Research Paper

JIS

Journal of Information Science
1–13

DOI: 10.1177/0165551518761013
journals.sagepub.com/home/jis
SAGE

Bibliometrics of sentiment analysis literature

Abdalsamad Keramatfar
University of Qom, Iran

Hossein Amirkhani
University of Qom, Iran

Abstract

This article provides a bibliometric study of the sentiment analysis literature based on Web of Science (WoS) until the end of 2016 to evaluate current research trends, quantitatively and qualitatively. We concentrate on the analysis of scientific documents, distribution of subject categories, languages of documents and languages that have been more investigated in sentiment analysis, most prolific and impactful authors and institutions, venues of publications and their geographic distribution, most cited and hot documents, trends of keywords and future works. Our investigations demonstrate that the most frequent subject categories in this field are *computer science, engineering, telecommunications, linguistics, operations research and management science, information science and library science, business and economics, automation and control systems, robotics* and *social sciences*. In addition, the most active venue of publication in this field is *Lecture Notes in Computer Science* (*LNCS*). The United States, China and Singapore have the most prolific or impactful institutions. A keyword analysis demonstrates that *sentiment analysis* is a more accepted term than *opinion mining*. Twitter is the most used social network for sentiment analysis and *Support Vector Machine* (*SVM*) is the most used classification method. We also present the most cited and hot documents in this field and authors' suggestions for future works.

Keywords

Bibliometrics; keyword analysis; opinion mining; sentiment analysis; Twitter

1. Introduction

Whenever we need to make a decision, we often seek out others' opinions. This is true not only for individuals but also for organisations and governments [1,2]. This way we want to benefit from others' experiences to avoid faults and earn more profits. If we can do this process successfully, we can make more efficient decisions. With today's web technologies, each person can express his opinions easily, and therefore, it is possible to benefit from the bulk of existing opinions to gain insights for decision-making. Sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, appraisals, attitudes and emotions towards entities and their aspects expressed in (usually) written texts [1]. The entities can be products, services, organisations, individuals, events, issues or topics. The goal of this analysis is to automatically extract from natural language the overall sentiment (emotional direction or feeling) of a word, phrase, sentence or combination thereof. This type of analysis has a lot of applications in the industry and government, as well as for individuals [3]. The sentiment analysis gives companies and individuals the ability to monitor their reputations in different social media sites and get timely feedbacks about their products and actions. Marketing managers, public relations (PR) firms, campaign managers, politicians, equity investors and online shoppers are the direct beneficiaries of the sentiment analysis technology [4,5].

To provide a clear perspective of this important field for researchers, we provide a bibliometric study of the sentiment analysis in this article. Bibliometrics is a field of study that attempts to use bibliographic data of publications and their

Corresponding author:
Abdalsamad Keramatfar, University of Qom, Tehran, Iran.
Email: keramatfar.a.s@gmail.com

citation relations to evaluate and reveal the structure of research. Previous research [6–10] used bibliometric methods to analyse different subfields of computer science. In one of its sections, this paper [11] used bibliometrics to show the trend of the usage of 'machine learning' and 'fuzzy sets' with 'sentiment analysis'. After showing the sustained growth of the former and the lesser growth and lesser number of publications of the latter, the paper concludes that further investigation about the potential use of fuzzy sets in the sentiment analysis problem is rewarding. Recently, data mining methods have been used in bibliometrics to generate better representation of science too [12]. In this article, we employ different bibliometric methods to gain insights about patterns in global sentiment analysis studies from the following perspectives: characteristics of scientific outputs, subject categories and major journals, language and geographic distribution, temporal trends of keywords and future works. The rest of this article is organised as follows. In Section 2, we introduce the setup of our study and the statistics of the used database. Section 3 provides the findings of the study, including languages and most investigated languages, most impactful authors, institutions, countries, venues of publications, must-read and hot publications, most important keywords and future works. This article concludes in Section 4.

2. Study setup and statistics

Web of Science (WoS) is the most frequently used database in bibliometric studies [13]. A previous research [11] also indicates that the sentiment analysis documents are similar in WoS and Scopus – another major database. So we use only WoS in this research. To collect the related documents in the field of sentiment analysis, we use the following phrase as our search query:

ts = 'opinion mining' or ts = 'sentiment analysis'

By *ts* we enforce the engine to search in the topics (title, abstract and keywords) of documents. Some handwritten Python scripts are also used for data gathering, such as Scimago Journal Rank[1] (SJR) of journals from Scimago, discovering language distribution in data set and extracting fields of records such as authors and institutions. In addition, some bibliometric software are used, such as BibExcel [14] and VOSviewer,[2] and Microsoft Excel 2010 is used to extract some statistics of combining different fields by pivot tables and calculation of the correlation between variables.

There are 3225 documents in WoS that contain 'sentiment analysis' or 'opinion mining' in their title, abstract or keywords. These documents received 10,466 citations. This set includes proceedings papers (~68%), journal articles (~30%), reviews and editorial materials, book reviews, book chapters, corrections and reprints (~2%). Since 'sentiment analysis' is a subfield of 'computer science', these statistics are in line with the main venue of publications in computer science, that is, conference proceedings [15]. The average citations for proceedings, journal articles and review papers are 1.2, 7 and 12, respectively. This shows that the journal articles and review papers have more impact than proceedings in this field.

Every source covered by the WoS core collection is assigned to at least one subject category. There are 66 different subject categories related to our investigated documents, but the following 10 subject categories contain more than 90% of documents. The 'computer science' as a super field of 'sentiment analysis' contributed to more than half of the documents (~52%). Other major subjects are engineering (~19%), telecommunications, linguistics, operations research, and management science (~10), information science and library science, business and economics, automation and control systems, robotics and social sciences (~10%). Note that the main focus of research in computer science, engineering, telecommunications, robotics, automation and control systems is to develop systems and algorithms [16], in which researchers use the machine learning methods as a core component. Although sentiment analysis originated from computer science, in recent years, it has spread to management and social sciences because of its importance to business and society as a whole [1]. Management science and economics study more business-related problems such as sale prediction [17]. On the contrary, linguistic documents usually concentrate on lexicon-based sentiment analysis that constructs and uses lexicons for sentiment analysis. Information and library sciences documents have a mixed nature; some documents in this category are related to social issues [18], while others are methodological [19]. Social sciences usually study social or political issues such as election forecasting [20].

