YÖNETİM BİLİŞİM SİSTEMLERİ DERGİSİ

http://dergipark.ulakbim.gov.tr/ybs/

Yayın Geliş Tarihi: 30.05.2016 Cilt:1, Sayı:3, Yıl:2016, Sayfa 29-39

Yayına Kabul Tarihi: 15.08.2016 ISSN: 2148-3752

Online Yayın Tarihi: 05.10.2016

LINGUISTIC STUDIES ON TWEETS GATHERED FROM MUĞLA REGION: A PRELIMINARY STUDY

Feriştah Dalkılıç

Enis Karaarslan

Ali Hurriyetoglu

Abstract

Citizens or visitors of a city can supply significant information with their social media posts by using mobile devices. These data can give information about complaints, touristic attractions, emergency situations etc. Social media analysis will be beneficial for smart city and smart management concept. This study is a first attempt to analyze and understand this touristic Muğla region by using social media. During this study, a sample dataset is formed by collecting the tweets that were sent from the Muğla region. Linguistic studies are implemented in tweets which are in Turkish language. Various techniques, statistical language and characteristics are used. Preliminary study revealed main topics about the region, user and hashtag types. We consider this analysis as a first step to a more detailed and complete study for this region.

Keywords: social media analysis, tweet, smart city, crowd sensing

MUĞLA BÖLGESINDEN TOPLANAN TWEET VERILERI ÜZERINDE DILBILIMSEL ÇALIŞMALAR: ÖN ÇALIŞMA

Öz

Bir şehrin sakinleri ve ya ziyaretçileri, mobil cihazlar aracılığıyla sosyal ağlar üzerinde önemli miktarda veri üretmektedirler. Bu veriler; şikâyetler, acil durum halleri, turistik eğlence programları gibi konularda bilgi sağlayabilmektedir. Sosyal medya analizinin, akıllı şehir ve akıllı yönetim sistemleri için faydalı olacağı düşünülmektedir. Bu çalışma, turistik bir bölge olan Muğla yöresini sosyal medya kullanarak analiz etmek için bir ilk girişimdir. Bu çalışma süresince, Muğla yöresinden atılan tweetler toplanarak örnek bir veri seti oluşturulmuştur. Türkçe tweetler üzerine dilbilimsel bir çalışma uygulanmıştır. Çeşitli teknikler, istatistiksel dil ve karakteristikler kullanılmıştır. Ön çalışma, sosyal ağda kullanılan dilin özelliklerini, yöre hakkındaki ana konuları, kullanıcı ve hashtag tiplerini ortaya sermiştir. Bu çalışma, yöre hakkında daha ayrıntılı ve tam bir analiz için ilk adımı oluşturmaktadır.

Anahtar Kelimeler: sosyal medya analizi, tweet, akıllı şehir, kalabalık algılama

INTRODUCTION

Smart city concept is built on smart management of all city structures like power, water, and transportation etc. for a safe, secure and efficient usage of resources. Several technologies can be used for this concept such as sensors, electronics, computerized systems, networks and communication systems etc. (Bowerman et al., 2000).

We are living in a knowledge society. Citizens or visitors of a city share the up-to-date events and their emotions to these events by using social media (Twitter, Facebook, Instagram etc.) environments. These environments have publicly open, free and real time data. These data can be analyzed to obtain complaints, touristic attractions and emergency situations of the cities. The produced knowledge can be used for smart management of the cities.

Analysis of social media data and converting it to benefit is one the most important research topics today. The big data and cloud environments are becoming widespread everyday facilitated by the fact that these can be set up with a less economical cost than before. These systems make social media data processing and storage easier. Social media analysis is used for journalism, tourism and commercial applications such as public research about the companies. It's also possible to use it for scientific research and the public interest purposes such as collecting information about natural disasters or national security. Retrospective and instantaneous detection can be done with social media analysis; also predictions about the future can be made in light of the information obtained.

Starting with a small dataset, this research aims to reach some preliminary results which will be used in the following studies. In the first section, related work will be given. Next, the usage of Social Media for Smart City and Smart Management is discussed. Then the implementation and experimental results will be given. Lastly, detected results and the possible future work will be discussed.

