Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi
Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **20** (2020) 025503 (276-286) AKU J. Sci. Eng. **20** (2020) 025503 (276-286)
DOI: 10.35414/akufemubid.667397

Araştırma Makalesi / Research Article

Sosyal Medyanın Gönüllü Coğrafi Veri Olarak Kullanımı ve Sosyal Medya Verilerinden Coğrafya Sözlüğü Üretimi

Ayşe Giz GULNERMAN[1*], Himmet KARAMAN[1]
[1] *Istanbul Teknik Üniversitesi, İnşaat Fakültesi, Geomatik Mühendisliği Bölümü, Istanbul.*

*Sorumlu yazar e-posta: *gulnerman@itu.edu.tr ORCID ID: http://orcid.org/0000-0002-9163-6068*
e-posta: karamanhi@itu.edu.tr ORCID ID: http://orcid.org/0000-0003-4923-3561

Geliş Tarihi: 30.12.2019 Kabul Tarihi: 24.03.2020

Anahtar kelimeler
Sosyal Medya (SM); Mekansal Veri Madenciliği; Gönüllü Coğrafi Bilgi (GCB); Coğrafi Ayrıştırma; Coğrafya Sözlüğü

Öz

Gelişen web ve akıllı mobil teknolojileri ile sosyal medya platformları yaygınlaşmıştır. Son 10 yılda bu platformlardaki aktif kullanıcı sayısının artması veri üretimine de yansımıştır. Sosyal medya platformları aracılığı ile üretilen mekansal veri doğrudan ya da dolaylı kullanımlarla afet yönetimi, pazarlama, politika gibi geniş çerçevede katkılar sunmaktadır. Bu veri geleneksel gönüllü coğrafi bilgi projelerinde üretilen verinin aksine yapılandırılmamış ve çoğunlukla belirli bir amaç için projelendirilmeden üretilen karma bir veridir. Bu nedenle veri üzerinde çalışılacak konuya göre metin analizleri ile filtreleme yapmanın yanında verinin mekansal tarafını ele almak için coğrafi etiketleme ve referanslama konusunda ön işleme yapmayı gerektirmektedir. Bu makalenin amacı, gönüllü coğrafi bilginin bir alt başlığı olan sosyal medya verilerinin mekansal veri olarak kullanımını değerlendirerek, metinlerden coğrafi bilgi çıkarımı yaklaşımlarını tanıtmaktadır. Coğrafi ayrıştırmada ihtiyaç duyulan coğrafya sözlüğü üretimi için bir metodoloji sunmaktadır. Sunulan metodoloji İstanbul ve Londra için üretilen tweetlerde test edilmiş ve ilgi noktalarının tespitinde özellikle bina bazında temsil edilen alanlar için başarı sağlamıştır. Bu çalışma, doğal dilden bağımsız ve coğrafi tekrarlılığa dayalı coğrafi veri elde etme metodolojisi ile literatüre katkı sağlamaktadır.

Use of Social Media as a Volunteered Geographic Data and the Gazetteer Production from Social Media Data

Keywords
Social Media; Spatial Data Mining; Volunteered Geographic Information; Geo-parsing; Gazetteer

Abstract

Social media platforms became widespread thanks to the developments in web and smart mobile technologies. Produced data volume has tremendously increased with the growing number of active users in these platforms in the last decade. Spatial data generated through social media platforms, that is in-/directly produced, contribute to diverse topics such as disaster management, marketing, and policy. This data, unlike the general voluntary geographic information, is unstructured and undirected for a project or for a specific purpose. Therefore, it requires pre-processing and filtering for text analysis according to the subject to be studied, and evaluation for direct or indirect spatial data for geospatial analysis. The aim of this article is to introduce and discuss the use of social media data as a subtitle of voluntary geographic information over geo-parsing approaches. This article also presents a methodology for the production of a gazetteer, which is required for geo-parsing techniques. The proposed methodology in this study is tested with the tweets generated within Istanbul and London areas and it is succeeded especially in the detection of point of interest that is representing the buildings. This study contributes to the literature of geographic data retrieval with the methodology, which is independent of natural language and based on the geographic data repetitiveness.

1. Giriş

Gönüllü Coğrafi Bilgi (GCB), belli bir konuda uzmanlık gerekmeksizin, gönüllüler tarafından belirli yazılımlardaki araçlar kullanılarak üretilen coğrafi veridir (Goodchild 2007). Turner (2006) bu üretim şeklini yeni-coğrafya (neogeography), bu veriyi

üreten gönüllüleri ise yeni-coğrafyacılar olarak adlandırmıştır. GCB literatürden özetle 3 başlıkta ele alınır; 1- Halk Katılımlı Coğrafya (Schroeder 1996), 2- Salt Harita Üretimi, 3- Sosyal Medya (Gulnerman et al. 2016, Hecht et al. 2011). Halk katılımlı coğrafya çoğunlukla kent planlamada proje paydaşlarının (mahalle sakini, ziyaretçi, iş yeri sahibi vb.) öneri ve taleplerini gözetmek amacı ile tasarlanmış web tabanlı ya da masaüstü coğrafi yazılımlar ile coğrafi veri üretimine dayanmaktadır (Sieber 2006). Salt harita üretimi örneklerinde ise, gönüllüler çeşitli çevrimiçi platformlarla sadece haritalama yapmaktadır. Salt harita üretimi türünde GCB temsillerinden en ünlüsü Open Street Map (OSM) platformu 2004 yılında çevrimiçi olarak kurulmuş, gönüllüler tarafından üretilen coğrafi verileri açık veri olarak sunmaktadır. GCB başlıklarından Sosyal Medya (SM) ise sosyal medya platformları aracılığı ile doğrudan ya da dolaylı olarak coğrafi nitelik içeren verileri kapsamaktadır.

GCB bu üç alt başlıktaki yöntemlerle, geleneksel veri toplama ve harita üretme yöntemlerine alternatif olmakla kalmamış, veri kapsamı, büyüklüğü, sürekliliği ve maliyeti gibi konularda öne geçmiştir. Özellikle sosyal medya, biyonik sensör olarak adlandırılan aktif kullanıcıları sayesinde, internetin kısıtlı olmadığı tüm dünya ülkeleri için, gerçek zamanlı bir mekansal izleme imkanı sunarak, GCB ’ye en büyük katkı sağlama potansiyeline sahiptir.

