ESTUDAM Bilişim Dergisi
Cilt 1, Sayı 2 , 10-13, 2020

Journal of ESTUDAM Information
Volume 1, Issue 2, 10-13, 2020

(Geliş Tarihi / ReceivedDate: 23.01.2020, Kabul Tarihi/ AcceptedDate: 19.04.2020)

Twitter'daki Verilere Metin Madenciliği Yöntemlerinin Uygulanması

Hatice Kübra ÜÇÜKKARTAL[*1]

[1]Eskişehir Osmangazi Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Böl., 26480, Eskişehir

Anahtar Kelimeler:
Sosyal Medya,
Metin Madenciliği,
Veri Madenciliği,
Veri Görselleştirme,
Duygu Analizi

Özet: Sosyal medya insanların birbirleriyle iletişimini daha etkileşimli hale getirmiştir. Bu etkileşimler sayesinde çok büyük miktarlarda veriler üretilmektedir. Bu sayede gün geçtikçe artan verilerin işlenmesi ve analiz edilmesine yönelik ihtiyaçlar ortaya çıkmaktadır. Bu çalışmanın amacı, sosyal medya platformu olan twitter üzerinden insanların tartıştığı konular hakkında fikir sahibi olmaktır. İnsanların görüşlerinin sonuçlarını göstermek metin madenciliği konusunda alt yapı oluşturmak ve veri görselleştirme yöntemleriyle de verileri daha anlamlı hale getirmek istenmiştir. Bu çalışmada Twitter üzerinden toplanan verilere metin madenciliği yöntemleri uygulanmıştır. Metin madenciliğini, veri bilimi ve veri görselleştirme araçlarıyla birleştirerek veriler kolay anlaşılır hale getirilmiştir. Duygu analizi ile insanların yaptıkları paylaşımların pozitif, negatif veya nötr olma durumu analiz edilmiştir.

Applying Text Mining Methods to Twitter Data

Keywords:
Social Media,
Text Mining,
Data Mining,
Data Virtualization,
Sentiment Analysis

Abstract: Social media has made people's communication more interactive. Thanks to these interactions, huge amounts of data are produced. In this way, the needs for processing and analyzing increasing data are emerging. The aim of this study is to have an idea about the issues that people discuss via the social media platform twitter. It was aimed to show the results of people's opinions, to create a substructure on text mining and to make the data more meaningful with data visualization methods. In this study, text mining methods were applied to the data collected on Twitter. By combining text mining with data science and data visualization tools, data is made easy to understand. Sentiment analysis was used to analyze whether people's posts were positive, negative or neutral.

1. GİRİŞ

Sosyal medya insanların birbirleriyle iletişimini daha etkileşimli hale getirmiştir. Bu etkileşimler sayesinde çok büyük miktarlarda veriler üretilmektedir. Bu sayede gün geçtikçe artan verilerin işlenmesi ve analiz edilmesine yönelik ihtiyaçlar ortaya çıkmaktadır.

Twitter'da paylaşılan içerikler yapısal olmayan veri türlerindendir. Metin tabanlı olan bu yapısal olmayan veriler analiz etmek için, veri madenciliği algoritmalarında kullanılabilecek yapısal formata dönüştürülmesi gerekir. Bu verileri yapısal veri formatına dönüştürmek ve analiz etmek için metin madenciliği yöntemleri uygulanır.

Metin madenciliği, doğal dil işleme ve veri madenciliğinin birlikte kullanılmasıdır[1]. Doğal dil işleme, Natural Language Processing (NLP) olarak bilinen yapay zeka ve dilbilimin alt kategorisidir. İnsanların konuştuğu dillerin işlenmesi ve kullanılması amacı ile araştırma yapan bilim dalıdır[2]. Veri madenciliği, büyük veri yığınlarını anlamlı ve faydalı bilgiye dönüştürme işlemidir[3].

Bu çalışmanın amacı, sosyal medya platformu olan twitter üzerinden insanların tartıştığı konular hakkında fikir sahibi olmaktır. İnsanların görüşlerinin sonuçlarını göstermek metin madenciliği konusunda alt yapı oluşturmak ve veri görselleştirme yöntemleriyle de verileri daha anlamlı hale getirmek istenmiştir.