RELATED WORK

Many studies have been developed using Twitter data for a variety of purposes. Most of the studies are on sentiment analysis. Mohammad and Kiritchenko (2015) used hashtags to capture fine emotion categories from tweets. Cavazos-Rehg et al. (2016) examined depression-related content in Twitter to glean insight into social networking about mental health. Kunneman et al. (2014) experimented the majority-based and machine-learning methods for the identification of future event start dates from Twitter streams. Serrano et al. (2015) presented a simulation tool implementing several models of rumor diffusion in Twitter. Shendge et al. (2015) worked on real time Tweet analysis for event detection and reporting system for Earthquake. Tumasjan et al. (2010) presented a study uses the context of the German federal election to investigate whether Twitter is used as a forum for political deliberation and whether online messages on Twitter validly mirror offline political sentiment.

Understanding a city through lens of social media has been studied by Bakıcı et al. (2013). Social event detection (Ilina et al., 2012), traffic event detection (Anantharam et al., 2015), touristic support (Leung et al., 2013; Quercia et al., 2014) are the most prominent recent approaches. Every bit of information in the aforementioned lines and the ones yet to be discovered contributes to understanding and management of a city. Social media analysis can also be used for crowd sensing and some implementation examples are given in some studies like (Roitman et al., 2012). A recent work (Preece et al., 2015) proposes a method to get active response from the tweeters.

Multilingual Analysis of twitter data is important and as far as our knowledge, there are not many academic studies which focus in analyzing tweets in Turkish language. In an

interesting study (Zielinski, 2012), multilingual (Romanian, Greek and Turkish) analysis of Twitter data is implemented to detect human responses during earthquakes. In a recent study (Demirci, 2014), micro-blog entries and special usage of symbols and conveniences are studied. This work states that a new data set of Turkish tweets for emotion analysis is constructed.

SOCIAL MEDIA ANALYSIS FOR THE SMART CITY

Smart city and smart management depends upon collecting real-time data. Common interest of the citizens can be learned by collecting data from the sensing devices which is called crowd sensing. There are different types of crowd sensing; social media analysis can be used as opportunistic and social crowd sensing where mobile devices are mostly used and there is minimal involvement of the user (INFOSEC, 2014). Such a system should consist of some tools which will collect data from social media, filter unnecessary data and noise in the collected dataset, transform into a uniform format, store data in a database, and make analysis to produce meaningful information. The last step's success depends on the content of the data and the algorithms used to analyze this data. Different data sources like physical sensors can be integrated to this system (Roitman et al., 2012).

EXPERIMENTAL

In this study, a Python script has been developed to collect tweets. Dataset (500 MB of Twitter data posted in Muğla, geolocated tweets) has been collected between dates February 22 and March 6 in 2016. Twitter data contains many attributes such as tweet id, creation time, coordinate, place, the user who posted the tweet, text, entities, etc. This data has been retrieved in JSON format. Fields, which do not have any data, have been eliminated and resulting dataset has been stored into a Mongo database.

Firstly, it's expected to get an idea about the user count, mostly tweeting users, hashtag count, most used hashtags, temporal distribution of the tweets, similarities and differences between Tweet and traditional language use.

Before starting linguistic studies, body texts of the tweets have been retrieved and data cleaning process has been performed on this data. 14,501 tweets which have auto generated content (such as "I'm at Muğla, Türkiye", "Just posted a photo @ Muğla" etc.) have been eliminated completely. Body texts of the remaining tweets have also been filtered to eliminate web links, hashtags and user ids.

The resulting filtered text was used to form a Twitter corpus which has 29 characters of Turkish alphabet and the space character. All words containing Q, W, X characters which don't belong to Turkish alphabet have been eliminated completely. A Turkish corpus which had been formed by collecting a large amount of Turkish newspaper articles (Örücü, 2009) has been used for the linguistic comparisons. The properties of these two corpora are given in Table 1.

Table 1. Properties of Twitter and Turkish corpora.

	Twitter Corpus	**Turkish Corpus**
Word count	666,903	105,863,484
Character count	5,214,629	776,755,254
Size (MB)	4.97	857

The Twitter corpus is an example of contemporary Turkish language and has an extensive word variety. Several analyses based on letters and words were made on the corpus, which will be given in the following sub sections.

N-gram Analysis

N-gram analysis is an effective technique in modeling language data and widely used in computational linguistics, computational biology and data compression. In this context, most common letter and word n-gram analysis has been performed on both Twitter and Turkish corpora.

Most common letter n-grams

Letter n-gram analysis has been performed (n = 1 to n = 6) by using virtual corpus (Kit and Wilks, 1998) algorithm. The most frequently used 30 letter n-grams for $1 \leq n \leq 3$ and $4 \leq n \leq 6$ are given in Table 2 and Table 3, respectively. (# is used for space character in the content.)

