# A SOCIAL MEDIA BIG DATA MINING FRAMEWORK FOR DETECTING SENTIMENTS IN MULTIPLE LANGUAGES

MUSTAFA COŞKUN

BOĞAZİÇİ UNIVERSITY

2018

# A SOCIAL MEDIA BIG DATA MINING FRAMEWORK FOR DETECTING SENTIMENTS IN MULTIPLE LANGUAGES

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Management Information Systems

by

Mustafa Coşkun

Boğaziçi University

2018

## DECLARATION OF ORIGINALITY

I, Mustafa Coşkun, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date .....May 31, 2018.....

ABSTRACT

A Social Media Big Data Mining Framework

for Detecting Sentiments in Multiple Languages

The popularity of social media platforms has generated a new social interaction environment thus a new collaboration network among individuals. These platforms own tremendous amount of data about users' behaviors and sentiments. One of these platforms is Twitter, which provides researchers data potential of benefit for their studies. Based on Twitter data, in this study a multilingual sentiment detection framework is proposed to compute European Gross National Happiness (GNH). This framework consists of a novel data collection, filtering and sampling method, and multilingual sentiment detection algorithm for social media big data, and tested with nine European countries (United Kingdom, Germany, Sweden, Turkey, Portugal, Netherlands, Italy, France and Spain) and their national languages over six-year period. The reliability of the data is checked with peak/troughs comparison for special days from Wikipedia. The validity is checked with a group of correlation analyses with OECD Life Satisfaction survey reports', currency exchanges, and national stock market time series data. Then, the European GNH map is drawn for six years. Lastly, an exploratory study for determining the relationships between users' Twitter account features (number of tweets, number of followers etc.) and happiness polarities are analyzed. Main aim of this study is to propose a novel multilingual social media sentiment analysis framework for calculating GNH for countries and change the way of OECD type organizations' survey and interview methodology. Also, it is believed that this framework can serve more detailed results (e.g. daily or hourly sentiments of society in different languages).

# ÖZET

## Çok Dilde Duygu Tespiti için
## Bir Sosyal Medya Büyük Veri Madenciliği Çerçevesi

Sosyal medya platformlarının popülaritesi yeni bir sosyal etkileşim ortamı yaratmış ve böylece bireyler arasında yeni bir işbirliği ağı oluşturmuştur. Bu platformlar; kullanıcı davranışları ve duyguları hakkında yoğun miktarda veriye sahiptir. Bu platformlardan biri araştırmacılara çalışmalarında verilerinden yararlanma potansiyeli sunan Twitter'dır. Twitter verilerine dayanarak, bu çalışmada Avrupa Gayri Safi Ülke Mutluluğunu (GSÜM) hesaplamak için çok dilli bir duygu algılama uygulama çerçevesi önerilmiştir. Bu uygulama çerçevesi, yeni bir veri toplama, filtreleme ve örnekleme yöntemini ve sosyal medya büyük verileri için çok dilli bir duygu algılama algoritmasını sunmakta ve 9 Avrupa ülkesinin (İngiltere, Almanya, İsveç, Türkiye, Portekiz, Hollanda, İtalya, Fransa ve İspanya) 6 yıllık verisi ve ulusal dillerinde test edilmiştir. Verilerin güvenilirliği, Wikipedia'daki özel günler için en yüksek/düşük duygu seviyesi karşılaştırmasıyla kontrol edilmiştir. Geçerlilik ise OECD Yaşam Memnuniyeti anket raporları, döviz kurları ve ulusal borsa verilerinin bir grup korelasyon analizi ile kontrol edilmiştir. Sonrasında, 6 yıllık dönem için Avrupa GSÜM haritası çizilmektedir. Son olarak, keşif amaçlı bir çalışma kapsamında, kullanıcıların Twitter hesabı özellikleri (tweet sayısı, takipçi sayısı vb.) ve mutluluk polariteleri arasındaki ilişkiler analiz edilmiştir. Bu çalışmanın temel amacı, ülkeler için GSÜM'yi hesaplamak ve OECD tipi kuruluşların anket ve görüşme yöntemini değiştirmek için yeni bir çok dilli sosyal medya duygu analizi çerçevesi sunmaktır. Ayrıca, bu çerçevenin daha ayrıntılı sonuçlar verebileceğine inanılmaktadır (örneğin, toplumun farklı dillerde günlük ya da saatlik duyguları).

# CURRICULUM VITAE

NAME: Mustafa Coşkun

## DEGREES AWARDED

PhD in Management Information Systems, 2018, Boğaziçi University
MA in Management Information Systems, 2013, Boğaziçi University
BA in Computer Education and Instructional Technologies, 2004, Middle East Technical University

## AREAS OF SPECIAL INTEREST

Social media and big data analysis, mobile and cloud computing, artificial intelligence, instructional technology and educational data mining

## PROFESSIONAL EXPERIENCE

IT Specialist, Author, İzmir National Education Department, 2017 – present
Quality Controller, Middle East Technical University BilgeİS Project, 2016 – present
IT Teacher, Ministry of National Education, 2004 – present
Vice Director, İstanbul Bahcelievler Ş.O.Y. Vocational High School, 2011 – 2013

## AWARDS AND HONORS

Üstün Başarı Belgesi, İzmir Governor, 2018
Best Paper Award, Information and Communication Technologies in Organizations and Society / ISC Paris Business School, 2016
TÜBİTAK PhD Scholarship, 2013-2017
Highest Honors List, Boğaziçi University, 2013
Üstün Başarı Belgesi, Bahçelievler Governer, 2012
Aylıkla Ödüllendirme, Ministry of National Education, 2008
Takdir Belgesi, Havran Governor, 2007
Teşekkür Belgesi, Havran National Education Department, 2007
Honors List, Middle East Technical University, 2004

## GRANTS

Boğaziçi University Research Fund, Grant no. 10600, 2013 – 2018

PUBLICATIONS

*MA Thesis*

Coşkun, M. (2013). A Web Based Multi-Criteria Decision Support System for Department Selection Process of Vocational High School Students, Boğaziçi University.

*Journal Articles*

Coşkun, M., & Ozturan, M. (2018). #europehappinessmap: A Framework for Multi-Lingual Sentiment Analysis via Social Media Big Data (A Twitter Case Study). Information, 9(5), 102. (ESCI-doi:10.3390/info9050102)

Bozanta, A., Coskun, M., Kutlu, B. (2018). Usage factors of location-based social applications: the case of foursquare. *International Journal of Web Based Communities* (Scopus-doi: 10.1504/IJWBC.2018.10010850).

Bozanta, A., Coskun, M., Kutlu, B., Ozturan, M. (2017). Relationship between stock market indices and google trends. *The Online Journal of Science and Technology (TOJSAT), Vol: 7, No: 4, pp. 168-172.* (ISSN: 2146-7390)

Coşkun, M. & Ozturan, M. (2016). A framework for data collection, analysis and evaluation of students' computer interaction in laboratory courses. *FormaMente: Rivista Internazionale Di Ricerca Sul Futuro Digitale*. *51-65.* (ISSN: 1970-7118)

Coşkun, M. & Mardikyan, S. (2016). Predictor factors for actual usage of online evaluation and assessment systems: a structural equation model (SEM) study. *Education and Science, 41*(188). (SSCI-doi: 10.15390/EB.2016.6579)

Coşkun, M. & Ozturan, M. (2016). How do we react@ social media? #catchthemoment. *Yönetim Bilişim Sistemleri Dergisi*, *1*(3), 282-293. (ISSN: 2148-3752)

Kutlu, B., Bozanta, A., Coşkun, M., Dişçi, D. (2016). Foursquare kullanımına etki eden faktörler: Türkiye örneği, ss. 81-93, *Boğaziçi Üniversitesi Matbaası, İstanbul*. (ISBN: 978-975-518-394-7)

Durahim, A. O., & Coşkun, M. (2015). # iamhappybecause: Gross national happiness through twitter analysis and big data. *Technological Forecasting and Social Change*, *99*, 92-105. (SSCI-doi:10.1016/j.techfore.2015.06.035)

Özturan, M., Bozanta, A., Basarir-Ozel, B., Akar, E., & Coşkun, M. (2015). A roadmap for an integrated university information system based on connectivity issues: case of Turkey. *International Journal of Management Science & Technology Information*, (17). (ISSN: 1923-0265)

*Conference Proceedings*

Coskun, M., Hakyemez, T.C., Coşkun, B. (2018). Chess is a social issue more than a two-player strategy dame. *International Multidisciplinary Conference on Education, Arts, Law, Business & Politics (MEALP-18), Amsterdam, Netherlands, 3-4 February 2018* (ISBN:972-91-615-6571-3).

Bozanta, A., Coşkun, M., Bayraktar, B. K. & Ozturan, M. (2016). Relationship between stock prices and google trends. *Proceedings of International Science and Technology Conference, Vienna, Austria, 1023.* (ISSN: 2146-7382)

Bozanta, A., Coşkun, M. & Bayraktar, B. K. (2015). Towards the new trends in recommendation systems for location-based social networks: a survey of variables and algorithms. *2. Ulusal Yönetim Bilişim Sistemleri Kongresi, Erzurum, Turkey – October 2015*

Coşkun, M. & Nasir, V. A. (2014). Cloud computing: a program, a course, a chapter or a topic? *Proceedings of Advances in Business-Related Scientific Research Conference, Venice, Italy* (ISBN: 978-961-6347-53-2)

Coşkun, M. & Badur, B. (2012). Predicting vocational high school students' proficiency exam scores from student's course grades: comparison of the three supervised data mining models. *18th International-Business-Information-Management-Association Conference, MAY 09-10, 2012, pp. 1880-1885. (ISBN: 978-0-9821489-7-6,* WOS*: 000317549801068)*

# ACKNOWLEDGEMENTS

It has been a longer journey than it was supposed to be. What kept me on track was my motivation to share findings with people and be a good model to my daughter. But, I was not alone throughout this hard road. This PhD dissertation was conducted with the contribution and endless support of many people and institutions. It would indeed be churlish of me not to acknowledge them and give my heartfelt thanks.

First and foremost, I would like to thank my thesis advisor, Prof. Meltem Özturan, for all kinds of support she has provided, for her valuable guidance and for her patience. She not only gave me wise counsel but supported and believed in me when those things were sorely needed. I think it was Mina Urgan who said that "a library and a good professor is enough for a perfect PhD study" and I got both from my glorious advisor.

I am also indebted to the members of my thesis supervising committee, Prof. Aslı Sencer and Prof. Adem Karahoca. They always helped me re-define myself for my academic agenda, when I lost my way. I would also like to extend my gratitude to the other jury members, Prof. Birgül Kutlu Bayraktar and Prof. Vahap Tecim, for agreeing to be jury members and for their valuable contributions.

This study was supported by National Scholarship Program for PhD Students (2211-A) of The Scientific and Technological Research Council of Turkey (TÜBİTAK). Thus, I would like to thank TÜBİTAK for supporting my academic studies and my thesis.

I would like to thank faculty (family) members of Boğaziçi University Management Information Systems Department for their priceless contributions to my long journey from the very beginning of MA to the end of the PhD. Throughout my

whole training, those of my brilliant teachers contributed not only to my academic development but also to developing an understanding of work ethic and coping strategies to survive within the academic environment. I must also offer a special mention to my colleagues (friends) in the department, especially to my little sister Dr. Aysun Bozanta, who participated in the study directly and made superhuman efforts and provided enormous friendship.

Big data analysis needs the best hardware and materials, if you intend to conclude universally valuable findings. I also want to thank to the Boğaziçi University Research Fund (10600) for supplying the best equipment for my deep learning analysis containing both social and engineering structures.

As a semi-teacher and semi-student during my MA and PhD studies, I needed help from my colleagues in the schools where I work. I am also grateful to my colleagues in Department of Information Technologies in İstanbul Bahçelievler Ş.O.Y. Vocational School, İzmir Gültepe Secondary School and İzmir National Education Department for their precious support and friendship, without whose sympathy I would not have been able to finish my study.

Nothing would be possible without the unconditional support and love of my family. First, I am deeply grateful to my mother, Gönül Coşkun, for all her love and support and putting me through the best education possible. Also, special thanks to my father, Oğuz Coşkun, as I hope, for watching and being proud of me from heaven. I also thank to my sister, Eda Cora, for loving me with unconditional positive regard. And my sister, Bahar Karateke, and brothers, Serhat Cora and Murat Karateke, who I had later in my life, gratefully; I thank them for providing full support for me and believing me. Additionally, I would like to thank my second parents, Nurcan and Mehmet İnce, for growing my love and trusting me.

And, of course, to my daughter, Nehir Ela… I hope you read these words whenever you need encouragement to find your best way, because I have done every good thing in my life in order to be a good role model for you.

Finally, to Bircan, my wife – my best friend, my companion on every step of this exhausting journey – thank you. She listened to countless ideas about software programming, algorithms, big data mining etc. which are completely out of her academic area. She, at all times, found a way to bury the inapplicable ones gently and made me recognize new good ones. Whatever is good in this dissertation is due in great measure to her.

*I dedicate this dissertation to my endless love, Bircan.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER 1

# LITERATURE REVIEW

Peace[2] at home,
peace[2] in the world.
[+5, -1]
M. Kemal Atatürk

The rise of social publishing technologies lead to open data access for researchers. Today, several social media platforms let researchers to gather valuable public data for free and to conduct their studies based on those data (Fuchs, 2017; Hanna, Kee, & Robertson, 2016). Also, social media usage has diffused widely in societies with fresh statistical data showing high penetration rates (Bello-Orgaz, Jung, & Camacho, 2016). As it is mentioned by Quan-Haase and Young (2010) users have a tendency to hold new media and adopt them as a part of their communication repertoire. To some degree this is an advantage at the current stage of studying social media, as it leaves much room for exploring approaches to address research questions (Ellison, Steinfield, & Lampe, 2007; Mayr & Weller, 2017).

These studies are inspiring about the social media and big data analysis concepts, then, to this respect, a deep literature review was done on social media and big data analysis concepts at the beginning of this study. First of all, common and popular social media platforms and the studies about them were investigated. Then, chronological trending topics analysis was done on them.

## 1.1 Popular social media platforms, studies and trending topics

Facebook, Instagram and Twitter are the most popular ones of the social media in the World Wide Web. Thus, investigating these studies in the literature was the first

stage of this study to gather common and popular social media studies and trending topics among academics.

### 1.1.1 Facebook

Facebook, as the most popular social media platform (Greenwood, Perrin, & Duggan, 2016), let users publish their events, feelings and actions etc., via text, image, and video messages. This social media platform is available for tracking web browsing and users' browsing histories, monitoring social media activities and so on (Fuchs, 2017). Thus, there appears thousands of studies (in social sciences, engineering etc. topics) among this platform in the academic databases.

For instance, Menon (2012) studied Facebook's big data structure and how they put up with big data. Additionally, Thusoo et al. (2009) examined warehousing problem of big data on Facebook and proposed a simple "map/reduce" framework. These early studies were followed with several big data studies (Boyd & Crawford, 2012; John Walker, 2014; McAfee & Brynjolfsson, 2012; Wu, Zhu, Wu, & Ding, 2014) of Facebook data. Today, coping with Facebook big data and its complexness (Brandtzaeg, 2017; Godwin-Jones, 2017; Wells & Thorson, 2017), mining for social analysis (Gil de Zúñiga & Diehl, 2017; Zook et al., 2017), capturing value from big data (Hartmann et al., 2016), concerning ethical issues (Mittelstadt & Floridi, 2016) and critical thinking on validity of big data on Facebook (Panger, 2016) are still being considered deeply.

Sentiment detection is also one of the most popular topics among social media studies performed using Facebook data with different sentiment detection algorithms and software. For instance, Ahkter and Soria (2010) analyzed the Facebook status posts to make a classification of users by the help of Neuro-

Linguistic Programming (NLP) information mining. Similarly, Cvijikj and Michahelles (2011) examined the user generated comments of Facebook branding pages for designing an opinion mining with their proposed sentiment detection algorithm. As the recent popular studies; Meire, Ballings, and Van den Poel (2016) analyzed 17,697 Facebook status updates to evaluate the added value of leading and lagging information in sentiment analysis and Tian, Galery, Dulcinati, Molimpakis, and Sun (2017) proposed that Facebook Reactions are a good data source for indicating the overall sentiment of the entire message as well as the sentiment of the emoji. The data mining methods (called differently such as models, algorithms etc.) mainly used on Facebook studies can be listed as natural language processing (Kumar et al., 2016; Ortigosa, Martín, & Carro, 2014; Trinh, Nguyen, Vo, & Do, 2016), lexicon-based analysis (Ngoc & Yoo, 2014; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) and classification (Akaichi, Dhouioui, & Pérez, 2013; Hamouda & Akaichi, 2013; Terrana, Augello, & Pilato, 2014; Troussas, Virvou, Espinosa, Llaguno, & Caro, 2013).

### 1.1.2 Instagram

Instagram is also a popular social media platform, which was launched in October 2010 and in which people publish photos and followers of them comment on the images (Hu, Manikonda, & Kambhampati, 2014). Additionally, users can share their photos on the other social media platforms by connecting the Instagram accounts of them to their accounts on other social media sites. Since it is very new comparing to other popular social media platforms number of studies is relatively fewer.

The early studies on this platform mainly worked on the difference of it, how to use it and social movement from other platforms to this new media (Frommer,

2010; Salomon, 2013; Systrom, 2010). Then, the advantages of this platform on commercial and business purposes (Linaschke, 2011), such as consumer production (McCune & Thompson, 2011), shaping organizational image-power (McNely, 2012) and e-marketing (called insta-marketing) were studied. Afterwards, literature shows that, social and cultural studies about Instagram platform were published. Lee, Lee, Moon, and Sung (2015) studied on usage of pictures instead of words and motives on this manner. Also, Sheldon and Bryant (2016) stated motives for its usage and examined the relationship to narcissism. Additionally, Manikonda, Hu, and Kambhampati (2014) stated a quantitative study about user activities, demographic information, social media structure and user-generated content.

At the same time, sentiment analysis started to be popular on this platform. Gunawardena, Plumb, Xiao, and Zhang (2013) studied the Instagram hashtags with another sentiment discovery algorithm to categorize human judgments by the help of Naive Bayes classifiers. Moreover, Silva, de Melo, Almeida, Salles, and Loureiro (2013) made a mining Instagram's published images with pattern recognition and showed that opinion detection is not only done by text analysis but can also be achieved by other multimedia such as photos. Thus, Y. Wang, Wang, Tang, Liu, and Li (2015) proposed a unique unsupervised sentiment analysis (USEA) framework for social media photos. In addition, Ranaweera and Rajapakse (2016) discovered the tourist perceptions locally about Sri Lanka. Thus, sentiment analysis software and frameworks for image processing begin to be an emerging topic in the literature with newly proposed tools and methods (AbdelFattah, Galal, Hassan, Elzanfaly, & Tallent, 2017; Katsurai & Satoh, 2016).

### 1.1.3 Google+

Google's first social networking platform was Orkut, which was launched in 2004 and, quickly became to be popular only in Brazil (Kugel, 2006). The next platform was called Buzz, in which service took data from users' Gmail contacts without their consent, and this progress made followers intrusive. Google's struggles about social media projects was summarized by Pariser (2011) that Facebook is good at managing people's relationships, Google is good at managing information relationships. But, when Facebook and Twitter dominate the social media market, Google+ was introduced in June 2011 and reached a significant growth (Gonzalez, Cuevas, Motamedi, Rejaie, & Cuevas, 2013).

Although the number of studies is the least on this platform, because of its user number size is very high, especially from South American countries, user generated content analysis studies are found in the literature on this platform. For instance, (Russell, 2013) indicated the difference of Google+ on mining social media concept in his book comparing this platform with other popular media. In addition, Messias, Magno, Benevenuto, Veloso, and Almeida (2015) focused on migrating trend of this new platform and used support vector machine (SVM) technique for investigating which features of Brazilian users are relevant to classify them as a possible emigrant. User characterization has become very popular analysis method for this platform and literature includes (Casas et al., 2014; Cunha et al., 2014; Cunha, Magno, Gonçalves, Cambraia, & Almeida, 2013; Dumba, Golnari, & Zhang, 2016) several studies on this issue.

There are also sentiment analysis studies in which the user generated content of this platform is compared to other platforms (Heimbach, Schiller, Strufe, & Hinz,

2015; Kharde & Sonawane, 2016). On the other hand, to the best of our knowledge, there is not a specific sentiment analysis study on this platform.

### 1.1.4 Twitter

In addition to those social media platforms examined in the previous sections, it can be clearly specified that Twitter is the most accessible data source for the social media researchers with its related Application Programming Interfaces (APIs) which make it easy to gather data from that platform. As an early study on Twitter, Java, Song, Finin, and Tseng (2007) showed their observations on the microblogging phenomena by exploring the geographical and topological features of Twitter's social network. Over the past decade, significant development has been made on sentiment analysis techniques that extract predictors of public mood from social media content. For instance, Mishne and Glance (2006) tried to predict movie sales by a blogger sentiment. Gilbert et al. (2010), Gilbert and Karahalios (2010) and Y. Liu, Huang, An, and Yu (2007) stated their own sentiment detection algorithms for estimating stock market changes.

As the recent local works in Turkey, Ertugrul, Onal, and Acarturk (2017) explored the effect of regression usage on confidence scores for sentiment analysis of Turkish tweets and Ayata, Saraçlar, and Özgür (2017) worked on four different sector tweet datasets to compare word embedding model, SVM and random forests classifiers.

The results showed that although 50% of the users of Twitter are from the Asia Pacific region, when the ratios of users to country populations are examined, Twitter is in fact mostly used in European and North American countries (Statista, 2018). Also, it has been the most accessible data source for the social media

researchers because of its related Application Programming Interfaces (APIs) that lead an easy data collection progress. After discussing social media studies, a common trend on user generated content analysis was detected and also was seen that sentiment analysis on these platforms is an emerging topic. Thus, Twitter sentiment analysis studies in the literature were deeply analyzed.

## 1.2 Twitter sentiment analysis studies

After discussing social media studies, a common trend on user generated content analysis was detected. Also, it can be concluded that sentiment analysis on these platforms is an emerging topic. In addition, it is obvious that Twitter is the most popular platform for sentiment analysis with its (mostly) free accessible structure to the user data. Thus, Twitter sentiment analysis studies in the literature were deeply analyzed. There are two groups of studies in the Literature based on Twitter sentiment analysis. The ones in the first group are based on user analysis (classification, clustering etc.) via their tweets (see Table 1) and the other group studies are based on tweets (idioms, emoticons wording etc.) to analyze sentiments of the communities (see Table 2). These studies are filtered from the deep literature analysis on Twitter sentiment analysis subject in different databases.

Table 1. User-Based Twitter Sentiment Analysis Studies, Details and Possible Contributions to This Study

| Source | Empirical Approach | Details and Possible Contributions |
|---|---|---|
| User-level sentiment analysis incorporating social networks (Tan et al., 2011) | Are the same type of users publish same tweets? User connection network was investigated. | Focused on user-level rather than tweet level sentiment calculation. Includes a new user data collection methodology. |
| Characterizing Geographic Variation in Well-Being Using Tweets. (Schwartz et al., 2013) | 1293 counties of USA and 1-year period tweets were analyzed. | Only done in English. Found own words which are frequent like "bored", "tired" etc. But a limited number of words were used. Geolocation codes of all tweets could not be found, instead self-expression of users was relied on. Results compared with the survey results and other data of USA government (socio economic level of counties, income, education level etc.) |
| Tracking gross community happiness from tweets. (Quercia, Ellis, Capra, & Crowcroft, 2012) | Study concluded socio economic results and tweets well-being are correlated. | 1-month period data was used for London and in English LICW dictionary. Only smile and only words were used but not both together. It is concluded that 2300 words dictionary is enough for calculating GNH because %80 of everyday language constitutes of those words. |
| Do all birds tweet the same?: characterizing twitter around the world. (Poblete, Garcia, Mendoza, & Jaimes, 2011) | To make a deep understanding on differences between countries and supports this understanding can be useful in many ways. | 2 languages were used (English & Spanish) 10 countries, USA, Brazil, Indonesia, Mexico, and South Korea etc. were analyzed with one-year period data. Clustered users and found their tweets and make sentiment analysis. Concluded average sentiment levels but not standardized them. |
| Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. (Baucom, Sanjari, Liu, & Chen, 2013) | 12 NBA games were analyzed with tweets. Geolocation coded tweets were used with limited number. | Used most frequent words for classification. Just positive and negative happiness was discussed not neutral (all tweets are classified). |
| Sentiment-based User Profiles in Microblogging Platforms. (Gutierrez & Poblete, 2015) | Users were classified in terms of sentiment behaviors. | 36000 users were analyzed in English. Users were not classified by their features SentiStrength dictionary (Thelwall et al., 2010) was used. It is proved that this dictionary (English) is better than others |
| World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans' tweets. (Yu & Wang, 2015) | Used sentiment analysis to explore U.S. football fans' emotional responses from tweets, e.g. the emotional changes after goals of their teams. Results showed that sports fans use negative emotions during matches. | Claims that big data analysis cannot be done for dictionary-based sentiment analysis. Just analyzed a period of time (cup 2014) and just for US soccer team fans. Geo-located tweets for the first game could not be collected. NRC software was used. 13 million tweets (7 million original and 6 million retweet) were analyzed. Emoticons were used. |
| Twitter verileri ile duygu analizi. (Akgül, Ertano, & Banu, 2016) | Lexicon and n-grams were used for analysis and found out that lexicon has more performance. | 7000 tweets were analyzed in Turkish with a unique dictionary. |

Table 2. Tweets-Based Twitter Sentiment Analysis Studies, Details and Possible Contributions to This Study

| Source | Empirical Approach | Details and Possible Contributions |
|---|---|---|
| Collective smile: Measuring societal happiness from geolocated images. (Abdullah, Murnane, Costa, & Choudhury, 2015) | Derived a smile index with using smiles in the tweets of user. Started with 16 months and 9 million geolocated tweets and filtered them. | Only focused on smiles. Language Inquiry Word Count (LIWC) was used. Because of non-English words focused on smiles. Concluded that Gross National Happiness term is essential and used by governments but there is not a cross study among nations. |
| From unlabeled tweets to Twitter-specific opinion words. (Bravo-Marquez, Frank, & Pfahringer, 2015) | Clustered tweets without any dictionary and found frequent words Validated with SentiStrength (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) software. | Tweet based analysis was used, users were not included. "Index" term was used for analyzing one country and English language. Concluded that human generated dictionaries are good at sentiment analysis |
| Talk of the city: Our tweets, our community happiness. (Quercia et al., 2012) | Clustered the topics of tweets by using frequent words. Monitored subject matter of tweets. | Examined 573 twitter profiles, taken another study's (Cheng, Caverlee, Lee, & Sui, 2011) data, used English Language and analyzed 200 thousand tweets. Found a relation between topics and socio-economic levels. |
| Role of emoticons for multidimensional sentiment analysis of twitter. (Yamamoto, Kumamoto, & Nadamoto, 2014) | Classified the emoticons and so calculated the sentiments with this. | Japanese tweets with limited number were used. Insisted on the usage of emoticons on sentiment analysis. |
| Emotional states vs. emotional words in social media. (Beasley & Mason, 2015) | Focused on dictionary-based tools and suggest that it may not be sufficient to infer how people feel. | LIWC dictionary in English language was used. Not only Twitter (448) but also Facebook (515) users were examined. |
| Contextual semantics for sentiment analysis of Twitter. (Saif, He, Fernandez, & Alani, 2016) | Presented SentiCircles, a lexicon-based approach for sentiment analysis on Twitter. | This software tried to learn from dataset and changes the strength of words. A new way of lexicon-based sentiment analysis but for tweet-level sentiment detection. |
| Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. (Rodrigues Barbosa et al., 2012) | Effect of hashtags on sentiment analysis were analyzed. | Data was collected on Brazilian president election in 2010. 10 million tweets were manually classified with 10000 hashtags. It was claimed that while hashtags can identify election feelings it is better to analyze tweets with sentiment analysis big data. |
| A polarity analysis framework for Twitter messages. (Lima, de Castro, & Corchado, 2015) | Five different types of classifiers were considered: Naïve Bayes (NB), SVM, Decision Trees (J48), and Nearest Neighbors (KNN). | Done in English language only. Emoticons were not used and %80 accuracy gathered. |

The results showed that Twitter data is frequently used for sentiment analysis and researchers have a tendency to categorize the analyzed users into sentiment groups. Specifically, the following ideas for this study can be summarized from the literature review on this topic and details listed in Table 1 and Table 2:

- Dictionary-based text analysis is the most common and approved method.
- Embedding emoticons to the analysis helps to improve results' quality.
- Geo-location usage is a conflict for these kinds of analyses.
- Standardization and filtering is the main problem for generalizability issues of the results. This problem should be focused most.
- Data collection methods were not mentioned in the studies. Generally, researchers mention the collected data but not how they collected them.
- Generally English language used in the studies. Also, if English is not used, researcher use only one dictionary and analyze one language for one country.

