Marmara Fen Bilimleri Dergisi 2017, 3: 94-101
DOI: 10.7240/marufbd.330585

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

R ile Sosyal Ağ Madenciliği

Social Network Mining with R

Ufuk SARIKAYA[1] (ORCID ID: 0000-0001-6951-6388), **Buket DOĞAN**[1] (ORCID ID: 0000-0003-1062-2439)
Abdulsamet AKTAŞ[1] (ORCID ID: 0000-0003-0746-7693)

[1] *Marmara Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, 34722, Kadıköy / İSTANBUL*

Öz

Birbirimizle ve çevremizle iletişim şeklimizi değiştiren ve her geçen gün kullanıcı sayısı artan sosyal ağlar, yapılan paylaşım ve aktiviteler ile çok fazla verinin ortaya çıkmasını sağlayan ortamlardır. Sosyal ağ madenciliği ile bu ortamda farklı biçimlerde ve büyük miktardaki verinin, veri madenciliği yöntemleri ile analiz edilmesi ve anlamlandırılması mümkün olmaktadır. Bu çalışmada sosyal ağ madenciliği uygulaması kapsamında Twitter ve WhatsApp verileri üzerinde R programlama ortamında yapılan analiz süreci ve sonuçları açıklanmaktadır.

Anahtar Kelimeler: Sosyal Ağlar, Sosyal Ağ Madenciliği, Twitter, WhatsApp, R

Abstract

Social networks that change the way we communicate with each other and around the world, increasing in number of users every day, are the platforms that allow a lot of data to be generated with sharing and activities. Through social network mining, it is possible to analyze and interpret data in different format and large amount of data by means of data mining methods. In this study, the process and results of the analysis made on the Twitter and WhatsApp data in the R programming platform are explained in the scope of social network mining application.

Keywords: Social Network, Social Network Analysis, Twitter, WhatsApp, R

I. GİRİŞ

İnternet, milyonlarca bilgisayar sisteminin birbirine bağlı olduğu, tüm dünya tarafından yaygın olarak kullanılan ve her geçen gün kullanımı artan bir iletişim ağıdır. Sosyal ağlar da bu devasa iletişim ağının en önemli bileşenlerinden biridir. Sosyal ağlar, bireylerin birbirleriyle gizlilik çerçevesinde özgürce ve kolaylıkla iletişim kurabildiği, anlık ve çift taraflı veri paylaşımına olanak sağlayan, içerisinde milyarlarca veri barındıran iletişim aracıdır. Sosyal ağlar, bireylerin internet üzerinden birbirleriyle etkileşim halinde olmasını sağlayan ağlara verilen genel addır. Sosyal ağlar sayesinde bireyler birbirlerine metin, fotoğraf, video ve mesaj gönderebilir, birbirlerinin gönderilerine yorum ve beğeni yaparak kendi aralarında etkileşimde bulunabilirler. En yaygın kullanılan sosyal ağlardan bazıları; Twitter, Instagram, WhatsApp Messenger ve Facebook'tur. Günümüzde Twitter'da yaklaşık 256 milyon aktif kullanıcı bulunmakta ve her dakika ortalama 98 bin tweet atılmaktadır[1].Twitter'da iletiler 140 karakterle sınırlıdır. Twitter bu iletilere tivitler (tweets) adını vermiştir. WhatsApp Messenger, akıllı telefonlar için geliştirilen, platformlar arası çalışma özelliğine sahip bir mesajlaşma ve arama uygulamasıdır ve günümüzde yaklaşık 1.3 milyar kullanıcısı vardır[2].

Günümüzde bu kadar yaygın kullanılan sosyal ağların içerisinde bulunan metin, görsel ve video gibi farklı biçimlerdeki büyük veri üzerinde veri madenciliği ile analizler yaparak, bu ortamdaki yapısal kalıpları ortaya çıkarmak, ilişkileri ve dinamik durumu modellemek mümkün olmaktadır [3].

R programlama dili Yeni Zelanda Auckland Üniversitesi'nden Ross Ihaka ve Robert Gentleman tarafından geliştirilmiştir. Açık kaynak kodludur ve kullanıcıların eklediği özel fonksiyonlar veya çok özel araştırma alanlarına ait paketlerle kolayca geliştirilebilirdir. Grafik kullanıcı arayüzü için ise yaygın olarak R Studio programı kullanılmaktadır [4]. R ortamı, esnek bir istatistiksel analiz aracı olarak farklı formatlardaki verilerin aktarılmasını ve veri üzerinde istenilen

Sorumlu yazar/Corresponding Author: Buket Doğan, Tel: 0216 3365770/1228, e-posta: buketb@marmara.edu.tr
Gönderilme/Submitted: 24.07.2017, **Kabul/Accepted:** 12.10.2017

değişikliklerin yapılabilmesini ve uygun istatistiksel yöntemlerin kullanılabilmesini sağlamaktadır. Ayrıca bu ortam klasik istatistik testleri, zaman serileri analizi, sınıflandırma, kümeleme gibi analizleri büyük veriler üzerinde gerçekleştirme imkânı sunmaktadır. R programlama, diğer birçok istatistiki hesaplama dilinden daha kuvvetli bir nesneye yönelik programlama kabiliyetine de sahiptir [5].

1.2. İlgili Çalışmalar

Sosyal ağlarda kullanıcı sayısındaki artış beraberinde işlenmemiş büyük bir veriyi de getirmiştir. Bu verilerin işlenmesi ve daha kullanışlı bir hale getirilmesi için sosyal ağ analizi yapılması gerekmektedir. R programlama ortamı bu tip veri madenciliği analizlerinin yapılması için imkânlar sağlamaktadır. Bu başlık altında sosyal ağ analizi ve R programlama ile ilgili literatürdeki önemli çalışmalara değinilmektedir.

