük boyutlara ulaşmıştır. Bilgisayarlar hayatımıza girmeden önce dokümanlarla ilgili her türlü işlem elle yapılmaktayken bugün internette 2 milyardan fazla web sayfası olduğu göz önüne alınırsa bu bilgilere eski yöntemlerle ulaşmak imkansız denebilir (Döven,2013).

Verilerin boyutu günden güne katlanarak artmaktadır. Hemen hemen her tür kurum ve kuruluş verilerini elektronik olarak saklamaktadır. İnternet üzerinde bloglar, sosyal medya ağı ve elektronik postalar gibi dijital kütüphane, veri ambarı ve diğer metinsel bilgiler büyük veri yığınları akışı vardır. Bu büyük miktardaki veriden önemli bilgiler elde eden uygun kalıpları ve yaklaşımları belirlemek güçtür. Metin madenciliği metinsel veri kaynaklarından ilgi çekici ve önemli bilgiler keşfetmek için kullanılan bir süreçtir. Metin madenciliği bilgi çıkarımı, veri madenciliği, makine öğrenimi, istatistik ve bilgisayarlı dilbilime dayanan multi disiplinliner bir alandır. Şekil 8, metin madenciliğinin Venn diyagramını ve diğer alanlarla olan etkileşimini göstermektedir (Talib ve diğ.,2016).



Şekil 2.5. *Metin Madenciliği ve Diğer Alanlarla Etkileşimi* (Talib ve diğ. ,2016)

Günümüzde internet kullanımının yaygınlaşmasıyla birlikte bireylerin de sosyal mecralarda yaptıkları kişisel paylaşımlar oldukca fazladır. Bu durum da internet verilerini metin madenciliğinin merkezi haline getirmektedir (Zontul ve Aydın; 2017).

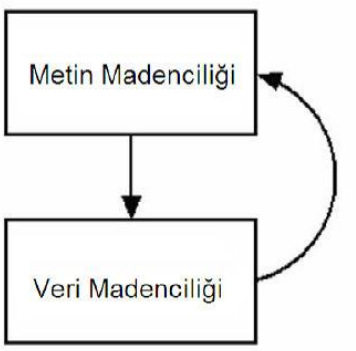
İnternet ortamında bulunan veriler büyük oranda yapısal değildir. Bu tür yapısal olmayan verilerin veri madenciliği ile işlenmesi mümkün değildir. Metin madenciliği, metin koleksiyonlarından bilgiye ulaşır, bireysel metinlerden bilgi çıkarır ve veri tabanlarından bilgi keşfeder. Metin madenciliği, veri madenciliğinin veya veritabanında bilgi keşfinin uzantısı olan farklı bir uygulama olarak görülebilir. Metinlerin işlenip anlamlı bilgiler sunabilmeleri için kullanılan bir uygulamadır. Metinlerin işlenmeye hazır hale gelmesi metin madenciliğiyle, işlenmesi ise veri madenciliği ile gerçekleştirilir (Karaca, 2012).

Metin madenciliği, veri madenciliği ile aynı analiz yaklaşımını ve tekniklerini kullanır. Ancak veri madenciliği yapılandırılmış veri gerektirirken metin madenciliği yapılandırılmamış verilerdeki kalıpları keşfetmeyi amaçlar. Metin madenciliğinde girdi, yapılandırılmamış veya yarı yapılandırılmış olan bir metin kümesidir. Örneğin bir metnin başlık, yazar, yayın tarihi ve kategorisi gibi birkaç yapılandırılmış bölümü olabilir. Fakat, yüksek bilgi değerine sahip içerik kısmı yapılandırılmamış bileşenlerden oluşur. Geleneksel veri madenciliği ile bu kısımlardan bilgi almak zordur (Dalmolen, 2010).

Metin madenciliğinin ilk adımlarından biri, metinleri sayısal temsillerine dönüştürmektir ki bu daha sonra veri setinde standart veri madenciliği yöntemlerinin kullanımına izin verir (Ayre, 2006).

Weiss ve diğerlerine göre (2005), Metin madenciliğini destekleyen ana temalardan biri, metnin sayısal verilere dönüştürülmesidir. Bu nedenle ilk veri farklı olsa da, bazı ara aşamalarda veriler klasik bir veri madenciliği işlemine taşınır. Yapılandırılmamış veriler yapılandırılır.

Yapısal veriler, bulundukları yapı içerisinde organize edilebilen ve tanımlanabilen veriler için kullanılır. En yaygın kullanılan yapısal veri kaynakları SQL (Structured Query Language) ve Access gibi veri tabanlarıdır. Bunlar sütun ve satır düzeyinde tanımlanıp verilere kayıt bazlı ulaşılabilir. Yapısal olmayan verilerin ise tanımlanabilir bir yapısı yoktur. En bilinen türleri; resim dosyaları, word, text dokümanları, web logları ve elektronik postalardır. Birçok kurumun verileri yapısal değildir. Metin madenciliği çok büyük kapasitedeki belgelerin analiz edilerek gizli yapıların elde edilmesi şeklinde tanımlanabilir (Dolgun ve diğ. 2009).



Sekil.2.6. *Süreçler arasındaki ilişki* (Çelikyay, 2010)

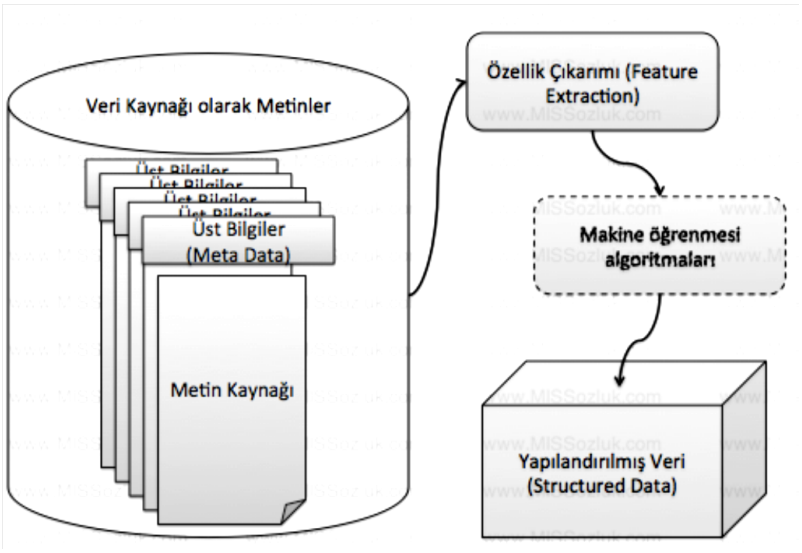
Şekil 2.6. ' da görüldüğü gibi, metin ve veri madenciliği arasında interaktif bir iliski vardır. Metin madenciliği süreçleriyle elde edilene yapısal veri, veri madenciliği modellerinde kullanılmakta ve elde edilen sonuclar daha sonra metnin yapısının incelenmesinde kullanılmaktadır (Çelikyay,2010).

Metinlerin işlenerek anlamlı bilgilere ulaşılması için, veri ön işleme ve özellik çıkarımı vb. olarak adlandırılan bazı aşamaların gerçekleştirilmesi gerekmektedir. Bu adımlardan sonra yapısal olmayan veriler, metin madenciliğinin kullanılacağı ve bilgisayarlar tarafından işlenen yapısal bir yapıya dönüştürülür. Böylece büyük yığınlar halinde bulunan veriler içerisinde gizli önemli bilgilere ulaşılmış olur. Oluşturulan değerli bilgiler kullanılarak, kurum ya da kuruluşların kullanabileceği farklı sonuçlara varılabilir. Metin

madenciliği yöntemlerinin temelinde matematiksel ve istatiksel yöntemler yer alır. Metin madenciliği, yazar tanıma, metin sınıflama, fikir madenciliği, duygu analizi, anahtar kelime çıkarımı, başlık çıkarımı gibi farklı alanlarda da sıklıkla kullanılmaktadır (Kılınç ve diğ. , 2016).

Metin madenciliği 1980' lerde ortaya çıkmasına rağmen, teknolojik ilerlemelerle gelişmiştir. Metin madenciliği isminin yanı sıra "metin veri madenciliği" , "metin analizi" , "kavram madenciliği" veya "web madenciliği" olarak da adlandırılır. Web madenciliği, yapılandırılmamış web içeriklerini yapılandırılmış içeriklere dönüştürerek sayfa kalıpları ve web istatistikleri gibi web siteleri hakkındaki verileri analiz eder. Metin madenciliği, veri madenciliğinin bir alt bölümüdür ve dijital ortamlarda dil, ses ve görsel olarak saklanan ve işlenmeye hazır halde bulunan yapılandırılmamış verilerle ilgilenir. Metin madenciliği ve veri madenciliği arasında etkileşimli bir ilişki vardır. Metin madenciliğinden elde edilen yapılandırılmış veriler, veri madenciliği modelleri kullanılarak değerlendirilir ve bulgular metinsel yapıyı analiz etmek için kullanılır (Ergün, 2017).

Metin madenciliği çalışmaları, metin kaynaklı alanyazında diğer bir çalışma alanı olan doğal dil işleme (natural language processing, NLP) çalışmaları ile pek çok kez birlikte yaplmaktadır. Doğal dil işleme çalışmaları daha çok yapay zeka altındaki dil bilim bilgisine dayalı çalışmaları içermektedir. Metin madenciliği çalışmaları ise daha çok istatistiksel olarak metin üzerinden sonuçlara ulaşmayı amaçlamaktadır. Metin madenciliği çalışmaları esnasında çoğu zaman doğal dil işleme kullanılarak özellik çıkarımı da yapılır. Genel olarak klasik bir metin madenciliği çalışması Şekil 2.7. ' de özetlendiği gibidir.



Şekil 2.7. *Genel olarak Metin Madenciliğinin Adımları*

2.3.1.Tarihsel Gelişimi

Metin madenciligi konusundaki ilk bulgular 1960' lı yıllarda ham metinlerin yer aldığı ilk bilgisayar sistemlerinin geliştirilmesiyle görülmeye başlamıştır. "Anahtar kelime ile arama" olgusuna odaklanan modellere 1980' lerin ortalarına kadar rastlanmamıştı. O dönemlerde yapılan metin madenciliği analizleri manuel olarak yapıldığı için çok zahmetli ve zaman alıcı olmuştur. Yapay zeka ailesinin bir öğesi olan Doğal Dil işleme sürecinden 1990' lardan sonra bahsedilmeye başlanması bu modellerin de görülmeye başlamasını sağlamıştır. Teknolojik gelişmeler ise aktif bir şekilde son yıllarda hızlı bir şekilde artmaya başlamıştır. Günümüzde kullanılan metin madenciliği yöntemlerindeki metodlar bu süreçte geliştirilen metodlardır (Bot, 2007; Melek, 2012).

Metin madenciliği; veriden bilgi çıkarma, yapılandırılmış veriler üzerinde veri madenciliği yöntemlerini uygulama, denetimli ve denetimsiz ögrenme, istatistiksel ve dilbilimi gibi konuları da kapsayan multidisipliner bir konudur. Veriler daha çok (%80 ve daha fazlası düşünülmektedir) metin şeklinde saklanmaktadır. Otomatik özetleme konusunda çok ilgi gören makalenin yazarı H.P. Luhn (1958), bahsi geçen makalede " önemli kelimelerin çözme gücü" ne değinmiştir.

Metin madenciliği ruhuna ve " bilginin dogal tanımlama ve örgütlemesinin frekanslar ve kütüphanedeki kelimelerin dağılımlarının analizinden gelebilecegine" değinen Lauren B. Doyle (1961) de ilgili metotlardan bahsetmiştir. (Burada kullanılan kütüphane kelimesi ana kısım veya toplanan bilgi kastedilmektedir.) Don R. Swanson (1988) ise bilimsel alanyazının "arastırma (exploration), korelasyon ve sentez" e değer bir olgu şeklinde düşünülmesinin önemini vurgulamıştır (Melek, 2012).

2.3.2.Veri Yapıları

Genellikle; toplanabilen, iletilebilen, sayılabilen ve işlenebilen gerçekleri, fikirleri ve kavramları tanımlayan veri her araştırma için oldukça önemlidir. Veri sözcüğü, Latince kökenlidir. Bilgiler, kullanıcıya sunulan anlamlı bilgilere dönüştürülür. Veriler yapılandırılmış veriler ve yapılandırılmamış veriler olarak iki şekilde kullanılır (Pasin, 2018).

i) Yapılandırılmış veri

Yapılandırılmış veri, biçimlendirilmiş bir ambarda, tipik bir veritabanı olarak düzenlenmiş verilerdir. Böylece veriler daha etkili işlem ve analizler için erişilebilir yapılabilir. Yapılandırılmış veriler, veri madenciliği araçlarıyla kolayca sıralanabilen ve işlenebilen, isimlendirilmiş sütun ve satırlarla gösterilebilen, genellikle metin dosyaları olan

bilgilerdir. Bu, her şeyin tanımlandığı, etiketlendiği ve erişiminin kolay olduğu mükemmel organize edilmiş bir dosya olarak görselleştirilebilir. Tablo 2.1. yapılandırılımış veriye örnektir (Pasin,2018).