2019 Eylül ayı verilerine göre, sosyal medya platformları 2.3 milyardan fazla aktif kullanıcıya sahiptir. Bu platformlar, konum kütüphaneleri aracılığı ile coğrafi etiketleme (paylaşılan verinin metin içeriğine coğrafi isimlerin eklenmesi) özelliği sağlamakta, ayrıca mobil aygıtlarda yer alan GNSS (Global Navigation Satellite System) antenleri ile de konum bilgisini (enlem, boylam) gönderilere ekleme imkanı sunmaktadırlar. Yapılan araştırmalarla, sosyal medya verilerinin büyük bir kısmında konum bilgisinin doğrudan paylaşılmadığı öne sürülmüştür (Hecht and Stephens 2014). SM paylaşımlarında, yukarıda bahsedilen şekillerde doğrudan coğrafi veri paylaşımının yanı sıra metin içerisinde dolaylı konumsal veri olması muhtemeldir. GCB‘ye sağlanan verinin arttırılması için dolaylı coğrafi veri içeren bu metin içeriklerinden konum bilgisinin çıkarılması üzerine çalışmalar önemlidir. Dolaylı verideki coğrafi sözcükleri ayrıştırmak doğal dil işleme (NLP) yaklaşımları ile mümkün olabilmekte ve ayrıştırılan bu sözcükler coğrafya sözlükleri sayesinde konumlanabilmektedir.

Bu çalışma, sosyal medya verilerinin, GCB olarak kullanımı ve kullanım kapasitesini arttırmak için var olan coğrafi ayrıştırma tekniklerini ele almaktadır. Makalenin materyal metot bölümünde, sosyal medya verilerini elde etmek için kullanılan araç ve yöntemler ve dolaylı konum verisi üzerinde coğrafi ayrıştırma için kullanılan tekniklere yer verilmiştir. Buna ek olarak, coğrafi ayrıştırma için ihtiyaç duyulan coğrafya sözlüğünün sosyal medya verilerinden elde edilebilmesi için izlenecek metodoloji üçüncü başlıkta tanımlanmıştır. Bulgular bölümünde uygulanan metodolojinin çıktıları, karşılaşılan özel durumlarla birlikte sunulmaktadır. Sonuç ve tartışma bölümünde ise verinin kalitesi ve güvenilirliği ele alınarak, veri kullanımının potansiyeli ve gelecek çalışmalarda geliştirilmesi gereken konulardan bahsedilmektedir.

2. Materyal ve Metot

Akıllı teknolojilerin ve konum servislerinin yaygınlaşması ile sosyal medya platformları kullanıcılar açısından haber alma/verme, sosyal iletişim kurma, reklam ve pazarlama gibi çeşitli nedenlerle tercih edilirken, bu platformlar da aktif kullanıcı sayısındaki artış ile büyük, anlık ve sürekli bir veri kaynağı haline gelmiştir. Kullanıcıların sosyal medyadan beklentilerindeki farklılığa göre, kullanılan platformlarda farklılık gösterebilmektedir. Örneğin günlük aktiviteleri ve yakın çevresi ile ilgili paylaşım yapacak kullanıcılar tarafından Facebook, fotoğraf düzenleme ve görsel paylaşımda bulunan kullanıcılar tarafından Instagram, haber verme/alma, politik görüş paylaşma/destekleme ya da iş takibi gibi daha resmi konular için ise Twitter daha çok tercih edilmektedir. Twitter, iletişim uzmanları tarafından yeni medya olarak da isimlendirilmekte ve bu konuda iletişim alanında özellikle politik konuları da içine alan geniş bir yelpazede kullanılmaktadır (Tufekci and Wilson 2012, Unver 2017).

Twitter ulusal ve uluslararası alanda en çok tercih edilen 10 sosyal medya platformu arasında yer almaktadır. Twitter aktif kullanıcılarının ürettiği

verilerle, geleneksel medya kaynaklarına alternatif olabilmekte, bu özelliği ile acil durumlarda, olay niteliği ve konumunun tespiti (McCreadie et al. 2013, Sakaki et al. 2010); sağlık araştırmalarında, salgın haritaları (Signorini et al. 2011); kent araştırmalarında, trafik kazaları ve yoğunluğu (D'Andrea et al. 2015, Gong et al. 2015, Hasby and Khodra 2013, Ishino et al. 2012, Kosala and Adi 2012); psikoloji alanında, terör saldırısı sonrası duygu durumunun haritalanması (Lin and Margolin 2014) gibi birçok çalışmaya mekansal veri sağlamaktadır. Bu gibi önemli kullanım alanları nedeni ile bu çalışmada sosyal medya platformları arasından Twitter ele alınmıştır.

Türkiye, 9 milyona yakın aktif kullanıcısı ile Twitter'da en çok kullanıcısı bulunan beşinci ülkedir (Int Kyn. 1). 2006 yılında kurulan Twitter, yer-referanslama (geo-referencing) özelliğini 2009 yılında tanıtmıştır. Bu özellik sayesinde yapılan paylaşımlar coğrafi olarak konumlanabilmektedir. 2010 'da coğrafi etiketleme özelliğini kullanıma açan platform, yer imleri kütüphanesi ile ilgi noktalarının (Point of Interest) paylaşımlara eklenebilmesini sağlamıştır (Moffitt 2017).

2.1. Twitter Verisi Elde Etme Yöntemleri

Twitter, kullanıcılar tarafından üretilen ve paylaşılan verilerin %1'ini Standart API (Int Kyn. 2) aracılığı ile ücretsiz olarak sunmaktadır. Twitter'dan veri indirmek için birçok uygulama bulunmaktadır ve bu yöntemlerden önde gelenleri incelenmiş ve aşağıda anlatılmıştır.

1- "twitteR" (Gentry 2016), üretilmiş verilerin 10 güne kadar geriye sorgulanmasına olanak veren R programlama dilinde yazılmış bir pakettir. Sorgulamalar "kullanıcı adı", "anahtar kelime", "dil" gibi seçenekler üzerinden yapılabilmektedir. Sorguya yanıt olarak, kullanıcı adı, tweet, zaman bilgisi ve etiketlenmişse coğrafi konum detayı döndürmektedir.

2- "Tweetcatcher Desktop" (TCD) (Cribbin et al. 2015), bir masaüstü yazılımı olup, Twitter verilerinin elde edilmesi ve görselleştirmesi için daha çok sosyal araştırmalarında kullanılan bir yazılımdır. Bu kullanımın nedeni, programın daha çok metin tabanlı analizlere yönelik çalışmalara izin vermesidir.