Alrubaian M. ve arkadaşları [6] twitter sosyal ağ kullanıcılarının doğru ve güvenli bir bilgi verip vermediğini yeni bir metrik sistemi geliştirerek analiz etmişlerdir. Kullanıcıların, verilen bir konu ile ilgili sosyal medyada belirli bir yetkinliği olup olmadığı araştırılmıştır. Daha sonra kullanıcının konu hakkında ne düşündüğünü anlayabilmek için duygu analizi yapılmıştır. Bu iki analiz sistemi birleştirilerek, var olan analiz modellerine birkaç yeni özellik daha ekleyip yeni bir metrik sistemi yapılmıştır. Veri toplama işleminde Twitter'ın akış uygulması(streaming API) ve arama uygulaması(search API) kullanılmıştır. Arama uygulaması Twitter'daki bir kullanıcının en son attığı 3200 tweeti getirmiştir. Veri seti oluşturmak ve istatistiksel bir çalışma yapmak için alınan tivitler bir tabloya alınmıştır. Elde olan veriler metrik sisteme göre değerlendirilmiş ve bir kullanıcı profili oluşturulmuştur. Daha sonra kullanıcı ve verdiği bilgi birlikte değerlendirilip güvenli olup olmadığına karar verilmiştir.

Zhang, K.P ve arkadaşları [7] sosyal medya platformu üzerindeki aktivitelerine göre kullanıcılara çevrimiçi marka reklamı yapmak için bir kullanıcı seçim sistemi yapmayı amaçlamışlardır. Analiz yapmak için sosyal ağların en yaygın olanlarından yani Facebook'tan yaklaşık 2.1 terabyte veri toplanmıştır. Facebook'dan verileri toplamak için Facebook Graph uygulaması kullanılmıştır. Toplanan veriler ile iki farklı ağ oluşturulmuş ve bu iki ağ üzerinde çalışılmıştır. Veri boyutu çok büyük olduğundan Zhang, K.P ve arkadaşları[7] veri seti üzerinde çalışabilmek için dağıtık programlama mantığını kullanan algoritmalar geliştirmişlerdir. Oluşturulan ağlar sayesinde markalar ve kullanıcılar arasındaki bağlantıyı analiz etmişlerdir. Ağ özelliklerini ve kullanıcı topluluklarını inceledikten sonra, hedef kullanıcı kitlesini bulabilmek için, yapılan analize göre ağlara bir etiket veren sistem tasarlamışlardır. Bu sistemin bir parçası olarak belirli bir marka ile ilgili olan diğer markaları bulabilmek için hiyerarşik kullanıcı algılama algoritması geliştirmişlerdir.Analiz sonucu oluşturulan setteki markaları "*odaklanmış marka (focal brand)*" olarak anlamlandırmışlardır. Facebook verileri ile deneyler yapılmış ve tanımladıkları odaklanmış marka'nın hedef kullanıcı kitlesini belirlemede önemli bir performans gelişimi sağladığını göstermişlerdir.

Iglesias, J. A. ve arkadaşları [8] kullanıcının seçtiği herhangi bir yerleşim birimi için, Twitter kullanıcı profilini otomatik olarak analiz etmeyi gerçekleştiren bir program yapmayı amaçlamışlardır. Gerçekleştirilen program uçdeğerlerin teşhisi, profillerin kümelenmesi ve sınıflandırılması gibi işlemleri de yapabilmektedir. Iglesias, J. A. ve arkadaşları verileri toplayabilmek için Twitter API uygulamasını kullanmışlardır. Toplanan veriler belirli bir ön eleme işleminden geçirilmiştir. Kalan veriler ile kullanıcı profilleri oluşturularak uçdeğer çıkarımı, kümeleme ve sınıflama işlemleri yapılmıştır. Analiz işlemi bittikten sonra kullanıcıya istediği bilgiler ayrı ayrı pencerelerde gösterilmiştir.

Sosyal ağlar, yerel ağızlardan farklı olan yeni iletişim yollarının ortaya çıkmasına sebep olmuştur. Bunun en önemli nedenlerinden birisi zamanın ve alanın kısıtlı olmasıdır. Bu yeni iletişim yollarından birisi de Twitter'da belirli bir konu üzerine yazılan tivitleri etiketlemek için kullanılan, "#" sembolü ile işaretlenen konu veya kelimelerdir. Bu kelime veya konular genel olarak "*hashtag"* olarak adlandırılır. Abascal-Mena, R ve arkadaşları[9] ortaya çıkan bu yeni iletişim yolunu kullanarak sosyal-semantik toplulukları tespit etmeyi amaçlamışlardır. Çalışmalarında veri seti olarak Twitter'dan aldıkları tivitleri kullanmışlardır. Verileri toplayabilmek için R programlama dilinde kendilerinin geliştirdikleri bir bilgisayar programı kullanmışlardır. Oluşturulan ağlar üzerinde gerekli çalışmalar yapıldıktan sonra tespit edilen toplulukları göstermek amacıyla açık kaynak kodlu yazılım olan "*GEPHI*" kullanılmıştır.

Kılınç ve arkadaşları akademik yayınlar üzerinde metin madenciliği yöntemlerini kullanarak, akademik makalelerin sınıflara tasnif etme başarısını ölçmeyi amaçlamışlardır. Veri setini oluşturabilmek için geliştirmiş oldukları bir yazılım aracı yardımıyla, akademisyenlerin kendi yapmış oldukları yayınları paylaştıkları web sitesi olan www.researchgate.net adresinden 50 farklı dergiden 2000 adet makaleye ait özet bilgilerini toplamışlardır. Bu veri setleri üzerinde çalışabilmek için R programlama dili ve R studio programı kullanılmıştır. R dili ile R studio ortamında veri seti, K-En Yakın Komşu (KNN) algoritması kullanılarak sınıflandırılmıştır. Çalışma sonucunda %96,67 oranında doğruluk değeri bulunarak yayınların hangi sınıfa ait olduğunu tespit etmişlerdir [10].

