MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

YABANCI DİZİLERİN ALTYAZI VE TWITTER YORUMLARININ METİN MADENCİLİĞİ İLE İNCELENMESİ

YÜKSEK LİSANS TEZİ

Zahide ÇELİKSU

Anabilim Dalı: İstatistik

Programı: İstatistik

Tez Danışmanı: Yrd. Doç. Dr. Elif Özge ÖZDAMAR

EYLÜL 2017

MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

YABANCI DİZİLERİN ALTYAZI VE TWITTER YORUMLARININ METİN MADENCİLİĞİ İLE İNCELENMESİ

YÜKSEK LİSANS TEZİ

Zahide ÇELİKSU

Anabilim Dalı: İstatistik

Programı: İstatistik

Tez Danışmanı: Yrd. Doç. Dr. Elif Özge ÖZDAMAR

EYLÜL 2017

Zahide ÇELİKSU tarafından hazırlanan YABANCI DİZİLERİN ALTYAZI VE TWITTER YORUMLARININ METİN MADENCİLİĞİ İLE İNCELENMESİ adlı bu tezin YÜKSEK LİSANS tezi olarak uygun olduğunu onaylarım.

Yrd. Doç. Dr. Elif Özge Özdamar

Tez Yöneticisi

Bu çalışma, jürimiz tarafından İSTATİSTİK Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

Başkan : Yrd. Doç. Dr. Elif Özge Özdamar

Üye : Doç. Dr. Semra Erpolat Taşabat

Üye : Prof. Dr. Müjgan Tez

Bu tez, Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygundur.

iii

ÖNSÖZ

İstatistik biliminin inceliklerini daha iyi kavrayabilmek adına çıktığım bu yolda bitiş çizgisini görmüş, yüksek lisans eğitimimin sonuna gelmiş bulunuyorum. Bu zorlu yolda;

İçimdeki potansiyeli ortaya çıkaran, benimle birlikte akşamlarını çalışarak geçiren değerli danışmanım Yrd. Doç. Dr. Elif Özge Özdamar başta olmak üzere;

Kelimeler üzerinde çalışırken her bir sorumu yanıtlayan ve yardım çağrılarıma her zaman olumlu yanıt veren İTÜ Bilgisayar ve Bilişim Fakültesi asistanlarından ve İTÜ NLP sözlük geliştiricilerinden biri olan değerli Tuğba Pamay'a;

Twitter verilerini test etmede ve tweetleri düzenlemede yardımlarını esirgemeyen Migros Genel Müdürlük A.Ş.'nin Veri Ambarı ve İnşaat departmanlarında bulunan değerli çalışma arkadaşlarıma;

İhtiyaç duyduğum her anda desteğini sonuna kadar gösteren başta Zeynep Ece Kızıltepe ve MehtapYentur olmak üzere bütün arkadaşlarıma;

Aldığım her kararda arkamda duran ve beni cesaretlendiren aileme teşekkür ederim.

Zahide Çeliksu

YABANCI DİZİLERİN ALTYAZI VE TWITTER YORUMLARININ METİN MADENCİLİĞİ İLE İNCELENMESİ

ÖZET

Bu çalımada amaç, belirlenen yabancı dizilerin Türkçe altyazı ve twitter yorumlarını açık kaynak R programı ile metin madenciliği açısından incelemektir.

Çalışmanın birinci bölümünde veri madenciliği, veri madenciliği uygulama alanları, veri madenciliği modelleri ve veri işleme sürecinden bahsedilmiştir.

İkinci bölümde metin madenciliği ve metin madenciliği metodolojisine yer verilmiş ve metin madenciliğinde en sık tercih edilen kümeleme analizi ayrıntılı olarak ele alınmıştır. Bu çalışmada metin madenciliğinin "bag of words" yaklaşımı ele alınmıştır.

Çalışmanın üçüncü bölümünde ise uygulamaya yer verilmiştir

Son bölüm olan dördüncü bölümünde ise elde edilen sonuçlar paylaşılmıştır.

TEXT MINING OF FOREIGN TV SERIES SUBTITLES AND TWITTER COMMENTS

ABSTRACT

The main purpose of this study is to examine Turkish subtitles and twitter comments of selected foreing TV series with open source programming languageR, in terms of text mining.

The first part, defitinion of data mining, data mining application areas, data mining models and data processing are mentioned.

In the second part, text mining and text mining methodology is given and the most preferred clustering analysis in text mining is discussed in detail. In this work, the "bag of words" approach to text mining is discussed.

The third part of the study is the application chapter.

In the forth part the results obtained are shared.

İÇİNDEKİLER

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

1. GİRİŞ

Son yıllarda gelişen teknoloji ve buna bağlı olarak insanlığın değişen ihtiyaçları; ölçülen, depolanan ve analiz edilen verinin yapı ve boyutunun değişim göstermesine neden olmuştur. Yüzyılın başında bilimsel çalışmalar en fazla birkaç yüz gözlemden oluşan veri setleri ile gerçekleştirilirken, günümüzde bankacılık işlemleri, gsm operatörleri, dna dizilimleri ya da uydu verileri ile "büyük veri" kavramı insanların gündelik hayatlarında karşılaştıkları sıradan bir kelime halini almış ve büyük verinin analizi ile gerçekleştirilen son ürünler giderek artan bir ivme ile kullanılır hale gelmiştir.

Önceden kiloByte boyutundaki veri setleri dijital olarak tek bir tabloda depolanabilirken, günümüzde yottaByte boyutuna ulaşılmış, ve buna bağlı olarak da verinin depolanma sistemleri gelişme göstermiştir.

Üretilen ve depolanan verinin giderek büyümesi, verinin modellenmesi ve veri içindeki "saklı" bilgiye ulaşılmasına aracı olan tekniklerin de değişmesine neden olmuştur. Farklı kaynaklarda depolananmış, statik olmayan ve büyük boyutlu veri, günümüzde "Veri Madenciliği", "Makine Öğrenmesi" ve "Büyük Veri" alanlarını ortaya çıkarmıştır.

Günümüzde şirketler, bünyelerinde ürettikleri ve çok sayıda veri tabanından depolayabildikleri veriden anlamlı, doğru, güvenilir ve hızlı bir şekilde bilgiye ulaşmak gerektiği için veri madenciliği yöntemlerine başvurulmaktadır. Veri madenciliği ile kurum, kuruluş veya kişisel verilerden hızlı bir şekilde istenilen bilgi elde edilebilecek hale gelmişlerdir.

Veri Madenciliği ile büyük miktardaki veriler içerisinden önemli olanlarını bulup çıkarılması için verilerin yapısallaştırılarak işlenmesi gerekmektedir. Ancak günümüzde veri tabanları sadece sayısal veriden oluşmayabilmektedir. Günümüzde sayısal verilerin yanısıra ses, metin, fotoğraf, video gibi mecraların da analiz edilmesinin bir gereklik olduğu aşikardır.

Bu çalışma; müşteri şikayeti, sosyal medya, görüş, spam olarak değerlendirilebilecek epostalar veya her türlü dijital ya da dijitalleştitilmiş dokümanı analiz ederek bilgi çıkarmayı amaçlayan araştırıcı ve analistlerin kullandığı "Metin Madenciliği" ni ele almaktadır. Metin Madenciliği, araştırma konusu olan verinin sadece doküman ya da dokümanlardan oluştuğu ve "metin" üzerinde analizlerin gerçekleştirildiği bir alandır.

Veri Madenciliği uygulamaları çoğunlukla yapısal verileri üzerinde gerçekleştirildiğinden, sadece metinden oluşan ve yapısal olmayan verilerin yapısal verilere dönüştürülmesi gerekmektedir. Bu durumda metin madenciliği devreye girmektedir. Metin madenciliği metin formatındaki verileri kullanarak yapısal olmayan verileri yapısallaştırır ve metinlerden nümerik değerler elde ederek bilgiye ulaşılmasını sağlar.

Çalışmanın birinci bölümünde veri madenciliği, veri madenciliği uygulama alanları, veri madenciliği modelleri ve veri işleme sürecinden bahsedilmiştir.

İkinci bölümde metin madenciliği ve metin madenciliği metodolojisine yer verilmiş ve metin madenciliğinde en sık tercih edilen kümelemeanalizi ayrıntılı olarak ele alınmıştır. Bu çalışmada metin madenciliğinin "bag of words" yaklaşımı ele alınmıştır. Bu yaklaşımda metin çeşitli ön işlemlerden geçirileren kelime kökleri elde edilmekte, ve bu köklerin metin içerisinde tekrarlanma sıklığı olan frekansları üzerinde uygun olan veri madenciliği analizleri yapılmaktadır. Çalışmada bu yaklaşımının seçilmesinin nedeni, R programının Türk Dili'ni desteklememesi ve daha da önemlisi metin madenciliğinde diğer yaklaşım olan Doğal Dil İşlemenin Türk Dili üzerinde gerçekleştirilmesi için Türkçe bir sözlüğün tam anlamıyla oluşturumuş olmamasından kaynaklanmaktadır.

Çalışmanın üçüncü bölümünde ise uygulamaya yer verilmiştir. Günümüz TV izleyici kitlesi, özellikle genç kitle, Türk dizilerine kıyasla yabancı dizi izlemeyi tercih etmektedir. Bunun başlıca sebeplerinden biri dizi sürelerinin yerli dizilere göre oldukça kısa olmasıdır. 40-45 dakikalık süreler arasında değişkenlik gösteren diziler farklı türler ve farklı konularda pek çok dizi alternatifi mevcut olması diğer tercih sebebi olarak sayılabilir. Hem dizilere harcanan emek hem kullanılan bütçe, hem de oyunculuklar açısından izleyiciyi tatmin etmeleri, yerli dizilerin ise kısır döngü haline gelen benzer senaryoları ve uzun reklam araları izleyi tercihlerini

belirlemektedir. Bu bölümde yabancı dizilerin altyazıları ve Twitter'den elde edilen yabancı dizi yorumları metin madenciliğinde en sık kullanılan yöntem olan kümelene analizi ile irdelenmiştir. Açık kaynak R programının Türkçe konuşma dili üzerindeki etkinliğini araştırmak amaçlanmıştır.

Son bölüm olan dördüncü bölümünde ise elde edilen sonuçlar paylaşılmıştır.

2. VERİ MADENCİLİĞİ

Veri madenciliği, büyük miktardaki gözlenmiş verilerden kuralların, örüntülerin ve modellerin ortaya çıkarılmasıdır. Bir başka ifade ile veri madenciliği, veri tabanları veya veri ambarlarında yer alan yığın veri içindeki gizli örüntüleri ve ilişkileri bulmak için istatistiksel algoritmaları ve yapay zeka yöntemlerini kullanan karmaşık bir veri arama yeteneği olarak tanımlanabilir. Veri madenciliği; aynı zamanda bilgisayar bilimini, makine öğrenmesini, veri tabanı yönetimini, matematiksel algoritmaları ve istatistiği birleştiren disiplinlerarası bir alandır. Veri madenciliğini farklı araştırmacılar tarafından,

- Veri madenciliği büyük veri kümeleri içinde saklı olan, faydalı bilgilerle genelde tahmin edilemeyen eğilim ve ilişkilerin keşfedilmesi için bir eleme faaliyetidir [1].
- Veri madenciliği veritabanı sahibi için büyük miktardaki veriden bilinmeyen ilişki ve düzenlerin keşfedilmesi ile faydalı ve net sonuçlar elde etmeyi hedefleyen seçme, araştırma ve modelleme sürecidir [2].
- Veri madenciliği, bilinmeyen ilişkilerin bulunması ve verinin değişik şekillerde özetlenmesi için gözlemsel verilerin, veri sahibi için anlaşılır ve yararlı olacak şekilde analiz edilmesidir [3]

olarak ifade edilmektedir.

Veri madenciliği, veritabanındaki bilgi keşfi sürecinin bir adımıdır. Bilgi keşfi sürecindeki adımlarını şu şekilde sıralayabiliriz. Bu süreçler interaktif olup gerektiği durumlarda sıralaması değişmektedir.

1. **Veri temizleme:** Gürültülü ve tutarsız verileri çıkarmak
2. **Veri bütünleştirme:** Farklı veri kaynaklarını birleştirmek
3. **Veri seçme**: Uygulanacak analizle ilgili olan verileri belirlemek
4. **Veri dönüşümü:** Verinin veri madenciliği tekniğinden kullanılabilecek hale dönüşümünü gerçekleştirmek

5. **Veri madenciliği**: Verideki örüntülerini yakalayabilmek için teknikleri uygulamak
6. **Bilgi sunumu:** Mmadenciliği yapılmış olan elde edilmiş bilginin kullanıcıya sunumunu gerçekleştirmek.

Veri madenciliğinde örüntü tanıma faaliyetleri üç temel sınıfta toplanabilir. Bunlar; keşif (discovery), tahmin edici modelleme (predictive modelling) ve adli analizdir (forensic analysis). Keşif, bir veri yığınındaki gizil örüntüleri önceden belirlenmiş bir fikir veya hipotez olmadan ortaya çıkarma sürecidir. Başka bir ifade ile verilerin içinde saklı olarak bulunan, hangi ürünlerin birlikte satıldığı veya hangi grup müşterilerin hangi zaman aralıklarında bir hizmeti kullandıkları gibi davranışları ortaya çıkarmaya yarar. Tahmin edici modelleme, ortaya çıkardığı örüntüler ile geleceği tahmin etmede kullanılmaktadır. Başarılı bir kredi verme işlemi veya bir hata olasılığı belirleme işlemi tahmin edici modelleme ile gerçekleşebilmektedir. Adli analiz ise ortaya çıkarılmış örüntülerin, kural dışı veya anormal veri elemanlarını bulmak için kullanılması süreci olarak tanımlanabilir[4].

Veri madenciliği, kavramsal olarak 1960'lı yıllarda, bilgisayarların veri analiz problemlerini çözmek için kullanılmaya başlamasıyla ortaya çıkmıştır. Veri madenciliği kavramı ortaya atılmadan önce, veri taraması (data dredging) ve veri yakalanması (data fishing) gibi isimler kullanılmaktaydı. 1960'lı yıllarda veri toplama ile başlayan bu süreç, 1970' lerde veritabanlarının oluşturulması ile devam etmiştir. 1990'lı yıllara gelindiğinde ise veri madenciliği ismi, Rakesh Aggrawal öncülüğünde bazı bilgisayar mühendisleri tarafından ortaya atılmıştır. Bundan sonra ise veri madenciliğine çeşitli yaklaşımlar getirilmeye başlanmıştır. Bu yaklaşımların kökeninde istatistik, makine öğrenimi (machine learning), veritabanları, otomasyon, pazarlama, araştırma gibi disiplinler ve kavramlar bulunmaktadır [5].