The order and frequencies of the 1-grams are observed almost identical for the two corpora. Also, 27 of the first 30 2-grams and 25 of the first 30 3-grams are common for these corpora.

Table 2. Most frequently used 30 letter n-grams ($1 \leq n \leq 3$) for Twitter and Turkish.

	Tweet		**Turkish**		**Tweet**		**Turkish**		**Tweet**		**Turkish**	
	1	**%**	**1**	**%**	**2**	**%**	**2**	**%**	**3**	**%**	**3**	**%**
1	#	14.331	#	13.629%	N#	2.231	N#	2.127	#Bİ	0.591	LAR	0.696
2	A	10.304	A	10.241%	#B	1.716	E#	1.789	EN#	0.562	#Bİ	0.595
3	E	7.836	E	8.011%	A#	1.623	#B	1.656	LAR	0.536	LER	0.547
4	İ	7.231	İ	7.457%	E#	1.611	AR	1.621	#YA	0.512	AN#	0.515
5	N	6.063	N	6.341%	R#	1.543	A#	1.614	AN#	0.497	İN#	0.487
6	R	5.465	R	6.029%	AR	1.499	R#	1.545	İN#	0.496	İR#	0.480
7	L	4.993	L	5.526%	İ#	1.437	İ#	1.529	#KA	0.465	EN#	0.469
8	M	3.940	I	4.134%	AN	1.336	LA	1.481	LER	0.431	ERİ	0.464
9	K	3.872	K	4.017%	LA	1.276	AN	1.412	AR#	0.429	DA#	0.463
10	I	3.745	D	3.679%	ER	1.257	ER	1.355	#DE	0.417	#YA	0.456
11	D	3.658	M	3.201%	İN	1.176	İN	1.258	YOR	0.387	BİR	0.451
12	Y	3.168	T	3.050%	EN	1.151	LE	1.244	#OL	0.375	#DE	0.429
13	S	3.087	Y	3.009%	#A	1.128	#D	1.170	DA#	0.365	#KA	0.428
14	T	2.784	S	2.713%	#D	1.116	DE	1.105	İR#	0.348	ARI	0.427
15	U	2.760	U	2.642%	#K	1.104	#K	1.079	DE#	0.343	DE#	0.420
16	O	2.411	O	2.294%	#S	1.086	I#	1.063	ERİ	0.339	YOR	0.364
17	B	2.300	B	2,204%	M#	1.058	#A	1.001	ER#	0.338	IN#	0.358
18	Z	1.668	Ü	1.627%	DE	1.036	EN	1.000	#HA	0.327	#BU	0.356
19	Ü	1.423	Ş	1.387%	LE	0.996	IN	0.984	IN#	0.324	AR#	0.355
20	Ş	1.278	Z	1.311%	K#	0.986	DA	0.951	#BE	0.322	#VE	0.352
21	G	1.264	G	1.095%	#Y	0.928	K#	0.924	İM#	0.317	#OL	0.344
22	H	1.203	H	0.928%	#G	0.902	#Y	0.900	#BU	0.306	#BA	0.335
23	C	0.938	Ç	0.922%	MA	0.891	#S	0.879	#BA	0.304	ARA	0.322
24	V	0.829	V	0.876%	YA	0.891	YA	0.867	BİR	0.297	NDA	0.309
25	Ç	0.826	Ğ	0.870%	I#	0.858	MA	0.841	ARI	0.297	#GE	0.307
26	P	0.745	C	0.854%	DA	0.818	İR	0.840	#SE	0.292	ER#	0.287

27	Ğ	0.720	P	0.766%	Bİ	0.805	Bİ	0.794	#GE	0.286	N#B	0.277
28	Ö	0.581	Ö	0.698%	IN	0.756	#G	0.786	#SA	0.279	İNİ	0.270
29	F	0.495	F	0.432%	#İ	0.735	İL	0.769	N#B	0.276	#HA	0.263
30	J	0.083	J	0.056%	AL	0.735	KA	0.768	NE#	0.264	İLE	0.259
∑		**100**		**100**		**34.68**		**35.35**		**11.33**		**12.09**

Table 3. Most frequently used 30 letter n-grams (4≤n≤6) for Twitter and Turkish.