### 1.3 Cultural well-being and life satisfaction studies

Since the sentiment analysis on social media has been found a trending topic among information systems and cyber-psychology researchers (C.-J. Wang, Wang, & Zhu, 2013), this study aims to focus on this topic. However, analyzing public sentiments is not a new methodology for scientist. Thus, a deep literature review was done on cultural well-being and life satisfaction studies to find out general tendency and possible contribution areas. Table 3 (adapted from study of Exton, Smith, and Vandendriessche (2015)) shows the filtered studies and details of these studies that noted down during this literature review.

Table 3. Cultural Well-Being and Life Satisfaction Studies, Details and Possible Contributions to This Study

| Source | Empirical Approach | Details and Possible Contributions |
|---|---|---|
| The French unhappiness puzzle: The cultural dimension of happiness. (Senik, 2014) | Survey study was conducted for life evaluation. 0-10 scale for happiness index was used. | 7 wealthy European countries, each with >15% migrants in the sample. Comparison of migrant and native experiences, controlling for a range of life circumstances. |
| Subjective well-being and culture across time and space. (Rice & Steele, 2004) | Survey study was conducted. 1-4 scale for happiness index was used. | People living in US from 20 (mainly European) different nations are analyzed. |
| Nations with More Dialectical Selves Exhibit Lower Polarization in Life Quality Judgments. (Minkov, 2009) | Survey study was conducted. 1-4 scale for happiness index and 1-10 for life satisfaction was used. | Happiness of 90 countries and life satisfaction of 82 countries are drawn. None of the cultural values variables tested contributed significantly to the prediction of life satisfaction. |
| International Evidence on the Social Context of Well-Being. (Helliwell, Barrington-Leigh, Harris, & Huang, 2009) | Survey study was conducted. 0-10 scale for happiness index was used. | Applied a huge sample, 125 countries. Using regression model, country fixed effects predicted. |
| Do Danes and Italians Rate Life Satisfaction in the Same Way? (Angelini, Cavapozzi, Corazzini, & Paccagnella, 2014) | Survey study was conducted. "How satisfied are you with your life in general?" 5-point scale. | 10 European countries. Exploring vignettes for adjusting individual differences in response ratios. |
| Comparing Happiness across the World: Does Culture Matter? (Exton et al., 2015) | Reports reviews "the main barriers to interpreting national differences in subjective well-being, noting the challenge of distinguishing between cultural bias and cultural impact". | Reasons for intending to compare happiness across countries are explored. Draws the OECD picture. It is extremely stated in the report that: "Measurement error is a fact of life in survey data" and "Comparability of survey data across countries, however, requires more than just a common methodological framework". |
| Cultural dimensions in management and planning. (Geert Hofstede, 1984) | Explains the scope of cultural differences in more than 50 countries. Discusses how those differences effect the management validity techniques. | Discusses the technical side is less culture-dependent than human side. Concludes that planning is a part of management and cultural differences is functional. |
| Cultures and organizations: Software of the mind (Vol. 2). (G Hofstede, Hofstede, & Minkov, 1991) | Explained detailly the cultural differences between countries from different dimensions. | Discusses the key conceptions of intercultural dynamics Supply the insightful advice for organizations and individuals. |
| A cross-media sentiment analytics platform for microblog. (Chen, Chen, Cao, & Ji, 2015) | Argued public sentiment analysis system is necessary under such a circumstance. | Used Chinese Weibo as the data gathering platform. Used 5 different algorithms for sentiment analysis in only one language and for one country. Visualized the result for better understanding. |

As a result of the analysis of cultural well-being, cultural differences and life satisfaction studies, it can be concluded that this kind of studies are generally made with survey (Diener et al., 2000; Helliwell et al., 2009; Minkov, 2009; Rice & Steele, 2004), interview and these types of qualitative methods (Angelini et al., 2014; Diener et al., 2000; Exton et al., 2015). On the other hand, social media sentiment analysis studies have a tendency to be popular in the literature. Combining these two types of studies, it is believed that, a multicultural and multidimensional study, which compares and contrasts the cultural and lingual differences, would contribute to information systems and social sciences literature. Moreover, if a novel methodology for data collection and analysis can be conducted, this framework can be useful for not only social media analysis but also for researchers working on social media.

## 1.4 Twitter user characteristics analysis studies

Another concept on the literature is social media user characteristics analysis. While several different features can be derived, main Twitter user profile features which can be accessed by appropriate data collection APIs can be listed as:

- Number of followers
- Number of friends (followees)
- Number of tweets
- Number of "favorited" tweets
- Date of account creation
- Screen name
- Account description

Also, tweets have their own features as:

- Number of “favorited” users
- Date of tweet published
- Language of tweet
- Number of “retweeted” users
- User of tweet (publisher)

Since Twitter is the most appropriate and popular data source for social media analysis and has a better understanding of consumers; it is an obligation for organizations to examine Twitter user profiles. For this reason, it would be helpful to conduct a deep literature review on analysis of Twitter users’ characteristics for this study. The filtered studies from this review, their empirical approaches and details together with their possible contributions to this study are listed in Table 4.

As a result of this part of literature review, it can be concluded that investigating the relationship between the user characteristics and sentiments on social media is totally a very new concept. Therefore, a social media sentiment analysis study should focus not only on finding out the sentiments of users but also on exploring the possible relationships of Twitter user characteristics and sentiments.

Table 4. Twitter User Profile Analysis Studies

| Source | Empirical Approach | Details and Possible Contributions |
|---|---|---|
| Inferring nationalities of Twitter users and studying inter-national linking. (W. Huang, Weber, & Vieweg, 2014) | Tried to find nationalities of users from user profile features. | Sample size is 400 users. One language (English) is used. Conducted in Qatar (since it is a multinational region). Claimed that user profile features include valuable data. |
| Analyzing and predicting viral tweets. (Jenders, Kasneci, & Naumann, 2013) | Tried to make an API to detect whether a tweet is viral or not. | Focused on content analysis in German. Analyzed user profiles in terms of mentions retweets etc. Derived a prediction method. Resulted that user profile info is valuable for detecting their characteristics. Not analyzed the interconnection of these features. |
| A Machine Learning Approach to Twitter User Classification. (Pennacchiotti & Popescu, 2011) | Tried to detect political orientation, ethnicity and network structure for particular business by leveraging observable information. | Focused on user attributes such as networking, how you tweet and where you tweet. Classified tweets in terms of used words. Concluded twitter features are enough for finding user behaviors. |
| Classifying latent user attributes in twitter. (Rao, Yarowsky, Shreevats, & Gupta, 2010) | Tried to detect age, gender, regional origin and political orientation from tweets. | Derived a classification model with SVM. At the conclusion part, a user classification was advised with possibly related features of Twitter. Also, it is concluded that language usage is important for user classification analysis |
| Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter. (Lahuerta-Otero & Cordero-Gutiérrez, 2016) | Objective of the study is to explore influencers on Twitter and examine the characteristics of tweets. | Tried to find the popular people (business case) 3853 users posting on Japanese car firms, Toyota and Nissan, were analyzed with 30000 tweets from 13th to 25 April 2015 Characteristics of influencers are defined as "They clearly define their feelings and opinions (either positive or negative) when tweeting, Influencers have on average number of people they follow etc." |
| Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. (Luo, Cao, Mulligan, & Li, 2016) | Investigated the demographic information of users and compared with human mobility measures. | 7967 Chicago users were analyzed. Tried to find age, gender, name and surname mobility differences. Concluded as a feature study recommendation to use Twitter characteristics for this geography study. |

## 1.5 Ethics on social media studies

The last part of the literature review is about ethical issues. Actually, the ethical aspects of using social media data for researches are still not clearly defined even though the structure of the data is publicly available. Therefore, it is not surprising to see a wide range of methodologies to cope with ethical constraints from the literature among the studies.

Related to ethical issues, it is seen that six studies (Braithwaite, Giraud-Carrier, West, Barnes, & Hanson, 2016; Coppersmith, Dredze, Harman, & Hollingshead, 2015; Guan, Hao, Cheng, Yip, & Zhu, 2015; Lv, Li, Liu, & Zhu, 2015; O'Dea et al., 2015; Tsugawa et al., 2015) were approved by their authors' corresponding Institutional Review Boards (IRB) and five (De Choudhury, Counts, Horvitz, & Hoff, 2014; Guan et al., 2015; P. Liu, Tov, Kosinski, Stillwell, & Qiu, 2015; Park et al., 2015; Park, Lee, Kwak, Cha, & Jeong, 2013) declared receiving requested consent from participants prior to data analysis. However, Youyou, Kosinski, and Stillwell (2015) stated that IRB approval is not essential for using the data. Similarly, in their study, Chancellor, Lin, Goodman, Zerwas, and De Choudhury (2016) did not collect IRB approval, since that manuscript used Instagram data which does not contain personal data. In addition, several studies (Burnap, Colombo, & Scourfield, 2015; Coppersmith, Ngo, Leary, & Wood, 2016; Harman & Dredze, 2014; X. Huang et al., 2014; O'Dea et al., 2015) showed that, anonymizing user profile data is another practical method for ethical constraints of social media studies. For instance, changing the names and usernames in tweets with other texts is a method of anonymization (Coppersmith et al., 2016; Ertugrul et al., 2017; Harman & Dredze, 2014). Based on these studies, it can be stated that, there is still not a common approach for handling ethical issues among researchers.

It is obvious that this issue is still unstable and fuzzy (Wongkoblap, Vadillo, & Curcin, 2017) but in order to contribute social media studies and science itself, public data should be used for scientific manner. Thus, the best method for decreasing the potential of ethical problems and using personal information is to anonymize the collected datasets and filtering the result to state a common (public) result (e.g. Gross National Happiness – GNH). Wilkinson and Thelwall (2011) suggested that academics must not quote messages directly or use the public URLs of messages, because they would be used to find the users who publish them.

## 1.6 Results of literature review

To conclude, a deep literature review was conducted in this study to discover trending topics, possible contributions of the related studies and their future recommendations to form a basis for the research of this study. Therefore, to the best of author' knowledge, it can be concluded that;

- There is not a multi-lingual framework for Twitter sentiment analysis.
- Lexicon-based (dictionary-based) sentiment analysis is still most popular instead of machine learning, classification and clustering.
- Multicultural comparison of social media data on sentiment analysis has not been done yet.
- Data collection is the least mentioned part in articles, while proposing a novel method for this issue can be very supportive for the academics.
- The user Twitter features such as follower count, friends count, Twitter age, number of Tweets have not been taken into account yet in terms of possible relations with user sentiments.

- Whereas business effect and value are mentioned in several studies, the result of a multicultural sentiment analysis and GNH map of a continent have not been considered by the researchers yet.
- Some of the dictionaries aimed to be used in this study are mentioned in some studies but have not been used all together yet (possibly because of huge work requirement).
- Big data studies are becoming very popular on sentiment analysis but have not been defined well yet.
- English is very popular and people usually use LICW dictionary, but, except for a few local small-scale studies, other languages have not been examined with big data analysis to detect sentiments.
- Validation and accuracy of findings is not a concept for sentiment analysis studies while it should be.
- Anonymizing users' information, converting the info with other texts and filtering out results to conclude a general result are the frequent methods for ethical consideration on social media studies.

# CHAPTER 2

# RESEARCH QUESTIONS AND SCIENTIFIC VALUE

Appreciation[1] is a wonderful[2] thing;
It makes what is excellent[1] in others belong to us.
[+5,-1]
Voltaire

In this chapter, the main research questions which inspire this study are shown with objectives and detailed explanations.

## 2.1 Main research questions

Throughout a deep literature review to find out trending topics on information systems and social media studies, "big data" concept was appeared as very popular among scholars. But the literature constitutes of local and very limited studies due to the fact that handling with big data for social analysis is not easy. On the other hand, Twitter is the most appropriate data platform for the social media analysis and for applying current data mining techniques for sentiment analysis which is popular but a very new area of science. Generally, in physiological sciences and public institutes sentiment analysis reports are tried to be prepared with survey and interview methods. Therefore, the first main research question stated for this study is:

"Is the Twitter social media big data appropriate for the sentiment analysis (instead of surveys or interviews) to draw a happiness map of Europe?"

Afterwards, literature showed that the user categorization and classification for specific aims (e.g. retweeting analysis, election result prediction, popularity examination, and cultural tendency investigation) are also a newly interested topic. On the other hand, there is not a social media model driven by big data in

multicultural and multilingual domain in the literature yet. Thus, the second main research question stated for this study is:

"Is the Twitter social media big data appropriate for building a Twitter Social Media Model with social media variables where <happiness> is the dependent variable?"

## 2.2 Objectives

After stating the main research questions of the study, the objectives were uttered. Since the data collection methodology is the least mentioned and touched part of big data studies in the social media researches, it is believed that a novel data collection methodology can be drawn for further studies and researchers. Thus, combining the first main research question with "designing a Twitter social media big data sentiment analysis algorithm", following objective was stated:

"To design, develop, implement and evaluate a framework for multi-lingual sentiment analysis via Twitter social media big data for calculating Gross National Happiness (GNH) levels of European Countries"

Furthermore, the second research question lead the second objective for the study connected with the results of first objective. The second objective of the study is:

"To build a new Twitter Social Media Model consisting of main social media variables where <happiness> is the dependent variable"

## 2.3 Detailed research questions

As it is mentioned before, the purposes of this research are to test a new sentiment analysis framework (sentiment detection algorithm, social media data collection and filtering methodology) by examining tweets of users from European countries (with different languages) to derive a happiness map and to build a Twitter social media

model by examining relationships between Twitter user characteristics and their sentiments. For the first purpose, the stated sentiment analysis framework will be implemented for determining the happiness polarities of European citizens through answering the following four research questions:

- Is there face validity when the polarities determined by sentiment analysis framework are compared with Exchange Rates and Stock Market Index?
- Is there convergent validity when the GNH results of the sentiment analysis framework and GNH survey results of Organization for Economic Cooperation and Development (OECD) report are compared?
- Is there data reliability when the peaks/troughs of the graphs of sentiment analysis framework are compared with specific dates obtained from news archives?
- What are the GNH polarities of European countries according to the proposed Twitter sentiment analysis framework?

The first two questions are about the validation and third question is about the reliability of sentiment detection algorithm and social media data collection and filtering methodology (social media data analysis framework at all). The fourth question is for exploring the GNH polarities of related countries as a result of the proposed algorithm.

Related to the second purpose of the study, after creating a "happiness" variable via Twitter social media big data, the following research question will be examined.

- What are the relationships between users' Twitter account characteristics (e.g. number of tweets, friends' count) and their happiness levels?

Figure 1 summarizes the last research question where “happiness” is the dependent variable and independent variables are user characteristics.

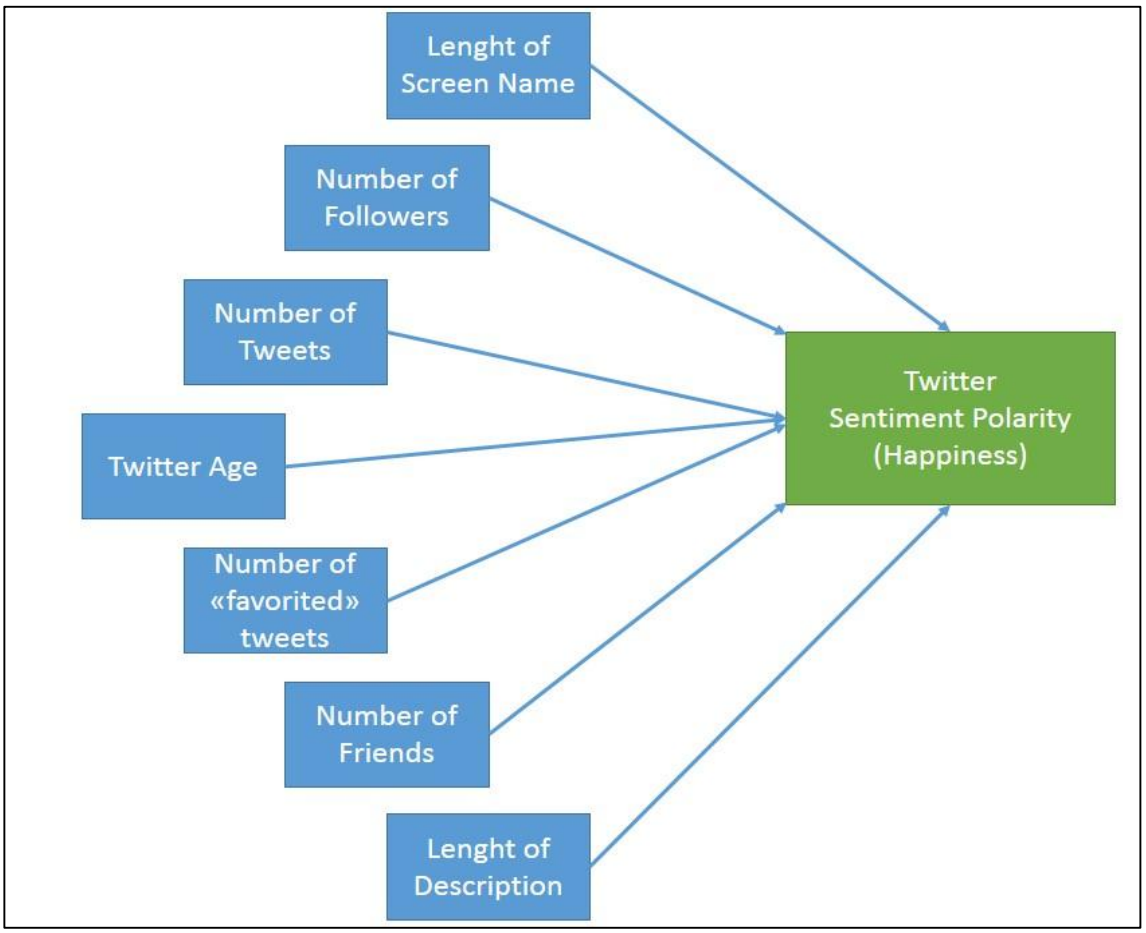



Figure 1. Social media user characteristics and happiness relationships

# CHAPTER 3

# MATERIALS AND METHODOLOGY

Love[2] comforteth[1] like sunshine[1]
after[»] rain[-1].
[+5,-3]
William Shakespeare

Ferraro (2007) defines the term "framework" as; "in general, a framework is a real or conceptual structure intended to serve as a support or guide for the building of something that expands the structure into something useful". In addition, according to Duncan (1996) "a framework may be for a set of functions within a system and how they interrelate". Combining these, a framework is more inclusive than a protocol and more rigid than a structure. Thus, the original scientific contribution of this study is intended to "propose an integrated framework (not only an opinion mining algorithm but also data collection and sampling techniques) via social media big data analysis for sentiment calculation". In addition to all the sequences of this proposed integrated framework (design, development, implementation and evaluation), a unique data collection methodology and a purposive sampling technique for the sentiment analysis social media big data are proposed as demonstrated in the following subsections.

## 3.1 Design of sentiment analysis algorithm

As promising and emerging research area, text mining for sentiment analysis has been widely studied (Ahmad & Almas, 2005; Chaovalit & Zhou, 2005; Xu, Liao, & Li, 2008; Yuan, 2003), where sentiment analyses are applied for text classification tasks (R. Huang & Hansen, 2007; Jain, Ginwala, & Aslandogan, 2004). Li and Wu

(2010) summarize that existing sentiment calculation methodologies can be categorized into two types: machine learning based approaches (Chaovalit & Zhou, 2005; Xu et al., 2008) and semantic orientation based approaches (Turney & Littman, 2003; Xu et al., 2008; Yuan, 2003). While text sentiment analysis is very popular, the literature has a big gap on multicultural and multilingual studies.

Moreover, in the literature, sentiment analysis studies can be grouped into two main categories: "supervised" and "unsupervised". The pre-defined words and their polarities are used in one, on the other hand, the other classifies the most frequent words and drives a dictionary with them. However, Thelwall, Buckley, and Paltoglou (2012) state that similar results and accuracy rates are achieved from those two methods, since big data eliminates the noise of data and extreme cases which cause differences between those two methods. Due to these facts, predefined dictionaries are used in this study instead of creating new dictionaries from the data set, which would possibly take the progress, workload and timeline of the project to unmanageable levels.

Also, it is stated by Thelwall et al. (2012) that dual output for the sentiment analysis of blogging short texts concludes more accurate results. Therefore, the polarity of the tweets in the algorithm is better to be calculated not in one (binary classification) dimension but in two dimensions (positive, negative).

The language dictionaries for the algorithm is taken from a short text sentiment analysis tool (SentiStrength) created by Thelwall et al. (2010). This tool was developed for short text analysis and is still on testing phase due to accuracy concerns. The sentiment dictionaries of this study are found to be appropriate for our study since some of the dictionaries were previously tested with different platforms for single language studies (Durahim & Coşkun, 2015; Garas, Garcia, Skowron, &

Schweitzer, 2012; Giannopoulos, Weber, Jaimes, & Sellis, 2012; Grigore & Rosenkranz, 2011; Kucuktunc, Cambazoglu, Weber, & Ferhatosmanoglu, 2012; Thelwall & Buckley, 2013; Vural, Cambazoglu, Senkul, & Tokgoz, 2013; Zheludev, Smith, & Aste, 2014).

Additionally, Pfitzner, Garas, and Schweitzer (2012) concludes that the sentences ending with a "question mark (?)" should not be included in sentiment analysis. Because those texts do not represent the feelings of the people who write them. To this end, in this study the tweets ending with or including "question mark" are removed in the algorithm.

In addition to these common text sentiment analysis concepts, which are used by different social median platforms than Twitter (blogs, Facebook etc.) in different languages, it is a well-known fact that limiting the polarity scale for analyzed texts in a range (e.g. -5 to +5) is appropriate for balancing the standard deviation of total (or filtered) score.

Moreover, Rudra, Chakraborty, Ganguly, and Ghosh (2017) state that idioms usage in Twitter is 9% and this percentage would increase from 21.88% (in mentioned network user groups) to 49.57% (in subscription network user groups). Thus, since the dictionaries of the algorithm include idioms, idiom looking up operations are embedded in the proposed algorithm to increase the validity and accuracy.

Lastly, the booster and negating words have an effect on the polarities of the neighbor words. Therefore, this calculation is included in the algorithm, too.

As a conclusion of these sentiment analysis literature survey and using the text mining dictionaries a new sentiment analysis algorithm was developed as shown with pseudocodes in Figure 2.

1) For each tweet
   i. If the tweet is a question
      - ending with a question mark
      - using question words from dictionary

      check the "question" flag of it in database
   ii. If the tweet has any irony term (from dictionary) check the "irony" flag of it in database
   iii. Explode sentences to words, punctuations and emoticons
   iv. Find the idioms in the tweets, calculate their sentiments
   v. For each word
      - Find the sentiment polarity (in a range of -5 to +5) from dictionary
      - If whole word could not be found in dictionary, check root from rooting words (dictionary contains words with roots, for instance, instead of "ağlamak", dictionary contains "ağla*")
      - If the word is "booster word" (from dictionary) increase (or decrease) the following word's sentiment polarity
      - If the word is "negating word" (from dictionary) flip the following word's sentiment
   vi. For each emoticon find the sentiment polarity from emoticon dictionary
   vii. Store negative and positive sum of polarities

2) Check all database and crop all polarities more than 5 to 5 for positive sentiment score and less than -5 to -5 for negative sentiment score

3) Check all database and change 0 to -1 for negative sentiment score and change 0 to +1 for positive sentiment score

Figure 2. Proposed sentiment analysis (polarity calculation) algorithm

## 3.2 GNH calculation for countries

In the second half of the 20th century, Bhutan, as a South Asian country stated that 'Gross National Happiness is more important than Gross National Product (GDP)" and started a visionary study at the governmental (kingdom) level for calculating GNH (Priesner, 1999). This country level work can be accepted as the first GNH dominated study under United Nations, because the previous studies were mainly based on GDP. After Bhutan come up with GNH idea, organizations began to conduct researches in the early 1960s, and apparently took the pace in the 1970s (Galay, 2004). Coming from the core idea of measuring happiness in country level, Paulson (2017) argues that, "rather than pursuing economic growth as a means to all good ends, it may be useful to focus efforts directly toward chosen outcomes of well-being, measured not in monetary terms, but in terms valued by individual socio-

cultures". Therefore, since social media can be defined as a screaming platform for society to define public feeling, the idea of "a country level happiness measurement methodology can be defined and tested via social media big data" inspired this study. For the calculation of GNH for European calculation, a unique (but combined by a deep literature survey) algorithm was created.

