ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

# The Evolution of Sentiment Analysis of Multiple Research Paper

[1] Jyoti
MCA, M. D. University, Rohtak

*Abstract:* - **Sentiment analysis is one of the fastest growing research areas in computer science, making it challenging to keep track of all the activities in the area. We present a computer-assisted literature review, where we utilize both text mining and qualitative coding, and analyze 6996 papers from Scopus. We find that the roots of sentiment analysis are in the studies on public opinion analysis at the beginning of 20th century and in the text subjectivity analysis performed by the computational linguistics community in 1990's. However, the outbreak of computer-based sentiment analysis only occurred with the availability of subjective texts on the Web. Consequently, 99% of the papers have been published after 2004. Sentiment analysis papers are scattered to multiple publication venues, and the combined number of papers in the top-15 venues only represent ca. 30% of the papers in total. We present the top-20 cited papers from Google Scholar and Scopus and a taxonomy of research topics. In recent years, sentiment analysis has shifted from analyzing online product reviews to social media texts from Twitter and Facebook. Many topics beyond product reviews like stock markets, elections, disasters, medicine, software engineering and cyberbullying extend the utilization of sentiment analysis.**

## I. INTRODUCTION

''The pen is mightier than the sword'' proposes that free com- munication (particularly written language) is a more effective tool than direct violence [1]. Sentiment analysis is a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language [2]. Traditionally, sentiment analysis has been about opinion polarity, i.e., whether someone has positive, neutral, or negative opinion to- wards something [3]. The object of sentiment analysis has typically been a product or a service whose review has been made public on the Internet. This might explain why sentiment analysis and opinion mining are often used as synonyms, although, we think it is more accurate to view sentiments as emotionally loaded opinions. Ancient works in East and West mingle with these subjects. ''The Art of War'' has a chapter on espionage that deals with spy recruiting and betrayal, while in the beginning of ''Iliad'' the leader of Greeks Agamemnon tries to gauge the fighting spirit of his men. Voting as a method to measure public opinion on policy has its roots in the city state of Athens in the 5th century BCE [5]. Efforts in capturing public opinion by quantifying and measuring it from questionnaires have appeared in the first decades of twen- tieth century [6], while a scientific journal on public opinion was established in 1937 [7].

We have seen a massive increase in the number of papers focusing on sentiment analysis and opinion mining during the recent years. According to our data, nearly 7000 papers of this topic have been published and, more interestingly, 99% of the papers have appeared after 2004 making sentiment analysis one of the fastest growing research areas. Although the present paper focuses on the research articles of sentiment analysis, we can see that the topic is getting attention in the general public, as well. Fig. 1 shows the increase in searches made with a search string ''sentiment analysis'' in Google search engine. We observed that the first academic studies measuring public opinions are during and after WWII and their motivation is highly political in nature [8,9]. The outbreak of modern sentiment anal- ysis happened only in mid-2000's, and it focused on the product reviews available on the Web, e.g., [3]. Since then, the use of sentiment analysis has reached numerous other areas such as the prediction of financial markets [10] and reactions to terrorist attacks [11]. Additionally, research overlapping sentiment analysis and natural language processing has addressed many problems that contribute to the applicability of sentiment analysis such as irony detection [12] and multi-lingual support [13]. Furthermore, with respect to emotions,1 efforts are advancing from simple polarity detection to more complex nuances of emotions and dif- ferentiating negative emotions such as anger and grief [15].

An

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

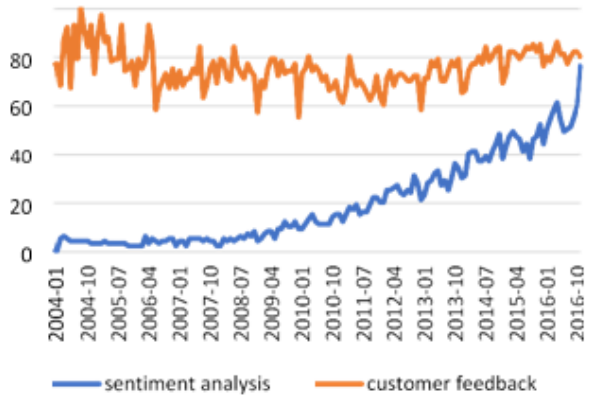


***Fig. 1. Google Trends (www.google.com/trends) data showing the relative popu- larity of search strings ''sentiment analysis'' and ''customer feedback''.***

academic literature review can only focus on one particular area of sentiment analysis as it typically includes between 10 and 100 studies, e.g., a recent systematic review of the prediction of financial markets with sentiment analysis reviewed 24 papers [10]. To overcome the challenges caused by the increasing number of articles about sentiment analysis, we present a computer-assisted literature review and a bibliometric study of sentiment analysis. Studies like the one we are presenting should be helpful when working in an area with large volumes of literature.