	Tweet		**Turkish**		**Tweet**		**Turkish**		**Tweet**		**Turkish**	
	4	**%**	**4**	**%**	**5**	**%**	**5**	**%**	**6**	**%**	**6**	**%**
1	#BİR	0.278	#BİR	0.427	#BİR#	0.2009	#BİR#	0.3396	YORUM#	0.1103	#İÇİN#	0.0799
2	LAR#	0.218	BİR#	0.351	YORUM	0.1152	LARIN	0.1749	#İNSAN	0.0693	LARIN#	0.0631
3	BİR#	0.209	LARI	0.323	ORUM#	0.1120	LERİN	0.1556	#İÇİN#	0.0687	#TÜRKİ	0.0625
4	LARI	0.201	LERİ	0.289	#İÇİN	0.0865	INDA#	0.1259	İYORUM	0.0625	TÜRKİY	0.0624
5	LERİ	0.186	#VE#	0.250	LARIN	0.0859	LARI#	0.1228	#KADAR	0.0608	ÜRKİYE	0.0624
6	İYOR	0.183	YOR#	0.220	ANLAR	0.0853	LERİ#	0.1106	KADAR#	0.0603	LARINI	0.0613
7	#BEN	0.171	ERİN	0.210	#ÇOK#	0.0814	#İÇİN	0.1074	#DEĞİL	0.0584	N#BİR#	0.0612
8	LER#	0.166	#BU#	0.207	İYORU	0.0808	İNDE#	0.1023	#GİBİ#	0.0579	INDAN#	0.0585
9	#BU#	0.164	INDA	0.206	LARI#	0.0784	İYOR#	0.0965	#GÜZEL	0.0479	LERİNİ	0.0563
10	DEN#	0.156	LAR#	0.200	#OLMA	0.0778	#TÜRK	0.0936	#KENDİ	0.0456	#DAHA#	0.0560
11	YORU	0.153	ARIN	0.197	LERİN	0.0771	İNİN#	0.0914	#DAHA#	0.0451	İ#BİR#	0.0527
12	YOR#	0.152	NDA#	0.184	#BEN#	0.0756	N#BİR	0.0849	#ZAMAN	0.0432	#GİBİ#	0.0524
13	NLAR	0.143	NİN#	0.163	#VAR#	0.0722	NDAN#	0.0823	#ALLAH	0.0430	LERİN#	0.0514
14	ANLA	0.137	İNDE	0.160	#DEĞİ	0.0712	İÇİN#	0.0815	DEĞİL#	0.0418	#DEĞİL	0.0500
15	#VE#	0.131	İYOR	0.160	#İNSA	0.0703	ININ#	0.0762	OLSUN#	0.0394	#KENDİ	0.0471
16	RUM#	0.129	DEN#	0.157	İNSAN	0.0693	IYOR#	0.0744	#OLSUN	0.0379	#KADAR	0.0455
17	ORUM	0.121	DAN#	0.156	İÇİN#	0.0692	#DEĞİ	0.0742	ANLAR#	0.0369	LARAK#	0.0442
18	#SEN	0.119	LER#	0.151	LERİ#	0.0689	ARIN#	0.0711	GÜZEL#	0.0367	KADAR#	0.0433
19	ERİN	0.117	NIN#	0.145	INDA#	0.0676	#OLMA	0.0676	#SONRA	0.0357	#SONRA	0.0433
20	DAN#	0.115	ERİ#	0.144	İYOR#	0.0668	ARINI	0.0670	İNSANL	0.0340	#BAŞKA	0.0430
21	IYOR	0.113	ARI#	0.143	#GİBİ	0.0661	ANLAR	0.0646	YORUZ#	0.0340	ASINDA	0.0429
22	#GEL	0.112	#BAŞ	0.137	#AMA#	0.0631	ERİNİ	0.0630	ALLAH#	0.0333	E#BİR#	0.0426
23	#VAR	0.111	İNİ#	0.136	KADAR	0.0626	#ÇOK#	0.0630	#DİYE#	0.0331	ERİNDE	0.0424
24	#YAP	0.110	#DE#	0.134	NLAR#	0.0617	#OLDU	0.0627	N#BİR#	0.0328	OLDUĞU	0.0416
25	ARIN	0.108	NLAR	0.134	#KADA	0.0610	TÜRKİ	0.0626	#BENİM	0.0322	#OLDUĞ	0.0414
26	#NE#	0.108	#OLA	0.133	ADAR#	0.0604	RKİYE	0.0625	LARIN#	0.0322	İNDEN#	0.0407
27	INDA	0.105	İNE#	0.132	DEĞİL	0.0591	NLARI	0.0624	LARINI	0.0320	YORUM#	0.0406
28	NDA#	0.105	INI#	0.131	GİBİ#	0.0581	ÜRKİY	0.0624	#HAYAT	0.0314	OLARAK	0.0404
29	İNE#	0.100	#DA#	0.131	#BENİ	0.0570	ARAK#	0.0622	NSANLA	0.0310	#OLARA	0.0404
30	#İÇİ	0.099	NDE#	0.122	#OLDU	0.0561	ANIN#	0.0619	SANLAR	0.0307	#KARŞI	0.0403
∑		**4.318**		**5.63**		**2.3175**		**2.83**		**1.3584**		**1.51**