### 3.3 GNH-TD: Gross national happiness calculation algorithm

Big data of this proposed social media analysis framework constitutes of daily tweets of the users coming from chosen countries (filtering options is explained in related section). GNH of each country is calculated by considering the daily tweets' happiness polarities of the users of the related country; therefore, for GNH Calculation in Tweets Domain (GNH-TD) operation, tweets of users are analyzed in terms of their polarities. For GNH Calculation in Tweets Domain (GNH-TD) operation, tweets of users are analyzed in terms of their polarities. In his novel GNH study, Kramer (2010) used Linguistic Inquiry and Word Count (LIWC) dictionary, and stated that this dictionary has different number of positive and negative words (and also their polarities are different at total). Thus, it is claimed that the potential for positive and negative word use is not equivalent. To cope with this problem and to generate a metric that is applicable and slightly independent of dictionary and language, adapting from Kramer (2010), the idea of "how much standard deviation away from mean?" is used in the GNH-TD algorithm. The proposed GNH-TD calculation algorithm is shown in Figure 3.

For each country, the following daily sentiment calculation algorithm is applied to find its daily GNH for 2404 days (from 2010 to 2015).

1) Sum the positive and negative daily polarities of tweets separately as ∑d+ and ∑d- where d+ is the positive, d- is the negative polarity of a tweet

2) Count number of tweets of each day (#d)

3) Find daily-average positive (μd+) and daily-average negative polarity (μd-)
   i. $\mu d+= \frac{\sum d+}{\#d}$
   ii. $\mu d-= \frac{\sum d-}{\#d}$

4) Find meta-standard deviation of daily positive and negative polarities
   i. σ+ : positive standard deviation
   ii. σ- : negative standard deviation

5) Find meta-average of daily positive and negative polarities
   i. μ+ : positive meta average
   ii. μ- : negative meta average

6) Find daily sentiment polarity for each day in a certain year
   i. $GNH = \frac{(\mu d+) - (\mu+)}{\sigma+} + \frac{(\mu d-) - (\mu-)}{\sigma-}$

Figure 3. Proposed GNH-TD calculation algorithm

At the end of the GNH calculations, the results are expected to be as in Table 5 for each country.

Table 5. Sample of GNH Calculation Results

| DATE | GNH-TD |
|---|---|
| 01-01-2010 | -1.146832669 |
| 02-01-2010 | 0.536102271 |
| 03-01-2010 | -1.541369300 |
| 04-01-2010 | -0.948914150 |
| 05-01-2010 | -0.705499462 |
| 06-01-2010 | -0.130058176 |
| 07-01-2010 | 0.380860574 |
| 08-01-2010 | -1.205014146 |
| 09-01-2010 | -1.620073852 |
| 10-01-2010 | 0.804391905 |
| 11-01-2010 | 0.033231117 |
| 12-01-2010 | 1.108196372 |
| ……. | |

## 3.4 Design of social media big data collection method

The proposed sentiment calculation algorithm for social media texts is intended to be used for measuring GNH for countries. On the other hand, the proposed framework which is uniquely designed for this study is not only composed of only sentiment calculation algorithm (GNH-TD) but also includes a novel social media big data collection technique which is designed to be used for Twitter platform. This method consists of three main steps. These steps are defined in the following part.

### 3.4.1 Accessing and collecting trend topics (TT)

In their study, Zheng, Han, and Sun (2017) discussed the location prediction methods on Twitter researches. They state that possible methods for location-based studies can be as listed below:

- Accepting users' self-declared profiles for location
- Aggregating geo-tags attached with users' tweets
- Choosing the most frequent city involved in the geotags
- Choosing the first valid geotag, and convert it to an administrative region, a cell, or coordinates
- Choosing the geometric median of the geo-tags

In addition to these, due to possible privacy concerns, empty and noisy information also appear in user profiles. Therefore, stating a location and accessing users (public) of that location is not a proper way. Also, if the research is about GNH calculation for a country, collecting users from Twitter who are from a specific country is not directly applicable with Twitter APIs. Similarly, although Twitter allows programs and devices to state geo-codes while publishing their tweets, the ratio of this geo-code usage is very low. For instance, Vieweg, Hughes, Starbird, and

Palen (2010) analyzed tweets generated during a flood event. Even this analysis was done with tweets about a flood event (among a homogenous group) it was found that 6% of the tweets contained location information. Due to this fact, collecting tweets and classifying them in terms of their possible geo-codes is not appropriate and efficient method. Instead, a complex but more efficient and valid method is designed in data collection phase of proposed framework.

As it is a common fact for data researchers, Twitter APIs let researchers, developers and practitioners gather data relevant to their works at no cost. Programmers can utilize the APIs which can be categorized related to their objectives as: a) REST API, which is popularly used for designing web APIs to use pull strategy for data retrieval, and b) Streaming API, which is used for continuous stream of public data with a push strategy. At this point, REST API method is advised to be used to collect data.

In order to gather a sample of active users, since it cannot be accessed in a direct way with just a single REST API, a TT search API ("GET trends / place") is suggested to be executed. This TT API works for different weeks for random sampling. Executing this API within the limit (100 APIs / hour) and from each execution, gathering 10 trend topics together with their characteristics, a dataset of approximately 300 unique trend topics (hashtags) for each country are intended to be collected using their "where on Earth identifier" (woeid). The data includes the following features;

- TT name,
- TT created at,
- TT search query,
- TT URL values.

### 3.4.2 Accessing users from TT and filtering bot (automatic) accounts

Collecting TT for a given "woeid", TT search query feature is advised to be used in "GET search / tweets" API to get 200 recent tweets about each TT. This API would help to collect tweets and unique users' features of those tweets. The following variables can be stored for each user account using the "json" format of the API.

- Account ID
- User name
- Screen name
- Number of followers
- Number of friends (followees, number of people s/he follows)
- Number of tweets
- Number of "favorited" tweets
- Account description
- Account creation time

After accessing the users and storing the account information of them, filtering phase begins. One of the most problematic issues in big data collection methodologies is detecting and filtering bot accounts (automatic -computer controlled- accounts used for publishing tweets with commercial or political purposes). In the proposed framework, eliminating the accounts (users) whose number of tweets are 2 standard deviations away from the mean of the users from that country is used to cope with this problem. Also, private accounts whose tweets cannot be collected via related APIs should be dropped from resultant dataset before the tweets collecting phase. This process is also important for ethical constraints in social media researches. Thus, filtering out private accounts is embedded to the framework.

### 3.4.3 Collecting tweets of chosen users

The last phase of the social media data collection part of the framework is collecting users' tweets. Tweets of the sample users can be gathered by "GET statuses / user_timeline" API. But, since this API is limited to collect 100 most recent tweets, a back-iterative API methodology executing with "max_id" option should be applied. In this method, firstly users last 100 tweets are collected. Then, the id of the last tweet is taken and it is given to the "GET statuses / user_timeline" API as "max_id" parameter for collecting last 100 tweets before the one whose id is this max_id.

Figure 4 summarizes the social media data collection phase of proposed framework in pseudocodes.

1) Collect TTs
   i. Collect Twitter TTs of each country with GET trends/place API using "woeid"
   ii. 10 for each iteration of API

2) Collect users and filter
   i. Collect 200 recent tweets written on each TT with GET search/tweets API
   ii. Collect authors (users) of those tweets and their Twitter properties (info)
   iii. Drop the ones whose account created before the beginning of intended research period
   iv. Eliminate the ones whose language is not the language of the related country
   v. Filter out the ones whose accounts are private
   vi. Calculate the standard deviation and mean of the number of tweets of remaining users
   vii. Remove the ones whose number of tweets are 2 standard deviation away from the mean
   viii. Randomly choose the given (selected quota) number of users

3) Collect tweets of users
   i. For each user
      - If first iteration of loop
        Collect users' 100 recent tweets with GET statuses/user_timeline API
        Else
        Collect users 100 tweets before max_id with GET statuses/user_timeline API
      - Take the oldest tweets' id and set as max_id
   ii. Stop loop if last collected tweet's creation time is before beginning of intended research period
   iii. Delete tweets before beginning of intended research period

Figure 4. Social media data collection methodology

# CHAPTER 4

# IMPLEMENTATION AND EVALUATION OF THE FRAMEWORK

Of all things, I liked[2] books best[+1].
[+4,-1]
Nikola Tesla

In this chapter the implementation of the proposed data collection methodology, filtering and sampling options, and the application of the novel sentiment analysis algorithm are stated. Additionally, the "happiness" calculation of the users for the exploratory analysis which states the relationship between users' social media characteristics and their happiness is defined.

## 4.1 Choosing countries for sample and collecting tweets

According to the official web site of the European Union (EU), www.europe.eu, the EU is a unique politic and economic partnership among 28 countries that together cover much of the continent. In addition to those 28 member-countries, there are seven candidates and two potential candidates. Since the main aim of this study is to draw a happiness map of European citizens with multilingual sentiment analysis framework of Twitter data, the main criteria for choosing a country for the framework are set as follows:

- the country should be open to Twitter usage with no bans or censorship
- there should be only one national language spoken within the country and that language must exist in our sentiment analysis dictionaries.

Based on these criteria, the eleven countries given in Figure 5 and Table 6 were chosen for the study.

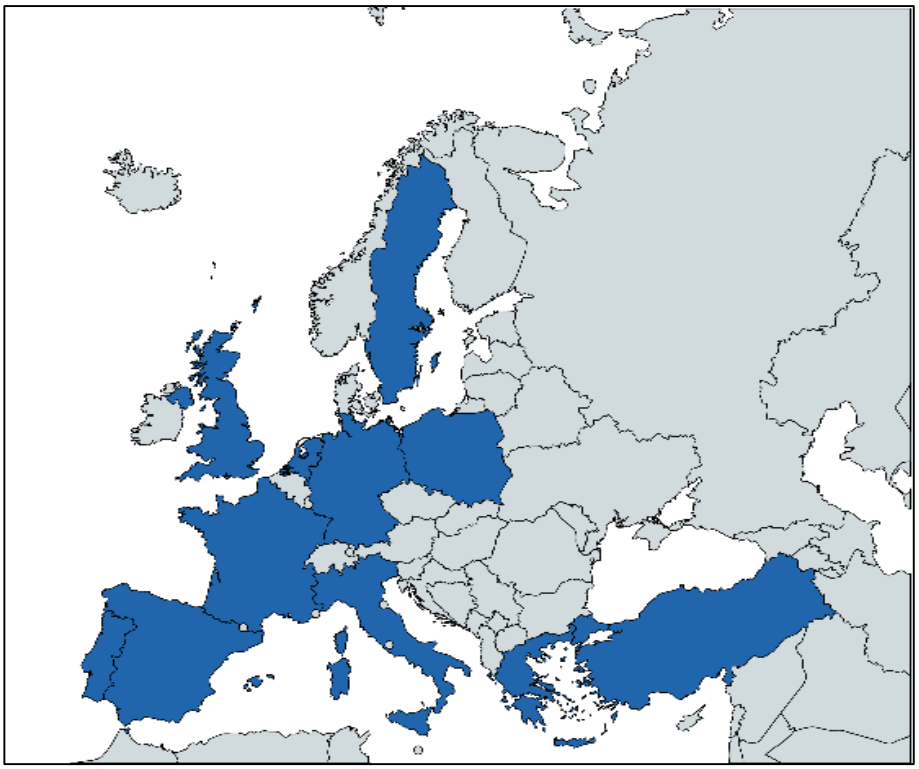


Figure 5. The eleven European countries chosen for the analysis

Table 6. The Eleven European Countries and Their Main Languages Chosen for the Analysis

| | Country | Language |
|---|---|---|
| 1 | Germany | German |
| 2 | Netherlands | Dutch |
| 3 | France | French |
| 4 | Greece | Greek |
| 5 | Italy | Italian |
| 6 | Portugal | Portuguese |
| 7 | Sweden | Swedish |
| 8 | Poland | Polish |
| 9 | Spain | Spanish |
| 10 | Turkey | Turkish |
| 11 | United Kingdom | English |

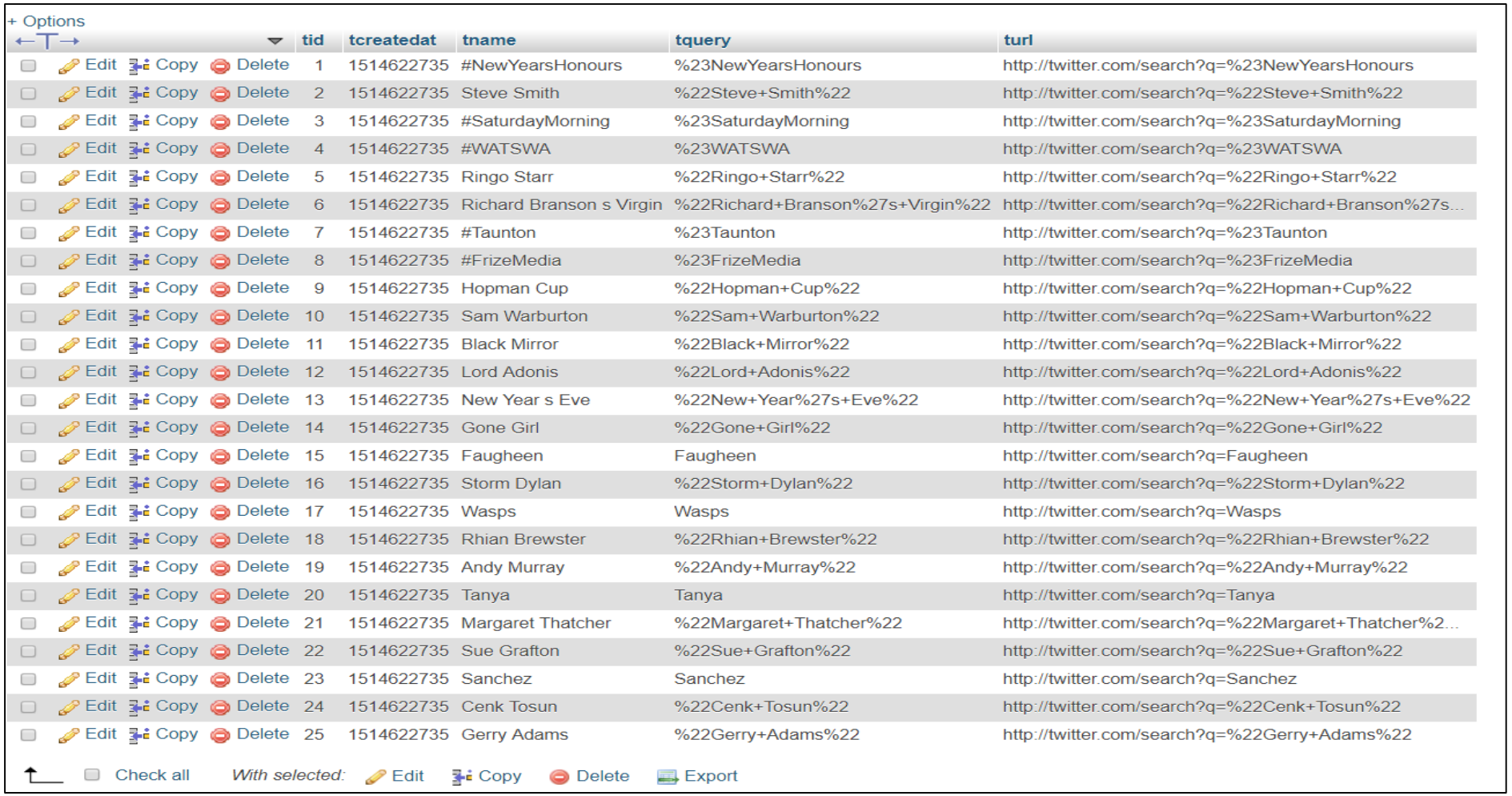
As it is mentioned in the data collection method, a REST API for tending topics (TT) with geocodes of those countries was executed for a period of time (three weeks to two months depending on countries). At the end of the TT API execution for 11

countries representative samples of users (in 1/5000 ratio with respect to populations (Krejcie & Morgan, 1970)) were accessed.

Figure 6 shows a sample of "trending topics" table and Figure 7 shows a sample of "users" table in the database.

Table 7. Sample Frame and Number of Accessed Users

| | | Sample Frame | | | Twitter Users (Sample) | | |
|---|---|---|---|---|---|---|---|
| | Country | Internet Users (Stats, 2014) (A) | Total Country Population (Stats, 2014) (B) | Number of Trend Topics Accessed | Total Accessed Users | Using National Language and Created Before 01/01/2010 (C) | Ratio to Total Population (B) / 5000 |
| 1 | Germany | 71,727,551 | 82,652,256 | 1,688 | 1,208,375 | 19,868 | 16,530 |
| 2 | United Kingdom | 57,075,826 | 63,489,234 | 789 | 119,335 | 15,856 | 12,698 |
| 3 | France | 55,429,382 | 64,641,279 | 3,750 | 1,679,862 | 15,414 | 12,928 |
| 4 | Italy | 36,593,969 | 61,070,224 | 2,660 | 1,308,952 | 14,487 | 12,214 |
| 5 | Turkey | 35,358,888 | 75,837,020 | 6,189 | 1,075,541 | 17,709 | 15,167 |
| 6 | Spain | 35,010,273 | 47,066,402 | 1,611 | 446,803 | 13,058 | 9,413 |
| 7 | Poland | 25,666,238 | 38,220,543 | 1,669 | 1,161,760 | 1,139 | 7,644 |
| 8 | Netherlands | 16,143,879 | 16,802,463 | 288 | 194,570 | 5,663 | 3,360 |
| 9 | Sweden | 8,581,261 | 9,631,261 | 1,488 | 1,119,278 | 2,777 | 1,926 |
| 10 | Portugal | 7,015,519 | 10,610,304 | 142 | 180,674 | 3,370 | 2,122 |
| 11 | Greece | 6,438,325 | 11,128,404 | 1,060 | 792,048 | 721 | 2,226 |
| TOTAL | | 355,041,111 | 481,149,390 | 21,334 | 9,287,198 | 110,062 | 96,230 |

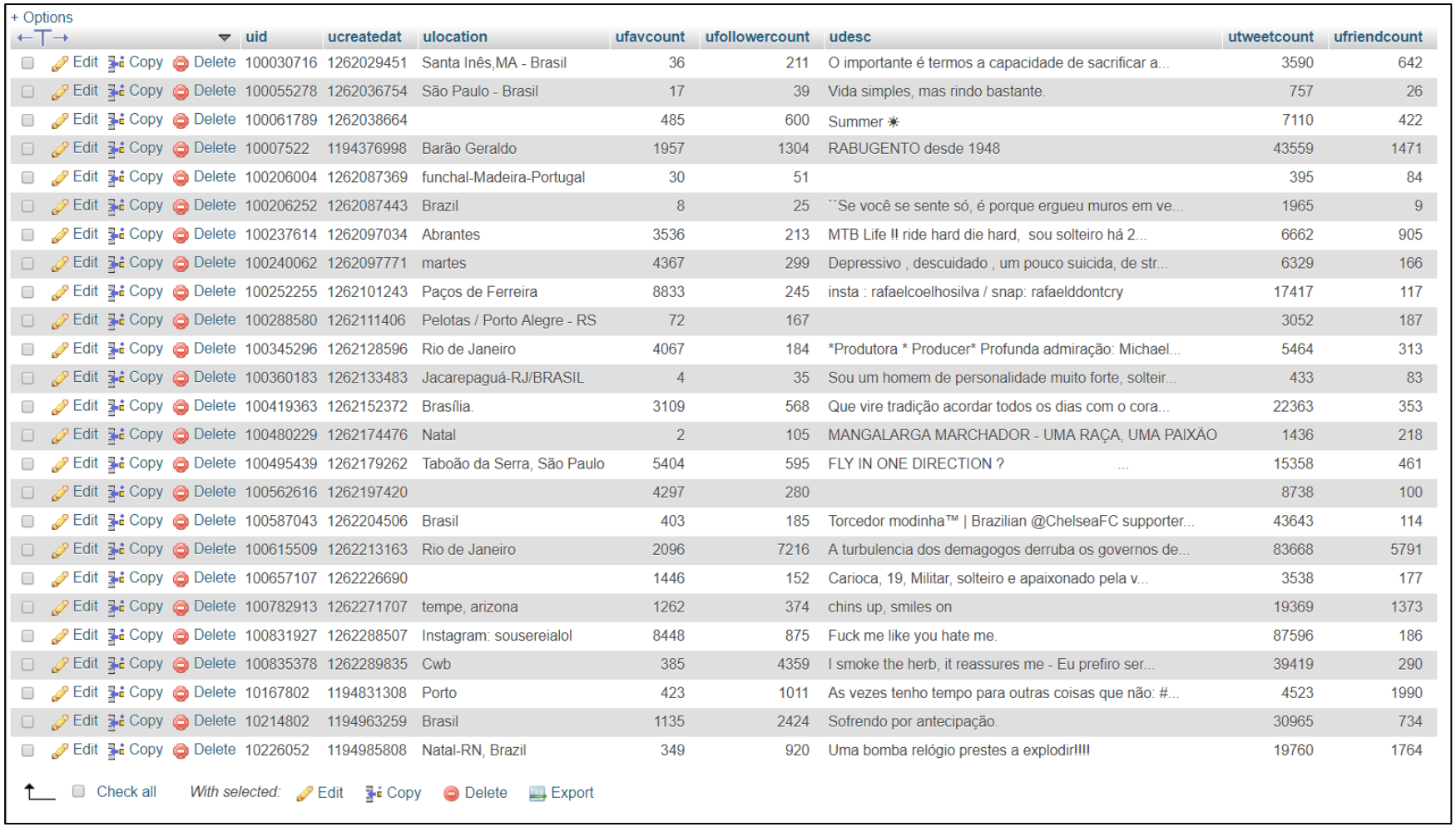


+ Options

| | tid | tcreatedat | tname | tquery | turl |
|---|---|---|---|---|---|
| Edit Copy Delete | 1 | 1514622735 | #NewYearsHonours | %23NewYearsHonours | http://twitter.com/search?q=%23NewYearsHonours |
| Edit Copy Delete | 2 | 1514622735 | Steve Smith | %22Steve+Smith%22 | http://twitter.com/search?q=%22Steve+Smith%22 |
| Edit Copy Delete | 3 | 1514622735 | #SaturdayMorning | %23SaturdayMorning | http://twitter.com/search?q=%23SaturdayMorning |
| Edit Copy Delete | 4 | 1514622735 | #WATSWA | %23WATSWA | http://twitter.com/search?q=%23WATSWA |
| Edit Copy Delete | 5 | 1514622735 | Ringo Starr | %22Ringo+Starr%22 | http://twitter.com/search?q=%22Ringo+Starr%22 |
| Edit Copy Delete | 6 | 1514622735 | Richard Branson s Virgin | %22Richard+Branson%27s+Virgin%22 | http://twitter.com/search?q=%22Richard+Branson%27s... |
| Edit Copy Delete | 7 | 1514622735 | #Taunton | %23Taunton | http://twitter.com/search?q=%23Taunton |
| Edit Copy Delete | 8 | 1514622735 | #FrizeMedia | %23FrizeMedia | http://twitter.com/search?q=%23FrizeMedia |
| Edit Copy Delete | 9 | 1514622735 | Hopman Cup | %22Hopman+Cup%22 | http://twitter.com/search?q=%22Hopman+Cup%22 |
| Edit Copy Delete | 10 | 1514622735 | Sam Warburton | %22Sam+Warburton%22 | http://twitter.com/search?q=%22Sam+Warburton%22 |
| Edit Copy Delete | 11 | 1514622735 | Black Mirror | %22Black+Mirror%22 | http://twitter.com/search?q=%22Black+Mirror%22 |
| Edit Copy Delete | 12 | 1514622735 | Lord Adonis | %22Lord+Adonis%22 | http://twitter.com/search?q=%22Lord+Adonis%22 |
| Edit Copy Delete | 13 | 1514622735 | New Year s Eve | %22New+Year%27s+Eve%22 | http://twitter.com/search?q=%22New+Year%27s+Eve%22 |
| Edit Copy Delete | 14 | 1514622735 | Gone Girl | %22Gone+Girl%22 | http://twitter.com/search?q=%22Gone+Girl%22 |
| Edit Copy Delete | 15 | 1514622735 | Faugheen | Faugheen | http://twitter.com/search?q=Faugheen |
| Edit Copy Delete | 16 | 1514622735 | Storm Dylan | %22Storm+Dylan%22 | http://twitter.com/search?q=%22Storm+Dylan%22 |
| Edit Copy Delete | 17 | 1514622735 | Wasps | Wasps | http://twitter.com/search?q=Wasps |
| Edit Copy Delete | 18 | 1514622735 | Rhian Brewster | %22Rhian+Brewster%22 | http://twitter.com/search?q=%22Rhian+Brewster%22 |
| Edit Copy Delete | 19 | 1514622735 | Andy Murray | %22Andy+Murray%22 | http://twitter.com/search?q=%22Andy+Murray%22 |
| Edit Copy Delete | 20 | 1514622735 | Tanya | Tanya | http://twitter.com/search?q=Tanya |
| Edit Copy Delete | 21 | 1514622735 | Margaret Thatcher | %22Margaret+Thatcher%22 | http://twitter.com/search?q=%22Margaret+Thatcher%2... |
| Edit Copy Delete | 22 | 1514622735 | Sue Grafton | %22Sue+Grafton%22 | http://twitter.com/search?q=%22Sue+Grafton%22 |
| Edit Copy Delete | 23 | 1514622735 | Sanchez | Sanchez | http://twitter.com/search?q=Sanchez |
| Edit Copy Delete | 24 | 1514622735 | Cenk Tosun | %22Cenk+Tosun%22 | http://twitter.com/search?q=%22Cenk+Tosun%22 |
| Edit Copy Delete | 25 | 1514622735 | Gerry Adams | %22Gerry+Adams%22 | http://twitter.com/search?q=%22Gerry+Adams%22 |

Check all *With selected:* Edit Copy Delete Export

Figure 6. Sample of trending topics table of database



+ Options

| | uid | ucreatedat | ulocation | ufavcount | ufollowercount | udesc | utweetcount | ufriendcount |
|---|---|---|---|---|---|---|---|---|
| Edit Copy Delete | 100030716 | 1262029451 | Santa Inês,MA - Brasil | 36 | 211 | O importante é termos a capacidade de sacrificar a... | 3590 | 642 |
| Edit Copy Delete | 100055278 | 1262036754 | São Paulo - Brasil | 17 | 39 | Vida simples, mas rindo bastante. | 757 | 26 |
| Edit Copy Delete | 100061789 | 1262038664 | | 485 | 600 | Summer ☀ | 7110 | 422 |
| Edit Copy Delete | 10007522 | 1194376998 | Barão Geraldo | 1957 | 1304 | RABUGENTO desde 1948 | 43559 | 1471 |
| Edit Copy Delete | 100206004 | 1262087369 | funchal-Madeira-Portugal | 30 | 51 | | 395 | 84 |
| Edit Copy Delete | 100206252 | 1262087443 | Brazil | 8 | 25 | ``Se você se sente só, é porque ergueu muros em ve... | 1965 | 9 |
| Edit Copy Delete | 100237614 | 1262097034 | Abrantes | 3536 | 213 | MTB Life ‼ ride hard die hard, sou solteiro há 2... | 6662 | 905 |
| Edit Copy Delete | 100240062 | 1262097771 | martes | 4367 | 299 | Depressivo , descuidado , um pouco suicida, de str... | 6329 | 166 |
| Edit Copy Delete | 100252255 | 1262101243 | Paços de Ferreira | 8833 | 245 | insta : rafaelcoelhosilva / snap: rafaelddontcry | 17417 | 117 |
| Edit Copy Delete | 100288580 | 1262111406 | Pelotas / Porto Alegre - RS | 72 | 167 | | 3052 | 187 |
| Edit Copy Delete | 100345296 | 1262128596 | Rio de Janeiro | 4067 | 184 | *Produtora * Producer* Profunda admiração: Michael... | 5464 | 313 |
| Edit Copy Delete | 100360183 | 1262133483 | Jacarepaguá-RJ/BRASIL | 4 | 35 | Sou um homem de personalidade muito forte, solteir... | 433 | 83 |
| Edit Copy Delete | 100419363 | 1262152372 | Brasília. | 3109 | 568 | Que vire tradição acordar todos os dias com o cora... | 22363 | 353 |
| Edit Copy Delete | 100480229 | 1262174476 | Natal | 2 | 105 | MANGALARGA MARCHADOR - UMA RAÇA, UMA PAIXÃO | 1436 | 218 |
| Edit Copy Delete | 100495439 | 1262179262 | Taboão da Serra, São Paulo | 5404 | 595 | FLY IN ONE DIRECTION ? ... | 15358 | 461 |
| Edit Copy Delete | 100562616 | 1262197420 | | 4297 | 280 | | 8738 | 100 |
| Edit Copy Delete | 100587043 | 1262204506 | Brasil | 403 | 185 | Torcedor modinha™ \| Brazilian @ChelseaFC supporter... | 43643 | 114 |
| Edit Copy Delete | 100615509 | 1262213163 | Rio de Janeiro | 2096 | 7216 | A turbulencia dos demagogos derruba os governos de... | 83668 | 5791 |
| Edit Copy Delete | 100657107 | 1262226690 | | 1446 | 152 | Carioca, 19, Militar, solteiro e apaixonado pela v... | 3538 | 177 |
| Edit Copy Delete | 100782913 | 1262271707 | tempe, arizona | 1262 | 374 | chins up, smiles on | 19369 | 1373 |
| Edit Copy Delete | 100831927 | 1262288507 | Instagram: sousereialol | 8448 | 875 | Fuck me like you hate me. | 87596 | 186 |
| Edit Copy Delete | 100835378 | 1262289835 | Cwb | 385 | 4359 | I smoke the herb, it reassures me - Eu prefiro ser... | 39419 | 290 |
| Edit Copy Delete | 10167802 | 1194831308 | Porto | 423 | 1011 | As vezes tenho tempo para outras coisas que não: #... | 4523 | 1990 |
| Edit Copy Delete | 10214802 | 1194963259 | Brasil | 1135 | 2424 | Sofrendo por antecipação. | 30965 | 734 |
| Edit Copy Delete | 10226052 | 1194985808 | Natal-RN, Brazil | 349 | 920 | Uma bomba relógio prestes a explodir!!!! | 19760 | 1764 |

Check all *With selected:* Edit Copy Delete Export

Figure 7. Sample of users table of database

As it can be seen in Table 7 there are 38 million people in Poland in which 25 million are internet users. For the ration of 1/5000, we should access 7644, but, while we accessed 1,161,760 users with 1669 trend topics for about two months execution of the API, only 1139 are using Polish and created before 2010 and this number is far less than 7644 (1/5000 of the total population). This situation was the same for Greece, so, those countries were dropped from the sample. Moreover, the ratio of the sample over the total population of the countries is (Total C/Total B) 0.02287%) which can be concluded as a generalizable sample ratio. At the end those tweets of 110,062 sample users were totally gathered by "GET statuses / user_timeline" API. Since this API can access to 100 most recent tweets, a back-iterative API which executes with "max_id" option was applied as mentioned in Figure 4. This way, all the tweets of every user were collected from 01.01.2010 00:00 GMT to 31.12.2015 23:59 GMT. The yearly number of tweets per country are listed in Table 8.