Tablo 2.1. *Yapılandırılmış Veri Örneği* (Pasin, 2018)

Yıl	Nüfus	Yıllık Artış (%)
1990	56,473,653	2,29
2000	67,804,543	2
2007	70,586,256	0,58
2008	71,517,100	1,31
2009	72,561,312	1,48
2010	73,722,988	1,6
2011	74,724,269	1,35
2012	75,627,384	1,2
2013	76,667,864	1,37
2014	77,695,904	1,34
2015	78,741,053	1,34

ii)Yapılandırılmamış veri

Gerçek dünya verilerinin % 90' ı yapılandırılmamış verilerdir. Yapılandırılmamış veriler, elektronik tablo sayfaları, veritabanı tabloları veya diğer doğrusal veya sıralı veri kümeleri gibi daha az düzenli formu olan verileri ifade eder. Örneğin, e-posta, yapılandırılmamış metin verilerinin güzel bir örneğidir (Weiss ve diğ. ,2005; Pasin, 2018). En yaygın yapılandırılmamış veri türlerinden biri metinlerdir. Yapılandırılmamış metin, word belgeleri, PowerPoint sunumları, anket cevapları, bloglardan ve sosyal medya sitelerinden gönderiler dahil olmak üzere çok çeşitli biçimlerde oluşturulur ve toplanır. Diğer yapılandırılmamış veri türleri ise görüntüler, ses ve video dosyalarını içermektedir. Tıpkı metinler gibi, günlük hayatımızda sık sık resim, ses ve videolarla karşılaşmaktayız. İnternetin yaygınlaşması bu verilerle daha sık karşılaşılmasına sebep olmuştur (Pasin, 2008).

2.3.3.Metin Madenciliği ve Veri Madenciliğinin Karşılaştırılması

Metin madenciliğinin ve veri madenciliğinin hem benzer hem de farklı yönleri vardır. Önemli örnekleri ve eğilimleri işaret eden kuralları ve belli başlıklar ile ilgili kaydadeğer

özellikleri ortaya çıkarmada metinsel verileri kullanmaları benzer yönleridir. İkisi de çok büyük veri yığınları ile çalışır ve anlamlı bilgiler bulmayı amaçlar.

Veri madenciliği, esas olarak homojen ve evrensel olarak nitelendirilebilecek rakamları analiz etmektedir. Metin madenciliği ise metinsel belgeleri, elektronik postaları, sosyal medya gönderileri vb gibi heterojen veri türlerini işlemektedir. Veri madenciliğinin tersine, metin madenciliği yapılandırılmamış veya yarı yapılandırılmış verilerle çalışır. Değişik metinsel verilerden otomatik bir şekilde bilgi keşfi yaparak metinde daha önce fark edilmemiş bir bilginin bilgisayar tarafından ortaya çıkarılmasına Metin madenciligi denir. Metin madenciliğinin amacı metinsel veri yığınlarındaki eğilimleri fark etme ve daha önce hiç karşılaşılmamış bilgiyi keşfetmektir. Metin madenciliği bilgiyi yönetme, bilgiye erişme ve bilgi analizine esnek yaklaşımlar sunmaktadır. Temeli teknolojik gelişmelere, olasılık teorisine, yapay zeka ve istatistiğe dayanmaktadır(Gao ve diğ., 2005; Pasin, 2018; Consoli, 2010; Lau ve diğ. , 2005).

Aşağıdaki tabloda veri madenciliği ve metin madenciliğinin karşılaştırılması özetlenmiştir:

Tablo 2.2. *Veri Madenciliği ve Metin Madenciliğinin Karşılaştırılması* (Consoli, 2010; Pasin, 2018)

	Veri Madenciliği	**Metin Madenciliği**
Analiz nesnesi	Sayısal&Kategorik	Metinsel veri
Veri yapısı	Yapısal	Yapılandırılmamış veya yeri yapılandırılmış
Ortaya Çıkış	1994	2000
Amaç	Sınıflandırma ve tahmin	Bilgi çıkarımı

2.3.4. Metin Madenciliğinin Uygulama Alanları

Metin madenciliği, metinlerin bulunduğu her alanda kullanılabilir. Bu konuda Döven(2013), Karaca(2012) ve Çeliksu(2017) aşağıdaki alanları örnek olarak göstermişlerdir:

Müşteri ilişkileri yönetimi,

- Sahtekarlık tespiti,
- Sağlık alanı,
- Pazar araştırmaları,
- Metinlerden bilgi çıkarımı,
- Belge özetleme,
- Belge sınıflandırma,
- Benzer içerik belirleme,

- Web içerikleri sınıflama,
- Yazar tanıma ,
- Soru-cevap .

Pasin (2018), günümüzde metin madenciliğinin birçok alanda kullanılmakta olduğunu; gelecekte kullanım alanları daha da artacağını; bu alanlardan en yaygın kullanılanların veri madenciliği, istatistik ve doğal dil işleme süreci olduğunu vurgulamıştır.

2.3.5. Metin Madenciliğinde Kullanılan Teknikler

Bilgisayarlara bir metinin nasıl analiz edileceğini, anlaşılıp üretileceğini öğretmek için teknolojiler doğal dil işleme özelliği ile üretilir. Metin madenciliği sürecinde bilgi çıkarımı, özetleme, sınıflandırma, kümeleme ve görselleştirme gibi teknolojiler kullanılmaktadır (Gaikwad ve diğ., 2014; Pasin, 2018).

2.3.5.1. Bilgi Çıkarımı

Bilgi çıkarımı; varlıklar, varlıklar arasındaki ilişkiler ve varlıkları yapılandırılmamış kaynaklardan tanımlayan öznitelikler gibi yapılandırılmış bilgilerin otomatik olarak çıkarılmasını ifade eder. Bilgi çıkarımı, genellikle bir metin üzerinde doğal dil işleme sürecini kullanarak belirli kriterler hakkında bilgi edinmeyi amaçlamaktadır. Amaç, büyük miktarda veriyi otomatik olarak işleyen ve insan müdahalesini en aza indiren bir yazılım oluşturmaktır. Bilginin çıkarılabileceği alan genellikle yazılı metinlerdir, ancak bu metinlerin bulunduğu alan değişebilir. Örneğin veritabanları, internetteki belgeler veya taranmış metin bu verilerin kaynağını oluşturabilir (Pasin, 2018).

2.3.5.2. Özetleme

Metin özetleme, belirli bir metnin kullanıcı için yararlı bilgiler sağlayan kısaltılmış bir örneğini otomatik olarak oluşturma işlemidir. Büyük organizasyon veya şirketlerde, araştırmacıların tüm belgeyi okumak için zamanları yoktur, bu nedenle belgeyi özetler ve önemli noktalarını vurgularlar. Özet, bilgilerin önemli bölümlerini içeren, uzunluğu azaltan ve genel anlamı orijinal metindekiyle aynı olan bir veya daha fazla metinden üretilen bir metriktir (Gupta ve Lehal, 2009; Pasin, 2018).

2.3.5.3. Sınıflandırma

Sınıflandırma, metinlerin içeriklerine göre önceden tanımlanmış konulara göre ayrılmasıdır. Bir metin belsegi koleksiyonu, her bir belge için doğru konuyu veya konuları bulma sürecidir. Naive Bayes olasılık yöntemi, k en yakın komşu algoritması, karar ağaçları,

yapay sinir ağları ve genetik algoritmalar gibi istatistiksel sınıflandırma teknikleri, metni kategorize etmek için kullanılabilir.

Yüksek gelirli, orta gelirli ve düşük gelirli olmak üzere üç hedef grup sınıflandırmaya örnek olarak verilebilir. Girdi veya tahmin değişken kümesinde olduğu gibi Veri madenciliği modeli de hedef değişkenler ile ilgili bilgi veren büyük veri kümelerini analiz eder (Lorese, 2005; Dang ve Ahmad, 2014; Pasin, 2018).

2.3.5.4. Kümeleme

Kümeleme, bir araştırmada çalışılan birimleri benzerliklerine göre belirli gruplar içinde toplayarak sınıflandırmaya, birimlerin ortak özelliklerini ortaya çıkarmaya ve bu sınıflar hakkında genel tanımlamalar yapmaya olanak sağlar. Kümeleme analizinin hedefi, gruplanmamış verileri benzer özelliklerine göre gruplamak ve araştırmacının yararlı, faydalı özetlenmiş bilgiler elde etmesine yardımcı olmaktır.

Sınıflandırma tekniğinde sınıflar önceden tanımlanmıştır, ancak kümeleme tekniğinde veriler belirli bir sınıf olmadan analiz edilir. Genel olarak, sınıflar bilinmedikleri için başlangıçta hazır değillerdir. Kümelemede analiz yapılırken sınıf sayısı artırılabilir. Veriler, sınıflar içindeki benzerliği en üst düzeye çıkararak sınıflandırılır veya gruplandırılır (Dinler, 2014; Pasin, 2018).

2.3.5.5. Görselleştirme

İnsan beyni görsel bilgileri metinsel bilgileri işlemekten daha iyi işler, bu nedenle çizelgeleri, grafikleri ve tasarım öğelerini kullanarak veri görselleştirme, eğilimleri ve istatistikleri çok daha kolay açıklamanıza yardımcı olabilir. Görselleştirme, bir nesnenin, sahnenin, kişinin veya görsel algıya benzer bir soyut kavramın zihinsel bir görüntüsü veya görsel bir temsilidir (Ergün, 2017; Pasin, 2018).

Görselleştirmenin birçok tanımı vardır, ancak literatürde en çok atıfta bulunulan, "bilişi güçlendirmek için verilerin bilgisayar destekli, etkileşimli, görsel temsillerinin kullanılması" dır, burada bilişten kastedilen insan algısının gücü veya basit bir deyişle bilginin edinimi veya kullanımıdır (Teyseyre ve Campo, 2009; Pasin, 2018).

2.3.6. Metin Madenciliği İle İlgili Yazılımlar

SAS,R programlama dili, ODM, RapidMiner, STATISTICA ve SPSS mevcut bazı metin madenciligi yazılımlarıdır (Melek, 2012). En çok tercih edilen yazılımlar ise R, RapidMiner ve SAS ' dır.

2.3.7. Metin Madenciliğinin Uygulama Adımları

Veri madenciliğinde olduğu gibi, metin madenciliğinde de verilerden anlamlı bilgiler elde edebilmek için verilerin çeşitli aşamalardan geçirilmesi gerekmektedir. Verilerin doğru analiz edilmesi, algoritmaların doğru bir şekilde kullanabilecek hale getirilmesi, doğru sonuçlar elde edilebilmesi için çok önemlidir (Döven, 2013).

Metin madenciliği, anlamlı analiz araçları kullanılarak belirli bir sürede elde edilen metinleri anlamlı bilgilere dönüştüren bir süreçtir. İlk aşamada, metinler belirli veri tabanlarından veya arama motorlarından toplanır. Ortaya çıkan metin, anlamsız kelimelerden arındırılması için ön işlemden geçirilir ve son olarak veri madenciliği teknikleri kullanılarak elde edilen sonuçlar değerlendirilir (Pasin, 2018).

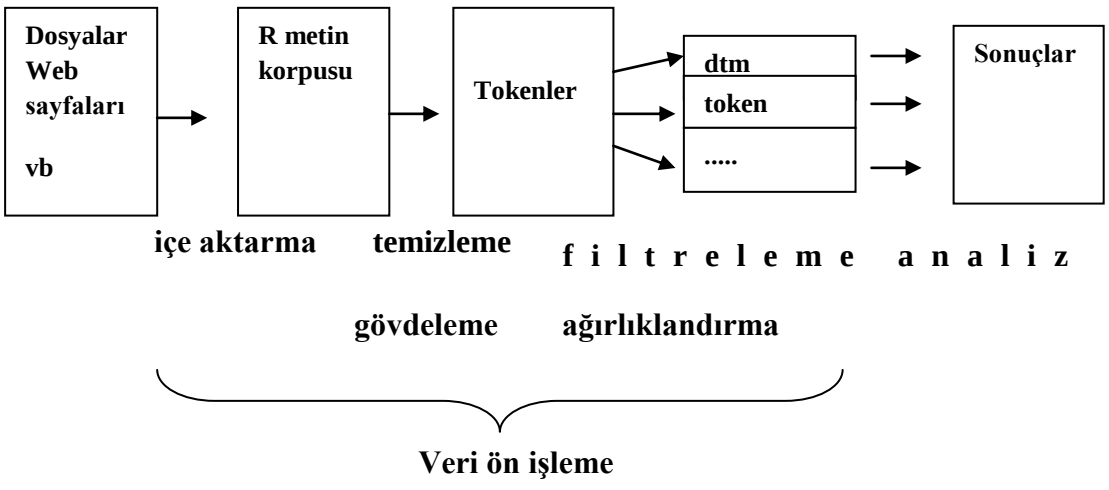
Metin madenciliğinde metin analizi için R programının kullanılması pek çok avantaj sağlamaktadır. R programı, ücretsiz, açık kaynaklı ve platformlar arası bir programdır. Çoğu programlama dilinin aksine R, özellikle istatistiksel analizler için tasarlanmıştır. Bu durum onu veri bilimi uygulamaları için uygun kılmıştır. R'ın hızla talep görmesinin sebeplerinden biri R terminolojisinde "paket" ismiyle kullanılan geniş yazılım kütüphanelerinin mevcut olmasıdır. Her paket, temel R dilinin ve çekirdek paketlerinin işlevselliğini genişletir, fonksiyonlara ve verilere ek olarak, genellikle paketin kullanımını gösteren örnek formunda formüller içermektedir. R programı ile metin analizi yapmanın en temel avantajlarından biri, farklı paketler arasında geçiş yapmanın veya bunları bir araya getirmenin genellikle mümkün ve nispeten kolay olmasıdır. Sonuç olarak, R' daki metin analizi için temel bilgileri öğrenmek, çok çeşitli gelişmiş metin analizi özelliklerine erişim sağlar (Welbers ve diğ. ,2017). Tablo 2.3. ' te R programında yer alan paketlerin isimleri, işlevleri ve alternatif paketler gösterilmektedir:

Tablo 2. 3. *R Programında Yer Alan Metin Analiz Paketleri Ve Fonksiyonları*

İşlem	**Örnek**	**Alternatifler**
Veri hazırlama		
Metni içeri aktarma	readtext	XML, readxl, pdftools,jsonlite, antiword
Dizi işlemleri	stringi	stringr
Ön işlem	quanteda	snowballC, tm, stringi, tokenizers,
Döküman-Terim matrisi		
Filtreleme ve Ağırlıklandırma		tm, tidytext, Matrix
Analiz		
Sözlük		tidytext, koRpus ,tm,corpustools
Denetimli Makine Öğrenmesi		kerasR ,RTextTools , austin
Metin İstatistikleri		koRpus, corpustools, textreuse
Denetimsiz Makine Öğrenimi	topicmodels	quanteda, text2vec, , austin, stm
İleri düzey konular		
İleri düzey NLP	spacyr	coreNLP, cleanNLP, koRpus
Kelime konumları ve Sözdizimleri	corpustools	quanteda, tidytext, koRpus

R programı ile metin madenciliği genel anlamda 2 temel adımda gerçekleştirilir. Bunlar veri ön işleme ve analiz süreçleridir. Veri ön işleme süreci içe aktarma, temizleme, gövdeleme, filtreleme ve ağırlıklandırma olmak üzere genel olarak 5 adımda gerçekleştirilmektedir. Bu süreç aşağıdaki tabloda gösterilmiştir:

Tablo 2.4. *R programında metin analizi adımları*



2.3.7.1. Veri Ön İşleme

Veri hazırlama, herhangi bir veri analizinin başlangıç noktasıdır. Yapılandırılmamış veya yarı yapılandırılmış veri setinden anlamlı sonuçlar çıkararak analiz aşamasına gelinceye kadarki süreci kapsayan veri ön işleme süreci beş aşamada gerçekleştirilmektedir (Welbers ve diğ. ,2017).