According to n-gram analyses (4≤n≤6) on Twitter and Turkish corpus, 19, 14 and 12 of first 30 entities are common for 4 to 6 grams, respectively. The similarities between n-gram series reduce while n is getting bigger.

Most common word n-grams

Most frequently used 20 word n-grams (1≤n≤2) in Twitter and Turkish corpus is given in Table 4.

Table 4. Top 20 word n-grams (1≤n≤2) in Twitter and Turkish corpora.

	Twitter		**Turkish**		**Twitter**		**Turkish**	
	Unigram	**%**	**Unigram**	**%**	**Bigram**	**%%**	**Bigram**	**%%**
1	BİR	1.402	BİR	2.491	CEZA#İNDİRİMİ	12.506	YA#DA	9.930
2	BU	1.144	VE	1.831	İÇERDE#BİN	12.056	BÖYLE#BİR	5.950
3	VE	0.912	BU	1.522	DIŞARDA#DOSYA	12.056	HEM#DE	5.930
4	NE	0.756	DE	0.986	BİN#İÇERDE	12.056	BİR#ŞEY	5.750
5	DE	0.632	DA	0.964	BİN#DIŞARDA	12.056	NE#KADAR	5.170
6	ÇOK	0.568	İÇİN	0.586	DOSYA#YARGITAYDA	12.041	BİR#DE	4.510
7	DA	0.562	ÇOK	0.462	YARGITAYDA#CEZA	12.041	BU#KADAR	4.220
8	BEN	0.527	NE	0.422	AŞKINDİĞERADI# MURATYILDIRIM	10.616	YENİ#BİR	3.900
9	O	0.508	DAHA	0.411	BU#KADAR	8.082	VE#BU	3.770
10	VAR	0.504	AMA	0.411	NE#KADAR	6.973	BÜYÜK#BİR	3.640
11	İÇİN	0.479	GİBİ	0.384	O#KADAR	5.368	EN#BÜYÜK	3.360
12	AMA	0.440	O	0.378	BİR#ŞEY	5.323	O#ZAMAN	3.290
13	KADAR	0.410	İLE	0.357	GÖNLÜMÜZÜNKRALI #MURATYILDIRIM	5.218	BU#KONUDA	3.290
14	GİBİ	0.404	EN	0.322	O#ZAMAN	4.738	O#KADAR	3.210
15	YA	0.392	KADAR	0.318	BİR#GÜN	4.124	ÖNEMLİ#BİR	3.210
16	EN	0.338	VAR	0.310	EMEKLİLİKTE#YAŞA	3.869	DAHA#DA	3.030
17	İYİ	0.335	OLARAK	0.296	YAŞA#TAKILANLAR	3.854	BEN#DE	3.010
18	YOK	0.334	Kİ	0.292	İYİ#GECELER	3.704	DE#BU	2.970
19	Bİ	0.332	HER	0.283	NE#DEMEK	3.659	BİR#BAŞKA	2.860
20	DAHA	0.315	DEĞİL	0.274	YA#DA	3.614	BAŞKA#BİR	2.800

First 20 word 1-grams of corpora have 15 common entities when 2-grams have only 6 common entities. Some of the 2-grams occurring in Twitter corpus show the effect of the hot topics on the language used.

Linguistic Features

In the fields of computational linguistics and natural language processing, some features such as Type/Token Ratio, Hapax Legomena Ratio, Index of Coincidence, Entropy and Redundancy are used to notice differences of languages, or different authors of same language. In this section these features are examined on Twitter and Turkish corpora.

