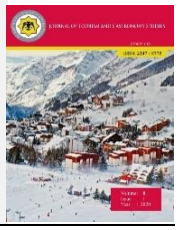
JOURNAL OF TOURISM AND GASTRONOMY STUDIES

ISSN: 2147 – 8775

Journal homepage: www.jotags.org



Metin Madenciliği ve Duygu Analizi Yöntemleri ile Sosyal Medya Verilerinden Rekabetçi Avantaj Elde Etme: Turizm Sektöründe Bir Araştırma (Gaining Competitive Advantage from Social Media Data with Text Mining and Sentiment Analysis Methods: A Research in Tourism Sector)

*** Ahmet BÜYÜKEKE** [a], **Alptekin SÖKMEN** [b], **Cevriye GENCER** [c]

[a] Adana Alparslan Türkeş Science and Technology University, Faculty of Business, Department of Management Information Systems, Adana/Turkey

[b] Hacı Bayram Veli University, Faculty of Economics and Administrative Sciences, Department of Business, Ankara/Turkey

[c] Gazi University, Faculty of Engineering, Department of Industrial Engineering, Ankara/Turkey

Makale Geçmişi

Gönderim Tarihi:08.01.2020

Kabul Tarihi:16.02.2020

Anahtar Kelimeler

Rekabetçi zekâ

Sosyal medya

Duygu analizi

Konu analizi

Öz

Kullanıcıların deneyim, görüş ve tavsiyelerini içeren sosyal medya verileri, seyahate yeni çıkacak olanların kararlarını etkileyen en önemli unsurlardandır. Bu nedenle, Türkiye ekonomisinde vazgeçilmez bir değere sahip olan turizm sektörü için geliştirilecek olan stratejilerde hem politika yapıcıların hem de otel işletmelerinin müşteri yorumlarını dikkate almaları, uygun yöntemlerle analiz etmeleri ve anlamlandırmaları gerekmektedir. Bu bilgiler ışığında gerçekleştirilen çalışmanın temel amacı, büyük sosyal medya verilerinden güncel metin madenciliği yöntemleriyle otel işletmeleri açısından rekabetçi bir zekâ oluşturulmasıdır. Uygulama alanı olarak Antalya bölgesinin seçilmesinin temel sebebi hem temel bir cazibe merkezi olması, hem de Türkiye'nin turizm başkenti olarak kabul edilmesidir. Veriler, Tripadvisor platformundan crawler geliştirilerek otomatik olarak toplanmıştır. Toplam yorum sayısı 212,435'tir. Duygu Analizi için; Lojistik Regresyon, Destek Vektör Makinesi ve Naive Bayes kullanılmıştır. Analiz sonucunda yorumlarının %80'inin olumlu, %20'sinin olumsuz olduğu bulunmuştur. Konu analizi sonucunda; Deneyim %26,70 ile birinci, Değer ve Eğlence %24,68 ile ikinci, Şikâyet %20,41 ile üçüncü sırada yer almaktadır. Diğer konular; %16,15 ile Temel Hizmetler ve %12,06 ile Yapılacak Şeyler'dir.

Keywords

Competitive intelligence

Social media

Sentiment analysis

Topic analysis

Abstract

Social media data, including traveler experiences, opinions, and recommendations, is one of the most important factors affecting the decisions of new travelers. Therefore, both policymakers and hotel businesses in the tourism sector need to consider customer reviews and analyze them appropriate ways in developing strategies. The main purpose of this study is to create a competitive intelligence for hotel businesses by using text mining methods from big social media data. The Antalya region has been selected as the application area. Data were collected automatically by developing a crawler from the Tripadvisor platform. The total number of reviews 212,435. For Sentiment Analysis; Logistic Regression, Support Vector Machine and Naive Bayes were used. As a result of the sentiment analysis, it was found that 80% of reviews were positive and 20% were negative. The subjects that arise as a result of the topics analysis are, respectively; Experience (26.70%), Value and Entertainment (24.68%), Complaints (20.41%), Basic Services (16.15%), Things to Do (12.06%).

Makalenin Türü

Araştırma Makalesi

** Sorumlu Yazar*

E-posta: ahmetbuyukeke@gmail.com (A. Büyükeke)

DOI: 10.21325/jotags.2020.550

GİRİŞ

Günümüzde konuklar ve gezginler yaşadıkları tatil deneyimini, Tripadvisor, Expedia, Yelp ve Booking gibi topluluk temelli sosyal ağlar sayesinde paylaşmaktadırlar (Leung, Law, Hoof, & Buhalis, 2013). Bu durumun doğal sonucu olarak da sosyal ağlarda çok büyük miktarlarda kullanıcı üretimli içerik oluşmaktadır. Bu içerikler ise seyahate yeni çıkacaklar için giderek artan düzeyde önemli bir bilgi kaynağı rolünü almaktadır (Xiang & Gretzel, 2010). Ayrıca gelişen sosyal medya teknolojileri yönetim faaliyetlerini kolaylaştıracak bilgi toplama, yönetme ve paylaşma için de büyük avantajlara sunmaktadır (Gretzel, Sigala, Xiang, & Koo, 2015). Hiç şüphesiz bu büyük miktarlardaki verilerin analizi, işletmeler için strateji geliştirme kapsamında eyleme geçirici önemli bilgiler kazanılmasını ve rekabet avantajı sağlayacaktır.

Sosyal medyanın turizmdeki etkisinin yükselmesi, bu alanda yapılan yeni araştırmaların giderek artmasına vesile olmuştur. Bununla birlikte turizm yazınında sosyal medya araştırmaları henüz başlangıç seviyesindedir (Zeng & Gerritsen, 2014). Özellikle Türkiye'de güncel metin madenciliği yöntemlerine dayalı sosyal medya verilerinden işletmeler için rekabetçi avantaj sağlayacak çalışmalara yazarlar tarafından yapılan araştırmalarda rastlanılmamıştır. Dünyada ise turizm sektöründe sosyal medya verileri kullanılarak özellikle konuk deneyimi ve müşteri memnuniyeti arasındaki ilişkiyi (Xiang, Schwartz, Gerdes, & Uysal, 2015), yorumlardaki duygu yoğunluğunu (olumlu, olumsuz) (Mankad, Han, Goh, & Gavirneni, 2016) ve otel performansı ve yorumlar arasındaki ilişkiyi (Xie, Zhang, & Zhang, 2014) ortaya çıkarmaya yönelik yapılan çalışmalarda son yıllarda hızlı bir artış meydana gelmiştir. Türkiye'de de bu alanda çalışmaların yapılması stratejik önem arz etmektedir. Turizm sektörü Türkiye ekonomisinin en değerli sektörlerinden birisidir. Özellikle döviz girdisi nedeni ile dış açıkların dengelenmesine ve işsizliğin azaltılmasına çok önemli katkı sağlamaktadır (Çımat & Bahar, 2003). Türkiye, gelen turist sayısına göre 2013, 2014 ve 2015'te Dünya genelinde 6. sırada, elde ettiği gelir olarak ise 11. sırada yer almaktayken, 2016'da gelen turist sayısı olarak 10. sıraya gerilemiştir. 2017'de ise tekrar 8. sıraya yükselmiştir (United Nations World Tourism Organization [UNWTO], 2017). Türkiye'de turizm sektöründe sosyal medya verilerini analiz etmek işletmelere müşterilerin karar verme davranışlarını anlamaya ve rekabet çevresindeki fırsatların ve tehditlerin farkına varmalarına imkân verecektir. Bu kapsamda; bu çalışmanın temel amacı, büyük sosyal medya verilerinden güncel metin madenciliği yöntemleri yardımıyla Türkiye'de ilk kez sektör açısından rekabetçi zekâ oluşturulmasıdır. Bu amaç doğrultusunda metin verilerinin makine tarafından işlenebilmesi için gerekli ön işlemler yapılacak, yorumların duygu yoğunluğunu bulmak için belli sınıflandırma algoritmaları kullanılacak ve ayrıca Latent Dirichelt Allocation (LDH) yöntemi ile yorumlarda konuşulan konular ortak başlıklar altında gruplandırılacaktır.

