Sentiment Analysis on TripAdvisor: Are There Inconsistencies in User Reviews?

Ana Valdivia[1(✉)], M. Victoria Luzón[2], and Francisco Herrera[1]

[1] Department of Computer Science and Artificial Intelligence, University of Granada,
18071 Granada, Spain
avaldivia@ugr.es, herrera@decsai.ugr.es
[2] Department of Software Engineering, University of Granada,
18071 Granada, Spain
luzon@ugr.es

Abstract. The number of online reviews has grown exponentially over the last years. As a result, several Sentiment Analysis Methods (SAMs) have been developed in order to extract automatically sentiments from text. In this work, we study polarity coherencies between reviewers and SAMs. To do so, we compare the polarity of the document evaluated by the user and the aggregated sentence polarity evaluated by three SAMs. The main contribution of this work is to show the flimsiness of user ratings as a generalization of the overall review sentiment.

Keywords: Sentiment Analysis · Opinion mining · Online reviews

1 Introduction

The concept of Sentiment Analysis (SA), also referred as Opinion Mining, has experienced an important growth through the last few years [15]. This topic has been established as a new Natural Language Processing (NLP) research branch which processes automatically written opinions so as to extract insights and knowledge. Moreover, the proliferation of the Web 2.0 and social networks has led to a huge amount of online recorded text. Users are free to express their opinions about products, places and experiences. This has implied a high development of SAMs for sentiment extraction ([21,22]).

TripAdvisor[1] is one of the most popular travel social network websites [12]. This Web 2.0 contains millions of written and ranked reviews about restaurants, hotels and attractions from a large number of travelers over the world. Tourists are able to plan their trip checking information, ranking list and experiences from others. In this website, users write opinions of 100 character minimum and rank them with 1 to 5 score (1 is representing a *Terrible* assessment and 5 an *Excellent* assessment). TripAdvisor has therefore become a rich source of data for SA research and applications.

[1] https://www.tripadvisor.com.

F.J. Martínez de Pisón et al. (Eds.): HAIS 2017, LNAI 10334, pp. 15–25, 2017.
DOI: 10.1007/978-3-319-59650-1_2

The purpose of this work is to study the robustness of the user's polarity comparing with three SAM polarities. For doing so, we analyze TripAdvisor reviews from three popular monuments in Spain: the Alhambra, the Sagrada Família and the Mezquita de Córdoba. We define the *User's Polarity* as the user rating. We then define the *SAMs' Polarities* by computing the sentiment on each sentence applying the corresponding method (*Syuzhet* [13], *Bing* [10] and *CoreNLP* [17]). In order to obtain an overall polarity, we aggregate sentence polarities by majority vote. Finally, we correlate the polarities.

The results show that there exists a latent inconsistency between the *User's Polarity* and the *SAMs' Polarities*. There is around 50% of correlation between positive sentiments.

The rest of this work is organized as follows: in Sect. 2 we describe the theory of SA thus far, including an introduction to the SA problem (Sect. 2.1) and the presentation of the three SAM that we select for this study (Sect. 2.2). After that, in (Sect. 3) we explain the developed methodology for our purpose. In there, we explain TripAdvisor structure (Sect. 3.1), how we scrap the web (Sect. 3.2) and the experiments layout (Sect. 3.3). Section 4 includes the analysis of results. We firstly explain the structure of datasets (Sect. 4.1) and then we present the numerical report (Sect. 4.2). Lastly, we present the conclusions and suggest future research lines in Sect. 5.

2 Sentiment Analysis

In this section we define the main concepts related to SA. In Sect. 2.1 we introduce the SA problem. Section 2.2 briefly describes the three sentiment tools used in this work.

2.1 The Sentiment Analysis Problem

Liu defines SA in [15] as the field of study that analyzes people's opinions toward products, services, organizations, individuals, events, issues, or topics in written text. SA is widely known as Opinion Mining, but recently has been popularized with the first bigram.