2.1 Veri Madenciliği Uygulama Alanları

Veri madenciliğinin uygulama alanları oldukça geniştir. Farklı bilim dallarında ve sektörlerde uygulama alanları, analiz edilen verinin yapısı ve boyutuna göre farklılaşmakdır. Veri madenciliğinin uygulama alanlarını kısaca aşağıdaki şekilde özetlenebilir:

- Pazarlama; müşterilerin satın alma alışkanlıklarının belirlenmesi, müşterilerin demografik özellikleri arasındaki bağlantıların bulunması, posta kampanyalarında cevap verme oranının artırılması, mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması, pazar sepeti analizi, müşteri ilişkileri yönetimi, müşteri değerlendirmesi, satış tahmini, çapraz satış analizi, mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerin oluşturulması.
- Bankacılık; farklı finansal göstergeler arasında gizli korelasyonların bulunması, kredi kartı dolandırıcılıklarının tespiti, kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi, kredi taleplerinin değerlendirilmesi, müşteri dağılımı, usulsüzlük tespiti, risk analizleri.
- Sigortacılık, yeni poliçe talep edecek müşterilerin tahmin edilmesi, sigorta dolandırıcılıklarının tespiti, riskli müşteri örüntülerinin belirlenmesi.
- Perakendecilik, satış noktası veri analizleri, alış-veriş sepeti analizleri, tedarik ve mağaza yerleşim optimizasyonu, hisse senedi fiyat tahmini, genel piyasa analizleri, alım-satım stratejilerinin optimizasyonu.
- Endüstri, kalite kontrol analizleri, lojistik, üretim süreçlerinin optimizasyonu olarak belirtilebilir [6].

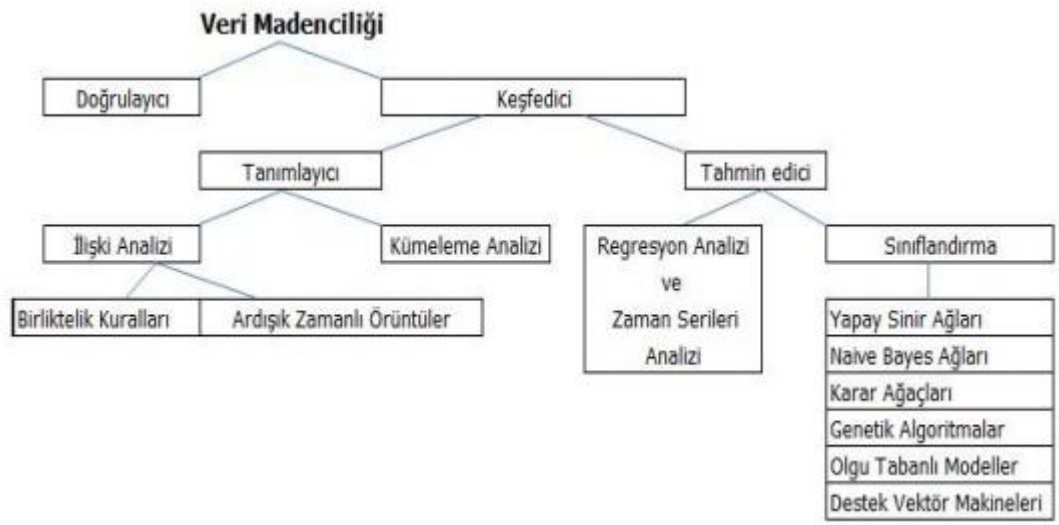
2.2 Veri Madenciliği Süreci

Veri madenciliği, aynı zamanda bir süreçtir. Gerçek dünyada verilerin büyük miktarlarda olmaları, kayıp olan veriler, yanlış işlenmiş ya da kodlanmış verilerin olması, hatalı ya da sapan değerler içeren gürültülü verilerin olması gibi nedenler dolayısıyla kaliteli ve kullanışlı veri madenciliği sonuçları elde edebilmek için veri madenciliği süreçleri uygulanmadan önce veri işleme tekniklerinin uygulanmasına ihtiyaç duyulur. Veri yığınları arasında, soyut kazılar yaparak veriyi ortaya çıkarmanın yanı sıra, bilgi keşfi sürecinde örüntüleri ayrıştırarak süzmek ve bir sonraki adıma hazır hale getirmek de bu sürecin bir parçasıdır. Üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda ne kadar etkin olursa olsun hiçbir veri madenciliği algoritmasının fayda sağlaması mümkün değildir. Bu sebeple, veri madenciliği sürecine girilmeden önce, analizlerin ilk şartı, iş ve veri özelliklerinin detaylı analiz edilmesidir[5,6,7].

- **Problemin tanımlanması:** Veri madenciliği çalışmalarında en büyük şart, problemin tanımlanması olarak bilinmektedir. Problemdeki amacın net bir şekilde ifadesinin yapılması gerekmektedir. Problemin hangi işletme amacı için yapılacağının ve elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceğinin tanımlanması en önemli aşamadır.
- **Verilerin hazırlanması:** Problem durumunun hazırlanmasından sonraki aşama olan verilerin hazırlanması; çalışmaya temel oluşturacak son verilere dönüştürülme aşamasıdır. Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. Verilerin hazırlanması, "toplama", "değer biçme", "birleştirme ve temizleme", "örneklem seçimi" ve "dönüştürme" aşamalarından oluşmaktadır.
- **Modelin kurulması ve değerlendirilmesi:** Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.
- **Modelin kullanılması:** Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir.
- **Modelin izlenmesi:** Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve yeniden düzenlenmesini gerektirecektir.

2.3 Veri Madenciliği Modeller

Literatürde veri madenciliğinde kullanılan modeller farklı şekillerde sınıflandırılabilmektedir. En temel anlamda bu modeller doğrulayıcı ve keşfedici olarak iki temel gruba ayrılabilir. Keşfedici modeller ise tanımlayıcı ve tahmin edici modeller olara iki ana grupta incelenebilir. Şekil 2.1 bu bakış açısıyla veri madenciliğinde kullanılan modellerin sınıflandırılmasını gösterilmektedir[7].



Şekil 2.1 Veri Madenciliği Modelleri

2.3.1 Doğrulayıcı ve Keşfedici Modeller

Doğrulayıcı modeller; araştırmacının bilge ve tecrübesi dahilinde, araştırmak istediği konu ilgili veya literatürün önerdiği bir hipotezin değerlendirilmesi ile ilgilenir. Bu modeller uyumluluk testi, ortalamaların t-testi, varyans analizi gibi geleneksel istatistiksel yöntemleri içerir. Bu yöntemler keşfedici veri madenciliği ile daha az ilişkilidir fakat keşfedici analiz sürecinde de gerekli görüldüğü durumlarda kullanılmaktadır. Keşfedici modeller ise veri kümesi içindeki örüntüleri yakalamak amacıyla kurulan tanımlayıcı ve tahmin edici olmak üzere iki ana grupta oluşturulan modellerdir[8].

2.3.1.1 Tanımlayıcı Modeller

Tanımlayıcı modeller; karar verme sürecine rehberlik amaçlı kullanılabilen, analiz edilen veri kümesinin altında yatan bilgilerin ortaya çıkmasını, yani veri kümesinde varolan örüntülerin tanımlanmasını sağlayan modellerdir. Kümeleme, birliktelik kuralları ve ardışık zamanlı örüntüler tanımlayıcı modellerdir.

2.3.1.1.1 Kümeleme Analizi

Kümeleme analizinin amacı, öznitelikleri açısından birbilerine benzer üyeleri olan farklı grupları veri içinde ortaya çıkarmaktır. Kümeleme analizinde, aynı grup üyelerinin birbirine benzer yani homojen, farklı grup elemanlarının ise birbirinden farklı yani heterojen olması beklenmektedir.

Kümeleme analizi veri madenciliğinde farklı amaçlarla kullanılabilmektedir: Oluşturduğu grafikler ile veri seti içindeki grup benzerliklerinin kolaylıkla görselleştirilmesi, veri seti içindeki aykırı gözlemlerin kolaylıkla tespit edilmesi, büyük verilerle çalışamayan algoritmalar için örneklem yaratması bunlar arasında sayılabilir [7,9].

Genel olarak kümeleme algoritmaları aşağıdaki gibi sınıflanmaktadır:

1. Hiyerarşik yöntemler
    1.1. Toplaşım kümeleme algoritmaları
        1.1.1. Tek bağlantı yöntemi
        1.1.2. Tam bağlantı yöntemi
        1.1.3. Ortalama bağlantı yöntemi
        1.1.4. Merkezi kümeleme(Centroid) yöntemi
        1.1.5. Ward yöntemi
        1.1.6. İki aşamalı yöntem
    1.2. Bölünür kümeleme algoritmaları
2. Hiyerarşik olmayan yöntemler
    2.1. K – Ortalamalar yöntemi
    2.2. Medoid yöntemi
3. Yoğunluk bazlı yöntemler
4. Grid bazlı yöntemler
5. Model bazlı yöntemler

Kümeleme analizi gruplara ayırma işleminde gözlemlerin birbirine olan uzaklıklarını veya birbirine olan benzerliklerini kullanır. Veri setindeki değişkenlerin ölçeğine göre Euclidean (Öklit), Karesel Euclidean, Pearson, Manhattan, Minkowski, Mahalanobis uzaklık ölçütleri, Açısal ve Kosinüs benzerlik ölçütleri ya da Jaccard, Ochiai ve Rao benzerlik katsayıları tercih edilmektedir[9].

2.3.1.1.2 Birliktelik Kuralları

Birliktelik kuralı, veri setinde bir arada sık olarak görülen yani eş zamanlı gerçekleşen ilişkilerin ortaya çıkarılması amaçlamaktadır. Birliktelik kurallarının analizi süreci market sepeti analizi olarak da adlandırılır. Market sepeti analizinde müşteri ile ilgili veri hareketlerinden gelecekte müşterinin nasıl bir tercih yapacağına dair sonuçlar tahmin edilmektedir. Çok sayıda verinin depolandığı bir veri tabanı içinde çeşitli nitelikler arasında hemen fark edilmeyen birtakım ilişkilerin ortaya çıkartılması stratejik kararların alınmasına yardımcı olabilmektedir. Ancak, bu ilişkilerin çok sayıda verinin içinden elde edilmesi basit bir süreç değildir. Bu süreç birliktelik kuralı madenciliği (association rule mining) olarak da adlandırılmaktadır [9].

2.3.1.2 Ardışık Zamanlı Örüntüleri

Ardışık zamanlı örüntüleri, birbirleri ile ilişkisi olan ve birbirini izleyen dönemlerde gerçekleşen olaylar arasındaki ilişkilerin tanımlanmasında kullanılır. Bir alışveriş sırasında müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi birliktelik kuralları ile bulunurken, birbirini izleyen alışverişlerde bu eğilimin belirlenmesi ardışık zamanlı örüntüler ile belirlenmektedir[7].

2.3.2 Tahmin Edici Modeller

Tahmin edici modeller; bilinen verilerden yararlanarak, bilinmeyen bir değeri tahmin etmeye çalışırlar. Veri kümesinden hareket ederek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak, yeni ve sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Bu tip modellerde, ayrıca en anlamlı verinin hangisi olduğu ve her bir değişkenin önemliliği belirlenir. Sınıflandırma, regresyon analizi ve zaman serileri analizi tahmin edici modellerdir.

2.3.2.1 Sınıflandırma

Sınflandırma modelleri veri madenciliği teknikleri arasında kümeleme ile birlikte en sık tercih edilen modellerdir ve denetleyici (supervised) öğrenme gerçekleştirirler. Resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konuları sınıflandırma tekniklerinin kullanıldığı alanlar olarak örneklendirilebilir. Veri madenciliğinde sınıflama, önceden tanımlanmış sınıfların birinde görülmeyen verileri sınıflandırmada kullanılabilen önceden sınıflandırılmış veri nesnelerinden bir model çıkarmayı gerektirmektedir.

Birçok sınıflandırma modeli mevcuttur. En çok kullanılan modeller, yapay sinir ağları (Neural Networks), karar ağaçları (Decision Trees), Bayes ağları (Bayesian networks), genetik algoritmalar, doğrusal (linear) ve olgu tabanlı (instance based) sınıflama modelleri, destek vektör makineleri (Support Vector Machines) ve Naive Bayes olarak sıralanabilir. Bu modeller kısaca şu şekilde tanıtabilinir[7,9,10]:

Yapay Sinir Ağları en basit tanımla, insan beyninin işleme mantığını temel alarak modelleme yapan algoritmalardır. Beyindeki sinirlerin çalışmasından esinlenilerek sistemlere öğrenme, hatırlama, bilgiler arasında ilişkiler oluşturma gibi yetenekleri kazandırmayı amaçlayan yapay sinir ağları, basit biyolojik sinir sisteminin çalışma şeklini simüle etmek için matematiksel model olarak tasarlanmışlardır. Simüle edilen sinir hücreleri çeşitli şekillerde birbirlerine bağlanarak ağı oluştururlar. Bu ağlar öğrenme, hafızaya alma ve veriler arasındaki ilişkiyi ortaya çıkarma kapasitesine sahiptirler. Yapay sinir ağlarının varsayımdan bağımsız olması bir avantaj olarak sayılabilirken, “black box” olarak tabir edilen teknikler arasıdadır. Yani analizcinin bilgi ve tecrübesine dayanır, parametre ayarları ve iteratif yapısından dolayı her denemede farklı sonuçlar doğurur.

Karar Ağaçları; sınıflandırma problemlerinde en çok kullanılan tekniklerden biri olup akış şemalarına benzemektedirler. Karar ağaçlarında kökten dallara doğru gidilerek sınıflandırma kuralları yazılır ve ağaç oluşturulur. Daha sonra veritabanındaki her bir kayıt bu ağaca uygulanır. Çıkan sonuca göre de kayıt sınıflandırılır. Karar ağaçlarında kullanılan birçok algoritmalar mevcuttur. Kurallar oluşturulurken hangi algoritmanın kullanılacağı önemlidir. Kullanılan algoritmaya göre ağacın şekli değişebilmektedir.

Bayes Ağları; Bayes teoremini kullanan istatistiksel sınıflandırıcı olup bir sınıflandırma sorununun olasılık terimleriyle açıklanabileceği varsayımına dayanır. Değişkenlere ait alt kümeler arasındaki koşullu bağımsızlıkları tanımlar.

Naive Bayes kolay uygulanabilir olduğu kadar üstün performansıyla kesikli değişkenlerden oluşan veri setlerinin sınıflandırmasında oldukça tercih edilen modellerdendir.

Genetik algoritmalar; doğal seçim ilkelerine dayanan bir arama ve optimizasyon yöntemidir. Geleneksel optimizasyon yöntemlerine göre farklılıkları olan genetik algoritmalar, parametre kümesini değil kodlanmış biçimlerini kullanırlar. Olasılık kurallarına göre çalışan genetik algoritmalar, yalnızca amaç fonksiyonuna gereksinim duyar. Genetik algoritmaların, fonksiyon optimizasyonu, çizelgeleme, mekanik öğrenme, tasarım, hücresel üretim gibi alanlarda başarılı uygulamaları bulunmaktadır.

Olgu tabanlı modeller; tahmin işlemi sırasında önceden derlenmiş soyut çıkarımlar yerine belirli, özel örnekler kullanır. Bu algoritmalar olasılık kavramlarını tanımlayan ifadeleri kullanabilirler çünkü örnekleri sınıflandırırken doğru eşleşmeyi sağlamak için benzer fonksiyonları kullanırlar

Destek vektör makineleri, sınıflandırmayı bir doğrusal ya da doğrusal olmayan bir fonksiyon yardımıyla yapar. Veriyi birbirinden ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanmaktadır.

