The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers

Mika Viking Mäntylä, mika.mantyla@oulu.fi, M3S, ITEE, University of Oulu, Finland
Daniel Graziotin, daniel.graziotin@informatik.uni-stuttgart.de Institute of Software Technology, University of Stuttgart, Germany
Miikka Kuutila, miikka.kuutila@oulu.fi, M3S, ITEE, University of Oulu, Finland

Research in sentiment analysis is increasing at a fast pace making it challenging to keep track of all the activities in the area. We present a computer-assisted literature review and analyze 5,163 papers from Scopus. We find that the roots of sentiment analysis are in studies on public opinion analysis at the start of 20th century, but the outbreak of computer-based sentiment analysis only occurred with the availability of subjective texts in the Web. Consequently, 99% of the papers have been published after 2005. Sentiment analysis papers are scattered to multiple publication venues and the combined number of papers in the top-15 venues only represent 29% of the papers in total. In recent years, sentiment analysis has shifted from analyzing online product reviews to social media texts from Twitter and Facebook. We created a taxonomy of research topics with text mining and qualitative coding. A meaningful future for sentiment analysis could be in ensuring the authenticity of public opinions, and detecting fake news.

General terms: Surveys and overviews
Subject descriptors: Information systems -> Information retrieval -> Retrieval tasks and goals -> Sentiment analysis
Keywords and phrases: sentiment analysis, opinion mining, bibliometric study, text mining, literature review, topic modelling, LDA, qualitative analysis

1 INTRODUCTION

"The pen is mightier than the sword" is a metonymic adage indicating that communication (particularly written language), or in some interpretations, administrative power or advocacy of an independent press, is a more effective tool than direct violence." [Anon 2016b]

Sentiment analysis is a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language [Liu 2009]. Traditionally, sentiment analysis has been about opinion polarity, i.e., whether someone has positive, neutral, or negative opinion towards something [Dave et al. 2003]. The object of sentiment analysis has typically been a product or a service whose review has been made public in the Internet. This might explain why sentiment analysis and opinion mining are often used as synonyms, although, it would be more accurate to view sentiments as emotionally loaded opinions.

The interest on other's opinion is probably almost as old as verbal communication itself. Historically, leaders have had interest on opinions of their subordinates to either prepare for opposition or to increase their popularity. Examples of trying to detect internal dissent can be found already at Ancient Greece's times [Richmond 1998]. 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 [Thorley 2004]. Efforts in capturing public opinion by quantifying and measuring it from questionnaires have appeared in the first decades of twentieth century [Droba 1931], while a scientific journal on public opinion was established in 1937 [Anon n.d.].

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, over 5,000 papers of this topic have been published and, more interestingly, 99% of the papers have appeared after 2005 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. Figure 1 shows the increase in searches made with a search string “sentiment analysis” in Google search engine.

![Google Trends](www.google.com/trends) data showing the relative popularity of search strings “sentiment analysis” and “customer feedback”.

We observed that the first academic studies measuring public opinions are during and after WWII and their motivation is highly political in nature [Stagner 1940; Knutson 1945]. The outbreak of modern sentiment analysis happened only in mid-2000’s, and it focused on the product reviews available on the Web, e.g., [Dave et al. 2003]. Since then, the use of sentiment analysis has reached numerous other areas such as the prediction of financial markets [Nassirtoussi et al. 2014] and reactions to terrorist attacks [Burnap et al. 2014]. 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 [Reyes and Rosso 2014] and multi-lingual support [Hogenboom et al. 2014]. Furthermore, with respect to emotions\(^1\), efforts are advancing from simple polarity detection to more complex nuances of emotions and differentiating negative emotions such as anger and grief [Cambria et al. 2015].

