YÖNETİM BİLİŞİM SİSTEMLERİ DERGİSİ
http://dergipark.ulakbim.gov.tr/ybs/

Yayın Geliş Tarihi: 30.05.2016
Yayına Kabul Tarihi: 15.08.2016
Online Yayın Tarihi: 05.10.2016

Cilt:1, Sayı:3, Yıl:2016, Sayfa 129-139
ISSN: 2148-3752

TOP 10 TURKISH UNIVERSITIES TWITTER ANALYSIS USER SENTIMENT ANALYSIS AND COMPARISON WITH INTERNATIONAL ONES

Mohammed ALSADI

Sevinç GÜLSEÇEN

Elif KARTAL

Abstract

During the last 5 yeras, the importance of social media is increasing in an amazing way. Social media gives individuals, companies, organizations and many others the oppurtunity to create, share, or exchange any kind of informations such as ideas, opinions, news, media and many others. With the huge improvements in Internet and Mobile sectors access to such websites is availiable from people’s smart phones. Since users are able to create and share different types of content via social media websites, such websites became as bank of data. Huge volume of useful and valuable information could be extracted from there. ,thus, they become as one of the primary source of information for both consumers and businesses.

Keywords twitter, user sentiment, opinion mining, text mining, features .

INTRODUCTION

Twitter is one of the well-known social media websites, founded in March 21, 2006. It allows users to follow other user’s accounts. Users are able to share what is called tweets up to 140 characters. Nowadays, due to the huge amount of information embedded inside in tweets, different analysis in various areas and sectors have been accomplished and the results are used by different companies or businesses.

Text mining (sometimes called text analytics) refers to the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output.

Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling. Sentiment analysis is a field of research that explores people’s opinions and thoughts towards different matters such as organizations, services, products, and events (Bing, 2012).In other words, it could be defined as determining the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state (the emotional state of the author when writing), or the intended emotional communication (the emotional effect the author wishes to have on the reader). Due to rapidly spread and highly accepted social networks, the role of sentiment analysis has been growing significantly. Nowadays, regardless to social media website, we can find a number of comments in each user’s post (Facebook case) or tweet (Twitter case). These comments made by different users on a point are very important especially for businesses. Mining this huge volume of user’s opinions provides information for understanding human behavior.

The main two methods of sentiment analysis, lexicon-based method (unsupervised approach) and machine learning based method (supervised approach), both rely on the bag-of-words. In the machine learning supervised method the classifiers are using the unigrams or their combinations as features. In the lexicon-based method the unigrams which are found in the lexicon are assigned a polarity score, the overall polarity score of the text is then computed as sum of the polarities of the unigrams.

During the process of sentiment analysis, there are two kinds of learning that usually could be used; supervised learning and unsupervised learning [2]. One of the well-known examples of the supervised learning is machine learning method in which a training set of data which is manually labeled is used. On the contrary, unsupervised learning does not need any training data [3]. An example of this approach is the lexicon-based method in which the collected dictionary is considered enough to reach a result. The aforementioned methods are different from each other, but many studies have shown the ability of combining or merging them. In [4] the authors discussed a way to combine them by using lexicon-based method to create labeled tweets which will be used later as a training data set for machine learning method.

In [5], a comparative study of existing techniques for sentiment analysis including machine learning and lexicon-based was provided. Results of the experiments showed that lexicon-based approach is very competitive since few effort is needed and this approach is not sensitive to the quantity and quality of training dataset.

In this research, the Turkey’s top ten universities twitter accounts will be analyzed and then compared to some outside universities from all over the world. Furthermore, a sentiment analysis will be made about what users write about the selected universities. In current days, social media is not just a platform for message exchange or writing comments for other's picture or events. The role and importance for them becomes much bigger. Nowadays, for example, students wish to know everything about their universities, such as activities, change in program, announcements and others through social media. The main goal of this research is check whether universities are able to satisfy the student's needs about social media and to which degree.

The rest of this paper is organized as the following. Section 2 gives details about the methodology used in this study. In section 3, the obtained results are introduced. Finally, the paper is concluded in section 4.