II. GERÇEKLEŞTİRİLEN UYGULAMALAR

Bu çalışmada gerçekleştirilen Twitter ve WhatsApp verileri üzerinde sosyal ağ madenciliği uygulamalarının R programlama ortamında nasıl gerçekleştirildiği ve sonuçları başlıklar halinde sunulmaktadır.

2.1. Twitter Uygulaması

Bu analiz işlemini gerçekleştirmek için; öncelikle, https://dev.twitter.com/ adresinde uygulama oluşturmak amacıyla twitter hesabıyla oturum açılır. Bu ekranda, Şekil 1'de görüldüğü gibi Create New App düğmesi kullanılarak uygulama detaylarına geçilmesi sağlanır. Oluşturulan uygulamadaki consumer key, consumer secret, access token ve access token secret değerleri, yapılacak projede kullanılacağı için kaydedilir.

R programı içerisinde ise twitter bağlantısının başarılı bir şekilde yapılması ve yapılacak metin madenciliği için gerekli olan 'ROAuth', 'RCurl' ve twitteR gibi paketler indirilerek bilgisayara yüklenir.

Geliştirilen yazılım aracılığı ile belirli bir tarih aralığına, atılan tweet'in diline ve özel etiket (hashtag)'e göre arama yapılması sağlanır. Twitter bağlantısı kurulup, girilen herhangi bir hashtag için alınan veri setinde metin temizleme, dönüştürme, parçalama, bazı bağlaç ve edat gibi gereksiz kelimelerin çıkarılması işlemleri yapılır. R programlamada seyrek geçen kelimelerin doküman koleksiyonundan kaldırılması sağlanır. Ardından en sık geçen kelimelerin frekansına göre bir kelime bulutu oluşturulması sağlanmıştır.

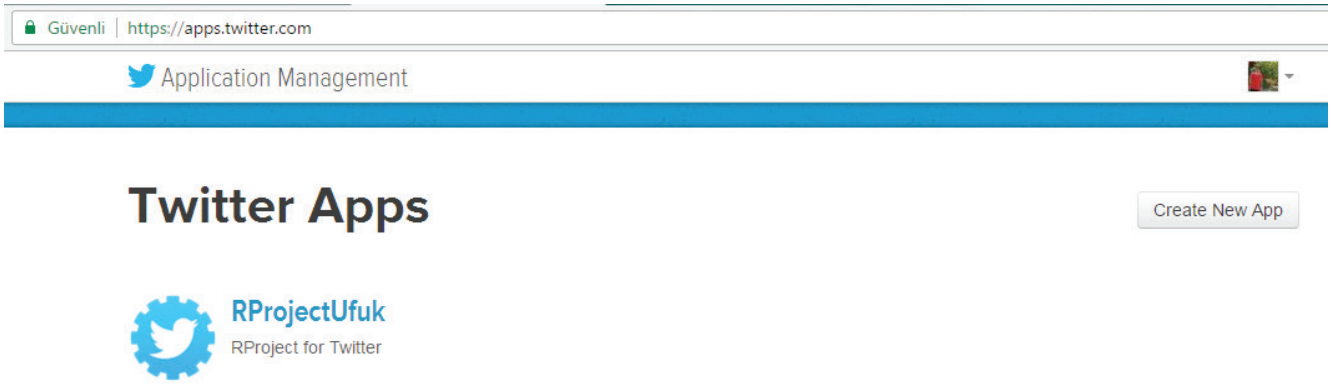
Gerçekleştirilen uygulama ile 01.07.2016 ve 01.07.2017 tarihleri arasında paylaşılan, içinde "*samsung*" ve "*apple*" kelimelerinin geçtiği, 1000 adet tweet program aracılığı ile iki ayrı veri setine kaydedilmiştir. Tweet'lerden oluşturulan doküman koleksiyonları içerisinde küçük harfe dönüştürme işlemi, noktalama işaretlerinin kaldırılması sağlanmıştır. R programlama çeşitli dillerde cümle sonu karakterlerini de kaldırabilmektedir. Uygulamada İngilizce olarak belirlenen dil için cümle sonu karakterlerinin removeWords(x,stopwords()) fonksiyonu ile kaldırılması sağlanmıştır. Yaşanabilecek Türkçe karakter problemi için de iconv(*metin*, "UTF-8", "UTF-8") fonksiyonu kullanılmaktadır.

TermDocumentMatrix fonksiyonu ile tweet'lerden alınan veri seti için bir terim belge matrisinin oluşturulması sağlanır. Burada kullanılan TermDocumentMatrix fonksiyonu, özellikle metin madenciliği çalışmalarında sıklıkla kullanılmaktadır. İki boyutlu bir matrisin bir boyutu metinlere diğer boyutu da terimlere ayrılır. Bu matriste, her metinde o terimden kaç tane olduğunun sayısı tutulmaktadır. Matris 1 ve 0 değerlerinden oluşmaktadır.

Ardından removeSparseTerms fonksiyonu belirtilen değerden daha az seyreklik değerine sahip kelimelerin kaldırılmasını sağlar. Seyreklik değeri 0 ile 1 arasında değişen bir parametredir. Tweet'lerde geçen kelimelerden oluşan belge matrisinde removeSparseTerms() fonksiyonu ile sparse parametresi 0.95 değeri ile seyrek geçen kelimelerin atılması, removeSparseTerms(dtm,sparse = 0.95) komutu ile sağlanır. Bu komuttaki sparse parametresi, belli değerin altında kalan kelimelerin çıkarılması için eşik değeri olarak kullanılır.