We think that the present article can offer an overview of sentiment analysis to newcomers and it may provide valuable to more seasoned scholars for educational purposes. To provide such an overview, we characterize the field of sentiment analysis by answering the following research questions that are typical in bibliometric studies, e.g., [16–19]:

RQ1: What is the number of papers in sentiment analysis? This question helps us in understanding the volume of work in sentiment analysis. We can also observe yearly trends that tell us about the history of the topic and can help in predicting the future.

RQ2: What is the number of citations in sentiment analysis? This question addresses the impact of sentiment analysis. Similarly to RQ1, we can observe yearly trends about the history of the topic and can also help in predicting the future. RQ3: What are the most popular publication venue for senti- ment analysis? This question shows the popular venues for publishing sentiment analysis studies. Understanding the different communities related to sentiment analysis helps understanding the entire field.

RQ3: What research topics have been investigated with senti- ment analysis? Given that the topic has rapidly grown very large, we use text clustering to get on overview of the different areas of sentiment analysis. Our text clustering approach originates from influential paper by Griffiths and Steyvers [20] titled ''Finding scientific topics''. We support the automated text clustering with manual qualitative anal- ysis and they jointly provide a thematic overview of the research topics in this field.

RQ4: What are the most cited original works and literature reviews in sentiment analysis and what research topics do these papers cover? Citations are often referred as the backbone of science. Investigation of the landmarks in sentiment anal- ysis can demonstrate the most interesting and impactful work in this area. Using citation counts has been previously done for example [21,22] that studied the 100 most cited papers published in Nature and in Software Engineering.

The paper makes the following six contributions. First, we show how attempts to understand public opinion at the start of 20th century through questionnaires and subjectivity analysis in computational linguistics gave birth to this topic. For decades, the research area was mostly ignored until massive amounts of opinions available in the Web gave birth to modern sentiment analysis. Second, we demonstrate how modern sentiment analysis has received a 50-fold growth in ten years between 2005 and 2016. The number of citations has grown along with the paper counts. Third, we show that most popular venues for sentiment analysis are series with large volumes: Lecture Notes In Computer Science, CEUR Workshop, ACM International Conference Proceeding Series. However, the venues with the highest shares of sentiment analysis papers adjusted by the total publication output are Procesamiento De Lenguaje Natural, International Conference Recent Advances In Natural Language Processing, and Communications In Computer And Information Science.

This paper is structured as follows. Section 2 explains our re- search methods. Beyond that our structure deviates from conven- tional research articles. The related work in the area of sentiment analysis is an output of the results in Section 3. Section 4 discusses the limitation of this work and Section 5 provides the conclusions.

## II. RESEARCH METHODS

In this section, we present the research methods that we used in order to answer our research questions. Section 2.1 explains our search strategy. Section 2.2 describes the quantitative analysis while Section 2.3 tells how we analyzed the publication venues. In Section 2.4, we explain how we studied the research topics with automated text analysis and Section 2.5 explains the manual qualitative classification we did on top of the automated analysis. Finally, Section 2.6 describes our analysis of the top-cited papers.

### 2.1 Searching literature

We used Scopus search engine for our literature search and data retrieval. Our results in Sections 3.1–3.6 are based on Scopus data. According to its developer Elsevier, ''Scopus is

the largest abstract and citation database of peer-reviewed literature: sci- entific journals, books and conference proceedings''.2 Thus, it

***Table 1***
***Search strings and the share of search hits in Scopus (with overlaps).***

Search strings and the share of search hits in Scopus (with overlaps).