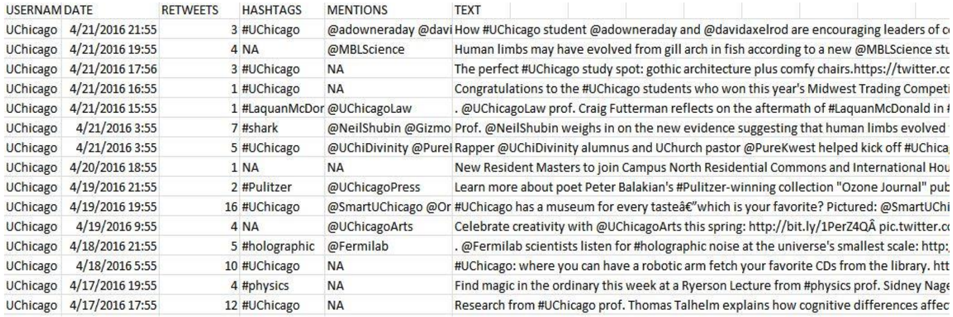
METHODOLOGY

According to URAP [6] report, the Turkey's top ten universities were announced. The universities included in this study are those announced in the report. They are: Middle East Technical University, Hacettepe University, Istanbul University, Bilkent University, Ankara University, Istanbul Technical University, Gebze Technical University, Ege University, Gazi University, Sabanci University. The tweets size of Gebze Technical University was just 65 tweets at the time of collecting data, so it has been replaced by KOC university. Data used for sentiment analysis in this research is gathered from the formal twitter account of the aforementioned universities. Furthermore, data from the following international universities' twitter accounts was collected. These universities are: 1) California Institute of Technology (CALTECH). 2) University of Chicago. 3)Harvard University. 4)Imperial College London. 5)Oxford University. 6)Princeton University. 7)Yale University 8) Royal Holloway, University of London. 9) Massachusetts Institute of Technology, and 10) University of Pennsylvania.

Data Collection

The data was collected during April and May 2016 using an open source java soruce [7] with some kinds of modification to suite the required case of our study. A snap of the used code is shown in Figure 2. Twitter username was used to search for the required data, for each university the maximum number of colleccted tweetes was about 3000. Each row of data contains : username, data, retweets, hashtags, mentions, and text. A sample is given in Figure 1.

Figure 1. Data Sample



USERNAM	DATE	RETWEETS	HASHTAGS	MENTIONS	TEXT
UChicago	4/21/2016 21:55	3	#UChicago	@adowneraday @davi	How #UChicago student @adowneraday and @davidaxelrod are encouraging leaders of c
UChicago	4/21/2016 19:55	4	NA	@MBLScience	Human limbs may have evolved from gill arch in fish according to a new @MBLScience stu
UChicago	4/21/2016 17:56	3	#UChicago	NA	The perfect #UChicago study spot: gothic architecture plus comfy chairs.https://twitter.cc
UChicago	4/21/2016 16:55	1	#UChicago	NA	Congratulations to the #UChicago students who won this year's Midwest Trading Competi
UChicago	4/21/2016 15:55	1	#LaquanMcDor	@UChicagoLaw	. @UChicagoLaw prof. Craig Futterman reflects on the aftermath of #LaquanMcDonald in #
UChicago	4/21/2016 3:55	7	#shark	@NeilShubin @Gizmo	Prof. @NeilShubin weighs in on the new evidence suggesting that human limbs evolved
UChicago	4/21/2016 3:55	5	#UChicago	@UChiDivinity @PureI	Rapper @UChiDivinity alumnus and UChurch pastor @PureKwest helped kick off #UChica
UChicago	4/20/2016 18:55	1	NA	NA	New Resident Masters to join Campus North Residential Commons and International Hou
UChicago	4/19/2016 21:55	2	#Pulitzer	@UChicagoPress	Learn more about poet Peter Balakian's #Pulitzer-winning collection "Ozone Journal" pub
UChicago	4/19/2016 19:55	16	#UChicago	@SmartUChicago @Or	#UChicago has a museum for every tasteâ€"which is your favorite? Pictured: @SmartUChi
UChicago	4/19/2016 9:55	4	NA	@UChicagoArts	Celebrate creativity with @UChicagoArts this spring: http://bit.ly/1PerZ4QÂ pic.twitter.cc
UChicago	4/18/2016 21:55	5	#holographic	@Fermilab	. @Fermilab scientists listen for #holographic noise at the universe's smallest scale: http:
UChicago	4/18/2016 5:55	10	#UChicago	NA	#UChicago: where you can have a robotic arm fetch your favorite CDs from the library. htt
UChicago	4/17/2016 19:55	4	#physics	NA	Find magic in the ordinary this week at a Ryerson Lecture from #physics prof. Sidney Nage
UChicago	4/17/2016 17:55	12	#UChicago	NA	Research from #UChicago prof. Thomas Talhelm explains how cognitive differences affec