1) Metni İçe Aktarma

R programında yapılandırılmamış veya yarı yapılandırılmış bir metin okutmak , herhangi bir R tabanlı metin analizinde ilk adımdır. Metinsel veriler çok çeşitli dosya formatlarında oluşturulabilir. R programı, .csv ve .txt uzantıları gibi metin dosyalarının okunmasını destekler ancak .json,. html ve . xml gibi biçimlendirilmiş metin dosyalarını işlemek için ve Word, Excel, PDF gibi karmaşık dosya formatlarını okumak için ek paketler gerekir. Bu farklı paketlerle farklı arayüzleri ve çıktıları çalışmak , özellikle aynı projede farklı dosya formatları bir arada kullanılıyorsa zor olabilir. Böyle sorunlar için uygun çözüm yolu **readtext** paketidir. Bu paket birçok veri türünü tek bir biçimde içe aktarmak için çeşitli içe aktarma paketlerini bir araya getirir. Çeşitli dosya formatlarından (örneğin, txt, csv, pdf) oluşan metinleri R' daki ham metin korpusuna okuma fonksiyonlarını gerçekleştirir. Korpus, farklı uzantılı dosya türlerinin işlenmek üzere R programına aktarıldıktan sonra toplandığı alana verilen addır.

Aşağıda verilen kod, internet kaynaklı .csv uzantılı bir dosyanın metin okuma fonksiyonuyla nasıl içe aktarılacağını göstermektedir (Ooms, 2014, Lang ve the CRAN Team, 2017,Ooms, 2017a, Wickham ve Bryan, 2017, Ooms, 2017b, Welbers ve diğ. ,2017).

```
library(readtext)
<-
"https://raw.githubusercontent.com/kbenoit/readtext/master/inst/extdata/csv/inaugCorpus.csv"

rt <- readtext(filepath, text_field = "texts")
```

Aynı fonksiyon, yukarıda belirtilen tüm dosya biçimlerini içe aktarmak için kullanılabilir ve dosya yolu, bir zip klasörü de olabilir. Pek çok durumda, belirtilmesi gereken tek şey, metinleri içeren alanın adıdır.

2) Temizleme

Metin analizi ile dijital metinler de işlenebilir. Dijital metin, dize adı verilen bir karakter sıralaması olarak gösterilir. R' da dizeler, dizelerin vektörleri olan "karakter" türleri adı verilen nesneler olarak gösterilir. En yaygın dize işlemleri, dizelerin bölümlerini birleştirmek, bölmek ve çıkarmak (topluca çözümleme(parsing) olarak adlandırılır) ve kalıpları bulmak veya değiştirmek için normal ifadelerin kullanılmasıdır.

R programında, bu tür metinlerle çalışmak için **stringi** paketi tercih edilir (Gagolewski, 2017). Çünkü stringi, noktalama işaretlerinin doğru kullanımı gibi konularda dil desteği için Uluslararası Bileşenler (ICU) kütüphanesini kullanmaktadır. Alternatif olarak, stringi paketini arka uç olarak kullanan, ancak daha basit bir sözdizimine sahip olan **stringr** paketidir.

R' daki çoğu fonksiyonda olduğu gibi, dizi işlemleri vektörleştirilir, yani bir vektörün her bir elemanına uygulanırlar. Vektör, metin dokümanlarının ifade edildiği bir alandır. Aradığınız bir metnin bu alandaki yerini bulabilirseniz ve bu vektöre en yakın dokümanlara erişebilirsiniz. Dokümanlar arası uzaklığı birbiri ile kıyaslayabilirsiniz (Gagolewski, 2017, Welbers ve diğ. , 2017).

stringi paketiyle işaretleme imleri, konuyla ilgisi olmayan aralıklar ve alfabe dışı karakterler kaldırılarak bazı temel temizleme fonksiyonlarını gerçekleştirilir.

R programındaki pek çok fonksiyonda olduğu gibi, stringi işlemleri vektörleştirilmiştir;yani bir vektörün her bir öğesi için gerçekleştirirlir. Dizelerin vektörlerinin manipülasyonu, R programında önerilen bir yaklaşımdır, çünkü her bir eleman üzerinde tek tek değişiklik yapmak ve onu R' da işlemek oldukça zordur (Welbers ve diğ. ,2017).

3) Gövdeleme, Normalleştirme ve Durak Kelimelerini Ayıklama (Tokenization, Lowercasing and Stemming, Removing Stopwords)

Çoğu hesaplamalı metin analizi yöntemi için,metnin tamamı, sözcükler veya bileşik sözcük gibi daha küçük, daha özel metin özelliklerine bölünmelidir. Ayrıca, birçok metin analiz tekniğinin hesaplama performansı ve doğruluğu, sözcük ve bileşik sözcükler normalleştirilerek veya "durak kelimeleri" kaldırılarak iyileştirilebilir. Durak kelimeleri edat, bağlaç vb olduğu önceden belirlenmiş ve bu nedenle analizden önce atılan sözcüklere denir.

Gövdeleme, çekimli fiilleri yalın hallerine dönüştüren bir algoritmadır. **Kökenine Döndürme (Stemming)** yöntemi sözcükleri basit hallerine çevirir. İsimlerden çoğul eklerin

atılması, çekim eklerinin fiillerden arındırılarak kök haline döndürülmesi gibi işlemler kökenine döndürme olarak adlandırılır.

Parçalara ayırma(parsing), metin analizi için çok önemlidir; çünkü metinler, anlamlı hesaplamalar yapmak için çok özgüldür.

Bir cümleyi sözcüklere bölmek için "quanteda paketi" kullanılır. Parçalara ayırma fonksiyonu, her bir metnin kelimelerini karakter vektörü olarak içeren bir liste verir.

Normalleştirme süreci; küçük harfe çevirme ve gövdeleme genel olarak kelimelerin daha tekdüze bir formata dönüştürülmesini ifade eder. Belirli bir analiz için, bir bilgisayarın iki sözcüğün biraz farklı bir şekilde yazılsa bile aşağı yukarı aynı anlama geldiğini tanıması gerekiyorsa, bu önemli olabilir. Diğer bir avantajı, kelime boyutunu yani analizde kullanılan tüm özellikleri küçültmesidir. Basit ancak önemli bir normalleştirme tekniği tüm metni küçük harflere çevirmektir.

quenteda paketinde **char_tolower** karakterleri küçük harflere çevirmek için ya da **tokens_wordstem** dizgeleri köklere ayırmak için kullanılır. Böylece büyük harf -küçük harf, tekil ve çoğul formlar arasındaki farklar kaldırılmış olur.

durak kelimeler/etkisiz kelime/durma kelimeleri/durdurma kelimeleri kaldırma, edat/ bağlaç gibi bir metnin içeriği hakkında nadiren bilgi veren kelimelerin filtrelenmesi, verilerin boyutunu azaltma, hesaplama yükünü azaltma ve bazı durumlarda doğruluğu arttırma fonksiyonunu sağlar. Bu sözcükleri kaldırmak için önceden tanımlanmış "durak kelime" listeleriyle bu sözcükler eşleştirilir ve silinir**. quenteda** paketinde stopwords fonksiyonu, belirli bir dilin karakterlerini durak kelimelerine dönüştürür.

Ön işleme adımlarının doğru sırada yapılmasına özen gösterilmelidir. Örneğin köklere ayırmadan önce durak kelimeleri kaldırılmalıdır. Aksi taktirde örneğin "dolayı" sözcüğü "dol" içine köklenecek ve "dolayı" kelimesiyle eşleştirilmeyecektir. quanteda paketinde kullanılan stopword eşleşmesi büyük / küçük harfe duyarlı olmasa da küçük harf dönüştürmede de sıralama dikkate alınmalıdır (Benoit ve diğ. , 2017; Welbers ve diğ. , 2017).

4)Döküman Terim Matrisi Oluşturma (DTM)

Yukarıda bahsedilen ön işleme tekniklerinin tümü, döküman terim matrisi oluştururken tek bir fonksiyonla uygulanabilir. Döküman terim matrisi (DTM), bir metin korpusunu (örn:bir metin koleksiyonu) temsil etmek için kullanılan en yaygın biçimlerden biridir. DTM, satırların belgeleri, sütunların terimleri ve hücrelerin her bir terimin her belgede

ne sıklıkta gerçekleştiğini gösterdiği bir matristir. Bu gösterimin avantajı, metin halinden rakamlara etkili bir şekilde dönüşerek verilerin vektör ve matris cebiri ile analiz edilmesine izin vermesidir.

R programında DTM sınıfları sağlayan en gelişmiş metin analiz paketlerinden ikisi **tm** ve **quanteda**' dır. İkisi arasında da , **tm** paketi daha yaygın kullanılır.

Dikkate değer başka bir alternatif **tidytext** paketidir. Bu paket döküman terim matrisi kullanmak yerine , aynı verileri , DTM' nin her bir (sıfır olmayan) değerinin , sütun belgesi, terim ve sayım ile bir satırda bulunduğu uzun bir formatta gösterir (Welbers ve diğ. , 2017).

5)Filtreleme ve Ağırlıklandırma

Bir korpustaki tüm kelimeler metin analizi için eşit oranda bilgi verici değildir. Bununla başa çıkmanın bir yolu döküman terim matrisinden bu kelimeleri çıkarmaktır. Önceki bölümlerde durak kelimelerini filtrelemekten bahsetmiştik; bunun yanı sıra durak kelime listesinde yer almayan fakat çok sık rastlanan, çok tekrar edilen ve korpular arasında farklılık gösteren kelimeler vardır. Ayrıca kategori tahmini ve konu belirleme için çok nadir rastlanan kelimeler de vardır. Bu, özellikle verimliliği artırmak için çok işe yaramaktadır. Çünkü kelime dağarcığının boyutunu (yani nadir rastlanan terimlerin sayısını) büyük ölçüde azaltır; aynı zamanda doğruluğu da artırır. Basit ama etkili bir yöntem de , minimum ve maksimum belge sayısı (veya oranı) için bir eşik kullanarak belge sıklıklarını (bir terimin geçtiği belge sayısı) filtrelemektir (Yang ve Pedersen, 1997; Griffiths ve Steyvers, 2004; Welbers ve diğ. , 2017).

Korpusta az bilgi verici terimleri kaldırmak yerine, alternatif bir yaklaşımla onlara değişken ağırlıklar atanır. Birçok metin analizi tekniği, terimlerin kullanılma sıklıklarını kullanmak yerine, tahmini bir bilgi değerini hesaba katacak şekilde ağırlıklandırıldığında daha iyi performans gösterir. Yeterince büyük bir korpus verildiğinde, bu bilgi değerini tahmin etmek için korpus'ta terimlerin dağılımı hakkında konuşabiliriz. Bu fonksiyonu gerçekleştiren popüler bir ağırlıklandırma fonksiyonu, corpustaki kelimeleri ağırlıklandıran **frekans-ters belge frekansı (tf-idf)** dır. Doküman frekans eşiği kullanma ve ağırlıklandırma, bir DTM üzerinde kolayca gerçekleştirilebilir. quanteda, sırasıyla belge sıklığı için **docfreq**, terim sıklığı için **tf** ve tf-idf için **tfidf** fonksiyonlarını içerir. Her fonksiyon, SMART ağırlıklandırma şemasını uygulamak için çok sayıda seçeneğe sahiptir. Bunlardan üst düzey bir paket olan quanteda**, dfm_weight** fonksiyonunu da sağlar (Manning ve diğ. , 2008; Welbers ve diğ. , 2017).

2.3.7.2.Analiz

Metin madenciliği analizi kapsamında Boumans and Trilling (2016) tarafından önerilen üç sınıflama vardır:

-sayma ve sözlük yöntemleri (Counting and Dictionary)

-denetimli makine öğrenmesi (Supervised machine learning)

- denetimsiz makine öğrenmesi (Unsupervised machine learning)

Bu yaklaşımlar tümden gelimden tüme varıma doğru sıralanmıştır. Tümdengelim, bu senaryoda, önceden tanımlanmış bir kodlama şemasının kullanımı anlamına gelir. Başka bir deyişle, araştırmacılar ne aradıklarını önceden biliyorlar ve sadece bu analizi otomatikleştirmeye çalışıyorlar. Tümdengelimli akıl yürütme kavramıyla olan ilişki, araştırmacının belirli kuralların veya öncüllerin doğru olduğunu varsaydığı (örneğin, olumlu duyguları belirten kelimelerin bir listesi) olduğu ve bu nedenle metinlerle ilgili sonuçlar çıkarmak için uygulanabileceğidir. Tümevarımda ise öncül bir kodlama şeması kullanmak yerine, bilgisayar algoritmasının kendisi bir şekilde metinlerden anlamlı kodlar çıkarır. Örneğin, en azından matematiksel olarak kelimelerin bir arada ortaya çıkışında kalıpları aramak ve bu kalıpları açıklayan gizli faktörleri (örneğin konular, çerçeveler, yazarlar) bulmaktır. Tümevarımsal akıl yürütme açısından, algoritmanın belirli gözlemlere dayalı geniş genellemeler oluşturduğu söylenebilir (Welbers ve diğ. ,2017).