Gerçekleştirilen uygulamada hiyerarşik kümeleme kullanılmıştır. Kümeleme işlemi için, ilk olarak Öklit uzaklık formülü kullanılarak kelimeler arası uzaklıklar bulunarak uzaklık matrisi oluşturulur. Hiyerarşik kümeleme işlemini gerçekleştirmek için kullanılan hclust fonksiyonu yığıştırma kümeleme metodudur ve bu fonksiyon yardımıyla kümeleme sonucunda elde edilen gözlemler dendrogram üzerinde gösterilebilmektedir. Dendrogramlar, veri grupları arasındaki bilgileri modellemek için kullanılırlar. Bu örnekte de kelimeler arasındaki bağlantıları ve kümeleme sonucunda elde edilen gözlemleri göstermek için kullanılmıştır.

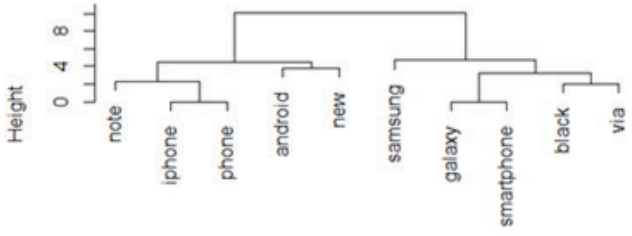
Dendrogramlar, birer ağaç gibi düşünülebilir. Ağacın en altında, yaprak denilen veri setindeki elemanlar bulunur. Yukarıya doğru çıkıldıkça birbirine benzer veya daha çok ilişkili



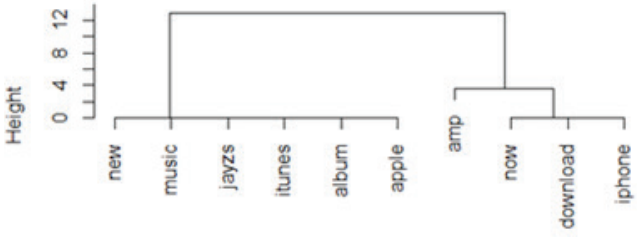
Şekil 1. Twitter İçin Uygulama Oluşturma Ekranı

olan elemanlar, dallar halinde daha üst düzeylerde birleşmeye başlar. Aşağıdaki dendrogram grafiklerinde yer alan "Height", birleşim yüksekliği olarak adlandırılır ve elemanların benzemezliğinin göstergesidir. Ağaçtaki daha uzun bir dal, daha az benzeyen elemanla birleşebilir. Birleşmiş iki elemanın birbirleriyle benzer olup olmadığı dal yüksekliğinden anlaşılır.

Şekil 2 ve Şekil 3'de "*samsung*" ve "*apple*" kelimelerinin geçtiği tweetlerden oluşan veri seti için oluşturulan dendrogram grafikleri yer almaktadır. Örneğin, Şekil 2'deki dendrogram grafiğine bakıldığında " "iphone" ve "phone" kelimelerinin birbirlerine benzer oldukları, bu çalışma için aynı tweet içerisinde kullanılma durumlarının diğer kelimelere göre daha yüksek olduğundan dolayı aynı yükseklik seviyesinde kümelendiği görülmektedir..



Şekil 2. Samsung Kelimesi İçin En Sık Kullanılan Kelimelerin Küme Dendrogramı

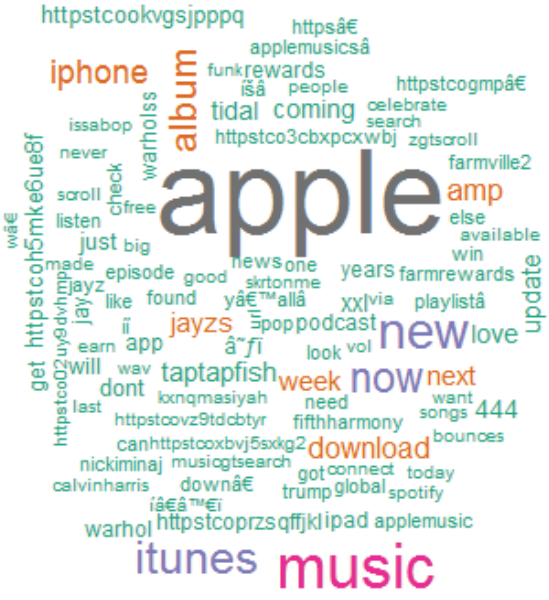


Şekil 3. Apple Kelimesi İçin En Sık Kullanılan Kelimelerin Küme Dendrogramı

Veri seti içerisindeki kelimelerin frekans değerlerine göre büyük, küçük ve renkli gibi çeşitli niteliklerin kullanılması ile kelimelerin bir bulut olarak görselleştirilmesi işlemine kelime bulutu çıkarım işlemi denir. Kelime bulutu sayesinde elverişsiz ve ham veriyi daha anlamlı bir hale getirmek mümkündür. Wordcloud komutu ile belge matrisinde frekans değeri istenilen değerin üzerinde olan kelimelerin yer alacağı, belirlenen sayıda kelime için kelime bulutunun oluşturulması sağlanır. Şekil 4'de Samsung kelimesi için ve Şekil 5'de ise Apple kelimesi için minimum frekansı 3 olan 100 adet olacak şekilde oluşturulan kelime bulutlarına ait ekran görüntüleri görülmektedir.