2.3.2.2 Regresyon ve Zaman Serileri Analizi

Regresyon analizi ve zaman serileri analizi kişisel yargılardan etkilenmeyen, objektif tahminler geliştirilebilmesi ve işletmelere doğru kararlar alabilmelerinde önemli avantajlar sağlamaktadır[7,9,10].

Tahmin edici model olarak kullanılan zaman serileri analizinde, tahmin edilecek değişkene ilişkin geçmiş veriler belirli bir veri seyri elde etmek üzere analiz edilmektedir. Bu nedenle tahmin etme sadece geçmiş verilerin bu amaçla analiz edilmesine ve yapılacak tahminlerde kullanılmasına dayanmaktadır. Bu özelliğinden dolayı zaman serileri analizi, değişmeyen koşullar altında daha etkin olmaktadır.

Regresyon analizinin kullanılması ise, değerleri tahmin edilecek değişkenle ilişkili olan diğer değişkenlerin belirlenmesini içermektedir. Bu değişkenler belirlendikten sonra geliştirilen istatistiksel model, tahmin edilecek değişken ile diğer değişkenler

arasındaki ilişkiyi tanımlamakta ve ele alınan değişkene ilişkin tahminler yapılmasında kullanılmaktadır.

3. METİN MADENCİLİĞİ

Metin madenciliği en kısa tanımla veri madenciliğinin dokümanların barındırdığı metinler üzerinde uygulanmasıdır. Günümüzde matbu olarak basılı dokumanların yanısıra dijital dokümanların boyutu oldukça fazladır, ve gün geçtikçe artmaktadır. Dijital olarak dokümanlar, internet ortamındaki dokümanlar, web sayfaları, e-postalar ve yazılı ortamlarda bulunan dokümanların dijital ortama aktarılmasıyla elde edilen metinler olarak örneklendirilebilir. Bu dokümanlar büyük ölçekte yapısal olmayan veri barındırmaktadır. Yapısal olmayan verilerin işlenmesi ve analiz edilmesi, sayısal verilere göre farklılık göstermektedir[11].

Metin madenciliğinin uygulamaları farklı isimler alabilmektedir. Örneğin sosyal medyada yer alan kısa metinler sosyal medya analizi, metinlerdeki duygu ve fikir ifade eden terimler ise duygu analizi(sentiment analysis) ve fikir analizi(opinion mining), sadece internet sitelerinin incelenmesi ise internet(web mining) analizi olarak isim alabilmektedir. Bu uygulamalarda çoğunlukla sınıdlandırma, kümeleme ve birliktelik analizleri kullanılmaktadır. Metinlerden oluşan veri setinin bir ağ yapısı barındırması durumunda, bu analizlere ek olarak ağ analizi de dahil edilmektedir.

Metin madenciliği, analiz edilen doküman üzerinde iki ana yaklaşımı barındırır. Bunlardan ilki bilgisayar bilimlerinin bir branşı olan doğal dil işlemedir. Doğal dil işlemede incelenen dokümanın yazıldığı dilin gramatik yapısı da gözönünde bulundurularak doküman bir bütün olarak analiz edilir. Türk Dili üzerinde bu yaklaşımın uygulanabilirliği, günümüzde yeterli yetkinlik seviyesinde değildir. “Bag of words” olarak adlandırılan diğer yaklaşım, dokümanları parçalayarak inceler, gramer ve sektans yapısını dikkate almaz[11,12]. Metin önce kelimelere, daha sonra kelimelerin köklerine ayrılır ve köklerin frekansları üzerinde analizler gerçekleştir. Türk Dili üzerinde bu yaklaşımını kullanan çalışmalar, özellikle sosyal medya analizinin de popülerleşmesiyle giderek ivme kazanmaktadır. Fakat her iki yaklaşımda da karşılaşılan en büyük problem, doküman üzerinde analizlerin gerçekleşmesi için gereken Tükçe bir sözlüğün henüz tam olarak oluşturulmamasından kaynaklanmaktadır. Bu konu üzerinden güncel son çalışmalar,

İstanbul Teknik Üniversitesi Bilgisayar be Bilişim Fakültesi'nde Doğal Dil İşleme Grubu'nun oluşturduğu İTÜ Doğal Dil İşleme Yazılım Zinciri web arayüzü [13] ve Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği bölümü Kemik Doğal Dil İşleme Grubu'nun geliştirdiği java tabanlı Kemik[14] isimli sözlük programıdır. Diğer akademik araştırmalar, araştırmacıların kendi oluşturdukları sözlükler üzerinde gerçekleştirilmektedir.

Veri madenciliğinde kullanılan sınıflandırma, kümeleme ve birliktelik analizleri metin madenciliğinde de tercih edilen analizlerdir, fakat verinin yapısı ve analizin amacı açısından farklılıklar ortaya çıkmaktadır. Örneğin veri madenciliğinde birliktelik analizi ile veri setinde kullanıcı tarafından belirtilmiş nesnelerin birbirleriyle ilişkisi ya da birlikte olma sıklıkları ortaya çıkartılır ve en fazla sepet analizinde kullanılır. Burada amaç, müşterinin aldığı ürünlerin bir arada bulunma olasılıklarının belirlenmesi ve birliktelik algoritması ile hesaplanan güven ve destek seviyeleriyle birlikte değerlendirilmesidir. Metin madenciliğinde ise müşterinin aldığı ürünlerin bir arada bulunma sıklığı yerine sözcüklerin belli koşullar altında birbirini izleme sıklığı ortaya çıkartılır. Buna bağlı olarak da bir dokümanın başlık analizi (topic modelling) yapılır ve/ya dokümanın içerdiği bilgi hakkında fikir sahibi olunur. Burada analiz edilen verinin yapısı önem kazanmaktadır. Analiz edilen veri, tek bir doküman ya da aynı anda birden fazla doküman olabilmektedir ve bu teknik farklılık, analizde kullanılabilecek algoritmaları da belirleyen en önemli unsurdur[11,12].

Herhangi bir kavramın metinde bir da belirli sayıda geçme olasılığının belirlenmesi ve bununla ilgili kurallar türetme çalışmaları metin madenciliğinde sınıflandırma çalışmalarına tipik bir örnektir. Örnek olarak "bulut" sözcüğünün bir metinde üç ve üzerinde geçmesi için gerekli kurallar çeşitli analizlerin yapılması ve metinler üzerinde algoritmaların konuşturulmasından sonra şu veya benzer bir şekilde oluşacak ya da algoritma tarafından üretilip kullanıcıya sunulacaktır.

Metin madenciliğinin amacı yapılandırılmamış bilgiyi işlemek, metinden anlamlı sayısal içerikleri çıkarma ve böylece çeşitli veri madenciliği algoritmaları için metinde içerilen bilgiye erişebilmektir. Bilgi, dokümanlarda bulunan kelimelerin özetlerinden türetilerek çıkarılabilir. Böylelikle bir dokümanın içerdiği kelimeler, veya kelime kümeler analiz edilebilir ya da birden fazla doküman aynı anda analiz edilerek dokümanlar arasındaki benzerlikler belirlenebilir[15].

3.1 Metin Madenciliği Uygulama Alanları

Metin madenciliğinin uygulama amaçları aşağıdaki gibi sıralanabilir[11,16]:

- **Enformasyon Getirimi (Information Retrieval):** Bu aşama ilgilenilen korpus hakkında ön bilginin toplandığı aşamadır. Örneğin metin madenciliği web üzerindeki veri kaynakları üzerinde yapılacaksa web sayfaları, adresleri veya dosya sistemi üzerindeyse dosyaların tarihleri, kullanıcı bilgileri, dosya isimleri, dizin bilgileri gibi bilgiler öncelikli olarak derlenir.
- **Doğal Dil İşleme Aşaması (Natural Language Processing):** Bu aşama bütün metin madenciliği aşamalarında kullanılmasa bile genelde özellik çıkarımı ve metinden bazı anlamsal bilgilerin elde edilmesinde sıklıkla başvurulan aşamadır. Örneğin, konuşma parçalarının etiketlenmesi (part of speech tagging) veya cümlebilimsel parçalama (syntactic parsing) veya diğer dilbilimsel işlemler doğal dil işleme aşamasında yapılır.
- **Adlandırılmış Varlık Tanıma (Named Entity Recognition):** Genellikle metin işleme aşamasında istatistiksel bazı özelliklerin çıkarılması için kullanılır. Örneğin, metnin içerisindeki kişi isimleri, yer isimleri, semboller, kısaltmalar v.s. bu yöntemle bulunur. Örneğin ‘’osmanbey’’ kelimesi, istanbulda bir semt ismi olabileceği gibi bir kişi ismi de olabilir. Adlandırılmış varlık tanıma çalışmalarında, hedeflenen kelime gruplarının metin içerisinden çıkarılması, sayılması, yoğunluğunun bulunması, etiketlenmesi gibi işlemler yapılabilir.
- **Örüntüsü Tanımlı Varlıkların Bulunması (Pattern Identified Entities):** Bazı durumlarda, metnin içerisinden özel bazı bilgilerin metin madenciliğine konu olması mümkündür. Örneğin e-posta adresleri, telefon numaraları, adresler, tarihler gibi bazı bilgileri özel olarak tespit edilmek istenebilir.
- **Eş Atıf (Coreference):** Bir varlığa işaret eden (atıf eden) isim kelime gruplarını ve diğer terimlerin bulunması/ayrılmasını hedefler.
- **İlişki, kural, olay çıkarımları:** Çeşitli amaçlarla metnin içerisinden bazı bilgilerin çıkarılması istenebilir
- **Duygu analizi (Sentimental Analysis):** Metinlerde geçen duygusal ifadelerin çıkarılmasını amaçlar.

Metin madenciliğinin uygulandığı alanları ise şu şekilde örneklendirilebilir:

- Müşteri ilişkileri yönetimi,
- Sahtekarlık tespiti,
- Sağlık alanı,
- Pazar araştırmaları,
- Metinlerden bilgi çıkarımı,
- Doküman özetleme,
- Doküman sınıflandırma
- Benzer içerikleri belirleme
- Web içerikleri sınıflandırma
- Yazar tanıma sistemleri ve Soru-cevap sistemleri.

3.2 Metin Madenciliği Metodolojisi

Veri madencileri tarafından yaygın olarak kabul edilen birkaç veri madenciliği süreci mevcuttur, ancak metin madenciliği için tam olarak kabul edilen bir süreç modeli mevcut değildir. Çalışmanın bu kısmında metin madenciliği için önerilen bir süreci anlatmaya ayrılmıştır.

Metin madenciliği uygulamaları kapsamları bakımından çok genel, amaçları bakımından ise çok çeşitlidir, dolayısıyla başarısını genel anlamda ifade etmek zordur. Diğer iyi kurulmuş yöntemlerle karşılaştırıldığında, metin madenciliği bilgi keşfi için nispeten yeni ve standartlaştırılmamış bir analitik yöntemdir. Metodolojisini anlatmak zordur. Bir metodoloji birbiriyle ilişkili birçok görevi içeren (örn. metinsel veri tabanlarından bilgi çıkarmak gibi) karmaşık süreçleri yürütmek ve yönetmek için çeşitli yöntemleri, araçları ve teknikleri kullanarak belgelenmiş ve bir şekilde standartlaştırılmış bir süreçtir. İyi tasarlanmış ve düzgün bir şekilde takip edilen ya da uygulanan bir metodoloji başarılı sonuçlar elde etmeye yardımcı olabilir[11].

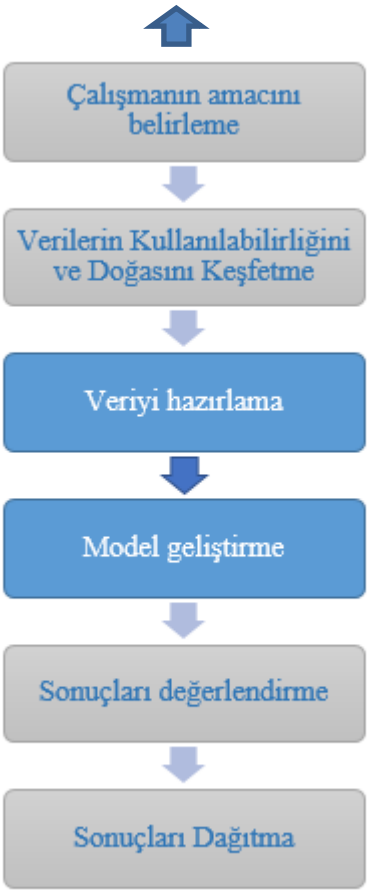
Metin madenciliği uygulamaları, öncelikle kişisel deneyim ve tercihler bazında deneme yanılma yöntemiyle sağlanır. Veri madenciliği yöntemleri nispeten olgunlaşırken herhangi bir alanda uygulamaların özünü yansıtan ve kabul edilen bir metin madenciliği yöntemi yoktur. Bu boşluğun en önemli sebepleri:

- Metin madenciliği farklı insanlar için farklı şeyler demektir. Hatta bunun tanımı ve kapsadığı şey çok kararsız ve tartışılabilir konulardır.
- Verilerin yapılandırılmamış yapısı çok farklı yelpazelerde keşfedici yollar açar.
- Bazıları yarı-yapılandırılmış (HTML ve XML dosyaları gibi) olmak üzere pek çok yapılandırılmamış veri türü vardır.
- Eldeki verilerin büyüklüğü erken örnekleme ve basitleştirme faaliyetlerini teşvik eder.

Bu bağlamda veri madenciliğinin yaygın olarak kullanılan işleyiş süreci olan CRISP-DM, bir standart oluşturulana kadar metin madenciliğinde de tercih edilebilir[11]. Açılımı Veri Madenciliği İçin Çapraz Endüstri Standart Süreci olan CRISP-DM'de bir veri madenciliği projesinin tüm yaşam döngüsü altı aşamadan oluşmaktadır:

- İşi anlama (çalışmanın amacını belirleme)
- Veriyi anlama (veri araştırması)
- Veriyi hazırlama (ön işleme, veriyi uygun bir temsil yöntemi aracılığıyla temsil etme, öznitelik seçimi)
- Modelleme
- Değerlendirme
- Dağıtım

Şekil 3.1'de metin madenciliği için CRISP-DM süreci görselleştirilmiştir.



Şekil 3.1 CRISP-DM Metin Madenciliği İşleme Süreci

3.2.1 Çalışmanın Amacının Belirlenmesi

Başka herhangi bir çalışmada olduğu gibi metin madenciliği çalışması da çalışmanın amacını belirlemekle başlar. Altta yatan sistemi, yapısını, sistem kısıtlamalarını ve mevcut kaynakları ayrıntılı bir şekilde değerlendirmek için genellikle alan uzmanlarıyla yakın bir ilişki içinde olmalıyız. Ancak o zaman çalışmanın yönünü yönetmek için gerçekçi hedef ve amaçlar geliştirebiliriz.