Kavramsal Çerçeve

Hizmetlerin çok daha standart hale geldiği turizm sektöründe rakiplere karşı avantaj sağlamak için müşteri taleplerindeki değişimi rakiplere oranla daha erken anlamak ve tepki oluşturmak önemlidir. Bu çalışmada Antalya il sınırlarında bulunan ve TripAdvisor'da listelenen otellere ait konuk yorumlarına odaklanılmıştır. Türkiye'de Antalya turizmin başkenti olarak kabul edilir (Aksu, Uçar, & Kılıçarslan, 2016). TripAdvisor ise araştırmacılar tarafından turizm alanında sosyal medya veri kaynağı olarak en çok tercih edilen dünyanın en büyük topluluk temelli çevrimiçi yorum platformudur (Xiang, Du, Ma, & Fan, 2017). 7,7 milyon konaklama yeri, havayolu şirketi, deneyim ve restoranı kapsayan geniş yelpazede işletme kaydı vardır ve kullanıcıların deneyimlerine dayalı 661 milyonu aşkın

yorum ve görüşü gezginlere sunmaktadır (tripadvisor.com, 2019). İlerleyen bölümde araştırmada geçen kavramlar kısaca açıklanacaktır.

Rekabetçi Zekâ: Firmaların çevresindeki tehditleri anlama ve fırsatları yakalama becerisidir. Ham bilgiyi ve veriyi eyleme geçirici zekâya dönüştürmek işletme liderleri için en kritik yönetim araçlarından biri olmaktadır. Kahaner (1997, s.16) Rekabetçi Zekâyı; "işletmenizin hedeflerini daha ileriye götürmek için, rakipler ile genel işletme trendleri hakkında sistematik bir şekilde bilgi toplamak ve analiz etmek" olarak açıklamıştır. Fleisher (2004) rekabetçi zekâyı; organizasyonların rakipleri ve rekabet ortamı hakkında harekete geçirici, uygulanabilir bilgi topladığı ve analiz ettiği sistematik bir süreç olarak görür. Ona göre işletmeler, ideal olarak bu süreçten performanslarını iyileştirmek için, karar verme ve planlama aşamalarında faydalanırlar.

Metin Madenciliği: Bilgi alma, makine öğrenmesi, veri madenciliği, istatistik ve hesaplamalı (computational) dilbilim alanlarını kullanan yeni disiplinler arası bir alandır (Gupta & Lehal, 2009). Metin madenciliği yapılandırılmamış metnin toplanması, ön işlemlerden geçirilmesi ve kelimelerin ilişkileri veya desenlerinin keşfi için kümeleme veya sınıflandırma algoritmalarının uygulanması ve en sonunda da görselleştirilmesi süreçlerinden oluşur (Younis, 2015). Özellikle metnin ön işlenmesi aşamasında Doğal Dil İşleme (DDİ) teknikleri kullanılmaktadır.

Duygu Analizi: Temel olarak olumlu ya da olumsuz duyguları ifade eden ya da ima eden görüşlere odaklanır (Liu, 2012). Duygu analizini konu alan birçok makale özellikle yorumların olumlu ya da olumsuz sınıflandırılması uygulamalarına yoğunlaşırken; aslında gerçek şu ki birçok yazarın duygu analizine bu dar çerçevede görev yüklemesi bu duruma sebep olmuştur. Bununla birlikte günümüzde birçok kişi duygu analizini, metinde geçen görüşlerin, duyguların ve nesnelliğin daha geniş anlamda hesaplamalı (computational) işlemleri olduğunu yorumlamaktadır (Pang & Lee, 2008). Duygu analizi çalışmaları iki farklı yöntem ile yapılmaktadır; 1) sözlük tabanlı yöntem, 2) makine öğrenmesi yöntemidir. Bu çalışmada makine öğrenmesi yöntemlerinden olan ve metin sınıflandırma da en çok tercih edilen (Ravi & Ravi, 2015) yöntemlerden 3 tanesi kullanılacaktır; Lojistik Regresyon, Destek Vektör Makinesi ve Naive Bayes.

Konu Analizi: Büyük miktarlardaki sosyal medya verilerinin, istatistiksel olarak benzerliklerinden yararlanılarak özetlenmesi veya gruplar altında listelenmesi, işletmeler için kısa zamanda yorumlar hakkında fikir sahibi olabilmelerine imkân verebilir. Bu amaçla metin verilerini gruplandırmak için standart hale gelen Gizli Dirichlet Dağılımı (Latent Dirichlet Allocatin-LDA) aracı kullanılmaktadır (Hong & Davison, 2010). Bu çalışmada da LDH yöntemi kullanılacaktır. LDH, bag of words (BOW- kelime kutusu) yöntemini kullanır (Blei, Ng, & Jordan, 2003). Dokumandaki bütün yorumlar kelime vektörü olarak temsil edilir. Bu yöntemde her farklı konu belli olasılıklarla bütün yorumlarda vardır. Toplam olasılık 1'e eşittir. En fazla olasılık değerine sahip olan konu, o yorum için baskın konu kabul edilir. Bir yorumda sadece bir konu konuşulduğu düşünülemez. Bütün kelimelerin de her konu için belli bir olasılık değeri vardır. Bu çalışmada yorumlar, baskın konu seçilerek gruplandırılmıştır.