Table 8. Yearly Tweet Numbers Collected for Countries

| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Germany | 1,594,312 | 2,167,848 | 3,446,841 | 5,579,979 | 11,038,771 | 14,500,870 |
| United Kingdom | 264,460 | 696,529 | 1,987,674 | 4,127,235 | 10,243,911 | 26,893,481 |
| France | 442,461 | 976,445 | 2,594,032 | 4,762,945 | 9,608,994 | 19,044,399 |
| Italy | 430,638 | 1,027,982 | 3,371,571 | 5,547,187 | 8,791,952 | 14,411,831 |
| Turkey | 141,592 | 595,032 | 2,162,091 | 5,097,690 | 7,707,706 | 11,430,470 |
| Spain | 174,285 | 626,509 | 1,748,940 | 3,613,995 | 8,966,293 | 25,190,433 |
| Netherlands | 330,863 | 907,171 | 1,543,801 | 2,384,925 | 4,686,031 | 7,841,054 |
| Sweden | 131,119 | 258,425 | 689,734 | 1,190,342 | 2,031,537 | 2,114,549 |
| Portugal | 116,946 | 250,063 | 390,444 | 840,847 | 2,372,739 | 6,754,129 |

To sum up, totally 255,842,103 tweets were collected to perform a sentiment analysis for nine countries for a six-year period.

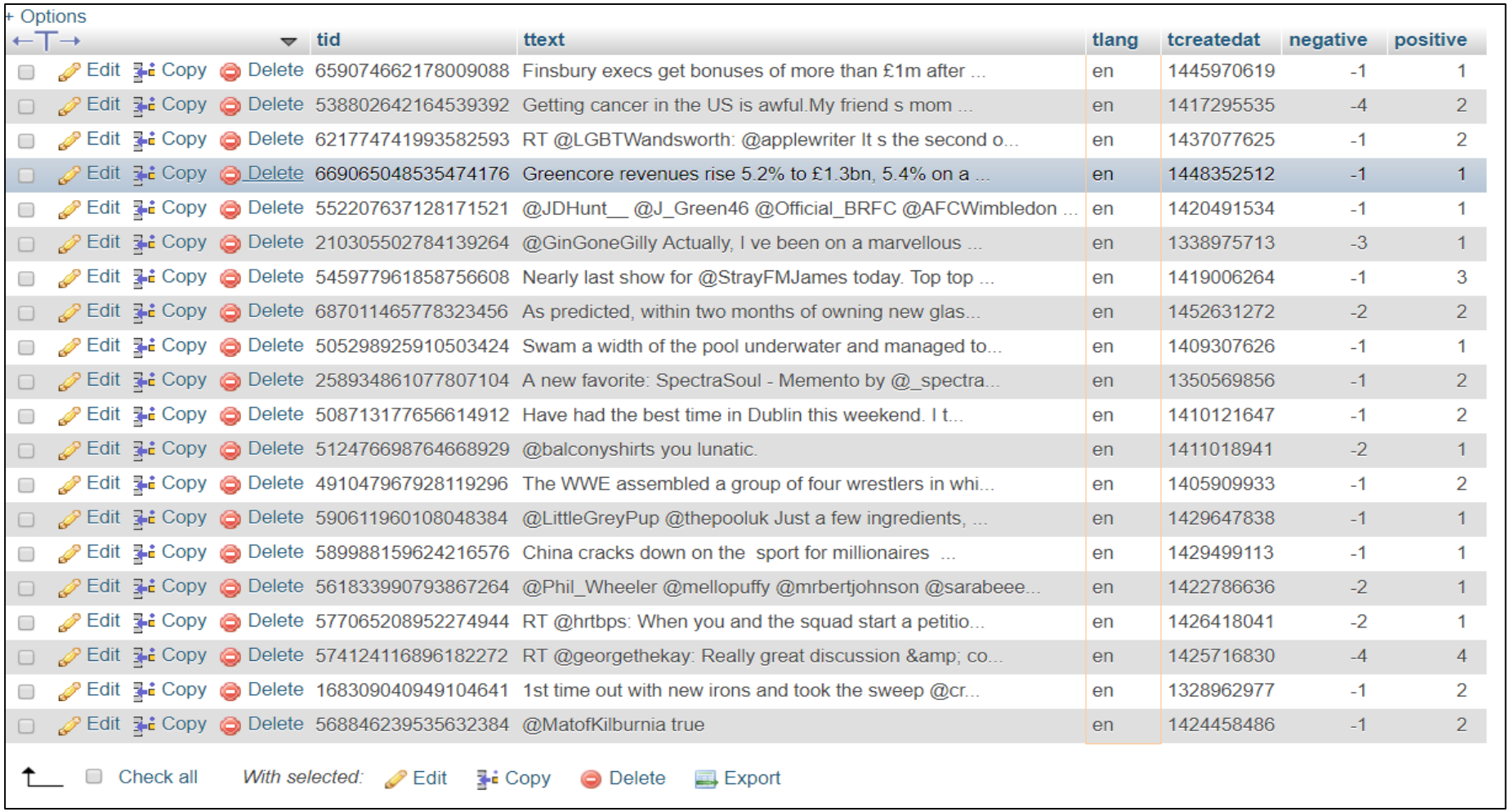
## 4.2 Sentiment analysis algorithm and polarity (GNH-TD) calculation

After collecting tweets of 110,062 users from nine countries, the proposed sentiment analysis algorithm (see Figure 2 and Figure 3) was applied for all tweets. A sample tweet is shown in Figure 8.

> Really good coverage of #ParisAttacks on itv. Clear and concise programme, but god! what an incredibly scary night. There are no words :_(

Figure 8. Sample sentiment analysis report of a tweet

This tweet was published in United Kingdom at 13.11.2015 about Paris Terrorist attacks. The word “good” has +1 weight in dictionary, but since there is “really” near before it, the weight of it is increased by algorithm to +2. The word phrase “but god” would have positive or negative feeling but in the same sentence there is “clear and concise” words (+1) thus the weight of “but god” was automatically stated as -1. Lastly, “scary” has a weight of -2 but the booster word “incredibly” increased its negative value to -3. Lastly, the emotion :_( has the polarity of −1 in the emoticon dictionary. At a result, the polarity of the tweet stated as +3 and −5. By this methodology all the tweets of the users were calculated and their polarities were stored into database. Then with the GNH-TD calculation algorithm (see Figure 3) daily GNH values of all countries were calculated (for 2191 days 6 years). The results of the sentiment analysis are stored to the database as shown in Figure 9.



+ Options

| | tid | ttext | tlang | tcreatedat | negative | positive |
|---|---|---|---|---|---|---|
| Edit Copy Delete | 659074662178009088 | Finsbury execs get bonuses of more than £1m after ... | en | 1445970619 | -1 | 1 |
| Edit Copy Delete | 538802642164539392 | Getting cancer in the US is awful.My friend s mom ... | en | 1417295535 | -4 | 2 |
| Edit Copy Delete | 621774741993582593 | RT @LGBTWandsworth: @applewriter It s the second o... | en | 1437077625 | -1 | 2 |
| Edit Copy Delete | 669065048535474176 | Greencore revenues rise 5.2% to £1.3bn, 5.4% on a ... | en | 1448352512 | -1 | 1 |
| Edit Copy Delete | 552207637128171521 | @JDHunt__ @J_Green46 @Official_BRFC @AFCWimbledon ... | en | 1420491534 | -1 | 1 |
| Edit Copy Delete | 210305502784139264 | @GinGoneGilly Actually, I ve been on a marvellous ... | en | 1338975713 | -3 | 1 |
| Edit Copy Delete | 545977961858756608 | Nearly last show for @StrayFMJames today. Top top ... | en | 1419006264 | -1 | 3 |
| Edit Copy Delete | 687011465778323456 | As predicted, within two months of owning new glas... | en | 1452631272 | -2 | 2 |
| Edit Copy Delete | 505298925910503424 | Swam a width of the pool underwater and managed to... | en | 1409307626 | -1 | 1 |
| Edit Copy Delete | 258934861077807104 | A new favorite: SpectraSoul - Memento by @_spectra... | en | 1350569856 | -1 | 2 |
| Edit Copy Delete | 508713177656614912 | Have had the best time in Dublin this weekend. I t... | en | 1410121647 | -1 | 2 |
| Edit Copy Delete | 512476698764668929 | @balconyshirts you lunatic. | en | 1411018941 | -2 | 1 |
| Edit Copy Delete | 491047967928119296 | The WWE assembled a group of four wrestlers in whi... | en | 1405909933 | -1 | 2 |
| Edit Copy Delete | 590611960108048384 | @LittleGreyPup @thepooluk Just a few ingredients, ... | en | 1429647838 | -1 | 1 |
| Edit Copy Delete | 589988159624216576 | China cracks down on the  sport for millionaires  ... | en | 1429499113 | -1 | 1 |
| Edit Copy Delete | 561833990793867264 | @Phil_Wheeler @mellopuffy @mrbertjohnson @sarabeee... | en | 1422786636 | -2 | 1 |
| Edit Copy Delete | 577065208952274944 | RT @hrtbps: When you and the squad start a petitio... | en | 1426418041 | -2 | 1 |
| Edit Copy Delete | 574124116896182272 | RT @georgethekay: Really great discussion & co... | en | 1425716830 | -4 | 4 |
| Edit Copy Delete | 168309040949104641 | 1st time out with new irons and took the sweep @cr... | en | 1328962977 | -1 | 2 |
| Edit Copy Delete | 568846239535632384 | @MatofKilburnia true | en | 1424458486 | -1 | 2 |

Check all *With selected:* Edit Copy Delete Export

Figure 9. Sample of tweets table in the database with sentiment results

## 4.3 Users' Twitter social media happiness calculation

To determine the relationship between users' Twitter social media characteristics and their happiness levels, simple linear regression analyses are used. The followings are the Twitter social media characteristics which are taken to be the independent variables:

- Length of screen name
- Number of followers
- Number of friends (followees, number of people s/he follows)
- Number of tweets
- Number of "favorited" tweets
- Length of account description
- Twitter age

The length of screen name, length of account description and Twitter age are calculated variables. These values (Twitter social media characteristics of users) belongs to the end of 2015 which is the end point of chosen time interval for the study. Thus, the dependent variable, "happiness" of the users is calculated up to this time for simple linear regression analyses.

The average happiness calculation algorithm for the users can be simply defined as in Figure 10.

1) Sum the positive and negative polarities of tweets of each user for 2404 days
   i. $\sum p+$ : positive polarity total
   ii. $\sum p-$ : negative polarity total

2) Count number of tweets of user (#t)

3) Find average positive ($\mu p+$) and average negative polarities ($\mu p+$) of user
   i. $\mu p+ = \frac{\sum p+}{\#t}$
   ii. $\mu p- = \frac{\sum p-}{\#t}$

Figure 10. User sentiment calculation algorithm

# CHAPTER 5

# ANALYSES OF THE FRAMEWORK

I can calculate the motion of heavenly[1] bodies,
but not[±] the madness[-1] of people.
[+3,-1]
Isaac Newton

In this chapter validity and reliability analysis of sentiment algorithm results are described. Then, cross sectional analyses of users' happiness levels and account features are captured.

## 5.1 Sentiment analysis

The main aim of this study is to state a framework for GNH calculation via social media big data. Thus, proper number of active users and their tweets are calculated with proposed data collection method. Afterwards, the novel sentiment calculation algorithm was applied to more than 250 million tweets. On the other hand, before stating the GNH values of the countries, the validity and reliability of the results and algorithm should be examined. In this perspective, first three research questions were asked and analyzed.

The first question is "Is there face validity when the polarities determined by sentiment analysis framework are compared with Stock Market Index and Exchange Rates?". To check face validity of the results, the historical data (from 1.1.2010 to 12.31.2015) of main stock market indices of the countries were collected from Yahoo Finance web site (https://finance.yahoo.com/). Also, Euro-Dollar (eur-usd), Euro-Pound (eur-gbp) and Pound-Dollar (gbp-usd) daily exchanges are collected. Then, bivariate correlations between the daily GNH-TD results, main stock indices

and monetary exchanges are examined with Pearson's Correlation statistical analysis. The results are shown in Table 9.

Table 9. Results of Face Validity Pearson's Correlation Analysis

| | | GNH-TD National Market Index | GNH-TD EUR-USD | GNH-TD GBP-USD | GNH-TD GBP-EUR |
|---|---|---|---|---|---|
| Germany | Pearson Correlation | -.731** | .498** | .059* | .589** |
| DAX | Sig. (2-tailed) | 0 | 0 | 0.019 | 0 |
| | n | 1527 | 1565 | 1565 | 1565 |
| United Kingdom | Pearson Correlation | -.603** | .627** | .124** | .714** |
| FTSE100 | Sig. (2-tailed) | 0 | 0 | 0 | 0 |
| | n | 1514 | 1565 | 1565 | 1565 |
| France | Pearson Correlation | -.537** | .494** | .079** | .572** |
| CAC40 | Sig. (2-tailed) | 0 | 0 | 0.002 | 0 |
| | n | 1537 | 1565 | 1565 | 1565 |
| Italy | Pearson Correlation | -.183** | .417** | -0.044 | .545** |
| FTSEMIB | Sig. (2-tailed) | 0 | 0 | 0.081 | 0 |
| | n | 1538 | 1565 | 1565 | 1565 |
| Turkey | Pearson Correlation | -.548** | .506** | -0.004 | .631** |
| BIST100 | Sig. (2-tailed) | 0 | 0 | 0.888 | 0 |
| | n | 1511 | 1565 | 1565 | 1565 |
| Spain | Pearson Correlation | -.268** | .503** | .054* | .597** |
| IBEX35 | Sig. (2-tailed) | 0 | 0 | 0.033 | 0 |
| | n | 1535 | 1565 | 1565 | 1565 |
| Netherlands | Pearson Correlation | -.687** | .551** | .184** | .584** |
| AEX | Sig. (2-tailed) | 0 | 0 | 0 | 0 |
| | n | 1537 | 1565 | 1565 | 1565 |
| Sweden | Pearson Correlation | -.641** | .469** | .056* | .551** |
| OMX30 | Sig. (2-tailed) | 0 | 0 | 0.026 | 0 |
| | n | 1506 | 1565 | 1565 | 1565 |
| Portugal | Pearson Correlation | .344** | .585** | .118** | .664** |
| PSI20 | Sig. (2-tailed) | 0 | 0 | 0 | 0 |
| | n | 1440 | 1565 | 1565 | 1565 |
| **. Correlation is significant at the 0.01 level (2-tailed).<br>*. Correlation is significant at the 0.05 level (2-tailed). | | | | | |

Results showed that all the GNH-TD of countries are significantly correlated with monetary exchanges and stock market indices.

The second research question is about convergent validity of the proposed framework: "Is there convergent validity when the GNH results of the sentiment

analysis framework and GNH survey results of Organization for Economic Cooperation and Development (OECD) report are compared?". To check this validity, OECD life satisfaction survey results of all countries were gathered from OECD Data Bank (http://stats.oecd.org). Since there are online four-year results (2012 to 2015) matching to our time interval, only 36 GNH measures (4 years - 9 countries) were examined again with Pearson's Correlation Analysis. Results of this analysis are shown in Table 10 and the scatter graph of these 36 cases is given in Figure 11.

Table 10. Convergent Validity Analysis Results

| | | OECD-Better Life Index | GNH-TD |
|---|---|---|---|
| OECD-Better Life Index | Pearson Correlation | 1 | .854** |
| | Sig. (2-tailed) | | .000 |
| | n | 36 | 36 |
| GNH-TD | Pearson Correlation | .854** | 1 |
| | Sig. (2-tailed) | .000 | |
| | n | 36 | 36 |
| **. Correlation is significant at the 0.01 level (2-tailed). | | | |

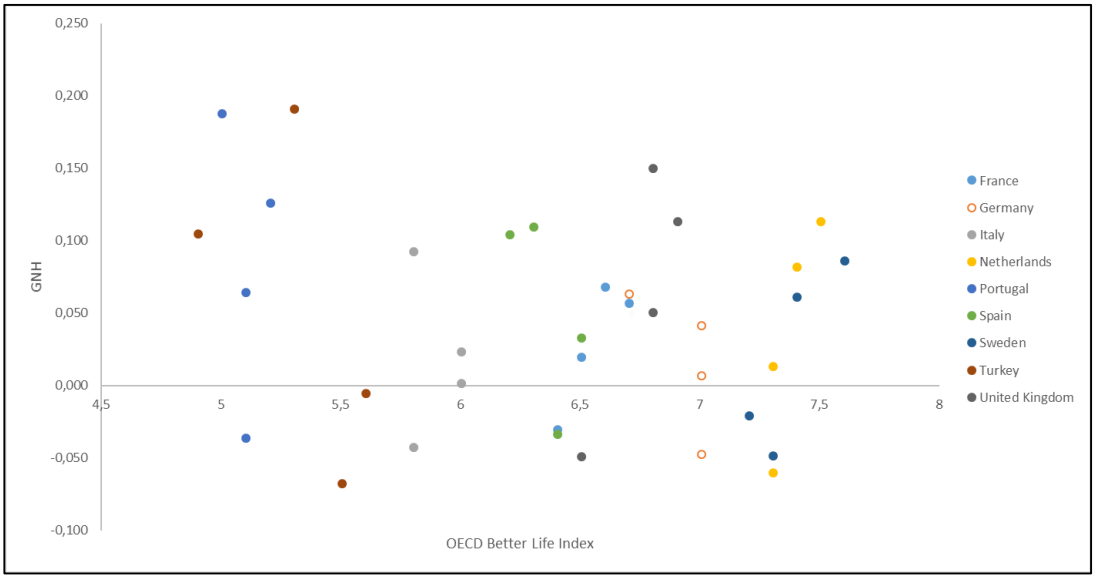



Figure 11. Scatter graph of 36 cases

Since there is a significant and high correlation between the GNH-TD results and OECD survey report, the convergent validity of the proposed framework is proved. Additionally, this result can be concluded as a replacement of this social media sentiment analysis framework to the OECD survey method for Life Satisfaction analysis among countries.

The third research question is about reliability of the dataset and results: "Is there data reliability when the peaks/troughs of the graphs of sentiment analysis framework are compared with specific dates obtained from news archives?". The common way for finding reliability of the sentiment analysis in literature is that the results are compared with findings from other sources such as news, archives, questionnaires, company secondary data, even manual provided and classified data etc. but, by this backward accuracy checking method, the real power of sentiment analysis cannot be detected. In other words, we cannot claim that our sentiment analysis results are accurate when we check the results with real data, because in this way we probably miss some real events to check. Therefore, a forward methodology for our accuracy check is more appropriate and valuable. In this forward method, first the data from the past are collected and the socially effective days of the countries for the selected time period were stated. For stating those days, the Wikipedia events pages were used (e.g. "2014 in Spain" with web address https://en.wiki.ng/wiki/2014_in_Spain, "2011 in Turkey" with web address https://en.wiki.ng/wiki/2011_in_Turkey). Then the results of the sentiment analysis are checked with this data in terms of how much of the past could be detected. In country level, the reliability of sentiment analysis framework was checked in terms of detecting those effective days. Since there are 2192 days in the chosen time interval and since GNH-TD found an aggregate happiness polarity value for all of

those days, a threshold value was needed for determining the socially important days where after would be called as “extraordinary” days. To this respect, threshold value was calculated as “two standard deviations away from mean”. The mean and standard deviation values of all the polarities of 2192 days and Positive and Negative Threshold values for all countries are listed in Table 11.

Table 11. Means and Standard Deviations of Polarities and Threshold Values

| Country | Mean (Negative) | Standard Deviation (Negative) | Negative Threshold | Mean (Positive) | Standard Deviation (Positive) | Positive Threshold |
|---|---|---|---|---|---|---|
| Germany | -1.2192 | 0.6364 | <-2.4920 | 1.3684 | 0.6307 | >2.6298 |
| United Kingdom | -1.4522 | 0.8271 | <-3.1064 | 1.5253 | 0.7507 | >3.0267 |
| France | -1.4936 | 0.8214 | <-3.1364 | 1.2965 | 0.5799 | >2.4563 |
| Italy | -1.1926 | 0.5490 | <-2.2906 | 1.2906 | 0.5678 | >2.4262 |
| Turkey | -1.1579 | 0.5092 | <-2.1763 | 1.2656 | 0.5526 | >2.3708 |
| Spain | -1.4155 | 0.7900 | <-2.9955 | 1.6825 | 0.9871 | >3.6567 |
| Netherland | -1.3444 | 0.6957 | <-2.7358 | 1.2974 | 0.6130 | >2.5234 |
| Sweden | -1.2122 | 0.5478 | <-2.3078 | 1.2930 | 0.5725 | >2.438 |
| Portugal | -1.3959 | 0.7635 | <-2.9229 | 1.3422 | 0.6323 | >2.6068 |

After examining deeply, the days of having negative aggregate polarities below negative threshold or upper positive threshold values, the detection accuracy of the proposed sentiment analysis framework is listed in Table 12.

Table 12. Detection Accuracy Results

| Country | Number of Event Days in Wikipedia Pages | Number of Days chosen from GNH-TD | Detection Accuracy |
|---|---|---|---|
| Germany | 105 | 74 | 70.48% |
| United Kingdom | 121 | 104 | 85.95% |
| France | 137 | 112 | 81.75% |
| Italy | 112 | 83 | 74.11% |
| Turkey | 163 | 146 | 89.57% |
| Spain | 72 | 52 | 72.22% |
| Netherland | 84 | 59 | 70.24% |
| Sweden | 69 | 57 | 82.61% |
| Portugal | 58 | 42 | 72.41% |

To sum up, since all the accuracy percentages are bigger than %70 threshold value (Rost & Sander, 1993), the reliability of the dataset and proposed GNH-TD framework is proved.

## 5.2 Simple linear regression analyses between Twitter users' happiness levels and social media characteristics

After validating the sentiment analysis algorithm, users' happiness level at the end of the chosen time period are calculated with the given algorithm in Figure 10. The main objective of these analyses is to find and analyze the possible significant relationships between Twitter users' social media account characteristics and their happiness levels, which is intended to be calculated by the proposed sentiment analysis framework.

The number of users for which 6-years period average happiness are calculated is given in Table 13 in country base.

Table 13. The Number of Users for Which 6-Years Period Average Happiness are Calculated

| Country | Number of Users |
|---|---|
| Germany | 18,955 |
| Spain | 11,804 |
| France | 14,884 |
| Italy | 14,117 |
| Netherlands | 5,536 |
| Portugal | 3,124 |
| Sweden | 2,719 |
| Turkey | 14,203 |
| United Kingdom | 14,551 |
| TOTAL | 99,893 |

At this point following Twitter social media characteristics of users are directly gathered from related APIs during the data collection phase:

- Number of followers
- Number of friends (followees, number of people s/he follows)
- Number of tweets
- Number of "favorited" tweets

On the other hand, the following Twitter user characteristics are calculated with "Length" function of MYSQL query language.

