T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

METİN MADENCİLİĞİ VE DUYGU ANALİZİ: IMDB EN İYİ ÜÇ FİLMİN TWİTTER YORUMLARININ ANALİZİ

Necla KAPUKAYA

YÜKSEK LİSANS TEZİ
İstatistik Anabilim Dalı
İstatistik Programı

Danışman
Dr. Öğr.Üyesi Elif TUNA

Şubat, 2023

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

METİN MADENCİLİĞİ VE DUYGU ANALİZİ: IMDB EN İYİ ÜÇ FİLMİN TWİTTER YORUMLARININ ANALİZİ

Necla KAPUKAYA tarafından hazırlanan tez çalışması 24.02.2023 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı İstatistik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Dr. Öğr.Üyesi Elif TUNA
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Dr. Öğr.Üyesi Elif TUNA, Danışman
Yıldız Teknik Üniversitesi _____

Dr. Öğr.Üyesi Doğan YILDIZ, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr.Üyesi Büşra ŞAHİN, Üye
Haliç Üniversitesi _____

Danışmanım Dr. Öğr.Üyesi Elif TUNA sorumluluğunda tarafımca hazırlanan Metin Madenciliği ve Duygu Analizi: IMDb En iyi Üç Filmin Twitter Yorumlarının Analizi başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Necla KAPUKAYA

İmza

Aileme
ve Engin'e

TEŞEKKÜR

Beni her konuda destekleyen ve yardımlarını esirgemeyen babacığıma, anneciğime, canım abim Hasan'a ve güzel kız kardeşim Hülya'ya; bilgisi, donanımı ve yardımı ile hep yanımda olan sevgili Engin'e; zamanını, ilgisini ve desteğini esirgemeyen canım arkadaşım Hilal'e, bu süreçte tüm tecrübeleriyle beni aydınlatan değerli hocam Elif Tuna'ya teşekkür ederim

Necla KAPUKAYA

İÇİNDEKİLER

ŞEKİL LİSTESİ

ÖZET

Metin Madenciliği ve Duygu Analizi: IMDb En iyi Üç Filmin Twitter Yorumlarının Analizi

Necla KAPUKAYA

İstatistik Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Dr. Öğr.Üyesi Elif TUNA

Çalışmada hedeflenen amaç Twitter Developer hesabı üzerinden R programlama dili kullanılarak IMDb "En İyi 250 Filmler" listesindeki ilk üç film olan Esaretin Bedeli, Baba ve Kara Şövalye filmlerinin Twitter gerçek kişi kullanıcılarının yorumlarına ait metin analizini elde etmektir. IMDb puanlamasına göre yüksek derecelendirilen filmlerin günlük konuşma metninde de duygu değerlerinin benzer olup olmadığı tespit edilmiştir. Çalışmada Twitter platformu üzerinden çekilen yapılandırılmamış verilerin R programlama dili kullanılarak metin ön temizleme işlemleri yapılmış, meta veri haline dönüştürülmüş ve elde edilen temiz veri üzerinden metin madenciliği yöntemleri olan tf-idf, N-gram modelleme ve duygu analizi incelemesi yapılmıştır. Her üç film için de elde edilen kelime bulutları ile kullanıcı yorumlarının duygu analizi incelenmiştir. Çalışmanın ilk bölümünde veri madenciliği, metin madenciliği, metin madenciliğinin uygulama alanlarından bahsedilmiştir. Çalışmamızda kullanılan tekniklere benzer çalışmaların literatür taraması yapılmıştır. Literatürde var olan çalışmaların katkısı ve inceledikleri uygulamadan bahsedilmiştir. Tezin amacına yer verilmiştir. Twitter verilerine ulaşmak için kullanılan Twitter Developer hesabının nasıl elde edildiğini, aşamalarını ve elde edilen kullanıcı hesabı ile ulaşılan API ve şifre detayları hakkında bilgi verilmiştir. İkinci bölümde metin madenciliği metodolojisine yer verilmiştir. Yapılandırılmamış olan metinlerin ön temizleme işleminin nasıl yapıldığına dair bilgi verilmiştir. Yapılandırılmış verilere metin madenciliği tekniklerinden tf-idf, N-gram yoluyla kelimeler arasındaki ilişki ve korelasyon hesaplaması tekniklerinin teorilerinden bahsedilmiştir. Elde edilen temiz veriye duygu analizi teknikleri uygulanmıştır ve kelime bulutları oluşturularak

görselleştirilmiştir. Çalışmanın üçüncü bölümünde uygulamaya ve son bölümde elde edilen sonuç ve değerlendirmeye yer verilmiştir.

Anahtar Kelimeler: Veri madenciliği, metin madenciliği, duygu analizi, makine öğrenmesi, twitter geliştirici hesabı

ABSTRACT

Text Mining and Sentiment Analyses: Analysis of Twitter Comments of IMDb Top Three Movies

Necla KAPUKAYA

Department of Statistics
Master of Science Thesis

Supervisor: Asst. Prof. Elif TUNA

The aim of this study is to obtain a text analysis of the comments of Twitter users (by filtering out bot account) on the first three movies in the IMDb "Top 250 movies" list, namely The Shawshank Redemption, The Gotfather and The Dark Knight, using the R programming language on the Twitter Developer account. It has been determined whether the emotional values of the movies that are rated high according to the IMDb scoring are similar in the daily speech text or not. In the study using the R programming language, the unstructured data captured from the Twitter platform were applied text pre-cleaning process, thereafter processes of transforming into metadata, text mining methods such as tf-idf, N-gram modeling and sentiment analysis were carried out on the obtained clean data. The word clouds obtained for all three movies and the sentiment analysis of user comments were examined. In the first part of the study, data mining, text mining, application areas of text mining are mentioned. A literature review of the studies similar to the techniques used in our study was conducted. The contribution of the existing studies in the literature and the application they examined are mentioned. The purpose of the thesis is mentioned. Information were given on how the Twitter Developer account that is used to access Twitter data was obtained, the stages of obtaining the account and the API and password details accessed with the obtained user account. In the second part, text mining methodology is mentioned. Information on how to pre-clean the unstructured text is given. Structured data, text mining techniques tf-idf, the relationship between words through N-gram and the theory of correlation calculation techniques are mentioned. The application of sentiment analysis techniques to the

clean data obtained is discussed and visualized by creating word clouds. In the third part of the study, the application is mentioned. In the application, unstructured text data of Twitter data for three separate movies were obtained. The obtained data has been made processable by text cleaning processes. Afterwards, metadata sizing and tf-idf analysis comments mentioned in the second part, bi-gram results from n-gram techniques and word clouds created by sentiment analysis are included. In the last section, the results and evaluation are summarized.

Keywords: Data mining, text mining, sentiment analysis, machine learning, twitter developer account

1
GİRİŞ

1.1 Literatür Taraması

Teknolojinin gelişmesinin ardından insanların sosyal medya platformlarında duygu ve düşüncelerini paylaşabildiği ortamlar gelişmiştir. Sosyal medya kullanımının yaygınlaşması, kullanıcı yorumlarının önem kazanması sonucu metin temelli verilerin analizine ihtiyaç duyulmuştur. Bu veri setleri büyük veri olarak tanımlanabileceğinden ihtiyaç duyulan bilginin elde edilmesi daha da zorlaşmaktadır. Büyük verilerden anlamlı sonuçlar elde etmek için araştırmacılar veri madenciliği yaklaşımlarından olan metin madenciliği tekniklerinden yararlanmışlardır. İnsanların duygu, düşünce ve hislerini anlamak için geleneksel yöntemlerden anketler ve diğer araştırma yöntemlerinin ötesinde, basılmış yazılı metinleri ele almak ve bu metinleri de büyük bir veri tabanı gibi analiz ederek sayısal değerlere dönüştürmek metin madenciliğinin temel fonksiyonudur. Metin madenciliği teknikleri kullanılarak; okuyarak analiz etmenin mümkün olmadığı büyüklükteki metinlerin analizi yapılabilmektedir. Yayınlanmış metinler, sosyal medya yorumları, bloglar, haber kaynakları gibi farklı kategorilerden oluşan kaynaklar veri setini oluşturmaktadır. Bu veri setleri yapılandırılmamış şekildedir. Yapılandırılmamış veri seti bilgisayar dilinin anlayamadığı doğal dil ile yazılmış metinlerdir. Yapılandırılmamış veri seti ve içeriğinde bulunan meta datalar (üst bilgiler) veri kaynağı olarak kullanılacak metinlerdir ve bu metinlerden nitelik çıkarımı ve makine öğrenmesi algoritmaları kullanılarak bilgisayar diline hazır hale gelen yapılandırılmış veri elde edilir. Yapılandırılmış metin kaynakları ile metin madenciliği fonksiyonları kullanılarak duygu analizi yapılabilmektedir.

Veri setinin metin madenciliği algoritmaları ile analiz edilmeden önce ön hazırlık sürecinden geçmesi gerekmektedir. Burada çoğunlukla yapısal olmayan veriler üzerinden işlem yapılacağından metni istenilen şekilde (genellikle kelime kelime) parçalayarak dizilere ayırma, metin içerisinde geçen anlamda değişiklik yapmayan kelimelerin atılması, metin içerisindeki noktalama ve sayıların çıkartılması, metindeki büyük küçük harf ayrımının ortadan kaldırılması, metinde geçen kelime eklerinin

atılarak kelime köklerinin kaydedilmesi şeklinde özetleyebileceğimiz ön hazırlık süreci tamamlanır. Veri ön işleme verilerin analiz edilebilmesi için en önemli adımlardan biridir. Sonraki aşamada kümeleme, duygu analizi, karar ağacı, rasgele orman gibi algoritmalar ile verinin analiz edilmesi mümkün olabilecektir. Çalışma R dili ile hazırlanacak olup, ön hazırlık sürecinde R dilinin açık kaynak kütüphanesinden faydalanılmıştır. Bu çalışmada IMDb puanlamasında en iyi üç filmin tweetleri elde edilmiş ve metin madenciliği algoritmaları ile duygu analizi yapılmıştır. Elde edilen sonuçlar değerlendirilmiştir.

Analitik veya veri alanındaki çalışmalar çok hızlı veri üretimine yol açmaktadır. Analistler genellikle tablo halindeki verileri işlemek üzere eğitilmiştir, fakat günümüzde hızla artan verilerin çoğu yapılandırılmamış metin verilerinden oluşmaktadır. Bilgisayarların günlük hayatın bir parçası olması dijital ortamda verilerin toplanması ve saklanmasını da beraberinde getirmiştir. Araştırmacılar dijital ortamlarda oluşturulan haber, duygu ve düşünce temelli kullanıcı yorumları, bloglar ve dergilerden oluşan metin içerikli verileri çalışmalarına konu edinmiştir.

Twitter verilerinin büyük veri olması, kolay ulaşılması ve konu bakımından çeşitlilik göstermesi araştırmacıları bu verileri işlemeye ve analizlerinde kullanmasına yönlendirmiştir. Literatür araştırmasında Twitter verileri kullanılarak yapılan metin madenciliği çalışmaları incelenmiştir. Yerli ve yabancı yapılan çalışmalar metin madenciliğinin gelişmesine önemli katkı sağlamıştır. [1] metin madenciliğinde en sık kullanılan kümeleme analizi “bag of words” yaklaşımını ele almıştır. Analiz yapısal olmayan dizi altyazıları ve aksiyon dizi türüne ait Twitter yorumları metin madenciliği yöntemleri ile öncelikle yapısal hale getirilmiş, sonra k-means kümeleme algoritması kullanılarak altyazılar kendi türünde anlamlı kümelere ayrılmıştır. [2] makine öğrenmesi teknikleri uygulayarak duygu analizi yapmıştır, karar ağaçları ve olasılığa dayalı algoritmalardan faydalanmıştır. İlgili çalışmada R yazılımı ile duygu analizi elde edilmiştir. Duygu analizi için Twitter’dan 24 Haziran 2018’de Türkiye’de yapılan Cumhurbaşkanlığı ve 27. Dönem Millet Vekilli Genel Seçimi’ ne ilişkin en yüksek oy alması beklenen aday ve partilere ilişkin bir ay boyunca atılan tweetler incelenmiştir. Tweetlerden hareketle duygu analizi yapılmış, bu amaçla duygu skoru hesaplanmış ve makine öğrenmesi ile hesaplamalar yapılmıştır. Elde edilen sonuçlar karşılaştırılmıştır. [3] doğal dil analizi tekniklerini Twitter verileri üzerinde uygulamıştır. Çalışmada Twitter verilerini olumlu, olumsuz ve nötr sınıflandırabilen uygun sınıflandırıcı oluşturulmasıdır. Sınıflandırıcı oluşturtulabilmesi için sözlük tabanlı ve n-gramlardan oluşan iki farklı doğal dil işleme. İzlenen iki yöntem için farklı özellik vektörleri oluşturularak sınıflandırıcılar dokuz farklı alandan toplanmış veriler ile eğitilmiştir. Telekomünikasyon, finans, sağlık, sigorta, spor, gıda, otomotiv, gayrimenkul ve siyaset konularından oluşan rasgele seçilmiş 8.321 tweet manuel

olumlu, olumsuz ve nötr olarak etiketlenmiştir. Analizlerde doğal dil işleme, Naive Bayes, destek vektör makinesi ve rastgele orman algoritmaları kullanılmıştır. Hem ayrı ayrı hem de bütün veriler birlikte analiz edilmiş ve bütün verilerin birlikte analiz edildiği durumda başarı ölçütleri kıyaslamasında, en iyi sınıflandırıcının sözlük tabanlı yöntem olduğu görülmüştür. [4] Twitter verileri ile hem sözlük hem de n-gram modeli kullanarak olumlu, olumsuz ve nötr etiketlemesi üzerine iki yöntem geliştirmiştir. Sözlük yöntemi, n-gram yöntemine göre daha başarılı sonuçlar vermiştir. [5] veri üzerinde sınıflandırma, kümeleme ve birliktelik analizi gibi veri madenciliği yaklaşımları uygulamıştır. Hesaplamalar sonucu oluşan terim frekansı ve döküman terim matrisleri ile metin sınıflandırması elde edilmiştir. İlgili çalışmada Rapid Miner yazılımı kullanılarak Karar Ağaçları, Yalın Bayes, 1-En Yakın Komşu ve Destek Vektör Makineleri sınıflandırma algoritmalarının performansları karşılaştırılmıştır. Liu ve arkadaşları [6], blog verilerinden satış performansının tahmin edilmesi için Sentiment-PLSA isimli metodu önermişlerdir. Pennacchiatti ve Popescu [7], Twitter kullanıcılarının davranış, çevre ve tweet gibi bilgilerinden yararlanarak kullanıcıların politik yönelim ve etnik kökenini tespit etmiştir. Twitter üzerinden politik analiz yapan başka çalışmalar da mevcuttur[8] - [9]. Bollen ve arkadaşları [10], tweetlerdeki ruh hali değişikliklerini ele alarak borsa tahmini yapmışlardır. Şimşek ve Özdemir[11], Türkçe tweetler ile borsa endeksi arasındaki ilişkiyi karşılaştırmışlardır. Birjali ve arkadaşları [12], Twitter verisi kullanarak intihar düşüncesi tahmini yapmıştır. Baj-Rogowska [13], Facebook yorumlarını kullanarak Uber hakkında analiz yapmıştır.