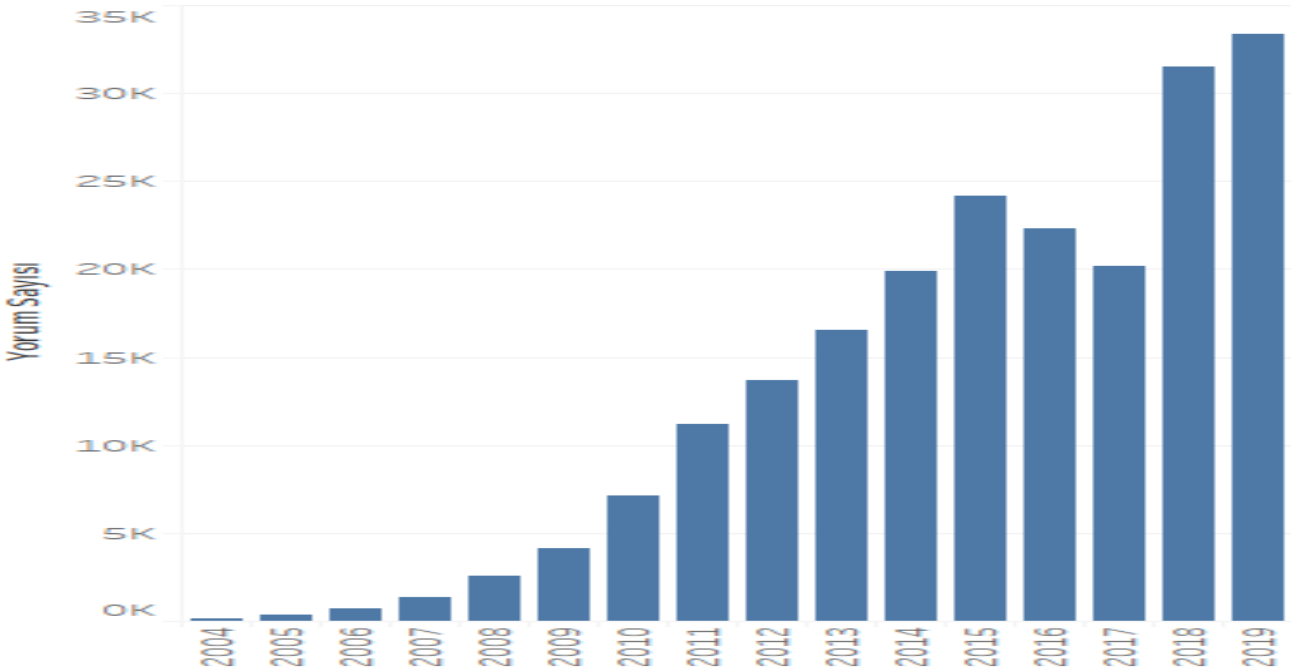
Yöntem

Araştırmada kapsamında Antalya yöresinde hizmet veren konaklama tesislerine ait sosyal medya yorumları Tripadvisor platformu üzerinden toplanmıştır. Yorumların otomatik olarak toplanabilmesi maksadıyla Python programlama dili ile bir crawler geliştirilmiştir. Geliştirilen crawler için herhangi bir zaman sınırlaması yapılmamıştır. Crawler çalıştırıldığı anda Antalya ilinde bulunan ve TripAdvisor'da işletme kaydı olan konaklama tesislerine ait tüm İngilizce yorumları toplamaktadır. Yaklaşık olarak 1 ay kadar çalıştırılmış crawler ile Aralık

2019'a kadar olan yorumlar toplanmıştır. Yorumların duygu yoğunluğunu bulmak için makine öğrenmesi yöntemlerinden SVM, LR ve NB yöntemleri kullanılmıştır. Konu Analizi için ise LDH yöntemi kullanılmıştır. Ayrıca yorumlardan anahtar kelimeler bulmak amacıyla kelime bulutları oluşturulmuştur.

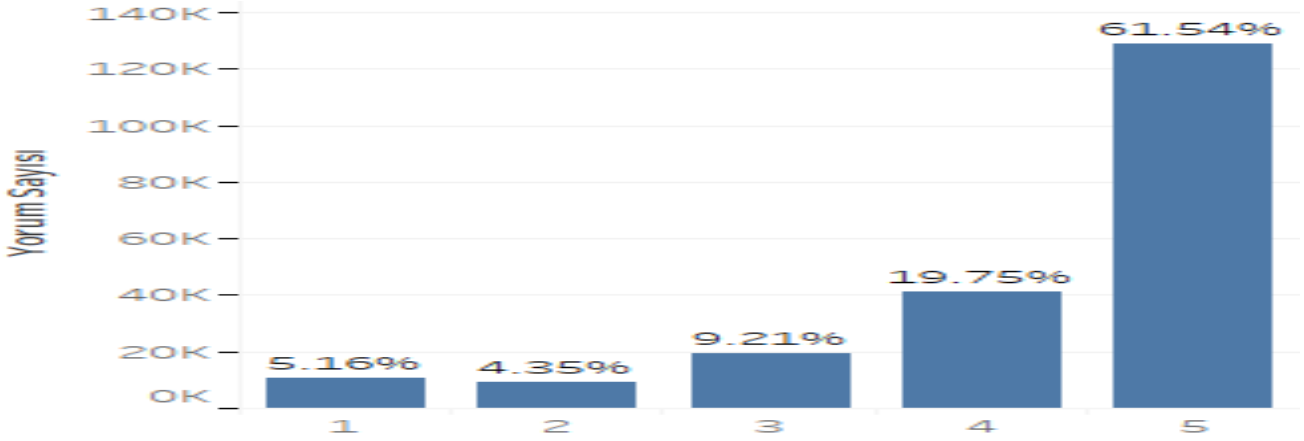
Veriler

Toplam yorum sayısı 212,435'tir. Toplam tesis sayısı 1,801'dir. Bazı yorumlarda İngilizcenin yanında farklı bir dilin de kullanıldığı gözlemlenmiştir. Bu durumda olan toplam 537 yorum langdetect isimli Python kütüphanesi ile otomatik olarak tespit edilmiş ve analizden çıkarılmıştır. Ayrıca bir işletmeye ait yorum sayısı 10'dan az ise çalışmanın güvenirliğini artırmak amacıyla bu işletmelere ait yorumlar çalışmadan çıkartılmış ve toplam yorum sayısı 209,171'e, toplam işletme sayısı ise 1072'ye düşmüştür. İşletme sayısında 729 gibi büyük miktarda azalma varken, çıkartılan işletmelerin ortalama yorum sayısı 4'tür. Kalan işletmelerin ortalama yorum sayısı 195'tir. Aşağıda şekil 1'de yıllara göre yorum sayısı dağılımı görülmektedir. Yorumlar otel ismi, kullanıcı ülkesi (belirtilmişse), yorum başlığı, yorum metni, yorum tarihi ve yorum rating değerini barındırmaktadır.