-Sayma ve sözlük yöntemleri(Counting and dictionary)

Sözlük yaklaşımı genel olarak, basit anahtar kelimelerden karmaşık ifadelere kadar, belirli kavramların metinlerde ne sıklıkta gerçekleştiğini saymak anlamına gelir. Bu tümdengelimli bir yaklaşımdır, çünkü sözlük hangi kodların ne şekilde ve nasıl ölçüldüğünü belirten bir öncül tanımlar ve bu verilerden etkilenmez. Sözlük kullanmak sayısal olarak basit ama güçlü bir yaklaşımdır.

Bir sözlüğü quenteda paketinde yer alan DTM' ye uygulayabilmek için ilk adım, sözlük fonksiyonunu kullanarak bir sözlük nesnesi oluşturmaktır. **Dfm_lookup** fonksiyonuyla, sözlük nesnesi, sütunların sözlük kodlarını temsil ettiği yeni bir DTM oluşturmak için bir DTM' ye uygulanabilir (Welbers ve diğ. , 2017).

-denetimli makine öğrenmesi (Supervised machine learning)

Denetimli makine öğrenmesi yaklaşımı, bir algoritmanın bir veri üzerinden gizli kalmış yapıları öğrendiği tüm sınıflandırma tekniklerini içerir. Genellikle, bu algoritmalar,

nasıl kodlama yapması gerektiğine dair yeterli örnekler verilirse, metinleri nasıl kodlayacağını öğrenebilir. Tümdengelim ve tümevarım kısımları vardır. Tümdengelim kısmı, araştırmacıların, tahmin etmeye veya ölçmeye çalıştığı kategorileri temsil eden eğitim verilerini sağlamasıdır. Bununla birlikte, araştırmacılar bu kodlara nasıl bakılacağı konusunda açık kurallar sağlamamaktadır. Tümevarım bölüm ise, denetlenen makine öğrenme algoritmasının bu kuralları eğitim verilerinden öğrenmesidir.

- denetimsiz makine öğrenmesi (Unsupervised machine learning)

Denetimsiz makine öğrenmesi yaklaşımlarında kodlama kuralları belirtilmez ve eğitim verisi kullanılmaz. Bunun yerine, bir algoritma metindeki belirli kalıpları tanımlayarak bir model ortaya çıkarır. Araştırmacının tek etkisi, belgelerin sınıflandırıldığı kategori sayısı gibi belirli parametreleri belirlemesidir.

Grimmer ve Stewart (2013) denetimli ve denetimsiz makine öğreniminin rakip yöntemler olmadığını, ancak farklı amaçları yerine getirdiklerini ve birbirlerini tamamlamada çok iyi kullanılabileceğini vurgulamışlardır. Belgelerin önceden belirlenmiş kategorilere yerleştirilmesi gerekiyorsa, denetimli yöntemler en uygun yaklaşımdır. Çünkü denetimsiz bir yöntemin, bu kategorileri yansıtan ve araştırmacının bunları nasıl yorumladığı ile ilgili bir sınıflandırma getirmesi olası değildir. Denetimsiz yöntemlerin avantajı, araştırmacıların dikkate almadığı kategorileri ortaya çıkarmasıdır.

2.3.8. Konu İle İlgili Yapılan Araştırmalar

Çepni (2014), "Metin madenciliği ve bir kompozisyon modelini "Taklit etme" stratejilerinin ikinci dilde yazılan öğrenci kompozisyonlarındaki kelime zenginliği, çeşitliliği ve öğrenci başarısı üzerine olan etkisi" isimli tezi ülkemizde eğitim alanında metin madenciliği konusunda yapılan tek çalışmadır. Çepni araştırmasında öğrencileri deney ve kontrol grubuna ayırarak çeşitli eğitimlerden geçirmiş ve sonrasında kompozisyon yazmalarını istemiştir. Komposizyonlar türkçe karakterlerden arındırılarak Kuzey Arizona Üniversitesi Corpus Lab'daki Biber (1993) etiketleme programı aracılığıyla etiketlenmek üzere ABD' ye gönderildi. Yazılan kompozisyonları kelime zenginliği açısından kıyaslamıştır.

Pasin (2018), "Türkçe metinler üzerinde metin madenciliği yöntemlerinin incelenmesi" isimli tez çalışmasında metin halindeki Türkçe verilerin sınıflandırılmasını amaçlamıştır. Türkçe köşe yazılarından oluşturduğu veri kümesini “Cinsiyet Tanımlama”, “Yazar Tanımlama” ve “Tür Belirleme” olmak üzere üç kategoride incelenmiştir. Sınıflandırma yaparken Naive Bayes metot ve bit skor ağırlıklandırılmış k-NN metotlarını

kullanmıştır. İki metodun doğruluk oranları karşılaştırılmış ve Naive Bayes metodunun daha doğru sonuçlar verdiğini gözlenlemiştir. Sınıflandırma için R programlama dili kullanılmıştır.

Taha (2011) "Metin madenciliği ile doküman demetleme" konulu tezinde bölünmeli kümeleme tekniklerini kullanarak İngilizce ve Türkçe metinlerde yer alan verileri belirli başlıklar altında kümelemiş ve gerekli bilgiyi elde etmek istemiştir. Çalışmasında metinlerin tamamı Terim Frekansı – Ters Doküman Frekansı (TF-IDF) vektörleri ile anlatılmış; metin madenciliği konusunda ise, geleneksel bilgiye ulaşma çalışmalarının eksik yönlerini gideren Latin Semantic Index (LSI) yöntemini kullanmıştır. Çalışmasında TF, TF-IDF ve LSI kullandığında K-Means ve K-Median algoritmalarının başarılarını karşılaştırmış ve K-Means algoritmasının kümeleme başarısının K-Median algoritmasından daha iyi çıktığı sonucuna varmıştır. Veri seti olarak Milliyet gazetesi veri seti ve literaturde sıklıkla kullanılan R8 ve WebKB-4 veri setlerini kullanmıştır. Çalışmasını Microsoft. Net ortamında C# dili kullanarak gerçekleştirmiştir.

Karaca (2012), " Metin madenciliği yöntemi ile haber sitelerindeki köşe yazılarının sınıflandırılması" tezinde, metin madenciliği yöntemi ile haber sitelerindeki köşe yazılarını sınıflandırmıştır. Veri madenciliği ve metin madenciliği konularını alt başlıkları ile vererek bir uygulama yazılımı geliştirmiştir. Geliştirdiği yazılımda eğitim ve test dokümanlarının alınmasından sınıflandırılmasına kadar olan bütün işlemleri gerçekleştirmiştir. Köşe yazılarındaki kelime köklerinin bulunması için ayrıca bir yazılım geliştirilmiştir ve 6 farklı gazeteden 25 yazar ile sistem eğitilmiştir.

Varol (2011) , "Metin madenciliği yöntemlerini kullanarak türkçe dökümanlarda tür ve yazar tanıma" tez çalışmasında yedi şairin şiirinin bulunduğu iki yüz on adet şiirden oluşan bir eğitim veri seti kullanmıştır. Şair tanıma problemi için iki yöntem izlenmiş; ilkinde eğitim ve test şairlerine ait her bir şiirin istatistiksel özellikleri, kelime zenginliğine bağlı özellikleri, dilbilgisi özellikleri, karakter n-gramları gibi bazı özellik vektörlerini çıkarmıştır. Bu vektörleri WEKA programında yer alan çeşitli sınıflandırma algoritmalarıyla işleyerek şair belirleme çalışması yapmıştır. İkinci kullanılan yöntemde Ng-İnd sınıflandırma yöntemini uygulamış ve bu iki yöntemden elde edilen sonuçlar sınıflandırma performansları açısından karşılaştırılmıştır.

Çelikyay (2010), "Metin madenciliği yöntemiyle Türkçe'de en sık kullanılan ve birbirini takip eden harflerin analizi ve birliktelik kuralları" adlı tezinde kullandığı metinlerin tamamını internetten elde etmiştir; gerekçesini de doğal bir sekilde yazılmıs metinlerin internet kaynaklarından daha kolay ulaşılabilecek başka bir kaynağın olmaması şeklinde açıklamıştır. Önce Türkçe metinlerde en sık kullanılan harfler, sonrasında Türkçe metinlerde

iki harfin birbirini takip etme sıklığını en son da Türkçe metinlerde üç harfin birbirini takip etme sıklığını incelenmiştir.

Çeliksu (2017), "Yabancı dizilerin altyazı ve twıtter yorumlarının metin madenciliği ile incelenmesi" konulu tezinde, yabancı dizilerin Türkçe altyazı ve twitter yorumlarını açık kaynaklı R programı ile metin madenciliği açısından incelemeyi amaçlamıştır. Analiz aşamasında, yapısal olmayan dizi altyazılarını ve aksiyon dizi türüne ait twitter yorumlarını metin madenciliği yöntemleriyle yapısal hale getirmiştir. Dizi altyazılarını ITU Türkçe Doğal Dil İşleme Yazılım Zinciri sistemini kullanarak rakamlardan, noktalama işaretlerinden, beyaz boşluklardan ve linklerden arındırarak kelime köklerine ayırmıştır. Eklerden arındırılan kökler R' da analiz sürecine sokularak kelimelerin koordinat sistemindeki konumları belirlenmiştir. Uzaklık ölçüsü olarak Öklit Uzaklığı kullanılmış ve k-means kümeleme algoritması kullanılarak altyazılar kendi türünde anlamlı kümelere ayrılmıştır.

Yıldız(2016), "Metin madenciliğinde anahtar kelime seçimi bir üniversite örneği " makalesinde, bir üniversitede kullanılan kurum ile ilgili şikayet, teşekkür, görüş ve öneri mesajlarının yazılabildiği ve bu mesajlara ilgili kurum tarafından cevap verilebildiği bir bilişim sistemine ait verileri kullanılmıştır. Veri seti yaklaşık 3961 mesajdan oluşmaktadır ve bu mesajlar metin madenciliği teknikleri kullanılarak ön işlemden geçirilmiştir. Ön işlem sonrasında elde edilen metinlerin içindeki önemli kelimeleri tespit etmek için tf-idf ve ki-kare istatistik algoritmasını kullanarak anahtar kelime seçimi yapmıştır.

Göker ve Tekedere (2017), "Fatih projesine yönelik görüşlerin metin madenciliği yöntemleri ile otomatik değerlendirilmesi " adında bir makale yayınlamıştır. Makalede FATİH projesine yönelik internet ortamında yer alan görüşler hazırladıkları yazılım ile önce metin madenciliği yöntemleri kullanılarak analize hazır hale getirilmiş; daha sonra metin madenciliğinde ve literatürde en çok kullanılan makine öğrenmesi algoritmalarından Naive Bayes, K-En yakın komşu (k-NN, IBk), Karar Ağaçları (J48), SMO ve RBF Network algoritmaları uygulanarak oluşturulan modellerin başarım ölçütleri karşılaştırılmıştır. Karşılaştırılan sınıflandırma algoritmalarının başarı yüzdelerinin %80 ve üzerinde olduğu tespit edilmiştir.

Abuzir (2018), " Metin madenciliğine dayalı öğrenci projesi değerlendirmede yenilikçi model" isimli çalışmasında, proje yönetimi ve değerlendirme dersinde öğrencilerin ilerlemesinin değerlendirilmesi için metin madenciliği tekniklerinin kullanılmasını önermektedir. Bu kapsamda öğrencilere 3 aşamalı bir proje hazırlama planı oluşturup her aşama sonunda öğrencilerden teslim alınan projeleri kelime zenginliği, koşullu ve koşulsuz ilişkiler ve kullanılan terimlerin yapıları bağlamında metin madenciliği yöntemiyle değerlendirmiştir. Bu çalışma sayesinde kompozisyon türü yazılı sınavlarda karşılaşılan

subjektif değerlendirmelerin ortadan kaldırılabileceği aynı zamanda öğretmenlerin yüklerinin hafifletileceği savunulmaktadır.

W.He (2013), " Veri madenciliği ve metin madenciliği kullanarak öğrencilerin çevrimiçi etkileşimlerini inceleme" adlı bir çalışma yapmıştır. Bu çalışmada veri ve metin madenciliği yöntemlerinin eğitim kurumlarına öğrencilerin öğrenme davranışlarının keşfedilmesi, görselleştirilmesi ve analiz edilmesini sağladığını savunmaktadır. Bu amaçla (live video streaming (LVS) learning environment) canlı öğrenme ortamlarından elde edilen iki faklı veri kümesi karşılaştırılmıştır. Bu iki farklı veri kümesi öğrencilerin canlı ortamdaki öğrenci-öğretmen ve öğrenci-öğrenci arasında gerçekleşen sosyal etkileşimleriyle ilgilidir. Analiz sonucunda öğrencilerin sordukları sorular ile final notları arasında bir korelasyon bulunmuştur. Ayrıca araştırma sonunda veri ve metin madenciliğinin beraber kullanımıyla öğrencilerin öğrenme davranışlarına dair değerli yapılara ulaşılabileceği önerilmektedir.

BÖLÜM III

YÖNTEM

3.1. Araştırmanın Yöntemi

Bu bölümde, araştırmanın modeli, veri toplama aracı ve tekniklerine ve veri analizi ile ilgili bilgiler bulunmaktadır.

