MANAS Sosyal Araştırmalar Dergisi 2020 Cilt: 9 Sayı: 3
MANAS Journal of Social Studies 2020 Volume: 9 No: 3
ISSN: 1694-7215
Research Paper / Araştırma Makalesi

Metin Madenciliği ile Shakespeare Külliyatının İncelenmesi

Sadullah ÇELİK[1]

Öz

Metin madenciliği, doğal dil metninde yer alan yapılandırılmamış (metin) verilerin çeşitli yöntem, araç ve tekniklerin kullanılarak analiz edilmesidir. Bugün, kurum ve kuruluşların çoğu, veri ambarlarında ve bulut platformlarında büyük miktarda veri toplamakta ve depolamaktadır. Bu veriler, birden fazla kaynaktan gelen yeni verilerin gelmesiyle birlikte, üssel olarak artmaya devam etmektedir. Şirketlerin ve kuruluşların geleneksel araçlarla büyük miktardaki metin verilerini depolaması, işlemesi ve analiz etmesi zordur. Bugün, gelişen Tableau gibi yazılımlar sayesinde bu problemler ortadan kalkmıştır. Bu çalışmanın amacı; metin madenciliği yöntemi ile Shakespeare eserlerindeki kahramanları ve olay örgülerini istatistiksel olarak saptamak ve edebiyat alanında çalışanlara bazı öngörüler sağlamaktır. Bu amaçla çalışmada, Tableau yazılımı kullanılarak Google BigQuery'nin alt yapısında bulunan Shakespeare veri setine kelime frekansları, görselleştirme ve kümeleme analiz yöntemi uygulanmıştır. Kümeleme analizi sonucunda "Hamlet" kelimesinin tüm eserlerin merkezinde yer aldığı ve Hamlet'in Shakespeare'in en önemli eseri olduğu bulunmuştur. Ayrıca, "Romeo ve Juliet" eserinde sırasıyla; "Romeo", "Juliet" ve "Love" en çok kullanılan kelimeler olarak bulunmuştur. Elde edilen bu bulgulardan eserin ana karakterlerinin "Romeo" ve "Juliet", konusunun ise "aşk" olduğu sonucuna varılmıştır.

Anahtar Kelimeler: Metin madenciliği, Tableau, kelime frekansları, Görselleştirme, K-means kümeleme

The Investigation of Shakespeare Corpus with Text Mining

Abstract

Text mining is the analysis of unstructured (text) data in natural language by using various methods, tools and techniques. Today, the most institutions and organizations collect and store large amounts of data in data warehouses and cloud platforms. These data continue to increase exponentially with the arrival of new data from multiple sources. It is difficult for companies and organizations to store, process and analyze large amounts of text data with traditional tools. Today, these problems have disappeared thanks to software like Tableau. The aim of this study is; to determine the characters and plot patterns in Shakespeare dataset by using text mining method and to give some predictions to the literature researchers. In this study, word frequencies, visualization and clustering analysis method was applied to Shakespeare dataset which is in Google BigQuery infrastructure by using Tableau software. As a result of the clustering analysis, it was found that "Hamlet" was at the center of all the works and Hamlet was the most important work of Shakespeare. In addition, in the work of "Romeo and Juliet" respectively; "Romeo", "Juliet" and "Love" were found to be the most commonly used words. It is concluded that the main characters of the work are "Romeo" and "Juliet" and "love" is the subject.

Key Words: Text mining, Tableau, Word frequencies, Visualization, K-means clustering

Atıf İçin / Please Cite As:

Çelik, S. (2020). Metin madenciliği ile Shakespeare külliyatının incelenmesi. *Manas Sosyal Araştırmalar Dergisi, 9*(3), 1343-1357.

Geliş Tarihi / Received Date: 08.05.2019 **Kabul Tarihi / Accepted Date:** 03.02.2020

[1] Dr. Arş. Gör. - Aydın Adnan Menderes Üniversitesi Nazilli İktisadi ve İdari Bilimler Fakültesi, sadullah.celik@adu.edu.tr - ORCID: 0000-0001-5468-475X

Giriş

Yirmi birinci yüzyılda bilişim ve bulut teknolojilerinde görülen gelişmeler, üretilen verinin miktarında, yapısında ve hızında büyük artışa sebep oldu. Hemen hemen her tür kurum, kuruluş ve iş sektöründe, verilerin çoğu metin formatında olup elektronik veri tabanlarında depolanmaktadır. İnternet üzerinden dijital kütüphaneler, depolar ve bloglar, sosyal medya ağı ve e-postalar gibi büyük miktarda metin verisi vardır (Sagayam vd., 2012, s. 1443). Bu büyük hacimli verilerden değerli bilgiler elde etmek için uygun modelleri ve eğilimleri belirlemek oldukça zordur (Padhy vd., 2012). Geleneksel veri madenciliği araçları ile bilgi toplamak çok fazla zaman ve çaba gerektirdiğinden, bu araçlar metinsel verileri işlemede yetersiz kalmaktadır (Talib vd., 2016, s. 414). Bugün gelişen Hadoop, Spark, R, Python ve Tableau gibi yazılımlar sayesinde bu sorun büyük ölçüde ortadan kalkmıştır.

Günümüzde bilgi, çağdaş iş ortamındaki en önemli kaynaklardan birisidir. Müşterileri, çalışanları ve diğer paydaşları hakkında yeterli bilgiye sahip olmadan herhangi bir şirketin başarılı olması zordur. Her gün, şirketler anket sonuçları, tweet'ler, çağrı merkezi notları, telefon kayıtları, çevrimiçi müşteri yorumları, e-postalar, sosyal ağ paylaşımları, tıbbi kayıtlar ve diğer belgeler gibi çeşitli kaynaklardan yapılandırılmış ve yapılandırılmamış metinler almaktadır. Bu kaynakların, doğru metin analizi araçları kullanmadan anlaşılması kolay değildir. Metin analizini elle gerçekleştirmek mümkündür, ancak manuel işlem etkisiz kalmaktadır. Geleneksel sistemler anahtar kelimeleri kullanır ve e-postalarda, tweet'lerde, web sayfalarında ve metin belgelerindeki dili okuyamaz ve anlayamaz. Bu nedenle şirketler, büyük hacimli metin verilerini analiz etmek için metin analizi yazılımlarını kullanmaktadır. Bu yazılımlar kullanıcıların uygun şekilde davranabilmeleri için metin verilerinden bilgi edinmelerine yardımcı olmaktadır. Bugün, metin madenciliği yetenekleri arasına; şirketlerin pazarlama, satış ve müşteri hizmetleri operasyonlarına otomatik yanıtlar vermek için kullandığı Yapay Zekâ sohbet robotları ve sanal aracılar da girdi.

Metin madenciliği, doğal dil metninden anlamlı bilgiler elde etmeye çalışan yeni bir alandır. Metin madenciliği, anahtar kavramları ve temaları yakalamak için veri analizi süreci olarak tanımlamakta ve yazarların bu kavramları ifade etmek için kullandıkları belirli kelimeler veya terimler hakkında önceden bilgi sahibi olmadan gizli ilişkileri ve eğilimleri ortaya çıkarmaktadır (Bose, 2018). Metin madenciliği, yarı yapılandırılmış ve yapılandırılmamış formatta saklanan doğal dil metni ile ilgilenmektedir (Weiss vd., 2010). Metin veri madenciliği olarak da bilinen metin madenciliği, veri madenciliği, makine öğrenimi, istatistik ve doğal dil işleme algoritmalarını içermektedir. Bu algoritmalar sayesinde yapılandırılmamış verilerden yüksek kaliteli, yararlı bilgiler çıkarmaya çalışılmaktadır. Metin analizleri ile sıklıkla kullanılan metin madenciliği, yapılandırılmamış verilerin makine kullanımı için işlendiği bir araçtır (Bose, 2018). Metin madenciliği, veri madenciliğinin veya (yapılandırılmış) veri tabanlarından bilgi keşfinin bir uzantısı olarak da görülmektedir.

Metin madenciliği, verilerdeki kavramları, kalıpları, konuları, anahtar kelimeleri ve diğer nitelikleri tanımlayabilen yazılım tarafından desteklenen büyük miktarda yapılandırılmamış metin verilerinin araştırılması ve analiz edilmesi sürecidir. Metin analizi, veri kümeleri arasında sıralama yapmak için metin madenciliği tekniklerinin kullanılmasıyla etkinleştirilen bir uygulamadır (Rouse, 2018). Günümüzde metin madenciliği, büyük veri platformları ve büyük yapılandırılmamış veri kümelerini analiz edebilen derin öğrenme algoritmaları sayesinde veri bilimciler ve diğer kullanıcılar için çok kullanışlı hale gelmiştir (Linguamatics, 2018).