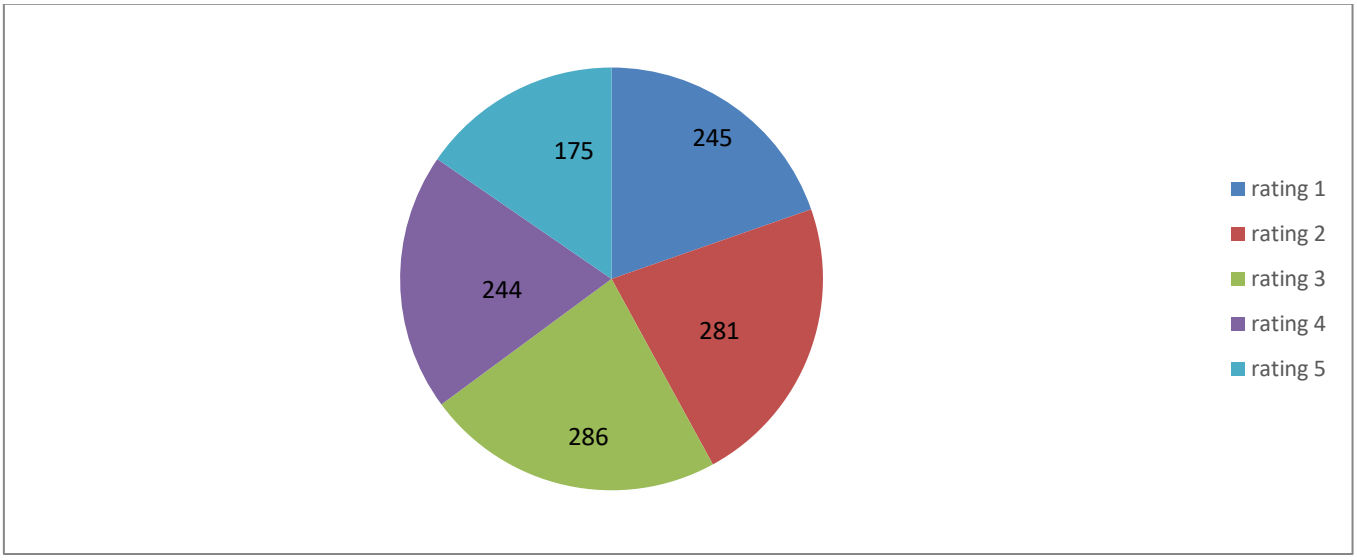
Şekil 1. Yıllara Göre Yorum Sayısı

Alt tarafta bulunan Şekil 2'de yorumların rating değerlerine göre dağılımı gösterilmektedir. Buna göre Antalya yöresi için müşteri yorumları genel olarak memnuniyet ifade etmektedir. Yorumların genel dağılımında, rating değeri 4 ve 5 olan yorumlar yaklaşık %81'dir. Bununla birlikte rating 3 değerine sahip yorumlar yaklaşık %9, rating 1 ve 2 değerine sahip olan yorumlar ise yaklaşık %10'dur. Analiz ve Bulgular bölümünde, makine öğrenmesi yöntemleri kullanılarak sınıflandırılmış tüm yorumların duygu yoğunluğu gösterilmiştir.



Şekil 2. Rating Değerlerine Göre Yorumların Dağılımı

Yorumların ortalama uzunluğu 206 kelimedir. En uzun yorum 5,383, en kısa yorum ise 4 kelime/emojiden oluşmaktadır. Şekil 3'de rating değerlerine göre yorumların ortalama uzunluğu gösterilmektedir. Şekilden de anlaşılacağı gibi memnuniyet belirten yorumlar daha kısa, memnuniyetsizlik içeren yorumlar daha uzundur. Özellikle en uzun yorumlar rating değeri 3 olan yorumlardır. Bu yorumlarda hem olumlu ifadelerin hem de olumsuz ifadelerin mevcut olabileceği düşünülmektedir.



Şekil 3. Rating Değerlerine Göre Yorumların Ortalama Kelime Sayısı

Antalya yöresinde bulunan turizm bölgeleri, TripAdvisor'da 8 ana bölgeye ayrılmıştır. Bu bölgelere göre toplam yorum sayıları ve genel rating değeri ortalamaları aşağıda tablo 1'de gösterilmektedir.

Tablo 1. Antalya Yöresi Turizm Bölgeleri TripAdvisor Yorum Durumları

Bölge Adı	**Yorum Sayısı**	**Rating Ortalaması**
Antalya	56,224	4,29
Belek	50,134	4,55
Side	32,231	4,13
Alanya	23,744	3,85
Kemer	23,291	4,27
Manavgat	10,430	4,14
Kalkan	8,411	4,53
Kaş	4,706	4,47

Verilerin ön işlenmesi

Yorumların makine tarafından işlenebilmesi için sayısal hale dönüştürülmesi gerekmektedir. Bu işlem için yorumlardaki bütün benzersiz kelimeler bulunarak her birine bir sayısal ID verilir ve her bir yorum kendisinde geçen kelimenin frekans değeri ile matriste temsil edilir. Kelimelerin frekans değerleri için tf-idf hesaplaması kullanılmıştır. Herhangi bir terimin matriste bulunabilmesi için en az 5 yorumda geçmesi, en fazla ise yorumların %70'inde görülmesi sınırlaması uygulanmıştır. Bu değerler, Scikit-learn kütüphanesinde bulunan GridSearchCV modülüne gönderilen belli parametreler ile hesaplanmıştır. Daha sonra Scikit-learn kütüphanesinde bulunan CountVectorizer modülü kullanılarak DTM oluşturulmuştur. Matris boyutunun çok fazla olmaması için öncelikle yorumlardaki bütün noktalama işaretleri ve sayısal karakterler çıkartılmış ve kalan karakterler küçük harfe çevrilmiştir. Ardından stop-words uygulaması ile cümle anlamını çok fazla etkilemeyen kelimeler yorumlardan çıkartılmış ve aynı köke sahip (örneğin; use, useful gibi) kelimeler kök alma yöntemi ile (stemming) köklerine indirilmiştir. Tüm bu önişlemler için Python programlama dili ile geliştirilen (NLTK, Scikit-learn, re) kütüphaneler kullanılmıştır. Önişlemlerden sonra oluşturulan DTM'nin toplam boyutu 9,809'dur.