3.2. Araştırma Modeli

Milli Eğitim Bakanı Ziya SELÇUK, Covid-19 pandemisi döneminde başlayan uzaktan eğitim süresince Twitter hesabından paylaşmış olduğu mesajlar üzerinde metin madenciliği yöntemiyle çıkarımlar yapmak amaçlanmıştır.

3.3. Veri Toplama Aracı

Araştırmada kullanılan veriler Twitter sosyal medya aracından alınmıştır.

3.4. Veri Toplama Süreci

23 Mart 2020 tarihinde başlayan uzaktan eğitim sürecinin başlamasıyla Miili Eğitin Bakanı tarafından paylaşılan mesajlar 17 Ekim 2020 tarihine kadar düzenli olarak eksiksiz bir şekilde toplanmıştır.

3.5. Veri Analizi

Bu kesitsel tanımlayıcı araştırma, Türkiye Milli Eğitim Bakanı Ziya SELÇUK' un Covid-19 pandemisi süresinde 23 Mart 2020 - 17 Ekim 2020 tarihleri arasında paylaştığı mesajların içerik analizini Metin madenciliği yöntemi ile yapan nitel bir araştırmadır.Analiz aracı olarak açık kaynak kodlu bir yazılım olan R- 3. 6. 1. tercih edilmişitir.Bununla beraber elde edilen kelime sayıları ile nicel araştırma yöntemi olan X^2 analizi uygulanmıştır.

BÖLÜM IV

BULGULAR VE YORUMLAR

4.1.Ham verinin ön işleme süreci

4.1.1.Verinin R programına aktarılması

Veri ön işleme aşamasının ilk adımı olan metni içe aktarma da işleme başlamadan önce R programında verilerin işlenmesi ve analizinde kullanılacak " readtext" , " quanteda" ve " stringi" paketleri açılmıştır.

Gerekli olan paketler açıldıktan sonra işlem için oluşturulan veri seti R programına okutulmuştur.İlgili kodlar şekil 4.8. ' de verilen R Console' da verildiği gibidir.

```
> ZS<-readtext("c:/users/samsung/desktop/ZS.docx")
> ZS
readtext object consisting of 1 document and 0 docvars.
# Description: df[,2] [1 x 2]
  doc_id  text
  <chr>   <chr>
1 ZS.docx "\"(Türkiyede\"..."
>
```

Şekil 4.8. *Verinin R Programına Aktarılması*

4.1.2.Verinin Temizleme işlemi

Veri ön işleme bölümünün ikinci adımı olarak programa okutulan veri html kodlarından ve gereksiz boşluklardan temizlenmiştir.

```
> ZS <- stri_replace_all(ZS, "", regex = "<.*?>")
> ZS<- stri_trim(ZS)
> ZS
[1] "(Türkiyede ilk vaka 11 Mart ta görüldü.Ara tatil 16 Mart a çekildi.23 Mart$
>
```

Şekil 4.9. *Verinin Temizleme İşlemi*

4.1.3.Verinin gövdelere ayrılması ve normalleştirilmesi

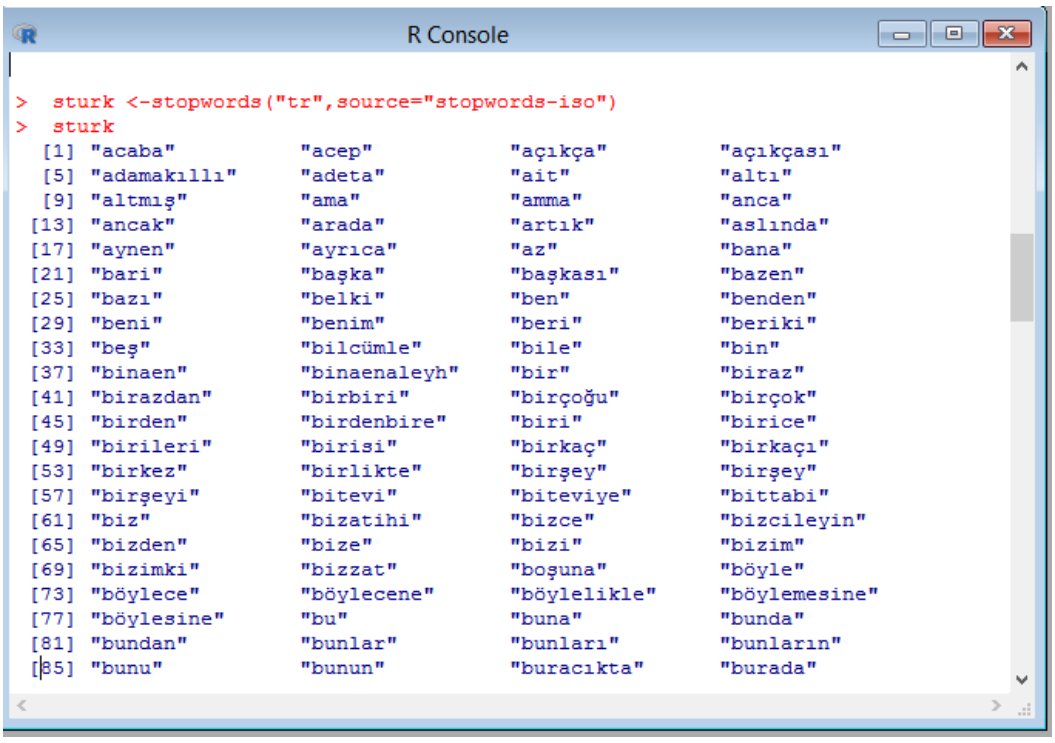
Bu aşamada veri seti önce tokenlerine ayrılmıştır. Yani, her bir cümle tek tek kelimelere ayrılmıştır.

```
> ZS <- stri_trans_tolower(ZS)
> toks  <- tokens (ZS)
> toks <-tokens_tolower(toks)
> toks <-tokens_wordstem(toks)
> ZS
[1] "(türkiyede ilk vaka 11 mart ta görüldü.ara tatil 16 mart a çekildi.23 mart$
> 
```

Şekil 4.10. *Verinin Gövdelere Ayrılması ve Normalleştirilmesi*

Tokenlerine ayrılan veri seti tokenler arasındaki büyük/küçük harf farklılıklarını ortadan kaldırmak için küçük harflere çevrilmiştir.

Gövdeleme ve normalleştirme aşamalarından sonra edat ve bağlaç gibi durak kelimeler korpustan temizlenmiştir. R programlama dili Türkçe edat/bağlaçları içermediği için öncelikle ona Türkçe edat/bağlaçlar tanıtılmıştır.



Şekil 4.11. *Türkçe Edat/Bağlaçların Programa Tanıtılması*

Edat/bağlaçlar tanıtıldıktan sonra korpustan durak kelimeleri temizlenmiştir.

```
> tokens_remove(toks,sturk)
Tokens consisting of 1 document.
text1 :
 [1] "("          "türkiyed"   "vaka"       "11"         "mart"
 [6] "ta"         "görüldü.ara" "tatil"     "16"         "mart"
[11] "a"          "çekildi"
[ ... and 14,792 more ]
```

Şekil 4.12. *Edat/Bağlaçların Korpustan Temizlenmesi*

4.1.4. Döküman Terim matrisinin oluşturulması

Oluşturulan DTM de satır korpusu, sütunlar da bir terimin korpusta ne sıklıkta gerçekleştiğini göstermektedir. Bu gösterimin avantajı, metin halinden rakamlara etkili bir şekilde dönüşerek verilerin vektör ve matris cebiri ile analiz edilmesine izin vermesidir.

```
> dtm <-dfm(toks)
> dtm
Document-feature matrix of: 1 document, 5,748 features (0.0% sparse).
       features
docs    ( türkiyed ilk vaka 11 mart ta görüldü.ara tatil 16
  text1 3        1  27    2  6    8  2           1     5  4
[ reached max_nfeat ... 5,738 more features ]
> |
```

Şekil 4.13. *Döküman Terim Matrisinin Oluşturulması*

4.1.5.Filtreleme ve Ağırlıklandırma

Filtreleme ve ağırlıklandırma bir veri setinde durak kelime sayılmayan ama metin hakkında çok az bilgi veren ya da çok sık tekrarlanan kelimelerin ağırlıklandırılarak filtrelenmesi için kullanılır. Çalışmamızda önceden belirlenen anahtar kelimeler sözlük sayma analizine tabi tutulacağı için bu adım uygulanmamıştır.

4.2 .Verinin analiz edilmesi

Milli Eğitim Bakanı Ziya SELÇUK' un Twitter hesabından toplanan veriler 31 Ağustos kesme tarihi kabul edilerek öncesi ve sonrası şeklinde ikiye bölünmüştür. Sözlük sayma analizinde veri seti sürecin tamamı, 31 Ağustos öncesi ve 31 Ağustos sonrası olarak üç şekilde ele alınmış ve dönemler seçilen anahtar kelimelerin frekansları ve dönem içi yüzdelikleri açısından karşılaştırılmıştır.

Analize başlamadan önce ikiye bölünen veri seti "önce" ve "sonra" şeklinde R programına okutulmuştur.

```
> önce<-readtext("c:/users/samsung/desktop/önce.docx")
> önce
readtext object consisting of 1 document and 0 docvars.
# Description: df[,2] [1 x 2]
  doc_id    text
  <chr>     <chr>
1 önce.docx "\"yede ilk v\"..."
> sonra<-readtext("c:/users/samsung/desktop/sonra.docx")
> sonra
readtext object consisting of 1 document and 0 docvars.
# Description: df[,2] [1 x 2]
  doc_id     text
  <chr>      <chr>
1 sonra.docx "\"Ders zilim\"..."
>
```

Şekil 4.14. *İkiye Bölünen Verinin Programa Okutulması*

Okutulan iki veri setine de ilk veriye uygulanan temizleme, gövdeleme ve normalleştirme basamakları uygulanmıştır.

```
> önce<-readtext("c:/users/samsung/desktop/önce.docx")
> önce <- stri_replace_all(önce, "", regex = "<.*?>")
> önce<- stri_trim(önce)
> önce <- stri_trans_tolower(önce)
> önce
[1] "yede ilk vaka 11 mart ta görüldü.ara tatil 16 mart a çekildi.23 mart ta uz$
> sonra<-readtext("c:/users/samsung/desktop/sonra.docx")
> sonra <- stri_replace_all(sonra, "", regex = "<.*?>")
> sonra<- stri_trim(sonra)
> sonra <- stri_trans_tolower(sonra)
> sonra
[1] "ders zilimiz çaldı. ben okuldayım. öğretmenlerimiz okulda. çocuklarımız ek$
>
```

Şekil 4.15. *Programa Okutulan Verilerin Temizleme, Gövdeleme Ve Normalleştirme Basamakları*

R programına yüklenen üç veri setinin (sürecin tamamı,31 Ağustos öncesi ve 31 Ağustos sonrası) de DTM si hesaplanmıştır. Şekil 4. 16. ' da, text1 sürecin tamamını, text2 31 Ağustos öncesini ve text3 31 Ağustos sonrasını göstermektedir.

```
> text <-c(ZS,önce,sonra)
>   dtm <-dfm(text,tolower = TRUE,stem = TRUE,remove = stopwords("tr",source="s$
> dtm
Document-feature matrix of: 3 documents, 5,540 features (28.4% sparse).
       features
docs    ( türkiyed vaka 11 mart ta görüldü.ara tatil 16 a
  text1 3        1    2  6    8  2           1     5  4 5
  text2 2        0    1  6    8  2           1     5  4 5
  text3 0        0    1  0    0  0           0     0  0 0
[ reached max_nfeat ... 5,530 more features ]
>
```

Şekil 4.16. *Programa Okutulan 3 Veri Setinin DTM Hesaplaması*

Analiz edilen ve döküman terim matrisi hesaplanan veri kümesi, tablo 4. 5. ' te gösterilmektedir.