3- KNIME (Int Kyn. 3), twitter verisinin elde edilmesi için "Knime Twitter Nodes" paketini içeren bir platformdur. Knime ayrıca "Sosyal Medya Duygu Analizi" ve metin madenciliği için "Doğal Dil İşleme (NLP)" araçları ve kütüphanelerini içermektedir.

4- Carto (Int Kyn. 4), internet tabanlı bir platformdur. Sürükle – bırak analizleri ve mekânsal veri görselleştirme imkanları sunmaktadır. Ancak ücretsiz kullanımı hem bireysel hem de kurumsal kullanıcılar için sınırlıdır.

5- NodeXL (Int Kyn. 5), Twitter verisini indirmek ve işlemek üzerine temel ve profesyonel sürümlerini sunmakta olan bir araçtır. Bu araç metin analizlerinde ve kullanıcılar arası ilişki ağlarının tespiti ve analizi üzerine ağ veri tipini işleme kabiliyetine sahiptir.

6- Geo Tweets Downloader (GTD) (Int Kyn. 6, Gulnerman et al. 2016), bir masaüstü programıdır. Bu program konumsal filtrelemeye izin vermekte ve sadece coğrafi konumlanmış tweetleri toplamaktadır.

Bu uygulamaların yanı sıra twitter ve diğer sosyal medya platformları verilerini elde etme, işleme için birçok yazılım ve yöntem Ryerson Üniversitesi'nin oluşturduğu "Sosyal Medya Veri Yönetimi İçin 50'den Fazla Araç Kiti" (Social Media Data Stewardship 2017) (Int Kyn. 7) başlıklı çalışmadan incelenebilir. Yukarıda bahsi geçen uygulamaların dışında, Twitter verileri belirli kriterler belirtilerek ücretli olarak da edinilebilmektedir. (Int Kyn. 8). Bunun yanında Twitter'da üretilen anlık veri "FireHoseAPI" aracılığı ile Twitter tarafından yine ücretli olarak sunulmaktadır (Int Kyn. 9).

Sosyal medya verileri ile yapılacak çalışmaların farklılıklarına göre veri toplama yöntemi ve detayı da değişmektedir. Planlanan çalışmanın amacına göre en uygun veri elde etme yönteminin seçilmesi gerekmektedir. Örneğin acil durumlara müdahale amaçlı olay tespiti yapılabilmesi için veri anlık olarak elde edilmeli iken, alınan politik bir karar sonrası

sosyal medyadaki tepkiler üzerine duygu analizi için geriye dönük günlük ya da haftalık veri elde etmek gerekmektedir.

2.2. Coğrafi Veri Ayrıştırma Teknikleri ve Çalışmaları

Metin içerisinde konum bildiren kelimelerin tespit edilmesi ve coğrafi etiketlerle ilişkilendirilmesi coğrafi ayrıştırma olarak adlandırılmaktadır. Coğrafi ayrıştırma özellikle sosyal medya uygulamalarında mekansal veri miktarını arttırmak ya da coğrafi konuma sahip verinin metin verisi ile coğrafi tutarlılığının araştırılması için önemlidir.

Coğrafi ayrıştırma metodolojilerinde sözlük tabanlı veya bağlamsal tabanlı olmak üzere iki temel yaklaşım vardır. Sözlük tabanlı uygulamalarda yöntem coğrafya sözlüğü kullanılarak veri üzerinden sorgu yapılmasına dayanmaktadır. Bu yaklaşımda coğrafya sözlüğünün zenginliği ve kalitesi coğrafi ayrıştırma başarısına doğrudan etki etmektedir. Bağlamsal temelli yöntemler ise doğal dil işleme yöntemleri ile sağlanabilmektedir. Bağlamsal yöntemlerde performans, doğal dile göre, dile özgü üretilmiş doğal dil temizleme, işleme vb. paketlerin performansına göre farklılık göstermektedir. Coğrafi veri ayrıştırma üzerine algoritma çalışmaları, masaüstü uygulamaları ve çevrimiçi servislerinden öne çıkanlar aşağıdaki şekilde sıralanabilir.

1- Keller et al. (2009) 4000 anahtar sözcükle metin içerisinde coğrafi ayrıştırma yaparak salgın haritalarının çevrimiçi sayfalardan üretilmesini sağlamışlardır.
2- Gelernter ve Balaji (2013) yazım hatası, kısaltma çevrimi, lehçe tespiti, mecaz anlam tespiti gibi ön işleme adımlarının ardından, konum belirten edatların, yön bildiren kelimelerin, uzaklık belirten birimlerin tespitlerini gözeterek algoritmalarını geliştirmişlerdir.
3- Geoparsepy, anlam bulanıklığı üzerine detaylı çalışmaları da içeren bir masaüstü uygulamasıdır. Bu masaüstü uygulaması, açık veri kaynağı olan OSM verisini coğrafi referanslama amaçlı kullanmak üzere tasarlamıştır (Middleton et al. 2018).
4- Geoparser.io (Int Kyn. 10) bir uygulama programlama arayüzü mantığıyla çalışmakta ve R programlama dili ile API sorgulamalarına izin vermektedir. Bu servis aracılığı ile coğrafi ayrıştırma her bir kullanıcı için aylık 1000 sorgu ücretsiz olarak sunulmaktadır.
5- Yavuz ve Abul (2016), yer belirten dil elemanları (ek (-de, -da), isim (okul, hastane), fiil (gidiyorum, dönüyorum)) ile durum tanımlamaları yaparak konum içeren sosyal medya mesajlarının tespitini yapmayı hedefleyen bir çalışma yürütmüştür.

Yukarıda da bahsedildiği üzere, coğrafi veri ayrıştırma ve coğrafi konumlama yapabilmek için coğrafya sözlüğüne, doğal dil işleme paketlerine ve bunun yanı sıra referanslama yapabilecek servislere ihtiyaç duyulmuştur. Her coğrafi bölge için coğrafya sözlüğü bulunmamakta ve doğal dil işleme algoritmaları dile özgü olduğu için her dil için istenen başarımı sağlayamamaktadır.

Bu nedenle, yukarı sunulan çalışmaların bir çoğu dil bağımlı ön işleme adımları sunduğu için her dilde başarıya ulaşamamaktadır. Bunun yanısıra, Keller et al. (2009) ve Geoparsepy örneklerinde olduğu gibi coğrafya kütüphanelerine dayanan tespit algoritmalarında, coğrafi kütüphanenin zenginliği önemli rol oynar. Yapılan bir çalışmada Geoparsepy yazılımı için önerilen coğrafya sözlüğünün güncellenmemesi nedeniyle, Birleşik Krallık ve ABD ülkeleri sınırları dışındaki kalan veriler için çok düşük oranda coğrafi eşleşme yakaladığı görülmüştür (Gulnerman et al. 2018). Bu çalışmada, coğrafi ayrıştırmadaki coğrafya sözlüğü ve coğrafi konumlama servis ihtiyacının sosyal medya verilerinden karşılanması amaçlanmıştır.