The area of sentiment analysis has become so large that any individual researcher would face several issues when keeping track of all the activities in the area and the information overload. An academic literature review can only focus on one particular area of sentiment analysis as it typically includes between 10 to 100 studies, e.g., a recent systematic review of the prediction of financial markets with sentiment analysis reviewed 24 papers [Nassirtoussi et al. 2014]. 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., [Garousi and Mäntylä 2016; Du et al. 2015; Ferreira et al. 2014; Kozak et al. 2015]:

\(^{1}\) See [Munezero et al. 2014] for the differences between feelings, opinions, sentiment, and emotions in the context of natural language processing.
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 sentiment 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 sentiment 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 [Griffiths and Steyvers 2004] titled “Finding scientific topics”. We support the automated text clustering with manual qualitative analysis 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 analysis can demonstrate the most interesting and impactful work in this area. Using citation counts has been previously done for example [Van Noorden et al. 2014; Garousi and Fernandes 2016] that studied the 100 most cited papers published in Nature and in Software Engineering.

The present paper presents the first bibliometric study of sentiment analysis and makes the following six contributions. First, we show how attempts to understand public opinion at the start of 20th century through questionnaires 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 100-fold growth in ten years between 2005 and 2015. 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. Newcomers have now an indication of venues that welcome the topic. Fourth, using word clouds we show that most prominent research activities have been related to the classification of online reviews and social media texts. We demonstrate using how the research activities have shifted from analyzing online product reviews, which were more common prior to 2014, to social media texts that dominate the studies in the years 2014-2016. Fifth, we demonstrate the value of mixing Latent Dirichlet Allocation (LDA, see section 2.4) topic modelling with qualitative coding for clustering all sentiment analysis papers and constructing a comprehensive classification of articles. This technique allows a deeper understanding of different research topics of an area, sentiment analysis in our case, by providing a tree like semantic structure. Sixth, we review the top-cited literature reviews and original works which demonstrates the hallmarks of sentiment analysis research.

This paper is structured as follows. Section 2 explains our research methods. Beyond that our structure deviates from conventional 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.

2 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. According to its developer Elsevier, “Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings.” Thus, it 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., [Harzing and Alakangas 2016; Wildgaard 2015; Mongeon and Paul-Hus 2016; S. Adriaanse and Rensleigh 2013]. We report in section 4 how Scopus was found to be the most reliable [S. Adriaanse and Rensleigh 2013] and comprehensive [Monegon and Paul-Hus 2016] 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 2,000 papers at a time. This enables further 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.

We used four search strings to search the literature, which we detail in Table 1. Two of our search string originated from the highly influential literature review on sentiment analysis by Pang and Lee [Pang and Lee 2008] that presented strings “Opinion mining” and “Sentiment analysis” as synonyms or near synonyms. Additionally, we noticed that authors were mixing our two original terms also using terms “Sentiment mining” and “Opinion analysis”. Thus, we included those search strings in our query. All of our search strings were targeted to the paper title, paper abstract, and paper keywords. A match any of our search string in either title, abstract or key words resulted in the inclusion of a paper.

Our search produced 5162 hits. Table 1 shows the number of hits produced by each search string individually in the diagonal. Overlap between search terms is visible in other cells, for example the junction of search strings between “Sentiment Analysis” and “Opinion mining” shows that 830 documents contained both search terms. The number in the parenthesis is the normalized Google similarity distance [Cilibrasi and Vitanyi 2007] (later relabeled to normalized web distance [Vitányi et al. 2010]) by Cilibrasi and Vitanyi for measuring the similarity of the results of two search string with scale from 0 (no similarity) to 1 (perfect similarity) from any search engine. From the Google / web similarity distance we can see that although strings “sentiment analysis” and “opinion mining” are generally accepted as synonyms they actually have the lowest similarity among our search terms suggesting that our inclusion of the two additional search terms is a well-motivated choice. Although, the addition of the two less common search string “opinion analysis” and “sentiment mining” produced only 157 additional papers combined.