- Length of Screen name
- Length Account description

Lastly Twitter Age variable is the most complex variable to be calculated in the dataset. The account creation time could be stored to database in UNIX time format because Twitter APIs' .json results turn back as this. This format is an integer number which represents "how many seconds it takes from 01.01.1970 00:00". Thus, for finding the Twitter age of users at the end of time interval (the time of all the other variables), the difference between "account creation time" and "1451606399" was calculated. The integer number "1451606399" represents the time of 31.12.2015 23:59:59 here. Then this number was divided to 86400 which means 1 day in seconds. Thus, the Twitter ages of users were found in "data" unit. For all of these calculations, MYSQL query shown in Figure 12 is used in database.

```
SELECT

    (1451606399-ucreatedat) / 86400 as Age ,
    ufavcount AS UserFavoritesCount  ,
    ufollowercount AS UserFollowersCount  ,
    length(udesc) AS LengthofUserDescription ,
    length(uscreenname) AS LengthofUserScreenName ,
    utweetcount AS UserTweetsCount ,
    ufriendcount AS UserFriendsCount ,
    uhappiness AS Happiness

FROM user
```

Figure 12. MYSQL query of Twitter social media characteristics

Figure 13 shows a sample of the results of the MYSQL query. As a result of the query the results of 99,893 users are collected and exported for simple linear regression analyses which are done in SPSS v23.

+ Options

| age | ufavcount | ufollowercount | length(udesc) | length(uscreenname) | utweetcount | ufriendcount | uhappiness |
|---|---|---|---|---|---|---|---|
| 189576948 | 36 | 211 | 99 | 9 | 3590 | 642 | 0.04074074 |
| 189569645 | 17 | 39 | 33 | 9 | 757 | 26 | -0.01259182 |
| 189567735 | 485 | 600 | 10 | 14 | 7110 | 422 | 0.03588144 |
| 257229401 | 1957 | 1304 | 20 | 9 | 43559 | 1471 | -0.30814677 |
| 189519030 | 30 | 51 | 0 | 8 | 395 | 84 | 0.22004357 |
| 189518956 | 8 | 25 | 68 | 11 | 1965 | 9 | 0.01183152 |
| 189509365 | 3536 | 213 | 95 | 13 | 6662 | 905 | 0.00272315 |
| 189505156 | 8833 | 245 | 48 | 13 | 17417 | 117 | -0.18093207 |
| 189494993 | 72 | 167 | 0 | 12 | 3052 | 187 | -0.07897098 |
| 189477803 | 4067 | 184 | 118 | 11 | 5464 | 313 | 0.28843826 |
| 189472916 | 4 | 35 | 123 | 14 | 433 | 83 | 0.36852590 |
| 189454027 | 3109 | 568 | 145 | 13 | 22363 | 353 | 0.00878569 |
| 189431923 | 2 | 105 | 45 | 14 | 1436 | 218 | 0.02956432 |
| 189427137 | 5404 | 595 | 124 | 12 | 15358 | 461 | 0.05534518 |
| 189408979 | 4297 | 280 | 0 | 12 | 8738 | 100 | -0.09887435 |
| 189401893 | 403 | 185 | 132 | 7 | 43643 | 114 | -0.09789994 |
| 189393236 | 2096 | 7216 | 75 | 7 | 83668 | 5791 | -0.34568670 |
| 189379709 | 1446 | 152 | 56 | 15 | 3538 | 177 | 0.01881331 |
| 189317892 | 8448 | 875 | 25 | 8 | 87596 | 186 | -0.06940554 |
| 189316564 | 385 | 4359 | 158 | 12 | 39419 | 290 | -0.06910443 |
| 256775091 | 423 | 1011 | 154 | 13 | 4523 | 1990 | 0.09851301 |
| 256643140 | 1135 | 2424 | 27 | 15 | 30965 | 734 | -0.13737923 |
| 256620591 | 349 | 920 | 41 | 15 | 19760 | 1764 | -0.15074906 |
| 256546478 | 3 | 24 | 0 | 9 | 39 | 73 | 0.07692308 |
| 256078225 | 274 | 84 | 124 | 6 | 3735 | 128 | -0.20582960 |

1 ▼ > >> | Number of rows: 25 ▼ Filter rows: Search this table

Figure 13. Sample of users' Twitter social media characteristics

# CHAPTER 6

# RESULTS AND FINDINGS

Let the beauty[2] of what you love[2]
be what you do.
[+5,-1]
Mevlana Rumi

In this chapter, two research questions of the study are answered. First, the sentiment analysis results of the Europe Countries during the chosen time period are shown. Then, the cross-sectional analysis results of users' social media account features and their happiness are listed.

## 6.1 Sentiment analysis results

The fourth research question of the study is about the GNH results of countries for chosen 6-year period: "What are the GNH polarities of European countries in accordance with the proposed Twitter sentiment analysis framework?". In order to answer this question, first of all, the yearly (average) results of the countries are found as in Table 14.

Table 14. Average GNH-TD of Countries for 6-Year Period

| Country | Average Sentiment Polarity |
|---|---|
| Germany | 0.040165 |
| Sweden | 0.040715 |
| France | 0.050131 |
| Netherlands | 0.055155 |
| Italy | 0.058553 |
| Spain | 0.085874 |
| United Kingdom | 0.104333 |
| Turkey | 0.105635 |
| Portugal | 0.132342 |

If these results are put in to a gradient color scale from light green (meaning lowest happy) to dark green (meaning highest happy), the resultant picture (Gradient Color GNH Map of Europe) would be as in Figure 14.

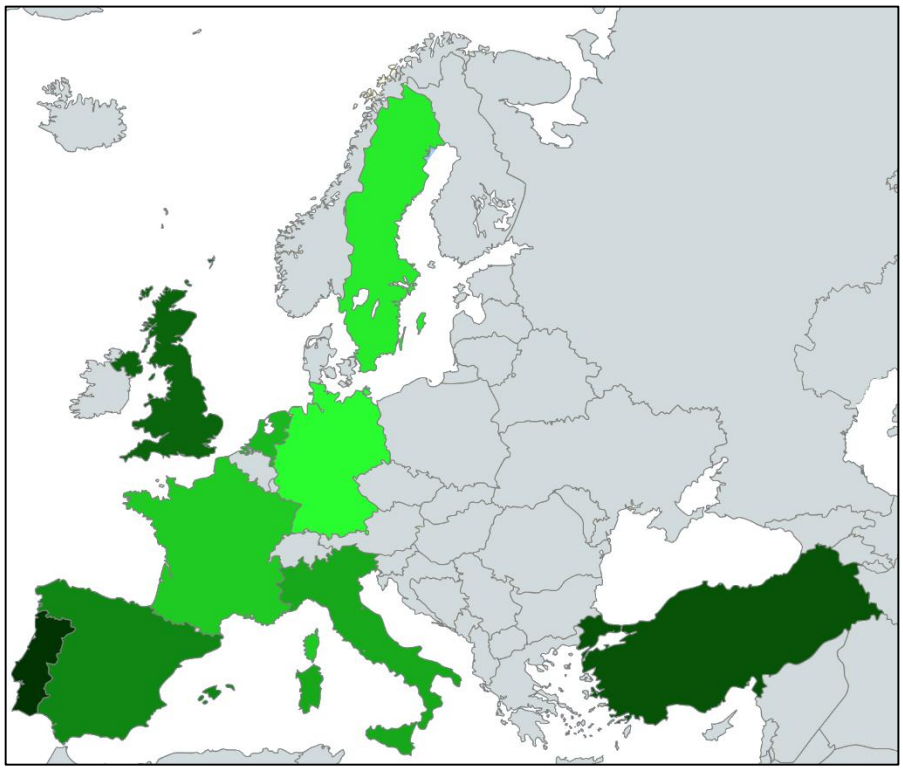


Figure 14. Gradient color map of GNH for 9 European countries between 2010 and 2015 (light green-lowest happy...dark green-highest happy)

But, this kind of aggregate figures are usually misleading. For avoiding this kind of misleading perspective, a detailed and comparable diagram should be designed. In Figure 15, yearly GNH-TD values for all countries are drawn.

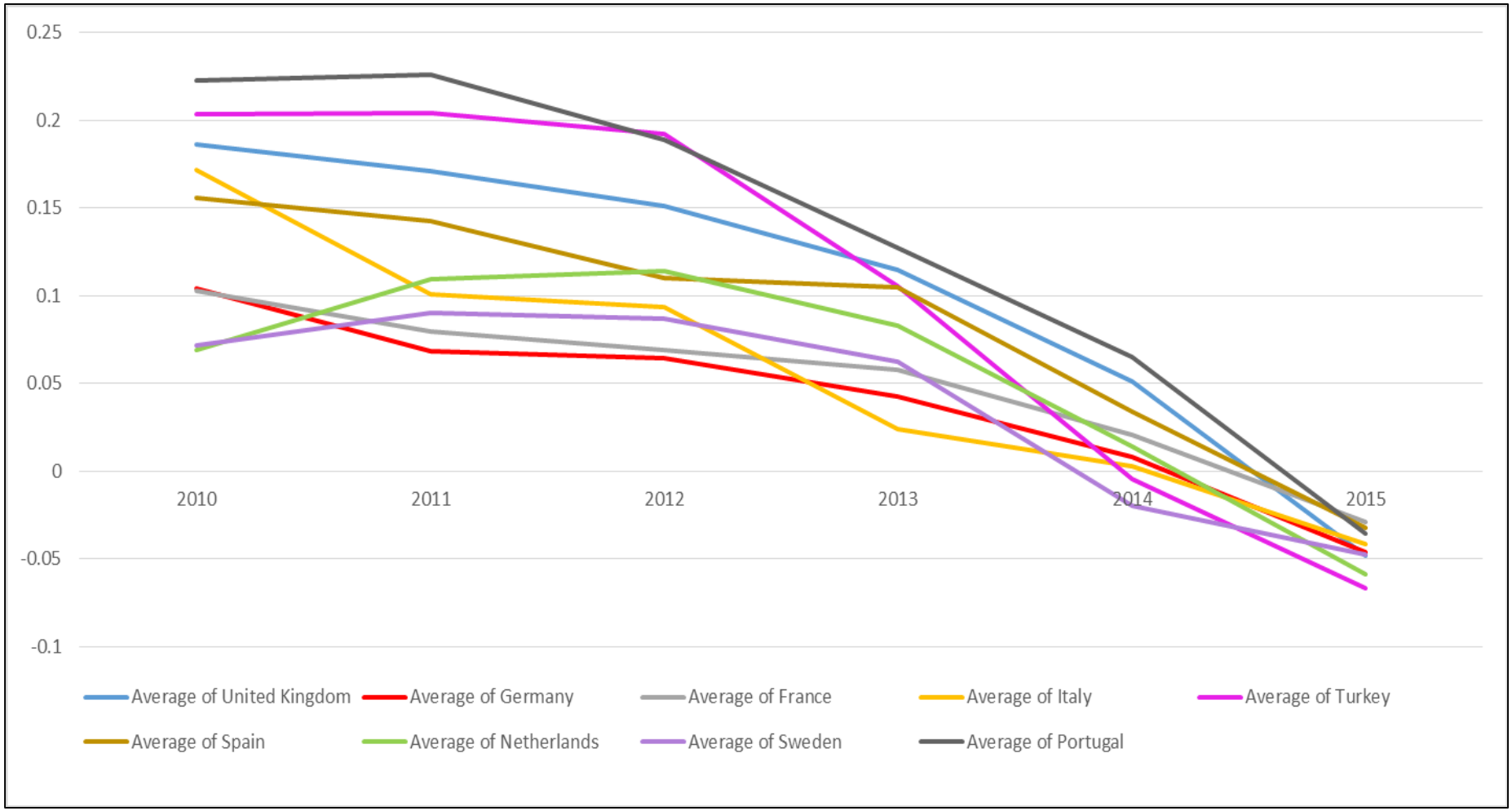



Figure 15. Yearly average GNH values of countries

This chart shows more detailed results and some of them can be listed as:

- A negativity trend appears in social media happiness of all countries through the six years period. This result is also approved by OECD Life Satisfaction results of countries, because those values are decreasing also year by year.
- France has changed its positive happiest level from 3rd unhappiest through six years.
- One of the most impressive results of the study, while Turkey starts with second highest (happiest) position in 2010 and in the second position in aggregate results (see Table 14); it is the unhappiest country among all at the end of 2015.

Before analyzing the countries one by one, in order to see the big picture of EU countries the total European daily sentiment results are determined as shown in Figure 16.

At first glance, a steady smooth trend appears from 2010 to the end of 2012 where afterwards a negative tendency arises.

When positive peaks of sentiment dates are considered, it is seen that the positive peaks are realized in Christmas Eve (24th December), Christmas Day (25th December) and New Year's First Day (1st January) for all years. The second positivity repeating event days (for all years) are Easter Days (4th April 2010, 24th April 2011, 8th April 2012, 31st March 2013, 20th April 2014) as an ordinary fact. But, since 2015 can be called as terrorism year which surrounded all the Europe, no Easter celebration appears in the Graph in 2015 April while it is a common fact for all other years. Lastly, a positive peak is shown in 13th May 2012 which is Mother's Day.

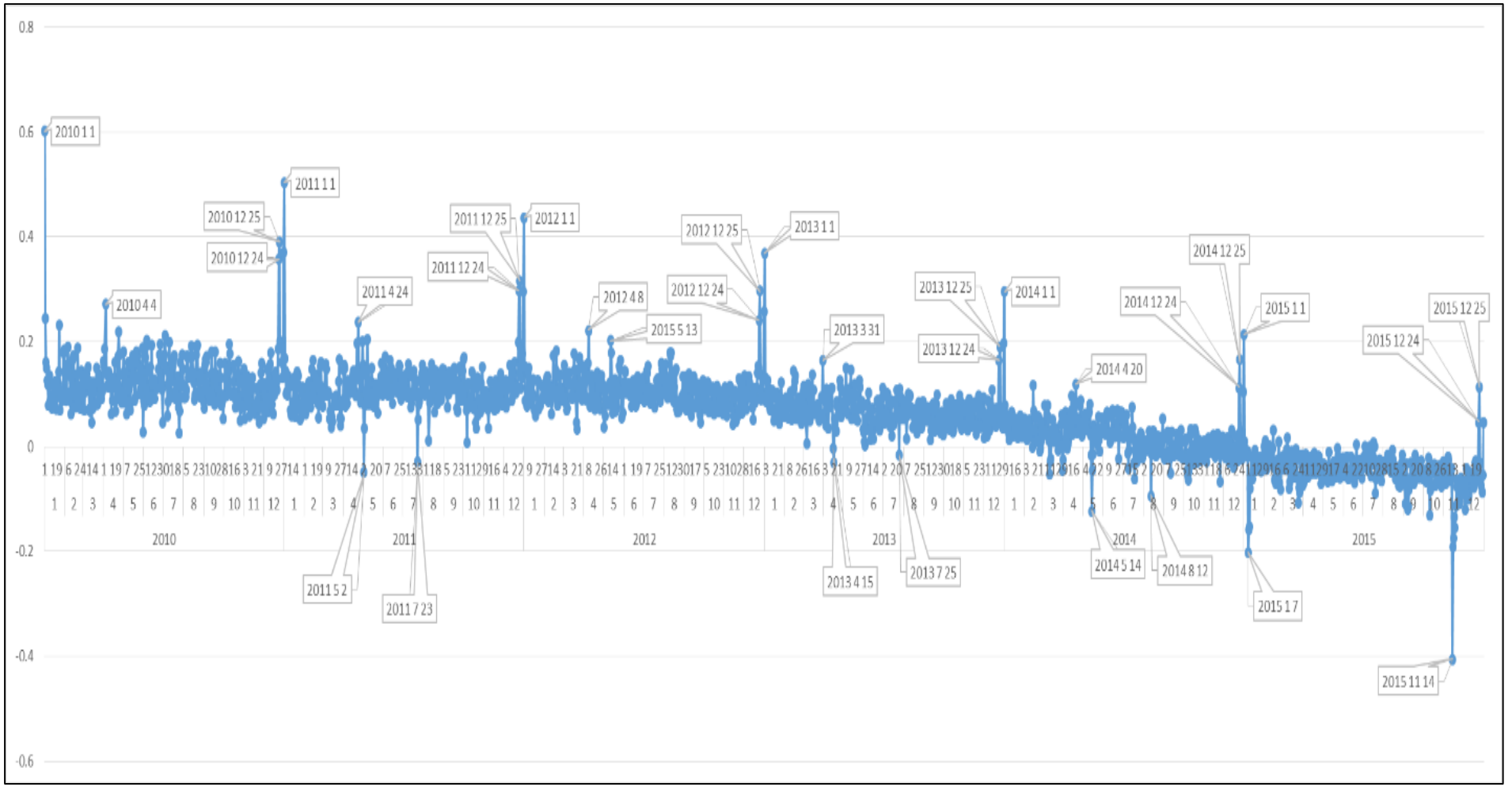



Figure 16. Daily sentiment polarities of EU countries from 2010 to 2015

Terrorist attacks and events have a dominance on the graph when the negative sentiment dates are analyzed. For instance, the most negative sentiment of Europe is 14th November 2015 in which terrorist attacks occurred in Paris (causing 137 deaths). Similarly, the Charlie Hebdo assault (7th January 2015) and Workers' Youth League (AUF)-run summer camp terrorist attack (23rd July 2011) are other negative sentiment days. 15th April 2013 is an interesting negative day; in this day there is not any negative event in Europe but a terrorist attack occurred in Boston Marathon in which there were lots of European participants (audiences and marathoners). Also, big accidental disasters (Soma mine explosion in Turkey on 14th May 2014 and train accident in Spain on 25th July 2013) appeared in the graph as most negative days. On the other hand, 12th August 2014 is also one of the unhappiest days as can be seen from the graph. On this day EUFA Super Cup final match was played between Real Madrid (the most popular club of Spain) and Sevilla (the oldest club of Spain) and Real defeated Sevilla (2-0). It seems that in Europe football fans did not like the victory of Real Madrid. Lastly, 2nd May 2011, as another negative day, is the day Osama Bin Laden died. At first glance, this day might be thought to be a positive day for Europe, but the tweets showed that it was a Remembrance Day for the thousands of innocents killed in the 9/11 event.

After the general analysis of Europe, the country datasets are graphed and examined specifically for stating country specific negative and positive peak dates. Findings, discussed in following sections, show that the algorithm is also successful for capturing the country-based social events as it is for EU countries in common.

### 6.1.1 Daily sentiment analysis for Germany

Figure 17 shows the daily sentiment polarities of Germany from January 1st 2010 to December 31st 2015. When the German dataset is investigated, 13th July 2014 when German national football team became world champion in Brazil is found to be one of the happiest days. Also, as a common fact, Saint Valentine's Day (14th February) appears as one of the happiest days in the scale for all years.

When we examine the negative days, interesting results are found. Firstly, as the unhappiest day of 2010, in July 24th, a massive stampede at the 2010 Love Parade in Duisburg killed 21 people and injured dozens (at least 500) more people. This Love Parade disaster affected German society very much and obviously led to a high position in GNH-TD results. Additionally, there are two nearly same degree negative days in 2011. In 27th March 2011, state elections were held in the Baden-Württemberg and Rhineland-Palatinate states. The negativity of the day was due to the fact that while Angela Merkel’s Christian Democrats had 39% and positioned the first, the total of Greens and Social Democrats became more than 40%, thus this result was concluded as “loss of Merkel”.

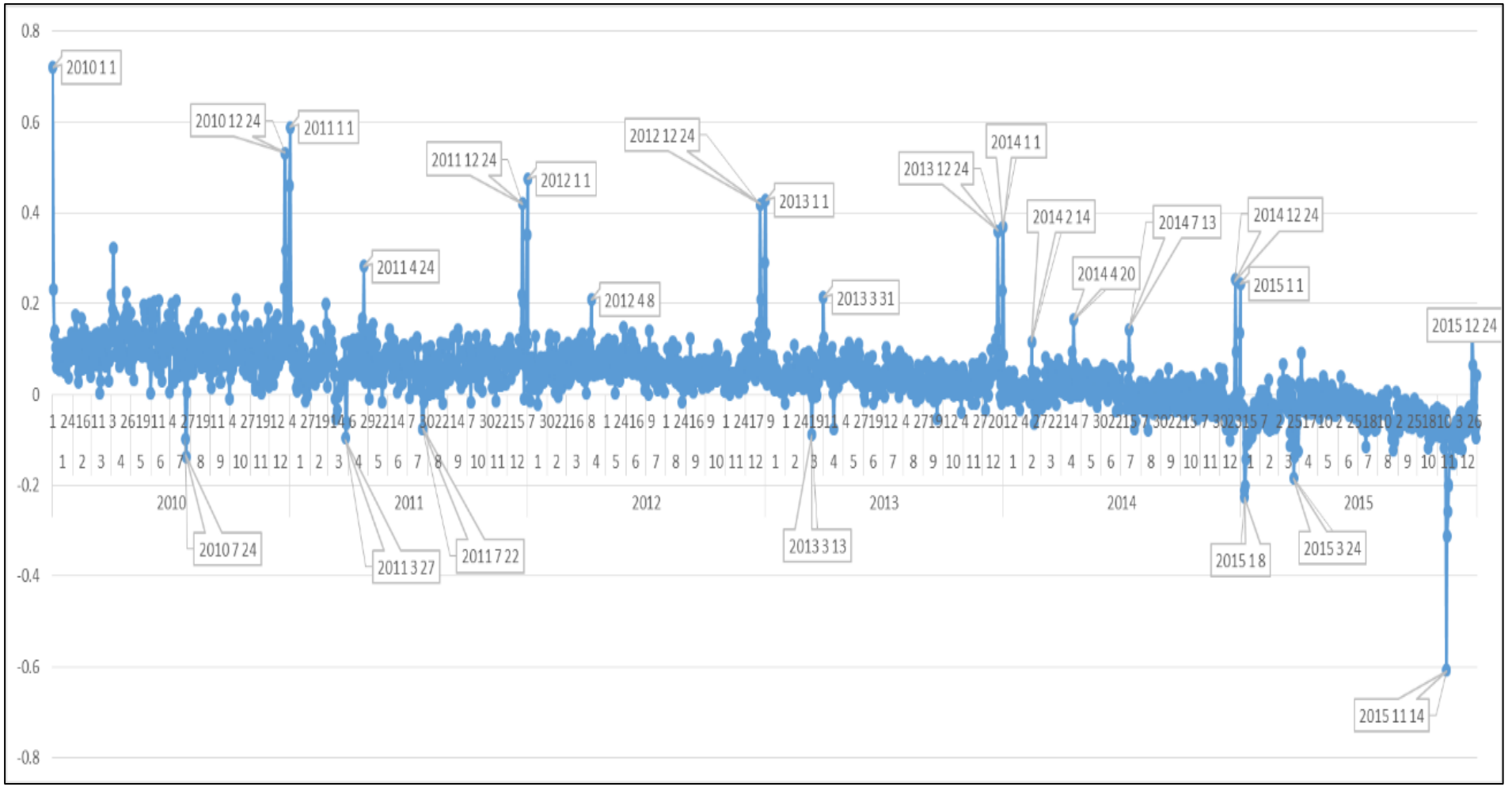



Figure 17. Daily sentiment polarities of Germany from 2010 to 2015

### 6.1.2 Daily sentiment analysis for Sweden

Figure 18 lists the daily sentiment polarities of the Sweden from 2010 to the end of 2015. The first interesting finding from Sweden dataset is that Swedish society does not react extraordinarily in social media and there is a smooth waving in their sentiment dataset. They don't react to 1st January as much as other EU countries and even their national day (June 6th) does not appear as extraordinarily positive in the scale when compared to other days. This emotionlessness for National Day of the country is very common for this country. It is a well-known fact that in 2004, the Swedish parliament started the discussions for making this day a public holiday in order to make society more interested in celebrating this day, but even the duration to end up with the decision for making it a public holiday took about one year. But the Midsummer Eve is an extraordinary social event (positive) for this country (e.g. 21st June 2013). As the general happiness tendency analysis, Swedish people has the happiest year in 2012 different from other countries.

The negative extraordinary days' analysis, as expected, showed that Paris attacks in 14th November 2015 has the most negative position in this country too. But, the negativity of 22nd July 2011 is more than 4th November 12015 in Sweden where 22nd July 2011 is the day of Norway terrorist attacks, and the reason of this high reaction may possibly be due to being a neighbor country of Norway is. As a result, it can be stated that, in Sweden terrorist attacks in Europe have negative effect on society while there are not widely peak positive days such as national days, religious events or sport events in society.

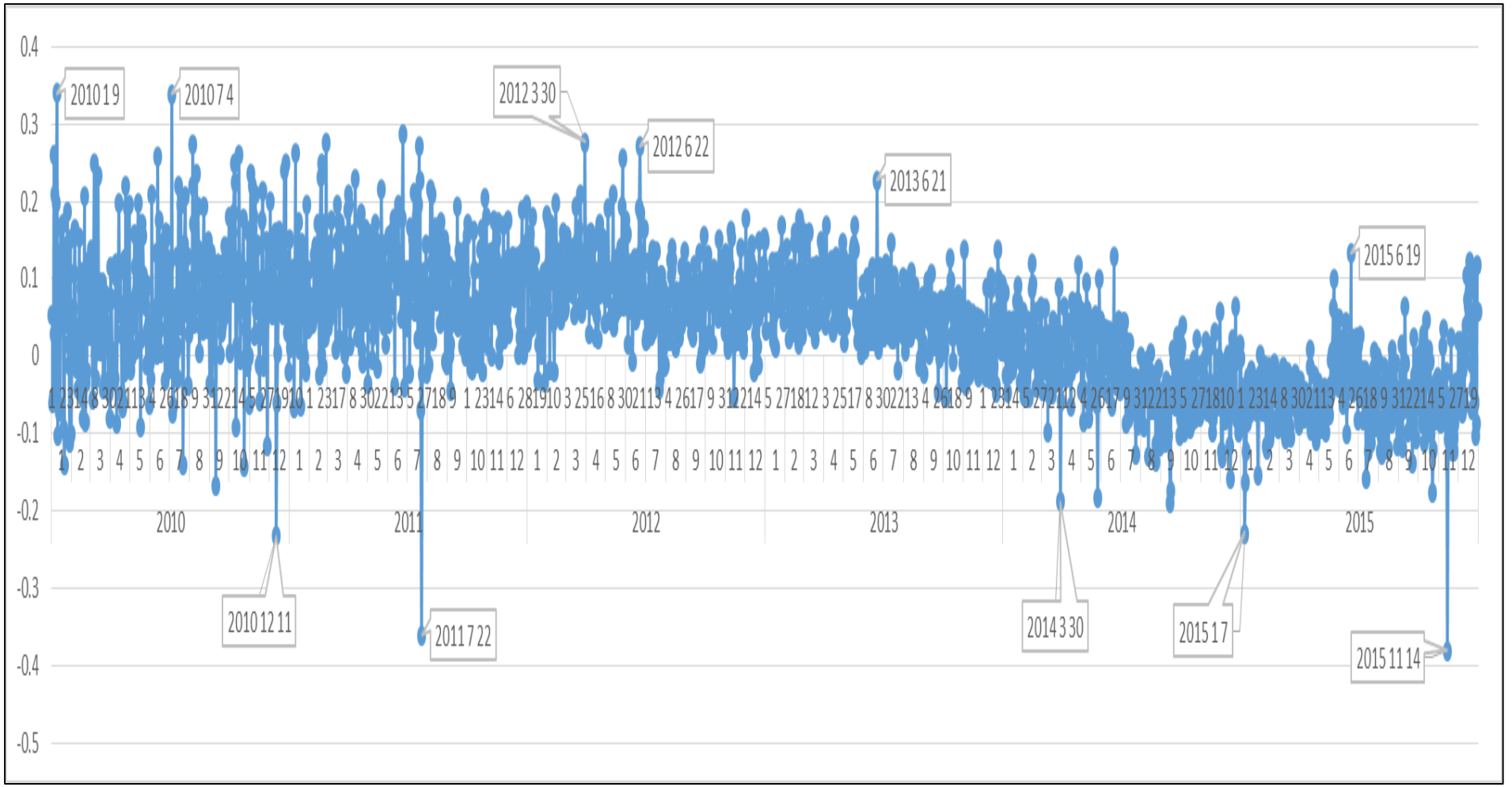



Figure 18. Daily sentiment polarities of Sweden from 2010 to 2015

### 6.1.3 Daily sentiment analysis for France

Daily GNH values of France from 2010 to 2015 are drawn in Figure 19. Daily GNH values of France showed that French people have a big tendency to celebrate New Year in 1st January. It is interesting that French people do not focus on Christmas (24th December) as much as others. This fact shows that in France, 1st January has a meaning of New Year than Christmas. On the other hand, extraordinary positivity of 23rd May 2010 showed a celebration of Whit Sunday. Also, 27th March 2011 is the Cantonal Election day in France and it seems the results gladden French people.