3. Coğrafya Sözlüğü Üretimi Metodolojisi

Bu çalışmada yöntem sırasıyla; coğrafi ayrıştırma, coğrafi etiketin konumunun belirlenmesi olarak iki temel aşamadan oluşmaktadır. Coğrafi ayrıştırma için, coğrafi olarak etiketlenmiş tweetler üzerinden 5 farklı cümle yapısı belirlenmiştir (Çizelge 1). Bu cümle yapıları göz önünde bulundurularak veriden coğrafi etiketler çıkarılmıştır.

Çizelge 1. Tweet içeriklerinde tespit edilen coğrafi etiketleme metin yapıları.

No	Örnek Tweet	Yapı	Coğrafi Etiket (CE)

1	I'm at Ortaköy in Beşiktaş, Istanbul https://t.co/...	**I'm at** [CE] **in**	Ortaköy
2	I'm at KoçSistem https://t.co/...	**I'm at** [CE] **https://**	KoçSistem
3	Endokrin ve Metabolizma (@Bakırköy Sadi Konukçu Hastanesi in Bakırköy, İstanbul) https://t.co/...	**(@**[CE] **in**	Bakırköy Sadi Konukçu Hastanesi
4	Yakmaya gidiyoruz @ Istanbul FSM Koprusu	**@**[CE] **https://**	Istanbul FSM Koprusu
5	At Hyde Park – https://t.co/..	**At** [CE] **–**	Hyde Park

R programlama diliyle her bir metin yapısı için coğrafi etiketleri ayıklayacak bir algoritma yazılmıştır. Bu sayede yapısal olarak tanımlanabilen tweet içeriklerinden coğrafi etiketlerin tespiti sağlanmıştır.

Sosyal medyada paylaşılan coğrafi etiketin konumla tutarlılığını tespit edebilmek ve etiketin tespit edilen konum doğruluğunun arttırılabilmesi için coğrafi etiketlerin mekansal tekrarı araştırılmıştır. Ayrıca veri üretiminde veri tekrarı coğrafi yer etiketinin kabulü için kural olarak tanımlanmasıyla, kişisel yer isimlendirmelerinin (evim, dükkanım, okulum vb.) göz ardı edilmesi sağlanmaya çalışılmıştır.

Sosyal medya verilerinin konum doğruluğu akıllı cihazlarda bulunan GNSS'in konum doğruluğu ile eş olarak kabul edilebilir. Bunun yanısıra paylaşım yaparken doğrudan bir mekanın içinde yer almadan belli bir mesafeden paylaşım yapıldığı gözetilirse konum doğruluğunda sapmaların olması olasıdır. Bu çalışmada, ilgi noktalarının doğruluğu için ilgi noktasının temsil ettiği alanın içerisine düşmesi yeterli olarak kabul edilmiştir.

Bu araştırmanın ilk adımı kent için mekansal bölgelerin oluşturulması, ikincisi ise coğrafi etiketlerin bu mekansal alt bölge çerçevelerinde rastlanma sayısının belirlenmesidir. Mekansal bölgeleme için kent yapısından bağımsız kareler ağı tabanlı bir alt bölgeleme sistemi yerine kentin yol ağı dikkate alınarak Voronoi alanları üretilmiştir. Voronoi alanları kareler ağının sabit büyüklüklerinin aksine kentte bulunan yapılaşma büyüklüklerini göz önünde bulunduran bir alt bölgeleme imkanı sağlamaktadır. Bu sayede, kent içerisindeki farklı büyüklükteki coğrafi elemanların (Kent parkları, Cafe, Spor Salonu) mekansal etiketinin izdüşümlerinin kentsel etki alanları içerisinde tespit edilebilmesi amaçlanmıştır. Mekansal bölgeleme ve coğrafi tekrarlılık metodolojisinin işlem adımları Şekil 1'de detaylı olarak sunulmaktadır.

Şekil 1. Coğrafi etiketlerin coğrafi konumlandırma metodolojisi adımları.

4. Bulgular

Sunulan metodolojinin adımları İstanbul ve Londra şehirleri için toplanan Twitter verilerine uygulanmıştır. Toplanan veriler her iki il için de 2 haftalık bir zaman dilimini kapsamaktadır.

Çizelge 2'de verildiği üzere işlenen verilerden İstanbul için 13.966, Londra' için 3.048 noktayı içeren coğrafya sözlüğü üretilmiştir. Tespit edilen coğrafi etiketlerin paylaşımlarda yer alma oranları kentlerin ilgi noktalarının popülerliği ve/veya ziyaretçi kapasitelerini ortaya çıkarmaktadır.

Çizelge 2. Veri ile ilgili genel bilgiler.

	İstanbul	Londra
Toplam Tweet Sayısı	251.654	96.256
Coğrafi etiket içeren tweet sayısı	158.920	32.998
Kullanıcı sayısı	34.746	15.103
Üretilen coğrafi etiket sayısı	13.966	3.048

Çizelge 3'te en yüksek sayıda paylaşımda yer alan 10 coğrafi etiket İstanbul ve Londra için listelenmiştir. Çizelgede yer alan en popüler etiketlerde kent isimleri ön plana çıkarken, bu konumlar daha çok, fazla sayıda insanın ziyaret ettiği havalimanı, alışveriş merkezi, üniversite, stat, park ve turistik çekim alanları gibi yerleri kapsamaktadır. Dikkat çeken bir başka konu, aynı yeri ifade eden etiketlerin kent isimleri, ülke isimleri gibi ek kelimeler (İstanbul, İstanbul Turkey, İstanbul Türkiye, Hyde Park, Hyde Park London vb.) nedeniyle sözlük içerisinde mükerrer kayda neden olmalarıdır. Bunun yanı sıra

yanlış yazımlar, kısaltmalar, dil encodinglerindeki karakter farklılıkları ve farklı dillerin bir paylaşımda kullanılabilmesi nedenleri ile etiketlerde hatalar ve tekrarlar da görülebilmektedir.

Çizelge 3. Veri ile ilgili genel bilgiler.