Table 1 Search strings, number of search hits, overlap between searches measured with the number of documents including both strings (the normalized web distance [Vitányi et al. 2010], aka Google similarity distance [Cilibrasi and Vitanyi 2007]).

|                | “Sentiment analysis” | “Opinion mining” | “Opinion analysis” | “Sentiment mining” |
|----------------|----------------------|------------------|--------------------|--------------------|
| “Sentiment analysis” | 3894                | 830 (0.15)       | 66 (0.34)          | 23 (0.38)          |
| “Opinion mining”    | 830 (0.15)          | 1767             | 49 (0.30)          | 19 (0.34)          |
| “Opinion analysis”  | 66 (0.34)           | 49 (0.30)        | 371                | 2 (0.39)           |
| “Sentiment mining”  | 23 (0.38)           | 19 (0.34)        | 2 (0.39)           | 81                 |

2 https://www.elsevier.com/solutions/scopus
2.2 Quantitative analysis of paper and citation counts

We performed quantitative analysis by plotting histograms of the paper and citation counts.

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. 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 [Team and others 2013]. 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 1348 different sources.

2.4 Word clouds and Topic modelling of research topics

Due to our high volume of articles (over 5,000) qualitative manual analysis of all the papers would have exceeded our resources. Therefore, we used the text mining and clustering with R language [Team and others 2013]. We did basic word clouds and dissimilarity word clouds using R package “wordcloud” [Fellows 2012].

We also used topic modelling to cluster our papers. In more detail, we used Latent Dirichlet Allocation (LDA) from R package “topicmodels” [Hornik and Grün 2011]. The origins of the LDA approach for scientific topic detection lie in the influential paper by Griffiths and Steyvers [Griffiths and Steyvers 2004] modelling scientific topics that was published in PNAS. Parts of the R computations are from Ponweiser [Ponweiser 2012] who replicated the study by Griffiths and Steyvers. Previously, we followed this approach when we analyzed software engineering literature [Garousi and Mäntylä 2016].

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 occasionally contains publisher names, e.g., IEEE or ACM. Second, we created and manipulated our corpus with R-package “tm” [Feinerer 2015]. 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, stemmed the document, 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 [Ponweiser 2012]. With the document-term-matrix, we performed term-frequency inverse-document-frequency (tf-idf) computations. Tf-idf improves raw term frequency computations by also considering the idf of each term that can be characterized as the amount of information each word carries. We used this information to remove the words having less than median tf-idf value from the document-term-matrix. An example of this step is show in [Hornik and Grün 2011]. In our experience, such tf-idf pre-processing reduces the computing time needed for creating LDA topics while improving the coherence the topics as less relevant words are not considered in the clustering. We used hyper-parameter settings with the default values suggested by Griffiths and Steyvers [Griffiths and Steyvers 2004], 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 to contain multiple words (beta). As in [Griffiths and Steyvers 2004] we have fixed small beta that should lead to a quite high number of fine grained

---

3 Articles published in journals are lately becoming as common as those in conferences, although it is still not the case yet [Anon n.d.].
topics. Such specific topics were then furthered analyzed with qualitative coding, see next section. Finally, similar to [Griffiths and Steyvers 2004] 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 modelling from the previous step suggested that 98 topics would be optimal, we opted to create 98 topics. Instead of reducing the number of topics, which would improve readability and understandability, we opted for a qualitative coding strategy. That is, we categorized the clusters using qualitative coding as presented in many qualitative analysis handbooks, e.g., [Patton 1990]. Our qualitative analyses were aided further by the investigation of the papers assigned to cluster and by the clustering of the LDA topics with hierarchical clustering of the topics and subsequent dendrogram visualizations. However, dendrograms were used as input only for the first author who performed the majority of the qualitative coding. First author was aided by the second author who did part of the qualitative coding and validated the qualitative coding performed by the first author. Finally, the first and second author jointly developed the qualitative coding strategy. During the qualitative coding phase the mind mapping tool Mindmup.com was used to enable real-time collaboration between the first and second author. Figure 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 modelling, 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 from Scopus. We also separated two groups of contributions: original research and literature reviews. We analyzed the top five papers based on citations, and then we evaluated the top cited paper of each of the past five years (2011-2015). These steps should ensure that we considered the most significant papers measured with the citation counts of all time and making sure that the recent papers having relatively but not absolutely high citation counts are also analyzed. We used this approach for both original research and literature reviews.