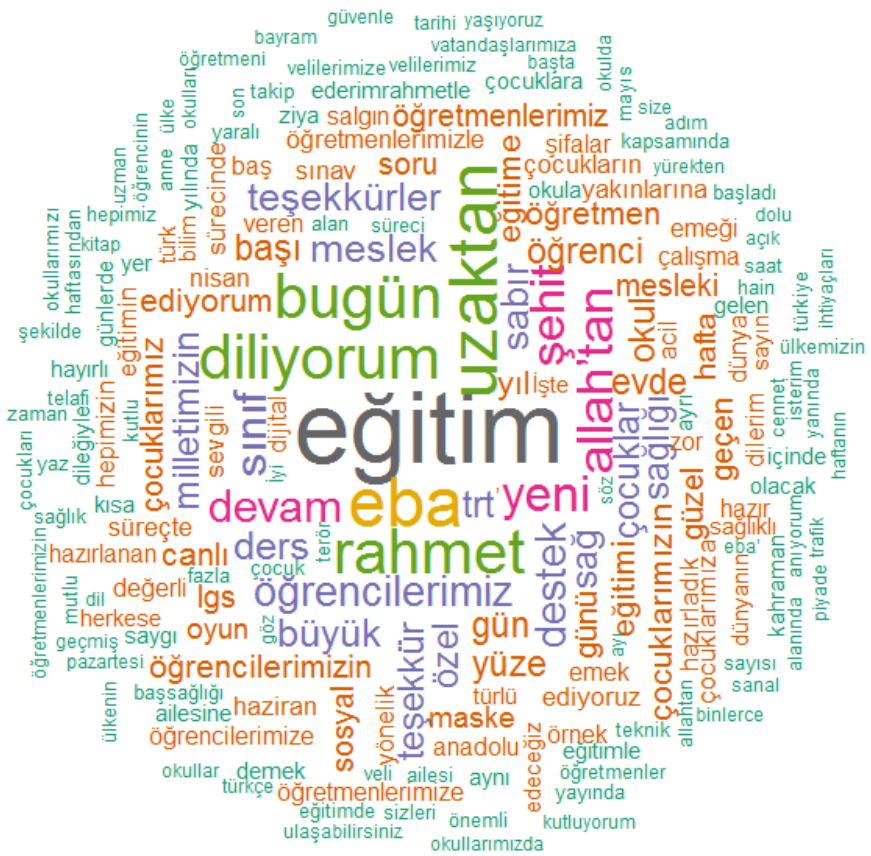
Tablo 4.5. *Veri Analizi Sonuçları*

	yks		telafi		eğitim	
	n	**%**	**n**	**%**	**n**	**%**
23 MART-17 EKİM	3		9		124	
23 MART-30 AĞUSTOS	3	100%	9	100%	84	67,74%
31 AĞUSTOS-17 EKİM	0		0		40	32,25%
	lise		**ilkokul**		**salgın**	
	n	**%**	**n**	**%**	**n**	**%**
23 MART-17 EKİM	7		13		14	
23 MART-30 AĞUSTOS	7	100%	7	54%	8	57%
31 AĞUSTOS-17 EKİM	0		6	46%	6	43%

	veli		lgs		eba	
	n	**%**	**n**	**%**	**n**	**%**
23 MART-17 EKİM	10		20		61	
23 MART-30 AĞUSTOS	7	70%	20	100%	37	60%
31 AĞUSTOS-17 EKİM	3	30%	0		24	40%
	oyun		**sınav**		**teknoloji**	
	n	**%**	**n**	**%**	**n**	**%**
23 MART-17 EKİM	18		17		4	
23 MART-30 AĞUSTOS	13	72%	16	94%	4	100%
31 AĞUSTOS-17 EKİM	5	28%	1	6%	0	

Araştırmada pandemi sürecinde eğitim-öğretim sürecinin değerlendirlmesinde anahtar olduğu düşünülen 12 kelime belirlenmiştir. Bu kelimeler "YKS", "telafi", "veli", "eğitim", "EBA", "LGS", "salgın", "oyun", "sınav", "teknoloji", "lise", "ilkokul" olarak belirlenmiştir. Bu kısımda sözlük sayma analiz yöntemiyle belirlenen bu kelimelerin 23 Mart 2020 tarihinden 17 Ekim 2020 tarihine kadar toplamda kaç kez kullanıldığı, 31 Ağustos' tan önce ve 31 Ağustos' tan sonra kaç kez kullanıldığı hesaplanacak ve araştırma sorularına cevap aranacaktır.

Öncelikle kelime bulutu tekniği ile Milli Eğitim Bakanı Ziya SELÇUK' un pandemi sürecinde paylaştığı mesajlar kelime bulutu tekniği ile görselleştirilmiştir. Sonra seçilen her bir anahtar kelime frekanları baz alınarak süreç içerisnde yorumlanmış. Son olarak da seçilen kelimeler 4 kategoride birleştirilerek SPSS programı ile ki-kare homojenlik testi yapılmıştır.



Şekil 4.17. *Kelime Bulutu*

Kelime bulutu tekniğinde frekansı en yüksek olan kelime en büyük punto ile görsellştirlmektedir. Diğer kelimelerde frekansları oranında büyükten en küçük puntoya doğru görselleştirilmektedir. Burada, Bakanın en çok kullandığı kelimenin "eğitim" olduğu göze

çarpmaktadır. Sonrasında "eba", "uzaktan", öğrenci" vb. gibi süreci yansıtan kelimeler göze çarpmaktadır.

Milli Eğitim Bakanı Ziya SELÇUK, uzaktan eğitimin başladığı tarih olan 23 Mart 2020' den 17 Ekim 2020 tarihine kadar 3 kez "YKS" kelimesini kullanmıştır. Bu kelimelerin üçünü de 31 Ağustos' tan önce kullanmıştır. YKS' nin 27-28 Haziran' da yapıldığı için 31 Ağustos' tan sonra bu kelimenin kullanılmaması analiz sonuçlarıyla örtüşmektedir.

"Telafi" kelimesi 9 kez kullanılmış ve tamamı 31 Ağustos' tan önce kullanılmıştır. Telafi eğitimi 31 Ağustos' ta başladığı için tamamının başlama tarihinden önce kullanılması Milli Eğitim Bakanının süreç başlayana kadar veli-öğretmen ve öğrencileri telafi eğitimine güdülemek için mesajlarında bir çok kez tekrarladığının kanıtıdır.

"Eğitim" kelimesi toplamda 124 kez kullanılmış, 84 tanesi 31 Ağustos öncesi 40 tanesi de sonrasında tekrarlanmıştır. Bakanın her mesajında eğitim kelimesini tekrarlaması eğitime verdiği önemin bi göstergesi kabul edilebilir.

"Veli" kelimesi toplamda 10 kez tekrarlanmış; 7 tanesi öncesinde 3 tanesi de sonrasında kullanılmıştır. Bakanın her iki dönmede de "veli" kelimesini kullanması toplum olarak yaşadığımız bu zorlu pandemi sürecinde velileri de eğitim -öğretim sürecine dahil etmenin gerekliliğini savunduğunun göstergesidir.

" LGS" kelimesi 20 kez tekrarlanmış ve tamamı da 31 Ağustos öncesinde kullanılmıştır. LGS' nin 20 Haziran' da yapılması ve sonuçlarının 10 Ağustos' ta açıklanması sonraki dönemde tekrarlanmamış olmasının makul bir sonucudur. YKS' nin 3 kez LGS' nin ise 20 kez tekrarlanmış olması LGS ye girecek öğrencilerin yaşlarının daha küçük olması, öğrencilerinin ve velilerinin daha çok desteğe ihtiyaç duydukları şeklinde yorumlanabilir.

"EBA" kelimesi toplamda 61 kez; 31 Ağustos öncesinde 37 kez sonrasında ise 24 kez kullanılmıştır. Pandemi nedeniyle eğitim- öğretime uzaktan eğitim yoluyla devam edildiği için EBA öğrenci-öğretmen ve velilerin hayatlarının ayrılmaz bir parçası haline gelmiştir. Bakanın mesajlarında bu kadar çok tekrarlaması da bu durumun bir göstergesidir.

"Lise" kelimesi toplamda 7 kez kullanılmış, tamamı da 31 Ağustos öncesinde kullanılmıştır. Bu durum da LGS kelimesiyle ve onun gibi 31 Ağustos sonrası kullanılmamasıyla ve 21 Eylül' de başlayan aşamalı ve seyreltilmiş yüzyüze eğitim sürecine liselerin dahil edilmemesiyle örtüşmektedir.

"İlkokul" kelimesi toplamda 13 kez kullanılmış, 7 tanesi 31 Ağustos öncesi 6 tanesi de sonrasında tekrarlanmıştır. 21 Eylül' de başlayan aşamalı ve seyreltilmiş yüzyüze eğitim sürecine okul öncesi ve 1. sınıflarla başlandığı için önceki ve sonraki süreçte ilkokul kelimesinin tekrarlanmış olması yaşanılan durumla örtüşmektedir.

"Salgın" kelimesi 14 kez kullanılmıştır. 31 Ağustos' a kadar 8 kez sonrasında ise 6 kez tekrarlanmış olması pandeminin hayatımızı ve eğitim-öğretimi ne denli etkilediğini göstermektedir.

"Oyun" kelimesi 31 Ağustos' a kadar 13 sonrasında 5 kez olmak üzere toplamda 18 kez tekrar edilmiştir. Bakanın pandemi sürecinden psikolojik olarak en çok etkilenen çocukları bu durumdan biraz kurtarmak için oyun oynamaya verdiği önem gözlenmktedir. 31 Ağustos' tan önce oyun kelimesinin daha çok tekrar edilmesi 4 Nisan 2020 tarihinden 1 Haziran 2020 tarihine kadar sokağa çıkma yasağı uygulanan çocukların evlerde sıkılmaması için oyun oynamalarına verilen önemi göstermektedir.

"Sınav" kelimesi toplamda 18 kez tekrarlanmıştır. 16 tanesinin 31 Ağustos' tan önce tekrarlanmış olması LGS ve YKS' nın belirtilen süreçte yapıldığının göstergesi olarak yorumlanabilir.

"Teknoloji" kelimesi toplamda 4 kez kullanılmış ve tamamı da 31 Ağustos'tan önce kullanılmıştır. 23 Mart'ta ilk kez uzaktan eğitim sitemine geçilmesi, bir çok veli-öğrenci ve öğretmenin EBA, canlı ders ve zoom uygulamalarına fazla aşina olmaması 31 Ağustos' tan önceki süreçte teknoloji kelimesinin tekrarlanmasının bir nedeni olarak düşünülebilir.

Sonraki aşamada seçilen 12 kelime 4 kategori altında toplanmış ve bu kategoriler arasında SPSS programıyla kikare homojenlik testi yapılmıştır.

Tablo 4.6. *Birleştirilen Kelime Kategorileri*

Kelimeler				**Kategori**
YKS	LGS	SINAV		ÖLÇME
TELAFİ	EĞİTİM	LİSE	İLKOKUL	EĞİTİM
VELİ	OYUN	SALGIN		SOSYALHAYAT
EBA	TEKNOLOJİ			TEKNOLOJİ

Milli Eğitim Bakanı Ziya SELÇUK' un uzaktan eğitim sürecinde 23 Mart 2020-30 Ağustos 2020 tarihleri ile 31 Ağustos 2020 -17 Ekim 2020 tarihleri arasında kullandığı kelimeler arasında anlamlı bir fark olup olmadığını belirlemek amacıyla iki değişken arasında kikare homojenlik testi yapılmıştır. Yapılan analiz sonucunda kullanılan kelimelerin dönemlere göre anlamlı bir biçimde farklılaştığı belirlenmiştir, $X^2(3)=16.251$, $p=0.001$.

Tablo 4.7. *Dönem*Kelime Kategorileri Çapraz Tablolama*

		Kelime Kategorileri				Toplam
		Ölçme	Eğitim	Sosyal hayat	Teknoloji	
Dönem 23 Mart-30 Ağustos	Sayım	39	107	28	41	215
	% Dönem içinde	18.1%	49.8%	13.0%	19.1%	100.0%
	% Kelime içinde	97.5%	69.9%	66.7%	63.1%	71.7%
	%Toplamda	13.0%	35.7%	9.3%	13.7%	71.7%
31 Ağustos-17 Ekim	Sayım	1	46	14	24	85
	% Dönem içinde	1.2%	54.1%	16.5%	28.2%	100.0%
	% Kelime içinde	2.5%	30.1%	33.3%	36.9%	28.3%
	%Toplamda	0.3%	15.3%	4.7%	8.0%	28.3%
Toplam	Sayım	40	153	42	65	300
	% Dönem içinde	13.3%	51.0%	14.0%	21.7%	100.0%
	% Kelime içinde	100.0%	100.0%	100.0%	100.0%	100.0%
	%Toplamda	13.3%	51.0%	14.0%	21.7%	100.0%

Tablo 4.8. *Ki-Kare Homojenlik Testi*

	Değer	Serbestlik derecesi (df)	İkik yönlü anlamlılık düzeyi (p)
Pearson Ki-kare	16.251 [a]	3	.001
En çok olabilirlik oranı	22.120	3	.000
Doğrusal bağlantı	9.629	1	.002

BÖLÜM V

SONUÇ VE ÖNERİLER

5.1.Sonuç

Ölçme konularına(YKS,LGS,Sınav) analiz edilen sürecin tamamında (23 Mart -17 Ekim tarihleri arasında) %13.3 oranında değinilirken bu oranın tamamına yakınına ilk dönemde değinilmiştir. Bunun sebebi gerek okullarda yapılan geçme-kalma sınav türlerinin gerekse YKS ve LGS gibi önemli sınavların haziran ve temmuz aylarına denk gelmesi olarak yorumlanabilir.

Eğitim konularına (telafi-lise-eğitim-ilkokul) tüm süreçte %51 oranında değinilmiştir.Birinci dönemde (%49.8) ve ikinci dönemde (%51) birbirine yakın oran gözlenmektedir. Bu durum her iki dönemde de paylaşılan tüm mesajların yarısının eğitim konularını içerdiği gözlenmektedir.

Sosyal hayat içerikli mesajlar (veli, oyun, salgın) her iki dönemin toplamında %14 oranındadır. İlk dönemde paylaşılan tüm mesajların %13 ünü ikinci dönemde atılan tüm mesajların ise %16,5 ini kapsamaktadır. Bu durum ilk dönemde yaşanılan ölçme değerlendirme yoğunluğunun ikinci dönemde azalması şeklinde yorumlanabilir. 31 Ağustos-17 Ekim sürecinde çocukları okula yeni başlayan ya da bir üst kademeye atlayan veliler, eğitim hayatına yeni atılan öğrencilerin orayntasyonu sürecinde oyunlaştırma yöntemine başvurma, pandemi ve kısıtlamalardan iyice bunalan çocukların yüz yüze eğitimin de başlamasıyla temazsız oyunlara verilen önem vb durumlar nedeniyle ikinci dönemde sosyal hayat içerikli mesajların artması beklenilen bir durumdur.

Teknoloji konulu (Eba, Teknoloji) mesajlar ilk dönemde % 19,ikinci dönemde %28, her iki dönemin toplamında atılan mesajların ise %21.7 sini kapsamaktadır. Pandemi nedeniyle 23 1Mart' ta ilk kez uzaktan eğitime geçilmiştir. Bu süreçte farklı uygulamalar üzerinden gerçekleştirilebilen canlı dersler her eğitim kademesinde zorunlu tutulmamış, dersler daha çok EBA TV den takip edilm veya öğrencilere ödev, video gönderme şeklinde tamamlanmaya çalışılmıştır. Fakat 31 Ağustos' tan itibaren gerek telafi eğitimleri süreci gerekse kademeli ve seyreltilmiş eğitim uygulamalarıyla tüm kademelerde canlı dersler zorunlu tutulmuş, canlı dersler EBA sistemine kaydedilerek açılmaya başlanmış ve EBA alt yapısı güçlendirilmiştir. Tüm bu yaşanan gelişmeler 31 Ağustos' tan sonraki süreçte teknoloji kullanımını arttırmış ve bu durum da paylaşılan mesajların artması şeklinde sonuçlanmıştır.