Günümüzde metin madenciliği birçok alanda yoğun olarak kullanılmakta ve bu kullanım her geçen gün daha da artmaktadır. Delibaş (2008), doğal dil işleme tekniklerini kullanarak Türkçenin biçimsel yapısını çözümlemeye çalışmıştır. Çalışmada, girilen bir Türkçe metnin yazım yanlışlarının bulunup bu yanlışların ayıklanması ve düzeltilmesi amaçlanmıştır. Çalışmadan elde edilen sonuçlar, daha önceki çalışmalara göre başarı oranının yüksek olduğunu göstermiştir. İlhan vd. (2008) doğal dil işleme ve metin madenciliği tekniklerinden yararlanarak kullanıcıdan alınan soruya en iyi yanıtı içeren metni bulmaya çalışmışlardır. Kullanıcıdan alınan soru, veri madenciliğindeki ön işleme aşamasından geçirilerek anahtar kelimeler belirlenmiş ve anahtar kelimenin metin içerisindeki önemine uygun cevap bulunmaya çalışılmıştır. Yapılan sorgulamalar sonucunda, veritabanında hazır bulunan anahtar kelimeler ile vektör uzayında gösterilen sorgu karşılaştırılmıştır. Elde edilen bulgulardan, veritabanında anahtar sözcüklerin hazır bulundurulmasının performansı arttırdığı sonucuna varılmıştır. Kaşıkçı ve Gökçen (2014) kullanıcılara e-ticaret sitelerinin bulunmasını kolaylaştırmak amacıyla yapmışlardır. Bu çalışmada kullanıcı tarafından belirtilen internet sitelerinin içeriği analiz edilmiş ve metin madenciliği kullanılarak bu sayfaların e-ticaret sitesi olup olmadığına karar veren bir uygulama geliştirilmiştir. Bu uygulama kullanıcıların e-ticaret sitelerinin bulunmasını kolaylaştırmayı hedeflemektedir. Naive Bayes ve k-En Yakın Komşu (KNN) sınıflandırma algoritmaları kullanılarak elde edilen bulgular karşılaştırılmıştır. Elde edilen sonuçlardan

Naive Bayes algoritmasının KNN algoritmasına göre daha iyi sonuç verdiği görülmüştür. Kılınç vd. (2016), KNN algoritmasının kullanarak akademik makalelerin kategorilere ayrılarak tasnif etme başarısını ölçmüşlerdir. Bunun için Research Gate üzerinde bulunan belirli akademik yayınların özetleri, R yazılımı kullanılarak elde edilmiş ve bu özetlerden bir veri seti oluşturulmuştur. Elde edilen sonuçlardan %96,67 oranında doğruluk değeri bulunarak makalelerin hangi kategorilere ait olduğu saptanmıştır. Arslan vd. (2017), personelin kurumsal e-posta hesaplarına gelen mesajlar incelenmiştir. Metin madenciliği ve sınıflandırma teknikleri kullanılarak yapılan çalışmada, e-posta sistemlerinin kurumsal uygulama ve iş süreçlerine dâhil edilmesi için yeni bir yöntem önerilmektedir.

Bu çalışmanın ikinci bölümünde, metin madenciliği hakkında bilgi verilerek metin madenciliğinde verinin işlenme aşamaları ve metin madenciliği uygulamaları hakkında bilgi verilmiştir. Üçüncü bölümde, çalışmada kullanılan; Shakespeare veri seti, kelime frekansları analizi, görselleştirme ve K-means kümeleme analizi yöntemi hakkında bilgi verilmiştir. Daha sonra, Tableau yazılımı kullanılarak Google BigQuery'nin alt yapısında bulunana Shakespeare veri setine bağlanılarak, kelime frekans analizi, görselleştirme ve kümeleme analizi yapılmıştır. Yapılan analizler sonucunda elde edilen sonçlar grafikler ve tablolar şeklinde verilmiştir. Sonuç bölümünde ise üçüncü bölümde elde edilen analiz sonuçları yorumlanmış ve metin madenciliğinin önemi hakkında bilgi verilmiştir.

Metin Madenciliği

Metin madenciliği aslında veri madenciliğine benzemektedir. Ancak metin madenciliğinde yapılandırılmış veri yerine metin verisi ile çalışılmaktadır. Metin madenciliğindeki ilk adım verileri düzenlemek ve yapılandırmaktır. Bu sayede veriler hem nitel hem de nicel analizlere tabi tutulabilmektedir.

Metin madenciliğinde amaç, metni kategorize etmek, kümelemek ve etiketlemek; veri kümelerini özetlemek; taksonomiler yaratmak ve kelime frekansları ve veri varlıkları arasındaki ilişkiler hakkında bilgi elde etmektir. Metin madenciliğinde analitik modeller, iş stratejilerini ve operasyonel eylemleri yönlendirmeye yardımcı olabilecek bulguları üretmek için kullanılmaktadır (Rouse, 2018). Bu bulguları elde etmek için ileri düzeyde matematik, istatistik, doğrusal cebir, optimizasyon, modelleme teknikleri ve gelişmiş yazılım araçlarını kullanabilme becerisine sahip olmak gerekmektedir.

Geçmişte, Doğal Dil İşleme algoritmaları temel olarak veri kümelerinde neleri arayacağımıza dair yön gösteren istatistiksel veya kural tabanlı modellere dayanıyordu. 2010'ların ortalarında ise daha az denetlenerek çalışan derin öğrenme modelleri, metin analizi ve diğer gelişmiş analitik uygulamalar için alternatif bir yaklaşım olarak ortaya çıkmıştır. Derin öğrenme, geleneksel makine öğrenimini desteklediğinden daha esnek ve sezgisel olan yinelemeli bir yöntem kullanarak verileri analiz etmek için sinir ağlarını kullanmaktadır (Rouse, 2018).

Günümüz de metin madenciliği, Google gibi bir arama motorundan bir metin içindeki öğeleri ve fikirleri anlamaya daha fazla imkân sağlamaktadır. Metin madenciliği, çok zor ya da zaman alıcı olan çok sayıda metin içerisinde bulunan kalıpları ve ilişkileri tanımlamayı sağlamaktadır.

Şekil 1'de görüldüğü üzere metin madenciliği; bilgi alma, bilgi çıkarımı ve veri madenciliği olmak üzere üç aşamadan oluşmaktadır.

Şekil 1. *Metin Madenciliği Aşamaları (Port, 2018)*

1. Aşama: Bilgi Alma: Metin veya veri madenciliğinin ilk aşaması bilgiyi almaktır. Bu aşama, önceden sayısallaştırılmış bir metinler kümesini tanımlamak için bir arama motorunun kullanılmasını veya yayınlarda ve makalelerde fiziksel metinlerin sayısallaştırılmasını gerektirebilmektedir. Bu da külliyat (corpus)'ın yararlı bir biçimde bir araya getirilmesi**ni** gerektirmektedir (Port, 2018).

2. Aşama: Bilgi Çıkarımı: İkinci aşama, anlamı tanımlamak için metnin işaretlenmesidir. Çoğu durumda bu, metin hakkında meta veriyi bir veritabanına (örneğin, yazar, başlık, tarih, baskı vb.) kaydetmeyi içerirken, bazı durumlarda ise metin içinde belirtilen tüm kişi adlarında veya konumlarda

anahtarlama içerebilmektedir. Bu süreç, arama motorlarının meta verileri oluşturanların önyargılarına dayanarak bilgi ve metinler arasındaki ilişkileri tespit etmelerini sağlamaktadır (Port, 2018).

3. Aşama: Veri Madenciliği: Veri madenciliği, yani veri kalıplarını çıkarmak için akıllı yöntemlerin uygulandığı önemli bir süreçtir. Bu son aşamada amaç, anlamı ortaya çıkaran ve araştırmacıların, keşfedilmesi zor olan yeni bilgileri keşfetmelerini sağlayan bilgi parçaları arasındaki ilişkiyi bulmaktır (Port, 2018).

Metin madenciliği, araştırmayı hızlandırabilen ve yeni sorular sormamıza ya da eskileri test etmemize imkan sağlayan bir araçtır (Port, 2018). Metin madenciliği, Tablo 1'dekine benzer ayırt edici görevleri içermektedir.

Tablo 1. *Metin Madenciliği Görevleri (Port, 2018)*

Görev	Anlam
Metin kategorizasyonu	Metinleri kategorilere ayırma
Metin kümeleme	Otomatik olarak alınan metin gruplarını anlamlı kategoriler listesine ayırma
Kavram/varlık çıkarma	Metin içindeki öğeleri kişilere, kuruluşlara, konumlara, parasal değerlere vb. gibi önceden tanımlanmış kategorilere yerleştirme ve sınıflandırma
Granüler taksonomiler	Birtakım nesneler olarak bilgilerin organizasyonu veya sınıflandırılması ve bir taksonomi olarak gösterilmesi
Duygu analizi	Kaynak materyallerde subjektif bilgilerin tanımlanması ve çıkarılması (örneğin, duygu, inançlar)
Belge özetlemesi	En önemli öğeleri içeren bir metnin kısaltılmış bir sürümünü oluşturma
Varlık ilişki modellemesi	Veri türleri arasındaki ilişkilerin otomatik olarak öğrenilmesi

Sonuç olarak, veri madencileri bir projenin başlangıcında neleri bulabileceklerini iyi bilmese de, metin madenciliği araçları sayesinde artık metin verisinin altında yatan benzerlikleri ve ilişkileri ortaya çıkarmak mümkündür. Örneğin, denetlenmeyen bir model, bir analistin yönlendirmesi olmaksızın metin belgelerinden veya e-postalardan bir grup konuya göre veriyi düzenleyebilmektedir.