Type/token ratio

Type-Token Ratio (TTR) is a measure of lexical diversity and shows the richness of vocabulary usage in a text. TTR is the ratio obtained by dividing the total number of different words (types) by the total number of words (tokens) occurring in a text as given in Formula 1. Table 5 shows the TTR values for Twitter and Turkish corpus. High TTR values indicate high degree of lexical variation.

$$TTR = \frac{Number\ of\ types\ \times 100}{Total\ number\ of\ tokens} \qquad (1)$$

Table 5. TTR values for Twitter and Turkish corpus

Corpus	Types	Tokens	TTR
Twitter	113,712	666,904	17.05
Turkish	1,291,005	105,863,484	1.22

Hapax legomena ratio

The Hapax Legomena Ratio (HR) is the ratio in percent between once-occurring types (hapax legomena) and the vocabulary size (types). This ratio is calculated by using Formula 2 and the HR values for Twitter and Turkish corpus are given in Table 6.

$$TTR = \frac{Number\ of\ types\ \times 100}{Total\ number\ of\ tokens} \qquad (2)$$

Table 6. HR values for Twitter and Turkish corpus

Corpus	Once Occurring Types	Types	HR
Twitter	70,293	113,712	61.82
Turkish	497,343	1,291,005	38.52

Index of coincidence

Index of coincidence (IC) was introduced as a statistical measure of text which distinguishes different languages or encrypted text from plain text (Friedman, 1922). IC shows the probability of two randomly chosen letters is being equal in a text. The Formula 3 is used to calculate IC; where f_i is the frequency of the i^{th} letter of the alphabet and N is the number of letters in alphabet. IC values for some languages are given in Table 7 (Menezes, 1996).

$$IC = \sum_{i=1}^{N} \frac{(f_i \times (f_i - 1))}{N(N-1)} \qquad (3)$$

When the Formula 3 is applied on both Twitter and Turkish corpus, same IC values are obtained as follows:

$IC_{Twitter} = (R_{\#})^2 + (R_A)^2 + \ldots + (R_J)^2 = (13.629\%)^2 + (10.241\%)^2 + \ldots + (0.056\%)^2 = 0.063$
$IC_{Turkish} = (R_{\#})^2 + (R_A)^2 + \ldots + (R_J)^2 = (14{,}331\%)^2 + (10{,}304\%)^2 + \ldots + (0{,}083\%)^2 = 0.063$

Table 7. IC values of some languages.

Language	French	Spanish	German	Italian	English	Russian	Turkish
IC	0.0778	0.0775	0.0762	0.0738	0.0667	0.0529	0.0630

This shows that the language used in social media has same characteristics with the formal Turkish language. Another result observed is that the corpus sizes do not affect the IC value of a language.

Entropy and redundancy

Entropy (H) is a measure which is used for lack of order or predictability. Information entropy is used to reveal the information distribution in a text of a language. Entropy is the lower bound to the number of bits per symbol required to encode a long string of text drawn from a language. Entropy values of n-gram series are calculated by using the Formula 4, where *x* is every n-gram observed in the corpus, *p* is the probability of n-gram.

$$\mathrm{H}(X) = -\frac{1}{n}\sum_{x \in X} \mathrm{p(x)} \log_2 \mathrm{p(x)} \qquad (4)$$

Linguistic Redundancy (R) can be described as the content which has the same information with different words. "Two tweets as redundant if they either convey the same information (paraphrase) or if the information of one tweet subsumes the information of the other (textual entailment) (Zanzotto, 2011). Redundancy of an n-gram series is calculated by taking difference of its entropy from maximum entropy value as shown in Formula 5.

$$\mathrm{R_n} = (\mathrm{H_0} - \mathrm{E}) = (\log_2 \mathrm{P} - \mathrm{E_n}) \qquad (5)$$

H_0 is the maximum value of the entropy when all the letters are considered equally probable in a language. $H_0 = \log_2 P = 4.9069$ bits (P = 30, 29 letters in alphabet and space) for the Turkish language.

Using single letter distributions given in Table 2, the H_1 values are calculated as 4.3517 and 4.3570 bits for Turkish and Twitter corpora, respectively. Per letter redundancy R_1 is 4.9069 - 4.3517 = 0.5552 for Turkish and 4.9069 - 4.3570 = 0.5499 bits for Twitter corpus. Table 8 shows the entropy and redundancy values for the order 1 to 6.