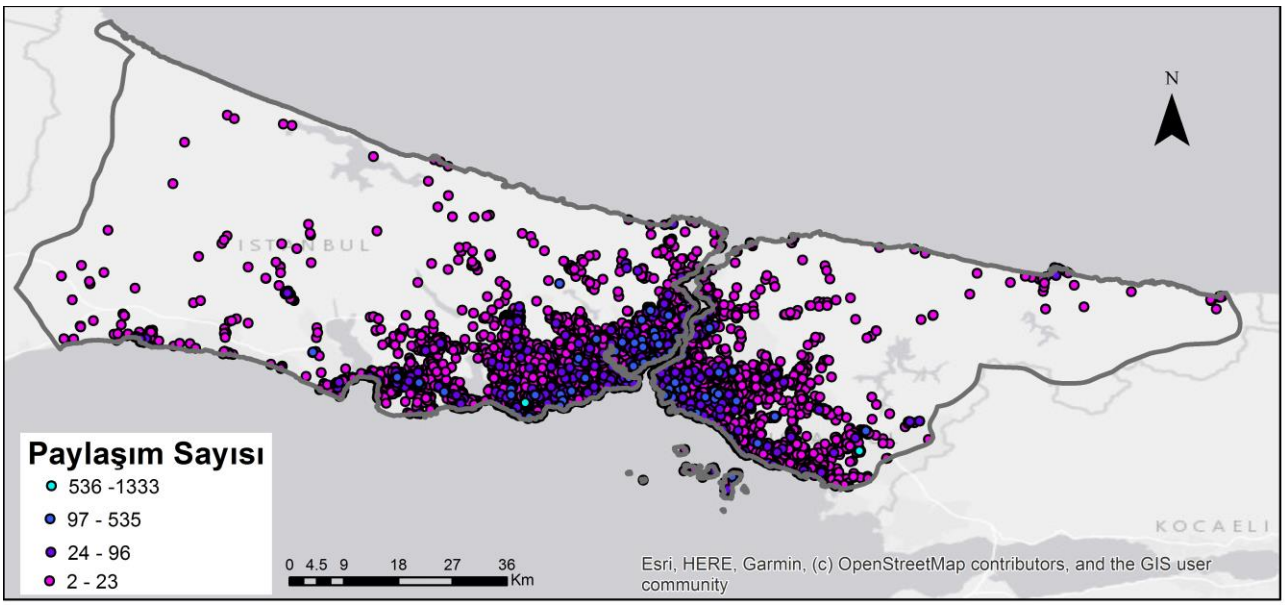
Sıra No	İstanbul		Londra	
	Etiket	Paylaşım Sayısı	Etiket	Paylaşım Sayısı
1	İstanbul, Turkey	1664	Wembley Stadium	422
2	İstanbul Sabiha Gökçen Uluslararası Havalimanı	1333	The O2	369
3	İstanbul Atatürk Havalimanı	771	London United Kingdom	256
4	Akasya Acıbadem	535	MCM London Comic Con	256
5	Bağdat Caddesi	437	Tower Bridge	243
6	Beykent Üniversitesi	437	Victoria Park	218
7	İstiklal Caddesi	412	Shoreditch	213
8	İstanbul Atatürk Havalimanı tavairports	308	London Bridge	201
9	Büyükada	263	We Are Festival	192
10	Cevahir	253	Buckingham Palace	188

Sosyal medya verilerinden üretilen coğrafi ilgi noktaları paylaşım sayılarına göre 4 sınıfla haritalandırılmıştır (Şekil 2, 3). Buna göre her iki il için de kent merkezindeki ilgi noktaları daha çok tweet içeriğinde yer almıştır. Bunun yanı sıra bir yerin ifadesinde özellikle Taksim, Trafalgar gibi meydan adları birden fazla noktada paylaşım içeriklerinde etiketlenmiştir (Şekil 2 (b), Şekil 3 (b)). Benzer konumda aynı yeri ifade eden ancak yazım farklılıkları nedeniyle etiketlerin tekrarlandığı da yine haritalardan görülebilmektedir.

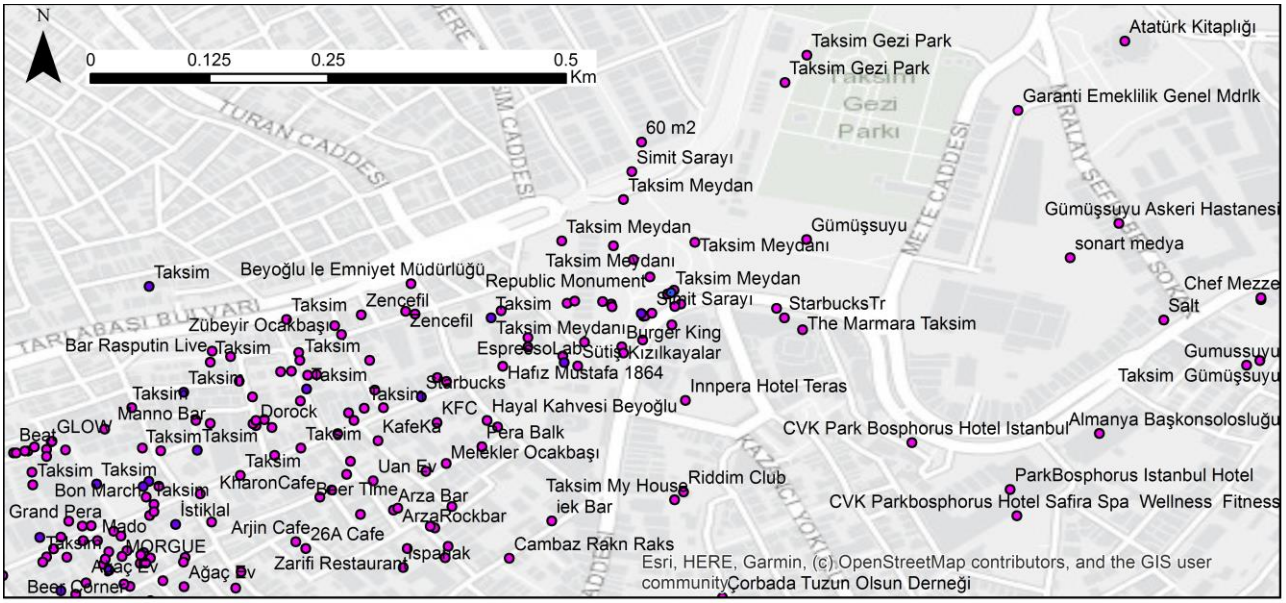
Dikkat edilmesi gereken diğer bir önemli detay ise isim tekrarı gibi görülen ancak zincir dükkan ve restoranlar gibi aynı isme sahip farklı mekanların yakın mesafede yer alabilmeleridir. Bu nedenle benzer isimlerin mekansal bölgeleme sınırları olmaksızın gruplanmasının anlamlı olmadığı gözlemlenmiştir. Bina ölçeğinde alanı temsil eden coğrafi ilgi etiketleri ise genellikle doğru olarak konumlanabilmiş ve tekrara rastlanmamıştır. Özellikle Londra için müze, galeri, elçilik gibi resmi binaları kapsayan etiketlemelerde metodolojinin başarılı sonuçlar verdiği görülmektedir. Bu çıktılar göz önünde bulundurulduğunda kent isimlerinin birçok noktada tekil ya da bir coğrafi etikete ek olarak işaretlendiği mekânsal olarak da tekrar görülmektedir. Bu durum veride gerekli olmayan bir detaya sebep olmakta, bu nedenle kent isimlerinin büyük ve küçük harf duyarlılığı da göz önünde bulundurularak tüm etiketler içerisinden silinmesi bir yol olarak görülebilmektedir. Kent ve ülke isimlerinin tüm etiketlerden silinmesi veriyi temizlemek adına pratik bir yöntem olarak görülebilirken, bu isimleri mekan adı içerisinde kullanan dükkan, restoran ya da resmi binalar için veride bozulmaya neden olabileceği de dikkate alınmalıdır. Bu nedenle bu tip bir temizleme için daha detaylı bir kural tanımına ihtiyaç vardır.