Şekil 4 ve Şekil 5'deki ekran görüntüleri incelendiğinde, boyut olarak en büyük görünen kelime (samsung ve apple) en çok tekrar eden kelimedir. Kelimelerin boyutu küçüldükçe, kullanım sayıları da azalmış demektir. Wordcloud komutu parametre olarak, words, freq, min.freq, max.words, random.order, colors niteliklerini alır. Bu niteliklerin tamamının kullanılması zorunlu değildir,isteğe bağlı değişebilir. Bu uygulamada, min.freq=3, max.words=100, random.order=T, colors= brewer.pal(8,"Dark2") parametreleri kullanılmıştır. Toplamda kullanılma sayısı 3'ten az olan kelimeler elenmiş olup, kelime bulutunda en fazla 100 kelime gösterilmiştir. Random.order= true (T) olduğu için kelimeleri rastgele çizmiştir.

Şekil 4. Samsung Kelimesi İçin Kelime Bulutu



Şekil 5. Apple Kelimesi İçin Kelime Bulutu

Brewer kütüphanesinde bulunan; Şekil 6'da görülen Dark2 isimli renk paletinde bulunan 8 farklı renk çeşidi kullanılacağı Colors = brewer.pal(8,"Dark2") komutu ile belirtilmiştir. Bu komut içerisinde farklı renk modelleri kullanılabilmektedir. Bu renk paletindeki düzene göre, kelimelerin frekansına söre en sık kullanılandan en az kullanılana göre bir renk atamasının yapılması sağlanmaktadır.

Örneğin bu çalışmadaki örnekte en sık geçen kelimeler için gri, en az geçen kelimeler için yeşil renk kullanılarak kelime bulutunda renklendirme gerçekleştirilmiştir.

Şekil 6. Wordcloud İçin Dark2 Renk Tonları

2.2. WhatsApp Analizi

Dünya üzerinde herkesin dilediği konumdan, herhangi bir ücret olmadan çevresiyle iletişimde kalmalarını sağlayan bu uygulama üzerinde yer alan sohbet geçmişinin analiz edilmesi R programlama ortamı ile mümkün olmaktadır. Gerçekleştirilen çalışmada özellikle grup mesajlarında hangi kullanıcının, hangi tarihte ne sıklıkta mesaj gönderdiği ve aktif olduğu, mesajların zamana bağlı olarak gösterdikleri dağılımı incelemek için bir uygulama geliştirilmiştir.

Bu işlem için öncelikle sohbet geçmişinin tamamı metin dosyası haline dönüştürülmesi gerekmektedir. Bu işlem için, WhatsApp uygulamasının ana ekranında **Seçenekler > Ayarlar > Sohbet Geçmişi > Sohbet geçmişini gönder** düğmesi aracılığı ile E-posta ile göndermek istediğiniz sohbeti seçilir. Herhangi bir grup veya kişisel sohbetler için bu işlemin yapılması mümkündür. Bu işlem sonrasında seçilen sohbet geçmişini belirlenen e-posta adresine metin dosyası olarak txt uzantılı olarak gönderilecektir.

Ardından gönderilen metin dosyası üzerinden, sohbetlerde kimin ne kadar mesaj attığını, hangi ay, gün ve saat dilimlerinde sıklıkla sohbet edildiğini, duygu analizini, sohbette en sık kullanılan kelimelerin grafiğini R programlama ile gerçekleştirilebilir.

Metin analizi iki aşamada gerçekleştirilmektedir. Öncelikle verinin temizlenmesi ardından görselleştirme yapılması gerekmektedir.

2.2.1. Veri temizleme

Veri temizleme işleminde zaman damgası veya özel karakter ile başlayan satırlar belirlenerek verinin standart, analiz edilebilir bir yapıya ulaşması sağlanmaktadır.

Veri temizleme aşamasında, eğer mesaj medya veya duygu ifade eden dijital karakter (emoji) gibi özel karakterler içeriyorsa bunlardan arındırılır. Ayrıca, grup sohbetleri içerisinde yer alan grubun oluşturulması, kişinin gruba eklenmesi veya gruptan ayrılması, grubun adının veya resminin değiştirilmesi gibi grup işlemleri de mesaj olarak geldiği için bunlar da temizlenmiştir. Şekil 7'de duygu ifade eden dijital karakterlerin ASCII koduna çevrilerek silinmesi için kullanılan kod parçası yer almaktadır.

```
1 Messages$Text.Cleaned <- sapply(Messages$Text.Cleaned,
2               function(row) iconv(row, "latin1", "ASCII",
3       sub = ""))
```

Şekil 7. Mesajların Duygu İfade Eden Dijital Karakterlerden Temizlenmesi

2.2.2. Veri işleme ve görselleştirme

Veri işleme aşamasında ise temizlenen verilerin (mesajların), mesajı gönderen kişi, mesajın gönderilme tarihi (gün ay yıl olarak), mesajın içeriği gibi nitelikleri çıkarılır. Bu niteliklere göre çeşitli R fonksiyonları ve kodlamalar kullanılarak veri analiz edilmektedir.