As mentioned before, twitter has limited the length of tweet length to 140 characters. This leads users to use acronyms, remove some letters from words, and use emoticons to express special meanings. Besides normal text a tweet could has 1) Emoticons : facial expressions pictorially represented by punctuation and letters. 2) Hashtags : using "#" symbol to mark topics. 3) Target / mention : using the "'@" symbol to refer to other users. For example the following tweet from

Istanbul University contains 2 target/mention which are *istanbuledutr* and *yunussoylet* ,whereas it does not contains a hashtag. Further more, it is usual to see a link in tweets.

> @istanbuledutr: İstanbul Üniversitesi Rektörü Prof. Dr. Yunus Söylet' in öğretmenler günü kutlama mesajı @yunussoylet http://www2.istanbul.edu.tr/?p=14435

Data Preprocessing

Data preprocessing is an important stage in data mining. In order to get high and accurate results, data is to be preprocessed. Data preprocessing includes : data cleaning, data integration, data reduction, and many others. In our case, hashtags, emoticons, targets, digits, unncecesary spaces, and any URL were removed.

Figure 2. Tweet text after some preprocessing

Before Preprocessing	Yarın #markethinkdays'de "Dijital Pazarlamaya Dair Yeni Şeyler Söylemek Lazım" paneli yer alacak. @BilkentMEC https://www.facebook.com/events/340522439382362/?ref=22 …
After Preprocessing	Yarın markethinkdays'de Dijital Pazarlamaya Dair Yeni Şeyler Söylemek Lazım paneli yer alacak

R program is used in this step. The complete data preprocessing overview is given in Figure 3.

Figure 3. Data preproceesing code in R.

```
# REMOVE ALL “@people”

Ttext = gsub("@\\w+", "", Ttext)

# REMOVE ALL punctuation

Ttext = gsub("[[:punct:]]","" , Ttext)

# REMOVE ALL NUMBERS

Ttext = gsub("[[:digit:]]", "", Ttext)

# REMOVE URLs

Ttext = gsub("http\\w+", "", Ttext)

# FINALLY, REMOVE ANY UNNECESSARY SPACES (white spaces, tabs etc)

Ttext = gsub("[ \t]{2,}", "", Ttext)

Ttext = gsub("^\\s+|\\s+$", "", Ttext)
```

Later, a list of meaningless words (English and Turkish) is removed. This list includes many words which do not have any effect on sentiment analysis. Example of such words is given below. At this point, the tweet text could be considered as a structured text that has a meaning and could be used to extract emotions.

Turkish	gün, var, o, sen, ben, tarih, ediyor, olsun, bugün, bizim, olan, için, ile
English	is, then, later, one, are, my, and,for,the,this,further, thus, ago, time

Used Approach

In this research, a number of tasks have been completed. Those tasks can be divided into two categories. First, analysing the collected tweet text for each university and a cloud of words for each university is presented, then a comparsison between Turkish universities and other universities is done. Second, user sentiment about what they are writing related to the involved universities.

Based on the obtained word cloud results from each university, a list of words related to science was not appeared, on the contrary the most used words obtained from collected tweets were generall words. A number of obtained word cloud for some of them is given below in the following figures.