Analiz ve Bulgular

Duygu Analizi

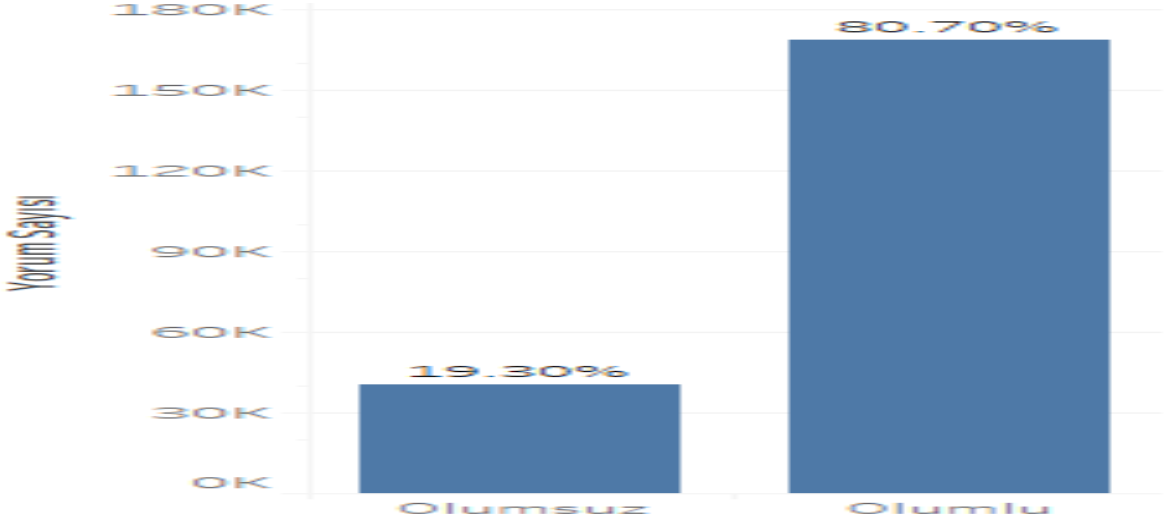
Toplanan yorumların duygu yoğunluğunu bulmak amacıyla metin madenciliğinde en çok tercih edilen sınıflandırıcı algoritmalardan; LR, DVM ve NB kullanılmıştır (Ravi & Ravi, 2015). Bu sayede hem bu algoritmaların performansları karşılaştırılarak konaklama işletmelerine ait konuk yorumlarını en iyi sınıflandıran algoritmanın tespiti yapılmış olacak, hem de Antalya yöresine ait turizm bölgelerinin duygu yoğunluğu ortaya çıkartılmış olacaktır. Yorumlardan sınıflandırıcı model oluşturmak maksadıyla dengeli bir eğitim ve test seti oluşturulmuştur. Yorumların rating değerlerine göre; 1 ve 2 olanları olumsuz (0), 4 ve 5 olanları da olumlu (1) kabul edilerek etiketleme yapılmıştır. Setin toplam büyüklüğü 39,800'dür. Yarısı rating 1 ve 2 değerine sahip yorumlardan, diğer yarısı da rating 4 ve 5 değerine sahip yorumlardandır. Setin %70'i eğitim seti olarak, kalan kısmı da test seti olarak kullanılmıştır. Kurulan modellerin başarısı değerlendirmek için Accuracy, Recall, Precision ve F1-skor değerleri kullanılmıştır. Modellerin test seti ile yapılan sınıflandırma başarıları tablo 2 görülmektedir.

Tablo 2. Sınıflandırma Algoritmaları Değerlendirilmesi

	Accuracy	**Recall**	**Precision**	**F1**
LR	0,95	0,95	0,96	0,95
SVM	0,95	0,95	0,95	0,95
NB	0,94	0,93	0,95	0,94

Not: Değerler yüzdelik ifade etmektedir.

Üç algoritmanın da sınıflandırma başarısı birbirine yakındır. Ayrıca sınıflandırma başarı oranları da oldukça yüksek gerçekleşmiştir. Tablo 2'deki değerlere göre yorumların sınıflandırılması için LR ve SVM tercih edilebilir. Bu çalışmada sınıflandırma için LR seçilmiştir. LR ile rating değeri 3 olan yorumlar da dahil olmak üzere tüm yorumların sınıflandırılması yapılmıştır. Aşağıda Şekil 4'te yorumların duygu yoğunluğu görülmektedir.



Şekil 4. Antalya Yöresi İçin Yorumların Duygu Dağılımı

Üst tarafta Şekil 4'te görülen, Antalya yöresine ait yorumların genel duygu yoğunluğu olmakla beraber, bu durum alt bölgelere göre değişmektedir. Bazı bölgelerde olumlu yorum oranı artarken bir kısım bölgelerde düştüğü gözlenmektedir. Aşağıda tablo 3'te bölgelerin duygu yoğunluğu yüzdelik olarak gösterilmektedir.

Tablo 3. Yorumların Bölgelere Göre Duygu yoğunluğu.

Bölge Adı	**Olumlu**	**Olumsuz**
Kalkan	**0,90**	0,10
Kaş	**0,89**	0,11
Belek	0,88	0,12
Antalya	0,81	0,19
Side	0,77	0,23
Kemer	0,80	0,20
Manavgat	0,75	0,25
Alanya	**0,68**	0,32

Tablo 3'ten de görüldüğü gibi en fazla olumlu yorum oranına Kalkan ve Kaş bölgesi sahiptir. Belek bölgesi de %88'in üzerinde olumlu yoruma sahiptir. Diğer taraftan Alanya ise %68 ile en az olumlu yorum oranına sahip olan bölgedir. Yorumların rating değerlerine göre duygu yoğunluğu tablo 4'te yüzdelik olarak gösterilmiştir.

Tablo 4. Rating Değerlerine Göre Antalya Yöresi Duygu yoğunluğu

Rating Değeri	**Olumlu**	**Olumsuz**
1	0.01	0.99
2	0.06	0.94
3	0.36	0.64
4	0.85	0.15
5	0.98	0.02

Makine tarafından, rating değeri 1 ve 2 olan yorumların %96,37'si olumsuz olarak sınıflandırılmıştır. Rating değeri 5 olan yorumların neredeyse tamamı olumlu iken, rating değeri 4 olan yorumların %15'inin olumsuz sınıflandırıldığı gözlemlenmiştir. Genel olarak rating değerleri ile sınıflandırma arasında yüksek bir korelasyon vardır. Rating değeri 3 olan yorumlar bölgenin geneli için olumsuz olarak değerlendirilebilir, fakat Kalkan ve Kaş bölgesinde rating değeri 3 olan yorumlar daha çok olumlu olarak sınıflandırılmıştır. Aşağıda Tablo 5'te bu durum görülmektedir.

Tablo 5. Rating değeri 3 olan yorumların bölgelere göre duygu yoğunluğu

Bölge Adı	**Olumlu**	**Olumsuz**
Kalkan	**52**	48
Kaş	**54**	46
Belek	38	62
Antalya	36	64
Side	36	64
Alanya	34	66
Kemer	32	**68**
Manavgat	32	**68**