Liu organizes this problem proposing that an *opinion* can be mathematically defined as a 5-tuple $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ where: e_i is the i-th opinion *entity*, i.e., the product, service, place, person, company or event which the opinion is addressed to; a_{ij} is the j-th *attribute*, a property related to the entity e_i; s_{ijkl} is the *sentiment* of the opinion towards an attribute a_{ij} of entity e_i by the opinion holder h_k at time t_l; h_k is the k-th *opinion holder* or the reviewer and t_l is l-th *time* when the opinion was emitted. Over this problem, the *sentiment* can be identified in different ways: polarity {positive, neutral, negative}, numerical rating {1, 2, ..., 5} or emotions {anger, disgust, fear, happiness, sadness, surprise}.

One other fact that makes this problem complex is that there exists several types of opinions [14]. *Regular opinions* express a sentiment about an aspect of an entity. On the other hand, *comparative opinions* compare two or more entities. *Subjective opinions* express a personal feeling or belief and thus are more likely to present sentiments. On the other side, *objective sentence* present factual information.

There exists three different levels of analysis to this problem. The first one is the *document level* and extracts the sentiment of the whole opinion. This is considered to be the simplest task. The next level is the *sentence level* which extracts a sentiment in each sentence of the text. Finally, the *aspect level* is the fine-grained level. This is the most challenging analysis because it extracts the sentiment related to its target (an entity or aspect's entity). Due to this fact, Aspect Based Sentiment Analysis (ABSA) has been widely studied over the literature. For example, Hu and Liu propose in [11] a methodology to extract product features from reviews. ABSA task has been repeatedly proposed in the International Workshop of Semantic Evaluation [19].

Because of the complexity of the SA, different task with different targets are related to this problem. The authors describe in [22] a total of 6 task:

1. **Sentiment Classification**: This is the most known task. The aim of Sentiment Classification is to develop models capable of detecting sentiment in texts. The first step is to collect text or reviews to set our analysis. After that, the sentiment is detected. It can be computed by the reviewer or computed with SAMs. Then, features are selected to train the classification model. In this step, text mining techniques are commonly used to extract the most significant features. Finally, machine learning or lexicon-based techniques can be used to address this problem [18].
2. **Subjectivity Classification**: This task is related to Sentiment Classification in the sense that the objective is to classify subjective and objective opinions. The purpose is to filter subjective sentences because they are more opinionated and thus can improve classification models.
3. **Opinion Summarization**: Also known as aspect based summary or feature based summary. It consists in developing techniques to sum up large amounts of reviews from people. The summarization should focus on entities or aspects and their sentiment and should be quantitative.
4. **Opinion Retrieval**: This is a retrieval process, which requires documents to be retrieved and ranked according to their relevance.
5. **Sarcasm and Irony**: This task is aimed to detect opinions with sarcastic or ironic expressions. As in Subjectivity Classification, the target is to delete these opinions from the SA process.
6. **Others**: Due to the fact that SA is a growing branch of knowledge, over recent years many new tasks have been appearing. Spam detection is one of the most popular.

2.2 Sentiment Analysis Methods (SAMs)

We define SAMs as those tools that are able to evaluate sentiments in text. There exists three main types of SAMs:

(a) Lexicon Dictionary Based Method: It mainly consists in creating a sentiment lexicon, i.e., words carrying a sentiment orientation. These methods can create the dictionary from initial seed words, corpus words (related to a specific domain) or combining the two. Frequently, the dictionary is fed with synonyms and antonyms.
(b) Machine Learning Based Method: It develops statistical models with classifier algorithms. These methods can be divided into super and unsupervised. The main difference is that the first group uses labeled opinions to build the model. One of the most important step in these methods is the feature extraction for representing the classes to be predicted.
(c) Hybrid Based Method: They combine both Lexicon Dictionary and Machine Learning approaches.