Figure 4. Istanbul University

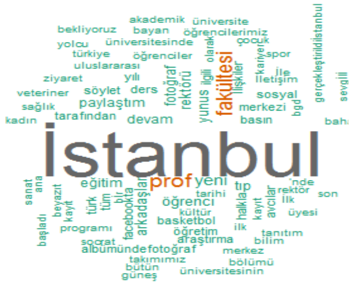


Figure 6. Ankara University

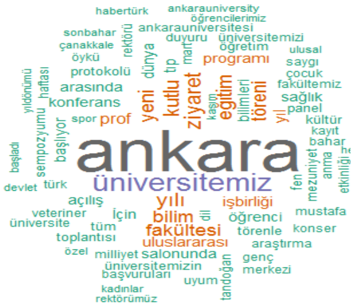


Figure 5. Bilkent University

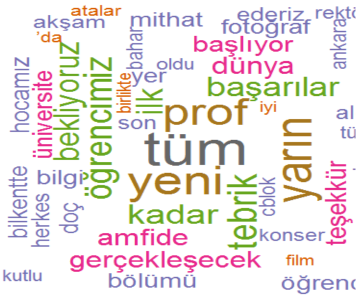


Figure 7. Ege University

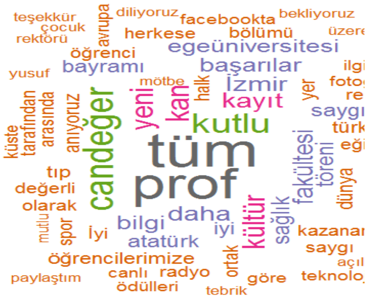


Figure 8. İstanbul Technical University

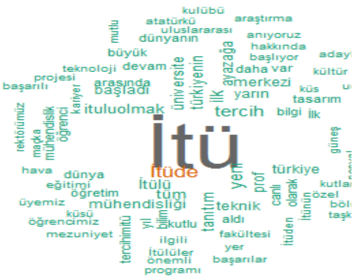


Figure 9. Hacettepe University

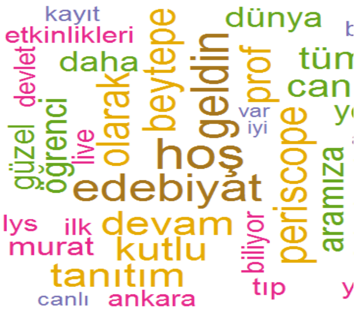


Figure 10. Gazi University

Figure 12. Sabanci University

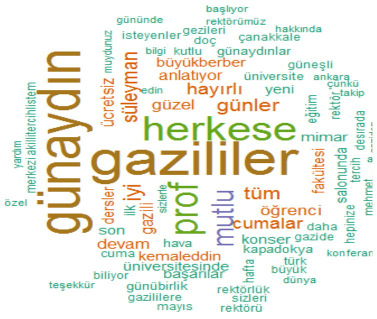


Figure 11. KOC University

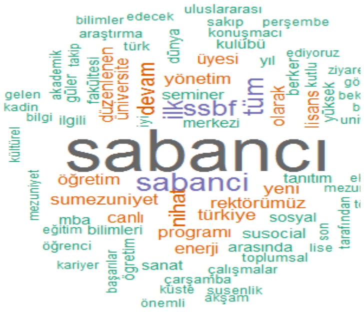
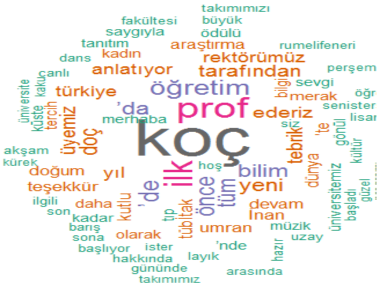
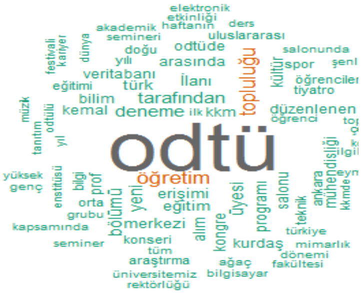


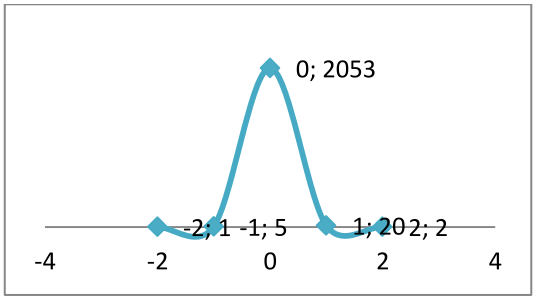
Figure 13. ODTÜ University





Since we use the formal accounts of targeted universities, it is not usual to find emoticons in tweets since they express to somewhat a formal thing. So the sentiment analysis of universities tweets was made by comparing the feature words extracted from tweet text with a prepared list of positive and negative words and then the value for each of the features is calculated. The obtained results show that almost 98% of universities tweets were neural and the remaining 2% is divided between positive and negative. For example for ODTU University, the sentiment analysis result is shown in Figure 14. Most of the obtained results show the same thing which is a neutral feeling.