| Search term | % of hits |
|---|---|
| ''Sentiment analysis'' | 68.5% |
| ''Opinion mining'' | 29.1% |
| ''Sentiment classification'' | 18.0% |
| ''Opinion analysis'' | 5.6% |
| ''Semantic orientation'' | 3.8% |
| Sentiwordnet | 2.7% |
| ''Opinion classification'' | 1.4% |
| ''Sentiment mining'' | 1.3% |
| ''Subjectivity analysis'' | 1.1% |
| Sentic | 1.0% |
| ''Subjectivity classification'' | 0.8% |

should offer the widest coverage of scientific literature that one can achieve with a single search engine. We are aware that there has been much discussion about which academic search database offers the least error-prone results and the highest coverage for knowledge discovery and researchers' assessment, e.g., [23–26]. Later in Section 4, we report how Scopus was found to be the most reliable [26] and comprehensive [25] academic search engine that is compatible with our aims. Scopus also offers advanced search engine features such as finding variant spellings. Another benefit of using Scopus is that it allows downloading paper titles, and abstracts in batches of 2000 papers at a time. This enables further offline analysis, for example in the form of text clustering. The batch export functionalities are limited in Google Scholar and Web of Science which thus become less usable for our purposes.

In more detail, we executed 11 queries corresponding to the search strings in Table 1 with Publish or Perish 5 software. We retrieved the top 100 results sorted by relevance for each string as Google Scholar does not allow sorting based on citations. As Google Scholar's relevance also considers citations, we are confi- dent that we were able to estimate the top cited sentiment analysis papers from although there is no way to guarantee this. These 1100 hits (11*100) were exported from Publish or Perish software to a CSV-file and merged and sorted for citations. This additional check should ensure that historical hallmark papers of sentiment analysis have not been missed when reconstructing the history of sentiment analysis.

**2.2 Quantitative analysis of paper and citation counts**

We performed quantitative analysis by plotting histograms of the paper and citation counts. Our analysis scripts are openly available.3

**2.3 Analysis of publication venues**

Given the large body of work we were interested in finding out the publication venues of the papers in order to further demarcate the publication area. Computer science disciplines publish the majority of research results into conference proceedings instead of journals.4 While journals rarely change their title name and vary in issue and volume numbers, we discovered that conferences proceedings names are not reliable over the years. In order to overcome this issue, the second author cleaned the venues that we retrieved from Scopus using R language [29]. The cleanup criteria included the deletion of the years from the conference names as well as their enumeration (e.g., 2015, 1st, 22nd, etc.). We also removed substrings related to the term proceedings (e.g., ''Proc. of the'', ''Proceedings of the''), because the term was also not used consistently over the years by the same conferences. After the cleanup, we found that overall the papers were distributed across 1526 different sources.

**2.4 Word clouds and Topic modeling with LDA of research topics**

Due to our high volume of articles (nearly 7000) qualitative manual analysis of all the papers would have exceeded our re- sources. Therefore, we used the text mining and clustering with R language [29]. We did basic word clouds and dissimilarity word clouds using R package ''wordcloud'' [30]. We used Latent Dirichlet Allocation (LDA) topic modeling to cluster our papers. LDA is a soft clustering algorithm developed for several scopes including text clustering. LDA approaches text clustering by acknowledging that each text document can be about multiple topics, e.g., a document describing cat foods would be about two more general topics, namely cats and foods. For example, a paper analyzing hotel reviews with a Bayesian algorithm could have the highest shares of words coming from two topics one containing words ''hotel'', ''reviews'', and the second containing words ''Bayesian'', ''algorithm''.

In more detail, we used LDA from R package ''topicmodels'' [31]. The origins of the LDA approach for scientific topic detection lie in the influential paper by Griffiths and Steyvers [20] modeling scien- tific topics that was published in PNAS. Parts of the R computations are from Ponweiser [32] who replicated the study by Griffiths and Steyvers. Previously, we followed this approach when we analyzed software engineering literature [16]. We followed the approaches and advices of the papers in the previous section and proceeded as follows. First, we removed all publisher's copyright information in the abstract. This text occa- sionally contains publisher names, e.g., IEEE or ACM. Second, we created and manipulated our corpus with R-