3.2.2 Verilerin Kullanılabilirliğini ve Doğasını Keşfetme

Çalışmanın amacı belirlendikten sonra spesifik çalışma bağlamında gerekli verilerin kullanılabilirliği, elde edilebilirliği ve uygulanabilirliğini değerlendirmek gerekir. Bu aşamanın bazı görevleri aşağıdaki gibidir:

- Metinsel veri kaynağının belirlenmesi
- Verilerin erişilebilirliğinin ve kullanılabilirliğin değerlendirilmesi
- İlk veri kümesinin toplanması

- Verilerin zenginliğinin araştırılması
- Verilerin nitelik ve kalitesinin değerlendirilmesi

3.2.3 Veriyi Hazırlama

Bu aşama veri madenciliği ile metin madenciliği arasındaki en önemli farklılıkları ortaya çıkarır. Şekil 2.1'de koyu renkle belirtilen veri hazılama ve model geliştirme aşamaları metin madenciğinin veri madenciliğine göre içerik olarak farklılaşan aşamalarını belirtmektedir[11].

Metin madenciliğinde veriyi hazırlamanın ilk adımı korpus oluşturmaktır. Korpus dil ile ilgili bir problemi analiz etmek için derlenen ilkeli bir metin koleksiyonudur.

Korpusu oluşturan tüm kelimeleri elde etmek için bir tokenization(dizgeciklere ayırma) süreci gerekmektedir. Korpus oluşturulduktan sonra ise inceleme yapabilmek için temizleme işlemlerine başlanır. Korpusu yapısal hale getirebilmek için metinlerde yer alan rakamların ve noktalama işaretlerinin metinden kaldırılması gerekir. Tekrarlı boşluklar ve beyaz boşluklar da korpustan kaldırılmalıdır. Ayrıca korpus yapısı web sayfalarından ya da HTML, XML gibi formatlardan derlenmişse tablo, şekil ve resimlerden de arındırılması gerekmektedir[17].

İlgili belge koleksiyonlarının bir araya getirilerek oluşturulduğu korpus içinde yer alan tüm farklı kelimeler belge topluluğunun sözlüğü olarak tanımlanır. Algoritmaların daha formal bir tanımını yapmak için sıklıkla aşağıdaki gibi gösterilen bazı terim ve değişkenleri tanımlamak gerekir. D doküman sayısı ve ve $T=\{t_1,\ldots, t_m\}$ sözlük olmak üzere *D*'de meydana gelen tüm farklı terimlerin kümesi mutlak sıklık frekansı $t \in T$ dökümanlarda $d \in D$ tf(d,t) olarak verilir [11].

Filtreleme (Filtering): Bu yöntem ile sözlükteki ve dolayısıyla dökümanlardaki kelimeler filtrelenebilir. En yaygın olarak kullanılan filtreleme yöntemi durma kelimeleridir *(stop words)*. Buradaki amaç edat, bağlaç, zamir gibi tek başına bir anlam ya da duygu durumu belirtmeyen ve içeriğe bu anlamda herhangi bir etkisi olmayan kelimelerin korpustan çıkarılmasıdır (örn, böyle, gibi, kadar). Bunun dışında dökümanlarda çok sık olarak geçen ve bu sebeple ayırt edici bir özelliği olmayan ya da çok az geçen ve istatistiksel olarak önemli bir etkisi olmayan kelimeler de filtrelenebilir[12].

Temel Yapıya Döndürme (Lemmatization): Bu yöntem herhangi bir zamanda çekimlenmiş fiilleri mastar halinde ve çoğul isimleri tekil formada haritalamaya çalışır. Bu işlem sözcüklerin cümle içerisindeki konumlarını, yapısını ve anlamsal olarak işlevlerini de bilmeyi gerektirdiğinden zor ve hataya oldukça açık bir yöntemdir. Bu nedenle uygulamada daha çok "kökenine döndürme" (stemming) işlemleri tercih edilir.

Kökenine Döndürme (Stemming): Bu yöntem sözcükleri basit hallerine çevirir. İsimlerden çoğul eklerin atılması, çekim eklerinin fiillerden arındırılarak kök haline döndürülmesi gibi işlemler stemming olarak adlandırılır. Türkçe için bu amaçla açık kaynak, platform bağımsız ve genel amaçlı bir Doğal Dil İşleme Kütüphanesi olan Zemberek geliştirilmiştir. Java ile çalışır. Zemberek kullanılarak Türkçe kelimelerde stemming kullanılabilir. Ayrıca bazı üniversiteler kendi stemming algoritmalarını geliştirmektedir. Örneğin Yıldız Teknik Üniversitesi'nin Kemik, İstanbul Teknik Üniversitesi'nin İTÜ NLP adında stemming algoritmaları vardır.

Korpus temizlendikten sonra modellemenin ilk aşaması için uygun bir temsil yöntemi seçilir. Metinlerin sözdizimsel yapısından ve anlamsal içeriklerinden daha fazla yararlanabilmek için çeşitli teknikler geliştirilmiştir. Bununla birlikte çoğu metin madenciliği uygulaması bir metin dökümanını metinde yer alan sözcüklerin kümesi olarak temsil etme fikri üzerinde geliştirilmiştir (bag-of-words, temsil yöntemi).

Sözcüklerin doküman içindeki önemlerinin de temsil edilmesine olanak sağlayan vektörel bir temsil şekli vardır (vector representation). Bu modelin adı Vektör uzayı modelidir (vector space model).

Vektör Uzayı Modeli

Vektör uzayı modeli büyük boyutlardaki veri dökümanlarının etkin olarak dizinlenmesi ve veri analizinin verimli bir şekilde yapılması için kullanılır.

Vektör uzayı modeli'nde doküman ve sorgular m-boyutlu vektörlerle temsil edilirler. Burada *m* sözlükteki terim sayısıdır. Vektör uzayı modelinde her bir doküman sayısal bir öznitelik vektörüyle temsil edilir: $w(d) = (w(d, t_1),\ldots,w(d, t_m))$. Vektörün her bir boyutunda ilgili terimin dökümanlardaki ağırlığı da yer almaktadır[11].

Vektör uzayı modeli içinde kelimelere ait sayısal bir değer olur. Örneğin bir terimin ilgili dökümanlarda yer alıp almaması (0,1) değerleri ile ifade edilebilir. Vektör uzayı modelininin basit bir örnek bir doküman üzerinde uygulaması aşağıdaki gibidir:

Çizelge 3.1 Örnek Metinler

ID	Metinler
1	İnsanlar şeytana tapmaktan ve insan kurban etmekten bahsediyor.
2	Basit insanların ünlü bir insanı tanımaları çok etkileyici olur.
3	Ama kafasının içinde fenalık fısıldayan şeytanlar vardı.
4	İfadenizde işinizin insanlara yalan söylemeyi öğretmek olduğunu söylediniz
5	Kurban ya da kurbanın yakını değil.

Yukarıdaki örnek metinler için anahtar kelimelerle oluşturulan sözlük ve terimsel tanımlamaları aşağıda görülmektedir.

Sözlük = {insan, şeytan, kurban, ünlü, kafa, fena, ifade, yalan, kötü}

Vektör uzayı modelinde sözcüklerin varlığı;

D1 = {1,1,1,0,0,0,0,0,0}

D2 = {1,0,0,1,0,0,0,0,0}

D3 = {0,1,0,0,1,1,0,0,0}

D4 = {1,0,0,0,0,0,1,1,0}

D5 = {0,0,1,0,0,0,0,0,0}

Vektör uzayı modelinde terimlerin frekansı;

D1 = {2,1,1,0,0,0,0,0,0}

D2 = {2,0,0,1,0,0,0,0,0}

D3 = {0,1,0,0,1,1,0,0,0}

D4 = {1,0,0,0,0,0,1,1,0}

D5 = {0,0,2,0,0,0,0,0,0}

şeklindedir.

Kural olarak, birbiriyle alakalı dökümanlarda sıkça yer alan ancak tüm korpusta nadir olarak karşılaşılan terimlerin ağırlığı da fazla olmalıdır [11,12]. O halde *d* dökümanındaki *t* terimi için *w* ağırlığı *w(d,t)* hesaplanırken terim frekansı (term

frequency - *TF*) ile ters doküman frekansı (inverse document frequency – *IDF*) çarpılmalıdır. Terim frekansından bazı kaynaklarda *Term Document Matris*, ters doküman frekansından ise *Document Term Matris* olarak bahsedilir. *TF(d,t)* aşağıdaki şekilde hesaplanabilir [11,12].

$$TF(d,t) = 1 \quad (2.1)$$

$$TF(d,t) = F(d,t) \quad (2.2)$$

$$TF(d,t) = 1 + \log_e F(d,t) \quad (2.3)$$

F(d,t), t'nin d içindeki frekansıdır. *TF* hesaplanırken denklem (2.3) tercih edilir.

IDF(t) için önerilen hesaplama şekli aşağıdaki denklemlerde gösterilmektedir [18].

$$IDF(t) = \frac{1}{F(t)} \quad (2.4)$$

$$IDF(t) = \log_e\left(1 + \frac{n}{F(t)}\right) \quad (2.5)$$

$$IDF(t) = \log_e\left(1 + \frac{F^m(t)}{F(t)}\right) \quad (2.6)$$

$$IDF(t) = \log_e\left(\frac{n-F(t)}{F(t)}\right) \quad (2.7)$$

F(t), t'yi içeren doküman sayısı; $F^m(t)$, en büyük f(d,t) değeri; n ise korpustaki doküman sayısıdır.

En tercih edilen şekliyle d dökümanındaki t teriminin ağırlığı:

$$W(d,t) = TF \times IDF = (1+\log_e F(d,t)) \times \log_e\left(1 + \frac{n}{F(t)}\right) \quad (2.8)$$

olarak ifade edilir.

3.2.4 Modeli Belirleme ve Geliştirme

W(d) içeriği denklem (2.8) aracılığıyla elde edildikten sonra benzerlik ölçüleri hesaplananır. Eğer metin madenciliğinde kümeleme algoritmalarının kullanılması amaçlanmışsa, iki doküman arasındaki benzerliğin ölçülmesi gerekmektedir.. Kümeleme analizinde benzerlik hesaplamak için çeşitli ölçüm yöntemleri mevcuttur fakat metin madenciliğinde doküman kümelemesi için Cosine ölçüsü kullanılmaktadır[12].

$$Cosine(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \, \|d_2\| \quad (2.9)$$

Denklem (2.9)'da · vektör nokta ürünü, $\|d\|$ ise *d* vektörünün uzunluğunu ifade etmektedir.

Benzerlik ölçüleri haricinde, uzaklık ölçüleri ile de kümeleme modeli kurulabilir. En çok kullanılan uzaklık ölçüsü Euclidean (Öklit) ölçüsüdür. Bu ölçüt, (2.10) eşitliği ile hesaplanmaktadır.

$$\text{dist}(d_1, d_2) = \sqrt{\sum_{k=1}^{n} |w(d_1, t_k)| - |w(d_2, t_k)|^2} \qquad (2.10)$$

Uzaklık ölçüsü belirlendikten sonra modellemeye geçilir. Metin madenciliğinde elde edilen terim frekans matrisinin oldukça büyük olması hiyerarşik bir kümeleme algoritmasının kullanılmasına engel oluşturur. Genellikle Mac Queen tarafından geliştirilen k-means (k-ortalama) algoritması tercih edilmektedir.

K-means'in atama mekanizması her verinin sadece bir kümede yer almasına izin verir. Bu nedenle keskin bir algoritmadır. Bu algoritma aşırı uç ya da gürültülü verilerden etkilenir. Algoritmada önceden belirlenen k adet kümenin hedefe konur ve sonrasında k tane ortalama değeri rasgele belirlenir. Bunlar kümelerin merkezleridir. Bu ortalama değerlerine göre veriler hangi değere daha yakınsa o kümeye dahil edilir. Bu işlemler merkezlerin değerleri değişmeyinceye kadar tekrar edilir [9,10].

K-means algoritmasının işlem basamakları aşağıdaki gibidir.

i. k adet nesne rasgele seçilir. Bu nesneler küme merkezlerini temsil eder. $M_1, M_2, .., M_k$ orta noktaları (2.11) ile verilen eşitlikle hesaplanır.

$$M_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ik} \qquad (2.11)$$

ii. Küme içi değişmeleri gösteren karesel hata formülü $e_1, e_2, .., e_k$ eşitlik (2.12)'deki gibi hesaplanır.

$$e_i^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2 \qquad (2.12)$$

k kümesini içeren bütün kümeler uzayı için hata kareler toplamı küme içindeki değişmelerin toplamıdır ve eşitlik (2.13)'deki gibi hesaplanır.

$$E_k^2 = \sum_{k=1}^{K} e_k^2 \tag{2.13}$$

iii. Her bir veri kendisine en yakın kümeye atanır.

iv. Verilerin hepsi en yakın kümelere atandığında tekrar k tane küme için merkezler hesaplanır.

v. Küme merkezlerinde değişme olmayıncaya kadar ii. ve iii. işlemler tekrarlanır [9].

3.2.5 Sonuçları Değerlendirme

Modeller oluşturulup veri analizi perspektifinden geçirildikten sonra, tüm işlemlerin doğru bir şekilde yürütüldüğünü doğrulamak ve değerlendirmek gerekmektedir. Ancak o zaman sonuçların paylaşılması aşamasına geçilebilir. Karar verme sürecinde hata yapılması olasılığı geri dönüşü olmayan zararlara neden olabilir. Sürecin bu şekilde kapsamlı değerlendirilmesi bu durumu büyük ölçüde hafifletmeye yardımcı olacaktır.

3.2.6 Sonuçların Sunulması

Modeller ve modelleme süreçleri doğrulama sürecinden başarılı bir şekilde geçtiğinde sunum aşamasına geçilebilir. Modelleri sunmak, araştırmanın bulgularını karar vericilere hitap edecek şekilde yazılması kadar basit ya da bu modeller etrafında yeni bir iş zekası sistemi kuracak kadar kompleks olabilir. Model sonuçları, daha iyi bir karar verme süreci gerçekleştirmek için tekrar tekrar kullanılabilir. Oluşturulan bazı modellerin doğruluk ve uygunluğu zamanla kaybolabilmektedir. Bu nedenle yeni verilerle periyodik olarak güncellenmesi gerekebilmektedir. Bu işlem, sıklıkla modeli yeniden yaratarak yeni bir analiz süreci başlatmaktan daha kazançlı olacağı aşikardır[11].

4. UYGULAMA

Bu bölümde çalışmanın uygulama aşaması açıklanmaktadır. Uygulama esnasında, analiz edilen dokümanın Türk Dili'nde olması, Türk Dili üzerinde çalışan araçların henüz geliştirilme aşamasında olması, R dili'nin metin madenciliğide Türk Dili'ni desteklememesi ve araştırmanın analiz edilecek veriyi geniş kapsamlı ele almasından kaynaklı veri toplama, temizleme ve düzenleme aşamalarında birçok öngörülemeyen problem ile karşılaşılmıştır. Bu problemler daha çok araç ve verilerin tutulduğu sunucu problemleri ile açık kaynaklı ve farklı araçların kullanılmasından kaynaklı protokol sorunlarıdır. Türk Dili üzerinde metin madenciliği çalışmaları yaygınlaştıkça benzer sorunların ortadan kalkacağı öngörülmektedir.