İlgili yazar: Hatice Kübra KÜÇÜKKARTAL, haticekubra26@gmail.com)

2. MATERYAL VE METOT

2.1. Projede Kullanılan Araçlar

Bu projede kullanılacak olan araçlar şu şekildedir;

- Twitter Api
- Python
- Jupyter Notebook

Twitter Api, twitter'dan verileri çekmek için kullanılacak olan arayüzdür.
Python, veri analizi, veri işleme makine öğrenmesi ve veri görselleştirme için kullanılacak olan programalama dilidir.
Jupyter Notebook, twitter'dan çekeceğimiz metinleri analiz edeceğimiz, python kodlarını çalıştıracağımız geliştirme ortamıdır.

Kullanılan kütüphaneler;

- Tweepy (Twitter Api)
- Pandas (Veri analiz aracı)
- Matplotlib (Veri görselleştirme)
- Seaborn (Veri görselleştirme)
- WordCloud (Veri görselleştirme)
- TextBlob (Metin verilerini işleme)
- NLTK (Doğal dil işleme paketi)
- Scikit-learn (Yapay öğrenme)

2.2. Veri Seti

Twitter'dan anlık olarak veri çekme işlemi Twitter API (Twitter Apps, 2019) ile gerçekleştirilmiştir. Twitter API kullanmak için Twitter Application oluşturulmuştur. Tweetlere erişim sağlayabilmek için gerekli izinler (key, secret) alınmıştır. Anahtar kelime olarak kullanıcının girdiği arama terimini içeren tweetler toplanmıştır. Bu işlemler için Python dilinin tweepy kütüphanesi kullanılmıştır. Twitter'dan çıkarılan yaklaşık 10000 İngilizce tweet üzerinde işlem yapılmıştır.

Ham tweetler Python dilinin Pandas kütüphanesi kullanılarak csv dosyasına yazılır. Daha sonra bu csv dosyasındaki verilere temizleme işlemi uygulamak için bazı metin madenciliği yöntemleri uygulanır.

2.3. Metin Madenciliği ile Ön İşleme

Metin madenciliği çalışmaları metni veri kaynağı olarak kabul eden veri madenciliği çalışmasıdır. Diğer bir tanımla metin üzerinden yapısallaştırılmış veri elde etmeyi amaçlar[12]. Metin Madenciliği, doküman koleksiyonlarının önişlemden geçirilmesi, ara sonuçların saklanması, ara sonuçların analiz edilmesi için çeşitli tekniklerin kullanılması ve nihai sonuçların görselleştirilmesi gibi aşamalardan oluşmaktadır[13].

Twitter'dan toplanan veriler metin tabanlı olduğu için metin madenciliği yöntemleri uygulanarak kullanılabilecek formata dönüştürülür.

Tweetler, sözdizimsel olarak iyi oluşturulmamıştır. Bu yüzden bazı ön işlemlerden geçirilmiştir. Bunlar;

- Tokenization
- Stopwords
- Lemmatization
- Twitter metinlerindeki büyük harfler küçük harflere dönüştürülmüştür.
- Metinler, noktalama işaretlerinden arındırılmıştır.
- Metin içeriğindeki URL, hashtag ve kullanıcı isimleri kaldırılmıştır.
- Metinlerdeki sayısal ifadeler kaldırılmıştır.
- TF-IDF (Term Frequency - Inverse Document Frequency) ağırlıklandırma işlemi yapılmıştır.

Tokenization, metinleri istenilen özelliklere göre parçalama işlemidir. Stopwords, bir dilde sık kullanılan etkisiz kelimeleri filtreleme işlemidir. Lemmatization, metindeki kelimeleri köküne indirgeme işlemidir. Twitter metinleri içerisinde '#' ile başlayan hashtag, '@' ile başlayan kullanıcı isimleri, URL formatında web sitesi adresleri, sayısal ifadeler ve noktalama işaretleri barınabilmektedir. Metinler bu karakterlerden temizlenmiştir. Bu işlemler için Python dilinin NLTK kütüphanesi kullanılmıştır.