Figure 1 shows the growing attention to the sentiment analysis field. The average annual growth of the documents is about 79%. Since 2002, research in sentiment analysis has been very active [1]. The field has grown rapidly to become one of the most active research areas in natural language processing (NLP), data mining and web mining. It is also widely studied in management sciences [1]. It should be noted that the reduction in the growth rate in 2016 is partly due to the fact that it takes time for new publications to be added to WoS. In terms of citations, it can be seen that ~50% of citations are to publications that are published within four recent years.

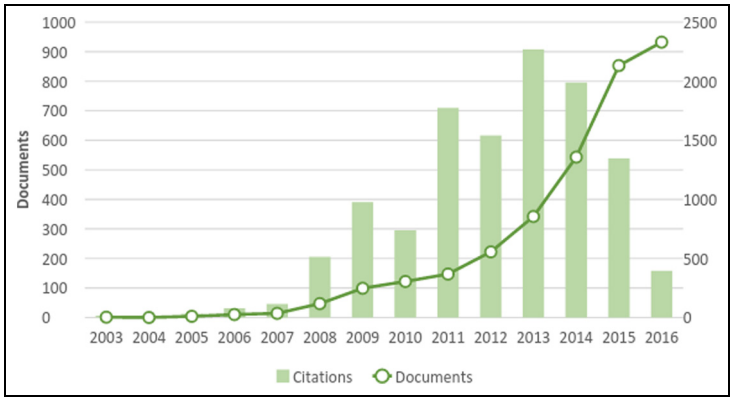

Figure 1. The number of documents and citations related to 'sentiment analysis' filed according to the WoS (2003–2016).

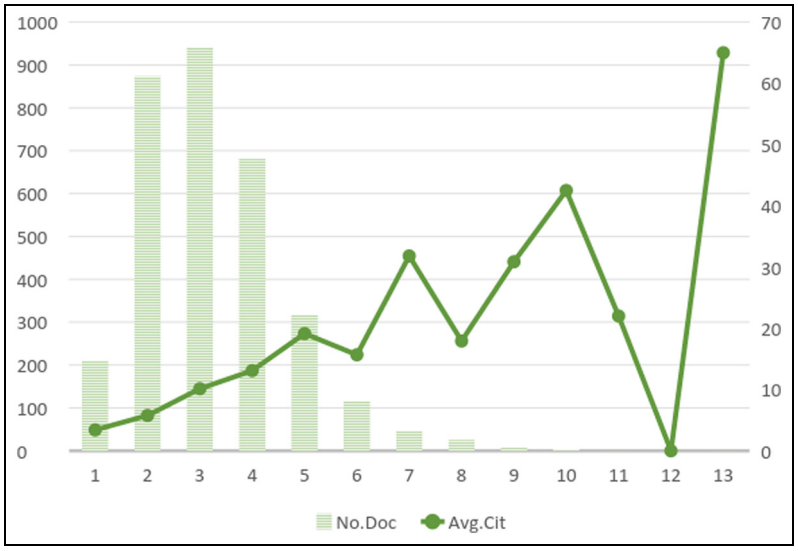

Figure 2. Distribution of the number of authors and their average number of citations.

As Liu [1] argues, the term 'sentiment analysis' first appeared in 2003 [21], but based on WoS, the first paper is Yi et al. [22] with an overlapping of authors between these two papers. We found the first mention of the term 'opinion mining' in 2006 [23].

In terms of co-authorship patterns, Figure 2 shows the distribution of the number of authors and their average citations. It can be seen that a great share of papers (946) have three authors, and the highest number of authors in a paper is 13. Single-author documents on average received ~3 citations, while multi-author ones on average received ~11 citations. There is also a correlation between the number of authors and the impact of papers (the correlation coefficient is ~0.6). This relation has also been observed by other researchers [24–27].

3. Results

3.1. Language of sentiment analysis papers and the most investigated languages

In this section, we first present statistics about languages of papers, and then we present languages on which sentiment analysis research has been done. This illustration can be useful for researchers to know in which languages there is room for further work in this field. Table 1 shows distribution of languages in our data. English is dominant (~99%) in these documents; this result has been reported in another study [28]. Spanish and Turkish show a high presence in this table above Chinese.

Table 1. Distribution of languages of documents in sentiment analysis papers.

Language of documents	No. of documents	Documents (%)
English	3190	98.91
Spanish	13	0.40
Turkish	8	0.25
Chinese	6	0.19
Portuguese	4	0.12
Rumanian	1	0.03
Estonian	1	0.03
French	1	0.03
Italian	1	0.03

Table 2. Distribution of documents investigated sentiment analysis in different languages (explicitly mentioned).

Language	No. of documents	Documents (%)	Language	No. of documents	Documents (%)
Chinese	228	35.35	Vietnamese	7	1.09
English	224	34.73	Romanian	7	1.09
Arabic	80	12.40	Kannada	5	0.78
Spanish	53	8.22	Malayalam	4	0.62
German	39	6.05	Urdu	4	0.62
Italian	34	5.27	Persian	3	0.47
French	22	3.41	Ukrainian	2	0.31
Thai	21	3.26	Hungarian	2	0.31
Portuguese	20	3.10	Nepal	2	0.31
Turkish	19	2.95	Pashto	1	0.16
Japanese	18	2.79	Marathi	1	0.16
Dutch	15	2.33	Konkani	1	0.16
Korean	14	2.17	Kurdish	1	0.16
Russian	13	2.02	Deccan	1	0.16
Greek	11	1.71	Tamil	1	0.16
Czech	10	1.55	Telugu	1	0.16
polish	8	1.24	Kazakh	1	0.16

For illustrating the languages that have been investigated more in sentiment analysis, we search different languages[3] in the title, keywords and abstract of documents. In this analysis, we only consider documents that explicitly mentioned the analysed languages. Table 2 shows amount of research in sentiment analysis in different languages. It can be seen that a lot of work has been done in Chinese and English (~70%). We test the correlation between the number of documents that investigated sentiment analysis in each of these languages and the percentage of people who speak the corresponding languages in the world. There is a high correlation (~0.9), which means that languages that have more native speakers have almost been more investigated in sentiment analysis. Among top list holders, Arabic has been investigated more than two languages that have more native speakers, that is, Spanish and Hindi.