Her dönemi kendi içinde değerlendirmek gerekirse 23Mart - 30 Ağustos döneminde, her eğitim kademesinde sınıf geçme-kalma değerlendirmeleri, liseye ve üniversiteye geçme ve yerleştirme sınav sonuçları Haziran-Temmuz ayına rastladığı için bu bu süreçte atılan ölçme konulu mesajlar (YKS,LGS,Sınav) %18; ilk kez uzaktan eğitime geçilen süreçte eğitimi aksatmamak için yapılan yoğun çalışmalar, yüz yüze derslerin telafisi vb konulu eğitim mesajları (telafi-lise-eğitim-ilkokul) % 50; veli ve öğrencilerin yaşadığı stres, sokağa çıkma yasadığından fazlasıyla etkilenen çocuklar için oyunun önemi, sosyal mesafe ve hijyen kuralları içerikli mesajlar, sosyal hayat mesajları (veli, oyun, salgın) %13; yüz yüze eğitime ara verilerek derslerin EBA TV ve canlı derslerle yürütlmeye çalışıldığı bu süreçte teknolojikonulu mesajlar (Eba,Teknoloji) %19 oranındadır.

31 Ağustos-17 ekim sürecinde ise tüm ölçme-değerlendirme ve yerleştirmeler tamamlandığı için ölçme konulu mesajlar (YKS, LGS, Sınav) %1.2 oranına düşmüştür ve bu beklenilen bir durumdur. Eğitim konulu mesajlar (telafi-lise-eğitim-ilkokul) her durumda odak noktası olduğu için bu dönemde de %54 oranındadır. Sosyal hayat konulu mesajlar (veli,oyun,salgın) % 16.5 ve teknoloji konulu mesajlar (Eba,Teknoloji) ise %28 oranındadır.

5.2.Öneriler

Araştırma sonuçları doğrultusunda öneriler aşağıda sunulmuştur:

1) Sözlük sayma analizi yöntemi ile öğrenci performans ödevleri, yazılı kağıtları gibi değerlendirmesi hem zaman alıcı hem de objektifliğin tehlikeye girdiği değerlendirme süreçlerinde hem zaman tasarrufu sağlanabilir hem de daha nesnel değerlendirmeler yapılabilir. Sözlük sayma analizinde yer aldığı gibi her bir soru için kategori ve kategori altındaki kelimeler değerlendirici tarafından belirlenerek hızla değerlendirme yapılabilir. Bu kapsamda hazırlanacak bir yazılım iş yükünü, yorgunluğu, subjektifliği ve dikkatsizliği ortadan kaldırabilir. Eşit, adil ve hızlı bir değerlendirmenin önünü açabilir.

2) Metin madenciliği yöntemlerinden biri olan kümeleme yöntemiyle gruplanmamış verileri benzerliklerine göre gruplanmaktadır. Veri setlerindeki kayıp değerlere müdahale bu şekilde sağlanabilir.

3) Rehberlik ve Psikolojik Danışmanlık servislerinde en çok karşılaşılan öğrenci problemlerinin analizinde metin madenciliği yöntemleri kullanılabilir.

4) Milli Eğitim Bakanı'nın uzaktan eğitim sürecinde paylaştığı mesajlardan en çok yorum yapılanlar ve yeniden paylaşılanlar üzerinde filtreleme ve ağırlıklandırma analizleri yapılabilir.

5) Uzaktan eğitime yönelik internet ortamında yer alan öğretmen - öğrenci - veli görüşlerinin analiz edilerek süreç hakkında başarılı bulunan ve ya eksik kalan konular hakkında fikir edinilebilir.

6) Farklı ülkelerin eğitim bakanlarının paylaştığı mesajlar analiz edilerek karşılaşılan benzer sorunlar ve ya farklılıklar tespit edilebilir.

Metin analizinde kullanılan makine öğrenmesi algoritmaları yapay zekaya dayanmaktadır. Milli Eğitim Bakanı Ziya SELÇUK' un da yapay zeka kullanımına ve veri madenciliğine verdiği önem 2023 Eğitim Vizyonunda dikkat çekmektedir. Eğitim vizyonunda geniş yer tutan "Veriye Dayalı Yönetimle" ülke çapında eğitimin sağlıklı bir şekilde yönlendirilmesi amacıyla geçmiş kararlara yönelik objektif değerlendirmeler ve geleceğe yönelik gerçekçi planlar yapılabileceği bunlar için de çeşitli ve büyük miktarda veri yığınlarının analiz edilerek birbirleriyle ilişkilendirilmesi, sürekli değişen şartlara göre güncellenmesi ve sebep sonuç ilişkisi yönünden anlamlandırılmasının önemine değinilmektedir. Bu gelişmeler göz önüne alındığında metin madenciliği uygulamalarının kısa bir zaman sonra eğitim ve öğretim alanında çok sık kullanılacağını göstermektedir.

KAYNAKÇA

Abuzir, Y. (2018). Innovative model for student project evaluation based on text mining. *International Journal of Research in Education and Science (IJRES), 4*(2), 409-419. DOI:10.21890.409481

Aksoy, E. (2014). Matematik Alanında Üstün Zekalı Ve Yetenekli Öğrencilerin Bazı Değişkenler Açısından Veri Madenciliği İle Belirlenmesi. Dokuz Eylül Üniversitesi, Yüksek Lisans Tezi.

Albayrak, M. (2008). Eeg Sinyallerindeki Epileptiform Aktivitenin Veri MadenciliğiSüreci İle Tespiti. Sakarya Üniversitesi, Doktora Tezi.

Aydın,S.(2007). Veri Madenciliği Ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama. Anadolu Üniversitesi, Doktora Tezi.

AYDOĞAN, F.(2003), e-Ticarette Veri Madenciliği Yaklasımlarıyla Müsteriye HizmetSunan Akıllı Modüllerin Tasarımı ve Gerçeklestirimi. Hacettepe Ün., Fen Bil. Ens., Yüksek Lisans Tezi.

Ayık,Y. Z. , Özdemir, A. ,Yavuz, U. (2007). Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 10(2):441-454

Ayre,L. (2006). *Data Mining for Information Professionals.* 23 Temmuz 2020 tarihinde, https://www.researchgate.net/publication/228386369_Data_Mining_for_Information_Professionals adresinden erişildi.

Bayraktar Çepni, S. (2014). Impact Of "Text Mınıng And Imıtatıng Strategıes" On Lexıcal Rıchness, Lexıcal Dıversıty And General Success In Second Language Wrıtıng. Karadeniz Teknik Üniversitesi, Yüksek Lisans Tezi.

Benoit, K. , Watanabe, K. , Nulty, P. , Obeng, A. , Wang, H. , Lauderdale, B. , & Lowe, W. (2017). *quanteda: Quantitative analysis of textual data* [Computer software manual] (R package version 0.99). Retrieved from http://quanteda.io Bioinformatics Track Delft University of Technology Delft.

Bot, J. (2007). *Text-mining in the Life-Sciences, an Exploration*. Netherlands: Bioinformatics Track Delft University of Technology Delft.

Boyacı, A. (2017). Öğretmenlerin Algılanan Örgütsel Destek Ve Örgütsel Özdeşleme Düzeylerinin Veri Madenciliği İle Analizi. Hitit Üniversitesi, Yüksek Lisans Tezi.

Bölükbaş, M .Ay. (2013). Çalışan Memnuniyetinin Veri Madenciliği İle İncelenmesi. Mimar Sinan Güzel Sanatlar Üniversitesi, Yüksek Lisans Tezi .

Consoli, D. (2010). A New Framework To Extract Knowledge By Text Mınıng Tools. *Bilgi Ekonomisi Ve Yönetimi Dergisi*,V(II).

Chan, N. L. & Denizci, G. B. (2011). Investigation of Social Media Marketing: How does the Hotel Industry in Hong Kong Perform in Marketing on Social Media Websites?. *Journal of Travel & Tourism Marketing*, 28:4(345-368).

Çeliksu,Z. (2017).Yabancı Dizilerin Altyazı Ve Twıtter Yorumlarının Metin Madenciliği İle İncelenmesi. Mimar Sinan Güzel Sanatlar Fakültesi, Yüksek Lisans Tezi.

Çelikyay, E. K. (2010). Metin Madenciliği Yontemiyle Turkcede En Çeşitli Değişkenler Açısından İncelenmesi. Yıldız Teknik Üniversitesi, Yüksek Lisans Tezi.

Çelikyay, E. K. (2010). Metin Madenciliği Yontemiyle Turkcede En Sık Kullanılan Ve Birbirini Takip Eden Harflerin Analizi Ve Birliktelik Kuralları. Beykent Üniveristesi, Yuksek Lisans Tezi.

Çevik, K. (2019). 3-6 Yaş Arasındaki Çocuklara Sahip Ebeveynlerin Çocuk Yetiştirme Tutumları İle Sosyal Medya Kullanımı Arasındaki İlişki. Aydın Üniversitesi, Yüksek Lisans Tezi.

Dalmolen, S. (2010). Defining Patterns İn Unstructured Manifests İn A Volatile Cross-Domain Environment. University Of Groningen, Master's Thesis.

Dang, S. , & Ahmad, P. H. (2014). A comparative study on text mining techniques. *Global Journal of Advanced Research*, 1(2), 128-134.

Dinler, M. (2014). Kümeleme analizi yöntemlerinin hayvancılık verilerinde karşılaştırmalı olarak incelenmesi. Bingöl University, Master's thesis.

Doğan, Ş. , Türkoğlu, İ. (2007). Karar Ağacı Yöntemini Kullanarak Tiroid Hormon Parametrelerinden Hipertiroidi Ve Hipotiroidi Teşhisi.*Doğu Anadolu Bölgesi Araştırmaları.*163-169.

Dolgun, M. Ö., Özdemir G. T., Oğuz D., 2009. Veri madenciliğinde yapısal olmayan verinin analizi: metin ve web madenciliği. *İstatistikçiler Dergisi*. (2), pp.48-58.

Döven, S. (2013). Metin Madenciliği İle Dokümanlar Arasındaki Benzerliklerin Bulunması. Bahçeşehir Üniveristesi, Yuksek Lisans Tezi.

Erestin, E. (2019). Sosyal Medya Kullanımı, Kişilik Özellikleri Ve Girişimcilik Niyeti İlişkisi: Üniversite Öğrencileri Üzerine Bir Araştırma. Gebze Teknik Üniversitesi, Yüksek Lisans Tezi.

Ergün, M. (2017). Eğitim Araştırmalarında Data Mınıng Ve Text Mınıng Tekniklerinden Yararlanma. *Elektronik Eğitim Bilimleri Dergisi*,6(12),180-189.

Erkul, E. R. (2009). Sosyal Medya Araçlarının (Web 2.0) Kamu Hizmetleri ve Uygulamalarında Kullanılabilirliği. *Türkiye Bilişim Derneği Dergisi*, *116*(1), 96-101.

Gagolewski, M. (2017). *R package stringi: Character string processing facilities* [Computer software manual]. Retrieved from http://www.gagolewski.com/software/stringi/

Gaikwad, S. V., Chaugule, A., & Pramod, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, 85(17), 43-44.

Gao, L. Chang, E. Han, S. (2005). *World Academy of Science, Engineering and Technology*. Retrived from http://www.waset.org/journals/waset/v8/v8-21.pdf (15.01.2011).

Gıudıcı, P.(2003). *Applied Data Mining Statistical Methods for Business and Industry*. Library of Congress Cataloging-in-Publication Data.

Göker,H.,Tekedere,H.(2017). Fatih Projesine Yönelik Görüşlerin Metin Madenciliği Yöntemleri İle Otomatik Değerlendirilmesi. *Bilişim Teknolojileri Dergisi*,10(3). Doı: 10.17671/Gazibtd.331041.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. In Proceedings of the National Academy of Sciences,5228–5235. doi:10.1073/pnas.0307752101

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. doi:10.1093/pan/mps028

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60-63.

Han, J ., Kamber, M. (2006). *Data Mining:Concepts And Techniques*.(2nd Edition). Morgan Kaufmann Publishers.

Hand, D. , Mannila, H. ,Smyth, P. (2001). *Principles of Data Mining*.The MIT Press.

He,W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*,29, 90–102.

Hvass, K.A. & Munar, A.M. (2012). The Takeoff of Social Media in Tourism. *Journal of Vacation Marketing*, 18, 93-103.

İmik Tanyıldızı, N. , Ateş, T.Y. (2018). Siyasi Parti Genel Başkanlarının 15 Temmuz Darbe Girişimi Sonrası Twitter Mesajlarına Yönelik İçerik ve Söylem Çözümlemesi. *ASSAM International Refereed Journal*, 10: 46-63.

Karabatak, M. (2008). Özellik Seçimi, Sınıflama Ve Öngörü Uygulamalarına Yönelik Birliktelik Kuralı Çıkarımı Ve Yazılım Geliştirilmesi. Fırat Üniversitesi, Doktora Tezi.

Karaca, M. F. (2012) .Metin Madenciliği Yöntemi ile Haber Sitelerinde Köşe Yazılarının Sınıflandırılması. Karabük Üniversitesi, Yüksek Lisans Tezi.

Karataş,S. (2019). Ortaokul 5. Sınıf Öğrencilerinin Matematik, Fen Bilimleri Ve Türkçe Dersleri Kazanımlarına Ulaşma Düzeylerinin İncelenmesi:Veri Madenciliği Çalışması (Afyonkarahisar Örneklemi). Kocatepe Üniversitesi, Yüksek Lisans Tezi.