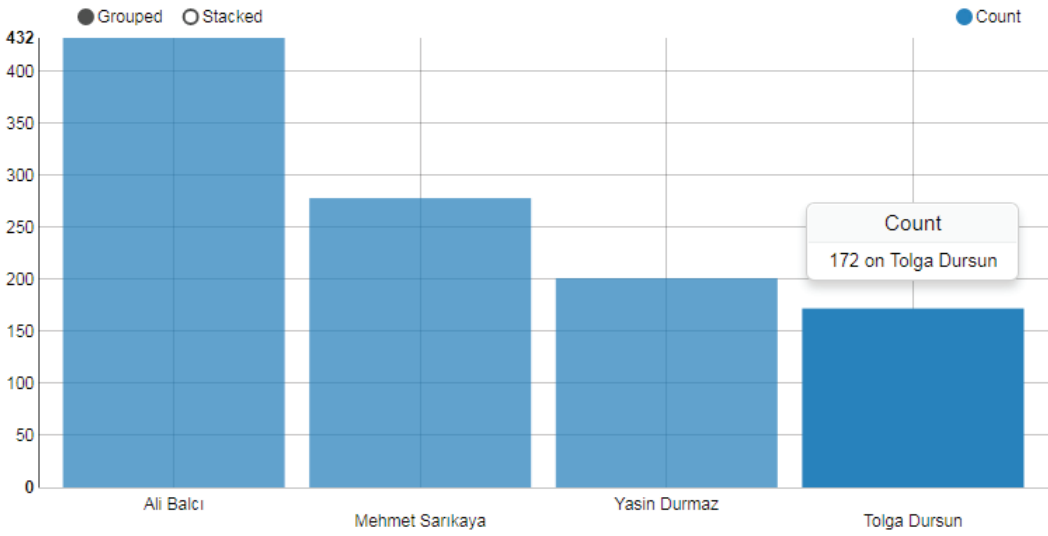
Şekil 8'de kişilerin toplam mesaj sayılarına göre dağılımlarını görselleştirmeyi sağlayan kod parçası bulunmaktadır. Öncelikle, her bir kişinin attıkları mesaj sayısını elde etmek için aşağıdaki kod parçasındaki ilk 4 satırda Person niteliğine göre count işlemi gerçekleştirilmiştir. Ardından 5.satırda R'ın grafik üretme komutu olan Plot komutu ile tüm temizlenmiş mesajları içeren Messages.Aggregated verisi için multiBarChart halinde görselleştirme yapılmıştır. Daha sonra görselin formatı ile ilgili işlemler yapılmıştır.

```
1 Messages.Aggregated <- Messages %>%
2 group_by(Person) %>%
3 summarise(Count = n()) %>%
4 arrange(desc(Count))

5 Plot.All <- nPlot(Count ~ Person, data = Messages.Aggregated,
6 type = "multiBarChart")
7 Plot.All$chart(reduceXTicks = FALSE)
8 Plot.All$xAxis(staggerLabels = TRUE)

9 Plot.All$yAxis(tickFormat = "#!\n function(d)
10 {return d3.format('.')(d)}\n!#")
11 Plot.All
```

Şekil 8. Gönderilen Mesajların Kişilere Göre Görselleştirilmesi İşlemi

Şekil 9 ve Şekil 10'da, örnek bir grup sohbeti için kişilerin attıkları mesaj sayılarının, toplam, aylık ve saatlik olarak grafikleri yer almaktadır.

Şekil 9'da grup sohbetindeki 4 kişinin gönderdikleri toplam mesaj sayıları çubuk grafik halinde gösterilmiştir.