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

package ‘‘tm’’ [33]. The corpus was created by merging the title and the abstract of each article. Then, we performed the following preprocessing steps: removed punctuation and numbers, made all the letters lower case, removed common stop words for English, and finally created a document-term-matrix aka a document-word-matrix, which describes the frequency of terms, while removing words with less than three letters and words occurring less than 5 times as in [32]. We used hyper-parameter settings with the default values suggested by Griffiths and Steyvers [20], i.e., alpha 50/k and beta 0.1. The hyper-parameters describe the Bayesian prior beliefs about how likely is each document to contain multiple topics (alpha) and how likely is each topic to contain multiple words (beta). As in [20] we have fixed small beta that should lead to a quite high number of fine grained topics. Such specific topics were then furthered analyzed with qualitative coding, see next section. Finally, similar to [20] we tuned for k, i.e., finding the optimal number of topics (k) using a log-likelihood measure.

### 2.5 Qualitative coding of automatically created research topics

Since our LDA topic modeling from the previous step suggested that 108 topics would be optimal, we opted to create 108 topics. As understanding 108 topics as a flat list is difficult, we used qualitative coding strategy to improve readability and understand- ability. That is, we categorized the topics using qualitative coding as presented in many qualitative analysis handbooks, e.g., [34]. During the qualitative coding phase the mind mapping tool Mindmup.com was used to enable real-time collaboration between the first and second author. Fig. 2 shows a snap-shot of the Mindmup.com tool as it was used during qualitative coding. One can also see some of the topics produced by the LDA topic modeling, i.e., bubbles starting with the text ‘‘Topic’’, and the emerging qualitative classification.

### 2.6 Analysis of the top-cited papers

In the final phase, the third author studied and summarized the top-cited papers in our data set. The classification created in the previous phase was utilized to show which parts of the classification had been studied in the most cited papers. To find out the top cited papers we used citation counts normalized for time from Scopus and Google Scholar.

## III. RESULTS

### 3.1. Number of papers and a brief history of sentiment analysis (RQ1)

The number of papers in sentiment analysis is increasing rapidly as can be observed from Fig. 3. The field has also changed in terms of content over the years. Prior to availability of mas- sive amount of text and opinions online, studies mainly relied on

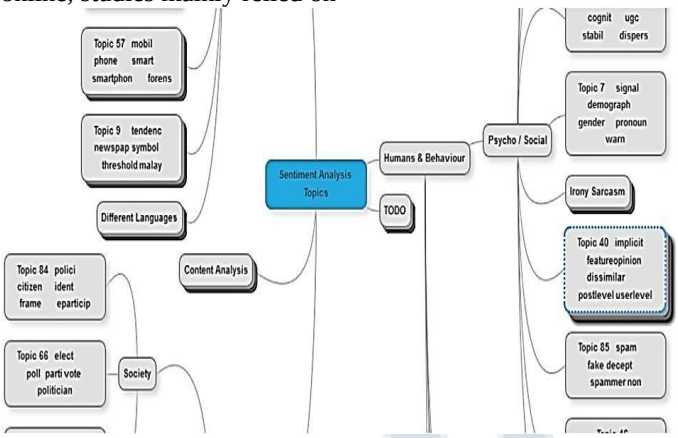


*Fig. 2. Mindmup.com tool was used to collaborate in quality coding of the topics.*

## IV. LIMITATIONS

We do realize that using our search strings in the title, abstract and key words has resulted in the inclusion of papers that only use sentiment analysis to motivate their work but we think those works are still valid as advances in, e.g., in sarcasm detection can boost advanced in sentiment analysis as well. However, we think that those papers are in the minority. This belief is based on the finding that among our top-cited literature reviews or original works, altogether a sample of 17 papers, we only found studies that were directly about sentiment analysis.

Our choice of employing the Scopus database with minor ad- ditions from Google Scholar could be considered a limitation of this study. The most widely employed academic databases are Google Scholar, Scopus, and the Web of Science [23]. e.g., [23–26]. As reported by Harzing and Alakangas [23], Google Scholar offers the most comprehensive coverage of the literature, but The results

of their study showed, among other results, that Scopus is capable to provide the highest field coverage with an overrepresentation of natural sciences and engineering and biomedical science, which is our case. Additionally, a major motivation for choosing Scopus was its ability to easily export high numbers of search results for further processing, a feature missing from Google Scholar and Web of Science. Thus, we believe that our choice of using Scopus augmented with some Google Scholar results, while not perfect, has been optimal for the task at hand. The software engineering backgrounds of both authors have likely influenced the qualitative classification result.