3.2. Most prolific and impactful authors

Table 3 shows the 10 most prolific and impactful authors in the sentiment analysis field. The ranking presented here is based on the *H*-index of researchers (A researcher's *H*-index is *H* if *h* of his or her N_p papers have at least *H* citations each, and other ($N_p - H$) papers have fewer than *H* citations each) [29]. The *H*-index is a bibliometric indicator that was originally proposed for evaluation of researchers and combines quantity and impact. For this reason, we used this sole indicator to assess researchers. Only citations to papers in this field are considered in calculating the *H*-index. We also calculated the correlations between authors' citations, documents and *H*-index. All of these correlations have coefficient above 0.6. Erik Cambria from Nanyang Technological University is the most prolific and impactful researcher in this field. He is founder of SenticNet, a Singapore-based university spin-off offering B2B sentiment analysis services, and currently he works there with a team on different aspects of affective computing and sentiment analysis, including subjectivity detection, aspect extraction, microtext analysis, anaphora resolution, named entity recognition, knowledge

Table 3. Most prolific and impactful researchers.

Researcher	Affiliation	Citations	Documents	*H*-index
E Cambria	Nanyang Technological University	573	42	15
A Hussain	Stirling University	301	30	12
M Thelwall	University of Wolverhampton	503	15	9
S Poria	Nanyang Technological University	245	16	9
B Liu	University of Illinois at Chicago	271	16	7
LA Urena-Lopez	University of Jaén	150	12	6
A Gelbukh	CIC-IPN	147	13	6
MT Martin-Valdivia	University of Jaén	163	15	6
C Havasi	Massachusetts Institute of Technology	248	6	6
F Frasincar	Erasmus University Rotterdam	70	18	6
G Paltoglou	European Commission	288	11	6
Q Li	City University of Hong Kong	116	12	6
A Balahur	European Commission Joint Research Centre	121	18	6

representation and reasoning, sarcasm detection and multimodal human–computer interaction (HCI). His top most used keywords are 'sentiment analysis' (19 times), 'opinion mining' (15 times), 'sentic computing' (13 times), 'NLP' (12 times), and 'artificial intelligence' (10 times). Here we can see that the most prolific and impactful researcher in the field has used 'sentiment analysis' more than 'opinion mining'. 'Sentic computing' [30,31] is a multi-disciplinary approach to NLP and understanding at the crossroads between affective computing, information extraction, and common-sense reasoning and exploits both computer and human sciences to better interpret and process social information on the Web. In sentic computing, the analysis of natural language is based on common-sense reasoning tools which enable the analysis of text not only at the document, page or paragraph level but also at the sentence, clause and concept level [32]. Sentiment analysis requires tackling many 'NLP' sub-tasks [33,34]; this is the reason for high usage of 'NLP' by him.

3.3. Most prolific and impactful institutions and countries

Table 4 shows 10 most prolific and impactful universities in this field. There is a dominance of the East and Southeast Asian countries' universities in publishing documents (five from China, two from Hong Kong and one from Singapore), but the United States and European countries' universities are dominant in citation ranking (three from the United States and three from European countries). It should be noted that Singapore has also two universities in top list of citation ranking. We also observed a correlation coefficient of ~0.65 among the number of publications and citations of all universities.

In total, 88 different countries contributed to these documents. China (647 documents), United States (447 documents), India (386 documents), Spain (170 documents), Italy (167 documents), England (124 documents), Germany (117 documents), Japan (105 documents), South Korea (105 documents) and Canada (92 documents) are the most active countries, and they contributed to ~68% of all the documents. Based on the number of citations, the United States (received 2910 citations), China (received 2045 citations), England (received 915 citations), Canada (received 873 citations), Germany (received 843 citations), Singapore (received 752 citations), Spain (received 739 citations), Italy (received 575 citations), Scotland (received 420 citations) and India (received 392 citations) are the top countries and their documents received ~73% of all citations.

3.4. Most prolific and impactful venues of publications

There are 534 venues that publish documents related to 'sentiment analysis'. Table 5 shows the most prolific and impactful venues of publications. About 24% of documents are published in these venues. The Lecture Notes in Computer Science (LNCS) published ~12% of all of the documents. In order to assess the venues, we also included these in Table 5. The SJR indicator is obtained from Scimago journal and country ranking that is a data platform based on Scopus data. It expresses the average number of weighted citations received in the selected year by the documents published in one particular journal in the three previous years [35]. In fact, SJR is a journal indicator similar to Impact Factor (IF), but it has a 3-year citation window and does not treat all citations equally. This means that citations from more reputed journals are considered more important than the citations from less reputed journals. There is a high correlation among these indicators [36].

Table 4. Most prolific and impactful universities.

Most prolific			Most impactful		
University	Country	Documents	University	Country	Citations
Tsinghua University	China	68	University of Wolverhampton	England	512
Chinese Academy of Sciences	China	56	Tsinghua University	China	414
Nanyang Technology University	Singapore	48	Nanyang Technology University	Singapore	357
Harbin Institute of Technology	China	40	Massachusetts Institute of Technology	United States	356
Beijing University of Posts and Telecommunications	China	38	National University of Singapore	Singapore	347
City University of Hong Kong	Hong Kong	33	Technical University of Munich	Germany	326
University of Illinois	United States	28	Simon Fraser University	Canada	310
Hong Kong Polytechnic University	Hong Kong	25	University of Illinois	United States	296
Erasmus University	Netherlands	25	University Arizona	United States	289
Beihang University	China	25	University of Stirling	Scotland	287

Table 5. Most prolific and impactful venues of publications.