Uygulama iki aşamadan oluşmaktadır. İki aşamada da R'ın Türkçe konuşma dili ile yazılmış dokümanları analiz etme kabiliyeti araştırıldığından, düzgün ifadeli yazı dili ile yazılmış dokumanlar yerine günlük konuşma dilini ifade edecek altyazılar ve twitter yorumları ele alınmıştır.

4.1 Veri ve Veri Ön İşlemleri

Tezin uygulama aşaması için IMDB'den komedi, aksiyon, dram, fantastik-bilimkurgu ve cinsel içerikli 293 dizini belirlenmiş ve sezon finali yapmış dizilerin son sezonlarınnın tüm bölümlerine ait altyazılar için www.turkcealtyazi.org sitesin kullanılmıştır.

Dizi türlerine hangi dizilerin dahil edileceğine karar verme amacıyla, Türkçe ve İngilizce dizi kanalları, altyazı siteleri ve dizi blogları incelenmiş, dahil edilecek dizi isimleri bellirlendikten sonra ilgili siteden indirilerek her dizinin her bölümü ayrı birer text dosyası olarak tutulmuştur.

Veri temizleme ve düzenleme işlemlerin ardından dizi kategorileri üzerinde yapılan ön kümeleme analizi sonucunda dram ve fantastik-bilimkurgu türlerinin uygulamadan çıkarılmasına karar verilmiştir.Analize alınan 263 dizi, ön işlemlerden sonra Aksiyon kategorisinde 81, Cinsellik kategorisinde 24, Komedi kategorisinde

ise 105 olmak üzere toplam 210 diziye indirgenmiştir. Çalışmada analize dahil edilen dizi kategori ve bu kategorilere ait dizi isimleri Çizelge 4.1'de verilmiştir.

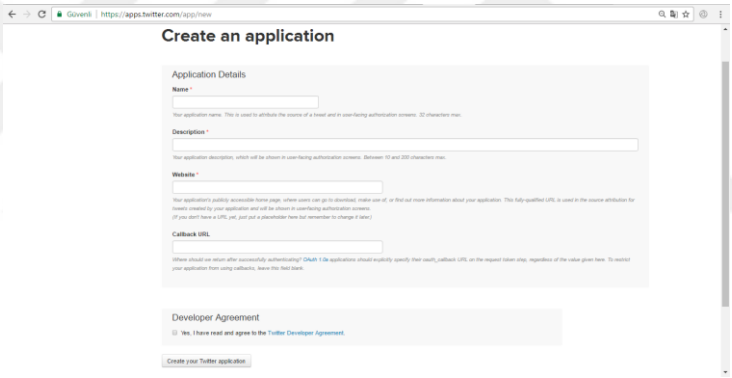
Çizelge 4.1 Analiz İçin Kullanılan Diziler ve Türleri

AKSİYON		
THE SOPRANOS	THE BORGIAS	CHUCK
GOTHAM	JUSTIFIED	LAST RESORT
THE WIRE	24	PERSON OF INTEREST
DEXTER	SHERLOCK	THE AMAZING RACE
BREAKING BAD	THE KILLING	VIKINGS
BOARDWALK EMPIRE	PRISON BREAK	HANNIBAL
ROME	NUMBERS	ARROW
SONS OF ANARCHY	THE MENTALIST	REVOLUTION
HOMELAND	THE GOOD WIFE	HAWAII FIVE-0
TRUE DETECTIVE	BURN NOTICE	PRETTY LITTLE LIARS
CSI MIAMI	WHITE COLLAR	CSI: CRIME SCENE INVESTIGATION
MARVEL'S AGENTS OF S.H.I.E.L.D.	NCIS: LOS ANGELES	ELEMENTARY
PERCEPTION	THE FOLLOWING	NIKITA
THE CLOSER	CRIMINAL MINDS	LAW & ORDER: SPECIAL VICTIMS UNIT
MAGIC CITY	THE AMERICANS	SCANDAL
MANHATTAN	REVENGE	BODY OF PROOF
CHUCK	LIE TO ME	TEEN WOLF
GOTHAM	SOUTHLAND	CSI: NY
DOCTOR WHO	HOMICIDE: LIFE ON THE STREET	THE TUDORS
NARCOS	LIFE ON MARS	CHICAGO FIRE
BLINDSPOT	QUANTUM LEAP	ALIAS
MR.ROBOT	MISSING	AGATHA CHRISTIE'S POIROT
THE SHIELD	MAKING A MURDERER	THE BOONDOCKS
BLUE BLOODS	MARCO POLO	BABYLON 5
THE NIGHT OF	FARGO	STRA TREK: DEEP SPACE NINE
STARGATE SG-1	DAREDEVIL	STARGATE: ATLANTIS
SPOOKS	THE BLACK DONNELLYS	DAMAGES
CİNSELLİK		
SPARTACUS: BLOOD AND SAND	SPARTACUS: GODS OF ARENA	MASTER OF SEX
GAME OF THRONES	SHAMELESS	TRUE BLOOD
THE ROYALS	NIP/TUCK	SHAMELESS
OZ	BANSHEE	NIP/TUCK
ENTOURAGE	BLUE MOUNTAIN STATE	BANSHEE
CALIFORNICATION	THE ROYALS	BLUE MOUNTAIN STATE
TRUE BLOOD	OZ	ORANGE IS THE NEW BLACK
CALIFORNICATION	ENTOURAGE	OUTLANDER

KOMEDİ		
WEEDS	SUITS	FREAKS AND GEEKS
THE SIMPSONS	MODERN FAMILY	PSYCH
FAMILY GUY	THE MIDDLE	THE OFFICE
HOW I MET YOUR MOTHER	BOB'S BURGERS	THE IT CROWD
MELISSA & JOEY	THE SECRET LIFE OF THE AMERICAN TEENAGER	SPACED
MOM	WORKAHOLICS	BORED TO DEATH
MIXOLOGY	PARENTHOOD	30 ROCK
HOT IN CLEVELAND	BOY MEETS WORLD	NEW GIRL
FUTURAMA	SPONGEBOB SQUAREPANTS	THAT '70S SHOW
SOUTH PARK	THE GOLDEN GIRLS	RAISING HOPE
SEINFELD	MIKE & MOLLY	NCIS: NAVAL CRIMINAL INVESTIGATIVE SERVICE
ARRESTED DEVELOPMENT	BLACKADDER GOES FORTH	BONES
THE BIG BANG THEORY	THE LEGEND OF KORRA	TWO AND A HALF MEN
RESCUE ME	CHAPPELLE'S SHOW	GLEE
IT'S ALWAYS SUNNY IN PHILADELPHIA	BLACK BOOKS	CASTLE
CURB YOUR ENTHUSIASM	COUPLING	2 BROKE GIRLS
EPISODES	THE COLBERT REPORT	SUBURGATORY
HOUSE OF LIES	NEWSRADIO	HART OF DIXIE
SCRUBS	DEAD LIKE ME	AMERICAN DAD
FRIENDS	PUSHING DAISIES	ADVENTURE TIME WITH FINN & JAKE
MIKE & MOLLY	THE VENTURE BROS.	COUGAR TOWN
BLACKADDER GOES FORTH	SUMMER HEIGHTS HIGH	NURSE JACKIE
THE LEGEND OF KORRA	3RD ROCK FROM THE SUN	THE MINDY PROJECT
PARTY DOWN	THE INBETWEENERS	SKINS
COMMUNITY	SAMURAI JACK	GO ON
PARKS AND RECREATION	AN IDIOT ABROAD	FULL HOUSE
ARCHER	EASTBOUND & DOWN	VEEP
EXTRAS	TRAILER PARK BOYS	SEX AND THE CITY
WILFRED	MARRIED WITH CHILDREN	AVATAR: THE LAST AIRBENDER
CHEERS	FRASIER	QI
GILMORE GIRLS	MY NAME IS EARL	YOUNG JUSTICE
FLIGHT OF THE CONCHORDS	THE FRESH PRINCE OF BEL-AIR	ROBOT CHICKEN
BOSTON LEGAL	THE OFFICE	FATHER TED
MONTY PYTHON'S FLYING CIRCUS	MYSTERY SCIENCE THEATER 3000	MR. BEAN
BLACK-ADDER II	BLACK ADDER THE THIRD	WHOSE LINE IS IT ANYWAY?

Uygulamanın ikinci bölümünde dört ay boyunca Twitter'da ilk aşamada analiz edilen dizilerin hashtagleri ile atılan Türkçe tweetler toplanmıştır. Türkçe tweetlerin R'da analizini gerçekleştirmek için öncelikle bir kullanıcı API(*Application Programming Interface)*'si oluşturulması gereklidir. API, bir uygulamanın diğer uygulamalarla kendisiyle etkileşime girebilmesi için sunduğu arayüz olarak tanımlanabilir. Diğer bir ifade ile A uygulamasının özelliklerini B uygulamasında da kullanabilmemizi sağlayan bir mekanizmadır. Bu aşamanın basamakları, R'da Türkçe twitter analizini amaçlayan araştırmacılar için detaylı olarak aşağıda sunulmuştur.

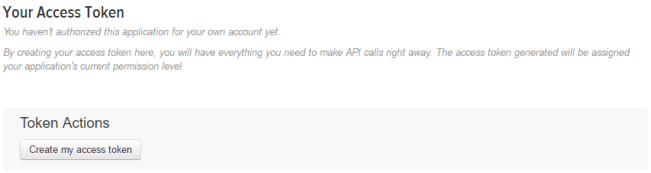
1. Araştırma amaçlı özel hesabınız dışında yeni bir twitter hesabı açmanız önerilir.
2. Twitter Apps adresinden (https://apps.twitter.com/) API oluşturulur(Şekil 4.1).
3. Twitter Apps'de API oluşturabilmek için ekranın sağ üst köşesindeki *Create New App* düğmesine tıklanır.



Şekil 4.1 Twitter Apps Penceresi

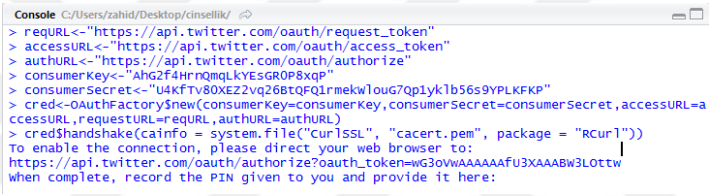
Açılan ekranda üst köşesinde kırmızı yıldız bulunan alanlar API'yi oluşturmak için doldurulması gereken zorunlu alanlardır. *Name* alanına API'ye vermek istenilen isim, *Description* alanına API'nin tanımı, *Website* alanına çekmek istenilen tweetlerin yer aldığı sayfanın linki yazılıp aşağıdaki *Developer Agreement* alanı işaretlendikten sonra *Create Your Application* butonuna tıklanarak API oluşturulur. API oluşturulduktan sonra *"Keys and Access Token"* sekmesine tıklanır. Burada API'ye erişeceğimiz esnada kullanacağımız anahtar bilgileri *(Consumer Key (API Key) ve Consumer Secret (API Secret)* yer almaktadır.

Anahtar bilgileri elde ettikten sonra tweet atılacak hesaplarda bu API'nin erişimi için yetki vermemiz gerekir. Bunun için de kendi hesabımıza ait olan uygulamamıza gereken izni verecek *Access Token*'i üretiriz.



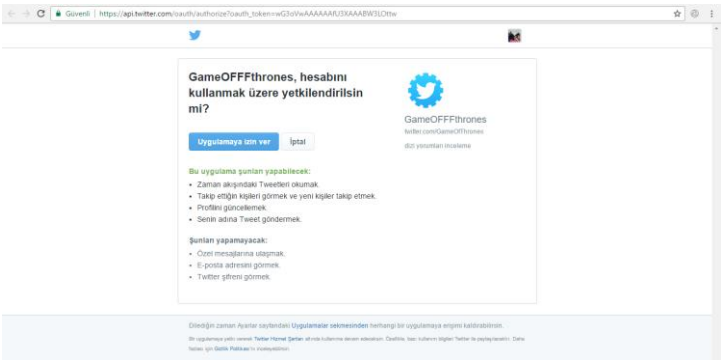
Şekil 4.2 Access Token Penceresi

4. *Access Token* ve *Access Token Secret* bilgileri de oluşturulduktan sonra R açılır ve tweetlerin R a aktarılması için *devtools, MASS, ROauth, TwitteR* paketleri yüklenir.



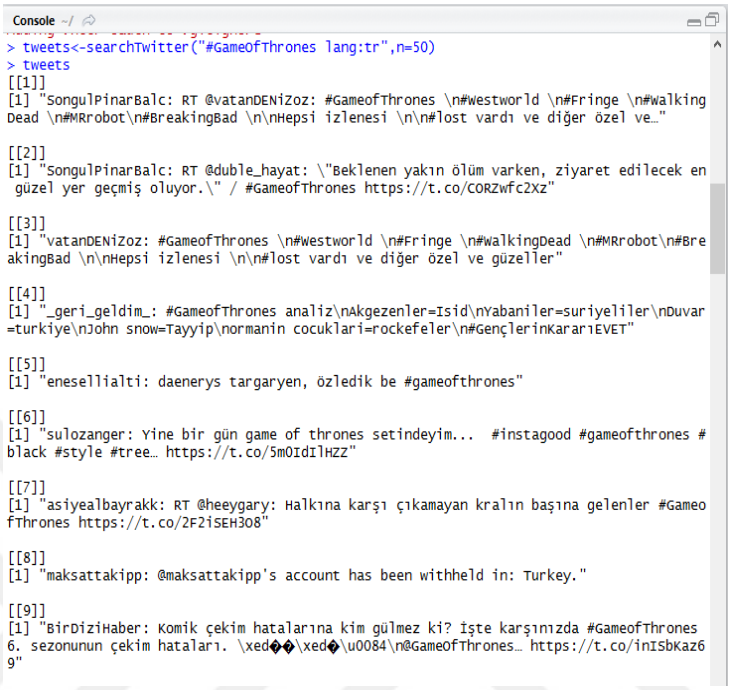
Şekil 4.3 R konsolu

5. API'ye ait URL, anahtar bilgiler ve API'ye entegre olabilmek için gereken Access Token bilgileri tanıtılır. Bağlantı kurabilmek için yazılan koddan sonra programın hesaba erişebilmesi için izin vermek için internet sayfası açılır.



Şekil 4.4 Access Token İzin Penceresi

6. Uygulamaya izin ver butonunu tıkladıktan sonra gelen kod R konsolunda *"record the PIN given to you and provide it here"* diye belirtilen yere girilir. Böylece R'a tweet çekmek için gereken adımları gerçeklemiş olur.