## V. DISCUSSIONS AND CONCLUSIONS

In this article, we presented a computer-assisted literature review, using automated text clustering with manual qualitative analysis, and a bibliometric study of sentiment analysis of 6,996 pa- pers. We investigated the history of sentiment analysis, evaluated the impact of sentiment analysis and its trends through a citation and bibliometric study, delimited the communities of sentiment analysis by finding the most popular publication venue, discov- ered which research topics have been investigated in sentiment analysis, and reviewed the most cited original works and literature reviews in sentiment analysis. We found that the citation counts have increased along with the paper counts. We found for example that the top-cited paper of sentiment analysis exceeds the citation counts of any paper published in a much mature and larger research area of software engineering. It is notable that the pool of papers used for sentiment analysis was only roughly 5000, while our past work on software engineering had nearly 70,000 papers in the pool. Thus, sentiment analysis is also making an impact at least when measured by the number of citations. We discovered that sentiment analysis papers are scattered to multiple publication venues and the combined number of papers in the top-15 venues only represent ca. 30% of the papers in total. Investigation of the research topics showed that sentiment analysis had used multiple data sources related to or coming from newspapers, tweets, photos, chats for example. We found that numerous data-analysis methods had been used and we classified them to three groups namely: machine learning, natural language processing and sentiment analysis specific methods. With respect to research goals, i.e., the targets of or issues with sentiment analysis we found numerous application areas like, movies, travel, health, argumentation, interaction with audience, elections, exper- tise, sarcasm, spam, dialects and so on. Investigation of changes in the research topics found that the most recent papers (2014–2016) had more focus on social media such Twitter and Facebook. Other topics which became popular in the recent years had been mobile devices, stock market, and human emotions. On the contrary, the papers published 2013 or prior to that had focused more on sentiment analysis in the context of pro- duce reviews, product features and analysis of political situations such as elections.

We produced a comprehensive taxonomy of the research topics of sentiment analysis with text mining and qualitative coding. We believe that studies like our can be act as broad overviews of areas that are too large to be investigated with traditional literature reviews.

Shortly before the submission of the present paper, another bibliometric study on sentiment analysis was published by Piryani et al. [66]. There are similarities but also many differences between our and their work. Similarities are:

Both studies are bibliometric studies on sentiment analysis. Both articles show roughly over tenfold growth trends in papers per year in a decade. Paper counts from 2005 to 2015 in Piryani et al. show growth trend of 1400% (98/7) while we show growth trend of 4139% (1490/36).

Both patterns analyze citation patterns but comparison can- not be made as we present citation counts and non-cited papers while they draw a map of citations.

Notable differences between the papers are:

We use Scopus supplemented with Google Scholar for top- cited papers as our data source. They use Web of Science. Our search identifies total of 6996 papers from Scopus while their work finds 697 papers.

Some differences in the search terms are that they also in- cluded terms like ‘‘affective computing’’ which we excluded as it often refers to sensor based measures. We have more detailed search terms about survey and text based senti- ment analysis that are missing in their work e.g. ‘‘opinion analysis’’ and ’’semantic orientation’’.

Both papers consider top publication venues with paper counts but their work only lists journals while our paper shows that only four of the top 15 publication venues are journals. Thus, considering only journals in their work can be seen as a limitation. Both papers show that the top journal in the area is Expert Systems With Applications, however, it is ranked as 9th overall behind many conference proceed- ings in our list. Decision Support Systems is identified in both, Piryani et al. rank it 4th while we rank it also in 4th in journals but only 13th overall.

Research topics are analyzed in both papers but with dif- ferent approaches. They used a totally manual approach while our larger corpus forced us to use computer based clustering in the first step after which we manually formed higher level research topics. Still, the identified topics have many similarities in terms of data sources (news, reviews, blogs, twitter), data analysis (machine learning, NLP), and application domains (finance, medical, advertisement, trav- eling). We present our classification as a tree like structure with more details in comparison to Piryani et al. As we used computer-based methods we were to independently arrive at the same but more accurate conclusions from larger data set and with lesser effort. We present a more thorough review of the history of senti- ment analysis.