Most prolific			Most impactful	
Title	Documents	SJR (2015)	Title	SJR (2015)
Lecture Notes in Computer Science	391	0.25	*Nature Climate Change*	9.562
Communications in Computer and Information Science	64	0.14	*MIS Quarterly*	6.984
Expert Systems with Applications	64	1.83	*Journal of Marketing Research*	5.764
Advances in Soft Computing	50	0.15	*Journal of Accounting Research*	5.733
Knowledge-Based Systems	46	2.14	*Frontiers in Ecology and the Environment*	5.205
Procedia Computer Science	44	0.31	*Journal of Consumer Research*	4.896
Decision Support Systems	44	2.26	*Management Science*	4.384
Information Processing and Management	26	0.89	*Marketing Science*	4.34
2014 IEEE International Conference on Data Mining Workshop	22	–	*Journal of the American Statistical Association*	3.447
Frontiers in Artificial Intelligence and Applications	20	0.16	*ACM Computing Surveys (CSUR)*	3.405

SJR: Scimago Journal Rank.

3.5. Must-read papers

Table 6 shows documents with highest number of citations. These documents are probably essential to be read by everyone who wants to do research in this field.

3.6. Hot publications

In this section, we present the most used documents in the last 180 days (search date has been 21 March 2017) based on the WoS Usage Count. This count measures the level of interest in a specific item on the WoS platform. It reflects the number of times the article has met a user's information need, demonstrated by clicking the link to the full-length article at the publisher's website (via direct link or open URL) or by saving the metadata for the later usage [54]. In this way, we can identify the publications that have recently attracted more attentions. We also observed a moderate correlation between this count and the number of citations (~0.44), which shows the importance of this count; this relation has recently been observed by Chi and Glänzel [55] too. Table 7 shows the top 20 documents based on the WoS Usage Count. The average age and citations of these documents are about 4 years and 57.5 citations, respectively. These numbers show the relative recency and impact of these documents.

The most frequently used keywords in these documents are 'sentiment analysis', 'social media', 'opinion mining', 'text mining', 'Twitter', 'social media analytics', 'big data', 'user-generated content' and 'machine learning'.

Table 6. Twenty most-cited papers.

Title	Year	Citations
Lexicon-based methods for sentiment analysis [37]	2011	280
Sentiment analysis in multiple languages: feature selection for opinion classification in web forums [38]	2008	160
Sentiment strength detection for the social web [39]	2012	155
New avenues in opinion mining and sentiment analysis [40]	2013	144
Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics [41]	2011	139
Techniques and applications for sentiment analysis [4]	2013	127
Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis [42]	2009	125
Sentiment analysis: a combined approach [43]	2009	122
Norms of valence, arousal, and dominance for 13,915 English lemmas [44]	2013	102
Opinion word expansion and target extraction through double propagation [45]	2011	102
Using text mining and sentiment analysis for online forums hotspot detection and forecast [46]	2010	95
Deriving the pricing power of product features by mining consumer reviews [17]	2011	92
Learning to identify emotions in text [47]	2008	90
An empirical study of sentiment analysis for Chinese documents [48]	2008	81
User generated content: the use of blogs for tourism organizations and tourism consumers [49]	2009	71
Data mining emotion in social network communication: gender differences in MySpace [18]	2010	68
More than words: social networks' text mining for consumer brand sentiments [50]	2013	65
A machine learning approach to sentiment analysis in multilingual web texts [51]	2009	63
Document-level sentiment classification: an empirical comparison between SVM and ANN [52]	2013	62
Sentiment knowledge discovery in Twitter streaming data [53]	2010	62

Table 7. Top 20 documents based on the WoS Usage Count (search date has been 21 March 2017).

Title	Usage Count	Citations	Year
More than words: social networks' text mining for consumer brand sentiments [50]	460	65	2013
Text mining for market prediction: a systematic review [56]	303	31	2014
Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics [41]	184	139	2011
Deriving the pricing power of product features by mining consumer reviews [17]	184	92	2011
Helpfulness of online consumer reviews: readers' objectives and review cues [57]	161	38	2012
Sentiment analysis on social media for stock movement prediction [58]	156	12	2015
Analysis and mining of online social networks: emerging trends and challenges [59]	149	4	2013
Document-level sentiment classification: an empirical comparison between SVM and ANN [52]	136	62	2013
Twitter brand sentiment analysis: a hybrid system using N-gram analysis and dynamic artificial neural network [60]	134	50	2013
A novel social media competitive analytics framework with sentiment benchmarks [61]	127	13	2015
Lexicon-based methods for sentiment analysis [37]	120	280	2011
Insights from hashtag #supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research [62]	118	15	2015
Unsupervised method for sentiment analysis in online texts [63]	111	4	2016
The effect of news and public mood on stock movements [64]	110	12	2014
Text mining of news-headlines for FOREX market prediction: a multi-layer dimension reduction algorithm with semantics and sentiment [65]	107	13	2015
The impact of social and conventional media on firm equity value: a sentiment analysis approach [66]	107	36	2013
Sentiment analysis of twitter audiences: measuring the positive or negative influence of popular Twitters [67]	105	31	2012
Using text mining and sentiment analysis for online forums hotspot detection and forecast [46]	105	95	2010
Opinion mining in social media: modeling, simulating, and forecasting political opinions in the web [68]	104	31	2012
Techniques and applications for sentiment analysis [4]	103	127	2013

Table 8. Most used keywords at two different time periods.

2007–2011			2012–2016			Share change (second column)
sentiment analysis	10.95%	149	sentiment analysis	12.52%	1433	14.40%
opinion mining	10.07%	137	opinion mining	5.28%	604	–47.56%
text mining	1.76%	24	twitter	1.95%	223	430.51%
machine learning	1.62%	22	machine learning	1.62%	185	0.00%
sentiment classification	1.32%	18	social media	1.61%	184	630.96%
natural language processing	1.25%	17	text mining	1.40%	160	–20.70%
data mining	1.03%	14	natural language processing	1.35%	155	8.45%
information retrieval	0.81%	11	sentiment classification	0.91%	104	–0.41%
text classification	0.73%	10	data mining	0.60%	69	–0.43%
web mining	0.66%	9	classification	0.54%	62	0.17%
support vector machine	0.59%	8	big data	0.49%	56	100.00%
information extraction	0.59%	8	feature selection	0.44%	50	0.00%
social networks	0.77%	11	social networks	0.86%	96	0.09%
nlp	0.44%	6	text classification	0.42%	48	–0.32%
polarity classification	0.44%	6	sentiwordnet	0.41%	47	0.19%
feature selection	0.44%	6	feature extraction	0.38%	44	0.01%
opinion extraction	0.37%	5	support vector machine	0.38%	43	–0.21%
ontology	0.37%	5	naive bayes	0.34%	39	0.19%
unsupervised learning	0.37%	5	information retrieval	0.33%	38	–0.48%
twitter	0.37%		sentiment	0.31%	36	0.24%

3.7. Most important keywords

Authors provide keywords as a summary of each article's content [69]. So in this section, we first present the most used keywords in the sentiment analysis literature and then investigate trends of keywords, comparing the keywords in two different time periods. Liu [1] argues that there has been some confusion among practitioners and even researchers whether the field should be called sentiment analysis or opinion mining. Our data show that 'sentiment analysis' (used 1599 times) is more popular than 'opinion mining' (used 750 times).