Metin Madenciliği Uygulamaları

Metin analitiği sektörü son birkaç yılda yüksek bir büyüme ve gelişme kaydetti ve gelecek yıllarda da önemli bir büyüme ve gelişme kaydetmesi beklenmektedir. Metin analitiğinin bu kadar çok benimsenmesinin en önemli sebeplerinden biri, işletmelerde rekabetin doğasını arttırmak ve şirketleri katma değerli çözümler aramaya zorlamaktır. Rekabetin artması ve tüketici bakış açılarının değişmesi ile birlikte organizasyonlar, rekabet gücünü artırmak için müşteri ve rakip verilerini analiz edebilecek çözümlere önemli yatırımlar yapmaktadır. Ana veri kaynağı, e-ticaret platformları, sosyal medya, anket, kişisel blog, yayınlanan makaleler vb. kaynaklardır. Üretilen bu verilerin büyük kısmı yapılandırılmamış formattadır, bu da şirketlerin bireylerin yardımıyla bu verileri analiz etmelerini zorlaştırmakta ve maliyeti arttırmaktadır. Veri üretimindeki hızlı büyüme ile birlikte bu zorluk, yalnızca büyük hacimli metin verilerini işleyebilen değil aynı zamanda daha iyi karar vermede yardımcı olan analitik araçların geliştirilmesine de yol açmaktadır. Metin analizi yazılımı, kullanıcının farklı kaynaklardan edinilen büyük hacimli veri kümelerinden anlam çıkarmasını sağlamaktadır (Mane, 2018).

Metin madenciliği, işletme sorularını yanıtlamak ve günlük operasyonel verimliliklerini optimize etmek ve otomotiv, sağlık ve finans sektöründe uzun vadeli stratejik kararları iyileştirmek için kullanılmaktadır. Sınıflandırma, varlık çıkarma ve duyarlılık analizi gibi teknikler, büyük hacimli yapılandırılmamış verilerdeki öngörüleri, kalıpları ve eğilimleri tanımlamak için kullanılmaktadır.

Bugün dünya genelinde kullanılan birkaç metin madenciliği uygulaması vardır. Bu uygulamalar aşağıdaki gibidir.

Risk Yönetimi: İş sektöründeki başarısızlığın ana nedenlerinden biri, uygun veya yeterli risk analizinin olmamasıdır (Rai, 2018). Ancak, metin madenciliği risk analizi sorununun doğru çözülmesine yardımcı olmaktadır. Finans sektöründe, metin madenciliği teknolojisine dayalı Risk Yönetimi Yazılımı, büyük veritabanlarının eksiksiz bir şekilde yönetilmesini sağlamaktadır (Bose, 2018). Metin madenciliği teknolojileri binlerce metin veri kaynağından ilgili bilgileri toplayabildiğinden ve elde edilen bilgiler arasında ilişkiler oluşturabildiğinden, şirketlerin doğru bilgilere doğru zamanda erişmelerini sağlayarak tüm risk yönetimi sürecini geliştirmektedir (Rai, 2018).

Bilgi Yönetimi: Büyük veri hacimlerini yönetmek çoğu zaman kısa sürede özel bilgileri bulmayı zorlaştırmaktadır. Sağlık sektörü bu konunun klasik bir örneğidir. Sağlık sektöründeki uzmanlar yeni

ürünleri geliştirmek için çok büyük miktardaki bilgi ile (örneğin genomik ve moleküler tekniklere, örneğin klinik hasta verilerinin hacimlerine) araştırma yapmak zorundadırlar. Burada, metin madenciliğine dayanan bilgi yönetimi yazılımı "aşırı bilgi" sorunu için açık ve güvenilir bir çözüm sunmaktadır (Bose, 2018).

Dolandırıcılık Tespiti: Metin madenciliği teknolojilerinin desteklediği metin analizleri, metin biçimindeki verilerin çoğunluğunu toplayan alanlar için büyük fırsatlar yaratmaktadır. Sigorta ve finans şirketleri bu fırsatları değerlendirmektedir. Bu şirketler metin analizlerinin sonuçlarını ilgili yapılandırılmış verilerle birleştirerek, talepleri hızlı bir şekilde işleme koyabilmekte ve sahtekârlıkları tespit edip önleyebilmektedir (Rai, 2018).

*Müşteri Hizmetleri Servisi***:** Metin madenciliği ve Doğal Dil İşleme, müşteri hizmetleri uygulamaları için yaygın olarak kullanılmaktadır (Bose, 2018). Şirketler, anket, müşteri geri bildirimi ve müşteri çağrıları gibi çeşitli kaynaklardan gelen metinsel verilere erişerek genel müşteri deneyimini geliştirmek için metin analizi yazılımlarına yatırım yapmaktadır. Metin analizi, şirketin yanıt süresini azaltmayı ve şikayetlerin ele alınmasına yardımcı olmayı amaçlamaktadır. Metin analizi ayrıca daha hızlı ve otomatik müşteri tepkisi için kullanılmakta ve çağrı merkezi işlemlerine olan bağımlılığı önemli ölçüde azalmaktadır (Rai, 2018).

Sosyal Medya Analizi: Sosyal medya platformlarının performansını analiz etmek için özel olarak tasarlanmış birçok metin madenciliği yazılım paketi bulunmaktadır. Bu paketler haberlerden, bloglardan, e-postalardan, vb. çevrimiçi olarak oluşturulan metinleri izlemeye ve yorumlamaya yardımcı olmaktadır. Ayrıca, metin madenciliği araçları, markanızın sosyal medyadaki yayınlarını, beğenilerini ve takipçilerinin sayısını analiz ederek, markanız ve çevrimiçi içeriğinizle etkileşime giren kişilerin tepkisini anlamanıza imkân sağlamaktadır. Bu analiz, hedef kitlesi için "neyin sıcak olduğunu ve neyin olmadığını" anlamayı sağlamaktadır (Rai, 2018).

Spam Filtreleme: E-postalar çoğu kurumda hala en resmi iletişim yolu olarak kabul edilmektedir. Ama sadece yirmi birinci yüzyıl spam'ında artan karanlık bir yan vardır. Posta kutusundaki her on e-postadan en az dokuzu spam'dır. Spam'lar yalnızca boşluk doldurmakla kalmaz, aynı zamanda virüsler ve dolandırıcılık için bir giriş noktası görevi görmektedir. Şirketler, daha önce kullanılan anahtar kelime eşleştirmelerine kıyasla akıllı metin analizi kullanarak, daha fazla spam e-postasını filtrelemek ve kullanıcıya daha sağlıklı bir deneyim sunmak için giderek daha fazla spam filtrelemeyi kullanıyor (Williams, 2018).

Veri Seti ve Yöntem

Bu çalışmada yeni nesil programlardan olan Tableau programı kullanılarak Google'ın altyapısında bulunan BigQuery'e bağlanılarak buradaki Shakespeare veri setine kelime frekansları, görselleştirme ve $K-means$ kümeleme analizi yöntemi uygulanmıştır. Bu teknikler sayesinde büyük miktardaki karmaşık verilerin basit grafik/tablo veya resimler şeklinde kolay anlaşılır hale getirilebileceği gösterilmiş ve Shakespeare eserlerindeki ana karakterler ile olay örgüleri saptanmıştır.

Veri Seti (Shakespeare)

Shakespeare veri seti, Google'ın alt yapısında bulunan BigQuery (Büyük Sorgulama) de bulunmaktadır. Bu veri seti, William Shakespeare'in eserlerinden oluşmaktadır. Bu veri seti Shakespeare'in eserlerinde geçen her bir kelimenin kaç kez kullanıldığını gösteren bir kelime dizinini içermektedir. Ayrıca bu veri seti, 164656 (6.13 MB) satırdan oluşmakta ve her satır, kelime, kelimenin kullanım sayısı ve kelimenin kullanıldığı eser gibi değişkenleri içermektedir.

Kelime Frekansları (Word Frequencies)

Kelime frekansları**,** veri setinde kelimelerin kaç kez kullanıldığını göstermektedir. Kelime frekansları, veri setinde en sık kullanılan kelimelerden en az kullanılana kadar matris terimlerinden derleme kullanılarak gösterilir (Maria, 2018).

Bir kelime vektörü, ilgili kelimeyi anlamsal olarak temsil eden yüksek boyutlu bir uzayda bir konumdur. Bu konumda, benzer anlamları olan kelimeler birbirine daha yakındır. Dolayısıyla, eşanlamlı kelimeler neredeyse aynı vektöre sahiptir ve birbirine yakındır. Aynı kavram cümlelere uygulanabilirken, benzer cümleler yüksek boyutlu bir uzayda birbirine daha yakındır.

Metin analizinde ham kelime frekansını hesaplamak (wf) için aşağıdaki Log-Frekanslar dönüşümü kullanılmaktadır.

$$f(wf) = 1 + \log(wf), \qquad wf > 0 \quad için \tag{1}$$

Bu dönüşüm ham frekansların ve daha sonra yapılacak olan hesaplamaları etkilemektedir.

Bir kelimenin bir dokümanda kullanılıp kullanılmadığını belirlemek için ikili (binary) frekanslar dönüşümü kullanılmaktadır.

$$f(wf) = 1, \quad wf > 0 \quad için \tag{2}$$

Bu matris dönüşümünde dokümanda eğer kelime varsa 1 yoksa 0 değeri girilmektedir.