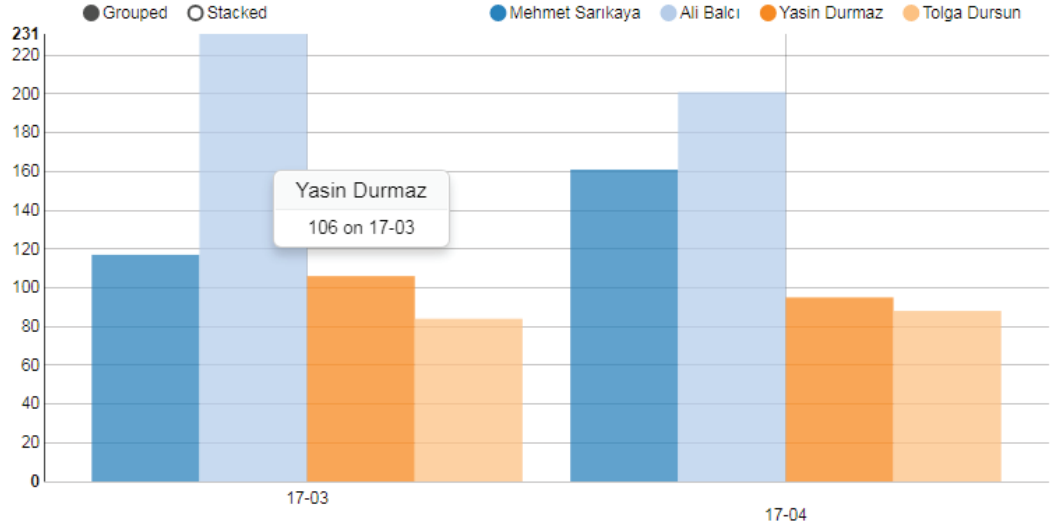


Şekil 9. Sohbetteki Kişilerin Attıkları Mesaj Sayıları

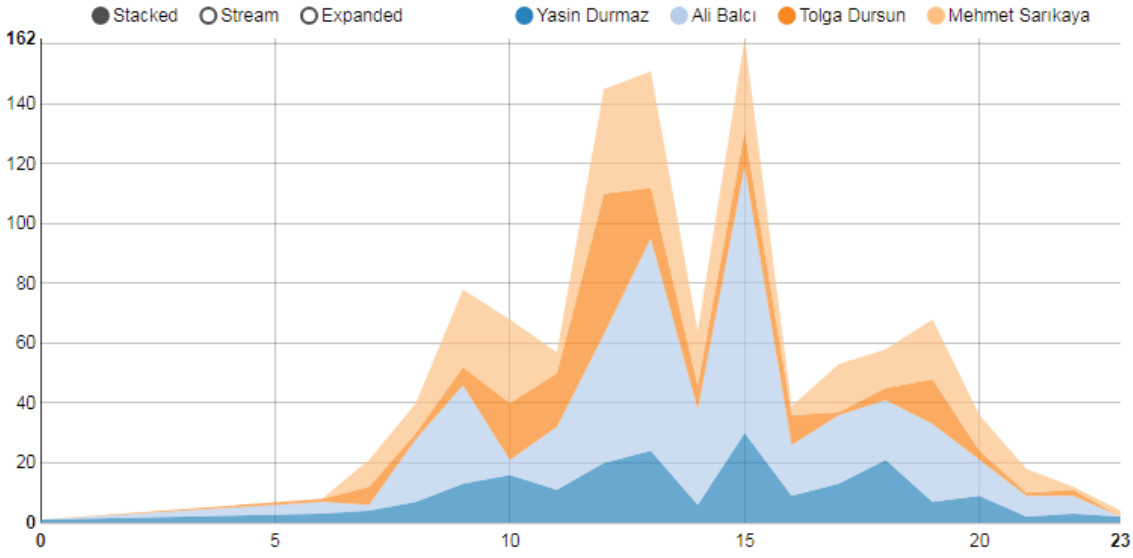
Grafiğin y ekseni gönderilen mesaj sayısını gösterirken, x ekseni ise mesaj gönderen kişileri göstermektedir. Çubukların üstüne fare ile gelindiğinde sayı ipucu (tooltip) olarak çıkmaktadır. Grafiğe göre en fazla mesaj gönderen kişi 432 mesaj ile Ali Balcı, 172 mesaj ile Tolga Dursun olmuştur.

Şekil 10'da ise grup sohbetindeki 4 kişinin gönderdikleri mesaj sayılarının yılın aylarına göre dağılımı çubuk grafik halinde gösterilmiştir. Grafiğin y ekseni gönderilen o ayda gönderilen mesaj sayısını gösterirken, x ekseni ise mesaj gönderme yılı ve ayını (2017 Mart ve Nisan ayları) belirtmektedir. Mesaj gönderen kişiler ise 4 farklı renkte gösterilmiştir ve her renk bir kişiyi temsil etmektedir. Çubukların üstüne fare ile gelindiğinde sayı ipucu (tooltip) olarak çıkmaktadır. Grafiğe göre Ali Balcı, 2017 Mart ayında, 2017 Nisan ayına göre daha fazla mesaj göndermiş ve her iki ayda da en fazla mesaj gönderen kişi olmuştur. Yasin Durmaz ise, 2017 Mart ayında 106 mesaj göndermiştir.

Şekil 11'de grup sohbetindeki 4 kişinin gönderdikleri mesaj sayılarının günün saatlerine göre dağılımı yığın (stack) grafik halinde gösterilmiştir. Grafiğin y ekseni gönderilen o saatte gönderilen mesaj sayısını gösterirken, x ekseni ise 24 saatlik dilim için mesaj gönderme saatini belirtmektedir. Mesaj gönderen kişiler yine 4 farklı renkte gösterilmiştir ve her renk bir kişiyi temsil etmektedir. Grafiğin üstüne fare ile gelindiğinde sayı ipucu (tooltip) olarak çıkmaktadır. Grafiğe göre grup üyeleri arasındaki en çok mesajlaşma saat 15:00 civarında yapıldığı görülmektedir. Saat 15:00'dan sonra mesajlaşma sayısında ciddi bir düşüş gözlemlenmektedir.



Şekil 10. Ay ve Kişilere Göre Atılan Mesaj Sayıları



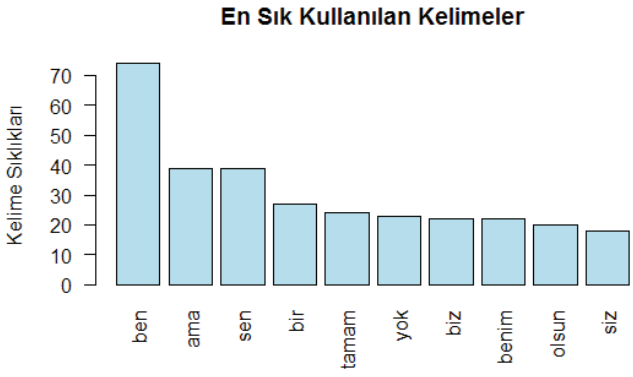
Şekil 11. Saat ve Kişilere Göre Atılan Mesaj Sayıları

Sohbette en sık kullanılan kelimelerin grafiğini üretmek için kullanılan komut parçası Şekil 12'deki gibidir. Kod bloğundaki ilk 10 satırda, R'ın metin temizleme paketi olan ve çeşitli fonksiyonlara sahip olan "tm" paketine ait "tm_map" fonksiyonu kullanılarak, her bir mesaj verisi üzerinde sayıların çıkarılması, tüm metnin küçük harfe çevrilmesi, boşlukların silinmesi, "@, /, \\" gibi özel karakterlerden temizlenmesi gibi işlemler yapılmıştır. Daha sonra, "rowSums" fonksiyonu ile her bir kelimenin sayısı çıkarılmış, "sort" fonksiyonu ile en sık kullanılan kelimeden en az kullanılana göre azalan sırada sıralanmış ve 14.satırda "d" adında "word" ve "freq" niteliklerini içeren bir veri seti oluşturulmuştur. Bu veri setinde, "word" mesajlardaki kelimeleri, "freq" ise her kelimenin kullanılma sayısını belirtmektedir. 15 ve 16.satırlarda ise önce en sık kullanılan ilk 10 kelime "head" komutuyla ekrana yazdırılmış ve "barplot" komutuyla da ekrana yazdırılan ilk 10 kelime çubuk grafik halinde görselleştirilmiştir.

```
1 docs <- Corpus(VectorSource(Messages$Text.Cleaned))
2 toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
3 docs <- tm_map(docs, toSpace, "/")
4 docs <- tm_map(docs, toSpace, "@")
5 docs <- tm_map(docs, toSpace, "\\|")
6 docs <- tm_map(docs, content_transformer(tolower))
7 docs <- tm_map(docs, removeNumbers)
8 docs <- tm_map(docs, removeWords, stopwords("english"))
9 docs <- tm_map(docs, removePunctuation)
10 docs <- tm_map(docs, stripWhitespace)
11 dtm <- TermDocumentMatrix(docs)
12 m <- as.matrix(dtm)
13 v <- sort(rowSums(m),decreasing=TRUE)
14 d <- data.frame(word = names(v),freq=v)
15 head(d, 10)
16 barplot(d[ 1: 10,]$freq, las = 2, names.arg = d[1:10,]$word,
        col ="lightblue", main ="En Sık Kullanılan Kelimeler",
        ylab = "Kelime Sıklıkları")
```