The 'Twitter', as a rich data source of this field, has been mentioned a lot of times (228 times). 'Machine learning' is another frequent keyword (208 times) because many of sentiment analysis systems use one type of machine learning methods – often classifications. Since a major part of works in the sentiment analysis is done on 'social media' (187 times) and 'social networks' (109 times), researchers used these keywords to indicate the type of analysis that can be done in social media, such as investigating profitability of banks [70] sentiment of people about issues or projects [71,72] and the effect of campaigns' contents on the electoral performance [73].

Since the existing research and applications of sentiment analysis have focused primarily on the written texts, it has been an active research area in 'NLP' [1] This is the reason for the high use of term 'NLP' (173 times) and related term 'text mining' (185 times). 'Sentiment classification' is another frequent keyword (123 times). It means using classification algorithms to classify sentiments of documents, sentences and aspects as positive, neutral or negative (or more fine-grained scales). In addition to the sentiment classification, there are other classifications in the field such as subjectivity classification – that is recognition of opinion-oriented language in order to distinguish it from objective language [28] – that goes under the wider topic 'classification' (67 times). 'Data mining' is another active field that conducts sentiment analysis research [74] (used 83 times).

To investigate the changes that occurred in the keywords related to this field, we analyse the most used keywords in two different time periods, 2007–2011 and 2012–2016. The reason for neglecting time period 2002–2006 is the small number of documents and keywords in these years. The keyword clouds are presented in Figure 3. Word clouds are frequently used to visually summarise text documents [75]. Size of each keyword in the cloud shows its relative use. Numerical information is presented in Table 8. Right column of this table shows the growth of these keywords in the second time period, relative to the first period. Red cells in the first column show keywords that do not exist in the top 20 list of the second time period, and green cells are the keywords that have appeared in top 20 list in the second time period. In other words, the red cells show the keywords that have lost their importance, and the green ones are keywords that are attracting more attention in recent years. One point that is clear from Table 8 is that the support vector machine

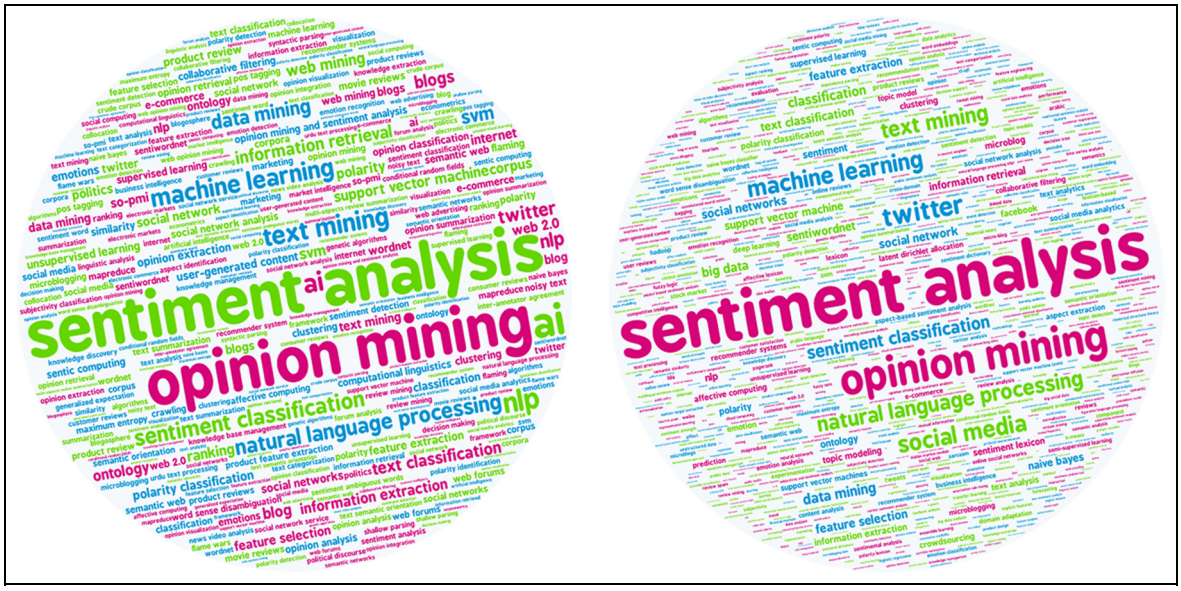

Figure 3. Word clouds of the keywords used in the sentiment analysis field in two different time periods.[4]

is probably the most used classification technique for sentiment analysis that is in line with Thelwall et al. [39] The next most prevalent technique is the Naive Bayes which is currently attracting more attention.

First, it can be seen that 'sentiment analysis' has attracted more attention in the second time period (~14%), whereas 'opinion mining' has witnessed a sharp decline (~48%). Moreover, the use of the 'sentiment' keyword has increased in the second time period (the last row in Table 8). Currently, Twitter and Facebook are focal points of many sentiment analysis applications [4], and this is reflected as a growth in using social media keywords in the second time period (positive growth of 'social media', 'Twitter' and 'social networks'). In addition, there is a positive growth in using 'NLP' and a negative growth in using 'text mining'. Some keywords such as 'web mining' and 'information extraction' have been removed from the top list of second time period, while some other keywords such as 'big data' and 'SentiWordNet' have appeared.