1.2 Tezin Amacı

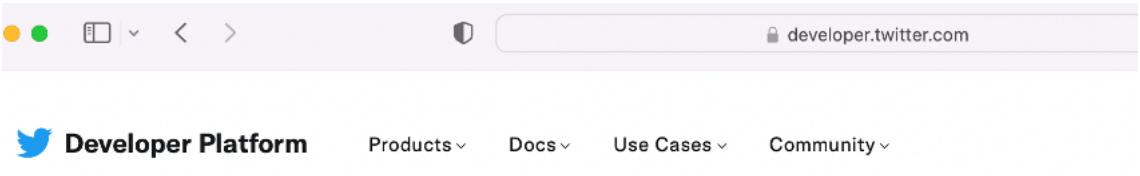
Bu çalışma kapsamında metin madenciliği tekniklerinden tf-idf ve n-gram yaklaşımları ile verideki metin ilişkisini görselleştirmek amaçlanmaktadır. Aynı zamanda Twitter Developer kullanıcı hesabı ve R programlama dili kullanılarak veri çekme konusunda açıklama yapılmaktadır. Çalışma kapsamında internet film veri tabanı olarak değerlendirilen IMDb Top Movies 250 [14] listesindeki ilk üç filmin Twitter kullanıcıları tarafından yazılan metinlerin analizi metin madenciliği yaklaşımları ile analizi incelenmiştir. IMDb filmler, diziler ve televizyon içerikleri hakkında bilgiler barındıran bir bilgi kaynağıdır. Haziran 2020 itibariyle veri tabanında yaklaşık 12 milyon yapım ve 7 milyon kullanıcı incelemesi bulunmaktadır. Elde edilen sonuçların görselleştirilmesi, yorumlanması ve kullanılan tekniklerin karşılaştırılması amaçlanmaktadır.

1.3 Hipotez

Çalışma kapsamında kullanıcıları gerçek kişilerden oluşan iki ayrı platform olan IMDb ve Twitter üzerinden paylaşılan beğenilerin benzerlikleri duygu analizi teknikleri ile test edilmektedir. IMDb puanlamasında beğeni bakımından yüksek puan ile derecelendirilen üç film için Twitter platformundaki kullanıcı yorumları ele alınmıştır. İlgili filmler hakkında paylaşılan tweet metinlerindeki duygu puanlamasında da pozitif duygu kelimelerinin ağırlıklı olduğu ortaya konulacaktır.

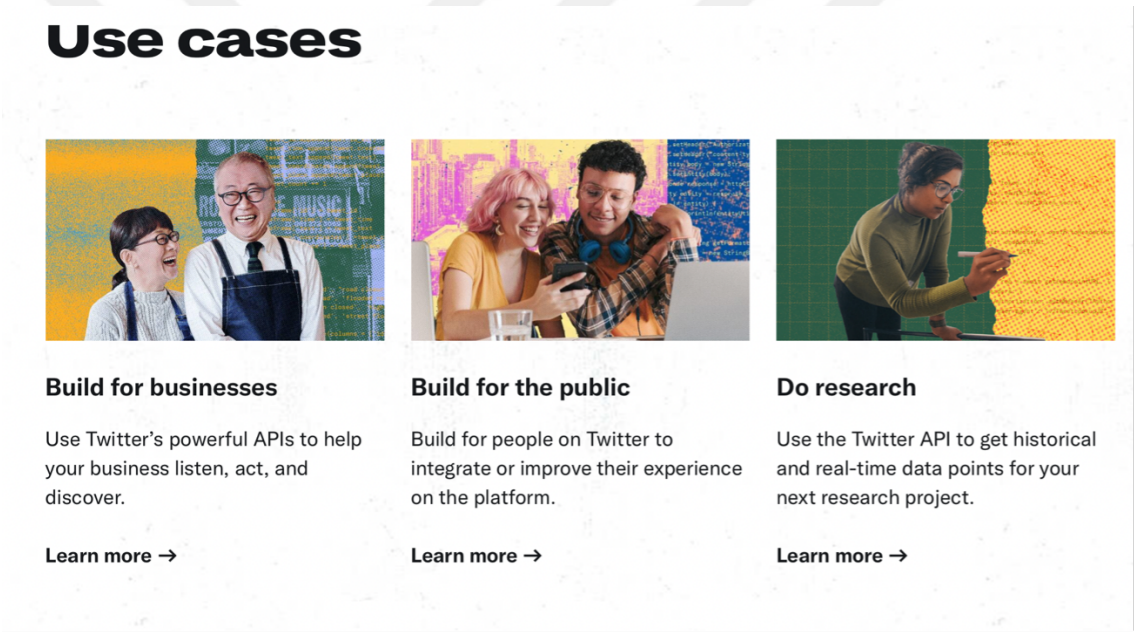
1.4 Twitter Developer

Twitter 2006 yılında Amerika'nın California eyaletinde oluşan bir sosyal paylaşım platformudur. Facebook ve İnstagram gibi ortamlardan sonra en çok kullanılan platformların başında gelen Twitter'da yazı, resim, video yüklemesi yapılabilmektedir. Gündemi güncel bir şekilde takip eden Twitter'da haber başlıkları trend konular arasında yer almaktadır. Sitenin ismi kuş cıvıltısı anlamında olan tweet kelimesinden gelir. Dünyanın her tarafında en çok kullanılan sosyal paylaşım sitelerinden biridir. Kullanıcıların gerçek zamanlı, herhangi bir konu hakkında duygu ve düşüncelerini paylaştığı, kullanıcı bilgileri, lokasyon, zaman, anahtar kelime, hasthtag gibi filtreler barındıran dijital sosyal ağdır. Üyeliği ücretsiz olan ve aylık kullanıcı sayısının 230 milyona, günlük tweet sayısının ise 500 milyona ulaştığı belirtilen sosyal ağ, dünyanın en büyük mikro bloglarından biri olarak değerlendirilmektedir. Bu yönüyle de metin madenciliği için büyük veri ambarı olarak görülmektedir. Nielsen ve Genart Medya tarafından 2014 yılında Türkiye'deki Twitter kullanıcılarının profilini ve alışkanlıklarını belirlemek için yapılan araştırma[15] sonuçlarına göre kullanıcıların büyük çoğunluğu erkektir. Kullanıcıların üçte biri 16-24, üçte biri 25-34 olmak üzere genç kullanıcı grubuna sahiptir. Yine aynı araştırmaya göre kullanıcıların yarısı üniversite mezunu bireylerdir. Twitter bilgilerinin büyük kitlelerce paylaşılabilmesi için şirketlere, geliştiricilere ve kullanıcılara API'ler aracılığıyla program dayalı erişim olanağı sunmaktadır. Bilgisayar programcılığında, Application Programming Interface kavramının kısaltması API olarak kullanılmakta olup, özetle üst düzeyde API'ler bilgisayar programlarının birbirleriyle "konuşma" yöntemidir, bu sayede bilgi akışı sağlayabilirler. API platformu kullanıcıların dünyayla paylaşmayı tercih ettiği herkese açık Twitter verilerine geniş çaplı erişim sağlar. API'lere erişmek isteyen birinin developer hesabı açması gerekmektedir. Developer Twitter hesabına erişmek için kullanıcı www.developer.twitter.com adresini kullanmaktadır.



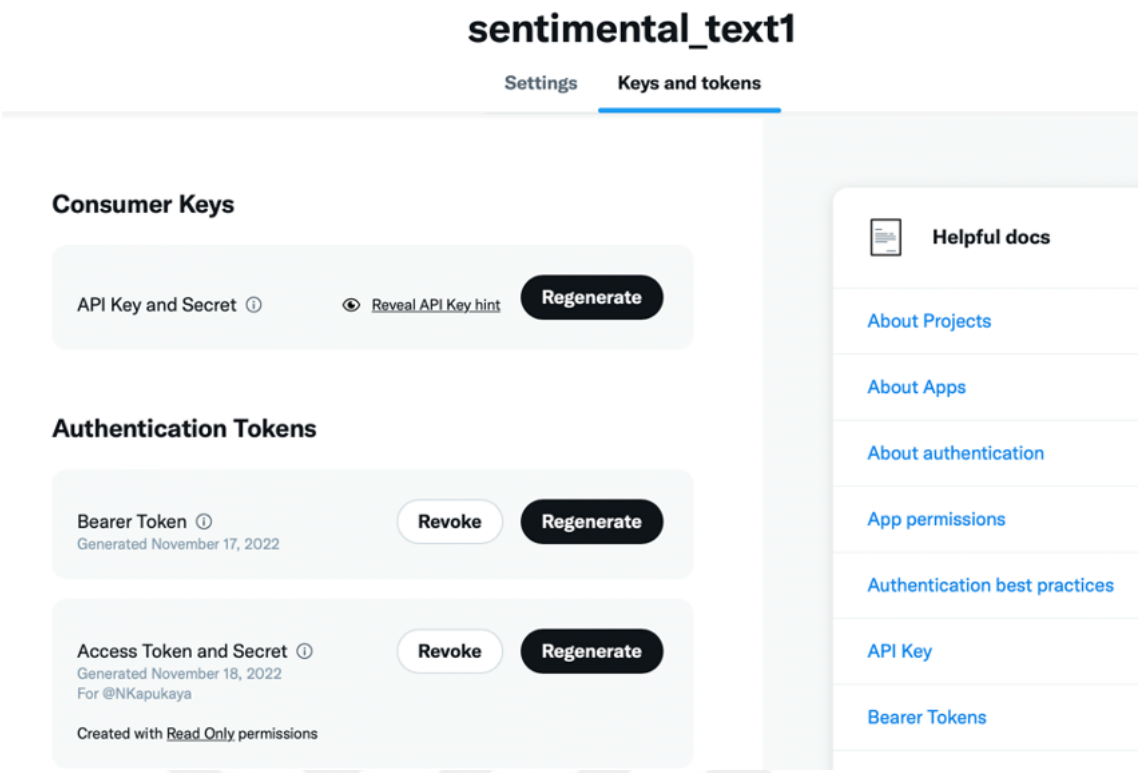
Şekil 1.1 Browser'da Developer Twitter giriş anasayfası

Başvurusu sırasında API'lerin ne amaçla kullanılacağına dair bilgiler paylaşılması istenilmektedir. Bu amaçlardan bazıları ise işletmelerin yorumlarının keşfedilmesine kaynak oluşturmak, akademik bir araştırma yapmak için veri toplamak, kamuyu ilgilendiren gündeme dair metin analizleri için veri toplamak, değerli becerileri öğretmek veya öğrenmek, dünyayı ve Twitter'ı daha iyi bir yer haline getirmek gibi maddelerden oluşmaktadır.



Şekil 1.2 Developer Twitter hesabının kullanım amaçlarına dair ana sayfadaki görsel

"Apply for a developer account" sekmesine tıklandığında developer hesabınızın amacı hakkında bilgi istenmektedir. Daha sonra projeye dair yazılı döküman istenmektedir. Bu bilgilerin Twitter yetkilileri tarafından değerlendirme süreci olumlu tamamlandıktan sonra, kişiye özel şifre bilgilerine erişim verilmektedir. Bu bilgiler sayesinde API'lerin Pyhton, R gibi programlama dilleri ile verilere ulaşım sağlanabilmektedir.



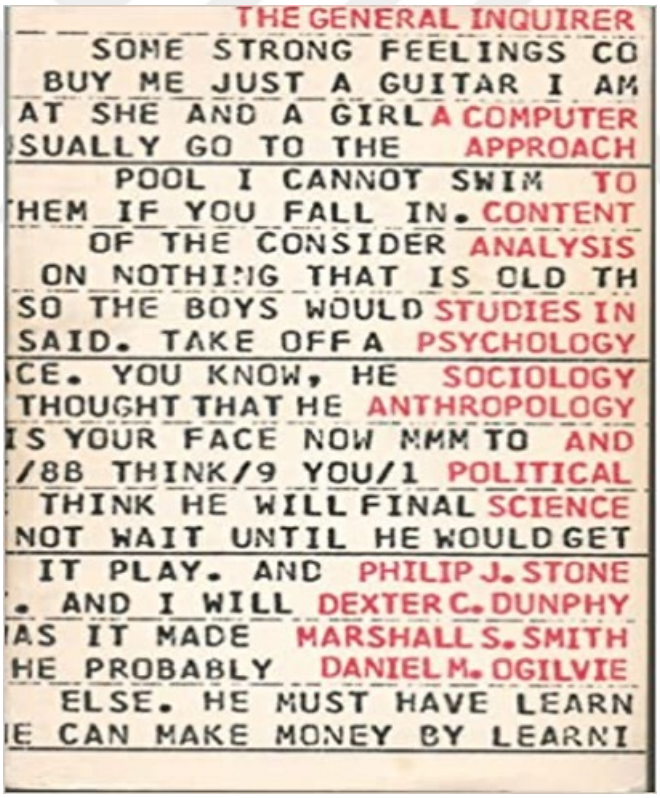
Şekil 1.3 Developer kullanıcı hesabının etkinleştirilmesi ile kişiye özel verilen şifre ve API detayları

Developer hesabının etkinleştirilmesi ile birlikte kişiye özel paylaşılan API detayları, "API Key and Secret" ve "Access Token and Secret" yeniden üret butonuna erişim sağlanmaktadır.

2
YÖNTEMLER VE TEKNİKLER

2.1 Metin Madenciliği Kavramı

İlk kez Stone Dunphy ve Smith tarafından doğal dille yazılmış metinler veri kaynağı olarak analiz edilmek üzere incelenmiştir (1966). Bu çalışma, metin verileri ile finansal ve ekonomik aktiviteler arasındaki ilişkiyi inceleyen ilk çalışmalardan biri olarak değerlendirilmiştir.



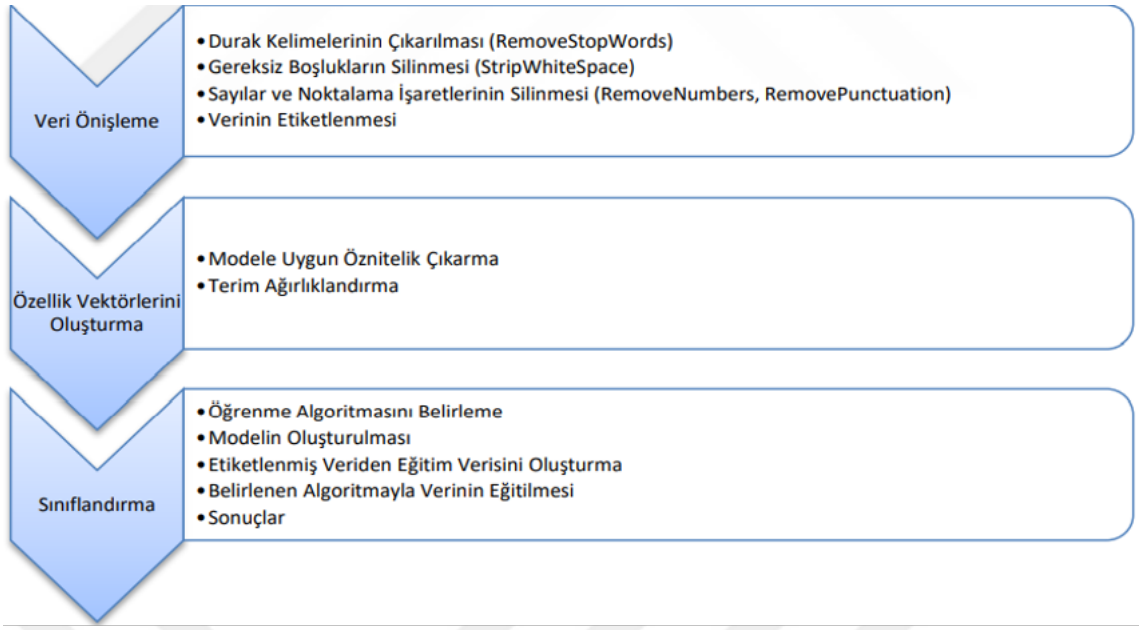
Şekil 2.1 İlk kez analizi yapılan metin verileri

İnsanların ne düşündüğünü, neler hissettiklerini anlamak için geleneksel yöntemler olan anketler ve diğer araştırma yöntemlerinin ötesinde, basılmış yazılı metinleri ele almak ve bu metinleri de büyük bir veri tabanı gibi analiz ederek sayısal değerlere dönüştürmek metin madenciliğinin temel fonksiyonudur[16, s. 128].