Thus, we define three examples of SAM methods based in the aforementioned:

1. Syuzhet: Syuzhet is a Lexicon Dictionary Based Method developed in the Nebraska Literary Lab under the direction of Matthew L. Jockers . Its dictionary is created from a collection of 165,000 human coded terms taken from corpus of contemporary novels [13]. Syzhet reports three polarity levels: {*negative*, *neutral*, *positive*}.
2. Bing: It is a Lexicon Dictionary Based Method developed by Hu and Liu at University of Illinois [10]. This dictionary contains around 6,800 words classified in positive or negative terms. In this case, the output of Bing is a numerical scale: {–1, 0, 1} representing {*negative*, *neutral*, *positive*}.
3. CoreNLP: CoreNLP is a Machine Learning Based Method created by the Stanford NLP Group. It is defined as an integrated toolkit capable of executing different NLP tasks [17]. One of these tasks is related to SA: they developed a deep learning algorithm, Recursive Neural Tensor Network (RNTN), which outperforms old methods in different metrics. This algorithm extracts sentences sentiment representing a phrase through word vectors and parse trees. Unlike Lexicon Dictionary Based Methods this algorithm is capable of capturing sentiment changes detecting scope negations and contrastive conjunctions like *but* (see attached Fig. 1). Finally, the output of this algorithm is a numerical scale: {0, 1, 2, 3, 4}. We set the polarities {0, 1} as *negative*, {2} as *neutral* and {3, 4} as *positive* to work with the same polarity levels.

3 Methodology

In this section we describe the setup of our experiments. The first part, Sect. 3.1, is focused on introducing TripAdvisor as our data source. In Sect. 3.2, we explain how we get the data from websites. Finally, an outline of the experiments is given in Sect. 3.3.

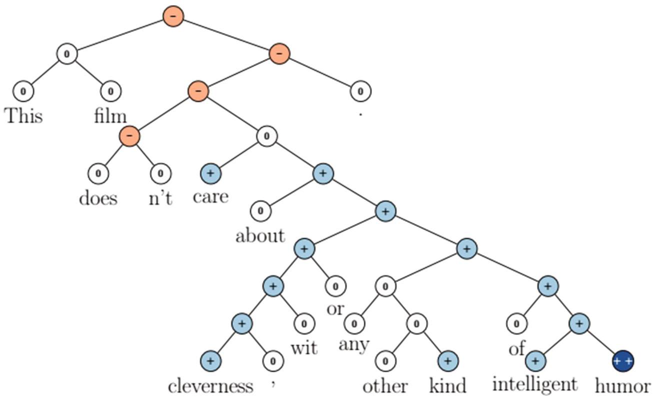

Fig. 1. Example of RNTN extracted from [23]. The algorithm predicts five sentiments, from very negative ({0}) to very positive ({4}). As it is shown in the example, the algorithm captures the negation which shifts the polarity of the whole sentence.

3.1 TripAdvisor

According to Wikipedia[2], TripAdvisor is an American travel website company, founded in 2000, providing reviews from travelers experiences about accommodations, restaurants and attractions.

TripAdvisor is considered one of the first web 2.0 adopter: its information and advice index is constructed from the accumulated opinions of million of tourists. For this reason, this website has made up the largest travel community, reaching 340 million unique monthly visitors reporting more than 350 million online reviews and opinions[3].

Due to these facts, this website has become a rich source for SA. Examples of works analyzing hotels reviews are [1,4–7,16,20,24]. Restaurant reviews are analyzed in [7,9,27].

However, one of the major concerns of user-generated content is the credibility of the opinions. Awaring of it, TripAdvisor has thought up several measures in order to avoid spam and fictitious reviews like allowing the use of commercial email addresses or posting warnings about the zero tolerance for fake opinions. Regarding to this, it has been carried out several studies for analyzing credibility and truthfulness of this website ([3,8,12,26]).

3.2 Web Scraping

We first describe the structure of TripAdvisor monuments websites to explain how do we scrap the data from the web.

[2] https://en.wikipedia.org/wiki/TripAdvisor.

[3] Source: TripAdvisor log files, Q1 2016.