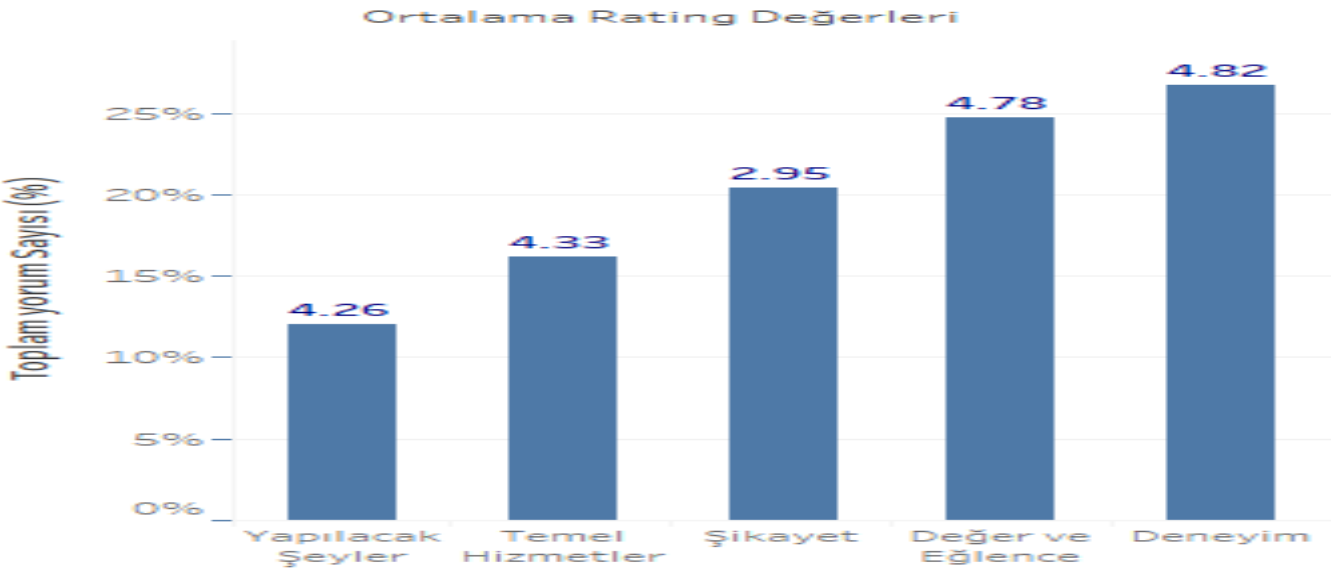
Konu Analizi

Yorumlarda geçen konuların çok fazla değişiklik göstermeyeceği düşünülebilir. Müşterilerin özellikle otel hizmetlerine dönük olarak memnuniyet ifadeleri veya eleştirileri, gezilecek yerler hakkında görüşleri ile bazı karşılaşılan özel durumların anlatılması beklenebilir. Tabi ki tüm yorumları tek tek incelemek mümkün değildir. Bu sebeple yorumların benzerliklerine göre gruplandırılması önem arz etmektedir. Bunun için LDH yöntemi kullanılmıştır. Bu yöntemde konu sayısı, analizi yapan kişi tarafından belirlenmektedir. En uygun konu sayısını belirlemek amacıyla 3'ten 10'a kadar rakamlar konu sayısını belirlemek için denenmiş, en küçük karmaşıklık (perplexity) değerine sahip olan 5 konu sayısı en uygun değer olarak bulunmuştur. Xiang vd. (2017) ve Mankad vd. (2016) tarafından yapılan çalışmalarda da konaklama tesislerine ait yorumların en uygun 5 farklı konu altında gruplandığı gösterilmiştir. Tablo 6'da konularda geçen en yüksek frekans değerine sahip ilk 30 kelime gösterilmektedir.

Tablo 6. Konularda En Sık Geçen Kelimeler

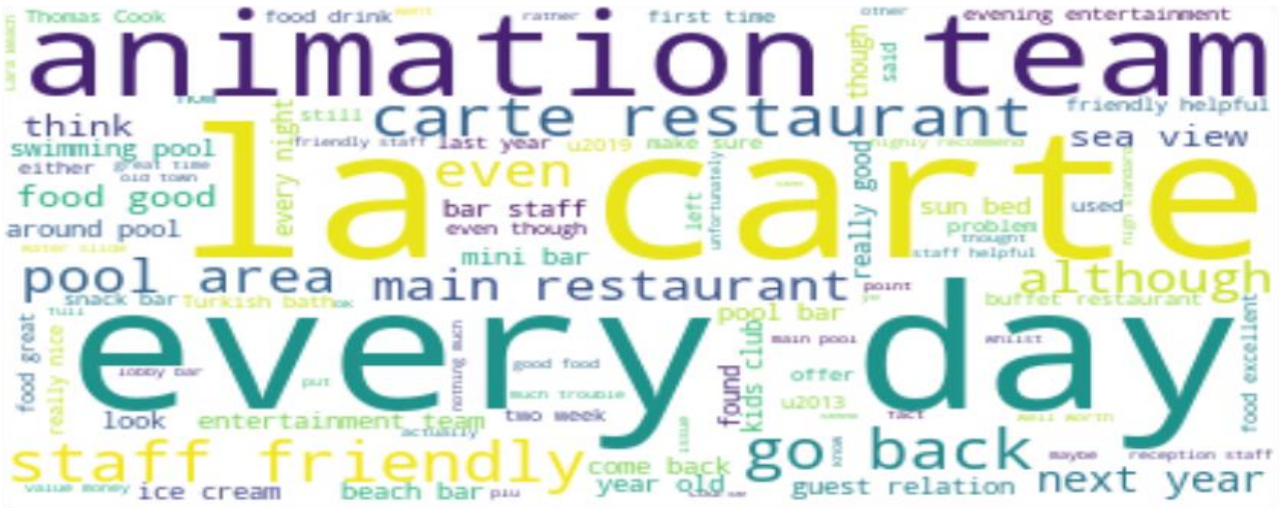
	Konu 1	**Konu 2**	**Konu 3**	**Konu 4**	**Konu 5**
1	say	great	lovely	pool	room
2	day	food	food	bar	walk
3	go	really	friendly	drink	area
4	people	family	holiday	restaurant	small
5	get	time	great	water	price
6	ask	kid	restaurant	day	beach
7	reception	team	excellent	area	place
8	guest	clean	return	bed	view
9	pay	amazing	visit	main	shop
10	time	good	clean	food	breakfast
11	speak	entertainment	recommend	night	minute
12	tell	pool	week	evening	bus
13	bad	love	helpful	eat	restaurant
14	book	year	stay	serve	old
15	leave	place	fantastic	towel	town
16	food	child	service	beach	location
17	night	day	definitely	plenty	quality
18	review	big	look	clean	close
19	look	animation	beautiful	available	hotel
20	give	activity	feel	free	resort
21	take	thank	bar	lunch	offer
22	thing	room	amazing	table	service
23	check	especially	welcome	lot	sea
24	arrive	perfect	wonderful	bit	turkish
25	star	people	go	fresh	trip
26	problem	night	book	hot	quite
27	come	holiday	choice	snack	local
28	bed	enjoy	brilliant	turkish	taxi
29	hour	friendly	year	choice	wifi
30	english	work	room	sunbed	away

Konulara ait kelimeler incelendiğinde bu konulara karşılık gelecek uygun başlıklar sırasıyla; Şikâyet, Değer ve Eğlence, Deneyim, Temel hizmetler, Yapılacak şeyler şeklinde verilmiştir. Yorumlarda geçen konular aşağıda şekil 5'te grafik olarak gösterilmektedir. Bölgenin genelinde en fazla konuşulan konu yaklaşık %27 ile Deneyim olurken, ikinci sırada Değer ve Eğlence konusu yer almaktadır. Bölgeler bazında incelendiğinde, Kemer ve Belek'te Değer ve Eğlence konusu; Antalya, Side ve Kalkan'da Deneyim konusu en fazla konuşulmuşken; Alanya ve Manavgat da ise Şikâyet konusu yaklaşık %29 ile en çok konuşulan konu olmuştur. Kaş'ta ise Yapılacak şeyler yaklaşık %49 ile birinci sıradadır. Bunun sebebi; buralardaki konaklama tesislerinin 4 ve 5 yıldızlı otellerden ziyade pansiyon, butik veya apart tarzı oteller olması ve konukların tesislerde vakit geçirmek yerine gezmeye daha fazla vakit ayırması olarak düşünülebilir.