Metin madenciliği teknikleri ile okuyarak analiz etmenin mümkün olmadığı büyüklükteki metinlerin analizi yapılabilmektedir. Yayınlanmış metinler, sosyal medya yorumları, bloglar, haber kaynakları gibi farklı kategorilerden oluşan kaynaklar veri setini oluşturmaktadır. Bu veri setleri yapılandırılmamış şekildedir. Yapılandırılmamış veri seti bilgisayar dilinin anlayamadığı doğal dil ile yazılmış metinlerdir. Yapılandırılmamış veri seti ve içeriğinde bulunan meta datalar (üst bilgiler) veri kaynağı olarak kullanılacak metinlerdir ve bu metinlerden özellik çıkarımı ve makine öğrenmesi algoritmaları kullanılarak bilgisayar diline hazır hale gelen yapılandırılmış veri elde edilir. Yapılandırılmış metin kaynakları ile metin madenciliği fonksiyonları kullanılarak duygu analizi yapılabilmektedir. Uygulamada kullanılan çeşitli veri kümeleri bulunmaktadır. Vikipedi'de ansiklopedik bilgi içeren sayfalar genellikle muntazam bir üslup ve objektif bir dille yazılmışlardır. Bu nedenle, referans belge olarak kullanılmaya uygun metinler olarak değerlendirilmektedir. Twitter iletileri kullanıcılar tarafından biçimsel açıdan düzgün olmayabilen ve düzenleme gerektiren veri seti kaynağıdır. 20-Newsgroups dökümanları, Karolina Üniversite'sinin Irvine(UCI) bünyesindeki Yapay Öğrenme ve Akıllı Sistemler Merkezi tarafından, bilimsel çalışmalarda kullanılması adına sağlanan veri kümelerinden biridir. Haber siteleri, bloglar, podcastler, vloglar, mikro bloglar da veri kümelerinden biri olarak değerlendirilmektedir.

2.2 Metin Madenciliği Aşamaları

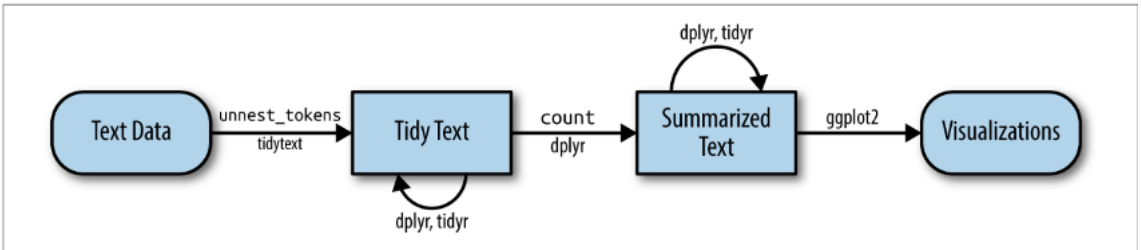
İlk aşama veri setini oluşturmak için ilgilenilen konuya dair yayınlanmış metinlere ulaşmaktır. Veri setine öncelikle ön işleme yapılmalıdır. Metinlerin kelime kelime ayrılması, kelimelerin anlamsal hesaplamalarının ve puanlamalarının elde edilmesi, kelimelerin eklerden ayıklanması, büyük küçük harflerin aynılaştırılması, ifadelerin temizlenmesi, imla hatalarını giderilmesi ve yazım hatalarını ortadan kaldırma işlemlerinin tamamına ön işleme denir. Ön işleme tamamlanınca nitelik belirleme hesaplamaları yapılır. Nitelik belirleme sürecinde; önişleme aşamasında temizlenen verilerdeki önemli kelimelerin tespiti ve ilişkili olmayan niteliklerin (özelliklerin) çıkarılması, az sayıda dökümanda yer alan özelliklerin ayıklanması, çok sayıda dökümanda yer alan özelliklerin azaltılması işlemleri yapılır. Nitelik seçimi ile büyük boyutlu verilerin boyutu azaltılmış, içinden niteliksiz terimler ayıklanmış, metnin analiz edilmeye hazır hale getirilir. Bu süreç sayesinde sınıflandırmanın başarılı bir şekilde yapılması olasılığını arttırmaktadır. Sınıflandırma aşaması aynı belgelerin birlikte sınıflandırılması süreci olarak ifade edilebilir. Temel amaç; metinleri anlamsal olarak, önceden belirlenmiş sınıflara otomatik olarak ayırmaktır. Metin madenciliğinin en önemli aşamalarından biri de elde edilen analiz sonuçlarının değerlendirilmesidir.



Şekil 2.2 Metin madenciliği aşamaları[17, s. 28]

2.3 Düzenli Metin Biçimi

Düzenli metin biçimi, her satırında bir hece bulunan tablo şeklinde tanımlanır. Bir hece, analiz için kullanmakla ilgilendiğimiz anlamlı bir metin birimi olan sözcüğe karşılık gelmektedir. Böylelikle metin simgelere bölünmüş olur. Bu satır başına token (one-token-per-row) yapısı şu anki analizlerde kullanılan string ya da döküman terim matrisi yapısının tersidir. Tidy text yapısında her satıra bir kelime depolanırken n-gram yapısında bir cümle ya da paragraf da depolanabilir. R Studio kütüphanesinde bulunan tidy text paketi metin verisini heceleştirmemize yardımcı olur. Veri manipülasyonu diplyr, tidyr ve broom gibi paketler kullanılarak yapılabilmektedir. Düzenli veri yöntemlerini kullanarak metin analizi yapmak aşağıdaki gibi görselleştirilmiştir

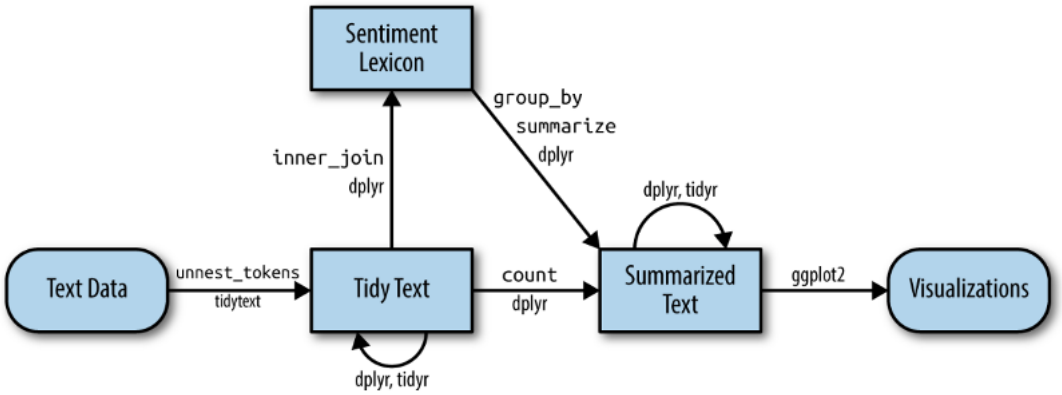


Şekil 2.3 Yapılandırılmış metin biçiminin görselleştirilmesi aşamaları[18, s. 9]

2.4 Düzenli Veri ile Duygu Analizi

Metnin duygusal içeriğine programlı bir şekilde yaklaşmak için metin madenciliği araçlarından faydalanılmaktadır. Duygu analizi metinlerde elde edilen tutum ve

görüşleri anlamanın bir yolunu sunmaktadır. Metin verileri düzenli bir veri yapısında olduğunda, duygu analizi bir iç birleşim olarak uygulanabilir. Bir kitabın seyri boyunca nasıl değiştiğini veya belirli bir metin için duygu ve düşünce içeriğine sahip hangi kelimelerin önemli olduğunu anlamak için duygu analizi kullanılabilmektedir.



Şekil 2.4 Yapılandırılmış metin verilerinin duygu analizi ve görselleştirilmesi aşamaları[18, s. 10]

2.4.1 Ağırlıklandırma : tf-idf

Metin madenciliği ve doğal dil işlemede temel soru: bir belgenin ne hakkında olduğunun nasıl ölçüleceğidir. Kelimenin bulunma frekansına göre metin içindeki önem ölçüsüne terim sıklığı(tf) ile ifade edilir. Bir başka yaklaşım, bir terimin ters döküman sıklığına (idf) bakmaktır.

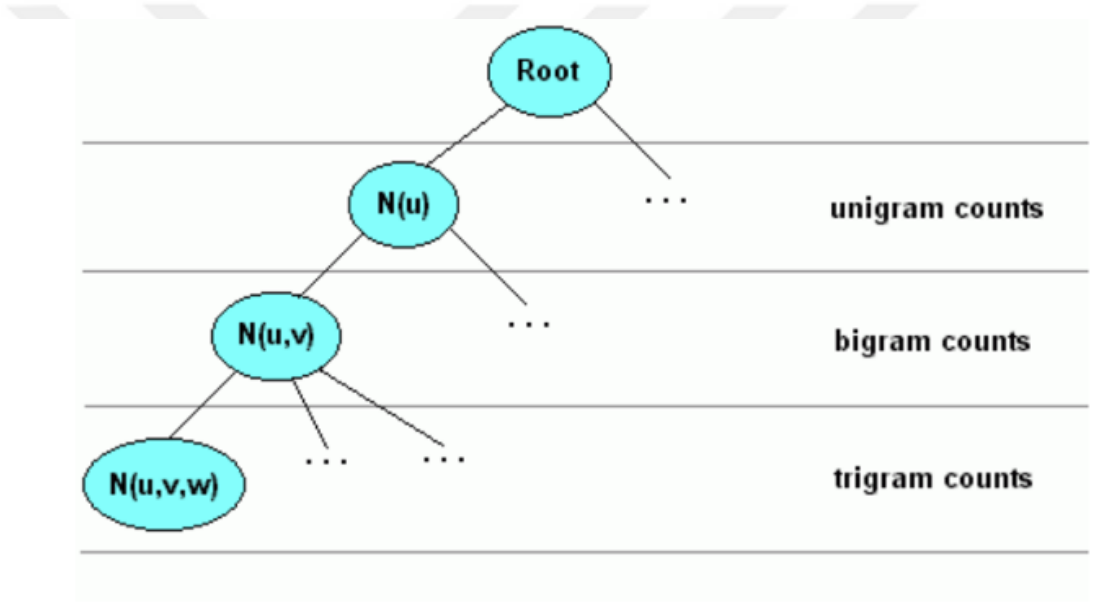
- Terim Sıklığı (TF): Metin içinde bir terimin kaç kez geçtiğini gösteren en temel ağırlıklandırma ölçüsüdür. Bu hesaplamada metinde tekrar eden kelimelerin daha önemli olduğu düşülmektedir.
- Ters Döküman Sıklığı (IDF): Ters döküman sıklığı kelimenin metin için ne kadar belirleyici olduğunu gösteren bir yaklaşımdır. Kelime tek bir metinde geçiyor ise IDF değeri yüksek hesaplanır. Fakat, kelime her metinde geçiyorsa veya hiçbir metinde geçmiyorsa IDF değeri sıfır olarak hesap edilir.

Metin madenciliğinde en çok kullanılan yaklaşım ise TF-IDF yaklaşımıdır. Bu yaklaşımda kelimenin hem kendi bulunduğu metin içindeki terim frekansı hem de diğer metinlerde bulunması önem belirtir.

2.4.2 Kelimeler Arasındaki İlişki: N-grams ve Korelasyon

N-gram yönteminde bir metin veri setinin ardışık n elemanlı alt kümelerinin olduğu tekniktir. Bu yöntem ile metinler içerisindeki gizli örüntüler keşfedilebilmektedir. Metinler için n-gram analizi dil algılama, konuşma tanıma, doğal dil işleme gibi alanlarda kullanılmaktadır. Ayrıştırma için kelimeler her n-gram n kelimeden oluşacak şekilde modellenir.

- bi-gram: -m, me, et, ti, in, n-
- tri-gram: -me, met, eti, tin, in-, n—
- four-gram: -met, meti, etin, tin-, in–, n—



Şekil 2.5 N-gram modelleri[19]

N-gram modelleri, doğal dil işleme alanında kullanılan bir dil modelleme tekniğidir. Bu teknik, bir metindeki ardışık kelime gruplarının sıklıklarını hesaplamak ve dil bilgisel yapıyı öğrenmek için kullanılır. N-gram modelleri, bir metnin ardışık n kelime grubunun sıklığını hesaplamak için kullanılır. N, 1'den başlayarak istenen herhangi bir sayıya kadar olabilir. Örneğin, bir 2-gram modeli, metindeki ardışık iki kelimelik grupların sıklığını hesaplar. Bu şekilde, bir 3-gram modeli, metindeki ardışık üç kelimelik grupların sıklığını hesaplar. N -gram modelleri, metindeki kelime sıklıklarını hesaplamak için kullanılır ve bu sayede bir metnin olasılığını tahmin etmek veya bir kelime verildiğinde bir sonraki kelimeyi tahmin etmek gibi çeşitli doğal dil işleme görevleri için kullanılabilirler.

N-gram modelleri, dil bilgisel yapıyı öğrenmek için kullanıldığından, otomatik dil tanıma, otomatik metin sınıflandırma, otomatik dil çevirisi ve metin oluşturma gibi doğal dil işleme görevlerinde yaygın olarak kullanılırlar. Örneğin, bir 2-gram modeli, "bir kedi" kelime grubunun sıklığını hesaplar ve eğer bu kelime grubu sık sık kullanılıyorsa, bu kelime grubunun "bir köpek" kelime grubundan daha olası olduğu sonucuna varabilir. Bu sayede, bir sonraki kelimeyi tahmin etmek veya bir metnin olasılığını hesaplamak için kullanılabilen bir dil modeli elde edilir. Bir metindeki tüm olası n-gram kombinasyonlarının sayısı çok büyük olabileceğinden n-gram modelleri, büyük veri setleri için hesaplama zorluklarına yol açabilir. Bu sorunu çözmek için, n-gram modelleri genellikle düşük sıklıkta olan n-gramları düşürür veya bir sözlük veya kelime dağarcığı kullanarak belirli kelime gruplarını sınırlar. Sonuç olarak, n-gram modelleri, doğal dil işleme alanında yaygın olarak kullanılan bir dil modelleme tekniğidir. Dil bilgisel yapıyı öğrenmek ve bir metnin olasılığını hesaplamak için kullanılabildikleri gibi, çeşitli doğal dil işleme görevlerinde de kullanılabilirler. N-gram modelleri, doğal dil işleme görevlerinde yaygın olarak kullanılmalarına rağmen, bazı dezavantajları da vardır. Bunlar şunlardır:

- Dil bilgisel Yapıdaki Bağımlılıkların Kısmi Öğrenimi: n-gram modelleri, dil bilgisel yapıdaki tüm bağımlılıkları öğrenemez. Bu nedenle, özellikle uzun cümlelerde yanlış sonuçlar verebilirler.
- Düşük Frekanslı N-Gramlar: Düşük frekanslı n-gramlar, veri setinde sadece birkaç kez göründüğü için model tarafından atlanabilir. Bu nedenle, nadir kelime gruplarının anlamı ve kullanımı hakkında bilgi kaybedilebilir.
- Kelime Sıralamasındaki Esneklik: n-gram modelleri, kelimelerin belirli bir sırayla kullanılacağını varsayar. Ancak, bazı durumlarda, kelime sırası değişebilir ve anlam hala korunabilir. Bu esneklik, n-gram modellerinin doğruluğunu azaltabilir.
- Genelleştirme Yeteneği: n-gram modelleri, veri setinde gördükleri kelime gruplarıyla sınırlıdır ve bu nedenle, genelleştirme yapmak zor olabilir. Bu nedenle, yeni veya farklı veri setleriyle çalışırken yanlış sonuçlar üretebilirler.

N-gram modelleri, doğal dil işleme alanında yaygın olarak kullanılmalarına ve birçok görevde başarılı sonuçlar vermiş olmalarına rağmen, bu dezavantajlar sebebiyle daha gelişmiş dil modelleme teknikleri geliştirilmektedir.