3 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 Figure 3. The field has also changed in terms of content over the years. Prior to availability of
massive amount of text and opinions online, studies mainly relied on survey-based methods and were focused on public or expert opinions rather than users or customers' opinions. The first paper that matched our search was published in 1940, and it was titled “The Cross-Out Technique as a Method in Public Opinion Analysis” [Stagner 1940]. In 1945 and 1947 three papers appeared that addressed measuring public opinions in post WWII countries that had suffered during the war (Japan, Italy, and Czechoslovakia) and they were all published in the journal Public Opinion Quarterly [Knutson 1945; Fegiz 1947; ADAMEC and Viden 1947].

In mid 90s, computer-based systems started to appear. The computing revolution started to be reflective in research, too. For example, a paper titled “Elicitation, Assessment, and Pooling of Expert Judgments Using Possibility Theory” was published in 1995, and it used a computer system for expert opinion analysis in the domain industrial safety that allowed for example a pooling of opinions [Sandri et al. 1995]. Still, the outbreak of modern sentiment analysis was nearly ten years away.

We can pinpoint the start of modern sentiment analysis to the year 2003 when two influential papers were published. The first one was titled “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews” [Dave et al. 2003] that was published in the WWW conference and has since been cited nearly 600 times according to Scopus. The paper positioned sentiment analysis in the context of product reviews available in the Web. The authors argued that automated analysis of such reviews are needed due to the high volumes of such reviews. The second article was presented in the International Conference on Data Mining. The paper was titled “Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques” [Yi et al. 2003]. The paper presented a sentiment analyzer that, according to the authors, unlike prior works was able to detect a sentiment with respect to some object, e.g., a digital camera, rather than giving sentimental scores to the entire text. In 2005 more influential papers appeared, with four of those papers currently having over 100 citations in Scopus, namely: “Recognizing contextual polarity in phrase-level sentiment analysis” [Wilson et al. 2005], “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales” [Pang and Lee 2005], “Using appraisal groups for sentiment analysis” [Whitelaw et al. 2005], and “Determining the semantic orientation of terms through gloss classification” [Esuli and Sebastiani 2005].

For years, the number of related published papers was extremely low. Figure 3 shows the published papers per year starting from the year 2000. By the year 2000 only 21 papers in total had been published, and by the year 2005 the number was 55. We can observe a momentum building up in 2005. By the end of 2010 the number of papers was 801 and, finally, by the end of 2015 the number of papers reached the value of 4,845. Our total pool of papers also includes early 2016 papers and the total count is at 5,163 papers. Our analysis shows that, although an understanding of opinions has always been important, what made opinion analysis a trending research topic was the possibility to automatically collect and analyze large corpuses of opinions with the help of text mining tools.
3.2 Publication Venues (RQ3)

The top 15 publication venues in terms of published sentiment analysis articles are in Table 2. The table provides the publication name, its type (P for proceedings, B for books, and J for journals), the number of sentiment analysis articles and their percentage with respect to the entire sentiment analysis dataset, and the ratio of the published sentiment analysis papers versus all published papers for the venue itself in 2015. The last column is bounded to the most recent year of complete data (2015) for publication because of Scopus limitations when retrieving data.

Despite of all our data cleanup efforts, we found out that Scopus enlists certain conference entries with the name of the proceedings series instead of the name of the conference itself. For example, the first four entries in Table 2 are proceedings series, which contain several different conferences, while the fifth entry represents the proceedings of a single conference. We first observe in Table 2 that most venues are conference proceedings (13 entries), while the remaining are journals (4 entries).

The most populated venue is Springer’s Lecture Notes in Computer Science (LNCS) with its subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics. LNCS publishes conference proceedings mainly, although the series disseminates nine transactions and other monographs such as Festschrifts, tutorials, and state of the art surveys. With 589 entries, LNCS has published almost four times sentiment analysis papers than the second most populated entry and it hosts 12.2% of all sentiment analysis papers indexed by Scopus. However, less than 1% of the 2015 LNCS papers were about sentiment analysis.