Figure 14. ODTÜ sentiment anlaysis



While analyzing tweets of international universities, we have noticed an obvious different. Almost all universities tweet text words cloud represent at least one word or phrase related to science. Analysis of those universities is given below.

Figure 15. Yale University

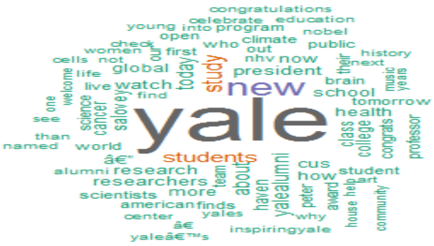


Figure 16. Princeton University

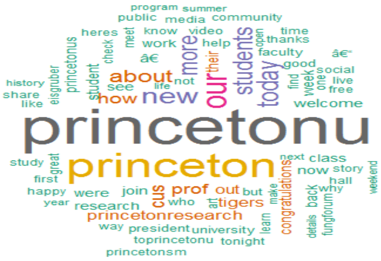


Figure 17. Oxford University

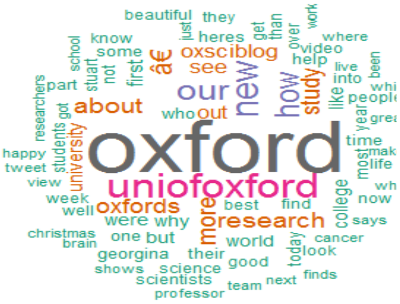


Figure 18. Imperial London

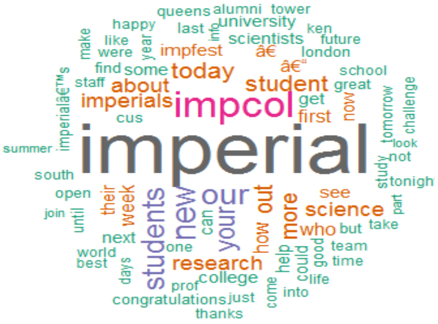


Figure 19. Harvard University

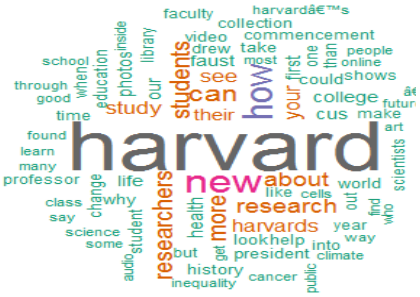
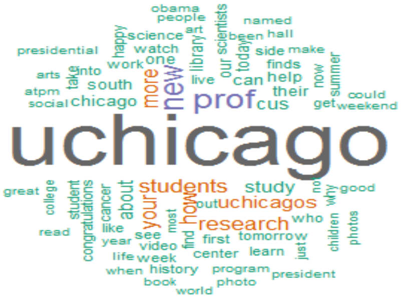
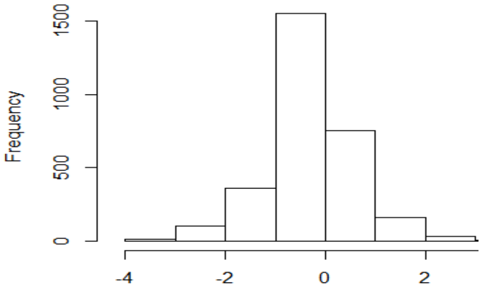


Figure 20 . Chicago University



It is clear that each of the aforementioned international universities write tweets related to schools which represent some kinds of activities that occur in that university. Furthermore, text sentiment of the above universities tweets was different from those in Turkey. For each university the text sentiment analysis shows different feeling ranging from emotional negative to emotional positive. One sample of text sentiment results is given in figure 21.

Figure 21. Harvard University Sentiment Results.



SENTIMENT ANALYSIS OF UNIVERSITIES

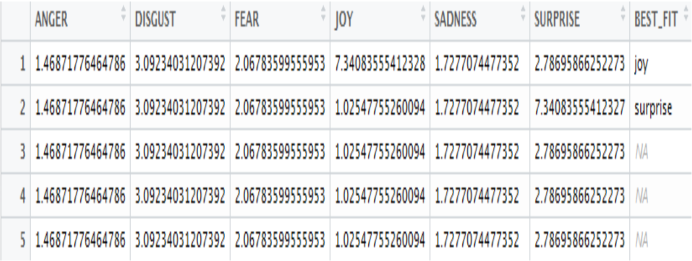
In this section sentiment analysis about what people write as a tweet about the previous mentioned universities. The tweets were collected with the same program [7], but instead of searching for a specific account and gather tweets from that account. We search for specific words and collect related tweets regardless to the writer of them.