3.7. Future trends

To explore future trends, we search through abstracts of our data in order to extract visions of researchers about future works of the field. We only explore papers that explicitly mention their suggestions in their abstracts. Table 9 shows these papers and their suggestions for future works.

4. Conclusion

In this article, we provided a perspective on the global trends in 'sentiment analysis' research. We performed a bibliometric analysis of scientific documents, distribution of subject categories, most prolific and impactful authors, languages, institutions, venues of publications and their geographic distribution, most cited and hot documents, trends of keywords and future works. Our analysis shows that there has been a significant growth in this field; the average annual growth rate is ~79%. 'Computer science', 'engineering', 'telecommunications', 'linguistics', 'operations research and management science', 'information science and library science', 'business and economics', 'automation and control systems', 'robotics' and 'social sciences' are the most frequent subject categories, and LNCS is the most active venue of publication in this field. Erik Cambria from Nanyang Technological University is the most prolific and impactful researcher with 42 documents (~1.3% of all documents) and 573 citations (~5.5% of all citations).

A small group of productive countries contributed to a substantial number of articles: the top 10 countries contributed to ~68% of the total documents and their documents received ~73% of citations. China has the leading position in the global sentiment analysis research, while in terms of citations the United States has the leading position. At an institution level, Tsinghua University is the most prolific one, and based on the number of citations, University of Wolverhampton

Table 9. Future works suggested by researchers.

Title	Year	Suggestions
Survey on aspect-level sentiment analysis [76]	2016	Semantically rich concept-centric aspect-level sentiment
Survey of text sentiment analysis [77]	2016	Cross-lingual sentiment analysis, multimodal aspect analysis and applying emotion detection to new applications
Survey on diverse facets and research issues in social media mining [78]	2016	Multiscale community detection
Review of intelligent microblog short text processing [79]	2016	Real-time online processing of big data
SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks [80]	2016	SentiHealth method could be instantiated as other disease-based tools during future works, for instance, SentiHealth-HIV, SentiHealth-Stroke and SentiHealth-Sclerosis
Emerging directions in predictive text mining [81]	2015	Resource-poor languages and multilingual texts
Web-based textual analysis of free-text patient experience comments from a survey in primary care [81]	2015	Whether more sophisticated methods of textual analysis (e.g. sentiment analysis, natural language processing) could add additional levels of understanding
Sentiment analysis for social media images [82]	2015	Exploring sentiment on signed social network

is in first place. There is a correlation between the number of authors and number of citations, indicating the higher quality level of collaborative documents. It can suggest to researchers in this field that they should work with other researchers and benefit from their counterparts' abilities. As a result of the keyword analysis, 'sentiment analysis' is a more accepted term than 'opinion mining', and some new keywords such as 'big data' are gaining more popularity in recent years due to growth of accessibility of web data and computational facilities. 'Twitter' is the most used social network in this field and has attracted a lot of attention in recent years. The rising use of 'Twitter' along with the emergence of 'social media' and the growth in 'social networks' keywords shows more concentration on such data sources recently. 'Support vector machine' is the most used classification method, and in recent years 'Naive Bayes' has attracted more attention. As a tool for lexicon-based sentiment analysis, 'SentiWordNet', published by Esuli and Sebastiani in 2006 [83], has attracted more attention, but based on the most used keywords, it can be seen that machine learning approach is more prevalent than lexicon-based approach. 'Unsupervised learning' has been removed from the top list that combined with the rising of 'classification' and can show that in recent years researchers have paid more attention to supervised learning.

As a limitation of our research, it should be noted that our results are restricted in some directions. The first limitation comes from limitations of database that we used because all our results are retrieved from WoS, but a research [84] shows that there is a high overlap between WoS and Scopus as another possible option for bibliometric study in natural science and Engineering. Also using Google Scholar as a source for bibliometrics involves issues such as data validity and ease of manipulation of citation data [84]. Another limitation is the phrases that we used, which may cause some data reduction. Because if a paper about sentiment analysis did not use the terms we used for search, it does not appear in our data set. We hope that this study will be useful for researchers who are trying to do a targeted research in the sentiment analysis field.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Abdalsamad Keramatfar https://orcid.org/0000-0001-6826-4692

Notes

1. www.scimagojr.com
2. www.vosviewer.com
3. List of languages is from here: https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
4. For the larger size and the interactive version see the following links: https://tagul.com/iimlfi598qcv/sentiment-analysis-2007-2011; https://tagul.com/9p178h39lc7c/sentiment-analysis-2012-2016