Terim frekansı (Term Frequency-TF), bir terimin bir doküman içerisindeki tekrar sıklığıdır. Her belgenin uzunluğu farklı olduğundan, bir belgenin uzun belgelerde daha kısa olanlardan çok daha fazla görünmesi muhtemeldir. Bu nedenle, TF genellikle belge uzunluğuna (yani, belgedeki toplam terim sayısına) normalizasyon yöntemi olarak bölünür. TF değerini hesaplamak için (3) formülü kullanılır.

$$TF_{(d,m)} = \frac{m\ kelimesinin\ d\ dokümanında\ geçme\ sayısı}{Dokümandaki\ toplam\ kelime\ sayısı} \tag{3}$$

Ters doküman frekansı (Invers Document Frequency-IDF) bir terimin tüm doküman koleksiyonu *(D)* içindeki önemidir (Coursehero, 2019). IDF'ye göre, terimin önemi, belge içerisindeki terimin kullanılma sıklığıyla doğru orantılıyken; tüm belge havuzu içerisindeki terimin kullanılma sayısıyla ters orantılıdır. Bir *D* belgesinde bulunan *i* teriminin ağırlığı (4) denklemindeki gibi hesaplanılır (Coursehero, 2019).

$$IDF_{(m)} = ln\frac{Vektör\ modelindeki\ toplam\ doküman\ sayısı}{İçerisinde\ m\ kelimesi\ bulunduran\ toplam\ doküman\ sayısı} \tag{4}$$

Düşük frekanslı terimlerin IDF skoru yüksek, yüksek frekanslı terimlerin IDF skoru düşüktür. Terim frekansı – ters metin frekansı (TF-IDF) değeri, az miktarda doküman içerisinde terim çok fazla geçiyor ise yüksek değer almaktadır. Eğer terim tüm dokümanlarda kullanılıyorsa TF-IDF değeri en düşük değerini almaktadır (Coursehero, 2019).

Metinde bulunan her bir kelime için TF ve IDF değerleri hesaplandıktan sonra (5) denklemindeki formül kullanılarak her bir kelimenin ağırlığı hesaplanır.

$$w_{(d,m)} = TF_{(d,m)} * IDF_{(m)} \tag{5}$$

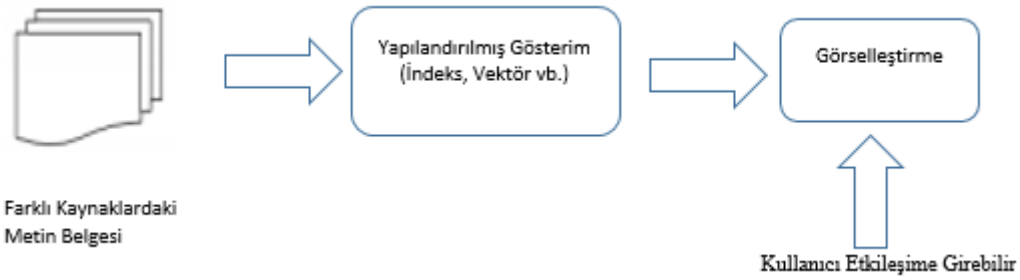
Yukarıdaki işlemler yapıldıktan sonra Doküman Terim Matrisi (Document Term Matrix-DTM) (6) denkelemi gibi oluşturulur.

$$\begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ \vdots \\ D_4 \end{array} \begin{array}{c} \begin{array}{ccccc} t_1 & t_2 & t_3 & \cdots & t_m \end{array} \\ \begin{bmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1m} \\ w_{11} & w_{12} & w_{13} & \cdots & w_{1m} \\ w_{11} & w_{12} & w_{13} & \cdots & w_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{m1} & w_{m2} & w_{m3} & \cdots & w_{mm} \end{bmatrix} \end{array} \tag{6}$$

Yukarıda verilen doküman terim matrisindeki D veri setinin içerisinde bulunan dokümanları, t dokümanlarda bulunan terimleri, w ise bu terimlerin ağırlıklarını göstermektedir (Bozyiğit, 2015).

Görselleştirme (Visualization)

Metin madenciliğinde görselleştirme yöntemi, ilgili bilgilerin keşfedilmesini iyileştirmekte ve basitleştirmektedir. Bireysel belgeleri veya belge gruplarını temsil etmek için, metin bayrakları belge gruplarını ve yoğunluk renklerini göstermek için kullanılmaktadır. Görsel metin madenciliği, büyük metin kaynaklarını görsel bir hiyerarşiye koymaktadır. Kullanıcılar, yakınlaştırıp ölçeklendirerek belge ile etkileşime girebilmektedir. Bilgi görselleştirme, terörist ağları tanımlamak veya suçlarla ilgili bilgi bulmak için hükümetler tarafından yoğun olarak kullanılmaktadır. Aşağıdaki Şekil 2, görselleştirme işleminde yer alan adımları göstermektedir (Gaikwad vd., 2014).



Şekil 2. *Görselleştirme (Gaikwad vd., 2014)*

Bilgi görselleştirme süreci üç aşamadan oluşmaktadır (Gaikwad vd., 2014):

1) Veri hazırlama aşaması, orijinal görselleştirme verilerinin ve orijinal veri kaynağının kararlaştırılması aşamasını içermektedir.

2) Veri analizi ve ayıklama aşaması, orijinal (kaynak) verilerden ihtiyaç duyulan görselleştirme verilerini analiz etme ve ayıklama aşamasıdır.

3) Görselleştirme haritalama aşaması, görselleştirme veri alanını görselleştirme hedefine eşlemek için belirli bir haritalama algoritması kullanılmaktadır.

Görselleştirme oluşturmak, verileri "görmenin" bir yoludur. Metin madenciliği görselleştirme, araştırmacıların belirli kavramlar arasındaki ilişkileri görmelerine yardımcı olabilmektedir. Veri görselleştirmesine örnek olarak, kelime bulutları, grafikleri ve haritaları vermek mümkündür.

Uzun bir metni okumak veya çok sayıda belgeye göz atmak, uzun zaman gerektirmektedir. Bunun yerine, sezgisel ve etkileşimli veri görselleştirmesi karar vericilerin analizin ortaya çıkardıklarını hemen anlamalarını ve daha sonra en çok ilgi çeken alanlara odaklanmalarını sağlamaktadır.

Metin madenciliği ve görselleştirme araçları, dokümanları, elektronik tabloları, raporları vb. açık tablolara veya grafiklere dönüştürerek analistlerin veri ve içeriğini kolayca keşfedip çalışmasına imkân sağlamaktadır.

Kümeleme (K-means)

Bölümlemeli kümeleme algoritmaların da giriş parametresi k alınarak n tane nesne k tane kümeye bölünür. Bu yöntem de iç içe geçmiş kümeler yerine tek-seviyeli kümeleri bulan işlemler yapılmaktadır (Jain vd., 1999). Bölünmeli kümeleme algoritmaları küme merkez noktasının kümeyi temsil etmesi esasına dayanmaktadır. Bu algoritmalar kolay uygulanabilir ve verimli olmasından dolayı iyi sonuçlar vermektedir (Işık, 2006:76).

$K-means$, en çok kullanılan denetimsiz öğrenme yöntemlerinden birisidir. Bu yöntemde her verinin sadece bir kümeye ait olmasına izin verilir (Evans, 2005). $K-means$ algoritması, n tane elemandan oluşan bir veri setini, giriş parametresi olarak alınan k tane kümeye bölümlemektedir. Burada amaç, bölümleme işlemi sonucunda elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerin

ise minimum olmasıdır. Bu çalışmada (7) denklemindeki Öklit uzaklığı formülü kullanılarak kümeleme yapılmıştır (Dinçer, 2006, s. 101).

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (7)$$

$K-means$ yönteminde küme merkezlerinin mutlaka bir başlangıcının olması gerekmektedir. Bu yöntemde rassal olarak seçilen k (küme sayısı) adet merkez nokta ile başlanılır. Bu yöntemde veri kümesinde bulunan her bir nokta kendisine en yakın olan merkez noktasının kümesine atanmaktadır.

Bu çalışmanın analizi Tableau programı kullanılarak yapılmıştır. $K-means$ yönteminde her küme, o kümedeki tüm noktaların ortalama değeri olan bir merkeze (centroid) sahiptir. Tableau'da, istenilen

sayıda küme sayısını bulmak için farklı k değerleri test edilmektedir. Bu işlem küme merkezlerinin değerleri sabit kalıncaya kadar devam edilir (Tableau, 2018).

Tableau, her bir k için $K-means$ kümelemesini hesaplamak için Lloyd'un algoritmasını kare Öklid uzaklıkları (squared Euclidean distances) ile kullanmaktadır. Her bir $k > 1$ için başlangıç merkezlerini belirlemek için ayırma işlemi kullanılmaktadır. Elde edilen kümelenme sadece küme sayısına bağlıdır ve sonuç olarak deterministiktir (Tableau, 2018).