3
UYGULAMA

Bu bölümde çalışmanın uygulama aşaması yer almaktadır. Uygulama esnasında kullanılan programlama dili R Studio olup, veri seti Twitter Developer hesabına bağlanarak R Studio üzerinden elde edilmektedir. Çalışma IMDB listesinde en çok izlenen üç film olan "Esaretin Bedeli" "The Godfather" "The Dark Knight" etiketi kullanılarak yazılan tweetlerin metin madenciliği teknikleri ile analizini kapsamaktadır. Çalışmada metin madenciliği analizlerinden faydalanmak üzere "textdata", "tm", "wordcloud", "RColorBrewer", "RWeka", "tidytext" faydalanılmıştır. RStudio kütüphanelerinden "R0Auth" ve "twitter" paketlerinden faydalanılmıştır. R0Auth paketi Twitter hesabının R Studio üzerinden erişim yetkisini sağlarken, TwitteR paketi Twitter API için ara yüz sağlamaktadır. Bu çalışmada Twitter API'ye erişim için "twitteR" paketi kullanılmıştır. Duygu analizi için "syuzhet" ve "sentiment" paketlerinden, kelime bulutu oluşturmak için "RColorBrewer", "wordcloud" paketlerinden faydalanılmıştır.

```
1  install.packages("rtweet")
2  install.packages("tidyverse")
3  install.packages("httpuv")
4  install.packages("twitteR")
5  install.packages("tidytext")
6  install.packages("textdata")
7  install.packages("tm")
8  install.packages("RColorBrewer")
9  install.packages("RWeka")
10 install.packages("sentimentr")
11 install.packages("syuzhet")
12 install.packages("NLP")
13 install.packages("wordcloud")
14 library(rtweet)
15 library(tidyverse)
16 library(httpuv)
17 library(twitteR)
18 library(tidytext)
19 library(tm)
20 library(sentimentr)
21 library(syuzhet)
22 library(wordcloud)
23
```

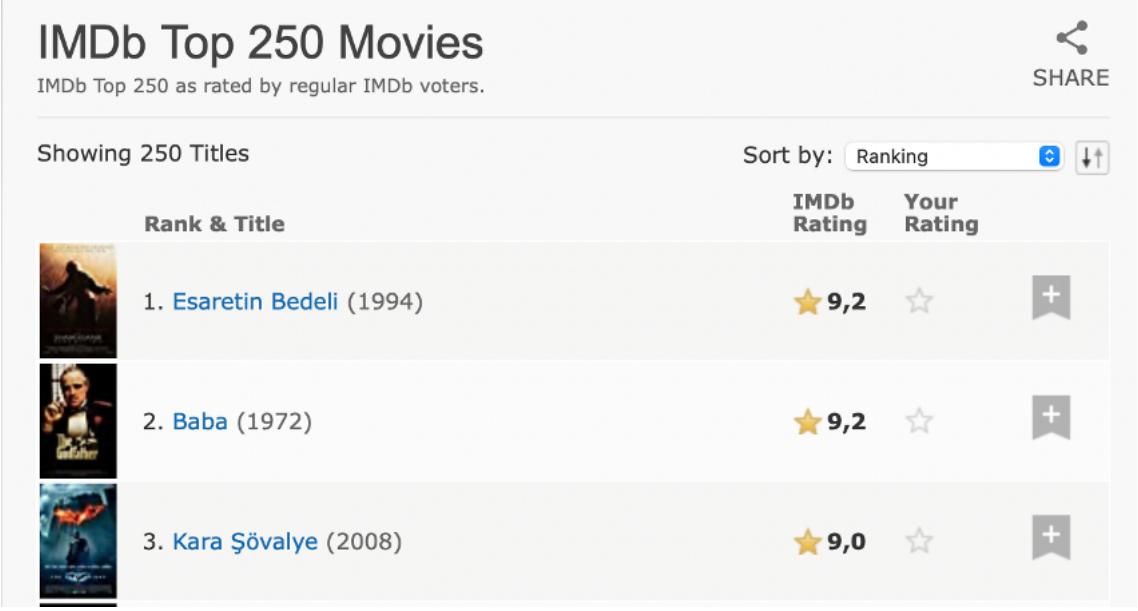
Source the contents of the active document

Şekil 3.1 Analizde kullanılan paketler ve kütüphaneler

```
appname <- "sentimental_text1"
key <- "[illegible]EqTDEH2uwxPmaHEoo"
secret <- "Q7zo5cE1oft[illegible]Xf1i9uT1DTDK1WoAT6cdsBN"
access_token <- "274[illegible]JqMN8UxkMzVXyt7K5ybqtzdmebv"
access_secret <- "nlPzR5EB20q84TcCCUh1IETVchdTI[illegible]izYLo"
setup_twitter_oauth(key, secret , access_token , access_secret )
```

Şekil 3.2 Kişiye özel paylaşılan şifreler

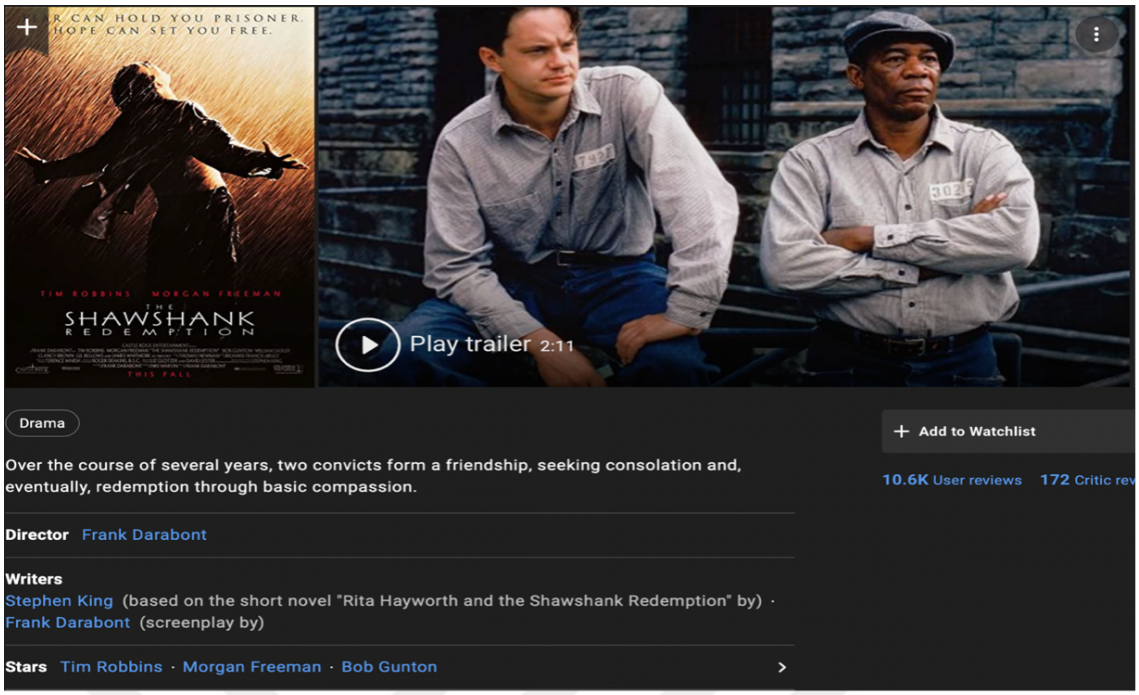
Veriye erişim sağlamak için Twitter Developer hesabının kişiye özel paylaştığı API detayları, key, secret, access token ve access secret şifreleri kullanılmıştır.



Şekil 3.3 IMDb en iyi iki yüz elli film listesindeki ilk üç film

1994 yapımı Esaretin Bedeli filminin IMBD kullanıcı puanlaması 9.3 ile ilk sıradadır.

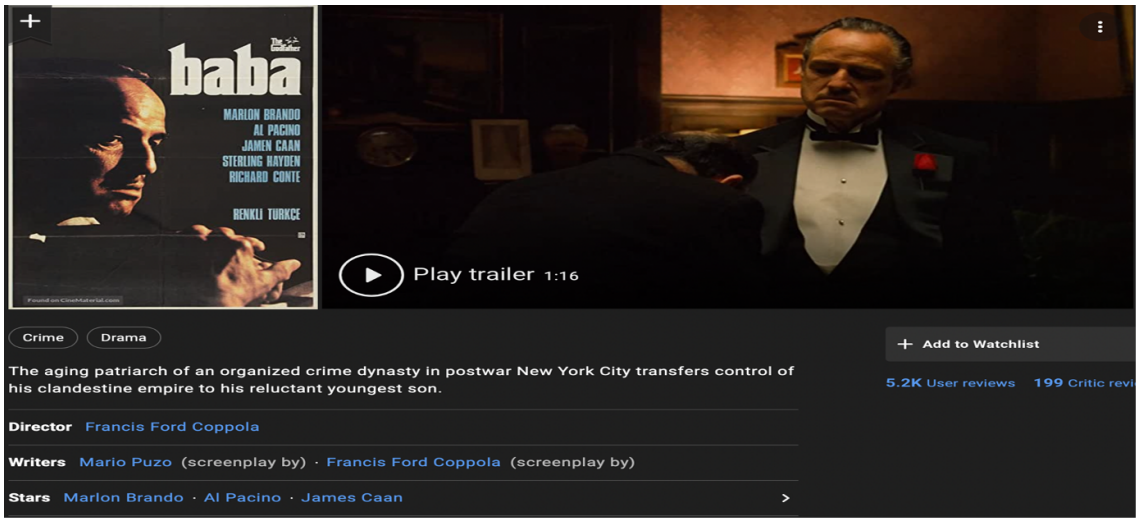
Şekil 3.4 Esaretin Bedeli filmine ait IMDb kullanıcı puanlaması



Şekil 3.5 Esaretin Bedeli filmine ait yönetmen, senarist, yıldız oyuncu bilgileri

1972 yapımlı Baba filmi 1,9 milyon kullanıcı puanlaması ile 9.3 puana sahip olup, 2. sıradadır.

Şekil 3.6 Baba filmine ait IMDb kullanıcı puanlaması

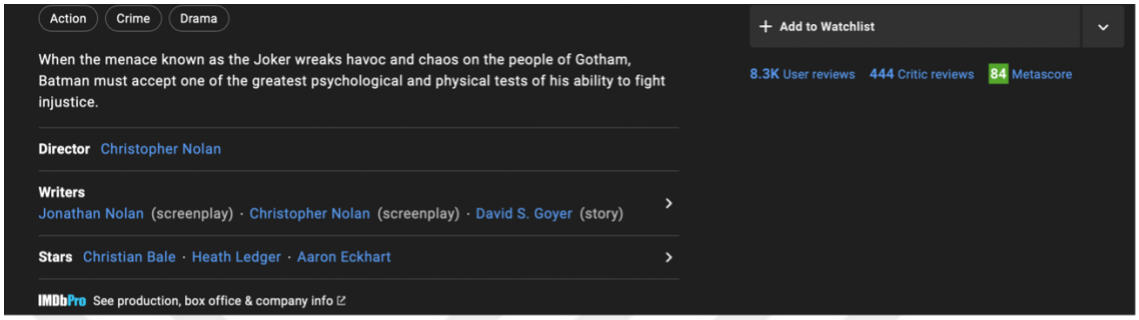


Şekil 3.7 Baba filmine ait yönetmen, senarist, yıldız oyuncu bilgileri

2008 yapımı The Dark Knight filmi IMDB puanlamasında 2,7 milyon kullanıcı

tarafından 10 üzerinden 9 puan almıştır.

Şekil 3.8 Kara Şövalye filmine ait IMDb kullanıcı puanlaması



Şekil 3.9 Kara Şövalye filmine ait yönetmen, senarist, yıldız oyuncu bilgileri

İlk filmin analizini aşağıdaki kod ile veri çağırılarak başlatılmıştır. 01 Ocak 2022 tarihinden itibaren "Esaretin bedeli" etiketi ile yazılan 2000 gözlemli metin verileri elde edilmiştir.

```
32
33  esaretin_bedeli <- searchTwitter(' The Shawshank Redemption ' , n=2000 , lang = 'en',since ='2022-01-01'  )
34  head(esaretin_bedeli)
35
```

Şekil 3.10 Esaretin Bedeli filmine ait 'The Shawshank Redemption" etiketi ile ingilizce dilinde yazılmış iki bin adet tweetin elde edilmesine dair kod

```
> esaretin_bedeli <- searchTwitter(' The Shawshank Redemption ' , n=2000 , lang = 'en',since ='2022-01-01'  )
> head(esaretin_bedeli)
[[1]]
[1] "TinDogPodcast: RT @lance73: @TinDogPodcast Shawshank Redemption\nTop Gun\nThe Green Mile"

[[2]]
[1] "ShawnUpchurch: \"Remember, Red. Hope is a good thing, maybe the best of things, and no good thing ever dies.\" ~Andy Dufresne,
Shawshank Redemption"

[[3]]
[1] "IndraKu42936167: RT @hvgoenka: Best movies of all time as per your votes (in order of your preference):\n\nINDIAN\n1. 3 Idiots\
n2. Sholay\n3. Mayabazar\n4. Gol Ma…"

[[4]]
[1] "lance73: @TinDogPodcast Shawshank Redemption\nTop Gun\nThe Green Mile"

[[5]]
[1] "thearunborana: Hope is a good thing &amp;amp; Good things never die.\n\n(Brooke from the Shawshank redemption )"

[[6]]
[1] "Blueblot: @Bennytouzine The Shawshank Redemption"

>
```

Şekil 3.11 Elde edilen iki bin tweetin ilk altı örneği

Veri setini oluşturan değişkenlerin isimleri için names komutundan faydalanılmıştır.

```
> names(esaretin_bedeli_df)
 [1] "text"          "favorited"     "favoriteCount" "replyToSN"
 [5] "created"       "truncated"     "replyToSID"    "id"
 [9] "replyToUID"    "statusSource"  "screenName"    "retweetCount"
[13] "isRetweet"     "retweeted"     "longitude"     "latitude"
>
```

Şekil 3.12 Veri setini oluşturan değişkenlerin isimleri

Veri setinin değişkenlerinden "text" başlığını inceleyeceğiz. Text verilerinden 1. text örneğini incelemek için head komutundan yararlanılmıştır. İlgili metinde kullanıcı ID bilgisi, tweet ve tweete dair link bilgisi yer almaktadır.

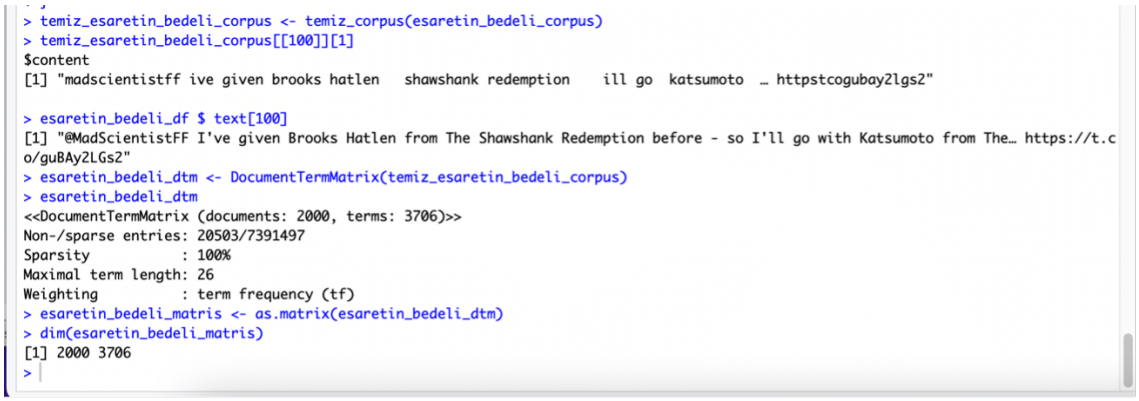
```
> esaretin_bedeli_text <-esaretin_bedeli_df$text
> head(esaretin_bedeli_text,1)
[1] "@narfault The Green Mile, Goodwill Hunting, Life is beautiful, Saving Private Ryan, The Shawshank Redemption,
Dil s… https://t.co/mktjuGf6qY"
```

Şekil 3.13 Veri setinin içindeki text değişkeninin yeni bir veri seti olarak tanımlanması

Bu metin verilerini incelenecek bir kütüphane haline getirmek için tm paketinin VectorSource ve VCorpus fonksiyonları kullanılır. VCorpus iç içe oluşturulmuş listeler olarak düşünülebilir. Bu listede esas incelenmek istenen metne ve bu metine bağlı meta dataya ulaşılır. Oluşturulan corpusu analiz edebilmek için öncelikle metin temizliği işlemi uygulanmalıdır. Tweetlerin içerisinde okunamaz karakterleri kaldırmak, büyük küçük harflerin hepsini küçük harfe dönüştürmek, noktalama işaretlerini, sayıları ve fazla boşlukları kaldırmak için veri seti ön işleme aşamasından geçer.

```
temiz_corpus <- function(corpus){
  corpus <- tm_map(corpus, stripWhitespace)
  corpus <- tm_map (corpus, removePunctuation)
  corpus <- tm_map(corpus, content_transformer(tolower))
  corpus <- tm_map(corpus, removeWords, c(stopwords("en")))
  return(corpus)
}
temiz_esaretin_bedeli_corpus <- temiz_corpus(esaretin_bedeli_corpus)
temiz_esaretin_bedeli_corpus[[100]][1]
```

Şekil 3.14 Metin temizleme işlemine dair oluşturulan fonksiyon



Şekil 3.15 Temizlenen veriyi döküman terim matris formatına dönüştürme ve TF uygulama kodu

Temizlenen corpuslar veri matrisi haline getirilmek üzere döküman terim matrisi fonksiyonundan faydalanılır. Düzenlenen döküman terim matrisi cümlelerdeki kelime frekanslarının hesaplanmasını elde etmemizi sağlayacaktır.