References

[1] Liu B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press, 2015.
[2] Sammut C and Webb G.I. Encyclopedia of Machine Learning and Data Mining, Second Edition. Berlin: Springer, 2017.
[3] Garrard WC. *Development of Tools for the Analysis of Messages in Controlled Social Network Environments*. Pittsburgh, PA: University of Pittsburgh, 2017.
[4] Feldman R. Techniques and applications for sentiment analysis. *Comm ACM* 2013; 56(4): 82–89.
[5] Banerjee H, Dey R, Chatterjee S, et al. (eds). Movie recommendation system using particle swarm optimization. In: *2017 8th annual industrial automation and electromechanical engineering conference (IEMECON)*, Bangkok, Thailand, 16–18 August 2017.
[6] Banshal SK, Uddin A, Singhal K, et al. Computer science research in India: a scientometric study. In: *2015 annual IEEE India conference (INDICON)*, New Delhi, India, 17–20 December 2015.
[7] Ibáñez A, Bielza C and Larrañaga P. Relationship among research collaboration, number of documents and number of citations: a case study in Spanish computer science production in 2000–2009. *Scientometrics* 2013; 95(2): 689–716.
[8] Lin CL, Yang HL, Yen WC, et al. Trend analysis of virtual community productivity based on SSCI database by bibliometric methodology. In: *2011 5th international conference on new trends in information science and service science (NISS)*, Macau, China, 24–26 October 2011.
[9] Ding Y. Semantic web: who is who in the field – a bibliometric analysis. *J Informat Sci* 2010; 36(3): 335–356.
[10] Yao J (ed.). A ten-year review of granular computing. In: *2007 GRC 2007 IEEE international conference on granular computing*, Fremont, CA, 2–4 November 2007.
[11] Appel O, Chiclana F and Carter J. Main concepts, state of the art and future research questions in sentiment analysis. *Acta Polytechnica Hungarica* 2015; 12(3): 87–108.
[12] Li M and Chu Y. Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis. *J Informat Sci* 2016; 43: 725–741.
[13] Yang K and Meho LI. Citation analysis: a comparison of Google Scholar, Scopus, and Web of Science. *Proc Assoc Informat Sci Tech* 2006; 43(1): 1–15.
[14] Persson O. BibExcel: a toolbox for bibliometricians Version 2008. *Int Soc Scientometr Informetr* 2008; 4: 8.
[15] Mattern F. Bibliometric evaluation of computer science–problems and pitfalls (Invited Talk). *SARIT*, https://www.vs.inf.ethz.ch/publ/slides/Mattern-Bibliometry-SARIT06.pdf; 2008.
[16] Lippi M and Torroni P. MARGOT: a web server for argumentation mining. *Expert Syst Appl* 2016; 65: 292–303.
[17] Archak N, Ghose A and Ipeirotis PG. Deriving the pricing power of product features by mining consumer reviews. *Manag Sci* 2011; 57(8): 1485–1509.
[18] Thelwall M, Wilkinson D and Uppal S. Data mining emotion in social network communication: gender differences in MySpace. *J Assoc Informat Sci Tech* 2010; 61(1): 190–199.
[19] Thet TT, Na JC and Khoo CS. Aspect-based sentiment analysis of movie reviews on discussion boards. *J Informat Sci* 2010; 36(6): 823–848.
[20] Tumasjan A, Sprenger TO, Sandner PG, et al. Election forecasts with Twitter: how 140 characters reflect the political landscape. *Soc Sci Comp Rev* 2011; 29(4): 402–418.
[21] Nasukawa T and Yi J (eds). Sentiment analysis: capturing favorability using natural language processing. In: *Proceedings of the 2nd international conference on knowledge capture*, Sanibel Island, FL, 23–25 October 2003.
[22] Yi J, Nasukawa T, Bunescu R, et al. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: *2003 ICDM 2003 third IEEE international conference on data mining*, Melbourne, FL, 19–22 November 2003.
[23] Smrž P (ed.). Using WordNet for opinion mining. In: *Proceedings of the third international WordNet conference*, South Jeju Island, Korea, 55–26 January 2006.
[24] Lokker C, McKibbon KA, McKinlay RJ, et al. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *BMJ* 2008; 336(7645): 655–657.
[25] Onodera N and Yoshikane F. Factors affecting citation rates of research articles. *J Assoc Informat Sci Tech* 2015; 66(4): 739–764.
[26] Annalingam A, Damayanthi H, Jayawardena R, et al. Determinants of the citation rate of medical research publications from a developing country. *Springerplus* 2014; 3(1): 140.

[27] Nourmohammadi H, Keramatfar A and Rafie M. What factors associated with citation impact in cell biology? In: *Collaboration in Science and in Technology* 2016, Nancy, 12–15 December.

[28] Ravi K and Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowled Based Syst* 2015; 89(Supplement C): 14–46.

[29] Bornmann L and Daniel H-D. Does the H-index for ranking of scientists really work? *Scientometrics* 2005; 65(3): 391–392.

[30] Cambria E and Hussain A. *Sentic Computing: Techniques, Tools, and Applications*. London: Springer Science & Business Media, 2012.

[31] Cambria E and Hussain A. *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. London: Springer, 2015.

[32] Cambria E, Das D, Bandyopadhyay S, et al. *A Practical Guide to Sentiment Analysis*. London: Springer, 2017.

[33] Li Y, Pan Q, Yang T, et al. Learning word representations for sentiment analysis. *Cognit Computat* 2017; 9: 848–851.

[34] Chaturvedi I, Poria S and Cambria E. Basic tasks of sentiment analysis. *Arxiv*. Epub ahead of print 18 October 2017. DOI: 10.1007/978-1-4614-7163-9_110159-1.

[35] SCImago Research Group. Description of Scimago journal rank indicator. http://www.scimagojr.com/SCImagoJournalRank.pdf; 2008.

[36] Rafie M and Keramatfar A. Self-citations of cell biology journals, *Science Evaluation*, 2016.

[37] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis. *Computat Linguist* 2011; 37(2): 267–307.

[38] Abbasi A, Chen H and Salem A. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. *ACM Trans Informat Syst* 2008; 26(3): 12.

[39] Thelwall M, Buckley K and Paltoglou G. Sentiment strength detection for the social web. *J Assoc Informat Sci Tech* 2012; 63(1): 163–173.

[40] Cambria E, Schuller B, Xia Y, et al. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Syst* 2013; 28(2): 15–21.

[41] Ghose A and Ipeirotis PG. Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. *IEEE Trans Knowled Data Eng* 2011; 23(10): 1498–1512.

[42] Wilson T, Wiebe J and Hoffmann P. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computat Linguist* 2009; 35(3): 399–433.

[43] Prabowo R and Thelwall M. Sentiment analysis: a combined approach. *J Informetr* 2009; 3(2): 143–157.

[44] Warriner AB, Kuperman V and Brysbaert M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res Met* 2013; 45(4): 1191–1207.

[45] Qiu G, Liu B, Bu J, et al. Opinion word expansion and target extraction through double propagation. *Comput Linguist* 2011; 37(1): 9–27.

[46] Li N and Wu DD. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decis Support Syst* 2010; 48(2): 354–368.

[47] Strapparava C and Mihalcea R. Learning to identify emotions in text. In: *Proceedings of the 2008 ACM symposium on applied computing*, Fortaleza, 16–20 March 2008.

[48] Tan S and Zhang J. An empirical study of sentiment analysis for Chinese documents. *Expert Syst Appl* 2008; 34(4): 2622–2629.

[49] Akehurst G. User generated content: the use of blogs for tourism organisations and tourism consumers. *Service Business* 2009; 3(1): 51.

[50] Mostafa MM. More than words: social networks' text mining for consumer brand sentiments. *Expert Syst Appl* 2013; 40(10): 4241–4251.