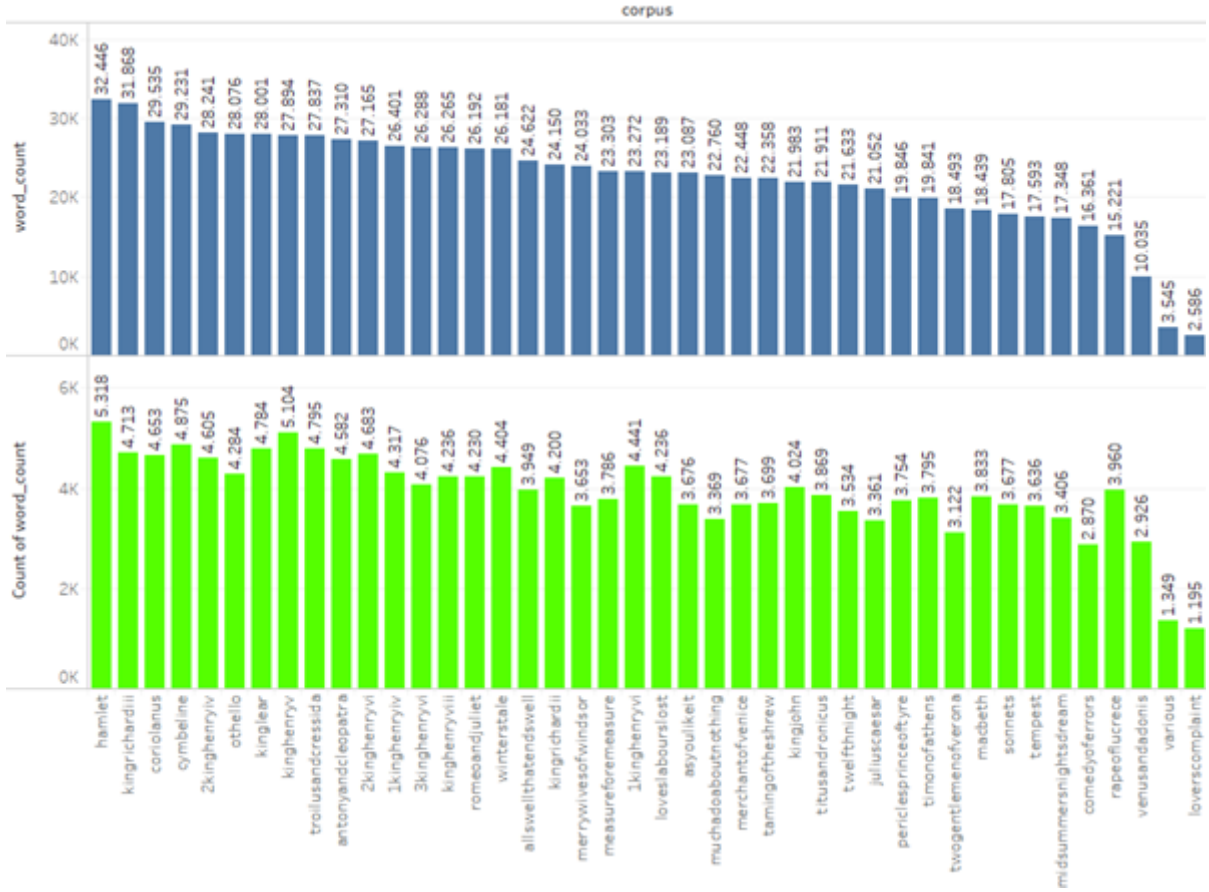
Tableau, küme kalitesini değerlendirmek için Calinski-Harabasz kriterini kullanmaktadır. Calinski-Harabasz kriteri şöyle tanımlanır:

$$\frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)} \tag{8}$$

Buradaki SS_B kümeler arası toplam varyansı, SS_W kümeler içi toplam varyansı, k kümelerin sayısını ve N toplam gözlem sayısını belirtmektedir. Ayrıca, (8) denklemindeki değer arttıkça, kümeler birbirine daha yakındır (küme içi varyans düşüktür) ve bireysel kümeler (küme arası varyans yüksek) daha belirgindir.

Bulgular

Metin madenciliği ve doğal dil işlemede temel soru, bir belgenin ne hakkında olduğunu ölçmektir. Bunu, belgeyi oluşturan kelimeleri analiz ederek yorum yapmak mümkündür. Bir kelimenin ne kadar önemli olabileceğini anlamak için terim sıklığına bakmak gerekmektedir. Bu amaçla çalışmada, öncelikle verilerin ön işleme aşaması yapılmıştır. Daha sonra doküman-kelime ve kelime-doküman matrisleri oluşturulmuştur. Elde edilen bu matrisler yardımıyla Shakespeare'in tüm eserlerinin kaçar kelime ve kelime çeşidinden oluştuğu Şekil 3'teki gibi bulunmuştur.

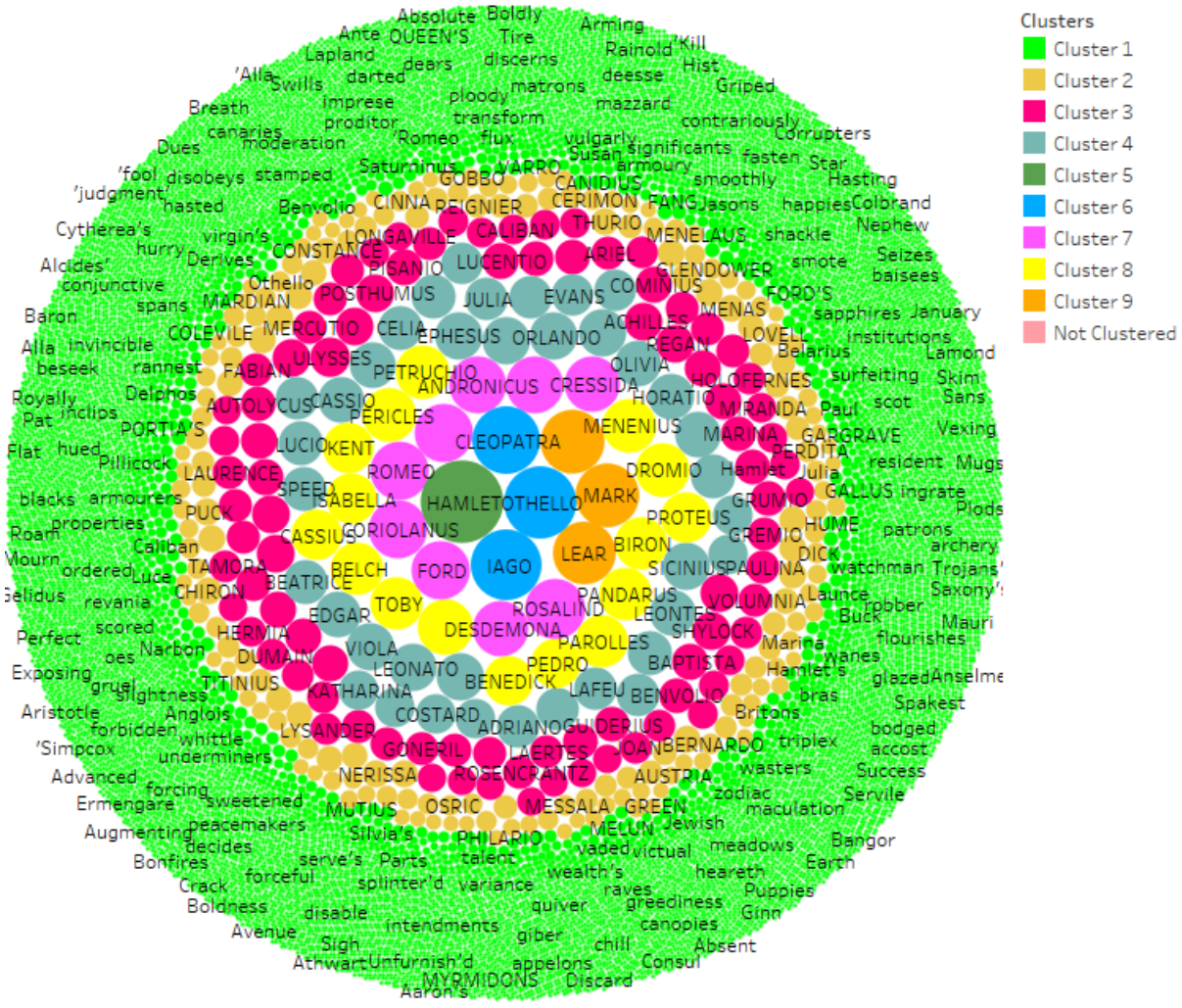


Şekil 3. *Shakespeare Eserlerinde Kullanıla Toplam Kelime Sayısı (Mavi) ve Kelime Çeşidi (Yeşil)*

Şekil 3'teki analiz sonuçlarına göre; Shakespeare'in en uzun eseri 32.446 kelime ile "Hamlet" tir. Yine en çok 5.318 farklı kelime kullanılan eser de "Hamlet"tir. Bir eserin kalitesinin ölçüsünü eserde kullanılan kelime çeşidi sayısı belirlemektedir. Bu bağlamda Hamlet, Shakespeare'in en çok kelime çeşidi kullandığı

eseri olduğundan hareketle yazarın söz varlığını (kelime hazinesi) yansıtan en önemli eseri olduğu söylenebilir.

Shakespeare'in tüm eserleri birlikte dikkate alındığında; yüksek frekanslı kelimelerin meydana getirdiği kelime bulutu Şekil 4'teki gibi bulunmuştur.