Şekil 4.5 R'a Aktarılan Tweetler

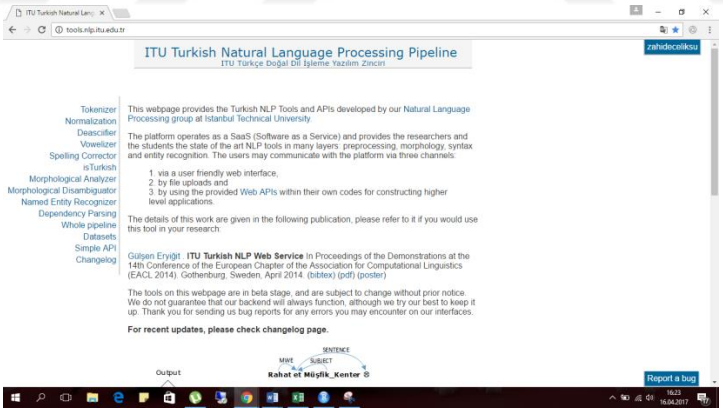
Böylelikle R'a tweetler aktarılır.

Ancak Twitter şirket politikası olarak tüm verisinin sadece %2'sini açık kaynak paylaşıma sunmaktadır. Bu da dizi için inlenecek hashtag'e ait en fazla 1 hafta geriye dönük tweet çekme olanağı sunmaktadır. Bu problemi gidermek, incelenecek tweet sayısını arttırmak amacıyla 4 ay süresi boyunca tweetler sanal bir makine üzerinden toplanmıştır. Böylelikle ilgili hashtagle ait tüm dillerde atılan tweetler bu süre zarfında toparlanmıştır. Sanal sunucu üzerinde toplanan tweetler C Sharp diline kodlama yapılarak Türkçe tweetler ayıklanmış ve tüm türlerde incelenen diziler için analiz edilecek tweetler elde edilmiştir. 4 aylık süre boyunca sadece aksiyon türündeki tweetler analiz edilecek yeterli sayıya ulaştığından tweet analizi uygulaması sadece bü türe ait diziler üzerinde gerçekleştirilmiştir. Çizelge 4.2'de Türkçe tweetleri elde edilebilen dizilerin isimleri verilmektedir.

Çizelge 4.2. Yorumları Çekilen Yabancı Aksiyon Dizileri

ARROW	MAKING A MURDERER
BLINDSPOT	MARCO POLO
BREAKING BAD	MR.ROBOT
CHICAGO FIRE	NARCOS
CRIMINAL MINDS	PERSON OF INTEREST
DAREDEVIL	PRETTY LITTLE LIARS
DEXTER	PRISON BREAK
ELEMENTARY	REVENGE
FARGO	SHERLOCK
GOTHAM	SONS OF ANARCHY
HANNIBAL	TEEN WOLF
HOMELAND	THE NIGHT OF

Tweetler ve altyazılar toplandıktan sonra metinleri geçen kelimeleri temizlemek ve köklerine ayırabilmek için İTÜ'nin Doğal Dil İşleme arayüzü kullanılmıştır R programı metinleri temizleme (rakamları silme, noktalama işaretlerini silme, beyaz boşlukları silme vs.), filtre kelimeleri (stopwords) çıkarma, kelime tamamlama (stemming) ve köklerine ayırma için gerekli algoritmalara sahiptir. Ancak NLP paketi sadece İngilizce ve Almanca dillerini desteklemektedir. Türkçe dil desteği olmaması sebebiyle bu adımları Türkçe Doğal Dil İşleme Yazılım Zinciri panelleri ile gerçekleştirilmiştir.



Şekil 4.6 İTÜ Türkçe Doğal Dil İşleme Yazılım Zinciri Penceresi

Text dosyası olarak tutulan metinleri düzenlemek ve kelime köklerini elde etmek için sırasıyla *Tokenizer*, *Normalization*, *Morphological Analyzer* ve *Morphological Disambugator* sekmelerini kullanılmıştır. Bu sekmelerin işlevleri aşağıdaki gibidir.

Tokenizer: Metni kelime bazında parçalara böler. Teknik anlamda dizgeciklere ayırır.

Normalization: Eksik harfleri tamamlar ve Türkçe yazım kurallarına uygun olarak yazılmayan karakterleri düzeltir (Örn; gitcem, büttttttüüüünnnn, zhnyt). Bu sekmeyi özellikle konuşma dilinin yazılı ifadesi olan tweetlerin temizlenmesinde oldukça kullanışlıdır.

Morphological Analyzer: Kelimeleri eş anlamlılarını da göz önüne alarak ek ve köklerine ayırır.

Morphological Disambugator: Kelimeyi metinde kullanıldığı anlamıyla kök ve eklerine ayrılmış olarak verir.

Metin madeniliğinin, analiz öncesi ön işlemlerini İTÜ Türkçe Doğal Dil İşleme Yazılı Zinciri sistemi ile gerçekleştirmek için şu işlemler sırasıyla uygulanır:

1. Herhangi bir metin editöründe (Notepad++ gibi) UTF-8 encodingi ile toparlanan dosyalar önce "Tokenizer" sekmesinde sayfanın altına yer alan *Upload File* alanına yüklenir ve *Analyze* butonuna basılır. Bir süre sonra sistem tarafından *tokenized* adında bir metin dosyası oluşturulur.
2. Bilgisayara indirilen bu dosya "Normalization" sekmesindeki alana yüklenir ve tekrar analiz edilir. Sistem *normalized* isimli yeni bir dosya oluşturur.
3. Bu dosya indirildikten sonra "Morphological Analyzer" sekmesine yüklenir. Sistem *MorpAnalyzer* isimli yeni bir dosya oluşturur
4. Bilgisayara indirilen bu dosya son olarak "Morphological Disambugator" sekmesindeki alana yüklenir ve tekrar analiz edilir. Sistem *disambugated* isimli yeni bir dosya oluşturur. İndirilen bu dosyada sisteme yüklenen dokümanlardaki metinlerin tüm kelimelerin kök ve ekleri mevcuttur.

Elde edilen son dosya bulunan kökler, Notepad++ yardımıyla eklerinden arındırılır. Her bir dizi türü altyazıları ve toplanan tweetler için bu işlemler tekrarlanır. Böylelikle metinler analize hazır hale gelir.

4.2 Analiz

Metin madenciliği için yaygın olarak kullanılan araçlardan biri olan R, yalın dili ve açık kaynak olması sebebiyle kullanıcıya büyük kolaylık sağlamaktadır. Ancak metin madenciliği uygulamalarında bazı paketlerin eski R versiyonlarında çalışması, yeni versiyon için güncellenmemi olması, dikkat edilmesi gereken bir durumdur. Metin madenciliği için kullanılan başlıca R paketleri aşağıdaki gibi listelenebilir:

Tm: Metin madenciliğinin temel kodlarını içerir.
Slam: Term Document Matrix üzerinde işlemler gerçekleştiren kodları içerir.
Wordcloud: Grafik çizdiren kodları içerir.

Metin madenciliğinde analiz edilen dokümana, latince sözlük anlamına gelen corpus denilmektedir. İTÜ Türkçe Doğal Dil İşleme Yazılım Zinciri sistemi ile metin madenciliğinde kullanılan ön işlemler gerçekleştirilerek analize hazır hale getirilen dokuman R'a aktarıldıktan sonra tm paketi kullanılarak tekrar R'ın ön işlemleri uygulanmıştır. R'da Türk Dili'ni destekleyen bir paketin henüz olmaması, İTÜ sistemi kullanılırken karşılaşılan aksaklıklar ve panel kodunun açık kaynak olmamasından kaynaklı bu sürecin R kodları Şekil 4.7'de belirtilmiştir.

```
#üzerinde çalışılacak metin programa çekilir
text1 <-read.table("aksiyon_tum.txt",header=FALSE,encoding="UTF-8")

#filtre kelimelerin yer aldığı dosya programa çekilir
myStopwords <- read.table("aksiyon_filtre.txt",header=FALSE,encoding="UTF-8")
myStopwords <- as.character(myStopwords$V1)
myStopwords <- c(myStopwords, stopwords())

#metin vektör yapısına dönüştürülür
text1 <- VectorSource(text1)

#korpus tanımlanır
text1 <- VCorpus(text1)

#korpustaki tüm harfler küçük harfe dönüştürülür
text1 <- tm_map(text1, content_transformer(tolower))

#korpustaki noktalama işaretleri kaldırılır, numaralar
# beyaz boşluklar ve filtre kelimeler kaldırılır
text1 <- tm_map(text1, removePunctuation)
text1 <- tm_map(text1, removeNumbers)
text1 <- tm_map(text1, stripWhitespace)
text1 <- tm_map(text1, removeWords, myStopwords)
text1 <- tm_map(text1,PlainTextDocument)
```

Şekil 4.7 R Kodları

Tm paketinin altında metni vektör yapısına dönüştüren, korpus olarak tanımlayan ve korpusu temizleyen kodlar yer alır. R'da stopwords(filtre kelimeler) her dil için otomatik olarak tanımlanmıştır ve bu genel bir bilgidir. Örneğin İngilizce için at, the, or, I, am, she, gibi kelimeler dil içinte çok fazla tekrar edildiğinden ve teknik

anlamda analize değer katan bir bilgi içermediğinden, analiz sürecinden filtre edilerek otomatik olarak çıkartılır. R'ın Türkçe stopwords fonksiyonu olmaması nedeniyle çalışmada her dizi kategorisi için filtre kelimeler belirlenmiş ayrı ayrı dosyalanmıştır.

Analiz edilen verinin sayısal veri yerine doküman olması,ve dokumandaki kelime köklerinin bir matrise çevrilerek analiz edilmesi bilgisayar RAM'ini oldukça işgal etmektedir. Bununla beraberinde Türkçe filtre kelimelerin çokluğu R'ın corpus oluşturmasını güçleştirmekte, hatta bazı durumlarda hata yaratmaktadır. Bunun önüne geçmek amacıyla filtre kelimeler bir döngü yardımıyla sisteme 500'erlik kelime grupları ile sokulmuş ve sorun yaşanmadan corpus'un oluşturulması sağlanmıştır. Şekil 4.8'de R kodları verilmiştir.

```
# ram kullanimi icin stopwords parcalara ayrildi
chunk <- 500
n <- length(myStopwords)
r <- rep(1:ceiling(n/chunk),each=chunk)[1:n]
d <- split(myStopwords,r)
for (i in 1:length(d)) {
  text1 <- tm_map(text1, removeWords, c(paste(d[[i]])))
}
```

Şekil 4.8 R Stopwords Döngüsü

Filte kelimerin temizlenmesiyle bir sonraki aşamaya geçilir. Bu aşamada metin dosyasında geçen her bir kelimenin frekansını veren matris *Term Document Matrisi* oluşturulur. Matrisin oluşturulması için gerekli R kodu Şekil 4.9'da gösterilmektedir.

```
tdm <- TermDocumentMatrix(text1)
freq.terms<-findFreqTerms(tdm,lowfreq=100)

term.freq<-rowSums(as.matrix(tdm))
term.freq<-subset(term.freq,term.freq>=100)
df<-data.frame(term=names(term.freq),freq=term.freq)
m <- as.matrix(tdm)

word.freq<-sort(rowSums(m),decreasing=T) #wordcloud
wordcloud(words = names(word.freq), freq = word.freq, min.freq = 3,
          random.order = F,random.color=TRUE,colors=rainbow(7))
```

Şekil 4.9 Term Document Matrisi

Çalışmada analiz edilen corpus oldukça büyük olduğundan, 100 kereden fazla tekrar eden kelimeler incelenmeye alınmıştır. Grafikler ve analizler bu frekansın üzeri kelime kökleri için oluşturulmuştur.

İTÜ Türkçe Doğal Dil İşleme Yazılım Zinciri sistemi ile bütün farklı kategorilere ait altyazılar tek bir doküman olarak elde edildiğinden kümeleme analizi tercih edilmiş

ve k-means kümeleme algoritması kullanılarak küme modeli oluşturulmıştır. Küme sayısına karar vermek için R'da *NbClust* paketi kullanılmıştır. Bu paket tanımlanan uzaklık türü ve algoritma dahilinde çeşitli indekslerden yararlanaran modelleme için en uygun küme sayısı önermektedir.

NbClust kelimelerin metinde yer alan orijinal sıralarını gözeterek küme merkezi ve terimler arasındaki mesafeyi hesaplar. Küme sayısına bu şekilde karar verir. Çalışmada NbClust algoritmasının, teknik olarak kaynaklanan sorunlardan dolayı -metnin İTÜ Türkçe Doğal Dil İşleme Yazılım Zinciri sunucusunda çıkan sorunlardan dolayı bazı dizilerde manuel köklere ayırma, temizleme ve alfabetik olarak sıralanmış elde edilmesinden – uygun sonuç vermediği gözlemlenmiştir. Bundan dolayı farklı kümeleme modelleri kullanılarak en anlamlı sonuç veren küme sayısı ile modellemede yapılmıştır. Aşağıda örnek bir kod Şekil 4.10'ile gösterilmiştir.

```
rownames(m) <- 1:nrow(m) #kmeans
cl <- kmeans(m, 2,nstart=25,algorithm = "Hartigan-Wong")
```

Şekil 4.10 R k-means fonksiyonu

4.3 Bulgular

Dizi kategorilerine göre grafik ve analiz bulguları aşağıda sunulmuştur:

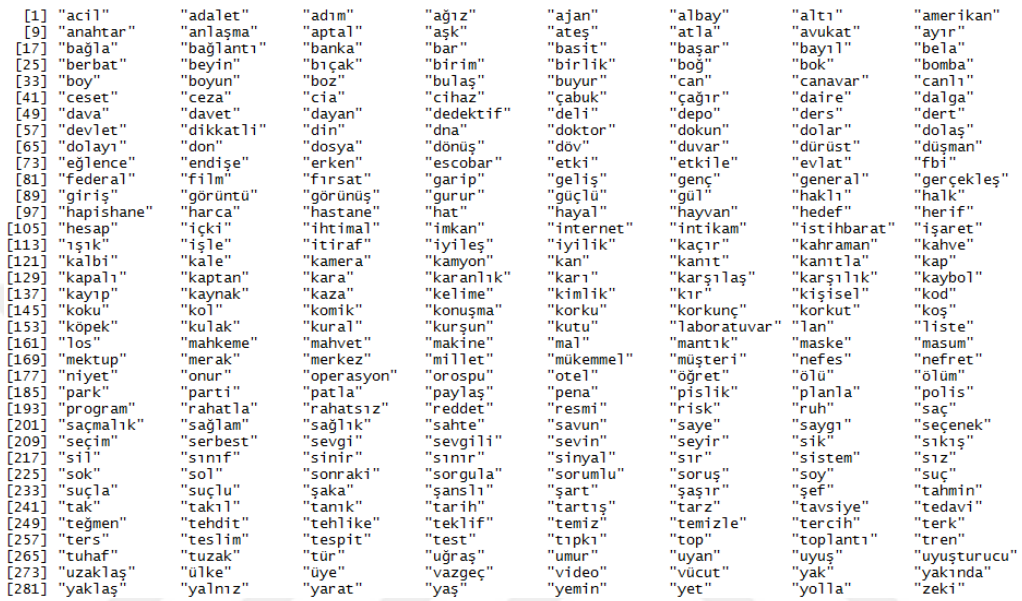
Aksiyon:

Aksiyon türü için çizdirilen wordcloud grafiği Şekil 4.11'de gösterilmektedir.