Şekil 5. Yorumların Konularına Göre Dağılımı

LDH yöntemi ile konu analizinin yanısıra, ayrıca yorumlar hakkında daha hızlı fikir sahibi olabilmek için kelime bulutları oluşturmak işletme yöneticilerine fayda sağlayacaktır. Kelime bulutunda göze çarpan kelimeler veya kelime grupları, oluşturulacak bir filtreleme sisteminde anahtar kelime olarak kullanılabilir. Bu şekilde, bölge bazında veya genel olarak hizmet isimleri ya da olumlu veya olumsuz duyguları ifade eden sıfatlar, zarflar, fiiller, isimler bir filtre amacıyla kullanılabilir. Aşağıda şekil 6'da tüm yorumlara ait kelime bulutu gözükmektedir.



Şekil 6. Yorumlara Ait Kelime Bulutu

Sonuç ve Öneriler

Bu çalışma ile Türkiye'de ilk defa turizm sektöründe makine öğrenmesi kapsamında güncel metin madenciliği yöntemleri kullanılarak sosyal medya verilerinden sektöre rekabet avantajı sağlayacak analizler gerçekleştirilmiştir. Elde edilen bulgular stratejik yönetim sürecinin çevre analizi adımında en önemli veri olan müşteri girdileri açısından eşsiz bir fırsat ve kaynak oluşturabilecektir. Analizler sonucunda Antalya yöresinde faaliyet gösteren konaklama işletmelerine ait yorumların duygu yoğunluğu ortaya çıkartılmıştır. Yorumların %80'inin olumlu olmasına rağmen, %20 düzeyinde olumsuz yorum da mevcuttur. Bu durum bölgelere göre de farlılıklar göstermektedir. Örneğin Kaş ve Kalkan için yorumların duygu ortalaması %90 iken, özelikle en düşük duygu ortalamasına sahip Alanya'da bu değer %66'dır. Ayrıca makine öğrenmesi yöntemlerine göre otomatik sınıflandırılan yorumlar rating değerlerine göre incelendiğinde; rating değeri 4 ve 5 olanların yaklaşık %94'ü, rating değeri 3 olanların yaklaşık %37'si, rating değeri 1 ve 2 olanların ise yaklaşık %4'ü olumludur. Yorumların rating değerleri tek başına yorumu değerlendirmek için yeterli olmayabilir. Örneğin rating değeri 5 olan bir yorum, makine tarafından olumsuz sınıflandırıldıysa ve konu analizi sonucu baskın konuşulan konu Şikâyet konusu çıktıysa, o yorumda memnuniyetsizlik belirten ifadelerin olduğu görülmüştür. Yorumlar uzunluklarına göre incelendiğinde rating değeri 3 olanların en fazla ortalama kelime sayısına sahip olduğu görülmüştür. Bu durum, rating değeri 3 olan yorumlarda memnuniyeti belirten ifadelerle birlikte memnuniyetsizliği belirten ifadelerin de (veya tersi) olduğu anlamına gelebilir. Bu açıdan otomatik duygu sınıflandırması, bu yorumlar hakkında doğru değerlendirme yapabilmek için önemli görülmektedir. Yapılacak pazar araştırmaları kapsamında bölgelerin kendi içinde değerlendirilmesi ve otellerin türlerinin ve sınıflarının göz önünde bulundurulması önemlidir. Örneğin yorumlar konularına göre makine tarafından otomatik gruplandırıldığında Belek, Side, Kemer gibi 5 yıldızlı otellerin daha fazla olduğu bölgelerde Değer ve Eğlence ile Deneyim konuları en fazla konuşulmuşken, Kaş ve Kalkan gibi ağırlıkla butik otellerin veya pansiyonların olduğu bölgelerde Yapılacak Şeyler konusu en fazla konuşulmuştur. Buradan, 5 yıldızlı otellerde müşterilerin daha çok otel içindeki eğlenceye dönük aktivitelere yoğunlaştıkları ve temel otelcilik hizmetlerinden bahsettikleri anlaşılabilir. Diğer taraftan Kaş ve Kalkan bölgesi yorumları için oluşturulan kelime bulutlarında terrace, balcony ve breakfast kelimelerinin birlikte ve sık geçmesinden bu bölgelerde konuklar tarafından balkonda sabah kahvaltısının beğenildiği düşünülebilir. Sosyal medya yorumları, anket ve söyleşi gibi yöntemlerle karşılaştırıldığında hem çok daha büyük miktardadır, hem de çok daha uygun maliyetlerle toplanabilmektedir. Ayrıca müşterilerin görüş ve düşüncelerini kendi istekleri ile paylaşmasının getirdiği bir objektiflik de bulunmaktadır. Turizm sektöründe sosyal medya verilerine dönük olarak uygulanan bu analizler sonucunda işletmelerin geliştirilecekleri stratejilerde, rekabet ortamına dair ihtiyaçları olan önemli bilgileri edinebilecekleri düşünülmektedir. Ayrıca gerçekleştirilecek pazar araştırmalarında bölgeye ait fırsatların veya tehditlerin farkına varılabileceği de düşünülebilir. Bu kapsamda elde edilen bulgular, bakanlık, sektör, bölge ve işletme düzeyinde rahatlıkla kullanılabilir ve eşsizdir. Gelecek çalışmalarda araştırmacıların konaklama tesislerinin türlerine ve sınıflarına göre metin madenciliği yapmalarının, rekabet analizi açısından daha etkin sonuçlar çıkaracağı düşünülmektedir.

KAYNAKÇA

Aksu, A., Uçar, Ö., & Kılıçarslan, D. (2016). Golf Tourism: A Research Profile and Security Perceptions in Belek, Antalya, Turkey. *International Journal of Business and Social Research, 6*(12), 1-12.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Çımat, A., & Bahar, O. (2003). Turizm sektörünün Türkiye ekonomisi içindeki yeri ve önemi üzerine bir değerlendirme. *Akdeniz İ.İ.B.F Dergisi, 6*, 1-18.

Fleisher, C. S. (2004). Competitive Intelligence Education: Competencies, Sources, and Trends. *Information Management Journal*, 56-62.

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: foundations and developments. *Electron Markets, 25*(3), 179-188. doi:10.1007/s12525-015-0196-8

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. *Proceedings of the first workshop on social media analytics* (s. 80-88). Washington DC: ACM.

Kahaner, L. (1997). *Competitive intelligence: how to gather analyze and use information to move your business to the top.* New York: Simon and Scuster.