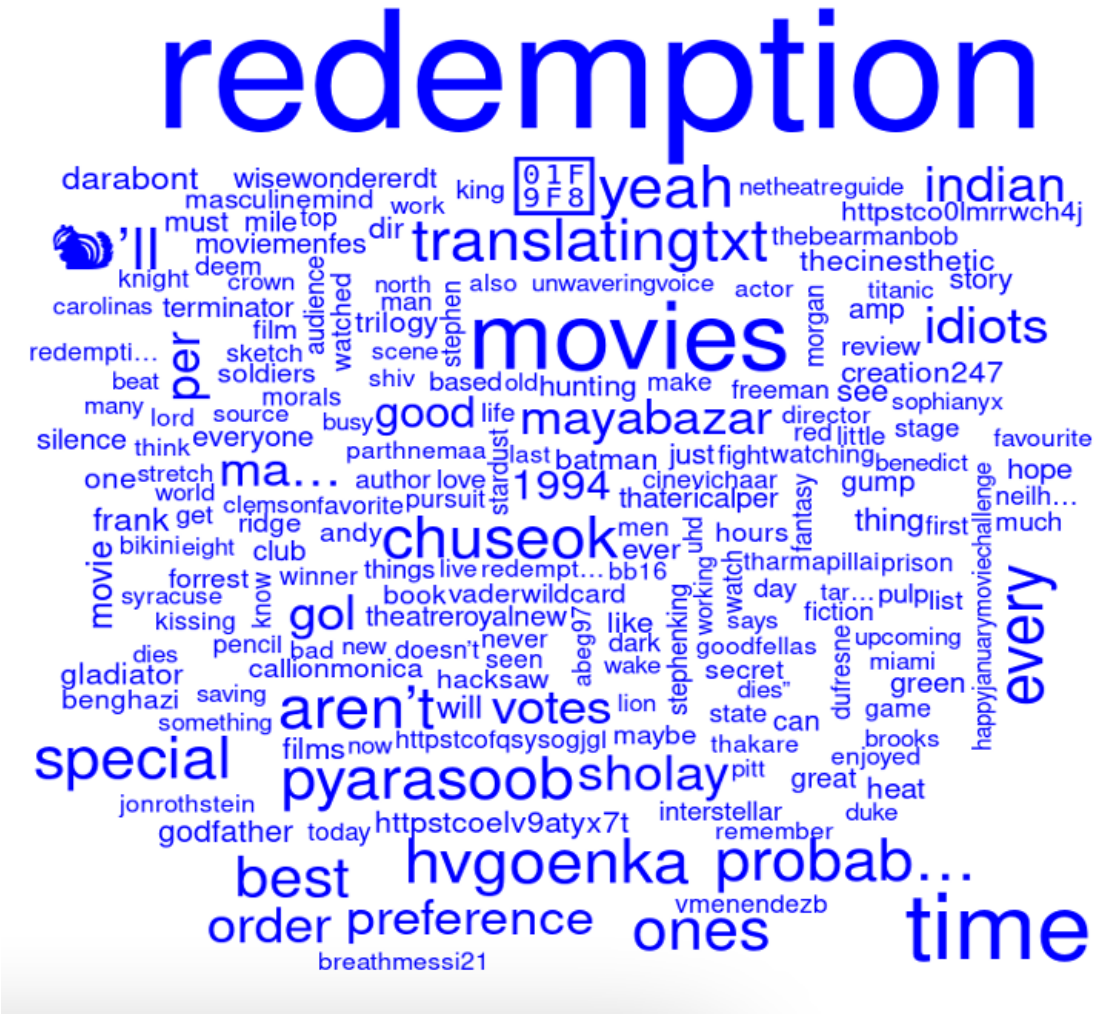
Şekil 3.16 Döküman terim matrisinin sıklık hesaplaması

Elde edilen frekanslarda filmin ismi olan "shawshank" ve "redemption" kelimeleri haricinde atılan tweetlerde 587 adet "movies" kelimesi, 566 adet "time" kelimesine, 317 adet "hvgoenka" kelimesine rastlanılmıştır. Ön işlemeden geçen verilerin analizinde kelime bulutlarından faydalanacaktır. Kelime bulutları, metin verilerindeki en sık kullanılan kelimelerin görselleştirilmesi için kullanılır. Bu nedenle, kelime bulutları metnin ana fikrini veya konusunu anlamamıza yardımcı olabilirler. Aşağıda kelime bulutlarını yorumlama adımları verilmiştir:

- Kelimelerin boyutu: Kelime bulutunda, kelimenin boyutu o kelimenin metinde ne kadar sık kullanıldığını gösterir. Büyük boyutlu kelimeler, metinde daha sık kullanılan kelimelerdir. Dolayısıyla, kelime boyutları metindeki kelimenin önemini belirtir.
- Kelimelerin renkleri: Kelime bulutundaki kelime renkleri, kelime boyutu gibi, kelimenin sıklığına göre ayarlanabilir. Ancak, kelime renkleri, farklı anlamlara veya konulara göre ayrılmış kelimeleri göstermek için de kullanılabilir.
- Kelimelerin konumu: Kelime bulutundaki kelime konumu, kelimenin ilişkili olduğu diğer kelimelerle gösterilir. Ayrıca, kelime bulutundaki kelime konumları, belirli bir kelimenin metindeki bağlamını belirleyebilir.

- Kelimelerin sıklığı: Kelime bulutları, metinde en sık kullanılan kelimeleri görselleştirir. Dolayısıyla, kelime bulutları, metnin anahtar kelimelerini ve önemli konularını belirlememize yardımcı olabilir.

Genel olarak kelime bulutları, metnin anahtar kelimelerini, sık kullanılan kelimeleri ve belirli bir konu veya konulara odaklanan kelimeleri görselleştirir. Dolayısıyla, bir kelime bulutunda en büyük boyuta sahip kelimeler genellikle metnin anahtar kelimeleridir ve metnin konusu veya odak noktası hakkında bir fikir verebilirler. Sıklık sayısına göre elde edilen kelimeler kelime bulutu aracı kullanılarak görselleştirme yapıldığında aşağıdaki sonuç elde edilmiştir.



Şekil 3.17 Döküman terim matrisinden oluşan sıklık sonuçlarına ait kelime bulutu

Kelime bulutu incelendiğinde sıklık tablosu ile oluşturulan bulutta Twitter kullanıcı yorumlarına göre en çok kullanılan kelimeler best, good, special gibi olumlu yorumlar, 1994 yapım tarihi, Andy, Frank gibi filmdeki karakter ismi ile karşılaşılmıştır. Sıklığa

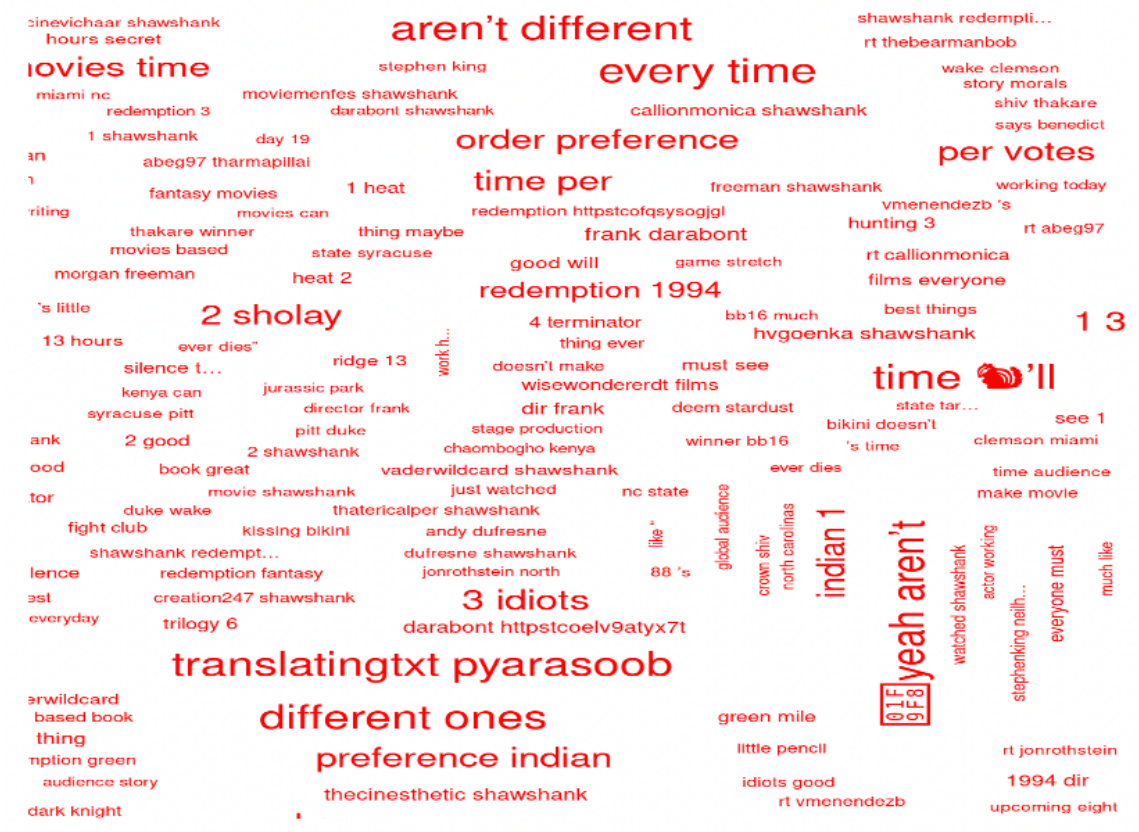
göre oluşturulan kelime bulutları ilgili film hakkında ilk gözlemi sunabilmektedir. Eğer kelime bulutu filmdeki senaryo metinlerinden oluşturulmuş olsaydı sıklığa göre oluşturulan buluttan yaklaşık olarak ana fikre ulaşılabilirdi. Sıklığa göre oluşturulan bulut konu hakkında fikir edinmemize olanak tanımaktadır. Kelime bulutları, metnin genel temasını hızlı bir şekilde anlamak için kullanışlıdır, ancak metnin ayrıntılarını veya nüanslarını belirlemek için yeterli olmayabilirler. Dolayısıyla, kelime bulutları, metnin tamamını anlamak için bir araç olarak kullanılabilirler, ancak tek başlarına yeterli değillerdir.

N-gram tekniğinden faydalanarak ikili kelime gruplarının sıklık tablosu oluşturulduğunda, veri setinde birlikte en çok kullanılan ikili kelimeler elde edilmiştir.

```
kelime_grubu_olustur <- function (x)
  NGramTokenizer (x, Weka_control(min=2 , max=2))


library(RWeka)
ikili.kelime <- DocumentTermMatrix(
  temiz_esaretin_bedeli_corpus,
  control = list(tokenize= kelime_grubu_olustur))
ikili.kelime
ikili.kelime.m <- as.matrix(ikili.kelime)
frekans.ikili.kelime <- colSums(ikili.kelime.m)
ikili.kelime.grup <- names(frekans.ikili.kelime)
wordcloud(ikili.kelime.grup, frekans.ikili.kelime, max.words = 200, colors= c("red", "darkgoldenrod1", "tomato"))
```

Şekil 3.18 N-Gram modellerinden bigram uygulamasına ait kod



Şekil 3.19 Bigram uygulamasından elde edilen sıklıkların kelime bulutu

İkili kelime gruplarından oluşan bulutta duygu analizine yer verilmediği için n-gram tekniğinde yorumlanabilir sonuçlar elde edilememiştir. Twitter kullanıcılarının metinleri günlük konuşma içeriğine sahip olduğundan ve konu bağlantısı zorunluluğu bulunmadığından n-gram yöntemi analizde anlamlı sonuçlar elde etmek için yeterli kalmamıştır. Metin verilerini anlamsal bütünlükte incelemek için duygu analizi tekniğinden yararlanılmıştır.

```
## sentimental analysis

sentimenttext <- as.character(esaretin_bedeli_df$text)
get_nrc_sentiment('happy')
get_nrc_sentiment('abuse')

S1 <- get_nrc_sentiment(sentimenttext)
View(S1)
sentimenttext_last <- cbind(esaretin_bedeli_df$text, S1)

barplot(colSums(S1), col = rainbow(10), ylab= 'Count' , main= 'Esaretin Bedeli Twitter duygu sözcüğü')
```

Şekil 3.20 Duygu analizi için Esaretin Bedeli veri setinin içindeki text değişkenine ait metinlerin duygu puanlaması ve NRC kütüphanesinden faydalanarak duygu analizi kodu

Text içeriğindeki kelimelerin sentiment değerleri tablo halinde elde edilmiştir. Tablo

sütunlarında tweetlerin içeriğinde bulunan “öfke”, “beklenti”, “nefret”, “korku”, “neşe”, “üzüntü”, “şaşkınlık”, “güven”, “negatif” ve “pozitif” on farklı duygu incelemesinin frekans değerlerinden oluşmaktadır.

Untitled1* × S1 ×

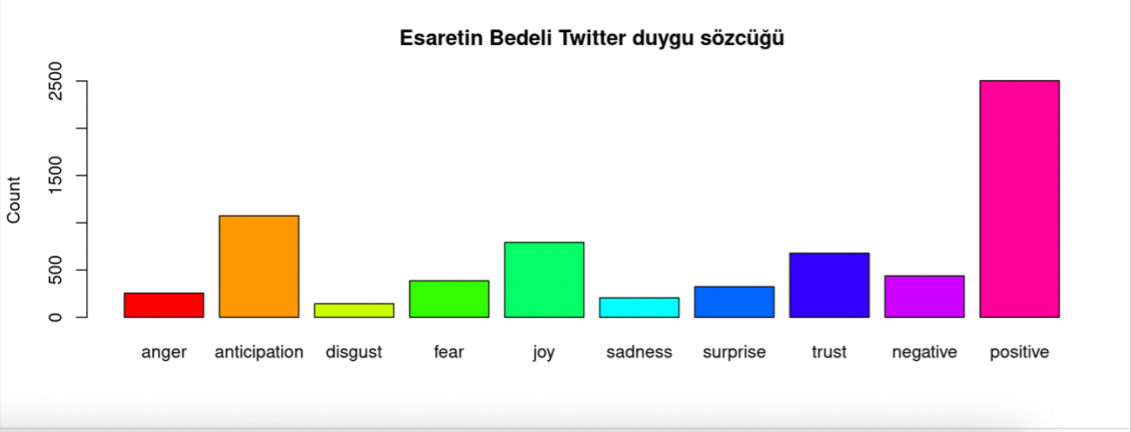
Filter

	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
1	1	1	0	1	1				1	3
2	0	2	0	0	2	0	2	2	0	3
3	0	1	0	0	0	0	0	0	0	0
4	1	1	0	1	1	0	0	2	1	3
5	0	2	0	1	2	1	2	2	1	3
6	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	1	0	2
10	0	0	0	0	0	0	0	0	0	1
11	0	1	0	0	1	0	0	0	0	1
12	0	0	0	0	0	0	0	0	0	1
13	0	2	0	0	2	0	2	2	0	2
14	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	0	0	0	0	0	1

column 8: numeric with range 0 - 4

Şekil 3.21 Duygu analizi ile elde edilen veri setinin tablo görseli

S1 tablosundaki frekansların çubuk grafiği analiz edildiğinde Şekil 3.21’de elde edilmiştir.