At this part, sentiment analysis of user's tweets is much simpler since tweet's text consists of the user's opinion about a specific thing. Thus the ability of finding emoticons or words that reflects the users feeling is very high. Before starting analysis the data was cleaned and preprocessed.

Sentiment package in R studio was used to analyze data. The tweet text would be classified into 7 categories representing emotions. These categories are: anger, disgust, fear, joy, sadness, surprise, and best fit. The classification process is done using naïve Bayes Classifier which is trained on what is called Carlo Strapparava and Alessandro Valitutti's emotions lexicon. The classifier depends on a long list of words in which each word is associated with an emotion. The process passes through the following steps:

- Prepare the tweet text for sentiment analysis by removing all unnecessary words or parts. This includes removing tags, mentions, punctuations, and many others. Later, the obtained words are changed to lower cases.
- For each tweet, emotion classification is done. Thus, for each row a data frame with seven columns will be created. This step is illustrated in Figure 22.

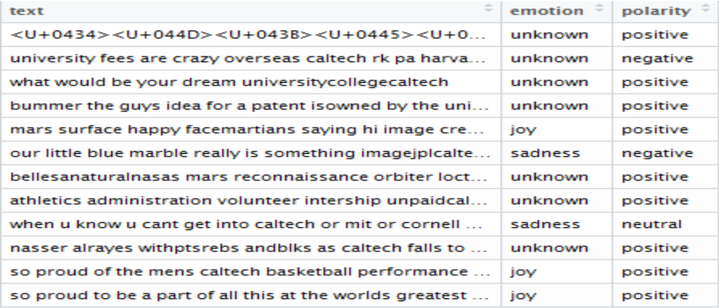
Figure 22 : Emotion Classification for each Tweet



	ANGER	DISGUST	FEAR	JOY	SADNESS	SURPRISE	BEST_FIT
1	1.46871776464786	3.09234031207392	2.06783599555953	7.34083555412328	1.7277074477352	2.78695866252273	joy
2	1.46871776464786	3.09234031207392	2.06783599555953	1.02547755260094	1.7277074477352	7.34083555412327	surprise
3	1.46871776464786	3.09234031207392	2.06783599555953	1.02547755260094	1.7277074477352	2.78695866252273	NA
4	1.46871776464786	3.09234031207392	2.06783599555953	1.02547755260094	1.7277074477352	2.78695866252273	NA
5	1.46871776464786	3.09234031207392	2.06783599555953	1.02547755260094	1.7277074477352	2.78695866252273	NA

- The most suitable emotion for the tweet is located in the seventh location in the data frame (best fit).
- Then a table as shown in Figure 23 is created. This table shows the emotion of the tweet text and the polarity.

Figure 23 . Best Emotion with Polarity for each Tweet



text	emotion	polarity
<U+0434><U+044D><U+043B><U+0445><U+0...	unknown	positive
university fees are crazy overseas caltech rk pa harva...	unknown	negative
what would be your dream universitycollegecaltech	unknown	positive
bummer the guys idea for a patent isowned by the uni...	unknown	positive
mars surface happy facemartians saying hi image cre...	joy	positive
our little blue marble really is something imagejplcalte...	sadness	negative
bellesanaturalnasas mars reconnaissance orbiter loct...	unknown	positive
athletics administration volunteer intership unpaidcal...	unknown	positive
when u know u cant get into caltech or mit or cornell ...	sadness	neutral
nasser alrayes withptsrebs andblks as caltech falls to ...	unknown	positive
so proud of the mens caltech basketball performance ...	joy	positive
so proud to be a part of all this at the worlds greatest ...	joy	positive

Based on these information we have analyzed the data collected from different users about the universities included in this research to know their emotions towards the universities. Samples of the obtained results are shown in the following figures.