Table 8. n^{th} order ($1 \leq n \leq 6$) entropy and redundancy values for Twitter and Turkish.

		n = 1	**n = 2**	**n = 3**	**n = 4**	**n = 5**	**n = 6**
Twitter	**H**	4.3570	3.9695	3.6512	3.3548	3.0941	2.8632
	R	0.5499	0.9373	1.2557	1.5520	1.8128	2.0437
Turkish	**H**	4.3517	3.9411	3.6034	3.2923	3.0342	2.8277
	R	0.5552	0.9658	1.3035	1.6146	1.8727	2.0792

User Analysis

The users play a crucial role in content generation on the social media. Their attitude toward content consumption and generation and interaction with each other differs drastically and determines characteristics of a social media platform. Therefore, we consider user analysis as a critical task in social media analysis. We analyze user behavior in two aspects, individual and group characteristics.

We noted 9,007 users in our data set. The mean, median, and standard deviation of the tweet count per user is 10.80, 2, and 46.84 respectively. This skewed tweet count per user distribution is caused by the outliers (users who tweet a lot more than most of the users). On the one hand, the most frequently tweeting user has 2,820 tweets in the data set, two times more than the second most tweeting user. On the other hand, 3,206 users have only 1 tweet in the data set.

Hashtag Analysis

Hashtags are used to form topics or discussions on social media. Tweets that have the same hashtag are assumed to be on the same track. In this section, we analyze hashtags both qualitatively and quantitatively.

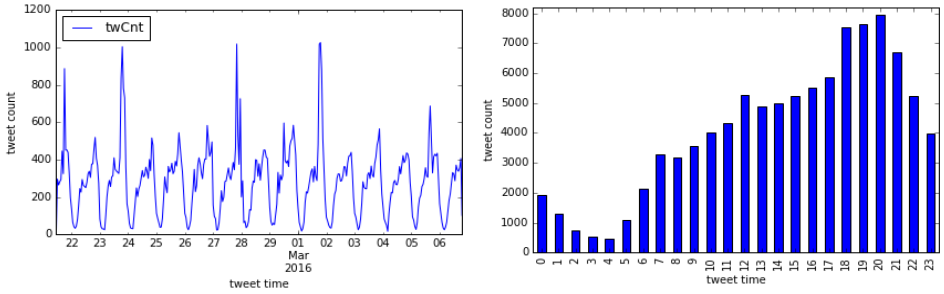
A tweet may contain one or several hashtags. In our dataset 13,578 tweets contain at least one hashtag. Although there are tweets that contain up to 14 hashtags, the number of hashtags per tweet is relatively stable. The mean, the median, and the standard deviation of the hashtag count per tweet is 1.67, 1.0, and 1.26 respectively.

The number of hashtag occurrence differs tremendously. The interests of the users in certain topics affect the use a hashtag. The popular TV program related hashtag 'GeceninKraliçesi' is used 3,216 times, which is almost two times more than the second most used hashtag `bedelineyse`.

Temporal Analysis

Every post has a timestamp on social media. This information enables short messages to be temporally contextualized. Moreover, the temporal distribution of group of tweets reveals significant signals. For instance, Figure 1 represents the daily rhythm of the social media use and times that generates relatively big peaks on the left. Time based tweet counts are given in the right figure, which demonstrates average rate of tweets during the day.

Figure 1. Temporal distribution and hourly average tweet account per hour during the day.



RESULTS AND DISCUSSION

As a result of linguistic studies, it has been observed that, the language used in social media has same characteristics with the formal Turkish language. However, word n-grams (n>1) differ with the effect of most tweeted topics.

Looking at the user analysis and given the big difference in tweet count per user, we suggest that the user dimension should be indispensable part of any social media analysis task. This basic example of user differences points us to delve into sophisticated interactions between the user and other social media message features, e.g., language use, position in social network.

Hashtag analysis showed that, using of hashtags facilitate forming topics and following a thread for social media users. On the other hand, hashtags provide a basis for effective automatic topic analysis. We can follow hashtags to understand main patterns of this rich and pervasive data source.

Temporal patterns informed us about tweeting behavior of the users living in this region. We observed the daily rhythm with a relative peak during lunch and dinner time, 12:00 and 20:00 o'clock respectively.

CONCLUSION AND FUTURE WORKS

In scope of this study, various linguistic techniques and statistical analysis have been applied on Twitter data gathered from Muğla region. Analyzing social media data gave us some clues about language characteristics, user and hashtag types. We consider this analysis as a first step to a more detailed and complete study for this region. It's planned to collect and store tweets for a long term to obtain more detailed information. A web interface where researchers will make real-time analysis of this collected data will be developed.