Leung, D., Law, R., Hoof, H. v., & Buhalis, D. (2013). Social Medıa In Tourısm And Hospıtalıty: A Lıterature Revıew. *Journal of Travel & Tourism Marketing, 30*(1-2), 3-22.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers.

Mankad, S., Han, H. "., Goh, J., & Gavirneni, S. (2016). Understanding Online Hotel Reviews Through Automated Text Analysis. *Service Science, 8*(2), 124-136.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*, 1-135.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14-46. doi:10.1016/j.knosys.2015.06.015

World Tourism Organization. (2018). *UNWTO Tourism Highlights.* Madrid: UNWTO. doi:https://doi.org/10.18111/9789284419876

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management, 31*, 179-188.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management, 58*, 51-65. doi:10.1016/j.tourman.2016.10.001

Xiang, Z., Schwartz, Z., Jr., J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120-130.

Xie, K. L., Zhang, Z., & Zhang, Z. (2014). The business value of online consumer reviews and management response to hotel performance. *International Journal of Hospitality Management, 43*, 1-12. doi:10.1016/j.ijhm.2014.07.007

Younis, E. M. (2015). Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study. *International Journal of Computer Applications, 112*(5).

Zeng, B., & Gerritsen, R. (2014). What do we know about social media in tourism? A review. *Tourism Management Perspectives*, 27–36.

Gaining Competitive Advantage from Social Media Data with Text Mining and Sentiment Analysis Methods: A Research in Tourism Sector

Ahmet BÜYÜKEKE

Adana Alparslan Türkeş Science and Technology University, Faculty of Business, Adana/Turkey

Alptekin SÖKMEN

Hacı Bayram Veli University, Faculty of Economics and Administrative Sciences, Ankara/Turkey

Cevriye GENCER

Gazi University, Faculty of Engineering, Ankara/Turkey

Extensive Summary

Today, guests and travelers share their holiday experiences through community-based social networks such as Tripadvisor, Expedia, Yelp, Booking (Leung, Law, Hoof, ve Buhalis, 2013). This results in huge amounts of user-generated content in social networks. These contents play an increasingly important source of information for travelers prior to travel (Xiang ve Gretzel, 2010). In addition, emerging social media technologies have great advantages for collecting, managing and sharing information that will facilitate management activities (Gretzel, Sigala, Xiang, ve Koo, 2015). In recent years, there has been a rapid increase in the studies conducted in the tourism sector to reveal the relationship between guest experience and customer satisfaction (Xiang, Schwartz, Gerdes, ve Uysal, 2015), the intensity of emotions in reviews (Mankad, Han, Goh, ve Gavirneni, 2016) and the relationship between hotel performance and reviews (Xie, Zhang, ve Zhang, 2014) by using social media data. However, studies that will provide competitive advantage for businesses from social media data in Turkey have not been observed in the research conducted by the author. It is important to carry out studies in this field in Turkey as well. The tourism sector is one of the most valuable sectors of the Turkish economy. Especially due to foreign exchange input, it contributes very importantly to the balance of external deficits and to reduce unemployment (Çımat ve Bahar, 2003). The main purpose of this study is to create competitive intelligence for hotel businesses with the help of current text mining methods from big social media data. Antalya region has been chosen as the application area of the research because Antalya is seen as the tourism capital of Turkey. The data consists of user reviews of all accommodation facilities with a business record on the Tripadvisor online social networking platform.

Methodology and Findings

In accordance with the purpose of this study, reviews of hotels serving in Antalya region were collected automatically through Tripadvisor with developed crawler. Total number of reviews 212.435. The total number of facilities is 1801. In some reviews, it has been observed that a different language is used in addition to English. A total of 537 comments in this case were automatically identified and removed from the analysis by the Python library called *langdetect*. If the number of reviews of a hotel is less than 10, those hotel reviews are also removed in order to increase the reliability of the study. in the last case, the total number of reviews decreased to 209.171 and the total number of hotels decreased to 1.072. While the number of hotels decreases by 729, the average number of reviews of the removed enterprises is 4. The average number of reviews for the remaining hotels is 195. Reviews include hotel name, user country (if specified), comment title, comment text, comment date and comment rating. When the

ratings are reviewed according to their rating values, the comments that have 4 and 5 rating values are approximately 81%. On the other hand, comments with 3 ratings are approximately 9%, while comments with 1 and 2 ratings are about 10%. The average length of comments is 206 words. The longest comment is 5,383 and the shortest is 4 words / emojis. Figure 3 (look at Şekil 3 above) shows the average length of comments based on rating values. As it can be understood from the figure, comments indicating satisfaction are shorter and comments containing dissatisfaction are longer. In particular, the longest comments are those with rating 3. It is thought that both positive and negative expressions may be present in these comments.

Preprocessing: Comments must be converted to digital form to be processed by the machine. For this process, all unique words are assigned a numerical ID and each comment is represented in the matrix by the frequencies of the word in it (bag-of-words method). The tf-idf calculation was used for the frequency of the words. Then Document Term Matrix (DTM) was created by using CountVectorizer module in Scikit-learn library. Fist, in order to avoid too much matrix size, all punctuation and numerical characters in the comments were removed and the remaining characters were lowercased. Along with words less than three characters, very common words called stop words were removed from the reviews. Words with the same root, such as use, useful, have been reduced to the same root with stemming process.

Analysis: Sentiment Analysis was performed to find the intensity of emotion of the reviews. For this analysis, Logistic Regression, Support Vector Machine and Naïve Bayes which are three of most preferred (Ravi & Ravi, 2015) methods in text classification was used. The classification success of these three algorithms is very similar. It can be seen in table 2 (Tablo 2) above. As a result of the Sentiment Analysis, it was found that 80% of reviews were positive and 20% were negative. This situation varies according to regions. For example, while the average of emotions of reviews for Kaş and Kalkan is 90%, it is 68% especially in Alanya which has the lowest average emotion. In addition, by doing Topic Sentence Analysis with LDA methods, the reviews were clustered under 5 topics (decided by calculating the perplexity[1] value); While *Experience* is the most talked about with 26.7%, *Value and Entertainment* is the second with 24.68% and *Complaint* is the third with 20.41%. Other topics mentioned in the reviews; *Basic Services* with 16.15% and *Things to do* with 12.06%.

Conclusion and Recommendation

Compared with methods such as surveys and interviews, social media interpretations are both much larger and can be collected at much more affordable costs. In addition, there is an objectivity of customers to share their opinions and thoughts with their own wishes. As a result of these analyses, which are applied to the social media data in the tourism sector, it is thought that the enterprises will acquire important information about the competitive environment in their strategies. In addition, it can be inference that the opportunities or threats of the region will be realized during the market researches. In future studies, it is thought that the separation of accommodation facilities by types and classes will yield more effective results in terms of competitor analysis.

[1] Perplexity is a statistical measure of how well a probability model predicts a sample. For optimal clustering, the number of topics in the model with a small perplexity value is preferred.