Şekil 4. *Shakespeare Eserlerinin Kelime Bulutu*

Şekil 4'teki her renk bir kümeyi göstermektedir. Şekil 4'te görüldüğü üzere "HAMLET" kelimesi 407 kez kullanılmış ve bu kelime Shakespeare'in tüm eserlerinin merkezinde konumlanmıştır. Bu bağlamda Shakespeare'in en iyi eserinin "HAMLET" olduğu söylenebilir.

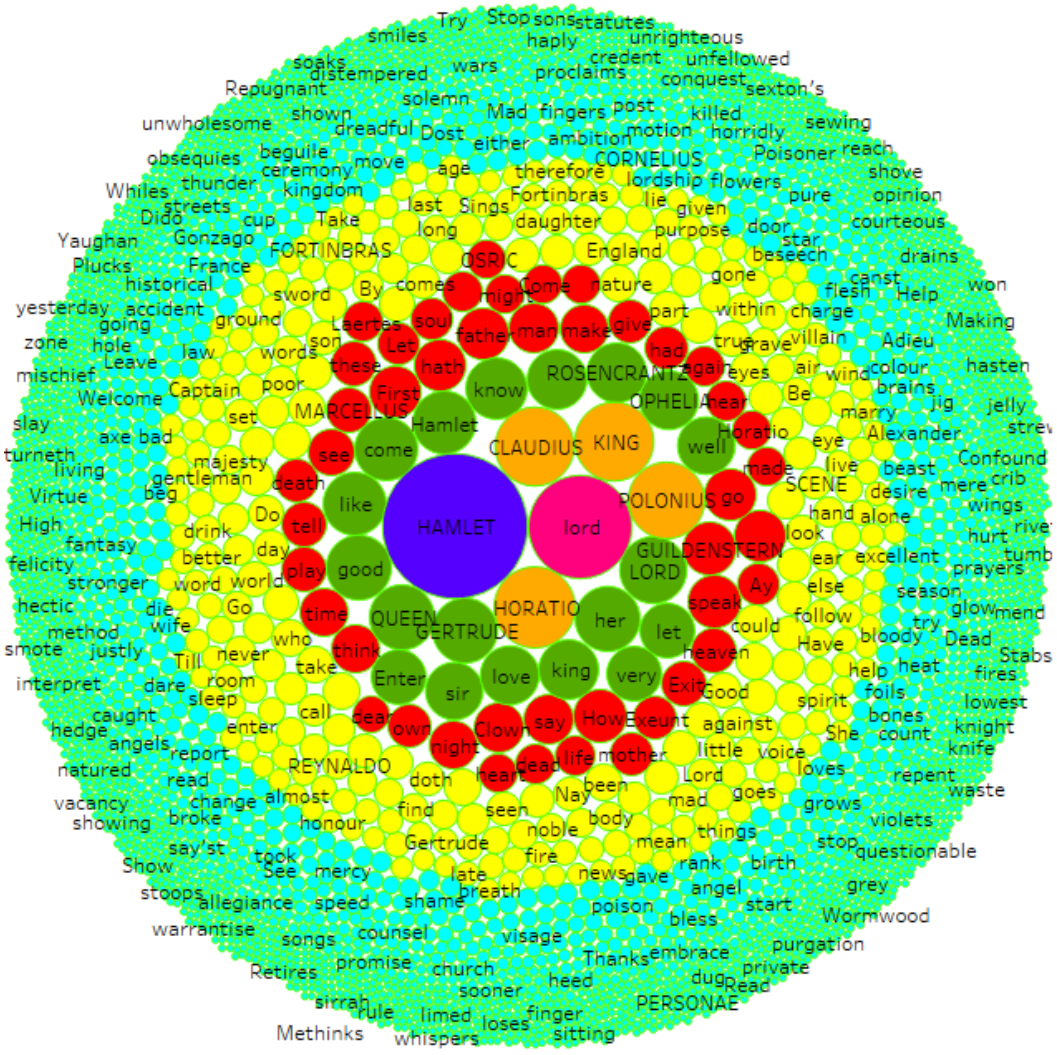
Kümeleme analizi sonucunda elde edilen kümelerin eleman sayıları ve merkezilik ölçüleri Tablo 2'de verilmiştir.

Tablo 2. *Kümelerin Merkezilik Ölçüleri*

Kümeler	Eleman Sayısı	Merkezilik Ölçüsü
Cluster 1	22704	0.00047907
Cluster 2	186	0.064754
Cluster 3	88	0.16743
Cluster 4	38	0.2922
Cluster 5	1	1.0
Cluster 6	3	0.72496
Cluster 7	9	0.49535
Cluster 8	16	0.40687
Cluster 9	3	0.5936

Tablo 2'deki sonuçlara göre; kümeleme analizi sonucunda toplam 9 küme bulunmuştur. Tablo 2 sonuçlarından en büyük kümenin 22775 kelimeden oluşan Cluster 1 ve en küçük kümenin 1 kelimeden (Hamlet) oluşan Cluster 5 kümesi olduğu bulunmuştur. Ayrıca merkezilik ölçüleri dikkate alındığında 1.0 merkeziliğe sahip olan Cluster 5 kümesinin tüm kümelerin merkezinde olduğu bulunmuştur.

Hamlet eserinde kullanılan kelimelerin meydana getirdiği kelime bulutu Şekil 5'teki gibi elde edilmiştir.



Şekil 5. *Hamlet Kelime Bulutu*

Şekil 5'e göre; kelime bulutunda en yüksek frekanslı kelime 407 kez tekrar eden "HAMLET" kelimesidir. Bu nedenle Hamlet'in eserde en baskın karakter olduğu söylenebilir. Aynı mantıkla eserin yardımcı karakterleri ve en çok geçen kelimelerden hareketle olay örgüsü belirlenerek eserin kurgusu, matematiksel ölçütlerle de saptanabilir. Ayrıca, Şekil 5'te görüldüğü üzere; kümeleme analizi sonucunda 7 farklı küme bulunmuş ve bu kümelerin her biri farklı renklerle gösterilmiştir.

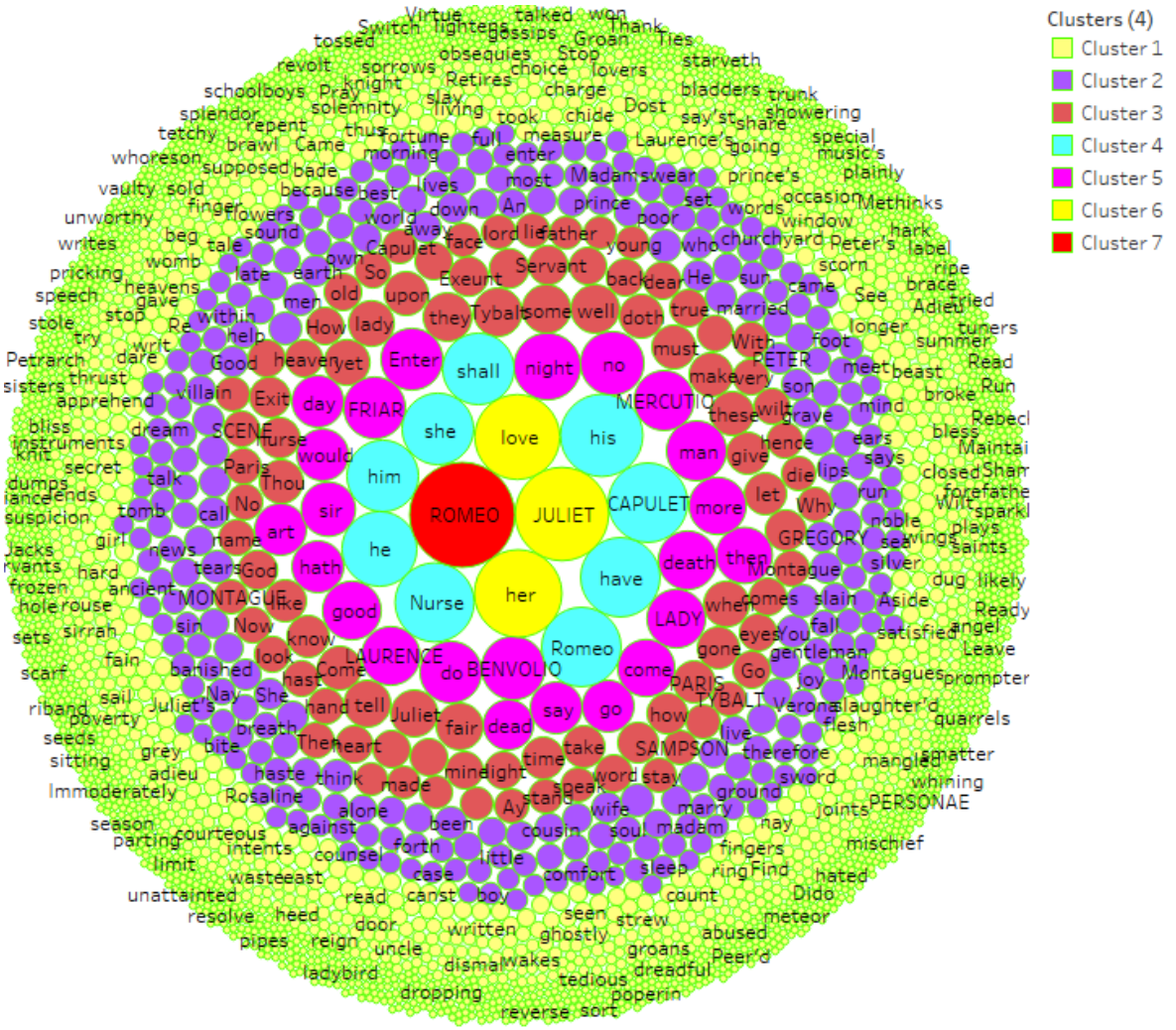
Kümeleme analizi sonucunda elde edilen kümelerin; eleman sayıları ve merkezilik ölçüleri Tablo 3'te verilmiştir.