Aynı mekansal bölge içerisinde yer alan ve isimleri tekrarlı olarak görülebilen Taksim, Taksim Meydanı, Taksim Meydan, Taksim Square gibi kayıtlar için metin benzerliği algoritmaları kullanılarak benzerlik düzeylerine göre veri eşleştirme, belirli kurallar çerçevesinde sağlanabilir. Yine bu öneri sonucunda yer isimleri benzer ancak farklı yerleri ifade eden verilerin kaybolması durumu ortaya çıkabilir. Bu duruma Londra'da aynı mekansal bölgede bir arada yer alan Bricklane (yerleşim yerinin adı), Bricklane Marketi, Bricklane Durağı örnek olarak verilebilir. Bricklane örneğinde metin benzerliğinin yüksek olarak çıkacak olmasına rağmen, verilerin eşlenmemesi gerekmektedir.

Büyük alanlar için etiket tekrarından ve alanların birçok mekansal bölgeyi kapsamasından önceden bahsedilmişti. Bu tekrarlar için nokta ifadesi yerine aynı etiketi taşıyan noktaların çevrelediği alan tabanlı veri üretimi üzerinde durulması coğrafi sözlüğün ifadesi için daha doğru öneriler doğurabilir. Bu çalışmada, metin verisi ile ilgili olan veri ön işlem adımları, verinin yapısallığının bozulmaması için kullanılmamıştır. Ancak bu yapısallık da göz önünde bulundurularak bazı dil karakter işlemleri, büyük küçük harf duyarlılığı ve mimik içeren şekillerin temizliği üretilen veri kalitesini arttırmaktadır.

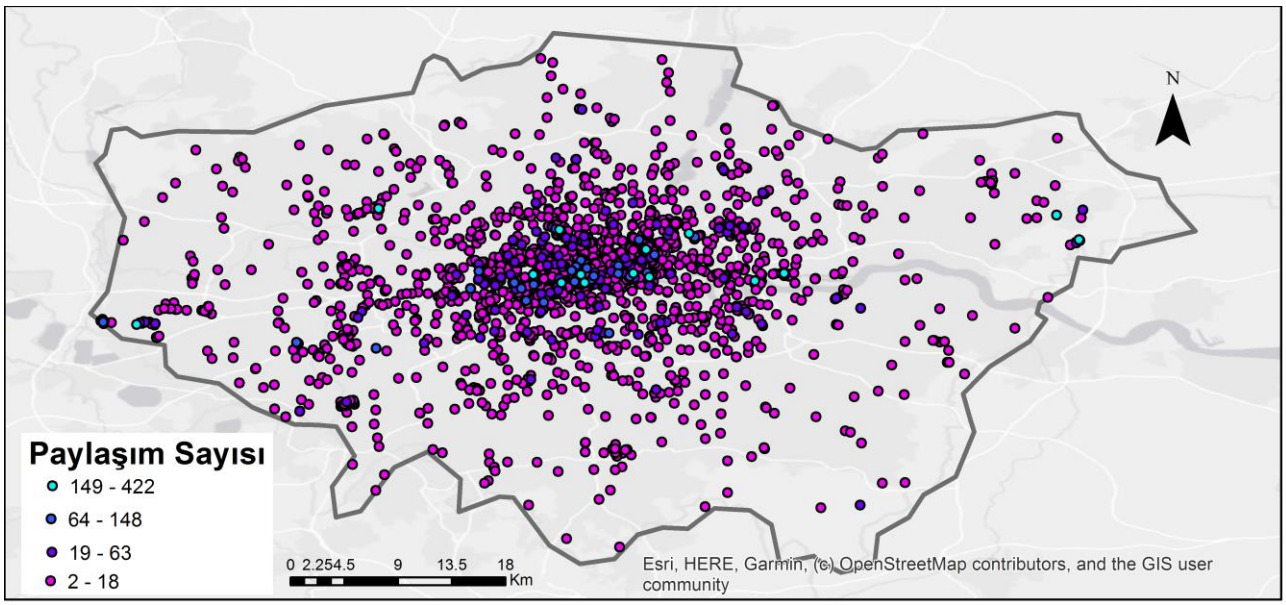


(a)