Figure 24. Sentiment Analysis of Tweets about Caltech University

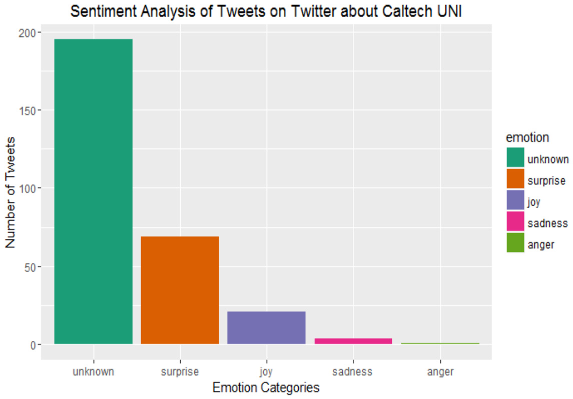
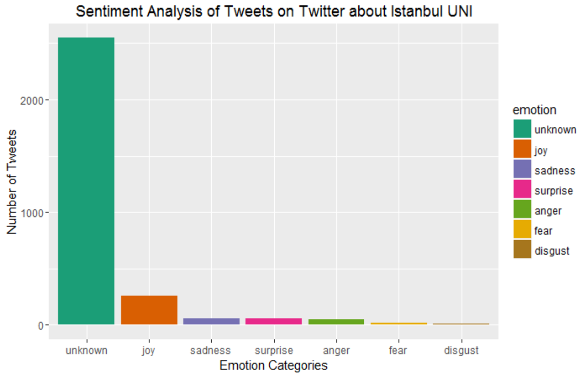


Figure 25. Sentiment Analysis of Tweets about IU.



DISCUSSION AND CONCLUSION

The obtained results have shown many points that have to be taken into consideration. First, while comparing the world cloud for Turkish universities and other universities from all over the worlds, it was clear that the Turkish universities do not pay a lot of attention to sharing their activities, conferences, meetings, or any others. Instead they just use Twitter account just for a platform for posting details about announcements most of the time. Despite the success recorded by many Turkish universities in different academic fields. They did not share such successes with the others through Twitter. On the other hand, the results show that the international universities use Twitter account for sharing their knowledge, research, development. Department, research, price, student, study, science, research and many other words appear in almost all the involved universities. Here, we are emphasizing that Twitter is not just an online platform for posting something. Universities should pay more attention to sharing their success through such websites.

Second, regarding to the result obtained about emotion analysis, we founds that the accuracy of the classifier was higher when classifying tweets written in English about international universities, while the accuracy of the classifier was lower when analyzing tweets in Turkish. The Accuracy is calculated by analyzing some tweets manually and then compare it with the result from the classifier. The accuracy was %68 for those tweets written in English, and 43.17% for those in Turkish. The reason for this could be because of the poor translation of positive and negative word lists. In further studies those list will be translated much more carefully.

REFERENCES

[1] Bing, L. (2012). Sentiment analysis: A fascinating problem. In Sentiment Analysis and Opinion Mining, pages 7–143. Morgan and Claypool Publishers.

[2] Tan, S., Wang, Y., & Cheng, X. (2008, July). Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 743-744). ACM.

[3] Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

[4] Khan, A. Z., Atique, M., & Thakare, V. M. (2015). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 89.

[5] Zhang, H., Gan, W., & Jiang, B. (2014, September). Machine Learning and Lexicon Based Methods for Sentiment Classification: A Survey. In Web Information System and Application Conference (WISA), 2014 11th (pp. 262-265). IEEE.

[6] URAP laboratuarı, 2015-2016 Turkey University Ranking, available online at http://tr.urapcenter.org/2015/2015_t9.php

[7] https://github.com/Jefferson-Henrique/GetOldTweets-java

[8] R Core Team (2015). R: A language and environment for statistical computing. RFoundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[9] Adrian A. Dragulescu (2014). xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7. https://CRAN.R-project.org/package=xlsx

[10] Jeff Gentry (2015). twitteR: R Based Twitter Client. R package version 1.1.9 .https://CRAN.R-project.org/package=twitteR

[11] Timothy P. Jurka (2012). sentiment: Tools for Sentiment Analysis. R package version 0.2. https://CRAN.R-project.org/package=sentiment

[12] Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

[13] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York,2009.

[14] Ian Fellows (2014). wordcloud: Word Clouds. R package version 2.5.https://CRAN.R-project.org/package=wordcloud.

[15] Duncan Temple Lang (2014). RJSONIO: Serialize R objects to JSON, JavaScript Object Notation. R package version 1.3-0. https://CRAN.R-project.org/package=RJSONIO