Potential use cases of preliminary results can be exemplified as uncovering most popular places for tourists, tracking deviations in language use to identify events or important discussions in Muğla, identifying best routes for tourists. In addition to the iterated use cases, any interaction among the aforementioned social media data dimensions has the potential to improve our understanding of social media which is becoming the reflection of the real-world.

This study focused on linguistic study in analyzing tweets in Turkish language. However, we believe that analyzing the tweets which are in other languages will provide us more insight of this touristic region and this is left for a future study. We also believe that, information systems should be developed where social media data will be integrated to other data sources for a better smart city management.

REFERENCES

Anantharam, P., Barnaghi, P., Thirunarayan, K., and Sheth, A. (2015). Extracting city traffic events from social streams. ACM Transactions on Intelligent Systems and Technology (TIST), 6(4), 43.

Bakıcı, T., Almirall, E., and Wareham, J. (2013). A smart city initiative: the case of Barcelona. Journal of the Knowledge Economy, 4(2), 135-148.

Bowerman, B., et al. "The vision of a smart city." 2nd International Life Extension Technology Workshop, Paris. Vol. 28. 2000.

Cavazos-Rehg, P.A., Krauss, M.J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., and Bierut, L.J. (2016). A content analysis of depression-related tweets. Computers in Human Behavior, 54: 351-357.

Demirci, S. (2014). Emotion Analysis On Turkish Tweets. M.Sc. Thesis, Middle East Technical University, Ankara.

Friedman, W. (1922). The Index of Coincidence and Its Applications in Cryptography. Publication No. 22. Geneva IL: Riverbank Publications.

Ilina, E., Hauff, C., Celik, I., Abel, F., and Houben, G.J. (2012). Social event detection on twitter. In Web Engineering (pp. 169-176). Springer Berlin Heidelberg.

INFOSEC Institute, Crowdsensing: State of the Art and Privacy Aspects, http://resources.infosecinstitute.com/crowdsensing-state-art-privacy-aspects/, 2014.

Kit, C., and Wilks, Y. (1998). The Virtual Approach to Deriving Ngram Statistics from Large Scale Corpora. International Conference on Chinese Information Processing Conference, Beijing, China, pp. 223—229.

Kunneman, F., Hürriyetoğlu, A., Oostdijk, N., and van den Bosch, A. (2014). Timely identification of event start dates from Twitter. Computational Linguistics in the Netherlands Journal, 4: 39-52.

Leung, D., Law, R., Van Hoof, H., and Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. Journal of Travel & Tourism Marketing, 30(1-2), 3-22.

Menezes, A., van Oorschot, P. and Vanstone, S. (1996). Handbook of Applied Cryptography. CRC Press.

Mohammad, S.M., and Kiritchenko, S. (2015). Using Hashtags to Capture Fine Emotion Categories From Tweets. Computational Intelligence, 31 (2): 301-326.

Örücü, F. (2009). Turkish Language Characteristics And Author Identification. M.Sc. Thesis, Dokuz Eylül University, İzmir.

Preece, A., Webberley, W., and Braines D. (2015). Tasking the tweeters: Obtaining actionable information from human sensors. SPIE Defense+ Security. International Society for Optics and Photonics.

Quercia, D., Schifanella, R., and Aiello, L.M. (2014). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In Proceedings of the 25th ACM conference on Hypertext and social media (pp. 116-125). ACM.

Roitman, H., et al. Harnessing the crowds for smart city sensing. Proceedings of the 1st international workshop on Multimodal crowd sensing. ACM, 2012.

Serrano, E., Iglesias, C.A., and Garijo, M. (2015). A Survey of Twitter Rumor Spreading Simulations. ICCCI 2015: 7th International Conference, September 21-23 2015, Madrid, Spain, pp. 113-122.

Shendge, G.V., Pawar, M.R., Patil, N.D., Pawar, P.R., and Bagul, D.B. (2015). Real time Tweet analysis for event detection & reporting system for Earthquake. IRJET: International Research Journal of Engineering and Technology, 2 (6): 706-712.

Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M. (2010). Predicting elections with Twitter:What 140 characters reveal about political sentiment. 4th International AAAI Conference on Weblogs and Social Media, May 23-26 2010, Washington, DC, pp. 178–185.

Zanzotto, F.M., Pennacchiotti, M., and Tsioutsiouliklis, K. (2011). Linguistic redundancy in twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Zielinski, A., and Bügel, U. (2012). Multilingual analysis of twitter news in support of mass emergency events. Proceedings of the 9th International ISCRAM Conference, April 22-25 2012, Vancouver, Canada.