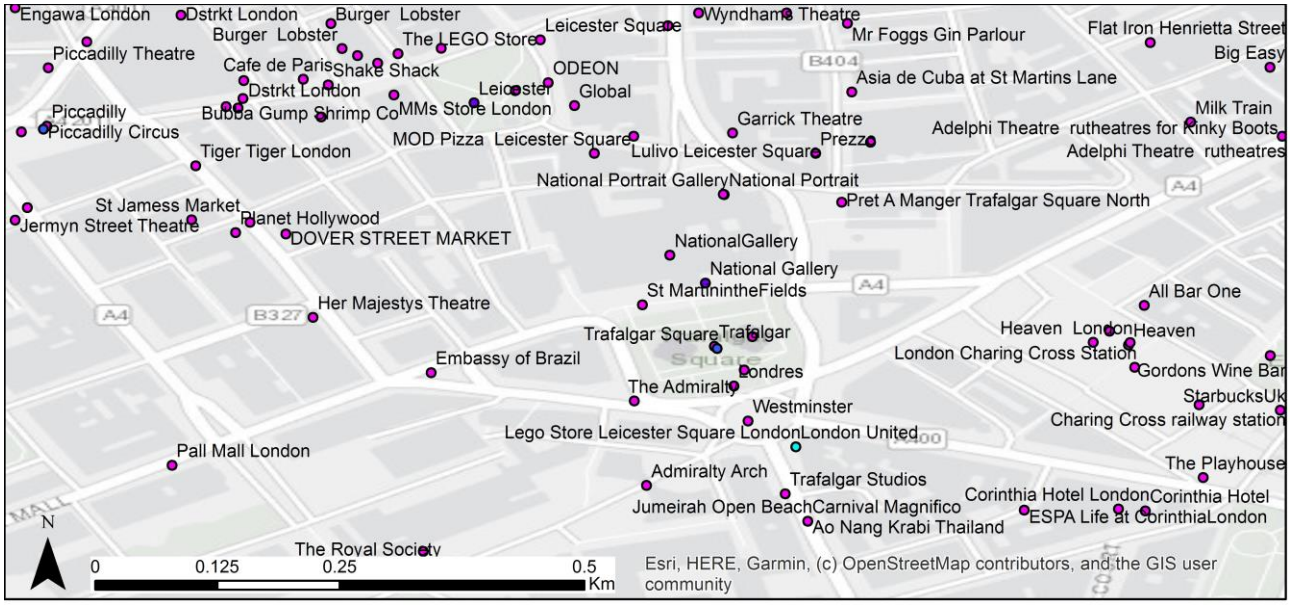


(b)

Şekil 2. İstanbul verisinden üretilen coğrafi ilgi noktaları (a) paylaşım sayısına göre sınıflandırılmış veri (b) Taksim meydanı ve çevresi etiket detayı.



(a)



(b)

Şekil 3. Londra verisinden üretilen coğrafi ilgi noktaları (a) paylaşım sayısına göre sınıflandırılmış veri (b) Trafalgar meydanı ve çevresi etiket detayı.

5. Tartışma ve Sonuç

Bu çalışmada, sosyal medya verilerinin konum tabanlı servisler ile birlikte mekansal veriye katkısı ve bu katkının kapasitesinin nasıl arttırılabileceği üzerinde durulmuştur. Sosyal medya platformları doğrudan ya da dolaylı olarak paylaşılan konum verileriyle geniş alanları kapsayan, sürekli ve düşük maliyetli bir veri kaynağıdır. Bu platformlar, özellikle doğal afetlerin yarattığı gibi acil durumların yönetilebilmesinde anlık veriyi büyük alanlar için hızla sunabilme kabiliyetiyle, acil müdahale için hayat kurtarıcı olabilme kapasitesine sahiptir. Düzensiz olan bu veri ile ilgili bu olumlu özelliklerin kullanılabilmesi için veri, ön işleme adımlarında filtrelenmeli, temizlenmeli, konumlandırılmalı ve ardından konusuna göre sınıflandırılmalıdır.

Sosyal medya verisi geleneksel haritalama yöntemlerinin aksine haritalama alanında uzman olmayan kişiler tarafından üretilebilmekte ve bu üretimde salt haritalama gibi bir amaç bulunmamaktadır. Bu nedenle bu verinin kullanılması yönünde bir başka tartışma da verinin güvenilirliği ve doğruluğu üzerinedir. Verinin sağlaması bu çalışmada olduğu gibi verinin birden fazla kullanıcı tarafından doğruluğunun teyit edilmesi ile öngörülebileceği gibi, hali hazırda

kullanıcı tabanlı güvenilirlik çalışmaları da literatürde yer almaktadır.

Sosyal medya kullanıcılarının konum bilgisi bahsinde ise kullanıcının "profil konumu", "dolaylı konumu" (tweet içeriğinde bahsettiği konum) ve "doğrudan konumu" (enlem, boylam)" ve "gerçek konumu" (paylaşım yaparken bulunduğu konum) gibi birden fazla konum söz konusudur. Profil bilgisi kullanıcın ikamet ettiği yer olarak değerlendirilirken, diğer üç konum bilgisi de her bir paylaşımın konumlandırılması için büyük önem arz edebilir. Bu üç konumun aynı olması verinin en tutarlı olduğu durumdur. Ancak "dolaylı konumun" tahmini için bu çalışmada da bahsedildiği gibi coğrafi ayrıştırma tekniklerinde başarıma, "gerçek konum" için ise kullanıcının paylaşım geçmişine yönelik analizlere ihtiyaç vardır (Gulnerman et al. 2018). Çoğu zaman bu kullanıcının gerçek konumundan emin olunamasa da dolaylı ve doğrudan konumun tutarlı olması, verinin içeriğinin gerçekten bahsi geçen konuyla ilgili olmasının doğrulanması ile veri anlamlı kılınmaktadır.

Bu çalışmada, sosyal medya verilerinin coğrafi kullanım kapasitesinin arttırılması ve doğruluğunun tartışılabilmesi için coğrafi ayrıştırma teknikleri ele alınmış ve coğrafi ayrıştırmada ihtiyaç olan coğrafya sözlüğünün üretimi için sosyal medya verisinin kullanımı üzerine bir yöntem önerilmiştir.

Gelecek çalışmalarda, bu çalışmadaki bulgular gözetilerek metodolojinin özellikle çeşitli metin işleme teknikleri ile geliştirilmesi amaçlanmaktadır. Geliştirilecek yeni yöntemin daha büyük bir veri kümesi ile değerlendirilerek daha kapsamlı bir coğrafya sözlüğü üretilmesi de planlanmaktadır. Üretilecek bu veri, sosyal medya verilerinin mekânsal tutarlılığının araştırılmasına altlık veri olarak kullanılabilecektir.

Bu çalışma, özellikle coğrafya sözlüğü yetersiz bölge ve şehirler için veri üretimine hizmet edecek bir yaklaşım sunmaktadır. Bu yaklaşım, doğal dilden bağımsız bir metin işleme yaklaşımı uygulamaktadır. Dilden bağımsız bu yaklaşım, olumlu yönlerinin yanı sıra doğal dil işleme ile ilgili bazı yardımcı özelliklerin kullanılmaması nedeniyle kısıtlı kalmıştır. Özellikle yazım hatalarının düzeltilmesi, eklerin giderilmesi için uygulanabilecek kök bulma (stemming) algoritmaları verideki doğruluğu arttırabilecektir.