We identify and present the top 20 cited articles of senti- ment analysis. So, what holds in the future of sentiment analysis? We as- sume that in the future the application areas of sentiment analysis will still increase and that the adaption with sentiment analysis techniques will become standardized part of many services and products. We see that the research methods will improve due to advances in natural language processing and machine learning. Additionally, we see a migration of currently mainly text based sentiment analysis methods towards the other affecting comput- ing

methods such as speech, gaze, and neuromarker analysis. How- ever, we are doubtful whether sentiment analysis can achieve a similar 50-fold increase in the number of papers in the next ten years as has occurred during the past ten years (2005–2015). This is based on the fact that this would result in having over 250,000 papers on sentiment analysis published by the year 2025.

Citizens of more established democracies can also benefit from sentiment analysis if it can detect opin- ionated or completely fake news articles that allegedly affected the presidential election of the USA [69]. Similarly, companies are influencing public opinion by generating fake news and organiza- tions to protect their interest against scientific facts like climate change [70]. Our analysis had a branch labeled as the Truth which contains research about detecting fake and spam opinions, see Fig. 14.

## REFERENCES

[1] The pen is mightier than the sword, Wikipedia, 22-Nov-2016. [Online]. Avail- able: https://en.wikipedia.org/w/index.php?title=The_pen_is_mightier_than_ the_sword{&}oldid=750939396. (Accessed 2 December 2016).

[2] B. Liu, Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Handbook of Natural Language Processing, Marcel Dekker, Inc. New York, NY, USA, 2009.

[3] K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519–528.

[4] J.A. Richmond, Spies in ancient Greece, in: Greece and Rome (Second Series), vol. 45, no. 01, 1998, pp. 1–18.

[5] J. Thorley, Athenian Democracy, Psychology Press, 2004.

[6] D.D. Droba, Methods used for measuring public opinion, Am. J. Sociol. (1931) 410–423.

[7] Public opinion quarterly. [Online]. Available: http://poq.oxfordjournals.org. (Accessed 2 December 2016).

[8] R. Stagner, The cross-out technique as a method in public opinion analysis, J. Soc. Psychol. 11 (1) (1940) 79–90.

[9] A.L. Knutson, Japanese opinion surveys: the special need and the special difficulties, Publ. Opin. Q. 9 (3) (1945) 313–319.

[10] A.K. Nassirtoussi, S. Aghabozorgi, T.Y. Wah, D.C.L. Ngo, Text mining for market prediction: A systematic review, Expert Syst. Appl. 41 (16) (2014) 7653–7670.

[11] P. Burnap, et al., Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack, Soc. Netw. Anal. Min. 4 (1) (2014) 1–14.

[12] A. Reyes, P. Rosso, On the difficulty of automatically detecting irony: beyond a simple case of negation, Knowl. Inf. Syst. 40 (3) (2014) 595–614.

[13] A. Hogenboom, B. Heerschop, F. Frasincar, U. Kaymak, F. de Jong, Multi-lingual support for lexicon-based sentiment analysis guided by semantics, Decis. Support Syst. 62 (2014) 43–53.

[14] M.D. Munezero, C.S. Montero, E. Sutinen, J. Pajunen, Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text, IEEE Trans. Affective Comput. 5 (2) (2014) 101–111.

[15] E. Cambria, P. Gastaldo, F. Bisio, R. Zunino, An ELM-based model for affective analogical reasoning, Neurocomputing 149 (2015) 443–455.

[16] V. Garousi, M.V. Mäntylä, Citations research topics and active countries in software engineering: A bibliometrics study, Comput. Sci. Rev. 19 (2016) 56– 77.

[17] H. Du, B. Li, M.A. Brown, G. Mao, R. Rameezdeen, H. Chen, Expanding and shifting trends in carbon market research: a quantitative bibliometric study, J. Cleaner Prod. 103 (2015) 104–111.

[18] M.P. Ferreira, J.C. Santos, M.I.R. de Almeida, N.R. Reis, Mergers & acquisitions research: A bibliometric study of top strategy and international business journals 1980–2010, J. Bus. Res 67 (12) (2014) 2550–2558.

[19] M. Kozak, L. Bornmann, L. Leydesdorff, How have the eastern european coun- tries of the former warsaw pact developed since 1990? A bibliometric study, Scientometrics 102 (2) (2015) 1101–1117.

[20] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (Suppl. 1) (2004) 5228–5235.