[51] Boiy E and Moens M-F. A machine learning approach to sentiment analysis in multilingual Web texts. *Informat Retrieval* 2009; 12(5): 526–558.

[52] Moraes R, Valiati JF and Neto WPG. Document-level sentiment classification: an empirical comparison between SVM and ANN. *Expert Syst Appl* 2013; 40(2): 621–633.

[53] Bifet A and Frank E. Sentiment knowledge discovery in twitter streaming data. In: *International conference on discovery science*, Kyoto, Japan, 15–17 October 2010.

[54] Web of Knoweldge, Usage Count, https://images.webofknowledge.com/WOKRS517B2.3/help/CSCD/hp_usage_score.html (accessed 10 March 2018).

[55] Chi P-S and Glänzel W. An empirical investigation of the associations among usage, scientific collaboration and citation impact. *Scientometrics* 2017; 112(1): 403–412.

[56] Nassirtoussi AK, Aghabozorgi S, Wah TY, et al. Text mining for market prediction: a systematic review. *Expert Syst Appl* 2014; 41(16): 7653–7670.

[57] Baek H, Ahn J and Choi Y. Helpfulness of online consumer reviews: readers' objectives and review cues. *Int J Elec Comm* 2012; 17(2): 99–126.

[58] Nguyen TH, Shirai K and Velcin J. Sentiment analysis on social media for stock movement prediction. *Expert Syst Appl* 2015; 42(24): 9603–9611.

[59] Bhat SY and Abulaish M. Analysis and mining of online social networks: emerging trends and challenges. *Wiley Interdis Rev* 2013; 3(6): 408–444.
[60] Ghiassi M, Skinner J and Zimbra D. Twitter brand sentiment analysis: a hybrid system using N-gram analysis and dynamic artificial neural network. *Expert Syst Appl* 2013; 40(16): 6266–6282.
[61] He W, Wu H, Yan G, et al. A novel social media competitive analytics framework with sentiment benchmarks. *Informat Manage* 2015; 52(7): 801–812.
[62] Chae BK. Insights from hashtag #supplychain and Twitter Analytics: considering Twitter and Twitter data for supply chain practice and research. *Int J Product Econom* 2015; 165: 247–259.
[63] Fernández-Gavilanes M, Álvarez-López T, Juncal-Martínez J, et al. Unsupervised method for sentiment analysis in online texts. *Expert Syst Appl* 2016; 58: 57–75.
[64] Li Q, Wang T, Li P, et al. The effect of news and public mood on stock movements. *Informat Sci* 2014; 278: 826–840.
[65] Nassirtoussi AK, Aghabozorgi S, Wah TY, et al. Text mining of news-headlines for FOREX market prediction: a multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Syst Appl* 2015; 42(1): 306–324.
[66] Yu Y, Duan W and Cao Q. The impact of social and conventional media on firm equity value: a sentiment analysis approach. *Decis Support Syst* 2013; 55(4): 919–926.
[67] Bae Y and Lee H. Sentiment analysis of Twitter audiences: measuring the positive or negative influence of popular twitterers. *J Assoc Informat Sci Tech* 2012; 63(12): 2521–2535.
[68] Sobkowicz P, Kaschesky M and Bouchard G. Opinion mining in social media: modeling, simulating, and forecasting political opinions in the web. *Govern Informat Quarter* 2012; 29(4): 470–479.
[69] Liu F, Lin A, Wang H, et al. Global research trends of geographical information system from 1961 to 2010: a bibliometric analysis. *Scientometrics* 2016; 106(2): 751–768.
[70] Tang C, Mehl MR, Eastlick MA, et al. A longitudinal exploration of the relations between electronic word-of-mouth indicators and firms' profitability: findings from the banking industry. *Int J Informat Manage* 2016; 36(6): 1124–1132.
[71] Estévez-Ortiz F-J, García-Jiménez A and Glösekötter P. An application of people's sentiment from social media to smart cities. *El Profesional De La Información* 2016; 25(6): 851–858.
[72] Jiang H, Qiang M and Lin P. Assessment of online public opinions on large infrastructure projects: a case study of the Three Gorges Project in China. *Environ Imp Assess Rev* 2016; 61: 38–51.
[73] Ceron A and d'Adda G. E-campaigning on Twitter: the effectiveness of distributive promises and negative campaign in the 2013 Italian election. *New Media Soc* 2016; 18(9): 1935–1955.
[74] Biyani P, Bhatia S, Caragea C, et al. Using non-lexical features for identifying factual and opinionative threads in online forums. *Knowled Based Syst* 2014; 69: 170–178.
[75] Heimerl F, Lohmann S, Lange S, et al. Word cloud explorer: text analytics based on word clouds. In: *2014 47th Hawaii international conference on system sciences (HICSS)*, Waikoloa, HI, 6–9 January 2014.
[76] Schouten K and Frasincar F. Survey on aspect-level sentiment analysis. *IEEE Trans Knowled Data Eng* 2016; 28(3): 813–830.
[77] Yang L, Zhu J and Tang S. Survey of text sentiment analysis. *J Comp Appl* 2013; 33(6): 1574–1607.
[78] Ambika P and Rajan MB. Survey on diverse facets and research issues in social media mining. In: *International conference on research advances in integrated navigation systems (RAINS)*, Bangalore, India, 6–7 May 2016.
[79] Huang W, Li Z, Zhang L, et al. Review of intelligent microblog short text processing. *Web Intellig* 2016; 14: 211–228.
[80] Rodrigues RG, das Dores RM, Camilo-Junior CG, et al. SentiHealth-Cancer: a sentiment analysis tool to help detecting mood of patients in online social networks. *Int J Med Informat* 2016; 85(1): 80–95.
[81] Indurkhya N. Emerging directions in predictive text mining. *Wiley Interdis Rev* 2015; 5(4): 155–164.
[82] Wang Y and Li B. Sentiment analysis for social media images. In: *2015 IEEE international conference on data mining workshop (ICDMW)*, Washington, DC, 14–17 November 2015.
[83] Esuli A and Sebastiani F. SentiWordNet: a high-coverage lexical resource for opinion mining. *Evaluation* 2007; 17: 1–26.
[84] Mongeon P and Paul-Hus A. The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics* 2016; 106(1): 213–228.