CEUR Workshop Proceedings (CEUR-WS) is a long-standing open access publication service in cooperation with Aachen University. CEUR-WS hosts exclusively proceedings of computer science workshops, which are free to publish and access for authors and readers, respectively. With 153 entries for CEUR-WS, we note that the two top venues (LNCS and CEUR-WS) account for almost half (49%) of the sentiment analysis papers of our top 15 rankings in Table 2. This indicates that the distribution of papers with respect to the venues is skewed towards the top two venues. However, the first two venues account for 15.4% of all papers on sentiment analysis that are indexed by Scopus. Less than 1% of the 2015 CEUR-WS papers were about sentiment analysis.

The ACM International Conference Proceedings Series (ICPS) is an ACM program for cost-effective publication of conference and workshop proceedings hosted into ACM digital library. Conference organizers submit the proceedings to ACM, which evaluates them for inclusion into ICPS. We note that ACM main conferences, such as the International Conference on Software Engineering, the ACM CHI Conference on Human Factors in Computing Systems, and SIGGRAPH conference are not published with ICPS. The main conferences in computer science have their own
proceedings, which are published by ACM, IEEE, or both. Less than 1% of the 2015 ICPS papers were about sentiment analysis.

Communications in Computer and Information Science (CCIS) is a Springer series of books devoted to the publication of conference proceedings of computer science conferences. As a way to differentiate from LNCS, CCIS declares to be interested mostly in topics related to theory of computing and information and communications science and technology. We observed that a difference in terms of topics does not seem to exist in the case of sentiment analysis. In 2015, 9% of the CCIS papers were sentiment analysis articles.

| Ranking | Name                                                                 | Type | Number of sentiment analysis papers | Percentage of sentiment analysis papers in total | Proportional share of sentiment analysis papers (Number of sentiment analysis papers in 2015 / Number of papers in total in 2015) |
|---------|----------------------------------------------------------------------|------|-------------------------------------|-----------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| 1       | Lecture Notes In Computer Science (including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics) | P, B | 589                                 | 12.2%                                         | 0.67% (145/21611)                                                                                                              |
| 2       | CEUR Workshop                                                      | P    | 153                                 | 3.2%                                          | 0.79% (29/3661)                                                                                                                |
| 3       | ACM International Conference Proceeding Series                     | P    | 103                                 | 2.1%                                          | 0.59% (19/3201)                                                                                                                |
| 4       | Communications In Computer And Information Science                 | P    | 95                                  | 2.0%                                          | 9.45% (26/275)                                                                                                                 |
| 5       | International Conference On Information And Knowledge Management   | P    | 83                                  | 1.7%                                          | 1.63% (4/246)                                                                                                                  |
| 6       | Advances In Intelligent Systems And Computing                      | P, B | 59                                  | 1.2%                                          | 0.92% (30/3245)                                                                                                                |
| 7       | EMNLP Conference On Empirical Methods In Natural Language Processing | P    | 59                                  | 1.2%                                          | 7.30% (23/315)                                                                                                                 |
| 8       | Expert Systems With Applications                                   | J    | 50                                  | 1.0%                                          | 1.29% (10/778)                                                                                                                 |
| 9       | Annual Meeting Of The Association For Computational Linguistics    | P    | 48                                  | 1.0%                                          | 3.09% (11/356)                                                                                                                 |
| 10      | International Conference On World Wide Web                        | P    | 45                                  | 0.9%                                          | 1.49% (2/134)                                                                                                                  |
| 11      | International Journal Of Applied Engineering Research              | J    | 41                                  | 0.8%                                          | 0.53% (3/576)                                                                                                                  |
| 12      | Procesamiento De Lenguaje Natural                                  | J    | 40                                  | 0.8%                                          | 28.57% (12/42)                                                                                                                 |
| 13      | International Conference Recent Advances In Natural Language Processing | P  | 36                                  | 0.7%                                          | 11.46% (11/96)                                                                                                                 |
| 14      | Decision Support Systems                                           | J    | 35                                  | 0.7%                                          | 2.31% (3/130)                                                                                                                  |
| 15      | IEEE International Conference On Data Mining                       | P    | 30                                  | 0.6%                                          | 2.05% (3/146)                                                                                                                  |
| 16      | ACM SIGKDD International Conference On Knowledge Discovery And Data Mining | P  | 27                                  | 0.6%                                          | 0.41% (1/241)                                                                                                                  |
| 17      | Procedia Computer Science                                          | P    | 27                                  | 0.6%                                          | 0.53% (18/3376)                                                                                                                |