Şekil 12. En Sık Kullanılan Kelimelerin Grafiğini Üretmek İçin Kullanılan Kod Parçası

Şekil 13'deki grafikte, grup sohbetinde kullanılan kelimelerden en sık kullanılan ilk 10 kelime ve kullanılma sayıları yer almaktadır. Grafiğe göre, en sık kullanılan kelime 74 defa ile "ben" kelimesi olmuştur.



Şekil 13. Mesajlarda En Sık Kullanılan Kelimeler ve Kullanılma Sıklıkları

III. SONUÇ VE ÖNERİLER

R programlama ortamı sosyal ağ analizi için etkin, hızlı ve esnek bir program geliştirme ortamı sağlamaktadır. Bu ortam aracılığı ile Twitter ve WhatsApp ortamlarından elde edilen verilerin analizinin nasıl yapılacağına dair gerçekleştirilen iki uygulama bu çalışma kapsamında açıklanmakta ve sonuçları ortaya konmaktadır. Farklı zaman aralıkları, anahtar kelimeler ve amaçlar için bu ortamın kullanılması mümkündür.

Bu çalışmada geliştirilen yazılım sayesinde Twitter'dan veriler alınmış ve bir veri seti oluşturulmuştur. Oluşturulan veri setinin kullanılabilmesi için R programlama dili ve R Studio platformu kullanılarak metin temizleme, dönüştürme, parçalama, bazı bağlaç ve edat gibi gereksiz kelimelerin çıkarılması gibi işlemler yapılmıştır. Eldeki işlenmiş veri setinin daha anlamlı ve işe yarar bir hale gelmesi için kelime bulutu ve dendrogram grafiği çıkarım işlemi yapılmıştır. Yapılan işlemler sayesinde kelimeler arasındaki ilişkiler daha açık bir şekilde gözlemlenmektedir.

R programlama dili ve R Studio platformu kullanılarak, WhatsApp üzerinden gerçekleştirilmiş görüşmelerde kimin ne kadar mesaj attığı, hangi ay, gün ve saat dilimlerinde sıklıkla sohbet edildiği ve en sık kullanılan kelimelerin grafik çıkarım işlemi gibi işlemler yapılmıştır.

Sosyal ağlarda bir zaman aralığında atılan veriler kullanarak, farklı cinsiyet ve yaş gruplarının ürün tercihleri ve yönelimleri ile ilgili analiz çalışmaları gerçekleştirmek, o alandaki mevcut durumun gerçekçi sonuçlarını ortaya koyabilmektedir. Sosyal ağlar bu tip satış pazarlama eğilimleri için kullanılabileceği gibi öğretim etkinlikler için de kullanılma kapasiteleri bulunmaktadır. Bir sonraki çalışmada öğrenci gruplarının eğitim-öğretim yarıyılı içerisinde, ders faaliyetlerde ve iletişimlerinde kullandıkları sosyal ağların analizinin gerçekleştirilmesi ve bu analiz sonuçlarına göre öğrenme güçlüğü yaşanılan durumların tespit edilmesinin sağlanması planlanmaktadır.

Kaynaklar

[1] Twitter number of users worldwide , https://www.statista.com/statistics/303681/twitter-users-worldwide/, (Temmuz 2017)

[2] Number of monthly active WhatsApp users worldwid, https://www.statista.com/statistics/260819/number-of-monthly-active-whatsapp-users/, (Temmuz 2017)

[3] Nasution, M. K., Sitompul, O. S., Sinulingga, E. P., & Noah, S. A. (2016). *An extracted social network mining.* Paper presented at the SAI Computing Conference (SAI), 2016.

[4] RStudio- Open source and enterprise-ready professional software for R., https://www.rstudio.com/ (Temmuz 2017)

[5] Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics, 5*(3), 299-314.

[6] Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M. M., & Alamri, A. (2017). Reputation-based credibility analysis of Twitter social network users. *Concurrency and Computation-Practice & Experience, 29*(7). doi: ARTN e387310.1002/cpe.3873

[7] Zhang, K. P., Bhattacharyya, S., & Ram, S. (2016). Large-Scale Network Analysis for Online Social Brand Advertising, Mis Quarterly, 40(4), 849-+.

[8] Iglesias, J. A., Garcia-Cuerva, A., Ledezma, A., Sanchis, A., & Ieee. (2016, Oct 09-12). *Social Network Analysis: Evolving Twitter.* Paper presented at the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, HUNGARY.

[9] Abascal-Mena, R., Lema, R., & Sedes, F. (2015). Detecting sociosemantic communities by applying social network analysis in tweets. [Article]. *Social Network Analysis and Mining, 5*(1), 17. doi: 10.1007/s13278-015-0280-2

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