TF-IDF, bir kelimenin doküman içerisinde ne kadar önemli olduğunu değerlendirmede kullanılan istatistiksel bir ölçüdür. Terim sıklığı (Term Frequency), bir doküman içerisinde geçen terim ağırlıklarını hesaplamak için kullanılan yöntemdir[14]. Ters doküman sıklığı (Inverse Document Frequency), birden fazla dokümanda kelimenin geçme sayısını bularak bu kelimenin terim olup olmadığını bağlaç vb (Stop Words) olduğu anlamaya çalışır. Bunun için Terimin Geçtiği Dokuman Sayısı / Dokuman Sayısı'nın logaritmasının mutlak değeri alınmaktadır[15]. Bu projede her bir tweet doküman olarak baz alınmıştır. Bu işlemler için Python dilinin Scikit-Learn kütüphanesi kullanılmıştır.

2.4. Veri Görselleştirme

Veri görselleştirme, soyut bilgilerin grafik biçiminde sunulmasını tanımlar. Normalde klasik formatta sunulan karmaşık ve dağınık haldeki verileri, kolay algılanabilir görseller ile rahatça anlaşılır ve yorumlanabilir hale getirmektir.

Görselleştirme metotlarından biri olan WordCloud, türkçe karşılığı kelime bulutu olarak geçmektedir. Kelime bulutları genellikle kelime sıklıklarına göre analiz sonrasında, farklı renk ve desenlerle oluşturulan metin görselleridir[16].

Twitter'dan çekilen tweetler temizleme ve ön işlemden geçirildikten sonra tweetlerde en çok geçen kelimeler, Pyhton içerisinde bulunan WordCloud kütüphanesi kullanılarak veriler görselleştirilmiştir.

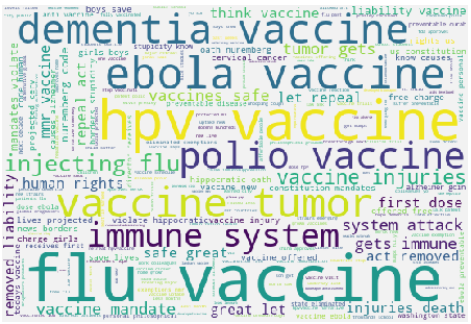
2.5. Veri Görselleştirme ile Duygu Analizi

Duygu analizi, fikir madenciliği olarak da bilinmektedir. Metinler içerisindeki gizli olan duygu, fikir ve düşüncelerin olumlu, olumsuz ya da nötr olup olmadığını belirleyen en yaygın metin sınıflandırma yöntemidir.

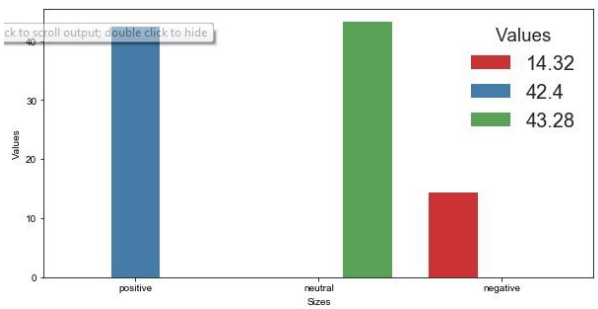
Duygu analizi için Python dilinin TextBlob kütüphanesi kullanılmıştır. Ve twitter metinlerindeki her bir kelimenin polarite skoru ve polaritesi bulunmuştur. Her bir tweet için bulunan polarite skoruna, pozitif, negatif ve nötr durumlarına göre değerlendirilmiştir. Tweetler, metinlerde geçmekte olan pozitif, negatif ve nötr kelime sayısına göre skor değeri almıştır. Daha sonra bu değerler Python dilinin Seaborn kütüphanesi kullanılarak veri görselleştirmesiyle daha anlaşılır hale getirilmiştir.