Tablo 3. *Kümelerin Merkezilik Ölçüleri*

Kümeler	Eleman Sayısı	Merkezilik Ölçüsü
Cluster 1	4820	0.0019597
Cluster 2	315	0.033466
Cluster 3	45	0.09595
Cluster 4	19	0.17617
Cluster 5	4	0.29618
Cluster 6	1	1.0
Cluster 7	1	0.50739

Tablo 3'teki sonuçlara göre; kümeleme analizi sonucunda 7 küme bulunmuştur. Cluster 6 kümesinin merkezilik ölçüsü 1.0 çıktığından tüm kümelerin merkezinde yer almaktadır. Buradan "HAMLET" kelimesinin eserdeki ana karakter olduğu ve tüm olayların onun etrafında geçtiği söylenebilir.

Shakespeare'in "Rome And Juliet" eseri için kelime bulutu aşağıdaki gibi bulunmuştur.



Şekil 6. *Romeo And Juliet Kelime Bulutu*

Şekil 6, 7 farklı renkten oluşmakta ve bu renklerin her biri farklı bir kümeyi göstermektedir. Şekil 6'daki analiz sonuçlarına göre; Shakespeare'in "Romeo And Juliet" eserinde en çok kullanılan kelimeler sırasıyla Romeo (322) ve Juliet (203) dir. Şekil 6'ya göre "love" (134) kelimesi de eserde yoğun geçen kelimelerden birisidir. Bu bağlamda sadece bu analizden hareketle eserde olayların Romeo (322) ve Juliet'in (203) çevresinde geliştiği ve ana olayın "aşk" olduğu söylenebilir.

Kümeleme analizi sonucunda elde edilen kelime sayıları ve merkezilik ölçüleri aşağıdaki Tablo 4'te verilmiştir.

Tablo 4. *Kümelerin Merkezilik Ölçüleri*

Kümeler	Eleman Sayısı	Merkezilik Ölçüsü
Cluster 1	3723	0.0039961
Cluster 2	298	0.055089
Cluster 3	90	0.13633
Cluster 4	9	0.53377
Cluster 5	23	0.30158
Cluster 6	3	0.71078
Cluster 7	1	1.0

Tablo 4'teki sonuçlara göre; kümeleme analizi sonucunda toplam 7 küme bulunmuştur. Tablo 4'e göre; en büyük küme 22775 kelimeden oluşan Cluster 1 ve en küçük küme 1 kelimeden oluşan Cluster 7 kümesidir. Merkezilik ölçüleri dikkate alındığında ise yarıçapı 1.0 olan Cluster 7 kümesinin tüm kümelerin merkezinde olduğu bulunmuştur.

Sonuç ve Tartışma

Metin madenciliği, yeni bir bilgisayar bilimi alanı olup bugün birçok farklı alanda kullanılmaktadır. Metin madenciliği, yapılandırılmamış verilerin hem ilişkisel hem de niceliksel olarak katlanarak artmasından dolayı giderek daha fazla kullanılmaktadır. Metin madenciliği, büyük miktardaki yapılandırılmamış veriyi erişilebilir ve kullanışlı hale getirmek için kullanılmaktadır. Böylece yalnızca bir katma değer üretmek için kullanılmaz, aynı zamanda Risk Yönetimi Yazılımı ve Siber Suç uygulamaları için de kullanılmaktadır. Metin madenciliği sınıflandırma, varlık çıkarma, duygu analizi ve görselleştirme gibi teknikler sayesinde metin içerisinde saklı olan yararlı bilgiyi ortaya çıkarmaktadır. Özellikle iş dünyası, büyük hacimli yapılandırılmamış verilerdeki içgörüleri, kalıpları ve eğilimleri ortaya çıkarabilmek için metin madenciliği yöntemini çok sık kullanmaktadır.

Metin madenciliği, mevcut bilginin çok daha verimli bir şekilde analiz edilmesini sağlamaktadır. Bilgi alma kabiliyeti, literatür taraması sürecinde alan bilgisinin kapsamını sağlamak için harcanan zamanı otomatik olarak azaltmaktadır. Örneğin, bugün biyomedikal alanlardaki bilimsel yayınların hacmi büyük olduğundan, bir araştırmacının, belirli bir problem için tüm ilgili kaynakları tanımlamak için külliyat (corpus)'ı analiz etmesi birkaç yıl gibi uzun bir süreyi alabilmektedir. İlgili materyali tanımlamak için metin madenciliğinin kullanılması ise gereken süreyi büyük ölçüde azaltmaktadır. Ayrıca, eğer metin metnindeki belgeler çıkarılan semantik bilgilerle eklenmiş ve daha sonra yeniden kullanıma hazır hale getirilmişse, anahtar kaynaklar daha çabuk bulunabilmektedir. Bu verimlilik tasarrufu, araştırmalarda kullanılan çok çeşitli elektronik araştırmalar için de geçerlidir.

Bu çalışmada öncelikle, yeni nesil programlardan Tableau yazılımı kullanılarak Google BigQuery'nin altyapısında bulunan Shakespeare veri setine bağlanmıştır. Daha sonra Shakespeare'in tüm eserlerinde kullanılan toplam kelime sayısı ve kelime çeşidi sayısı analiz edilerek sonuçlar grafikler şeklinde sunulmuştur. Analiz sonucunda Hamlet'in Shakespeare'in en uzun (32.446) ve en çok kelime çeşidinin (5.318) kullanıldığı eseri olduğu bulunmuştur. Yine tüm eserlere birlikte kümeleme analizi uygulandığında ve görselleştirildiğinde "Hamlet" kelimesinin merkezde olduğu ve merkezilik ölçüsü 1.0 bulunmuştur. Elde edilen bu sonuçlardan Hamlet'in Shakespeare'in en iyi eseri olduğu sonucuna varılmıştır. Yine Hamlet eserine tek başına kümeleme analizi yapılıp görselleştirildiğinde "Hamlet" (407) kelimesinin en çok kullanıldığı ve merkezde yer aldığı bulunmuştur. Elde edilen bu sonuçlardan eserin olay kurgusunun "Hamlet" etrafında geçtiği sonucuna varılmıştır. Son olarak "Romeo and Juliet" için kümeleme ve görselleştirme teknikleri kullanıldığında eserde en çok kullanılan kelimelerin sırasıyla "Romeo" (322), "Juliet" (203) ve "love" (134) olduğu bulunmuştur. Bu sonuçlardan hareketle eserdeki olayların "Romeo" ve "Juliet" çerçevesinde geliştiği ve eserin ana konusunun "aşk" olduğu sonucuna varılmıştır.

Bu çalışmada metin madenciliği yönteminin hemen her alanda hatta matematiğe en uzak alan gibi görülen edebiyatta da kullanılabileceği ve buradan hareketle bestseller olarak nitelenen romanların kahramanları ve olay örgüleri saptanarak bu eserlerin oluşturulduğu matematiksel ilişkilerin tespiti ile edebiyat alanında çalışanlara türlü öngörüler sağlanabileceği gösterilmiştir. Metin madenciliği sayesinde yazarın seçtiği kelimeler ve hangi kelimelerin birbiriyle kullandığı gibi karakteristik özellikleri saptayarak yazarlar arasında kıyaslama yapmakta mümkündür. Bununla birlikte Tableau gibi yeni nesil gelişmiş yazılımlar sayesinde büyük metin verilerinin kolayca analiz edilerek metin hakkında genel bilgilere ulaşmakta mümkündür. Bu yazılımlar sayesinde büyük metinler üzerinde çalışanlar zamandan tasarruf sağlamaktadırlar. Ayrıca bu yazılımlar sayesinde büyük metinler görselleştirilerek okuyucular tarafından görülemeyen karmaşık ilişki ve desenler kolayca saptanabilmektedir. Çalışmanın bir başka sonucu ise metin madenciliğinin hemen hemen her alana uygulanabileceğini göstermektedir. Her ne kadar bu çalışmada bir edebi metin analizi yapılmış olsa da aslında metin madenciliği bugün hemen hemen her alanda kullanılmaya başlandı. Nitekim bugün sosyal medya platformlarında üretilen veriler metin madenciliği sayesinde analiz edilmektedir. Bu veriler satış tahmininde, pazarlamada ve film reytinglerini saptamada gibi birçok alanda kullanılmaktadır. Şirketler bu sayede kar oranlarını büyük oranlarda arttırmaktadırlar.