Advances in Intelligent Systems and Computing (AISC), like LNCS and CCIS, is a recent series of Springer books that specialize in intelligent systems and intelligent computing. AISC contains mainly proceedings and few edited books.

The Conference on Empirical Methods in Natural Language Processing (EMNLP) is an international conference on natural language processing. EMNLP is organized by the SIGDAT special interest group for linguistic data, which is part of the Association for Computational Linguistics (ACL). The conference calls for papers related to SIGDAT community interests, and
sentiment analysis is a solicited topic. Similarly to CCIS, ENML had high share (7%) of published papers as sentiment analysis articles in 2015.

Expert Systems with Applications is an Elsevier journal, which publishes papers related to expert and intelligent systems applied in industrial settings (including education and public sector). The journal deals with any kind of expert and intelligent systems applied in any industrial area. Data and text-mining is a solicited topic.

The Annual Meeting of the Association for Computational Linguistics is an ACL Conference, which calls for papers about intelligent systems and their interactions with humans using natural language, computational and linguistic properties of language, speech recognition and translation, and information retrieval and extraction. The conference calls for contributions on sentiment analysis and opinion mining. The proceedings are published by ACL.

The International Conference on World Wide Web (WWW) provides a forum for discussion in regard to the standardization of World Wide Web associated technologies and their impact on society and culture. The conference calls for papers on Web mining, human factors, social network analysis, and content analysis. Several sentiment analysis papers are presented at the conference each year. The WWW proceedings are published by ACM.

The International Journal of Applied Engineering Research publishes articles in all engineering areas. The journal is published by Research India Publication publisher. We notice a high number of articles published per year and that the journal charges both for publishing and accessing the articles, which could be read for free only for one year after publication. The journal and its publisher were reported to be of predatory publishing nature [Anon 2014].

Procesamiento del Lenguaje Natural is an open access journal published by the Sociedad Española para el Procesamiento del Lenguaje Natural. The journal publishes articles on natural language processing. The journal welcomes articles in English as well in Spanish, although a bilingual abstract is required. The journal had the highest ratio of sentiment analysis papers in 2015 with a value of 28%.

The International Conference on Recent Advances in Natural Language Processing (RANLP) is an international conference organized by the Association for Computational Linguistics (ACL). Opinion mining and sentiment analysis are a welcomed topic for submission. 11% of the RANLP papers in 2015 was of sentiment analysis nature.

Decision Support Systems is an Elsevier journal. The journal welcomes submissions that are relevant to theoretical and technical issues in the support of decision support systems foundations, functionality, interfaces, implementation, and evaluation. While the journal does not call for data mining papers, papers related to sentiment analysis and opinion mining in the context of decision support systems are welcomed.

The International Conference on Data Mining (ICDM) is an IEEE conference which covers all aspects of data mining, including algorithms, software and systems, and applications. Opinion mining and sentiment analysis are naturally belonging to data mining, thus papers on those topics are solicited.

The SIGKDD International Conference on Knowledge Discovery and Data Mining is an ACM conference which, similarly to ICDM, welcomes contributions from data mining researchers and practitioners. The conference is organized by ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD).

Procedia Computer Science is Elsevier’s open access proceedings publishing system for computer science events. The series publishes proceedings of conference related to computer science and a small selection of workshops.