Şekil 4.11 Aksiyon Kategorisi Wordcloud Grafiği

Bu dizi türü yapılan analiz sonunda anlamlı iki kümeye ayrılmıştır. Birinci kümede 288, ikinci kümede 2491 terim yer almaktadır. Küme içi kareler toplamının genel kareler toplamına oranı %55.8'dir.

Birinci kümede "*ceset, bomba, dedektif, kan, polis escobar, uyuşturucu"* gibi birinci dereceden suç odaklı kelimeler bir araya toplanmıştır.



[1] "acil"	"adalet"	"adım"	"ağız"	"ajan"	"albay"	"altı"	"amerikan"
[9] "anahtar"	"anlaşma"	"aptal"	"aşk"	"ateş"	"atla"	"avukat"	"ayır"
[17] "bağla"	"bağlantı"	"banka"	"bar"	"basit"	"başar"	"bayıl"	"bela"
[25] "berbat"	"beyin"	"bıçak"	"birim"	"birlik"	"boğ"	"bok"	"bomba"
[33] "boy"	"boyun"	"boz"	"bulaş"	"buyur"	"can"	"canavar"	"canlı"
[41] "ceset"	"ceza"	"cia"	"cihaz"	"çabuk"	"çağır"	"daire"	"dalga"
[49] "dava"	"davet"	"dayan"	"dedektif"	"deli"	"depo"	"ders"	"dert"
[57] "devlet"	"dikkatli"	"din"	"dna"	"doktor"	"dokun"	"dolar"	"dolaş"
[65] "dolayı"	"don"	"dosya"	"dönüş"	"döv"	"duvar"	"dürüst"	"düşman"
[73] "eğlence"	"endişe"	"erken"	"escobar"	"etki"	"etkile"	"evlat"	"fbi"
[81] "federal"	"film"	"fırsat"	"garip"	"geliş"	"genç"	"general"	"gerçekleş"
[89] "giriş"	"görüntü"	"görünüş"	"gurur"	"güçlü"	"gül"	"haklı"	"halk"
[97] "hapishane"	"harca"	"hastane"	"hat"	"hayal"	"hayvan"	"hedef"	"herif"
[105] "hesap"	"içki"	"ihtimal"	"imkan"	"internet"	"intikam"	"istihbarat"	"işaret"
[113] "ışık"	"işle"	"itiraf"	"iyileş"	"iyilik"	"kaçır"	"kahraman"	"kahve"
[121] "kalbi"	"kale"	"kamera"	"kamyon"	"kan"	"kanıt"	"kanıtla"	"kap"
[129] "kapalı"	"kaptan"	"kara"	"karanlık"	"karı"	"karşılaş"	"karşılık"	"kaybol"
[137] "kayıp"	"kaynak"	"kaza"	"kelime"	"kimlik"	"kır"	"kişisel"	"kod"
[145] "koku"	"kol"	"komik"	"konuşma"	"korku"	"korkunç"	"korkut"	"koş"
[153] "köpek"	"kulak"	"kural"	"kurşun"	"kutu"	"laboratuvar"	"lan"	"liste"
[161] "los"	"mahkeme"	"mahvet"	"makine"	"mal"	"mantık"	"maske"	"masum"
[169] "mektup"	"merak"	"merkez"	"millet"	"mükemmel"	"müşteri"	"nefes"	"nefret"
[177] "niyet"	"onur"	"operasyon"	"orospu"	"otel"	"öğret"	"ölü"	"ölüm"
[185] "park"	"parti"	"patla"	"paylaş"	"pena"	"pislik"	"planla"	"polis"
[193] "program"	"rahatla"	"rahatsız"	"reddet"	"resmi"	"risk"	"ruh"	"saç"
[201] "saçmalık"	"sağlam"	"sağlık"	"sahte"	"savun"	"saye"	"saygı"	"seçenek"
[209] "seçim"	"serbest"	"sevgi"	"sevgili"	"sevin"	"seyir"	"sik"	"sıkış"
[217] "sil"	"sınıf"	"sinir"	"sınır"	"sinyal"	"sır"	"sistem"	"sız"
[225] "sok"	"sol"	"sonraki"	"sorgula"	"sorumlu"	"soruş"	"soy"	"suç"
[233] "suçla"	"suçlu"	"şaka"	"şanslı"	"şart"	"şaşır"	"şef"	"tahmin"
[241] "tak"	"takıl"	"tanık"	"tarih"	"tartış"	"tarz"	"tavsiye"	"tedavi"
[249] "teğmen"	"tehdit"	"tehlike"	"teklif"	"temiz"	"temizle"	"tercih"	"terk"
[257] "ters"	"teslim"	"tespit"	"test"	"tıpkı"	"top"	"toplantı"	"tren"
[265] "tuhaf"	"tuzak"	"tür"	"uğraş"	"umur"	"uyan"	"uyuş"	"uyuşturucu"
[273] "uzaklaş"	"ülke"	"üye"	"vazgeç"	"video"	"vücut"	"yak"	"yakında"
[281] "yaklaş"	"yalnız"	"yarat"	"yaş"	"yemin"	"yet"	"yolla"	"zeki"

Şekil 4.12 Aksiyon Kategorisi Birinci Küme Terimleri

İkinci kümede ise *"cinsel, intihar, işkence, kaçak"* gibi ikincil suç türleri bir arada görülmektedir.

[1] "açı"	"ada"	"adli"	"ağır"	"ağla"	"akıllı"	"alarm"
[8] "ambulans"	"amerikalı"	"aşırı"	"barış"	"baskı"	"başlangıç"	"batı"
[15] "bedel"	"belge"	"borç"	"bunca"	"cam"	"cehennem"	"cinsel"
[22] "çene"	"çoğu"	"dağıt"	"dans"	"davranış"	"delik"	"delil"
[29] "deniz"	"derin"	"detay"	"dil"	"diş"	"doğrula"	"doğu"
[36] "doğum"	"duygu"	"düşür"	"eğitim"	"elbise"	"eleman"	"engelle"
[43] "eşya"	"eyalet"	"fayda"	"fena"	"gaz"	"gerekli"	"gez"
[50] "görüşme"	"göt"	"gözük"	"güney"	"hani"	"hastalık"	"hızlı"
[57] "hizmet"	"hoşlan"	"ifade"	"ikna"	"ilaç"	"imzala"	"incit"
[64] "intihar"	"iptal"	"istasyon"	"istek"	"işkence"	"kaçak"	"kahret"
[71] "kana"	"kanun"	"kar"	"karşıla"	"kasa"	"kavga"	"kemik"
[78] "kenar"	"keyif"	"kilise"	"kıpırda"	"kira"	"kırmızı"	"kıyafet"
[85] "komiser"	"kov"	"kredi"	"kulüp"	"kuvvet"	"laf"	"lider"
[92] "madde"	"mavi"	"mekan"	"mermi"	"mesela"	"müdür"	"müsaade"
[99] "noel"	"onayla"	"öğrenci"	"özle"	"paket"	"pardon"	"pencere"
[106] "piç"	"randevu"	"renk"	"rica"	"rus"	"salon"	"sayı"
[113] "seri"	"sert"	"seviye"	"sıcak"	"sıkıntı"	"sıradan"	"site"
[120] "sonsuz"	"sus"	"süper"	"sürpriz"	"şiddet"	"şifre"	"tara"
[127] "tat"	"tebrik"	"teknoloji"	"tepe"	"umut"	"uyar"	"uzman"
[134] "vaka"	"varlık"	"veri"	"virüs"	"yarı"	"yerleş"	

Şekil 4.13 Aksiyon Kategorisi İkinci Küme Terimleri

Aksiyon türünde toplanan 6350 tweet ile elde edilen wordcloud grafiği Şekil 4.14'deki gibidir.

Şekil 4.14 Aksiyon Kategorisi Tweet Wordcloud Grafiği

Grafikten görüleceği üzere izleyiciler dizilerde yer alan şiddet unsurlarından ziyade dizi ve dizi karakterlerine duydukları hayranlığı dile getiren cümleler kurmuşlardır. Dizi türünde öne çıkan kelimeler ile izleyicilerin attığı tweetlerde ön plana çıkan kelimeler arasında benzerlik ya da uyum görülmemektedir.

Cinsellik:

Cinsel içerikli dizilerin analiziyle elde edilen wordcloud grafiği Şekil4.15'de gösterilmiştir.

Şekil 4.15 Cinsellik Kategorisi Wordcloud Grafiği

Bu kategoride ön plana çıkan kelimeler içinde seks terimleri dikkat çekmektedir. Analiz sonucunda bu dizi türü anlamlı iki kümeye ayrılmıştır. Birinci kümede 15, ikinci kümede 226 terim yer almaktadır. Küme içi kareler toplamının genel kareler toplamına oranı %65.8'dir.

Birinci kümede *"aşk, aile, kalp, seks, sevgili"* gibi aşk ve sevgi temalı kelimelerin bir arada bulunduğu görülmektedir.

```
[1] "acı"      "affet"   "aile"   "aşk"      "bebek"  "bok"   "kalp"   "kan"
[9] "lan"      "metres"  "seks"   "sevgili"  "şarap"  "yat"   "yatak"
```

Şekil 4.16 Cinsellik Kategorisi Birinci Küme Terimleri

İkinci kümede ise *"boşal, fahişe, kaltak, sok, tecavüz"* gibi cinsellik üzerinden şiddet içerikli kelimelerin bir arada toplandığı görülmektedir.

```
 [1] "arzu"     "ateş"     "avukat"      "ayarla"      "bağımlı"  "bedel"
 [7] "boşal"    "ceza"     "doğum"       "dokun"       "dol"      "eğlence"
[13] "erken"    "etkile"   "evlilik"     "fahişe"      "fransız"  "göğüs"
[19] "göt"      "hamile"   "hasiktir"    "ibne"        "içki"     "idam"
[25] "iğrenç"   "kahve"    "kaltak"      "kıç"         "kontrol"  "korkunç"
[31] "köpek"    "orospu"   "parmak"      "piç"         "risk"     "sarhoş"
[37] "seksi"    "sert"     "sıç"         "sikik"       "sikişme"  "sok"
[43] "soy"      "suçla"    "suçlu"       "şişe"        "taşak"    "tecavüz"
[49] "tedavi"   "tehlike"  "uyuşturucu"  "üniversite"  "vücut"    "yalvar"
[55] "zevk"     "zorla"
```

Şekil 4.17 Cinsellik Kategorisi İkinci Küme Terimleri

Komedi:

Komedi kategorisindeki dizilerin analiziyle elde edilen wordcloud grafiği Şekil4.18'de gösterilmiştir.

Şekil 4.18 Komedi Kategorisi Wordcloud Grafiği

Dizi kategorisinden de anlaşıldığı gibi ön plana çıkan kelimeler eğlence içeriklidir. Bu dizi türü anlamlı 3 kümeye ayrılmıştır. İlk kümede 9, ikinci kümede 20, üçüncü kümede ise 634 terim yer almaktadır.

Birinci kümede *"arkadaş, hayat, para"* gibi iyi hissettiren kelimeler bir arada görülmektedir.

```
[1] "arkadaş" "dost"    "düşün"   "erkek"   "hayat"   "hisset"  "ihtiyaç" "kadın"
[9] "para"
```

Şekil 4.19 Komedi Kategorisi Birinci Küme Terimleri

İkinci kümede *"aşk, ilişki, sevgili, eğlence"* gibi aşk temalı kelimeler bir arada görülmektedir.

```
 [1] "acı"     "akıl"    "aptal"   "aşk"     "ayrıl"   "eğlence" "evlen"   "film"
 [9] "hayal"   "içki"    "ilişki"  "karar"   "komik"   "kork"    "köpek"   "parti"
[17] "sevgili" "sik"     "şans"    "takım"
```

Şekil 4.20 Komedi Kategorisi İkinci Küme Terimleri

Üçüncü kümede ise genel ve günlük konuşma terimleri içeren kelimeler bir arada toplanmıştır.

```
  [1] "alerji"     "alkış"      "anahtar"    "araştır"    "asil"       "asyalı"
  [7] "avla"       "bakıcı"     "bale"       "belgesel"   "bıçak"      "binlerce"
 [13] "bulmaca"    "bulut"      "cilt"       "çaba"       "çap"        "çılgınlık"
 [19] "çim"        "dansçı"     "dar"        "değer"      "demir"      "deneyim"
 [25] "ders"       "dikkatli"   "dönüş"      "dürüst"     "ekip"       "elbise"
 [31] "emir"       "etek"       "etiket"     "fabrika"    "fakülte"    "feci"
 [37] "geleneksel" "gelin"      "geliş"      "gençlik"    "ger"        "gevrek"
 [43] "geyik"      "gitar"      "göbek"      "hamburger"  "hasta"      "havaalanı"
 [49] "havlu"      "herkül"     "heykel"     "hiçbiri"    "hindistan"  "hisse"
 [55] "hit"        "hizmetçi"   "iltifat"    "ırkçı"      "ister"      "kalın"
 [61] "kanat"      "kanka"      "kaplumbağa" "karış"      "kayıt"      "kemer"
 [67] "keşfet"     "klinik"     "koç"        "koleksiyon" "korsan"     "kov"
 [73] "köprü"      "kucak"      "kum"        "kupon"      "kutlu"      "küçümse"
 [79] "küvet"      "lokanta"    "lüks"       "maaş"       "mayo"       "mektup"
 [85] "meslek"     "milyar"     "mini"       "modern"     "mor"        "mücevher"
 [91] "müzikal"    "nakit"      "orijinal"   "pahalı"     "papel"      "piknik"
 [97] "poşet"      "rahat"      "rahatsız"   "rahip"      "randevu"    "rezalet"
[103] "saçı"       "sağlam"     "salon"      "sanatçı"    "satıcı"     "savun"
[109] "sebze"      "senatör"    "sepet"      "sinyal"     "sıvı"       "sızla"
[115] "sokak"      "süsle"      "şık"        "tara"       "tebrik"     "tel"
[121] "tercih"     "tıka"       "titre"      "tomun"      "topla"      "torun"
[127] "umutsuz"    "yastık"     "yığın"      "yurt"
```

Şekil 4.21 Komedi Kategorisi Üçüncü Küme Terimleri

5. SONUÇ

Analiz sürecinde, yapısal olmayan dizi altyazıları ve aksiyon dizi türüne ait twitter yorumları metin madenciliği yöntemleriyle öncelikle yapısal hale getirilmiştir. Dizi altyazıları ITU Türkçe Doğal Dil İşleme Yazılım Zinciri sistemi kullanılarak rakamlardan, noktalama işaretlerinden, beyaz boşluklardan ve linklerden arındırılmıştır ve kelime köklerine ayrılmıştır. Metin editörleri aracılığıyla eklerden arındırılan kökler R'da analiz sürecine sokulmuştur. Kelimelerin koordinat sistemindeki konumları belirlenmiş, uzaklık ölçüsü için Öklit Uzaklığı kullanılmıştır. Ardından k-means kümeleme algoritması kullanılarak altyazılar kendi türünde anlamlı kümelere ayrılmıştır.