The negative extraordinary days' analysis for France show that Paris attacks (4th November 12015) and Charlie Hebdo shooting (7th January 2015) are the unhappiest day of the six-year period. Another negative date is in 21st March 2012 where a bombing attack occurred in front of the Indonesian Embassy in Paris after president Sarkozy declares the operation done to arrest the author of the Toulouse murders. To sum up, it is ordinary to find out French society is unhappy in 2015, but the general negative tendency in the society from 2010 would be a trace for this kind of results, not only in France but also in all over Europe.

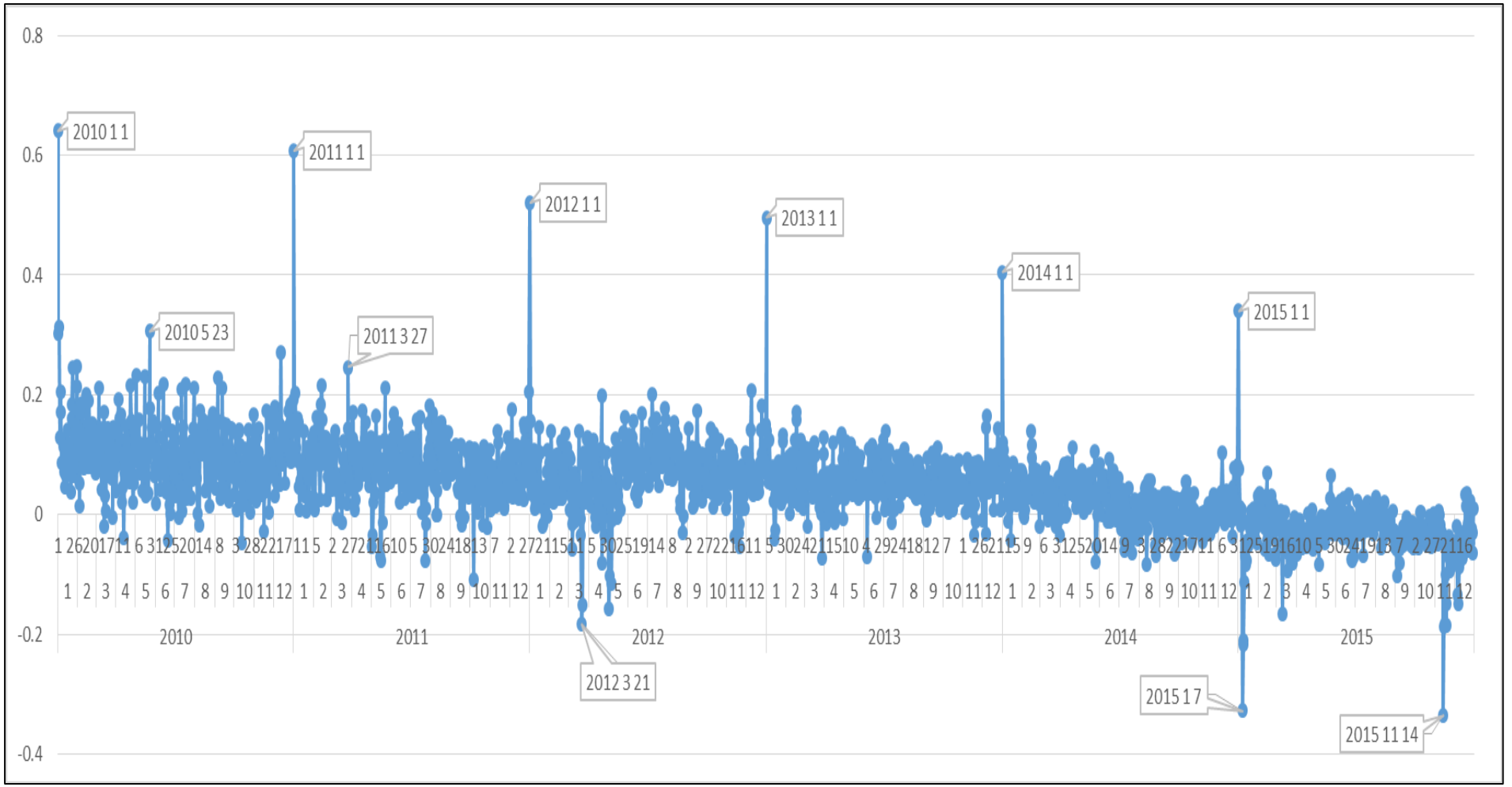



Figure 19. Daily sentiment polarities of France from 2010 to 2015

### 6.1.4 Daily sentiment analysis for Netherlands

Figure 20 shows the sentiment polarities from 2010 to 2015 in Netherlands. Netherlands' sentiment polarities state that 2012 is the happiest year for Netherland like Sweden, though a negative tendency of happiness in the six-year period.

The first negative day period in Netherlands was clearly on 24th February 2010 on which Queen Beatrix accepted the resignation of the Labor Party minister. However, in the June and August periods, new cabinet formation conversations were being done and positive and negative polarity days occurred depending on the direction of the discussions. On the other hand, in 12th July 2010 there was a negative polarity, and the reason for it was the World Cup defeat of Holland National Football Team by the Spanish team.

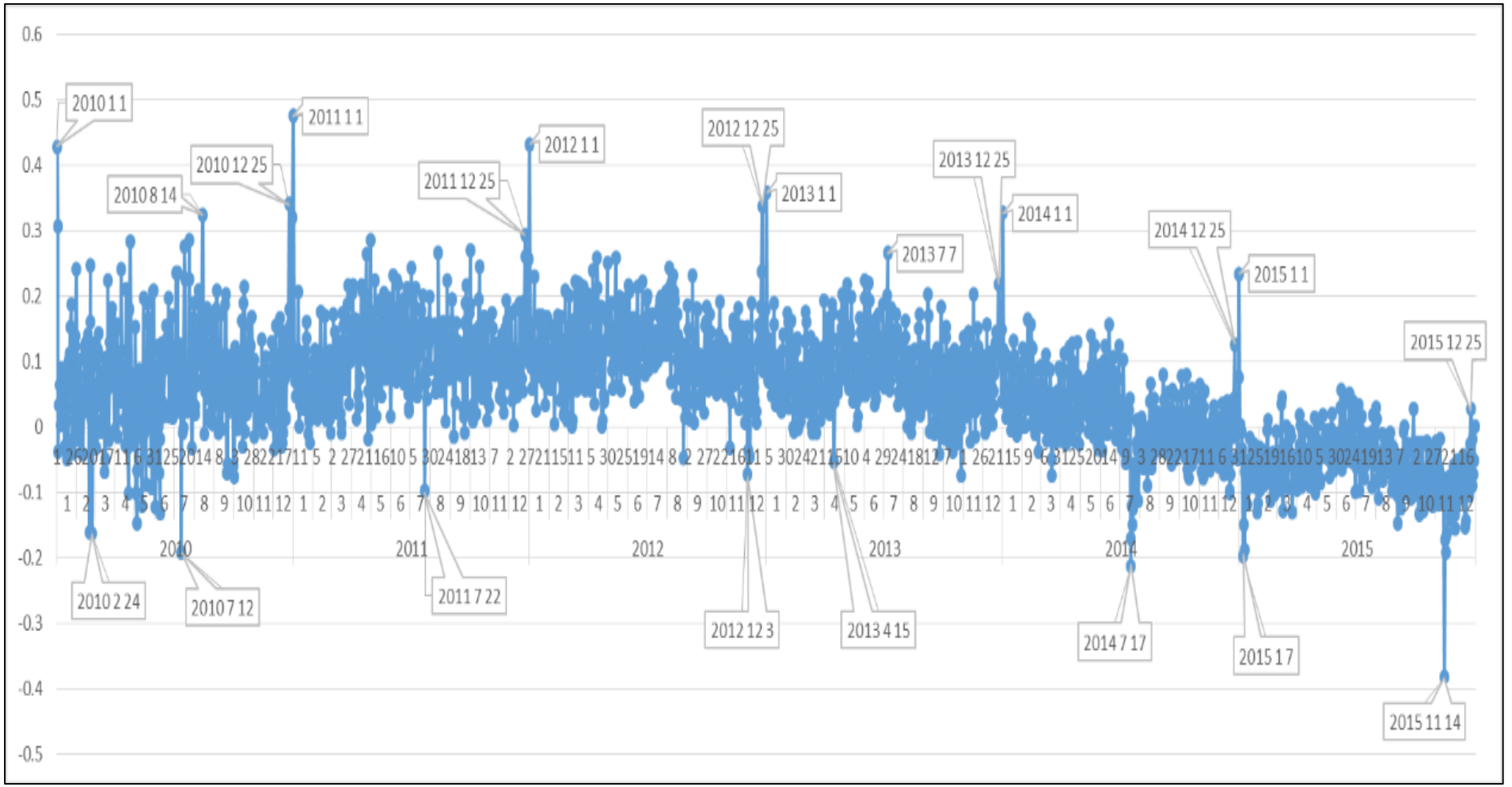



Figure 20. Daily sentiment polarities of Netherlands from 2010 to 2015

### 6.1.5 Daily sentiment analysis for Italy

While most of the other countries showed an increase in negativity in 2012, in this country negative flow GNH-TD is smooth for the six-year period (see Figure 21). On the other hand, while peaks of positive days appear in 2010 in high amount, these peaks immediately disappear after 2011. This can be an indicator for a rapid decrease of happiness in the Italy Report between 2012 and 2016 (Helliwell, Huang, & Wang, 2016). Another interesting finding for positive polarities is that, contrary to other EU countries, Italian citizens celebrate Christmas (25th December) instead of New Year (1st January).

The negativity on 19th May 2012 is because of Brindisi school bomb event which affected Italian society, who are not very familiar with terrorist attacks as other European countries, very much. At the same time the negative days show that, for all other terrorist attacks, Italy acts alike other European countries and feels unhappy.

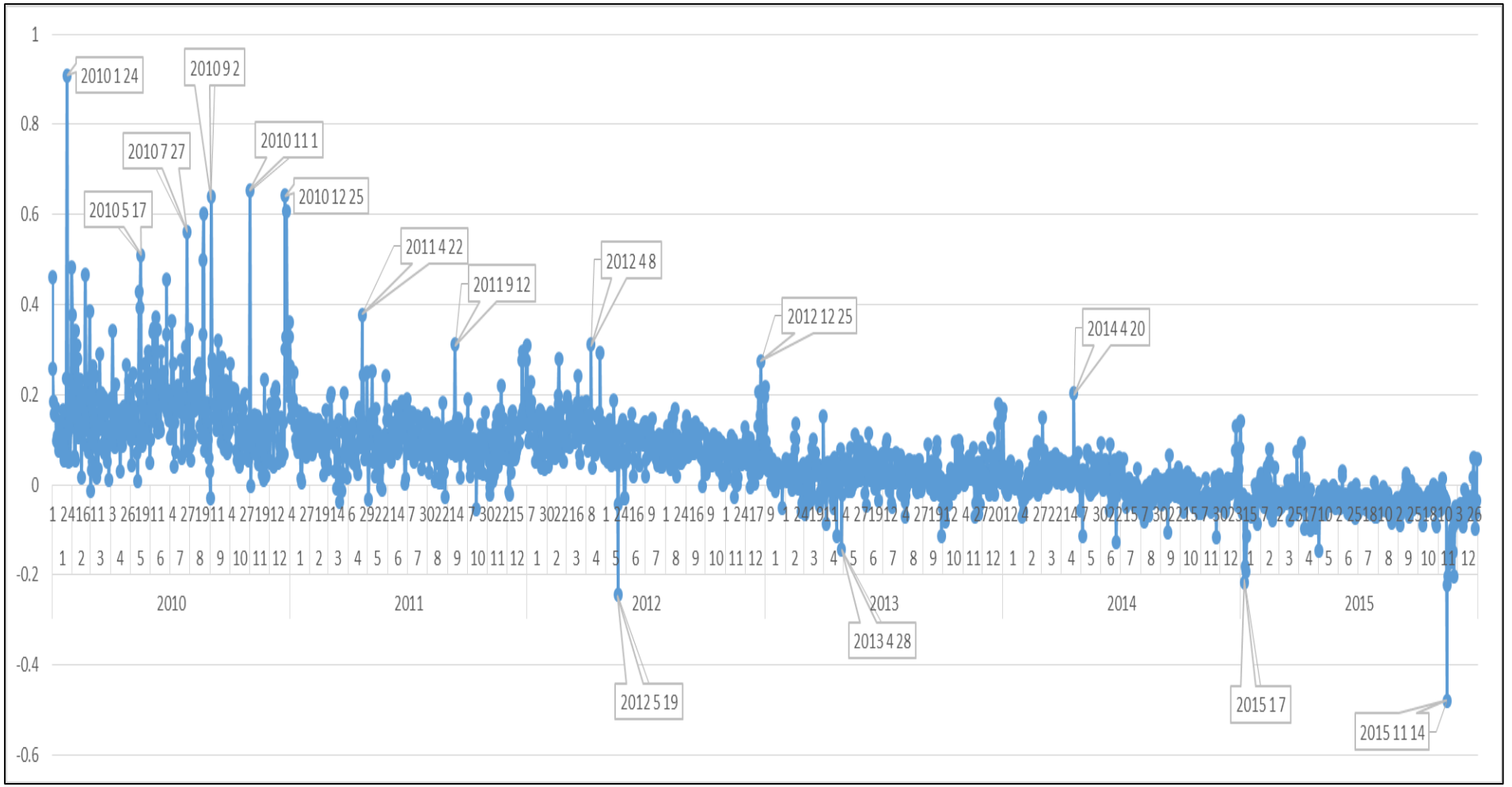



Figure 21. Daily sentiment polarities of Italy from 2010 to 2015

### 6.1.6 Daily sentiment analysis for Spain

In Spain, positive tendency on both Christ and New Year celebration time-periods appears very clear (see Figure 22). Additionally, just opposite to The Netherlands, in July 12th 2010, the national football team's victory shows a positive peak.

The negative days of Spain in the chosen time period is surprising since while 14th November 2015 terrorism event has a negative effect on the society the, other terrorist attacks in Europe do not indicate negativism. On the other hand, on 25th July 2013 on which the biggest train accident happened and dozens were killed is one of the negative peaks in Spain.

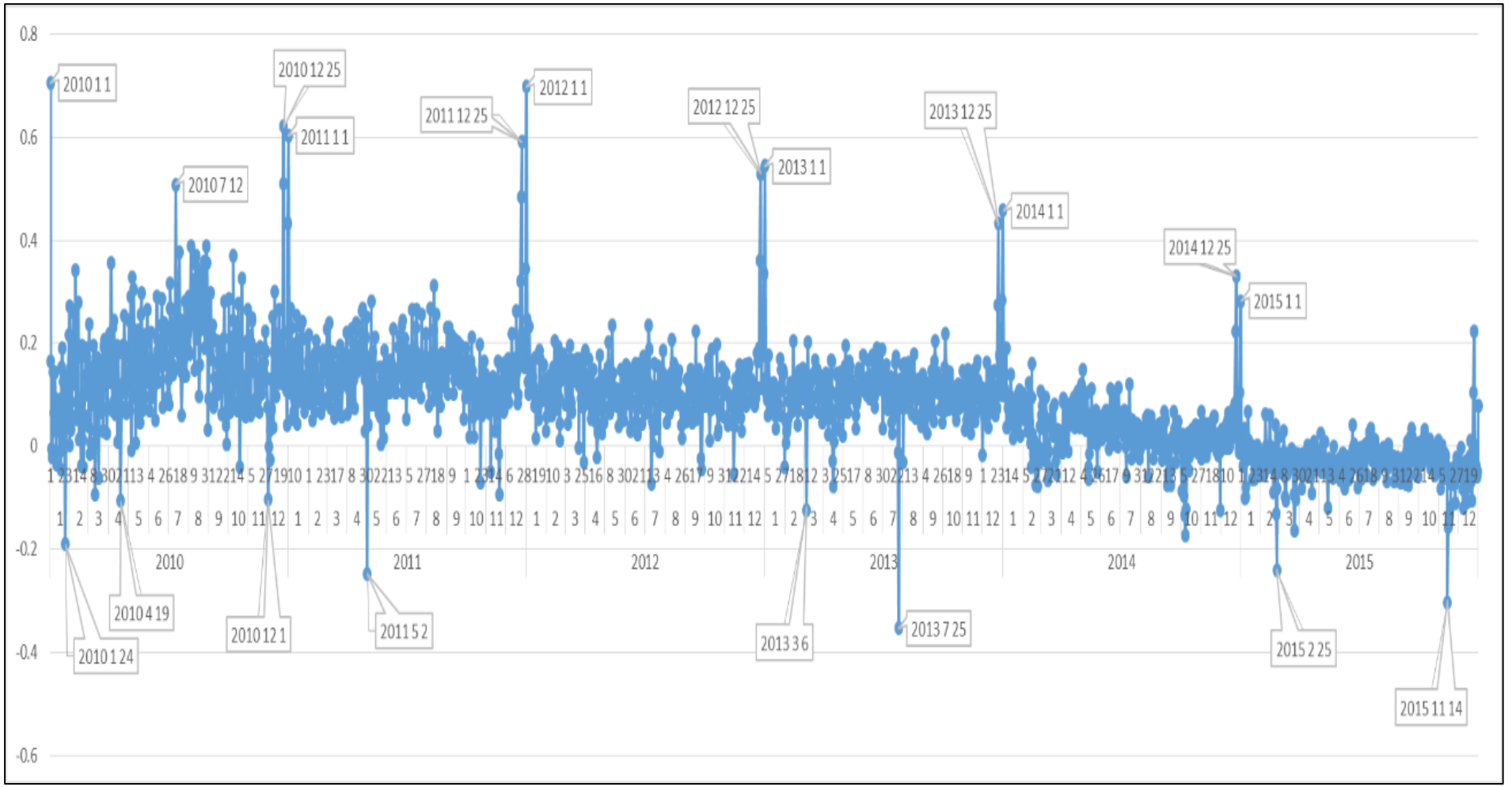



Figure 22. Daily sentiment polarities of Spain from 2010 to 2015

### 6.1.7 Daily sentiment analysis for United Kingdom

As still a part of the EU, United Kingdom society showed positive polarities on the Christmas and New Year celebrations, too. All other positive days are about football matches which supports the belief about the football focus of this society. (see Figure 23).

When the negative days are analyzed, the unhappiest day of all time interval is 14th November 2015, showing a big abhorrence to terrorism. On the other hand, like happiest days, most of other unhappiest days (e.g. 12th August 2014) are related to football events. Moreover, of course, 9th August 2011, as a domestic negative day, London riots and street fights appeared as one of the unhappiest day of 2011.

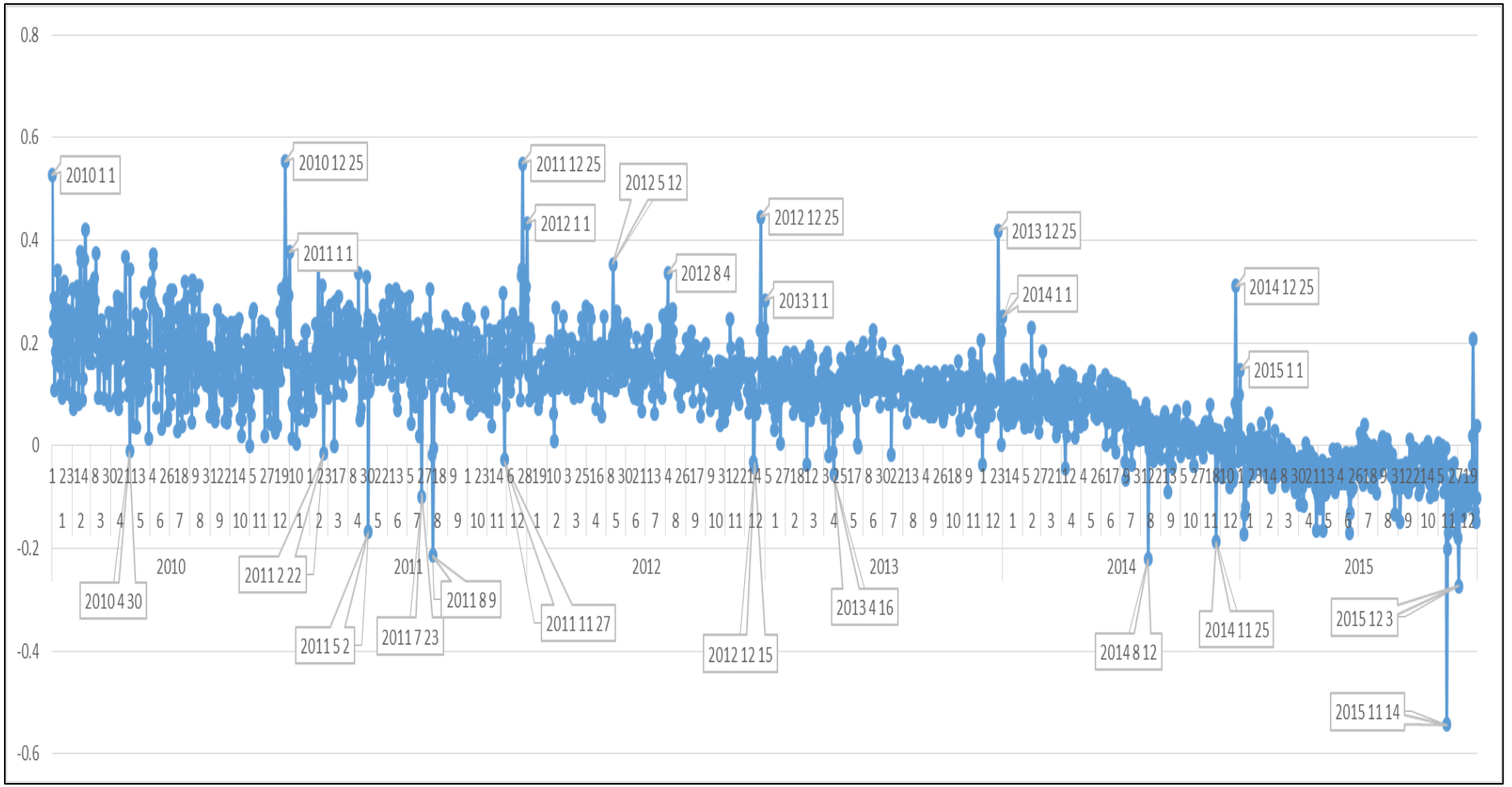



Figure 23. Daily sentiment polarities of United Kingdom from 2010 to 2015

### 6.1.8 Daily sentiment analysis for Turkey

Turkey is the only country that was chosen from the candidate countries pool of the EU. However, the results of this country (see Figure 24) shows the difference of EU citizenship. The positive GNH-TD days of Turkey is very different from other the EU countries. For instance, Turkish society does not celebrate New Year as a peak happy day of the year. Celebrating Christmas was not expected from this Muslim country but they use Gregorian calendar and New Year celebration would not be surprising. However, 15th November 2010 is a religious ceremony in İslamic World (Kurban) and 19th August 2012 has another ceremony (Ramadan); and these days have positive peaks. 30th August 2011 is Turkish national victory day and the positivity of this local celebration day is again natural.

The difference of Turkish society from EU countries is seen better in the negativity analysis of Turkish tweets. While 14th November 2015 is the unhappiest day for all other countries, it is not the unhappiest day in Turkey. Turkey has its own peak on 10th October 2015, where there was a terrific terrorist attack in capital city of Turkey (Ankara) resulting with more than 100 deaths and which was not too much considered as negative day by other EU countries. 14th May 2014 was the black day for Turkish citizens because of mine explosion in Soma with more than 200 deaths.