Bu çalışmada da görüldüğü üzere, sosyal medya, mekansal veri üretimine, hızlı, masrafsız ve kapsamlı bir kaynak oluşturma potansiyeline sahiptir. Bu önemli potansiyelin kullanımı coğrafi veri çıkarım yönteminin geliştirilmesi ile iyileştirilebilir. Ayrıca çıkarılan bu coğrafi verinin farklı analizlerle projelendirilmesi ile mekansal verideki olası değişimlerin takibi mümkün olabilir. Bu nedenle ileride yapılacak çalışmalar aşağıdaki adımlarla planlanmaktadır. İlk adım, ön işleme analizine metin bazlı benzerlik indekslerinin kullanımının dahil edilmesidir. Bu sayede olası yazım hatalarından kaynaklı coğrafi tekrarlılığı tespit edilemeyen ilgi noktalarının tespit edilmesidir. İkinci adım ise, bu çalışmada uygulanan metodolojinin daha uzun bir zaman aralığını kapsayan bir veri setinde uygulanmasıdır. Böylece daha çok ilgi noktasının tespiti mümkün olabilecektir. Üçüncü adımda ise, farklı zaman dilimleri için (aylık, sezonluk, yıllık vb.) elde edilen coğrafi etiketlerinin takibi ile veri güncelliği araştırmasıdır. Bu sayede adı değişen mekansal kullanımlar, yeni oluşan ilgi noktaları veya tamamen ortadan kalkan kullanımların tespiti yapılarak veri güncelliği sağlanabilir.

Teşekkür

Bu çalışma İstanbul Teknik Üniversitesi Bilimsel Araştırma Projeleri (BAP) programı kapsamında desteklenmiştir. (Proje Kodu: MDK-2017-40569 40569)

6. Kaynaklar

Cribbin T., Barnett J., Brooker P., Basnayake H., 2015. The Chorus Project Tweet Catcher. TCD 1.3.1. *http://chorusanalytics.co.uk/*.

D'Andrea E., Ducange P., Lazzerini B., Marcelloni F., 2015. Real-time detection of traffic from twitter stream analysis. *Intelligent Transportation Systems*, **16**:2269-2283.

Gelernter J., Balaji S., 2013. An Algorithm for Local Geoparsing of Microtext. *GeoInformatica*, **17(4)**:635-67.

Gentry J., 2016. R-Based Twitter Client. 1.1.9. *https://cran.r-project.org/web/packages/twitteR/twitteR.pdf.*

Gong, Y., Deng, F., & Sinnott, R. O., 2015. Identification of (near) Real-time Traffic Congestion in the Cities of Australia through Twitter. *In Proceedings of the ACM First International Workshop on Understanding the City with Urban Informatics*, 7-12.

Goodchild M.F., 2007. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *International journal of spatial data infrastructures research*, **2(2)**, 24-32

Gulnerman AG, Gengec NE, Karaman H. 2016. Review of Public Tweets Over Turkey Within a Pre-Determined Time. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. **3(1)**, 153-159.

Gulnerman, A. G., Karaman, H., Basiri, A., & Marsh, S., 2018. Credibility Index of the Social media Users to Detect Spatial Outliers. *RGS-IBG Annual International Conference*, Cardiff, UK.

Hasby M, Khodra ML., 2013. Optimal path finding based on traffic information extraction from Twitter. *In International Conference on ICT for Smart Society,* IEEE, 1-5.

Hecht B., Hong L., Suh B., Chi E.H., 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Hecht B., Stephens M., 2014. A tale of cities: Urban biases in volunteered geographic information. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

Ishino A., Odawara S., Nanba H., Takezawa T., 2012. Extracting transportation information and traffic problems from tweets during a disaster. *Proc IMMM,* 91-96.

Keller M., Freifeld C.C., Brownstein J.S., 2009. Automated vocabulary discovery for geo-parsing online epidemic intelligence, *BMC Bioinformatics*, **10 (1)**, 385.

Kosala R., Adi E., 2012. Harvesting real time traffic information from Twitter. *Procedia Engineering*, **50**,1-11.

Lin Y-R., Margolin D. 2014. The ripple of fear, sympathy and solidarity during the Boston bombings. *EPJ Data Science,* **3**,1-28.

McCreadie R., Macdonald C., Ounis I., Osborne M., Petrovic S., 2013. Scalable distributed event detection for twitter. *In 2013 IEEE international conference on big data*, 543-549.

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y., 2018. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems*, **36(4)**, 1-27.

Moffitt J., 2017. Tweet Metadata Timeline. *http://support.gnip.com/articles/tweet-timeline.html.*

Sakaki, T., Okazaki, M., & Matsuo, Y., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. *In Proceedings of the 19th international conference on world wide web,* 851-860.

Schroeder P., 1996. Criteria for the Design of a GIS/2. *Specialists' meeting for NCGIA Initiative 19. GIS and Society, Summer*.

Sieber R., 2006. Public participation geographic information systems: A literature review and framework. *Annals of the Association of American Geographers*, **96**, 491-507.

Signorini A., Segre A.M., Polgreen P.M., 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one*, **6(5)**.

Tufekci Z., Wilson C., 2012. Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square. *Journal of Communication*, **62**, 363-379.

Turner A., 2006. Introduction to neogeography. *O'Reilly*.

Unver A., 2017. What Twitter Can Tell Us about the Jerusalem Protests. *The Washington Post*, **28**.

Yavuz D.D., Abul O., 2016. Implicit Location Sharing Detection in Social Media Turkish Text Messaging. *In International Workshop on Machine Learning, Optimization, and Big Data*, *Springer*, 341-352.

İnternet kaynakları

1-https://www.statista.com/statistics/617136/digital-population-worldwide/, (30.08.2019)
2-https://developer.twitter.com/en/products/products-overview, (15.09.2019)
3-https://www.knime.com/blog/knime-twitter-nodes, (06.10.2018)
4-https://carto.com/connectors/twitter-maps/, (03.06.2017)
5-https://www.smrfoundation.org/nodexl/,(05.09.2019)
6-https://github.com/nagellette/geo-tweet-downloader, (10.05.2017)
7-https://socialmediadata.org/social-media-research-toolkit/, (10.10.2017)
8-https://www.podargos.com/,(10.10.2017)
9-https://support.gnip.com/sources/twitter/, (10.11.2019)
10-https://geoparser.io/, (10.12.2018)