All monuments websites are structured in the same way. On the top, it displays the total number of reviews, written in different languages, and a *Popularity Index ranking*. After that, the main page is divided in five sections: Overview, Tours & Tickets, Reviews, Q & A and Location. In the review section we find all the opinions written by TripAdvisor users. A review is formed by:

User Name: The name of the user in TripAdvisor.
User Location: The location of the user.
User Information: The total number of reviews, attraction reviews and helpful votes of the user.
Review Title: A main title of the text.
User Rating: The visit valuation of the user. It is expressed as a discrete number scale from 1 to 5 (from *Terrible* to *Excellent*).
Review Date: The reviewing time.
Review: The text of the opinion.

Finally, we develop a code in R software with `rvest` package which allow us to extract the TripAdvisor reviews from HTML and XML codes [25].

3.3 Experimental Setup

The aim of this work is to study the consistency of TripAdvisor's user ratings. To do so, we define *User's Polarity* from User Rating previously defined. We set the polarity as follows: from 1 to 2 is negative, 3 is neutral and from 4 to 5 is positive. Note that this is a document-level polarity.

The second step is to define *SAM Polarities.* We split in sentences each review and apply both *Syuzhet* and *Bing* methods with the `syuzhet` R package [13] and CoreNLP with the `coreNLP` R package [2].

In order to obtain a document-level polarity, we aggregate sentences scores by majority vote. In a tied event, the final result is neutrality.

Final step is to perform a quantitative analysis of the four polarities and study the correlation between them.

4 Experiment Results

In this section we insightfully describe the quantitative analysis of our experiments. A description of datasets is given in Sect. 4.1. After that, we discuss the results reporting some numbers and plots in Sect. 4.2.

4.1 The Data Sets

We base our experiments on TripAdvisor reviews of three monuments in Spain: the Alhambra, the Sagrada Família and the Mezquita de Córdoba. We base our study only in English reviews due to the fact that SAMs are mainly developed for this language. We create three data sets with reviews from July 2012 until June 2016. As it is shown in Table 1, we collect a total of 45,303 reviews. The monument of Barcelona holds a large amount of reviews during the selected period, 76.29 % of the total. Surprisingly, reviews of the Alhambra are in average larger than other reviews.

Table 1. Summary of text properties of the three data sets.

	Reviews	Words	Sentences	Word avg.	Sentence avg.
Alhambra	7,218	676,398	35,867	93.72	4.97
Sagrada Família	34,559	2,220,719	136,181	64.26	3.94
Mezquita de Córdoba	3,526	217,640	13,083	61.72	3.70

4.2 Analysis of Results

To begin with, Table 2 presents the polarity distribution of *User's Polarity* and the three *SAMs' Polarities*. The positive polarity predominates over the *User's Polarity* which means that TripAdvisor users are usually satisfied with their visit to these monuments. However, the three SAMs seem to detect more neutrality and negativity in the same reviews.

Table 2. Distribution of sentiments of monuments reviews.

User's polarity	Positive	Neutral	Negative
Alhambra	6,781	293	143
Sagrada Família	32,664	1,443	451
Mezquita de Córdoba	3,454	55	17
Syuzhet's polarity	Positive	Neutral	Negative
Alhambra	5,423	911	883
Sagrada Família	25,379	4,805	4,374
Mezquita de Córdoba	2,704	466	356
Bing's polarity	Positive	Neutral	Negative
Alhambra	3,310	1,252	2,655
Sagrada Família	16,541	6,644	11,373
Mezquita de Córdoba	1,918	642	966
CoreNLP's polarity	Positive	Neutral	Negative
Alhambra	3,154	1,143	2,920
Sagrada Família	17,561	6,007	10,990
Mezquita de Córdoba	1,992	577	957

Table 3 shows match rates between *User's Polarity* (columns) and *Syuzhet, Bing* and *CoreNLP Polarities* (rows). The shares represent the average of the three monuments due to the fact that the percentages are very similar (see Table 4).