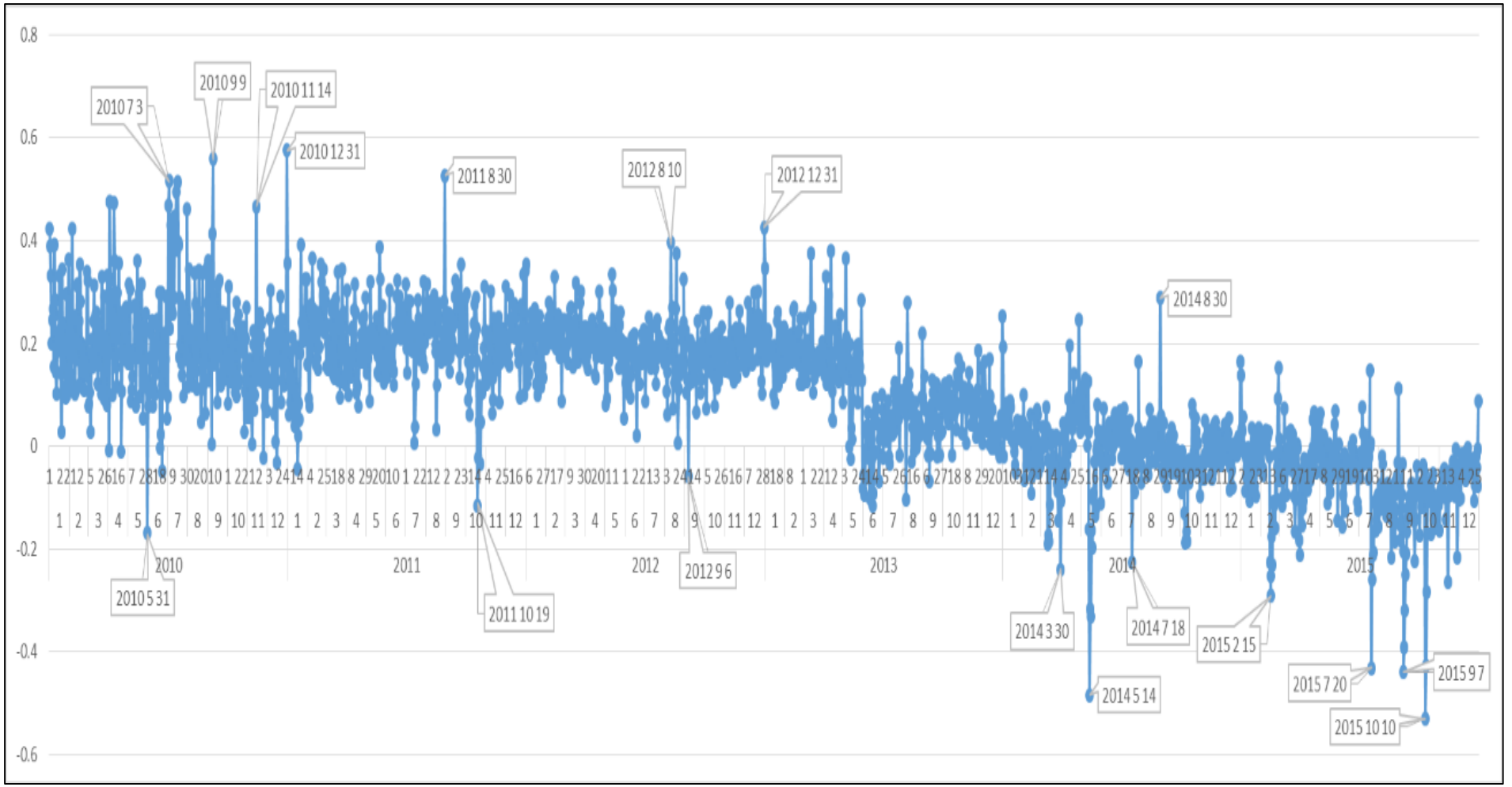



Figure 24. Daily sentiment polarities of Turkey from 2010 to 2015

### 6.1.9 Daily sentiment analysis for Portugal

Figure 25 shows GNH-TD values of the last country of the sample. When the Portugal dataset is analyzed, positive polarities are seen on 10th to 12th June of the years which are about Portugal Day celebrations.

The negative polarities show that as another EU country, Portugal shares the sadness of terrorist attacks. However, the results have a tremendous trough date (3rd May 2011) for this country, on which Portugal has reached an agreement with EU and IMF on 78 billion Euro financial rescue package, becoming the third Eurozone country to be bailed out of a sovereign debt crisis. This finding shows that economy is still one of the main factors of GNH for societies.

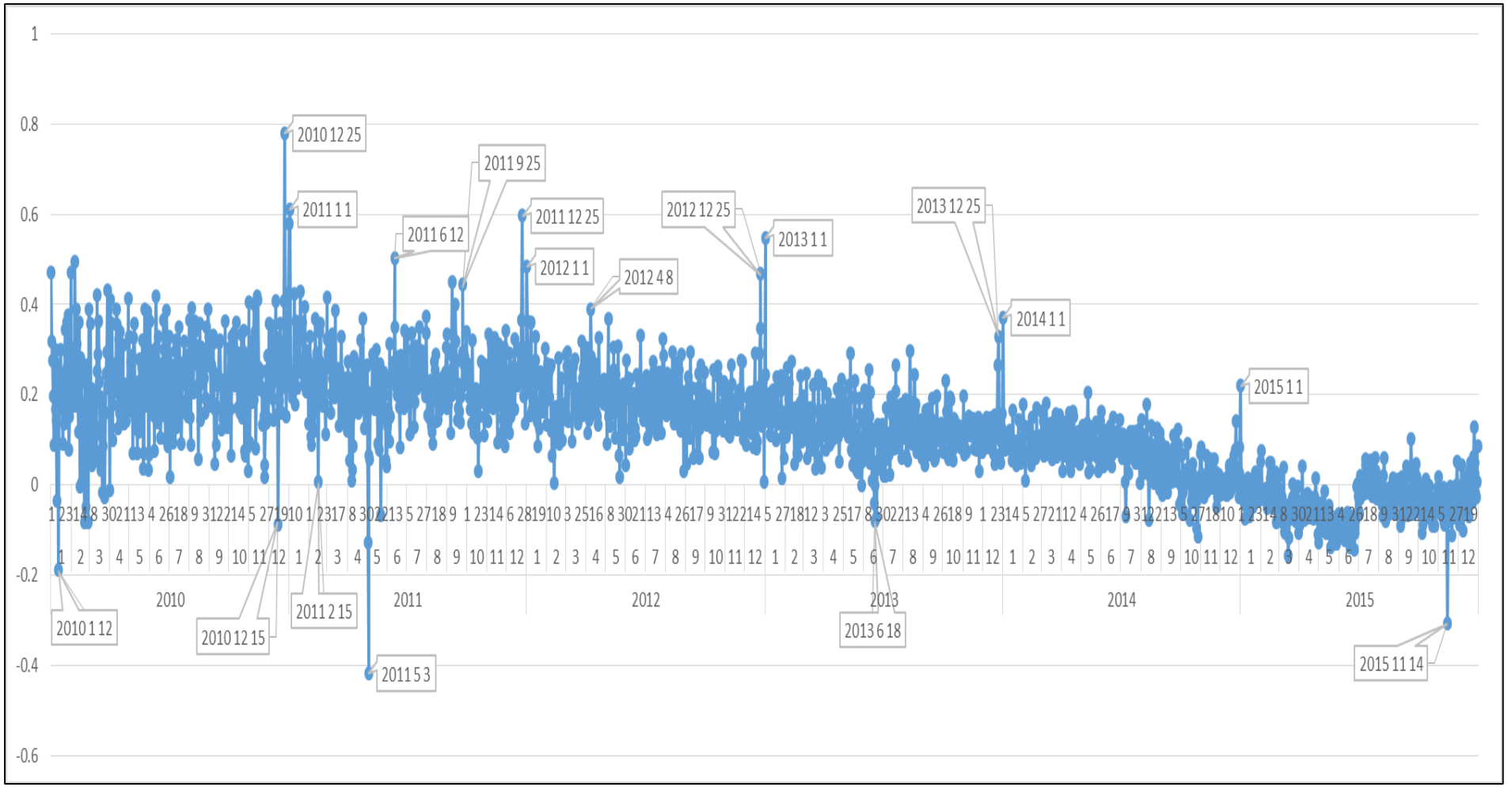



Figure 25. Daily sentiment polarities of Portugal from 2010 to 2015

## 6.2 Simple linear regression analyses result of Twitter social media characteristics and users' happiness levels

As it was mentioned in the Analyses Chapter, four direct and three calculated social media characteristics of 99,893 users from 9 countries are stored in the database for analyses. And the simple linear regression analyses between those variables and "happiness variable", which come from proposed sentiment analysis algorithm, were done.

### 6.2.1 Descriptive statistics

Descriptive results of the Twitter social media characteristics collected from 99.893 users between 1st Janurary 2010 to 31st December 2015 are listed in Table 15.

Table 15. Descriptive Statistics of Social Media Characteristics (1st January 2010 - 31st December 2015)

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| UserFavoritesCount | 99,893 | 0 | 504,698 | 1,597 | 7,229.368 |
| UserFollowersCount | 99,893 | 0 | 14,798,551 | 6,816 | 105,127.462 |
| UserTweetsCount | 99,893 | 386 | 1,640,653 | 12,861 | 29,440.234 |
| UserFriendsCount | 99,893 | 0 | 751,134 | 900 | 6,205.889 |
| Happiness | 99,893 | -2.2951 | 2.0000 | 0.078629 | 0.2260978 |
| LengthofUserDescription | 99,850 | 0 | 354 | 71 | 54.293 |
| LengthofUserScreenName | 99,893 | 2 | 16 | 10 | 2.728 |
| Age | 99,893 | 2,191 | 3,454 | 2,404 | 183.407 |

At first glance, the number of "LengthofUserDescription" variable's sample is not equal to 99893, it is 99850. This result is due to users who do not state any description (NULL) in their Twitter accounts.

Moreover, the average number of tweets per user in the sample is 12,861. This average value would seem very high, but when the average Twitter Account Age of the sample (2404 days) is taken into consideration, this average number of

tweets per account is not high. When the daily tweets average per account is found by dividing “average tweets per account” by “age of account (days)”; the results appears as 5.34 per day. If the proposed sampling and filtering methodology in this framework is remembered; the main aim was about capturing “active” accounts in order to calculate valid and representative GNH for countries. Thus, this amount of (5.34) daily average tweet publishing value is not surprising for the user dataset (Wikipedia, 2018).

In addition to these, the average number of followers of the sample (6816) and average number of friends of them (900) can be comparable. Since the Twitter calls “active users” as the ones who follow at least 30 accounts (Investopedia, 2018), average number of 900 is huge enough for capturing active users. Also, the massive difference between followers and friends (to whom people follow) shows that the created sample constitutes of “active users”.

### 6.2.2 Regression analyses results

As it is mentioned in Figure 1, relationships between the user characteristics and happiness level are analyzed using Simple Linear Regression Analyses for exploratory purposes. But, before the regression analyses, Pearson’s correlation analyses were done for finding the significance of the relationship between independent and dependent variables. Table 16 summarizes the results.

Table 16. Pearson's Correlation Summary

| | | User Favourites Count | User Followers Count | User Tweets Count | User Friends Count | Length of User Description | Age | Length of User Screen Name | Happiness |
|---|---|---|---|---|---|---|---|---|---|
| User Favourites Count | Pearson Correlation | 1 | 0.008 | 0.196 | 0.089 | 0.051 | 0.020 | 0.008 | 0.020 |
| | Sig. (2-tailed) | | 0.016 | 0 | 0 | 0.000 | 0 | 0.013 | 0 |
| | N | 99893 | 99893 | 99893 | 99893 | 99850 | 99893 | 99893 | 99893 |
| User Followers Count | Pearson Correlation | 0.008 | 1 | 0.172 | 0.167 | 0.026 | 0.037 | 0.003 | 0.002 |
| | Sig. (2-tailed) | 0.016 | | 0 | 0 | 0 | 0 | 0.410 | 0.451 |
| | N | 99893 | 99893 | 99893 | 99893 | 99850 | 99893 | 99893 | 99893 |
| User Tweets Count | Pearson Correlation | 0.196 | 0.172 | 1 | 0.194 | 0.145 | 0.067 | 0.026 | -0.059 |
| | Sig. (2-tailed) | 0 | 0 | | 0 | 0.000 | 0.000 | 0 | 0 |
| | N | 99893 | 99893 | 99893 | 99893 | 99850 | 99893 | 99893 | 99893 |
| User Friends Count | Pearson Correlation | 0.089 | 0.167 | 0.194 | 1 | 0.064 | 0.024 | 0.021 | 0.006 |
| | Sig. (2-tailed) | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | ,000 | 0.072 |
| | N | 99893 | 99893 | 99893 | 99893 | 99850 | 99893 | 99893 | 99893 |
| Length of User Description | Pearson Correlation | 0.051 | 0.026 | 0.145 | 0.064 | 1 | 0.116 | 0.076 | 0.043 |
| | Sig. (2-tailed) | 0 | 0 | 0 | 0 | | 0 | 0 | 0 |
| | N | 99850 | 99850 | 99850 | 99850 | 99850 | 99850 | 99850 | 99850 |
| Age | Pearson Correlation | 0.020 | 0.037 | 0.067 | 0.024 | 0.116 | 1 | -0.104 | -0.052 |
| | Sig. (2-tailed) | 0 | 0 | 0 | 0 | 0 | | 0.000 | 0 |
| | N | 99893 | 99893 | 99893 | 99893 | 99850 | 99893 | 99893 | 99893 |
| Length of User Screen Name | Pearson Correlation | 0.008 | 0.003 | ,026** | ,021** | ,076** | -,104** | 1 | 0.038 |
| | Sig. (2-tailed) | 0.013 | 0.410 | 0 | 0 | 0 | 0 | | 0 |
| | N | 99893 | 99893 | 99893 | 99893 | 99850 | 99893 | 99893 | 99893 |
| Happiness | Pearson Correlation | 0.020 | 0.002 | -0.059 | ,006 | 0.043 | -0.052 | 0.038 | 1 |
| | Sig. (2-tailed) | 0 | 0.451 | 0 | 0.072 | 0 | 0 | 0 | |
| | N | 99893 | 99893 | 99893 | 99893 | 99850 | 99893 | 99893 | 99893 |

As it is simply seen, all social media characteristics are significantly correlated with “happiness”, except for “User Followers Count”. Thus, this variable is excluded from the simple linear regression analysis. The results of regression analyses are summarized in Table 17.

Table 17. Simple Linear Regression Analyses Summary

| Variable | Model Summary | | | | ANOVA Summary | | Coefficients Summary | | |
|---|---|---|---|---|---|---|---|---|---|
| | R | R Square | Adjusted R Square | Std. Error of the Estimate | F | Sig. | Constant | B | Sig. |
| User Favourites Count | 0,020 | 0,001 | 0,001 | 0,226055 | 38,57 | 0 | 0,078 | 0,06145 | 0 |
| User Tweets Count | 0,059 | 0,004 | 0,004 | 0,225699 | 354,37 | 0 | 0,085 | -0,04566 | 0 |
| User Friends Count | 0,006 | 0,001 | 0,001 | 0,226095 | 3,24 | 0,072 | 0,078 | 0,02075 | 0,072 |
| Length of User Description | 0,043 | 0,002 | 0,002 | 0,225913 | 189,19 | 0 | 0,066 | 0,00102 | 0 |
| Age | 0,052 | 0,003 | 0,003 | 0,225789 | 274,42 | 0 | 0,234 | -0,06453 | 0 |
| Length of User Screen Name | 0,038 | 0,001 | 0,001 | 0,225936 | 144,20 | 0 | 0,047 | 0,00300 | 0 |

At first glance, the model summary in Table 17 states that all the simple linear regression results have very low R Square and Adjusted R Square values. These results can be concluded as “small” effect size (Cohen, 1992) and according to Sullivan and Feinn (2012) to cope with sample size should be enlarged. But, this is done by big data study and the sample size (99,893) is extremely enough. Then, what

would be the cause of small effect size and significant relations? At this point, Ford (2015) explains that R-Squared does not measure goodness of fit and predictive error, because it cannot explain how one variable explains another. Therefore, the dependent variable ("happiness") of the model's regression analyses should not be ignored. If all the things that might affect someone's "happiness" are taken into consideration, explaining a small part of this large variation is also very valuable.

The Analysis of Variance (ANOVA) summary in Table 17, additionally, states that with given F values all the regression models are significant in 95% confident level, except for User Friends Count (it is significant on 90% confident level). Thus, the coefficients of the models can be analyzed for stating the positive and negative relationships between users' account-characteristics and their happiness levels.

The coefficients with in the model show that following variables significantly increase the "happiness" of users on Twitter social media:

- User Favorites Count
- User Friends Count
- Length of User Description
- Length of User Screen Name

First, it can be specified unsurprisingly that, when a user's number of favorited tweets increase, her/his happiness also increases. In addition, the increase on number of friends (followees, number of people s/he follows) causes rise of happiness, possibly due to the fact that people feel happier when they become "socialized" with high number of friends.

As other interesting findings, the longer description and screen name mean more happiness in social media. While the recent studies about "bot account

detection" (Clark et al., 2016; Echeverria & Zhou, 2017; Varol, Ferrara, Davis, Menczer, & Flammini, 2017) focus on description and screen name characteristics of Twitter, their positive significant relationship with happiness has not been stated yet in the literature. These positive relationships would be derived by several psychological reasons such as being "extravert" or "introvert" as it is mentioned by Hunsinger, Isbell, and Clore (2012) and Rousseau (1996).

In addition to these positive relations, two characteristics have significantly negative affect on "happiness" of users. Those characteristics are "number of tweets" and "Twitter account age". First relationship means when the number of tweets increases the happiness of users decreases significantly. Bollen, Mao, and Pepe (2011) state in their social media study that public moods has a variation effect of number of tweets on social event days. Thus, this relationship may because of public effect on users. Or, this result may be due to being more relaxed to publish negative feelings after being more active (more tweet number) on Twitter.

Lastly, the Twitter age characteristics has a significantly negative effect on happiness. Thus, the older you are on Twitter (not actual age, but account age) the sadder tweets you publish. In 23 March 2016, Microsoft released an artificial intelligence chatter bot called "Tay" (acronym: "thinking about you") which was designed as a 19-year old American girl. She interacted with followers on Twitter and learned to be a social media user via conversations (Mason, 2016). On the other hand, within 16 hours (more than 40 million conversations and 96000 tweets) Microsoft suspended Tay's Twitter account, because she immediately began a "racist (Hitler fun), sexist, war supporter and enemy of human beings". If we consider that the selected users in the sample are "active" users (publishing average 5.34 tweets

every day), the older accounts mean more interactivity on Twitter. Thus, the negative effect of this Twitter age variable supports the results of failed Tay experiment.

# CHAPTER 7

# CONCLUSION AND DISCUSSION

Happiness[2] is when
what you think, what you say,
and what you do are in harmony[1].
[+4,-1]
Mahatma Gandhi

This Twitter social media big data analysis study is about learning from the past in country levels and detecting exploratory findings in multicultural and multilingual levels. With this perspective, a novel social media big data sentiment analysis framework, which consists of data collection, filtering, sampling and sentiment analysis algorithm, was conducted. In this respect, 11 countries have been chosen from Europe for Gross National Happiness Analysis with Twitter data. 2 countries (Poland and Greece) are dropped from this dataset due to not reaching the own language usage ratio (1/5000) of the total population in Twitter. After filtering, more than 110,000 active users from nine countries were accessed and their tweets from 1st January 2010 to 31st December 2015 were collected. After validating the algorithm results with convergent and face validity analyses, the proposed sentiment analysis framework was found to be reliable (greater than 70% threshold value, (Rost & Sander, 1993)) when checked with news archive. Lastly, with this validated and reliable algorithm, GNH are calculated and the results are discussed in general and on country domains. Investigating the results deeply, terrorist attacks and disasters (air crashes etc.) have naturally negative effects on society soul. Also, in Europe countries, society is affected by terrorist attacks not only in their country but also in other countries. This result concludes that in (especially for negative dates)

extraordinary situations, there still exists a “European Citizenship” concept. Also, unfortunately a tendency for increase in negative sentiment appears in GNH of all countries over the 6-year period.

When the proposed Twitter social media sentiment analysis framework is compared to alternative approaches, this framework can be found conspicuous with its following newly designed features:

- This framework does not only include sentiment analysis algorithm but also contains data collection, sampling and filtering methodology which are the main challenges of big data analysis (Nakov, Rosenthal, et al., 2016; Tole, 2013).
- The usability of the proposed framework meets a deficit, tested and validated for multiple languages, which was declared as future study recommendations of several studies (Giachanou & Crestani, 2016; Nakov, Ritter, Rosenthal, Sebastiani, & Stoyanov, 2016; Nakov, Rosenthal, et al., 2016).
- In addition to accessing the threshold value (70%) stated by Rost and Sander (1993), some of the accuracy of the stated framework results are more than the results of Pang, Lee, and Vaithyanathan (2002) (~75%) which high accuracy of machine learning method and results of (Poria, Cambria, Howard, Huang, & Hussain, 2016) (80%) in which feature- and decision-level fusion methods are used.
- Lastly, it can be stated that comparing to the survey-based methodology of GNH calculation by the global institutions (e.g. OECD), time series results (daily, monthly etc.) can be drawn and explained with this proposed framework. Thus, this promising framework can contribute the researchers for related specific social psychology studies.

As the exploratory analyses, after the “Twitter social media happiness” variable is created with the proposed framework, the relationship between this new variable and other collected Twitter social media user characteristics are analyzed. Results showed that the number of user’s favorited tweets, the number of friends, length of account description and length of screen name have significantly positive effects on happiness. On the other hand, the simple linear regression analyses conclude that when the number of tweets a user publishes increases, the happiness of her/him decreases. Lastly, the duration that a user spend on Twitter significantly decreases the happiness.

# CHAPTER 8

# LIMITATIONS AND FUTURE STUDY RECOMMENDATIONS

The main aim of this study is to find out societies' total sentiments in daily (even hourly) levels via Twitter social media big data analysis (GNH) both in country and user domains. Although the research has reached its aim, there were some unavoidable limitations. First, though there may be various variations in the dictionaries, it is assumed in the study that they are similar. Second, since OECD has declared only yearly based better life indices of countries starting from 2012, convergent validity is done only using 36 cases. Lastly, related to reliability, only one reference, Wikipedia, is used and is assumed that the source is valid.

As further studies, using the GNH results of the countries and/or the users, it can be recommended that the following researches can be done.

- The daily and hourly results using the algorithm of this study might be analyzed in a deeper way with the help of social psychologists in terms of socio-cultural effects.
- In order to enlarge the scope of cross-cultural analysis, the variations in the usage of most frequently used words, idioms and emoticons can be examined to detect exploratory differences between societies.
- Since machine translation of tweets for sentiment analysis is an alternative methodology for multi-lingual sentiment analysis (Chaturvedi, Cambria, & Vilares, 2016), the framework of this study can be integrated with machine translation via deep learning and fuzzy logic methodologies to determine GNH of countries on a common language

- GNH results to be obtained using the proposed framework of this study can be used for dynamic marketing strategy developments and also for global companies' supply management decisions.
- By using the same sentiment algorithm results, GNH calculation algorithm can be adapted for event domain to find event based GNH for countries.

# REFERENCES


AbdelFattah, M., Galal, D., Hassan, N., Elzanfaly, D. S., & Tallent, G. (2017). A sentiment analysis tool for determining the promotional success of fashion images on Instagram. *International Journal of Interactive Mobile Technologies, 11*(2).

Abdullah, S., Murnane, E. L., Costa, J. M., & Choudhury, T. (2015). Collective smile: Measuring societal happiness from geolocated images. In D. Cosley, A. Forte (Eds.) *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.* (pp. 361-374). New York, NY: ACM Press.

Ahkter, J. K., & Soria, S. (2010). Sentiment analysis: Facebook status messages. (Unpublished master thesis). Stanford University, CA. Retrieved January 3, 2017 from http://www-nlp.stanford.edu/courses/cs224n/2010/reports/ssoriajr-kanej.pdf

Ahmad, K., & Almas, Y. (2005). Visualising sentiments in financial texts? In E. Banissi, M. Sarfraz, J.C. Roberts, B. Loften, A. Ursyn, R.A. Burkhard, A. Lee, G. Andrienko (Eds.) *Proceedings of the 9th International Conference on Information Visualisation* (pp. 363-368). Piscataway, NJ: IEEE. doi: 10.1109/IV.2005.88

Akaichi, J., Dhouioui, Z., & Pérez, M. J. L.-H. (2013). Text mining facebook status updates for sentiment classification. In E. Petre, M. Brezovan (Eds.) *Proceedings Proceedings of the 17th International Conference on System Theory, Control and Computing* (pp. 640-645). Piscataway, NJ: IEEE.

Akgül, E. S., Ertano, C., & Banu, D. (2016). Twitter verileri ile duygu analizi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 22*(2), 106-110.

Angelini, V., Cavapozzi, D., Corazzini, L., & Paccagnella, O. (2014). Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual- Specific Scale Biases. *Oxford Bulletin of Economics and Statistics, 76*(5), 643-666.

Ayata, D., Saraçlar, M., & Özgür, A. (2017, May). *Turkish tweet sentiment analysis with word embedding and machine learning*. Paper presented at the 25th Signal Processing and Communications Applications Conference, Antalya, Turkey.

Baucom, E., Sanjari, A., Liu, X., & Chen, M. (2013). Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. In X. Liu, M. Chen, Y. Ding, M. Song (Eds.) *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing* (pp. 61-68). New York, NY: ACM Press.

Beasley, A., & Mason, W. (2015). Emotional states vs. emotional words in social media. In D. Roure, P. Burnap, S. Halford (Eds.) *Proceedings of ACM Web Science Conference* (p. 31). New York, NY: ACM Press.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45-59.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In N. Nicolov, J.G. Shanahan, L. Adamic, R.B. Yates, S. Counts (Eds.) *Proceedings of International AAAI Conference on Weblogs and Social Media* (pp. 450-453). Menlo Park, CA: AAAI Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662-679.

Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., & Hanson, C. L. (2016). Validating machine learning algorithms for twitter data against established measures of suicidality. *Journal of Medical Internet Research - Mental Health, 3*(2).

Brandtzaeg, P. B. (2017). Facebook is no "Great equalizer" A big data approach to gender differences in civic engagement across countries. *Social Science Computer Review, 35*(1), 103-125.

Bravo-Marquez, F., Frank, E., & Pfahringer, B. (2015). From unlabelled tweets to twitter-specific opinion words. In R.B. Yates, M. Lalmas, A. Moffat, B.R. Neto (Eds.) *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 743-746). New York, NY: ACM Press.

Burnap, P., Colombo, W., & Scourfield, J. (2015). Machine classification and analysis of suicide-related communication on twitter. In Y. Yesilada, R. Farzan, G.J. Houben (Eds.) *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 75-84). New York, NY: ACM Press.

Casas, D. C. d. L., Magno, G., Cunha, E., Andr, M., Almeida, V. (2014, June). *Noticing the other gender on Google+*. Paper presented at the 2014 ACM Conference on Web Science, Bloomington, IN.

Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016, February). *Quantifying and predicting mental illness severity in online pro-eating disorder communities.* Paper presented at the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, San Francisco, CA.

Chaovalit, P., & Zhou, L. (2005, January). *Movie review mining: A comparison between supervised and unsupervised classification approaches.* Paper presented at the 38th Annual Hawaii International Conference on System, Big Island, HI.

Chaturvedi, I., Cambria, E., & Vilares, D. (2016, September). *Lyapunov filtering of objectivity for Spanish sentiment model.* Paper presented at the Neural Networks (IJCNN), 2016 International Joint Conference, Vancouver, British Columbia, Canada.

Chen, C., Chen, F., Cao, D., & Ji, R. (2015). A cross-media sentiment analytics platform for microblog. In X. Zhou, A. Smeaton, Q. Tian (Eds.) *Proceedings of the 23rd ACM International Conference on Multimedia* (pp. 767-769). New York, NY: ACM Press.

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011, July). *Exploring millions of footprints in location sharing services*. Paper presented at International AAAI Conference on Weblogs and Social Media, Barcelona, Spain.

Clark, E. M., Williams, J. R., Jones, C. A., Galbraith, R. A., Danforth, C. M., & Dodds, P. S. (2016). Sifting robotic from organic text: a natural language approach for detecting automation on Twitter. *Journal of Computational Science, 16*, 1-7.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101.