Şekil 3.22 Esaretin Bedeli veri setindeki metinlerin duygu durumlarına ait grafik görseli

Esaretin bedeli etiketi ile elde edilen iki bin adet tweetin büyük bir çoğunluğunun pozitif kelimelerden oluştuğu gözlemlenmektedir. Yine grafikte duygu incelemesi yapıldığında “beklenti” ve “neşe” duygularının çoğunluğu gözlemlenmiştir. Filme dair negatif kelimelerle yazılan metinlerin sayıca az olduğunu söyleyebileceğimiz gibi, “öfke” “nefret” ve “üzüntü” duygularına dair frekansın da az olduğu gözlemlenmektedir. Bu grafik incelendiğinde IMDb puanlamasında ilk sırada olan

Esaretin Bedeli filminin, günlük dilde yazılan kullanıcı yorumlarının da benzer yönde ve ezici çoğunlukla pozitif içerikli olduğu söylenebilir. Pozitif ve negatif kelimelerin ayrımı kelime bulutu aracı ile incelenebilmektedir.

```
esaretin_bedeli_df$text <- gsub("[^0-9A-Za-z///' ]", "", esaretin_bedeli_df$text)
esaretin_bedeli_df$text <- gsub("http\\w+", "", esaretin_bedeli_df$text)
esaretin_bedeli_df$text <- gsub("rt", "", esaretin_bedeli_df$text)
esaretin_bedeli_df$text <- gsub("@\\w+", "", esaretin_bedeli_df$text)
emo_esaretinbedeli_ <- sentiment(esaretin_bedeli_df$text)
View(emo_esaretinbedeli_)
esaretin_bedeli_df$sentiment <- emo_esaretinbedeli_$sentiment
View(esaretin_bedeli_df)

positive_tweets_esaret <- head(esaretin_bedeli_df[order(emo_esaretinbedeli_$sentiment, decreasing= T), c(1,17)],25)
positive_tweets_esaret


positive_tweets_esaret <- head(unique(esaretin_bedeli_df[order(emo_esaretinbedeli_$sentiment, decreasing= T), c(1,17
positive_tweets_esaret
write.table(positive_tweets_esaret$text, file = "positive.txt", sep= "\n")

negative_tweets_esaret <- head(unique(esaretin_bedeli_df[order(emo_esaretinbedeli_$sentiment, decreasing= F), c(1,17
negative_tweets_esaret
write.table(negative_tweets_esaret$text, file = "negative.txt", sep= "\n")

pos_neg_tweets <- c(positive_tweets_esaret$text, negative_tweets_esaret$text)
View(pos_neg_tweets)
tweets_corpus <- Corpus(DirSource(directory = "tweet"))
summary(tweets_corpus)
```

Şekil 3.23 Metinlerin pozitif ve negatif kelimeler biçiminde ayrılması kodu

Untitled1* | emo_esaretinbedeli_

Filter

	element_id	sentence_id	word_count	sentiment
1	1	1	4	0.12500000
2	2	1	18	0.11785113
3	3	1	7	0.09449112
4	4	1	5	0.11180340
5	5	1	21	0.10910895
6	6	1	21	-0.10910895
7	7	1	18	-0.29462783
8	8	1	8	0.08838835
9	9	1	18	0.11785113
10	10	1	15	0.24528895
11	11	1	23	0.33362306

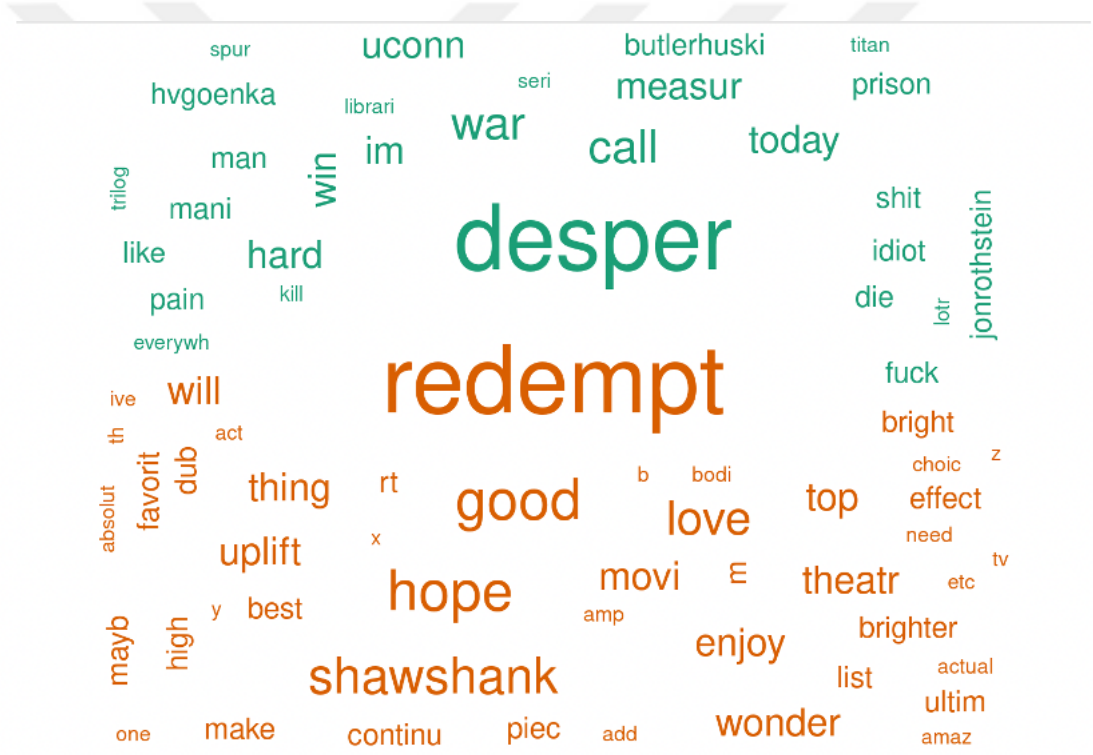
Showing 1 to 12 of 2,000 entries, 4 total columns

Şekil 3.24 Veri setindeki her bir kelime için duygu puanlaması

```
clean_tweets_corpus <- tm_map(tweets_corpus, tolower)
clean_tweets_corpus <- tm_map(clean_tweets_corpus, removePunctuation)
clean_tweets_corpus <- tm_map(clean_tweets_corpus, removeNumbers)
clean_tweets_corpus <- tm_map(clean_tweets_corpus, removeWords, stopwords())
clean_tweets_corpus <- tm_map(clean_tweets_corpus, stripWhitespace)
clean_tweets_corpus <- tm_map(clean_tweets_corpus, stemDocument)
str(clean_tweets_corpus)
ctc_tdm <- TermDocumentMatrix(clean_tweets_corpus, control = list(wordLengths=c(0,Inf)) )

ctc.matrix <- as.matrix(ctc_tdm)
head(ctc.matrix)
colnames(ctc.matrix) <- c( "Negative Tweets", "Positive Tweets")
comparison.cloud( ctc.matrix, max.words = 200 , random.order = F, )
```

Şekil 3.25 Veri temizliği işleminden sonra negatif ve pozitif kelimelerin aynı kelime bulutunda ayrılacak şekilde bulut oluşturma kodu



Şekil 3.26 Pozitif ve Negatif Kelimelerin Bulut Görseli

Pozitif ve negatif kelimelerden oluşan kelime bulutunda kırmızı kelimeler pozitif olanları yeşil olan kelimeler negatif olanları temsil etmektedir. Kelime bulutu daha da geliştirilmek istenirse ilgili duygu durumuna göre kelimelerin ayrımı yapılabilir. Buraya kadar yapılan analiz yalnızca Esaretin Bedeli filmine yöneliktir. Analizin bu kısmından sonra her üç film içinde yapılan kelime duygu durumuna yönelik kelime bulutlarının incelenmesi ile uygulama son bulacaktır. Her bir kelimenin duygu puanlaması için kodlarımız ve veri temizleme işlemimiz adım adım gösterilecektir.

```
f_clean_tweets <- function (tweets) {

  clean_tweets = sapply(tweets, function(x) x$getText())
  # remove retweet entities
  clean_tweets = gsub('(RT|via)((?:\\b\\W*@\\w+)+)', '', clean_tweets)
  # remove at people
  clean_tweets = gsub('@\\w+', '', clean_tweets)
  # remove punctuation
  clean_tweets = gsub('[[:punct:]]', '', clean_tweets)
  # remove numbers
  clean_tweets = gsub('[[:digit:]]', '', clean_tweets)
  # remove html links
  clean_tweets = gsub('http\\w+', '', clean_tweets)
  # remove unnecessary spaces
  clean_tweets = gsub('[ \t]{2,}', '', clean_tweets)
  clean_tweets = gsub('^\\s+|\\s+$', '', clean_tweets)
  # remove emojis or special characters
  clean_tweets = gsub('<.*>', '', enc2native(clean_tweets))

  clean_tweets = tolower(clean_tweets)

  clean_tweets
}
```

Şekil 3.27 Veri Temizleme İşlemi İçin Kullanılan Kodlar

"the godfather" etiketi ile ingilizce dilinde 01-01-2022 tarihinden itibaren yazılan 1200 adet tweet baba değişkenine, "the dark knight" etiketi ile ingilizce dilinde 01-01-2022 tarihinden itibaren yazılan 2000 adet tweet şövalye değişkenine atanmıştır.

```
62  baba <- searchTwitter('the godfather' , n=1200 , lang = 'en',since ='2022-01-01'  )
63  head(baba)
64
65  sovalye <- searchTwitter('the dark knight' , n=2000 , lang = 'en',since ='2022-01-01'  )
66  head(sovalye)
67
```

Şekil 3.28 Baba filmine ait 'the godfather' etiketi ile ingilizce dilinde yazılmış bin iki yüz adet tweetin ve kara şövalye filmine ait 'the dark night' etiketi ile ingilizce dilinde yazılmış iki bin adet tweetin elde edilmesine dair kod

Elde edilen tweetleri head komutuyla incelendiğinde, ilk altı tweet detayları aşağıda görseldeki gibidir. İlk tweet metnine detaylı bakıldığında tweete dair kullanıcı adı, retweet-tweet-mention ayrımı, metin ve link görülmektedir.

```
> baba <- searchTwitter('the godfather' , n=1200 , lang = 'en',since ='2022-01-01'  )
> head(baba)
[[1]]
[1] "Love4SSR6: RT @Rajshre73441992: @withoutthemind Public are outraged like never before as they
can relate to SSR having middle class roots &amp;amp; reached to…"

[[2]]
[1] "tekah2013: IFA Magazine team mourns the death of Paul Etheridge, the “godfather of financial p
lanning’   https://t.co/zYbJPRqlpN"

[[3]]
[1] "_M0rgaan_: So did they think they was gna change the Malcom x actor on godfather of Harlem and
nobody would notice. ??"

[[4]]
[1] "stephens_ben: @rickburin @CBuckthorn Second highest grossing film of its year, behind The Godf
ather and ahead of What's Up Doc? W… https://t.co/DTZpDNu4OH"

[[5]]
[1] "majorpsyche: Radiohead and the cure neutral milk hotel slaps I love American psycho and pulp f
iction and Trainspotting the godfa… https://t.co/PORpyPOLwi"

[[6]]
[1] "fakobthehonest: @thijsreuten EU should put #IRGCterrorists in terror list not only for Iranian
protesters, but for safety of every… https://t.co/8FJjAfCjOM"
```

Şekil 3.29 Baba filmine ait tweetlerin ilk altı örneği

```
> head(sovalye)
[[1]]
[1] "SpaceWizardPip: RT @Exalted_Speed: His reading of the Dark Knight Returns makes me want to see the entire movie dubbed like that. https://t.
o/cd7NZtWyZT"

[[2]]
[1] "Nianime_im: I felt Reigen really cared for Alina tho Ik his love is not so pure like how Mal does but I really really liked the… https://t.c
/LXOoErz61U"

[[3]]
[1] "sebasti30843494: @Superherotalk18 I like it better than the dark knight https://t.co/0KixiSIhI6"

[[4]]
[1] "DrKavarga: Watching The Dark Knight Rises on #HBOMax"

[[5]]
[1] "smokedblunt: i always think how he didnt get to see the masterpiece that is the dark knight and how much ppl loved him and i SOB every time"

[[6]]
[1] "KhalidElshhat: RT @BatmanShots_: The Dark Knight Rises (2012) https://t.co/EbN8DRwzFz"

>
```

Şekil 3.30 Kara Şövalye filmine ait tweetlerin ilk altı örneği

Her üç film için veri temizleme ve manipülasyonu yapılmıştır. Bir kullanıcı başka bir kullanıcının tweetini retweet yaptığında, aynı tweetin veride tekrar etmesine yol açmaktadır ve kelime sıklıklarını hesaplarken yanlış hesaplamaya sebep olmaktadır. Bu sorunu ortadan kaldırmak için Şekil 3.30'da “extract the timestamp” yorumu ile başlayan kodlarla retweet sebebiyle oluşan kopyaları verimizden çıkarıyoruz.

```
# First call the function defined above to clean the data
clean_tweets_esaret <- f_clean_tweets(esaretin_bedeli)
clean_tweets_baba <- f_clean_tweets(baba)
clean_tweets_sovalye <- f_clean_tweets(sovalye)



# extract the timestamp
timestamp_esaret <- as.POSIXct(sapply(esaretin_bedeli, function(x)x$getCreated()), origin="1970-01-01", tz="GMT")
timestamp_esaret <- timestamp_esaret[!duplicated(clean_tweets_esaret)]
clean_tweets_esaret <- clean_tweets_esaret[!duplicated(clean_tweets_esaret)]

timestamp_baba <- as.POSIXct(sapply(baba, function(x)x$getCreated()), origin="1970-01-01", tz="GMT")
timestamp_baba <- timestamp_baba[!duplicated(clean_tweets_baba)]
clean_tweets_baba <- clean_tweets_baba[!duplicated(clean_tweets_baba)]

timestamp_sovalye <- as.POSIXct(sapply(sovalye, function(x)x$getCreated()), origin="1970-01-01", tz="GMT")
timestamp_sovalye <- timestamp_sovalye[!duplicated(clean_tweets_sovalye)]
clean_tweets_sovalye <- clean_tweets_sovalye[!duplicated(clean_tweets_sovalye)]
```

Şekil 3.31 Her üç veri setine dair metin temizleme işlemi

Her üç film için farklı sözlüklerden duygu incelemesi yapılarak değişkenlere atanmıştır.

```
# Get sentiments using the four different lexicons
syuzhet_esaret <- get_sentiment(clean_tweets_esaret, method="syuzhet")
bing_esaret <- get_sentiment(clean_tweets_esaret, method="bing")
afinn_esaret <- get_sentiment(clean_tweets_esaret, method="afinn")
nrc_esaret <- get_sentiment(clean_tweets_esaret, method="nrc")
sentiments_esaret <- data.frame(syuzhet_esaret, bing_esaret, afinn_esaret, nrc_esaret, timestamp_esaret)

syuzhet_baba <- get_sentiment(clean_tweets_baba, method="syuzhet")
bing_baba <- get_sentiment(clean_tweets_baba, method="bing")
afinn_baba <- get_sentiment(clean_tweets_baba, method="afinn")
nrc_baba <- get_sentiment(clean_tweets_baba, method="nrc")
sentiments_baba <- data.frame(syuzhet_baba, bing_baba, afinn_baba, nrc_baba, timestamp_baba)

syuzhet_sovalye <- get_sentiment(clean_tweets_sovalye, method="syuzhet")
bing_sovalye <- get_sentiment(clean_tweets_sovalye, method="bing")
afinn_sovalye <- get_sentiment(clean_tweets_sovalye, method="afinn")
nrc_sovalye <- get_sentiment(clean_tweets_sovalye, method="nrc")
sentiments_sovalye <- data.frame(syuzhet_sovalye, bing_sovalye, afinn_sovalye, nrc_sovalye, timestamp_sovalye)
```