As can be hinted from the table, the analysis of the data revealed a clear disparity between polarities. *Syuzhet* shows a high correlation in the positive opinions, 76% of matching, while Bing and CoreNLP get around 50%. However,

Table 3. Correlation arrays between *User Rating* (rows) and SAMs (columns) sentiments. The numbers are the average of the three monuments

Syuzhet sentiment	Positive	Neutral	Negative
Positive	0.76	0.13	0.11
Neutral	0.60	0.18	0.22
Negative	0.42	0.23	0.35
Bing sentiment	Positive	Neutral	Negative
Positive	0.50	0.18	0.31
Neutral	0.35	0.18	0.48
Negative	0.09	0.14	0.77
CoreNLP sentiment	Positive	Neutral	Negative
Positive	0.52	0.16	0.32
Neutral	0.17	0.17	0.66
Negative	0.09	0.26	0.84

Table 4. Standard deviation of *User Rating* (rows) and SAMs (columns) sentiments. This table clearly shows that sentiments are equally distributed over the three monuments

Syuzhet sentiment	Positive	Neutral	Negative
Positive	0.02	0.01	0.01
Neutral	0.02	0.03	0.04
Negative	0.07	0.07	0.06
Bing sentiment	Positive	Neutral	Negative
Positive	0.04	0.01	0.04
Neutral	0.08	0.07	0.14
Negative	0.10	0.07	0.12
CoreNLP sentiment	Positive	Neutral	Negative
Positive	0.06	0.01	0.06
Neutral	0.03	0.06	0.08
Negative	0.05	0.30	0.09

this correlation decreases in the neutral and negative opinions where it detects a 35% of total negative reviews. Although *Bing* and *CoreNLP* show a low rate of detection in positive reviews, the match share on negative reviews is 77% and 84%, respectively. In neutral user's reviews distributions are more scattered. This has sense because when the sentiment is neutral is more frequent to write both equally, positive and negative sentences. However, the tendency is the same as before: *Syuzhet* tends to detect more positivity and *Bing* and *CoreNLP* more negativity.

One more fact is that although *Syuzhet* and *Bing* are both Lexicon Dictionary Based, their behavior analyzing sentiments is different.

5 Conclusions and Future Work

In this study we develop an empirical experiment for studying the consistency of user-based polarities. We download TripAdvisor reviews from three Spanish monuments. After that, we define the *User's Polarity* as the polarity given by the reviewer and the *SAMs' Polarities* as the most recurrent polarity of sentences in the review evaluated by three different SAMs. After obtaining the four polarity labels, we analyze its correlation.

Our analytic experiments show that there exists a low correlation between sentiment labels depending on the SAM. This led us to conclude that there exists disparity between polarities. Users may tend to write negative sentences on positive reviews, and vice versa. Therefore, we should recommend not to analyze reviews with the overall sentiment due to the polarity disparity in sentences. Additionally, *SAMs' Polarities* may wrongly assessed polarity in sentences due to the fact that they are not 100% precise. A study of the SAM's sentence labelling should be carried out in order to extend this work.

We suggest considering a classification method taking into account the sentiment extracted by SAMs. Moreover, we suggest to develop an opinion summarization methodology taking into account sentence sentiment instead of the overall sentiment. In this sense, more detailed insights can be discovered.

Acknowledgments. We thank the anonymous reviewers for their constructive feedback. This research has been supported by FEDER and the Spanish National Research Project TIN2014-57251-P.

References

1. Aciar, S.: Mining context information from consumers reviews. In: Proceedings of Workshop on Context-Aware Recommender System, vol. 201. ACM (2010)
2. Arnold, T., Tilton, L.: R packages. In: Arnold, T., Tilton, L. (eds.) Humanities Data in R. Quantitative Methods in the Humanities and Social Sciences, pp. 179–182. Springer, Heidelberg (2015)
3. Ayeh, J.K., Au, N., Law, R.: Do we believe in tripadvisor? examining credibility perceptions and online travelers attitude toward using user-generated content. J. Travel Res. **52**(4), 437–452 (2013)
4. Baccianella, S., Esuli, A., Sebastiani, F.: Multi-facet rating of product reviews. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 461–472. Springer, Heidelberg (2009). doi:10.1007/978-3-642-00958-7_41
5. Banić, L., Mihanović, A., Brakus, M.: Using big data and sentiment analysis in product evaluation. In: 2013 36th International Convention on Information & Communication Technology Electronics & Microelectronics (MIPRO), pp. 1149–1154. IEEE (2013)