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015, June). *From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses.* Paper presented at the CLPsych@ HLT-NAACL: Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, CO.

Coppersmith, G., Ngo, K., Leary, R., & Wood, A. (2016, June). *Exploratory Analysis of Social Media Prior to a Suicide Attempt.* Paper presented at the CLPsych@ HLT-NAACL: Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, San Diego, CA.

Cunha, E., Magno, G., Andr, M., Cambraia, C., Almeida, V. (2014, September). *How you post is who you are: characterizing google+ status updates across social groups*. Paper presented at the 25th ACM conference on Hypertext and Social Media, Santiago, Chile.

Cunha, E., Magno, G., Gonçalves, M. A., Cambraia, C., & Almeida, V. (2013, February). *A linguistic characterization of google+ posts across different social groups.* Paper presented at the 5th Workshop on Information in Networks, London, United Kingdom. Retrieved January 3, 2017 from https://www.researchgate.net/profile/Evandro_Cunha2/publication/270510750_A_linguistic_characterization_of_Google_posts_across_different_social_groups/links/54ac5a7f0cf23c69a2b7bb8b.pdf

Cvijikj, I. P., & Michahelles, F. (2011). Understanding social media marketing: a case study on topics, categories and sentiment on a Facebook brand page. In A. Lugmayr, H. Franssila, C. Safran, I. Hammouda (Eds.) *Proceedings of the 15th International Academic Mindtrek Conference: Envisioning Future Media Environments* (pp. 175-182). New York, NY: ACM Press.

De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. In S. Fussell, W. Lutters, M. Morris, M. Reddy (Eds.) *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 626-638). New York, NY: ACM Press.

Diener, E., Napa-Scollon, C. K., Oishi, S., Dzokoto, V., & Suh, E. M. (2000). Positivity and the construction of life satisfaction judgments: Global happiness is not the sum of its parts. *Journal of Happiness Studies, 1*(2), 159-176.

Dumba, B., Golnari, G., & Zhang, Z.-L. (2016). Analysis of a reciprocal network using Google+: Structural properties and evolution. In M. Thai, Z. Zhang, W. Wu (Eds.) *Proceedings of International Conference on Computational Social Networks* (pp. 14-26). Cham, Switzerland: Springer International Publishing.

Duncan, W. R. (1996). *A guide to the project management body of knowledge*. Sylva, NC : PMI Communications.

Durahim, A. O., & Coşkun, M. (2015). # iamhappybecause: Gross National Happiness through Twitter analysis and big data. *Technological Forecasting and Social Change, 99*, 92-105.

Echeverria, J., & Zhou, S. (2017). The Star Wars botnet with 350k Twitter bots. *arXiv preprint arXiv:1701.02405*.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer‐ Mediated Communication, 12*(4), 1143-1168.

Ertugrul, A. M., Onal, I., & Acarturk, C. (2017). Does the strength of sentiment matter? A regression based approach on Turkish social media. In F. Frasincar, A. Ittoo, L. Nguyen, E. Métais (Eds.) *Proceedings of Natural Language Processing and Information Systems: 22nd International Conference on Applications of Natural Language to Information Systems* (pp. 149-155). Cham, Switzerland: Springer International Publishing.

Exton, C., Smith, C., & Vandendriessche, D. (2015). Comparing happiness across the world: Does culture matter? *OECD Statistics Working Papers, 2015*(4), 0_1. Paris, France: Organisation for Economic Cooperation and Development (OECD) Publishing. doi:http://dx.doi.org/10.1787/5jrqppzd9bs2-en

Ferraro, J. (2007). *The strategic project leader: Mastering service-based project leadership*. Boca Raton, FL: CRC Press.

Ford, C. (2015). Is R-squared useless? Retrieved May 5, 2018 from http://data.library.virginia.edu/is-r-squared-useless/

Frommer, D. (2010). Here's how to use Instagram. Retrieved Ferbuary 5, 2016 from http://www.businessinsider.com/instagram-2010-11

Fuchs, C. (2017). *Social media: A critical introduction*: Thousand Oaks, CA: Sage Publications Inc.

Galay, K. (2004). Gross National Happiness and Development. *Proceedings of the First International Seminar on Operationalization of Gross National Happiness*. Thimphu, Bhutan: Centre for Bhutan Studies.

Garas, A., Garcia, D., Skowron, M., & Schweitzer, F. (2012). Emotional persistence in online chatting communities. *Scientific Reports, 2*, 402. Retrieved February 5, 2016 from https://www.nature.com/articles/srep00402.epdf

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys (CSUR), 49*(2), 28.

Giannopoulos, G., Weber, I., Jaimes, A., & Sellis, T. (2012). Diversifying user comments on news articles. In A. Bouguettaya, Y. Gao, A. Klimenko, L. Chen, X. Zhang, F. Dzerzhinskiy, W. Jia, S.V. Klimenko, Q. Li (Eds.) *Proceedings of International Conference on Web Information Systems Engineering* (pp. 100-113). Berlin, Heidelberg, Germany: Springer.

Gil de Zúñiga, H., & Diehl, T. (2017). Citizenship, social media, and big Data: Current and future research in the social sciences. *Social Science Computer Review, 35*(1), 3-9.

Gilbert, E., & Karahalios, K. (2010). Widespread worry and the stock market. In N. Nicolov, J.G. Shanahan, L. Adamic, R.B. Yates, S. Counts (Eds.) *Proceedings of International AAAI Conference on Weblogs and Social Media* (pp. 59-65). Menlo Park, CA: AAAI Press.

Godwin-Jones, R. (2017). Scaling Up and Zooming In: Big Data and Personalization in Language Learning. *Language Learning & Technology, 21*(1), 4-15.

Gonzalez, R., Cuevas, R., Motamedi, R., Rejaie, R., & Cuevas, A. (2013). Google+ or google-?: dissecting the evolution of the new osn in its first year. In D. Schwabe, V. Almeida, H. Glaser, R.B. Yates, S. Moon (Eds.) *Proceedings of the 22nd international conference on World Wide Web* (pp. 483-494). New York, NY: ACM Press

Greenwood, S., Perrin, A., & Duggan, M. (2016). Social media update 2016. *Pew Research Center, 11*. Washington, DC. Retrieved February 5, 2017 from http://www.pewinternet.org/2016/11/11/social-media-update-2016/

Grigore, M., & Rosenkranz, C. (2011, December). *Increasing the willingness to collaborate online: An analysis of sentiment-driven interactions in peer content production.* Paper presented at the Thirty Second International Conference on Information Systems, Shanghai, China.

Guan, L., Hao, B., Cheng, Q., Yip, P. S., & Zhu, T. (2015). Identifying Chinese microblog users with high suicide probability using internet-based profile and linguistic features: Classification model. *Journal of Medical Internet Research-Mental Health, 2*(2).

Gunawardena, N., Plumb, J., Xiao, N., & Zhang, H. (2013, June). *Instagram hashtag sentiment analysis.* Paper presented at the University of Utah CS530/CS630 Conference of Machine Learning, Salt Lake City, UT.

Gutierrez, F. J., & Poblete, B. (2015). Sentiment-based user profiles in microblogging platforms. In Y. Yesilada, R. Farzan, G.J. Houben (Eds.) *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 23-32). New York, NY: ACM Press.

Hamouda, S. B., & Akaichi, J. (2013). Social networks' text mining for sentiment classification: The case of Facebook'statuses updates in the 'Arabic Spring'era. *International Journal of Application or Innovation in Engineering & Management (IJAIEM), 2*(5), 470-478.

Hanna, B., Kee, K. F., & Robertson, B. W. (2016, 18-20 September). *Positive impacts of social media at work: Job satisfaction, job calling, and Facebook use among co-workers.* Paper presented at the SHS Web of Conferences, Kuala Lumpur, Malaysia.

Harman, G. A. C. C. T., & Dredze, M. H. (2014). Measuring post traumatic stress disorder in Twitter. In N. Nicolov, J.G. Shanahan, L. Adamic, R.B. Yates, S. Counts (Eds.) *Proceedings of International AAAI Conference on Weblogs and Social Media* (pp. 400-412). Menlo Park, CA: AAAI Press.

Hartmann, P. M., Hartmann, P. M., Zaki, M., Zaki, M., Feldmann, N., Feldmann, N., Neely, A. (2016). Capturing value from big data–a taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management, 36*(10), 1382-1406.

Heimbach, I., Schiller, B., Strufe, T., & Hinz, O. (2015). Content Virality on Online Social Networks: Empirical Evidence from Twitter, Facebook, and Google+ on German News Websites. In Y. Yesilada, R. Farzan, G.J. Houben (Eds.) *Proceedings of the the 26th ACM Conference on Hypertext & Social Media* (pp. 483-494). New York, NY: ACM Press

Helliwell, J. F., Barrington-Leigh, C. P., Harris, A., & Huang, H. (2009). *International evidence on the social context of well-being*. Cambridge, MA: National Bureau of Economic Research.

Helliwell, J. F., Huang, H., & Wang, S. (2016). The distribution of world happiness. *Canadian Institute for Advanced Research,* Toronto, Canada. Retrieved February 5, 2017 from https://s3.amazonaws.com/happiness-report/2016/HR-V1Ch2_web.pdf

Hofstede, G. (1984). Cultural dimensions in management and planning. *Asia Pacific Journal of Management, 1*(2), 81-99.

Hofstede, G., Hofstede, G., & Minkov, M. (1991). Cultures and organizations: Software of the mind. *Administrative Science Quarterly, 38*(1), 132-134. doi:10.2307/2393257.

Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What we Instagram: A first analysis of Instagram photo content and user types. In N. Nicolov, J.G. Shanahan, L. Adamic, R.B. Yates, S. Counts (Eds.) *Proceedings of International AAAI Conference on Weblogs and Social Media* (pp. 150-153). Menlo Park, CA: AAAI Press.

Huang, R., & Hansen, J. H. (2007, April). *Dialect classification on printed text using perplexity measure and conditional random fields.* Paper presented at the Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. Honolulu, HI.

Huang, W., Weber, I., & Vieweg, S. (2014). Inferring nationalities of Twitter users and studying inter-national linking. In L. Ferres, G. Rossi, V. Almeida, E. Herder (Eds.) *Proceedings of the the 25th ACM Conference on Hypertext & Social Media* (pp. 237-242). New York, NY: ACM Press

Huang, X., Zhang, L., Chiu, D., Liu, T., Li, X., & Zhu, T. (2014, December). *Detecting suicidal ideation in Chinese microblogs with psychological lexicons.* Paper presented at the Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom), Bali, Indonesia.

Hunsinger, M., Isbell, L. M., & Clore, G. L. (2012). Sometimes happy people focus on the trees and sad people focus on the forest: Context-dependent effects of mood in impression formation. *Personality and Social Psychology Bulletin, 38*(2), 220-232.

Investopedia. (2018). Monthly Active User (MAU) Criticism. Retrieved 2 May, 2018, from https://www.investopedia.com/terms/m/monthly-active-user-mau.asp

Jain, G., Ginwala, A., & Aslandogan, Y. A. (2004). An approach to text classification using dimensionality reduction and combination of classifiers. In A. Memon, N. Zhao (Eds.) *Proceedings of the Information Reuse and Integration, 2004. IRI 2004* (pp. 564-569). Piscataway, NJ: IEEE.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In H. Zhangi B. Mobasher, L. Giles, A. McCallum, O. Nasraoui, M. Spiliopoulou (Eds.) *Proceedings of 9th*

*WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (pp. 56-65). New York, NY: ACM Press

Jenders, M., Kasneci, G., & Naumann, F. (2013). Analyzing and predicting viral tweets. In D. Schwabe, V. Almeida, H. Glaser (Eds.) *Proceedings of 22nd International Conference on World Wide Web* (pp. 657-664). New York, NY: ACM

John Walker, S. (2014). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Houghton Mifflin Harcourt, Taylor & Francis.

Katsurai, M., & Satoh, S. i. (2016). Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In Z. Ding, W. Zhang, Z.Q. Luo (Eds.) *Proceedings of the Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference* (pp. 2837-2841). Piscataway, NJ: IEEE.

Kharde, V., & Sonawane, P. (2016). Sentiment analysis of Twitter data: A survey of techniques. *arXiv preprint arXiv:1601.06971*.

Kramer, A. D. (2010). An unobtrusive behavioral model of gross national happiness. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, G. Olson (Eds.) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 287-290). New York, NY: ACM Press

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30*(3), 607-610.

Kucuktunc, O., Cambazoglu, B. B., Weber, I., & Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for Yahoo! answers. In E. Adar, J. Teevan, E. Agichtein, Y. Maarek (Eds.) *Proceedings of the 5th ACM International Conference on Web Search and Data Mining* (pp. 633-642). New York, NY: ACM Press

Kugel, S. (2006, April 10). A web site born in US finds fans in Brazil. *New York Times.* Retrieved from https://www.nytimes.com/2006/04/10/technology/a-web-site-born-in-us-finds-fans-in-brazil.html

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. In M.F. Balcan, K. Weinberger (Eds.) *Proceedings of the 33rd International Conference on Machine Learning* (pp. 1378-1387). New York, NY: ACM Press

Lahuerta-Otero, E., & Cordero-Gutiérrez, R. (2016). Looking for the perfect tweet. The use of data mining techniques to find influencers on Twitter. *Computers in Human Behavior, 64*, 575-583.

Lee, E., Lee, J.-A., Moon, J. H., & Sung, Y. (2015). Pictures speak louder than words: Motivations for using Instagram. *Cyberpsychology, Behavior, and Social Networking, 18*(9), 552-556.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems, 48*(2), 354-368.

Lima, A. C. E., de Castro, L. N., & Corchado, J. M. (2015). A polarity analysis framework for Twitter messages. *Applied Mathematics and Computation, 270*, 756-767.

Linaschke, J. (2011). *Getting the most from Instagram*. Berkeley, CA: Peachpit Press.

Liu, P., Tov, W., Kosinski, M., Stillwell, D. J., & Qiu, L. (2015). Do facebook status updates reflect subjective well-being? *Cyberpsychology, Behavior, and Social Networking, 18*(7), 373-379.

Liu, Y., Huang, X., An, A., & Yu, X. (2007). ARSA: A sentiment-aware model for predicting sales performance using blogs. In W. Kraaij, A. Vries, C. Clarke, N. Fuhr, N. Kando (Eds.) *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 633-642). New York, NY: ACM Press

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography, 70*, 11-25.

Lv, M., Li, A., Liu, T., & Zhu, T. (2015). Creating a Chinese suicide dictionary for identifying suicide risk on social media. *PeerJ, 3*, e1455.

Manikonda, L., Hu, Y., & Kambhampati, S. (2014). Analyzing user activities, demographics, social network structure and user-generated content on Instagram. *arXiv preprint arXiv:1410.8099*.

Mason, P. (2016, March 26). The racist hijacking of Microsoft's chatbot shows how the internet teems with hate. *The Guardian*. Retrieved from

https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism

Mayr, P., & Weller, K. (2017). Think before you collect: Setting up a data collection approach for social media studies. *arXiv preprint arXiv:1601.06296*

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review, 90*(10), 60-68.

McCune, Z., & Thompson, J. (2011). *Consumer Production in Social Media Networks: A Case Study of the 'Instagram'iPhone App* (Unpublished master thesis). University of Cambridge, Cambridge, United Kingdom. Retrieved January 3, 2017 from http://thames2thayer.com/blog/wp-content/uploads/2011/06/McCune_Instagram_Dissertation_Draft.pdf

McNely, B. J. (2012). Shaping organizational image-power through images: Case histories of Instagram. In L. Ledbetter, J. Leydens (Eds.) *Proceedings of the Professional Communication Conference (IPCC)* (pp. 1-8). Piscataway, NJ: IEEE.

Meire, M., Ballings, M., & Van den Poel, D. (2016). The added value of auxiliary data in sentiment analysis of Facebook posts. *Decision Support Systems, 89*, 98-112.

Menon, A. (2012, March). *Big data@ facebook.* Paper presented at the Workshop on Management of Big Data Systems, Berkeley, CA.

Messias, J., Magno, G., Benevenuto, F., Veloso, A., & Almeida, V. (2015). Brazil Around the World: Characterizing and Detecting Brazilian Emigrants Using Google+. In M. Cristo, D. Oliviera (Eds.) *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web* (pp. 85-91). New York, NY: ACM Press

Minkov, M. (2009). Nations with more dialectical selves exhibit lower polarization in life quality judgments and social opinions. *Cross-Cultural Research, 43*(3), 230-250.

Mishne, G., & Glance, N. S. (2006, March). *Predicting Movie Sales from Blogger Sentiment.* Paper presented at the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Palo Alto, CA.

Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts *Science and Engineering Ethics, 22*(2), 303-341.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016, June). *SemEval-2016 task 4: Sentiment analysis in Twitter.* Paper presented at the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA.

Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., Zhu, X. (2016). Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation, 50*(1), 35-65.

Ngoc, P. T., & Yoo, M. (2014, February). *The lexicon-based sentiment analysis for fan page ranking in Facebook.* Paper presented at the Information Networking (ICOIN), 2014 International Conference, Phuket, Thailand.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions, 2*(2), 183-188.

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior, 31*, 527-541.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques.* In J. Hajic, Y. Matsumoto (Eds.) *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing* (pp. 79-86). New York, NY: ACM Press

Panger, G. (2016). Reassessing the Facebook experiment: critical thinking about the validity of Big Data research. *Information, communication & society, 19*(8), 1108-1126. doi: 10.1080/1369118X.2015.1093525

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London, UK: Penguin Press.

Park, S., Kim, I., Lee, S. W., Yoo, J., Jeong, B., & Cha, M. (2015). Manifestation of depression and loneliness on social networks: a case study of young adults on Facebook. In D. Cosley, A. Forte (Eds.) *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.* (pp. 557-570). New York, NY: ACM Press

Park, S., Lee, S. W., Kwak, J., Cha, M., & Jeong, B. (2013). Activities on Facebook reveal the depressive state of users. *Journal of Medical Internet Research, 15*(10).

Paulson, S. (2017). Degrowth: culture, power and change. *Journal of Political Ecology, 24*, 425-448.

Pennacchiotti, M., & Popescu, A.-M. (2011). A Machine Learning Approach to Twitter User Classification. In N. Nicolov, J.G. Shanahan, L. Adamic, R.B. Yates, S. Counts (Eds.) *Proceedings of International AAAI Conference on Weblogs and Social Media* (pp. 281-288). Menlo Park, CA: AAAI Press.

Pfitzner, R., Garas, A., & Schweitzer, F. (2012). Emotional Divergence Influences Information Spreading in Twitter. *Proceedings of International AAAI Conference on Weblogs and Social Media* (pp. 87-99). Menlo Park, CA: AAAI Press.

Poblete, B., Garcia, R., Mendoza, M., & Jaimes, A. (2011). *Do all birds tweet the same?: characterizing twitter around the world.* In B. Berendt, A. Vries, W. Fan, C. Macdonald, I. Ounis, I. Ruthven (Eds.) *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 1025-1030). New York, NY: ACM Press

Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing, 174,* 50-59.

Priesner, S. (1999). Gross national happiness–Bhutan's vision of development and its challenges. *Indigeneity and Universality in Social Science, 2*, 212-233. doi:10.11588/xarep.00000320

Quan-Haase, A., & Young, A. L. (2010). Uses and gratifications of social media: A comparison of Facebook and instant messaging. *Bulletin of Science, Technology & Society, 30*(5), 350-361.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). *Tracking gross community happiness from tweets.* In S. Poltrock, C. Simone, J. Grudin, G. Mark, J. Riedl (Eds.) *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 965-968). New York, NY: ACM Press

Ranaweera, E. H., & Rajapakse, C. (2016, January). *Instagram sentiment analysis: Discovering tourists' perception about Sri Lanka as a tourist destination*. Paper presented in International Research Symposium on Pure and Applied Sciences (IRSPAS 2016), University of Kelaniya, Sri Lanka.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010, October). *Classifying latent user attributes in twitter.* Paper presented at the 2nd International Workshop on Search and Mining User-Generated Content, Toronto, Canada

Rice, T. W., & Steele, B. J. (2004). Subjective well-being and culture across time and space. *Journal of Cross-Cultural Psychology, 35*(6), 633-647.

Rodrigues Barbosa, G. A., Silva, I. S., Zaki, M., Meira Jr, W., Prates, R. O., & Veloso, A. (2012, May). *Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment.* Paper presented at the CHI'12 Extended Abstracts on Human Factors in Computing Systems, Austin, TX

Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology, 232*(2), 584-599.

Rousseau, D. (1996, August). *Personality in computer characters.* Paper presented at the 1996 AAAI Workshop on Entertainment and AI/A-Life, Portland, OR

Rudra, K., Chakraborty, A., Ganguly, N., & Ghosh, S. (2017). *Pattern Recognition And Big Data* (pp. 767-788). Singapore: World Scientific.

Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. Sebastopol, CA: O'Reilly Media, Inc.

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management, 52*(1), 5-19.

Salomon, D. (2013). Moving on from Facebook: Using Instagram to connect with undergraduates and engage in teaching and learning. *College & Research Libraries News, 74*(8), 408-412.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Seligman, M. E. (2013). *Characterizing geographic variation in well-being using tweets.* In N. Nicolov, J.G. Shanahan, L. Adamic, R.B. Yates, S. Counts (Eds.) *Proceedings of International AAAI Conference on Weblogs and Social Media* (pp. 583-591). Menlo Park, CA: AAAI Press.

Senik, C. (2014). The French unhappiness puzzle: The cultural dimension of happiness. *Journal of Economic Behavior & Organization, 106*, 379-401.

Sheldon, P., & Bryant, K. (2016). Instagram: Motives for its use and relationship to narcissism and contextual age. *Computers in Human Behavior, 58*, 89-97.

Silva, T. H., de Melo, P. O. V., Almeida, J. M., Salles, J., & Loureiro, A. A. (2013, May). *A picture of Instagram is worth more than a thousand words: Workload characterization and application.* Paper presented at the Distributed Computing in Sensor Systems (DCOSS), 2013 IEEE International Conference, Cambridge, MA.

Statista. (2018). Distribution of Twitter users worldwide from 2012 to 2018, by region. Retrieved April 19, 2018 from https://www.statista.com/statistics/303684/regional-twitter-user-distribution/

Stats, I. L. (2014). Number of Internet Users. Retrieved August 24, 2014 from http://www.internetlivestats.com/internet-users

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of graduate medical education, 4*(3), 279-282.

Systrom, K. (2010). What is the genesis of Instagram. *Post on Quora by the CEO and cofounder of Instagram. Online available at: http://www.quora.com/Instagram/What-- is-- the-- genesis-- of-- Instagram*.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37*(2), 267-307.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P. (2011). *User-level sentiment analysis incorporating social networks*. Paper presented at the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA.

Terrana, D., Augello, A., & Pilato, G. (2014, June). *Facebook users relationships analysis based on sentiment classification.* Paper presented at the Semantic Computing (ICSC), 2014 IEEE International Conference Newport Beach, CA.

Thelwall, M., & Buckley, K. (2013). Topic- based sentiment analysis for the social web: The role of mood and issue- related words. *Journal of the Association for Information Science and Technology, 64*(8), 1608-1617.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology, 63*(1), 163-173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology, 61*(12), 2544-2558.

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Murthy, R. (2009). Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment, 2*(2), 1626-1629.

Tian, Y., Galery, T., Dulcinati, G., Molimpakis, E., & Sun, C. (2017). Facebook Sentiment: Reactions and Emojis. *SocialNLP 2017*, 11.

Tole, A. A. (2013). Big data challenges. *Database Systems Journal, 4*(3), 31-40.

Trinh, S., Nguyen, L., Vo, M., & Do, P. (2016). Lexicon-based sentiment analysis of Facebook comments in Vietnamese language *Recent Developments in Intelligent Information and Database Systems* (pp. 263-276)

Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J. (2013, July). *Sentiment analysis of Facebook statuses using naive bayes classifier for language learning.* Paper presented at the Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference, Piraeus-Athens, Greece.

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015). *Recognizing depression from twitter activity.* In B. Begole, J. Kim, K. Inkpen, W. Woo (Eds.) *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3187-3196). New York, NY: ACM Press

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS), 21*(4), 315-346.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010, May). *Microblogging during two natural hazards events: what twitter may contribute to situational*

*awareness.* Paper presented at the Sigchi Conference on Human Factors in Computing Systems, Paris, France.

Vural, A. G., Cambazoglu, B. B., Senkul, P., & Tokgoz, Z. O. (2013). A framework for sentiment analysis in turkish: Application to polarity detection of movie reviews in turkish *Computer and Information Sciences III* (pp. 437-445)

Wang, C.-J., Wang, P.-P., & Zhu, J. J. H. (2013). Discussing occupy wall street on twitter: longitudinal network analysis of equality, emotion, and stability of public discussion. *Cyberpsychology, Behavior, and Social Networking, 16*(9), 679-685. doi: 10.1089/cyber.2012.0409

Wang, Y., Wang, S., Tang, J., Liu, H., & Li, B. (2015, July). *Unsupervised sentiment analysis for social media images.* Paper presented at the International Conference on Artificial Intelligence, Buenos Aires, Argentina.

Wells, C., & Thorson, K. (2017). Combining big data and survey techniques to model effects of political content flows in Facebook. *Social Science Computer Review, 35*(1), 33-52.

Wikipedia. (2018). Monthly active users (MAU). Retrieved May 2, 2018 from https://en.wikipedia.org/wiki/Monthly_active_users

Wilkinson, D., & Thelwall, M. (2011). Researching personal information on the public web: Methods and ethics. *Social Science Computer Review, 29*(4), 387-401.

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2017). Researching Mental Health Disorders in the Era of Social Media: Systematic Review. *Journal of medical Internet Research, 19*(6), e228.

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 97-107.

Xu, D. J., Liao, S. S., & Li, Q. (2008). Combining empirical experimentation and modeling techniques: A design research approach for personalized mobile advertising applications. *Decision Support Systems, 44*(3), 710-724.

Yamamoto, Y., Kumamoto, T., & Nadamoto, A. (2014, December). *Role of emoticons for multidimensional sentiment analysis of Twitter.* Paper presented at the Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services, Hanoi, Vietnam.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *National Academy of Sciences, 112*(4), 1036-1040.

Yu, Y., & Wang, X. (2015). World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans' tweets. *Computers in Human Behavior, 48*, 392-400.

Yuan, S.-T. (2003). A personalized and integrative comparison-shopping engine and its applications. *Decision Support Systems, 34*(2), 139-156.

Zheludev, I., Smith, R., & Aste, T. (2014). When can social media lead financial markets? *Scientific Reports, 4*, 4213.

Zheng, X., Han, J., & Sun, A. (2017). A Survey of Location Prediction on Twitter. *arXiv preprint arXiv:1705.03172*.

Zook, M., Barocas, S., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Narayanan, A. (2017). Ten simple rules for responsible big data research. *PLoS computational biology, 13*(3), e1005399.