Kılınç, D. , Borandağ,E. ,Yücalar, F. , Tunalı,V. ,Şimşek, M. ,Özçift, A. (2016). *Marmara Fen Bilimleri Dergisi,* 3: 89-94. DOI: 10.7240/mufbed.69674

Kireççi,C.(2019). Üniversite Öğrencilerinde Sosyal Medya Kullanımının Dikkat Eksikliği Ve Hiperaktivite Bozukluğu Belirtileri,Sosyal Görünüş Kaygısı Ve Akademik Erteleme İle İlişkisinin İncelenmesi. Beykent Üniversitesi, Yüksek Lisans Tezi.

Lang, D. T., & the CRAN Team. (2017). *XML: Tools for parsing and generating XML within R and S-plus* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=XML

Larose, D. (2005). *Dıscovering Knowledge In Data An Introduction To Data Mining*. New Jersey: Published by John Wiley & Sons, Inc.

Lau, K. Lee, K. Ho, Y. (2005). Text Mining for the Hotel Industry. Cornell Unıversıty DOI:0.1177/0010880405275966 46 (3,) 344-362.

Lerman, K.(2007). Social İnformation Processing in News Aggregation . *IEEE Internet Computing*, 11(6),16-28.

Manning, C. D., Manning, C. D., Raghavan, P., Raghavan, P., Schütze, H., & Schütze, H. (2008). Introduction to information retrieval. *Cambridge, UK: Cambridge University Press.* doi:10.1017/cbo9780511809071

Melek,C.(2012). Metin Madenciliği Teknikleri İle Şirketlerin Vizyon İfadelerinin Analizi. Dokuz Eylül Üniversitesi, Yuksek Lisans Tezi.

Murathan,T. , Devecioğlu, S. (2018). Veri Madenciliği Ve Spor Alanındaki Uygulamaları. *Hacettepe Journal Of Sport Sciences*, 29 (3), 147–156.

Oğuzlar, A. (2005). Kümeleme Analizinde Yeni Bir Yaklaşım: Kendini Düzenleyen Haritalar (Kohonen Ağları) İ*ktisadi Ve* İ*dari Bilimler Dergisi,19(2),93-107.*

Onat, A. (2008). Veri Madenciliğinin Web Tabanlı Uygulamalarda İnsan Uyumluluklarının Tesbiti Üzerine Bir Çalışma. Selçuk Üniversitesi, Yüksek Lisans Tezi.

Ooms, J. (2014). *The jsonlite package: A practical and consistent mapping between json data and r objects* [Computersoftware manual]. Retrieved from https://arxiv.org/abs/1403.2805

Ooms, J. (2017a). *Extract text from microsoft word documents* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=antiword

Ooms, J. (2017b). *Text extraction, rendering and converting of pdf documents* [Computer software manual].Retrieved from https://CRAN.R-project.org/package=pdftools

Özbay,Ö.(2015). Veri Madenciliği Kavramı Ve Eğitimde Veri Madenciliği Uygulamaları. *Uluslararaı Eğitim Bilimleri Dergisi*,2(5), 262-272.

Özbay,Ö. (2015). Öğretim Yönetim Sistemi Üzerinde Üniversite (Lisans) Düzeyindeki Öğrenci Hareketliliğinin Veri Madenciliği Yöntemleriyle Analizi. Başkent Üniversitesi, Yüksek Lisans Tezi.

Özekes S. (2003). Data Mining Models and Application Areas. *İstanbul Commerce University Journal of Science.* 3, 65-82.

Özkan, M. F. , Türkmen, d. (2020). Sosyal Medyanın Siyasal İletişim Aracı Olarak Kullanılması: Twitter Ankara Milletvekilleri Örneği . *Bilim Armonisi Dergisi,* 3 (1): 5-15. doi: 10.37215/bilar .539861.

Pasin, E. (2018). Investıgatıon of Text Mınıng Methods on Turkısh Text. Dokuz Eylül Üniversitesi, Yüksek Lisans Tezi.

Savaş, S. ,Topaloğlu, N. ,Yılmaz, M. (2012).Veri Madenciliği Ve Türkiye'deki Uygulama Örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*,11(21), 1-23

Sever,H.,Oğuz,B.(2002). *Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım Kısım I: Eşleştirme Sorguları Ve Algoritmalar.* 01 Eylül 2020 tarihinde Https://Www.Researchgate.Net/Publication/26433812_Veri_Tabanlarinda_Bilgi_Kesfine_Formel_Bir_Yaklasim_Kisim_1 adresinden erişildi.

Şen, F. (2008). Veri Madenciliği İle Birliktelik Kurallarının Bulunması.Sakarya Üniversitesi, Yüksek Lisans Tezi.

Şengür, D. (2013). Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları İle Tahmini. Fırat Üniversitesi, Yüksek Lisans Tezi.

Şengür, D. (2013). Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları İle Tahmini. Fırat Üniversitesi, Yüksek Lisans Tezi.

Taha, M. S. (2011). Metin Madenciliği ile Döküman Demetleme.Gazi Üniversitesi, Yüksek Lisans Tezi.

Talib, R. , Hanif, M. H. , Ayesha, S. , Fatima, F.(2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*,7(11),414-418.

Tarhan, A. (2012). Büyükşehir Belediyelerinin Sosyal Medya Uygulamalarına Halkla İlişkiler Modellerinden Bakmak. *İletişim Kuram ve Araştırma Dergisi*, 35: 79-101

Taşdemir, M. (2012). Veri Madenciliği(Öğrenci Başarısına Etki Eden Faktörlerin Regresyon Analizi İle Tespiti). Dicle Üniversitesi, Yüksek Lisans Tezi.

Teyseyre, A. R., & Campo, M. R. (2009). An overview of 3D software visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(1), 87-100.

Uzun,Ö. , Yıldırım,V,Uzun,E.(2016). Dikkat Eksikliği Hiperaktivite Bozukluğu olan Ergenlerde Sosyal Medya Kullanım Alışkanlıkları ve Sosyal Medya Bağımlılığı, Benlik Saygısı ve Algılanan Sosyal Destek İlişkisi. *Turkish journal of family medicine and primary care*,10(3):142-147, DOI:10.21763/tjfmpc.16425.

Uzun,V.(2014). Semantıc Text Mınıng and an Applıcatıon ın Turkısh Documents. Dokuz Eylül Unıversity, Master's Thesis.

Varol, M. (2011) Metin Madenciliği Yöntemlerini Kullanarak Türkçe Dökümanlarda Tür Ve Yazar Tanıma. Süleyman Demirel Üniversitesi, Yüksek Lisans Tezi.

Weiss, S. M. , Indurkha, N. , Zhang, T. , Damerau, F.J. (2005). *Text mining: Predictive methods for analyzing unstructured information.* New York: Springer Science+Business Media, Inc.

Weiss, S.M., Indurkhya, N., & Zhang, T. (2010). *Information retrieval and text* mining: In fundamentals of predictive text mining (4th ed.). London.

Welbers, K. , Atteveldt, W. V. , Benoit, K. (2017). Communication Methods and Measures ,11(4) https://doi.org/10.1080/19312458.2017.1387238

Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization.* In Proceedings of the Fourteenth International Conference on Machine Learning (ICML) (pp. 412–420), Nashville, TN, July 1997.

Yıldız,O.(2016). Metin Madenciliğinde Anahtar Kelime Seçimi Bir Üniversite Örneği. *Yönetim Bilişim Sistemleri Dergisi*,2(3),29-50.

Zontul,M.,Aydın,G.(2017)Nosql Veri Tabanları Uzerinde Bir Metin Madenciliği Uygulaması. *Altınbaş Üniversitesi Mühendislik Sistemleri Ve Mimarlık Dergisi*,1(1),*103-113*

BİLDİRİM

Hazırladığım tezin tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin kağıt ve elektronik kopyalarının Akdeniz Üniversitesi Eğitim Bilimleri Enstitüsü arşivlerinde aşağıda belirttiğim koşullarda saklanmasına

izin verdiğimi onaylarım.

☐ Tezimin tamamı her yerden erişime açılabilir.

☐ Tezim sadece Akdeniz Üniversitesi yerleşkelerinden erişime açılabilir.

☐ Tezimin/Raporumun …….. yıl süreyle erişime açılmasını istemiyorum. Bu süre- nin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin tamamı her yer- den erişime açılabilir.

20/01/2021

Emine İÇÖZ

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı : Emine İÇÖZ

Doğum Yeri ve Tarihi : Antalya 16/11/1986

Eğitim Durumu

Lisans Öğrenimi : Dumlupınar Üniversitesi
Fen Edebiyat Fakültesi
İngiliz Dili ve Edebiyatı

Yüksek Lisans Öğrenimi : Akdeniz Üniversitesi Eğitim bilimleri Enstitüsü Eğitimde Ölçme ve Değerlendirme Programı

Bildiği Yabancı Diller : İngilizce

Bilimsel Faaliyetler : -

İş Deneyimi

Çalıştığı Kurumlar :

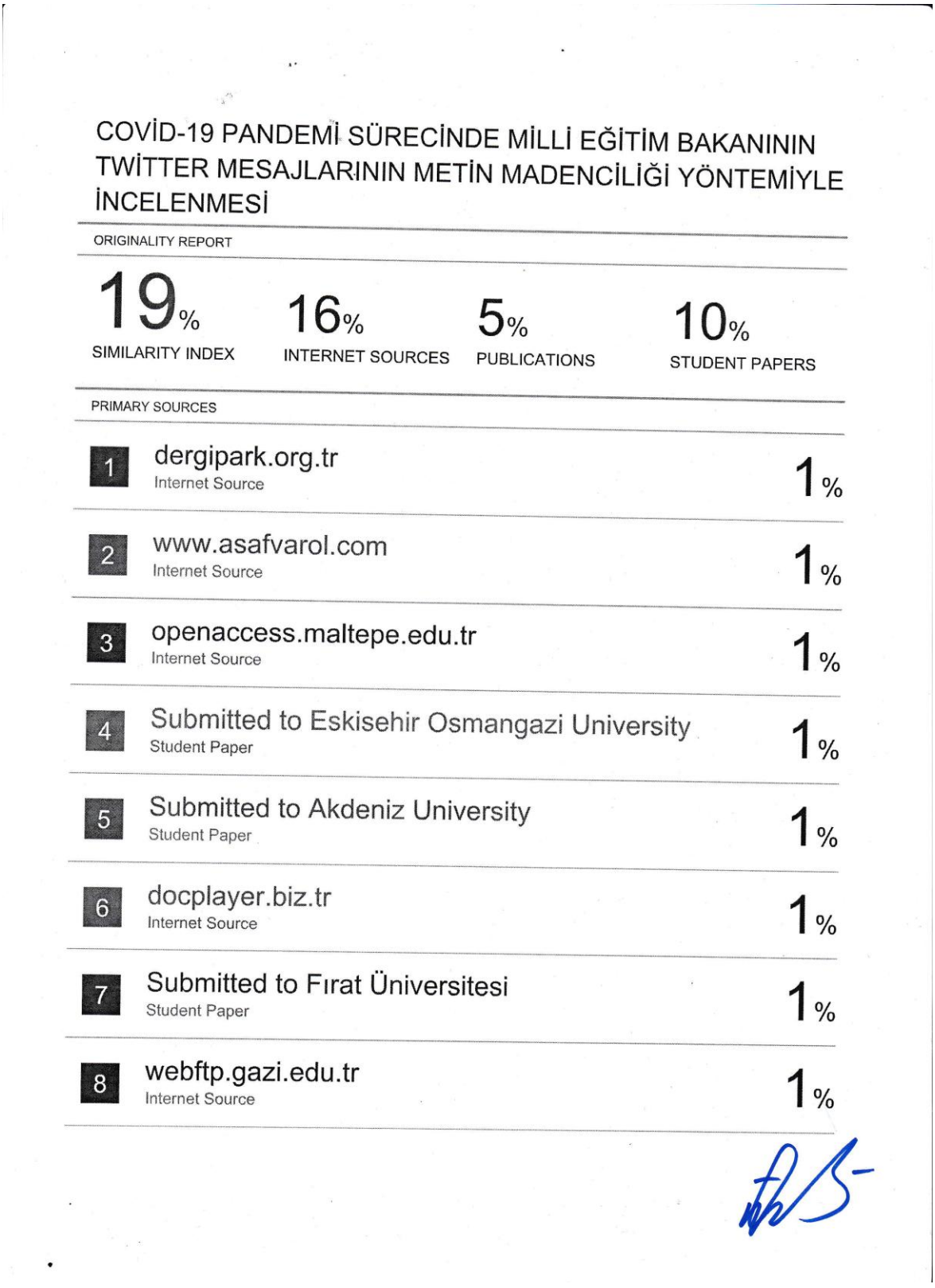
- Şanlıurfa Harran Cumhuriyet İlköğretim Okulu (2009-2010)
- Antalya Akseki Atatürk İlköğretim Okulu (2010-2015)
- Antalya Kepez Atatürk Anadolu Lisesi (2015-2016)
- Antalya Kepez Cengiz Topel İlkokulu (2016-Halen)

İletişim

E –Posta Adresi :eminekose2008@hotmail.com

Tarih :20/01/2021

İNTİHAL RAPORU



COVİD-19 PANDEMİ SÜRECİNDE MİLLİ EĞİTİM BAKANININ TWİTTER MESAJLARININ METİN MADENCİLİĞİ YÖNTEMİYLE İNCELENMESİ

ORIGINALITY REPORT

19%	16%	5%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dergipark.org.tr Internet Source	1%
2	www.asafvarol.com Internet Source	1%
3	openaccess.maltepe.edu.tr Internet Source	1%
4	Submitted to Eskisehir Osmangazi University Student Paper	1%
5	Submitted to Akdeniz University Student Paper	1%
6	docplayer.biz.tr Internet Source	1%
7	Submitted to Fırat Üniversitesi Student Paper	1%
8	webftp.gazi.edu.tr Internet Source	1%