3. BULGULAR

Twitter üzerinden, Twitter API ile kullanıcının girdiği arama terimi olan vaccine(aşı) içeren 10000 İngilizce tweet çekilmiştir. Tweetler gerekli ön işlemden geçirildikten sonra analiz aşamasına geçilmiştir. Pandas kütüphanesi kullanılarak tweetlerde geçen kelimeler ve kelimelerin sayısı bulunmuştur. Bu kelimeler WordCloud kütüphanesi kullanılarak görselleştirilmiştir. Şekil-1'de uygulanan çalışmanın sonucunda elde edilen kelimelerin görseli gösterilmektedir. Sonuca göre vaccine(aşı), flu(grip), hpv(HPV), tumor(tümör), ebola(Ebola), dementia(bunaklık) ve polio(çocuk felci) kelimeleri en çok geçen kelimelerden olmuştur. Bu sonuçlarla vaccine hakkında yazan kişilerin çoğunluğu hastalıklarla ilgili paylaşımlar yapmıştır. Şekil-2'de ise elde edilen polarite sonuçları gösterilmektedir. Sonuçlara göre tweetlerin yaklaşık %42'si pozitif, %14'ü negatif ve %43'ü ise nötr olduğu görülmektedir. Pozitif ifade mavi renkle, negatif ifade kırmızıyla ve nötr ifade ise yeşil renkle etiketlenmiştir.



Şekil 1. Tweetlerde en çok geçen kelimelerin sonucu



Şekil 2. Tweetlerin duygu analizi sonucu

4. SONUÇ

Bu çalışmada Twitter üzerinden toplanan verilere metin madenciliği yöntemleri uygulanmıştır. Metin madenciliğini, veri bilimi ve veri görselleştirme araçlarıyla birleştirilerek veriler kolay anlaşılır hale getirilmiştir. Duygu analizi ile insanların yaptıkları paylaşımların pozitif, negatif veya nötr olma durumu analiz edilmiştir.

KAYNAKÇA

[1] İnternet: H.,Takcı, Metin Madenciliği Çerçevesi, , http:// verimadencisi.blogspot.com/2012/10/metin-madenciligi-cercevesi.html, (Erişim Tarihi: 20.01.2020)

[2] İnternet: Doğal Dil İşleme, https://tr.wikipedia.org/wiki/Doğal_dil_işleme, (Erişim Tarihi: 20.01.2020)

[3] İnternet: Veri Madenciliği, https://tr.wikipedia.org/wiki/Veri_madenciliği, (Erişim Tarihi: 20.01.2020)

[4] İnternet: TWEEPY Kütüphanesi, http://www.tweepy.org/, (Erişim Tarihi: 20.12.2019)

[5] İnternet: PANDAS Kütüphanesi, https://pandas.pydata.org/, (Erişim Tarihi: 20.12.2019)

[6] İnternet: MATPLOTLİB Kütüphanesi, https://matplotlib.org/, (Erişim Tarihi: 20.12.2019)

[7] İnternet: SEABORN Kütüphanesi, https://seaborn.pydata.org/, (Erişim Tarihi: 20.12.2019)

[8] İnternet: WORDCLOUD Kütüphanesi, https://anaconda.org/conda-forge/wordcloud, (Erişim Tarihi: 20.12.2019)

[9] İnternet: TEXTBLOB Kütüphanesi, https://anaconda.org/conda-forge/textblob, (Erişim Tarihi: 20.12.2019)

[10] İnternet: NLTK Kütüphanesi, https://www.nltk.org/, (Erişim Tarihi: 20.12.2019)

[11] İnternet: SCİKİT-LEARN Kütüphanesi, https://scikit-learn.org/, (Erişim Tarihi: 20.12.2019)

[12] İnternet: Metin Madenciliği, https://tr.wikipedia.org/wiki/Metin_madenciliği/, (Erişim Tarihi: 22.01.2020)

[13] İnternet: Metin Madenciliği (Text Mining) Nedir, https://www.vtunali.com/tr/index.php/2009/10/metin-madenciligi-text-mining-nedir/, (Erişim Tarihi: 22.01.2020)

[14] İnternet: Onur, D. 2015. Term Frequency. https://medium.com/algorithms-data-structures/tf-idf-term-frequency-inverse-document-frequency-53feb22a17c6, (Erişim Tarihi: 22.01.2020)

[15] İnternet: Onur, D. 2015. Inverse Term Frequency. https://medium.com/algorithms-data-structures/tf-idf-term-frequency-inverse-document-frequency-53feb22a17c6, (Erişim Tarihi: 22.01.2020)

[16] İnternet: Hüseyin, D. 2014. Kelime Bulutu. https://huseyindemirtas.net/kelime-bulutu/, (Erişim Tarihi: 22.01.2020)