6. Duan, W., Cao, Q., Yu, Y., Levy, S.: Mining online user-generated content: using sentiment analysis technique to study hotel service quality. In: 2013 46th Hawaii International Conference on System Sciences (HICSS), pp. 3119–3128. IEEE (2013)
7. ElSahar, H., El-Beltagy, S.R.: Building large arabic multi-domain Resources for sentiment analysis. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9042, pp. 23–34. Springer, Cham (2015). doi:10.1007/978-3-319-18117-2_2
8. Filieri, R., Alguezaui, S., McLeay, F.: Why do travelers trust tripadvisor? antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. Tourism Manag. **51**, 174–185 (2015)
9. García, A., Gaines, S., Linaza, M.T., et al.: A lexicon based sentiment analysis retrieval system for tourism domain. Expert Syst. Appl. Int. J. **39**(10), 9166–9180 (2012)
10. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
11. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: AAAI, vol. **4**, 755–760 (2004)
12. Jeacle, I., Carter, C.: In tripadvisor we trust: rankings, calculative regimes and abstract systems. Acc. Organ. Soc. **36**(4), 293–309 (2011)
13. Jockers, M.: Package syuzhet (2016)
14. Liu, B.: Sentiment analysis and subjectivity. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, 2nd edn., pp. 627–666. Chapman and Hall/CRC, Boca Raton (2010)
15. Liu, B.: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, New York (2015)
16. Lu, B., Ott, M., Cardie, C., Tsou, B.K.: Multi-aspect sentiment analysis with topic models. In: 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), pp. 81–88. IEEE (2011)
17. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford coreNLP natural language processing toolkit. In: ACL (System Demonstrations), pp. 55–60 (2014)
18. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
19. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B.: Orphée de clercq, véronique hoste, marianna apidianaki, xavier tannier, natalia loukachevitch, evgeny kotelnikov, nuria bel, salud marıa jiménez-zafra, and gülsen eryigit. semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval, vol. 16 (2016)
20. Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: Kao, A., Poteet, S.R. (eds.) Natural Language Processing and Text Mining, pp. 9–28. Springer, London (2007)
21. Ribeiro, F.N., Araújo, M., Gonçalves, P., Benevenuto, F., Gonçalves, M.A.: Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. arXiv preprint arXiv:1512.01818 (2015)
22. Serrano-Guerrero, J., Olivas, J.A., Romero, F.P., Herrera-Viedma, E.: Sentiment analysis: a review and comparative analysis of web services. Inf. Sci. **311**, 18–38 (2015)
23. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C., et al.: Recursive deep models for semantic compositionality over a sentiment

treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), vol. 1631, p. 1642. Citeseer (2013)

24. Titov, I., McDonald, R.T.: A joint model of text and aspect ratings for sentiment summarization. In: ACL, vol. 8, pp. 308–316. Citeseer (2008)
25. Wickham, H.: rvest: Easily harvest (scrape) web pages. R package version 0.2. http://CRAN.R-project.org/package=rvest (2015)
26. Yoo, K.H., Lee, Y., Gretzel, U., Fesenmaier, D.R.: Trust in travel-related consumer generated media. In: Höpken, W., Gretzel, U., Law, R. (eds.) Information and Communication Technologies in Tourism, pp. 49–59. Springer, Vienna (2009)
27. Zhang, H.Y., Ji, P., Wang, J., Chen, X.: A novel decision support model for satisfactory restaurants utilizing social information: a case study of tripadvisor.com. Tourism Manag. **59**, 281–297 (2017)