Etik Beyan

"Metin Madenciliği İle Shakespeare Külliyatının İncelenmesi" başlıklı çalışmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş; toplanan veriler üzerinde herhangi bir tahrifat yapılmamış ve bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

Kaynakça

Arslan, H., Kaynar, O. ve Yüksek, A. G. (2015). Kurumsal kolektif süreçler için e-posta iletilerinden görev keşfi ve gerçek zamanlı görev yönetim sisteminin geliştirilmesi. *Bilişim Teknolojileri Dergisi*, *10*(4), 381-388.

Azzalini, A. ve Scarpa, B. (2012). *Data analysis and data mining: An introduction*. OUP USA.

Bose, B. (2018). Techniques and Aplications of Text Mining. https://www.digitalvidya.com/blog/techniques-applications-text-mining/, (Erişim Tarihi: 10.06.2018).

Bozyiğit, F. (2015). *Analyzing source code and detecting similarities* (M.Sc Thesis). Dokuz Eylül University, Izmir.

Coursehero (2019). Terim frekanı tf bir doküman içerisinde bir. https://www.coursehero.com/file/p14lar0/Terim-Frekans, (Erişim Tarihi: 20.01.2019).

Delibaş, A. (2008). *Doğal dil işleme ile Türkçe yazım hatalarının denetlenmesi* (Yüksek Lisans Tezi). İstanbul Teknik Üniversitesi, İstanbul, Türkiye.

Dinçer, E. (2006). *Veri madenciliğinde K-means algoritması ve tıp alanında uygulanması* (Yüksek Lisans Tezi). Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli, 101s.

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

Evans, S., Lioyd, J., Stoddard, G., Nekeber, J. ve Samone, M. 2005. Risk factors for adverse drug events. *The Annals of Pharmacotherapy, 39,* 1161-1168.

Gaikwad, S. V., Chaugule, A. ve Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17).

Hotho, A., Nürnberger, A. ve Paaß, G. (2005, May). A brief survey of text mining. In *Ldv Forum, 20*(1), 19-62.

Işık, M. (2006). *Bölünmeli kümeleme yöntemleri ile veri madenciliği uygulamaları* (Yüksek Lisans Tezi). Fen Bilimleri Enstitüsü, Marmara Üniversitesi, İstanbul.

İlhan, S., Duru, N., Karagöz, Ş. ve Sağır, M. (2008). Metin madenciliği ile soru cevaplama sistemi. *Elektronik ve Bilgisayar Mühendisliği Sempozyumu (ELECO), Bursa*, 26-30.

Jain, A. K., Murty, M. N. ve Flynni, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

Kaşıkçı, T. ve Gökçen, H. (2014). Metin madenciliği ile e-ticaret sitelerinin belirlenmesi. *Bilişim Teknolojileri Dergisi*, *7*(1).

Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M. ve Özçift, A. (2016). KNN algoritması ve r dili ile metin madenciliği kullanılarak bilimsel makale tasnifi. DOI: 10.7240/mufbed.69674

Linguamatics (2018). What is NLP Text Mining?. https://www.linguamatics.com/what-is-text-mining-nlp-machine-learning, (Erişim Tarihi: 20.12.2018).

Mane, S. (2018). What are the application of text mining?. https://www.quora.com/What-are-the-applications-of-text-mining/answer/Sayali-Mane-16, (Erişim Tarihi: 23.04.2018).

Maria, L. (2018). Understanding and Writing your first Text Mining Script with R. https://towardsdatascience.com/understanding-and-writing-your-first-text-mining-script-with-r-c74a7efbe30f, (Erişim Tarihi: 11.01.2018).

Padhy, N., Mishra, D. ve Panigrahi, R. (2012). The survey of data mining applications and feature scope. *arXiv preprint arXiv:1211.5723*.

Port (2018). An introduction to text mining. https://port.sas.ac.uk/mod/book/view.php?id=554&chapterid=325, (Erişim Tarihi: 22.12.2018).

Rouse, M. (2018). Text mining (text Analytics). https://searchbusinessanalytics.techtarget.com/definition/text-mining, (Erişim Tarihi: 20.12.2018).

Rai, A. (2018). What is Text Mining: Techniques and Applications. https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/, (Erişim Tarihi: 13.06.2018).

Seker, S. E., Mert, C., Al-Naami, K., Ozalp, N. ve Ayan, U. (2013). Correlation between the economy news and stock market in Turkey. *International Journal of Business Intelligence Research (IJBIR)*, *4*(4), 1-21.

Seker, S. E. (2015). Metin Madenciliği (Text Mining). *YBS Ansiklopedi*, *2*(3), 30-32.

Sagayam, R., Sriniyasan, S. ve Roshni, S. (2012). A survey of text mining: Retrieval, extraction and indexing techniques. *International Journal of Computational Engineering Research*, *2*(5), 1443.

Tableau (2018). Find Cluster in Data. https://onlinehelp.tableau.com/v2018.3/pro/desktop/en-us/clustering.htm#HowItWorks, (Erişim Tarihi: 25.12.2018).

Talib, R., Hanif, M. K., Ayesha, S. ve Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, *7*(11), 414-418.

Weiss, S. M., Indurkhya, N., Zhang, T. ve Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media.

Williams, J. (2018). 9 Best Applications of Text Data Mining and Analysisis. https://www.promptcloud.com/blog/9-best-examples-of-text-mining-analysis, (Erişim Tarihi: 06.08.2018).

EXTENDED ABSTRACT

Nowadays, the rapid development of internet technologies has led to the rapid increase in the number of shares made online and the creation of big data sets (Dolgun et al., 2009). A significant number of these data sets contain unprocessed and un-analyzed data in unstructured form. Texts, photos, videos, audio files are some of these data. Machine learning methods have been developed for processing unstructured data. These methods are used in various fields such as bioinformatics, system identification, high energy physics, market analysis, image processing (Kılınç et al., 2016).

Text mining can be defined as the process of generating structured texts containing information from unstructured texts. In order to obtain meaningful information by processing the texts, some steps, such as data preprocessing and feature extraction, must be performed. After these steps, the unstructured data can be converted into a structural format processed by text mining and processed by computers (Hotho et al., 2005). In this way, valuable information in large amounts of data is discovered (Azzalini, & Scarpa, 2012). By using the meaningful information produced, various results can be accessed by the institutions or organizations. There are mathematical and statistical methods on the basis of text mining methods. Text mining is also used in different fields such as author recognition, text classification, idea mining, emotion analysis, keyword subtraction, caption (Kılınç et al., 2016).

Text mining is a data mining study that considers the text as a data source. In other words, it aims to obtain structured data via text. For example, it aims at studies such as the classification, segmentation, exclusion of texts, the production of class particles, emotional analysis, text summarization, and entity relationship modeling (Seker, 2015, p. 30). In order to achieve these objectives, information mining methods such as information retrieval, syllable analysis, word frequency distribution, pattern recognition, labeling, information extraction, data mining and even visualization are used to achieve these objectives (Seker et al., 2013).

Text mining studies are often text-based and work together with natural language processing. Natural language processing studies mainly involve studies based on linguistics knowledge under artificial intelligence. Text mining studies aim to reach more statistical results. During text mining studies, feature extraction is often done by using natural language processing (Seker, 2015).

In this study, using the Tableau program, which is one of the new generation programs, connected to BigQuery in Google's infrastructure, the word frequency, visualization and K-means clustering analysis method was applied to Shakespeare data set. These techniques have shown that large amounts of complex data can be made simple in the form of simple graphs/tables or pictures.

The study is primarily linked to the Shakespeare data set, which is included in the Google BigQuery's infrastructure, using the Tableau software from next generation programs. Then, the total number of words and the number of words used in all of Shakespeare's works were analyzed and presented in graphs. The results of the analysis were found to be Hamlet, using Shakespeare's longest (32.446) and most vocabulary (5.318). Again, when all clustering analysis is performed and visualized together, the word "Hamlet" is in the center and the centrality measure is 1.0. It is concluded that Hamlet was the best work of Shakespeare. It was also found that "Hamlet" (407) was the most used and centralized method when the clustering analysis was performed and visualized by Hamlet alone. It is concluded from this result that the plot of the work is passed around Hamlet. Finally, when clustering and visualization techniques were used for "Romeo and Juliet", it was concluded that the most commonly used words were "Romeo" (322), "Juliet" (203) and "love" (134), respectively. Based on these results, it was concluded that the events in the work developed within the framework of "Romeo" and "Juliet" and the main event in the work was "love".

In this study, it has been shown that the text mining method can be used in almost every field, even in the literature which is seen as the most distant area to mathematics, and the protagonists and the lattices of the novels which are known as bestseller are determined by determining the mathematical relations in which these works are created and thus, it can be provided to the employees in the field of literature. However, it is shown that thanks to the new generation of advanced software such as Tableau, large text data can be easily analyzed and general information about the text can be obtained. Thanks to this software, those who work on large texts will be able to access more information in less time. In addition, thanks to the developing technological software, large texts have been visualized and it has been shown

that complex relationship structures cannot be easily seen by the readers. Another result of the study shows that text mining can be applied to almost every field. Although a literary text analysis was conducted in this study, it is possible to use the text mining method in almost every field today. Therefore, the analysis of such text data is of great importance not only for those working in literature but also for those working in many sectors. Thanks to the use of developing technological software in this context, it will be inevitable that new changes will be experienced in many business areas and sectors in the coming period.