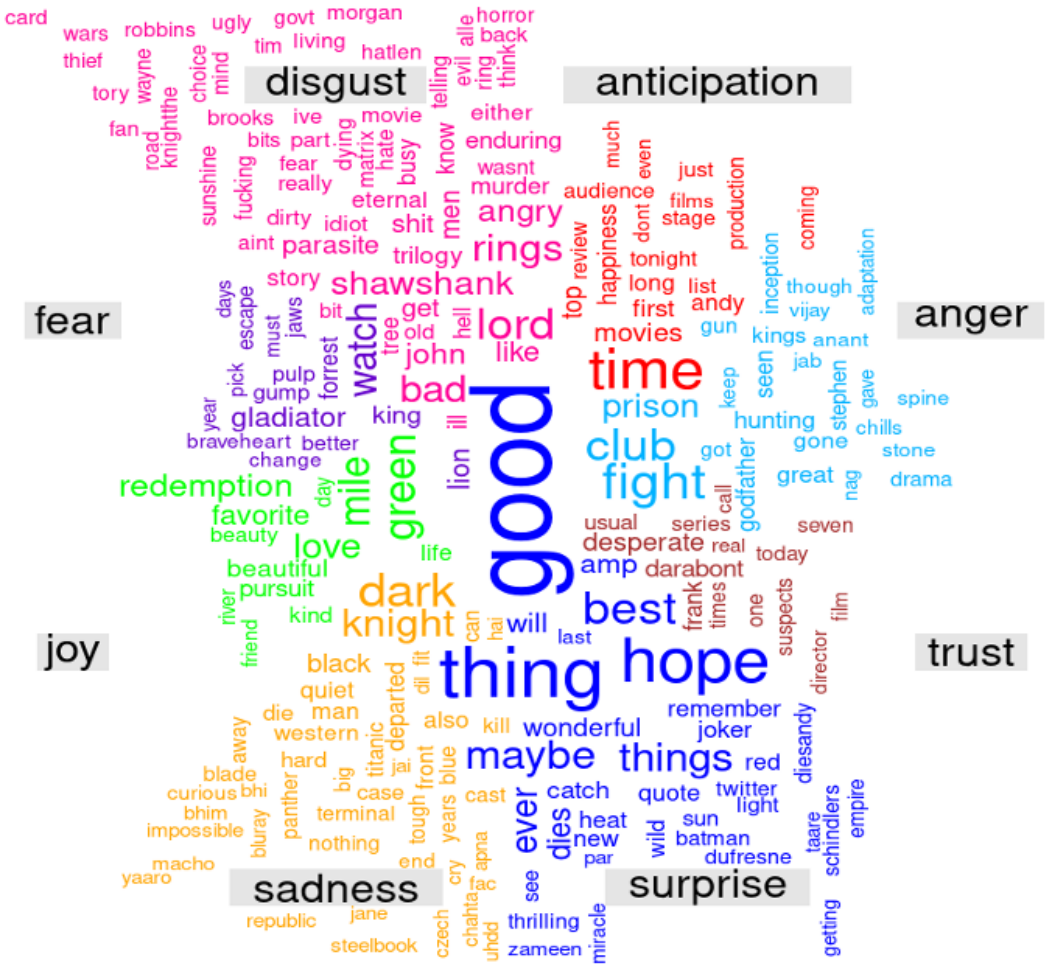
Şekil 3.32 Her üç filme ait BING, AFINN, NRC kütüphanelerinden faydalanarak duygu puanlaması kodları

NRC sözlüğünden faydalanılarak duygu çıkarımı yapılmıştır. Corpus ve döküman terim matrisleri elde edilerek ilgili sonuçların kelime bulutu görselleştirmesi yapılmıştır.

```
# get the emotions using the NRC dictionary
emotions_esaret <- get_nrc_sentiment(clean_tweets_esaret)
emo_bar_esaret = colSums(emotions_esaret)
emo_sum_esaret = data.frame(count=emo_bar_esaret, emotion=names(emo_bar_esaret))
emo_sum_esaret$emotion = factor(emo_sum_esaret$emotion, levels=emo_sum_esaret$emotion[order(emo_sum_esaret$count, decreasing = TRUE)

all_esaret = c(
  paste(clean_tweets_esaret[emotions_esaret$anger > 0], collapse=" "),
  paste(clean_tweets_esaret[emotions_esaret$anticipation > 0], collapse=" "),
  paste(clean_tweets_esaret[emotions_esaret$disgust > 0], collapse=" "),
  paste(clean_tweets_esaret[emotions_esaret$fear > 0], collapse=" "),
  paste(clean_tweets_esaret[emotions_esaret$joy > 0], collapse=" "),
  paste(clean_tweets_esaret[emotions_esaret$sadness > 0], collapse=" "),
  paste(clean_tweets_esaret[emotions_esaret$surprise > 0], collapse=" "),
  paste(clean_tweets_esaret[emotions_esaret$trust > 0], collapse=" ")
)
all_esaret <- removeWords(all_esaret, stopwords("english"))
# create corpus
corpus_esaret = Corpus(VectorSource(all_esaret))
#
# create term-document matrix
tdm_esaret = TermDocumentMatrix(corpus_esaret)
#
# convert as matrix
tdm_esaret = as.matrix(tdm_esaret)
tdm1_esaret <- tdm_esaret[nchar(rownames(tdm_esaret)) < 11,]
#
# add column names
colnames(tdm_esaret) = c('anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'trust')
colnames(tdm1_esaret) <- colnames(tdm_esaret)
comparison.cloud(tdm1_esaret, random.order=FALSE,
                 colors = c("#00B2FF", "red", "#FF0099", "#6600CC", "green", "orange", "blue", "brown"),
                 title.size=1, max.words=250, scale=c(2.5, 0.4),rot.per=0.4)
```

Şekil 3.33 Esaretin Bedeli filminin tweetlerine dair duygu analizi uygulaması kodları



Şekil 3.34 Esaretin Bedeli filmine dair tweetlerden duygu analizi ile elde edilen kelime bulutu

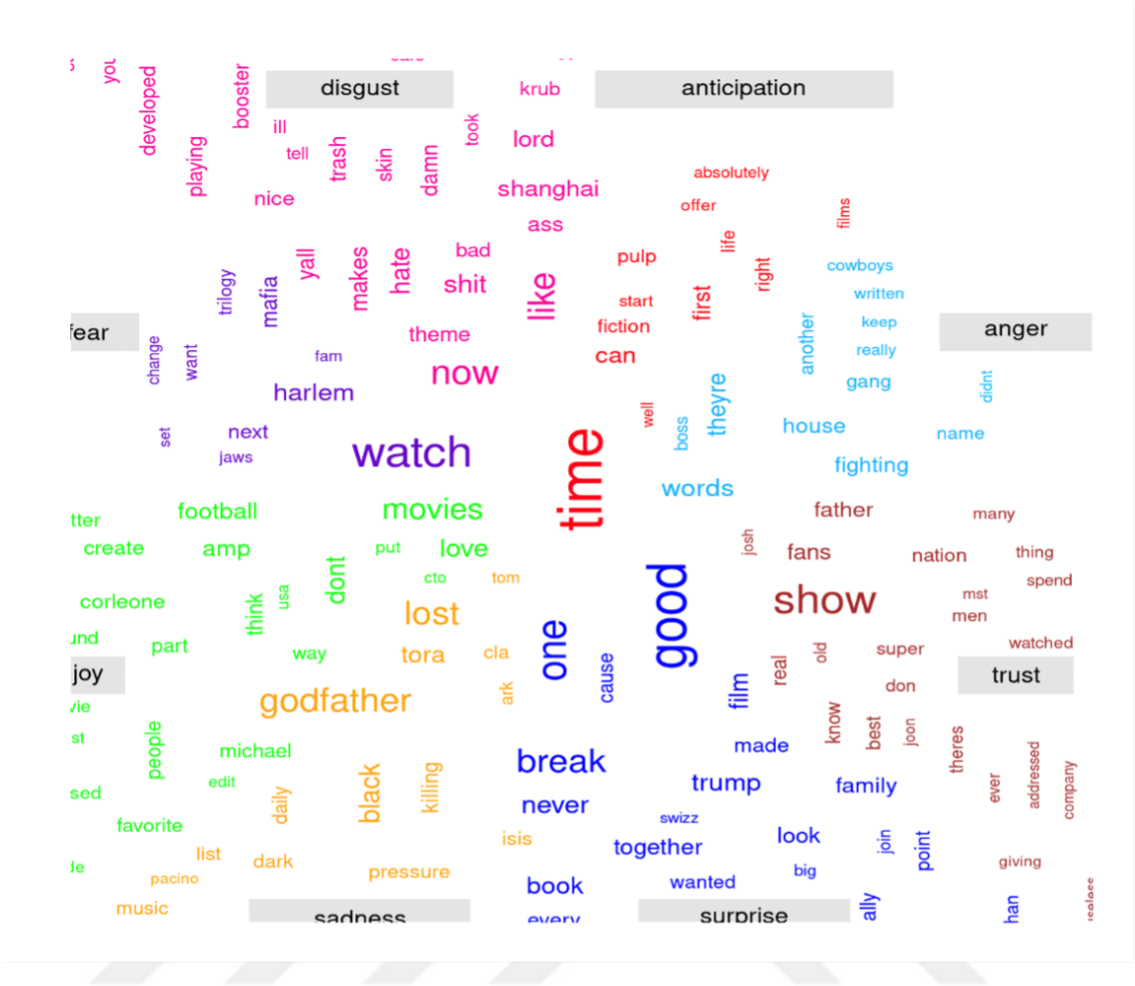
Esaretin Bedeli filminde Andy isimli başarılı bir bankacı cinayet ile yargılanması sonrası suçsuz olduğunu ispatlayamaz ve müebbet hapse mahkûm edilir. Hapishane duvarları arasında yaşamaya bağımlı hale gelmiş Red'in yaşamından esaretin benliğimizi, inançlarımızı ve insanlığımızı nasıl sarstığını anlatan filmde, esir edilmeye mahkûm olmayı reddeden ana karakterin, özgürlüğü için beslediği umudun nasıl bir kaçış macerasına dönüştüğü seyredilmektedir. Hapishanede yaşanan olumsuzlukları anlatan filme ait kelime bulutunda korku başlıklı kelimelerin frekansına bakıldığında dövüş, hapis, mücadele kelimelerini görmekteyiz. Kelimeler filmin ana konusu ile de benzerlik göstermektedir. Beklenti duygularını barından kelimeler incelendiğinde frekansa göre ağırlıklandırılan en önemli kelimenin zaman olduğu görülmektedir. Filmde bir replikte "Unutma Red umut iyi bir şeydir. Belki de en iyisi ve iyi şeyler asla ölmez." cümlesi geçmektedir. Bu bilgi ışığında şaşkınlık duygusu kelimeleri incelendiğinde frekans değerlerine göre en çok kullanılan kelimenin "good" ve

“thing” olduğu görülmektedir ve iyi şeyler anlamına gelen bu kelimelerin bahsi geçen repliğe atıfta bulunulması amaçlı tweetler yazıldığını yorumlayabiliriz. Yine bir başka replikte “Korktukça tutsak, umut ettikçe özgürsün.” cümlesi ışığında şaşkınlık duygu kelimelerinde en çok rastladığımız “hope” umut kelimesi anlamsal bütünlük oluşturmaktadır. Eğlence ile ilgili kelimelere bakıldığında “love” aşk kelimesi; üzüntü duygusu barındıran kelimeler incelendiğinde gece anlamına gelen “night” ve karanlık anlamına gelen “dark” kelimesinin frekansıyla orantılı olarak büyük gösterildiğini görülmektedir. Öne çıkan kelimelerin duygu beklentisiyle uyumlu olduğu, filmin konusu ile ilişkili olduğu söylenebilmektedir

```
emotions_baba <- get_nrc_sentiment(clean_tweets_baba)
emo_bar_baba = colSums(emotions_baba)
emo_sum_baba = data.frame(count=emo_bar_baba, emotion=names(emo_bar_baba))
emo_sum_baba$emotion = factor(emo_sum_baba$emotion, levels=emo_sum_baba$emotion[order(emo_sum_baba$count, decreasing = TRUE)])

all_baba = c(
  paste(clean_tweets_baba[emotions_baba$anger > 0], collapse=" "),
  paste(clean_tweets_baba[emotions_baba$anticipation > 0], collapse=" "),
  paste(clean_tweets_baba[emotions_baba$disgust > 0], collapse=" "),
  paste(clean_tweets_baba[emotions_baba$fear > 0], collapse=" "),
  paste(clean_tweets_baba[emotions_baba$joy > 0], collapse=" "),
  paste(clean_tweets_baba[emotions_baba$sadness > 0], collapse=" "),
  paste(clean_tweets_baba[emotions_baba$surprise > 0], collapse=" "),
  paste(clean_tweets_baba[emotions_baba$trust > 0], collapse=" ")
)
all_baba <- removeWords(all_baba, stopwords("english"))
# create corpus
corpus_baba = Corpus(VectorSource(all_baba))
#
# create term-document matrix
tdm_baba = TermDocumentMatrix(corpus_baba)
#
# convert as matrix
tdm_baba = as.matrix(tdm_baba)
tdm1_baba <- tdm_baba[nchar(rownames(tdm_baba)) < 11,]
#
# add column names
colnames(tdm_baba) = c('anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'trust')
colnames(tdm1_baba) <- colnames(tdm_baba)
comparison.cloud(tdm1_baba, random.order=FALSE,
                 colors = c("#00B2FF", "red", "#FF0099", "#6600CC", "green", "orange", "blue", "brown"),
                 title.size=1, max.words=250, scale=c(2.5, 0.4),rot.per=0.4)
```

Şekil 3.35 Baba filminin tweetlerine dair duygu analizi uygulaması kodları



Şekil 3.36 Baba filmine dair tweetlerden duygu analizi ile elde edilen kelime bulutu