Aksiyon tweetlerinin analizi için öncelikle Twitter'dan yorum çekmek için gerekli adımlar izlenmiştir. Her bir dizi için API Twitter developer hesabı aracılığıyla API oluşturulmuştur. API'lerin oluşturulmasının ardından önce R'da daha sonra sanal sunucu üzerinden toplanan tweetler Türkçe tweetleri ayırabilmek için C Sharp'ta gerekli işlemlerden geçirilmiştir. Elde edilen tweetler altyazılar ile aynı işlemlere sırasıyla sokulmuştur. Ancak hem aksiyon içerik türüne ait kelimelerin yorumlarda sık kullanılmamış olması hem de kümelerin anlamlı bir yapıda ayrışmaması sebebiyle kümeleme analizinden vazgeçilmiş, kelimelerin sadece wordcloud grafiği çizdirilmiştir.

Aksiyon dizilerinin kümeleme analizinden anlamlı iki küme elde edilmiştir. Birinci kümede katil ve ceset temalı birinci derecen suç teşkil eden kelimeler ve polis, ajan, dava gibi suçla mücadele içerikli kelimeler yer alırken ikinci kümede cinsellik ve darp gibi ikincil suçları temsil eden kelime grupları bir arada toplanmıştır. Toplam 2779 terimden 288'i birinci dereceden suçları temsil eden kümede toplanmıştır. Aksiyon dizilerinde suç ve suç unsurları ön plana çıkmaktadır. Şiddet içeriğinin dizi temasında fazlaca işlendiği görülmektedir. Buradan aksiyon dizilerinin yer yer küfür yada nefret söylemi içeriyor olması haricinde türüne ve içeriğine uygun altyazılara sahip olduğu sonucuna varılabilir.

Aksiyon dizi türüne ait Twitter yorumları genellikle dizilerin teması üzerine değil oyuncu ve sahnelerin ihtişamına duyulan hayranlığı ön plana çıkarmıştır. Dizi altyazıları ile atılan tweetler arasında bir benzerlik ya da uyum görülmemiştir.

Cinsel içerikli dizilerin altyazıları aynı yöntemlerle temizlenip analiz sürecine sokulmuş; k-means kümeleme analizi sonucunda anlamlı iki küme elde edilmiştir.

Terim sayısı az olan küme seks ve cinsellik odaklı ve yüksek sıklıklarla kullanılan terimlerden oluşurken terim açısından oldukça yoğun olan ikinci kümede özellikle kadını aşağılayan, nefret söylemleri ile cinsellik üzerinden şiddet içerikli kelimelerin yer aldığı görülmektedir. Bu sonuç cinsel içerikli dizilerin cinsellikten ziyade nefret ve şiddet odaklı olduğunun bir göstergesidir. Toplam 241 terimden sadece 15'i yalın cinsellik içermektedir. Cinsel içerikli dizilerde cinsiyetçilik ön plana çıkmaktadır. Kadın bireyden ziyade seks objesi gibi görülmektedir.

Komedi dizilerinin altyazıları aynı yöntemlerle temizlenip analiz sürecine sokulmuş; k-means kümeleme analizi sonucunda anlamlı üç küme elde edilmiştir. Birinci kümede hayat ve arkadaşlık temalı iyi hissettiren kelimeler yer almaktadır. Küme büyüklüğü 9 terim olan ve en az eleman sayısına sahip olan bu kümede yer alan kelimeler dizilerde en sık kullanılan kelimelerdir. İkinci küme aşk ve ilişkiler ile ilgili terimleri bünyesinde bulundurmaktadır. Toplam 663 terimin 20'si bu kümededir. Dizilerde yoğun olarak kullanılan kelimelerdir. Üçüncü küme ise günlük hayat ve konuşmaları ifade eden terimlerden oluşmaktadır. Bu kümenin dizi türü açısından diğer iki küme gibi belirleyici bir özelliği yoktur. Buradan komedi dizilerinin türüne ve içeriğine uygun altyazılara sahip olduğu sonucu çıkarılabilir.

Sonuç olarak dizi aksiyon ve komedi dizi türü içeriğine uygun senaryolarla hazırlanırken cinsel içerikli dizilerin saptırıcı ve kışkırtıcı senaryolarla beslendiği yargısına varılmıştır. Cinsel içerikli diziler sadece türü göz önünde bulundurularak bile belli bir yaş altında izleyici kitlesi için uygun değilken belli yaşın üstündeki izleyici kitlesi üzerinde de olumsuz etkiler bırakabilir. Cinsiyet açısından kutuplaştırabilir.

İnternet ortamnında ulaşıma açık olan ve her yaş kitlesinin izleyebileceği bu diziler için BTK ve TİB gibi organların bu tarz çalışmaları baz alarak ilgili siteye uyarı sembolleri eklemesiyle izleyici kitlesinin bilinçlendirilmesi sağlanabilir.

Ayrıca yapılan bu çalışmada kullanılan tüm analizler tüketim sektörü perakende sektörü vb topluma hizmet ve fayda sağlayacak tüm sektörlerde kullanılabilir. Hedef müşteri kitlelerini çözümlemek isteyen tüm bu sektör üyeleri ilgili analizler sayesinde henüz ulaşamadıkları sosyal medyada yer alan büyük veriden de yararlanabilir. Böylelikle yapılan tüm text mining çalışmaları aynı zamanda CRM çalışmalarına da önayak olup toplum istek ve görüşlerini bu açıdan da değerlendirerek fayda sağlamış ve rekabet piyasası içerisinde bir adım daha öne geçirmiş olacaktır.

Model için kullanılan yöntem çok fazla manuel müdahaleye maruz kalmıştır. Genellikle metin madenciliği alanında dökümanlar arasındaki benzerlik hiyerarşik kümeleme ile analiz edilir. Bu çalışmada ilgili dizi türüne ait tüm diziler ITU sistminin kullanımı sırasında çıkan aksaklıkların nedeniyle ayrı ayrı belgelenmemiştir. Türe ait tüm diziler tek bir metin belgesinde derlendikten sonra köklere ayırmak için arayüz adımlarından geçirilmiştir. Haliyle tek bir belgede ve tek bir kolonda yer alan kelimeler arasında birliktelik kuralları (hangi iki kelimenin ard arda kullanılması ve kullanım yüzdesi gibi) ve cosine benzerlik ölçüsü kullanılamamıştır. Yine tek bir kolonun olması sebebiyle R analiz sürecinde küme sayısı belirlemek için NbClust algoritması kullanılamamıştır. Dolayısıyla algoritmanın içerisinde yer alan 30 farklı indeksleme yönteminden bahsedilmemiştir. Kurulan model basit ancak işlevli bir modeldir. Tek bir türe ait diziler üzerinden her bir dizinin altyazıları farklı belgelerde yer alacak şekilde belgelerin sırayla ITU Türkçe Doğal Dil İşleme Yazılım Zinciri sisteminden geçirilmesi ve kelimelerin sırası bozulmayacak şekilde R analiz sürecinden geçirilmesi halinde küme modeli geliştirebilir ve farklı algoritmalar ile daha kararlı hale getirilebilir. Ancak kelimelerin sırasını bozmadan hatalı ayrılan kökleri ve İTÜ sisteminden inen dosyada yer alan anlamsız simgeleri temizlemek -java ve programlama bilgisi olmayanlar için- uzun ve sabır gerektiren bir süreçtir. Bu nedenle belge sayısını optimum sayıda seçmek önem arz eder.

KAYNAKLAR

[1] **Teorey t. J.,** (1998), *Database Modeling & Design*, Morgan Kaufmann Publishers, USA.

[2] **Giudici P.,** (2003), *Applied Data Mining: Statistical Methods for Business and Industry*, Wiley, England.

[3] **Hand D., Mannila H., Smyth P.,** (2001), *Principles of Data Mining*, MIT Press, Cambridge, MA.

[4] **Gökay Emel, G. Ve Taşkın, Ç.,** (2005), *Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması*, Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi Cilt:6 Sayı: 2, 224.

[5] **Fayyad U., Piatetsky-Shapiro G., Smyth P.,** (1996), *Knowledge Discovery and Data Mining: Towards a Unifying Framework*, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland.

[6] **Şen, F.**, (2008), *Veri Madenciliği ile Birliktelik Kurallarının Bulunması*, Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya.

[7] **Bölükbaş, M.A.,** (2013), *Veri Madenciliği Teknikleri Kullanılarak Çalışan Memnuniyetinin İncelenmesi*, Yüksek Lisans Tezi, Mimar Sinan Güzel Sanatlar Üniversitesi, İstanbul.

[8]**Nisbet R., Elder J., Miner G.,** (2009), *Handbook of Statistical Analysis and Data Mining Applications*, Elsevier Inc., Printed in Canada.

[9] **Kantardzic M.,** (2011), *Data Mining: Concepts, Models, Methods and Algorithms*, Wiley-IEEE, New Jersey.

[10] **Roiger, R. J.,** (2017), *Data Mining: A Tutorial-Based Primer*, CRC Press, North West.

[11] **Miner, G., Elder, IV John., Hill, T., Nisbet, R., Delen, D., ve Fast, A.,** (2017)**,** *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Academic Press, USA.

[12] **Berry, M.W., Kogan, J.**, () Text Mining Applications and Theory

[13] **İTÜ Doğal Dil İşleme Yazılım Zinciri**, http://tools.nlp.itu.edu.tr/, Erişim tarihi: 13/06/2017.

[14] **Kemik: Doğal Dil İşleme Grubu**, http://www.kemik.yildiz.edu.tr/, Erişim tarihi: 11/03/2017.

[15] **Statsoft**, http://www.statsoft.com, Erişim tarihi: 14/09/2016

[16] **Veri Madenciliği ders notları,** http://kergun.baun.edu.tr/veri_madenciligi_hafta11.pdf Erişim tarihi: 10/08/2016.

[17] **Veri Analitiği**, http://www.verianalitigi.org/metin-madenciligi/metin-madenciliginin-asamalari/, Erişim tarihi: 11/02/2017.

[18] **Jockers, M. L.,** (2014), *Text Analysis with R for Students of Literature*, Springer, London.

[19] **Wiedemann, G**., (2015), *Text Mining for Qualitative Data Analysis in the Social Sciences : A Study on Democratic Discourse in Germany*, Springer VS, Leipzig.

[20] **Srivastavasa, A. ve Sahimi M. (Ed.)**, (2009), *Text Mining: Classification, Clustering, and Applications*, Chapman & Hall/CRC, New West.

[21] **Hoffman, M., ve Chilshom, A. (Ed.)**, (2016), *Text Mining and Visualization : Case Studies Using Open-Source Tools*, CRC Press, New West.

[22] **Datacamp**, www.datacamp.com, Erişim tarihi: 11/06/2017

ÖZGEÇMİŞ

Zahide ÇELİKSU

ADRES: Örnek Mah. Erfelek Sok. No:5
Ataşehir, İstanbul
MEDENI HAL: Bekar
EHLIYET : Yok
TELEFON : +90 (553) 274 54 21
E-MAIL : zahideceliksu@gmail.com

PROFESYONEL DENEYIM

Migros Ticaret A.Ş., İstanbul
Veri Ambarı Uzmanı
Kasım 2015 –

İş Tanımı:

CRM Kampanya Değerlendirme Raporlarının SQL üzerinde tasarlanması ve otomatize edilmesi
İş Birimlerinden talep edilen raporların hazırlanması, analiz edilmesi
SPSS Modeler üzerinde müşteri ve ürün odaklı modelleme ve analizlerin yapılması

Digiturk, İstanbul
İş Geliştirme ve Veri Analizi Uzmanı
Mayıs 2014 – Kasım 2015

İş Tanımı:

Veri analizi yoluyla elde edilecek içgörü ile yeni iş alanlarının oluşturulmasına katkıda bulunma
İzleme ölçümü ve abone bilgisi gibi geniş çaplı veriler üzerinden izleme ve kuşak segmentasyonları ile tüketici hane penetrasyonu yapılması
Digiturk sitelerinin web trafik ölçümlemelerinin ve analizlerinin yapılması ve raporlanması
Yeni projelerin analitik altyapılarının tasarlanması ve datamart yapısına dahil edilmesi için IT ile koordinasyon kurulması

EGITIM BILGILERI

Mimar Sinan Güzel Sanatlar Üniversitesi, İstanbul
Yüksek Lisans – İstatistik *Eylül 2013 –*

Marmara Üniversitesi, İstanbul
Lisans- İstatistik *Eylül 2009 – Haziran 2013*

Canip Baysal Lisesi, Bolu
Süper Lise - Fen Bilimleri Bölümü
Eylül 2002 – Haziran 2006

YABANCI DIL DUZEYI

İngilizce : İyi Düzey

BILGISAYAR BECERILERI

MS Office	:	İleri düzey
SQL	:	İleri düzey
SPSS Modeler	:	İleri düzey
SPSS	:	İleri düzey
R	:	İleri düzey
SAS	:	İleri düzey
Matlab	:	İyi düzey

SEMİNER VE KURSLAR

2010 - 8. Uluslararası İstatistik Kolokyumu
2011 - Türkiye İnovasyon Zirvesi
2013 - 10. Uluslararası İstatistik Kolokyumu

PROJELER

Değerli ürünlerin ve mağazaların belirlenmesi için ürün ve mağaza segmentasyon modellemesi

Büyük çaptaki izleme verileri üzerinden abonelerin izleme hareketlerinden yola çıkarak Segmentasyon Analizi, Birliktelik Analizi vb. veri madenciliği yöntemleriyle tüketici hane profili tahmin etme ve tüketici davranışlarını anlama.

Digiturk web sitelerinin (ligtv.com.tr, turkmaxgurme.com, digiturk.com.tr) aylık web trafik ölçümlemelerini gerçekleştirme ve raporlama.

Yabancı Dizilerin Altyazı ve Twitter Yorumlarının Yazı Analizi (Textminig-YLisans Tezi)

Televizyon Dizilerinin Başarısını Etkileyen Faktörlerin Lojistik Regresyon Analizi ile İncelenmesi (Lisans bitirme tezi; 10. Uluslarası İstatistik Kolokyum'unda sunumu yapıldı.)

İris Çiçeği Türlerinin Diskriminant Analizi ile Sınıflandırılması

Ceza Rejimlerinin Suç Oranları Üzerindeki Etkilerinin Ridge Regresyon Analizi ile Belirlenmesi

Şizofreni Hastalığına Etki Eden Faktörlerin Ridge Lojistik Regresyon Analizi ile İncelenmesi

REFERANSLAR

Birol Yüceoğlu – Bilgi Teknolojileri Arge & Uygulama, Migros Ticaret A.Ş.
Telefon : 0536 614 15 52 e-mail : biroly@migros.com.tr

Banu Özaltun - Medya ve Kurumsal Pazar Araştırmaları, Digiturk Genel Merkez
Telefon : 0532 555 03 69 e-mail : banu.ozaltun@digiturk.com.tr

Yrd. Doç. Dr. Elif Özge Özdamar - Mimar Sinan Güzel Sanatlar Üniversitesi
Telefon : 0542 660 52 22 e-mail : erudio.ozdamar@gmail.com

Prof. Dr. Müjgan Tez - Marmara Üniversitesi İstatistik Bölüm Başkanı
Tel No: 0532 727 61 13 e-mail: mtez@marmara.edu.tr