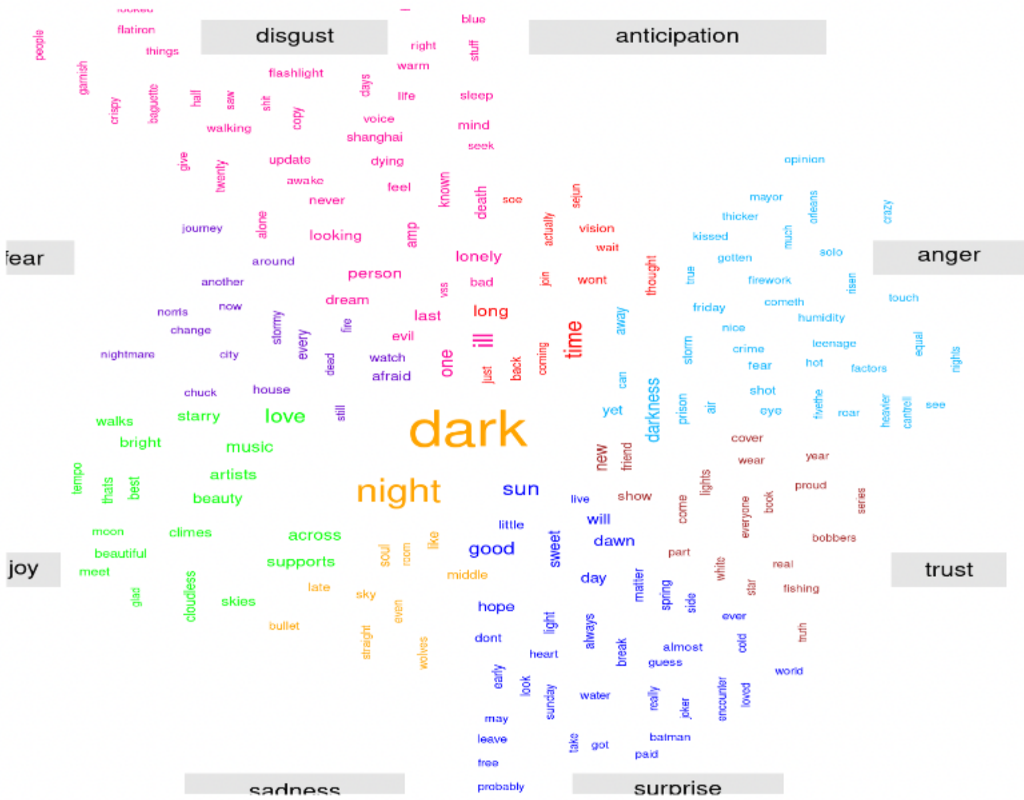
Baba filmine ait kelime bulutu analizinden önce konu incelendiğinde, filmde New York'ta yaşayan güçlü bir İtalyan mafya ailesinin hikayesi anlatılmaktadır. Baba Michael Corleone artık 60 yaşını geçmiştir ve babalığı devredecek bir varis aramaktadır. Bu sırada yıllardır başarmak istediği şey, aileyi bir suç örgütü konumundan çıkartıp yasal iş yapan bir aile haline getirme projesini yeniden ele almaktadır. Kelime bulutunda beklenti kelimeleri incelendiğinde Esaretin bedeli filminde olduğu gibi "time" kelimesi öne çıkmaktadır. Filmin içerisindeki bir replikte "Ailesi ile vakit geçirmeyen bir adam gerçek bir adam değildir." Cümlesi geçmektedir ve aile ve zaman konularına dikkat çekmektedir. Beklenti duygu kelimelerinde zaman ve şaşkınlık kelimelerinde aile ve birliktelik yer almaktadır. Korku duygu grubu incelendiğinde mafya kelimesi göze çarpmaktadır ve filmin konusu ile benzerlik göstermektedir. Şaşkınlık kelimelerinde "family" aile kelimesi konu bağlantısı sağlamaktadır. Üzüntü duygu grubuna bakıldığında "godfather" baba kelimesi görülmektedir ve dram türünde olan filmin isminin sadness duygusu başlığında ilişkilendirildiği de yorumlanabilmektedir. Güven duygusu başlığındaki kelimelerde "show" kelimesine rastlanılmaktadır ve baba filminde güven gösterimi

“Ben değerimi bilmeyen kişilere dostluk gösterecek bir insan değilim.” Cümlesinde de geçtiği gibi dostluğun, aile bağının ve yakın ilişkilerin görselleştirilmesi ile “Show” kelimesinin eşleştiği yorumlanabilir

```
emotions_sovalye <- get_nrc_sentiment(clean_tweets_sovalye)
emo_bar_sovalye = colSums(emotions_sovalye)
emo_sum_sovalye = data.frame(count=emo_bar_sovalye, emotion=names(emo_bar_sovalye))
emo_sum_sovalye$emotion = factor(emo_sum_sovalye$emotion, levels=emo_sum_sovalye$emotion[order(emo_sum_sovalye$count, decre

all_sovalye = c(
  paste(clean_tweets_sovalye[emotions_sovalye$anger > 0], collapse=" "),
  paste(clean_tweets_sovalye[emotions_sovalye$anticipation > 0], collapse=" "),
  paste(clean_tweets_sovalye[emotions_sovalye$disgust > 0], collapse=" "),
  paste(clean_tweets_sovalye[emotions_sovalye$fear > 0], collapse=" "),
  paste(clean_tweets_sovalye[emotions_sovalye$joy > 0], collapse=" "),
  paste(clean_tweets_sovalye[emotions_sovalye$sadness > 0], collapse=" "),
  paste(clean_tweets_sovalye[emotions_sovalye$surprise > 0], collapse=" "),
  paste(clean_tweets_sovalye[emotions_sovalye$trust > 0], collapse=" ")
)
all_sovalye <- removeWords(all_sovalye, stopwords("english"))
# create corpus
corpus_sovalye = Corpus(VectorSource(all_sovalye))
#
# create term-document matrix
tdm_sovalye = TermDocumentMatrix(corpus_sovalye)
#
# convert as matrix
tdm_sovalye = as.matrix(tdm_sovalye)
tdm1_sovalye <- tdm_sovalye[nchar(rownames(tdm_sovalye)) < 11,]
#
# add column names
colnames(tdm_sovalye) = c('anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'trust')
colnames(tdm1_sovalye) <- colnames(tdm_sovalye)
comparison.cloud(tdm1_sovalye, random.order=FALSE,
                 colors = c("#00B2FF", "red", "#FF0099", "#6600CC", "green", "orange", "blue", "brown"),
                 title.size=1, max.words=250, scale=c(2.5, 0.4),rot.per=0.4)
```

Şekil 3.37 Kara Şövalye filminin tweetlerine dair duygu analizi uygulaması kodları



Şekil 3.38 Kara Şövalye filmine dair tweetlerden duygu analizi ile elde edilen kelime bulutu

Kara Şövalye, halk için büyük bir tehdit oluşturan Joker'in ortaya çıkması ile kaosa dönen Gotham Sokakları'nın yeniden kurtarıcılığını üstlenen Batman'in hikayesini konu ediyor. Suç işleyenlerden arındırılan bir yer, bir zaman sonra yeniden tehdit altında kalabilir ve işte o zaman yeniden kolları sıvayacak olanların mücadelesi de daha keskin olarak hayata geçecektir. Batman, Teğmen Gordon ve Savcı Harvey Dent bir araya gelerek Gotham Sokakları'nda bu işi başarmış olsalar da ansızın ortaya çıkan Joker, işleri fena halde bozar. Onun dehası ile baş etmek kolay olmayacaktır. Gotham eski karmaşa dolu günlerin eşiğindedir. Batman yeniden kurtarıcılığa soyunurken kendi varlığının bulduğu anlamı da sorgulamaya başlar. O aslında suçluların sayısını azaltıyor mudur yoksa çoğaltıyor mudur bunu gerçekten anlamak isteyecektir. Filmin ismi ile çağırdığımız tweet metinlerinden oluşan verilerin kelime bulutu incelendiğinde üzüntü kelime grubunda "dark" kelimesinin frekansa bağlı olarak en çok karşılaşıldığı gözlemlenir. Şaşkınlık duygu grubundaki kelimelerde iyi, umut, güneş gibi olumlu kelimelere rastlanılmıştır. Öfke kelime grubunda karanlık, suç, korku gibi kelimelere rastlanılmıştır. Beklenti kelime grubunda zaman, düşünce, geri dönmek gibi kelimelere rastlanılmıştır. Bazı insanlar para gibi mantıklı şeylerin peşinde değildir. "Satın almak, korkutmak, anlaşmak ya da pazarlık etmek mümkün değildir. Bazı insanlar sadece dünyanın yandığını seyretmek ister." repliği ışığında fear(korku) duygusundaki kelimeler incelendiğine "fire" anlamına gelen yanma kelimesinin, "dead" ölüm kelimesinin, "afraid" korku kelimesinin konu bağlantısı ile ilişkili olarak karşımıza çıktığı söylenebilmektedir.

4
SONUÇ VE ÖNERİLER

Çalışma kapsamında veri madenciliği ve makine öğrenmesi konularının bir bölümü olan metin madenciliği yaklaşımlarına değinilmiştir. Metin madenciliği içinde analize başlamadan önce metin ön temizleme safhalarından bahsedilmiş ve uygulama kısmına R programında yazılan kodları da bulunan metin ön işleme aşamasına yer verilmiştir. Ön işlemenin çalışmaya etkisi metin içerisindeki kelimelerin bilgisayar dilinde okunabilir hale getirmektir. Büyük küçük harflerin dengelenmesi, ifadelerin temizlenmesi, analizde sıklık sayılarını artırmaması amacıyla kelime anlamı bulunmayan bağlaç, sıfat gibi durak kelimelerinin çıkarılması gibi aşamalara yer verilmiştir. Temizlenen verinin meta data formatına dönüştürülmesine hem teorik hem de uygulamada detaylıca yer verilmiştir. Meta data formatındaki temizlenen veriler metin madenciliği teknikleri ile işlenebilir ve analiz edilebilir hale getirilmiştir. Böylece verilerin tf-idf, n-gram incelemeleri yapılmıştır. Duygu analizi yaklaşımlarından yararlanarak verilerin pozitif kelimeler ve negatif kelimeler ayrımı yapılmıştır. Verilerin yine duygu analizi yaklaşımlarından yola çıkarak duygu puanlamaları ile korku, neşe, kızgınlık, beklenti, sürpriz gibi kategorilere ayrılarak bu kategorilerde en çok hangi kelimeler paylaşıldığı incelenmiştir. Çalışmamız bu yönde okunulamayacak kadar büyük metin kaynaklarının ana fikirlerine kolayca ulaşmamızı hedefleyecek çıktılar elde etmemizi sağlamıştır. Tüm bu analizlerde bar-plot grafiği, kelime bulutları gibi aracılarla görselleştirme sağlanmıştır. Çalışmamızda IMDb puanlamasına göre en iyi ilk üç filmin Twitter kullanıcı yorumlarının analizi konu edinilmiştir. Metin verilerinin kelime bulutu çıktılarının detaylı yorumlarına yer verilmiştir. Kelime bulutlarında çoğunlukla filme ait repliklere atıfta bulunulan tweetlere rastlanıldığı tahmin edilmiştir. Kelime bulutlarındaki kelimelerin duygu ayrımının içinde bulunduğu kelimelerin anlamları ile başarılı bir şekilde bütünlük oluşturduğu gözlemlenmiştir. Kelime bulutlarında frekans bakımından ağırlıklanmış kelimelerin boyut olarak büyük göründüğü gözlemlenmiş ve en çok rastlanan kelimelerin ilgili filmin konusu ile ilişkili olduğu yorumlanmıştır. Kelime bulutlarının inceleyenler için genel olarak filmin özetine dair fikir edinilmesine yardımcı olacağı gözlemlenmiştir. Bar-plot grafiğinde en çok puan alan Esaretin Bedeli filminin

Twitter kullanıcı yorumlarında da çoğunlukla pozitif kelimeler içerdiği gözlemlenmiş ve bu yönü ile IMDb puanlaması ile Twitter yorumlarının her ikisinin de gerçek kişilerce kullanıldığı göz önünde bulundurulursa analizlerinin de pozitif negatif ayrım bakımından benzer yönde olduğu yorumlanmıştır. Metin madenciliğinin büyük veri kapsamındaki veri setleri için birkaç adımda özetlenebilir ana fikre ulaşılması konusunda kolaylık sağlayabileceğine çalışmamız neticesinde ulaşılabileceği de yorumlanabilmektedir.Çalışma kapsamında kullanıcıları gerçek kişilerden oluşan iki ayrı platform olan IMDb ve Twitter üzerinden paylaşılan beğenilerin benzerlikleri duygu analizi teknikleri ile test edilmiştir. IMDb puanlamasında beğeni bakımından yüksek puan ile derecelendirilen üç film için Twitter platformundaki kullanıcı yorumları ele alınmıştır. İlgili filmler hakkında paylaşılan tweet metinlerindeki duygu puanlamasında da pozitif duygu kelimelerinin ağırlıklı olduğu tespit edilmiştir.

KAYNAKÇA

[1] Z. Çeliksu, "Yabancı dizilerin altyazı ve twitter yorumlarının metin madenciliği ile incelenmesi," M.S. thesis, Mimar Sinan Güzel Sanatlar Üniversitesi, İstanbul, Türkiye, 2017.

[2] Y. M. Kizilkaya, "Duygu analizi ve sosyal medya alanında uygulama," Ph.D. dissertation, Bursa Uludağ Üniversitesi, Sosyal Bilimler Enstitüsü, Bursa, Türkiye, 2018.

[3] M. Meral, "Twitter verilerini anlamsal sınıflandırma," M.S. thesis, Yıldız Teknık Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, Türkiye, 2014.

[4] E. S. Akgül, C. Ertano, B. Diri, "Twitter verileri ile duygu analizi.," *Pamukkale University Journal of Engineering Sciences*, vol. 22, no. 2, 2016.

[5] S. LORT TOSUN, "Transfer öğrenme tabanlı aktİf öğrenme metodu ile duygu analİzİ," M.S. thesis, Karabük Üniversitesi, Lisansüstü Eğitim Enstitüsü, Karabük, Türkiye, 2021.

[6] Y. Liu, X. Huang, A. An, X. Yu, "Arsa: A sentiment-aware model for predicting sales performance using blogs," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 607–614. DOI: 10.1145/1277741.1277845.

[7] M. Pennacchiotti, A.-M. Popescu, "A machine learning approach to twitter user classification," in *Proceedings of the international AAAI conference on web and social media*, vol. 5, 2011, pp. 281–288. DOI: 10.1609/icwsm.v5i1.14139.

[8] A. Tumasjan, T. Sprenger, P. Sandner, I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the international AAAI conference on web and social media*, vol. 4, 2010, pp. 178–185. DOI: 10.1609/icwsm.v4i1.14009.

[9] D. Watts, K. George, T. A. Kumar, Z. Arora, "Tweet sentiment as proxy for political campaign momentum," in *2016 IEEE International Conference on Big Data (Big Data)*, IEEE Access, 2016, pp. 2475–2484. DOI: 10.1109/BigData.2016.7840885.

[10] J. Bollen, H. Mao, X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011. DOI: 10.1016/j.jocs.2010.12.007.

[11] M. U. Şimşek, S. Özdemir, "Analysis of the relation between turkish twitter messages and stock market index," in *2012 6th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE Access, Tiflis,Gürcistan, 2012, pp. 1–4. DOI: 10.1109/ICAICT.2012.6398520.

[12] M. Birjali, A. Beni-Hssane, M. Erritali, "Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks," *Procedia Computer Science*, vol. 113, pp. 65–72, 2017. DOI: 10.1016/j.procs.2017.08.290.

[13] A. Baj-Rogowska, "Sentiment analysis of facebook posts: The uber case," in *2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)*, IEEE, Kahire, Mısır, 2017, pp. 391–395. DOI: 10.1109/INTELCIS.2017.8260068.

[14] *Imdb top 250 movies*, IMDb. [Online]. Available: https://www.imdb.com/chart/top/?ref_=nv_mv_250 (visited on 01/15/2023).

[15] N. ve Genart Medya, *Türkiye Twitter Kullanıcı İstatistikleri – 2014*, Dijital Ajanslar. [Online]. Available: https://www.dijitalajanslar.com/turkiye-twitter-kullanici-istatistikleri-2014/.

[16] S. Atan, "Metin madenciliği ile sentiment analizi ve borsa istanbul uygulaması," Ph.D. dissertation, Ankara Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara, Türkiye, 2016.

[17] M. Özdeş, "Büyük veri araçlarını kullanarak duygu analizi gerçekleştirimi," M.S. thesis, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü, Denizli, Türkiye, 2017.

[18] J. Silge, D. Robinson, *Text mining with R: A tidy approach*. California, USA: O'Reilly Basın, 2017.

[19] D. R. Michael K. Brown Andreas Kellner, *Stochastic Language Models (n-gram) Specification*, Stochastic Language Models (N-Gram) Specification. [Online]. Available: https://www.w3.org/TR/ngram-spec/ (visited on 01/15/2023).

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirisi

1. N. Kapukaya, E. Tuna, "İdam mahkumlarının son sözlerinin metin madenciliği algoritmaları kullanılarak R dili ile duygu analizi", Cukurova 8th International Scientific Researches Conference 15 - 17 April 2022 / Adana, TURKEY, pp.892-905