3.3 Citation patterns (RQ2)

As the number of papers grows so does the number of citations. We analyze the citation patterns as they are a backbone of science, which is well illustrated by the famous quote by Newton (“If I have seen further, it is by standing on the shoulders of giants”), which illustrates that only good prior works make research advances possible. We are also aware of the problems with citations
counts and that citations are not always used as part of building on top of previous work [Werner 2015]. Nevertheless, we consider them as useful for showing the growth in interest on sentiment analysis.

Figure 4 shows the cumulative total number of citations in comparison to the number of papers. The citations are counted for the papers that are published in the given year. Figure 5 shows the annual number of papers and the number of citations normalized for time. Since the accumulation of citations takes time, a normalization allows us to have more reliable data of the recent years. With respect to the year 2015 we assume that our time correction is insufficient since many of the years 2016 are still missing. Moreover, there is a delay between the publication date and the Scopus inclusion date. The fluctuation in the citation counts between years in Figure 5 is mainly due to single highly cited papers. For example, for the year 2008 one paper [Pang and Lee 2008] has 2169 citations out of the 8079 total citations made for papers published on that year.

![Figure 4 Cumulative number of papers and citations](image1.png)

![Figure 5 Annual number of papers and citations (normalized for time)](image2.png)
Figure 6 shows the number of citations given to each paper and the year when the paper was published. In the figure, we have used log scale transformation and we removed papers with zero citations to improve its readability. The figure also illustrates the recent growth this field has experienced in the past ten years.

Figure 7 shows a histogram of paper citations. We find that 57% of the papers (2962/5163) have not been cited at all, 12% papers (619/5163) had one citation while 31% (1582/5163) had two or more citations. For comparison purposes, we report that our recent study of the software engineering literature showed that 43% of papers had no citations, 14% had one citations, and 43% had two or more citations [Garousi and Mäntylä 2016]. In what follows, we compare the current results with relevant results of our previous study.
The difference in papers with no citations (sentiment analysis 57% vs. software engineering 43%) is simply due to the age of each research topic. Median publication year for papers in our previous software engineering data set was 2008 while in this paper the median is 2013. To adjust for age bias, we computed citation counts for papers aged between five and nine years for both data sets. This showed that in that age group software engineering has more papers with no citations (sentiment analysis 36% vs. software engineering 43%). Table 3 show the percentage of not cited papers with respect to paper age in years.

| Paper age (years) | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Sentiment analysis| 58% | 43% | 43% | 38% | 36% | 31% | 40% | 31% | 37% |
| Software engineering| 71% | 50% | 50% | 47% | 47% | 42% | 41% | 35% | 32% |

Table 4 shows the top-5 cited papers from each area with respective citations and publication years. Analysis that is more detailed showed how all of the top 30 cited papers in software engineering were published prior to 2005 while in the present study only two of the top 30 cited papers were published prior to 2005. It is interesting that the most cited paper in sentiment analysis has more citations than any paper in software engineering despite the only recent emergency of this research topic. This comparison further highlights the recency of sentiment analysis and opinion mining as a research topic.

| Rank | Title                                                                 | Year | Citations | Title and citation                                      | Year | Citations |
|------|-----------------------------------------------------------------------|------|-----------|----------------------------------------------------------|------|-----------|
| 1    | A metrics suite for object-oriented design [Chidamber]                | 1994 | 1,817     | Opinion mining and sentiment analysis [Pang and Lee 2008] | 2008 | 2169      |
3.4 Areas of research – Word clouds (RQ4)

Figure 8 shows the word cloud of all the paper titles. Figure 8 (a) shows how the key concepts of sentiment analysis are “social”, “online”, “classification”, “reviews”, “media” and “product”. We can also find more detailed topics and domains such as “movie”, “news”, “political”, “stock” and “financial” that refer to the business domains. Also, different languages can be seen with “Chinese” and “Arabic” being the most common. Similarly, different analysis methods are visible in word like “neural”, “fuzzy”, and “supervised”.

For an established research topic, one can show word clouds of different decades to highlight the differences. Due to the recency of our topic, we were not able to do that as the early years had very limited numbers of papers. Therefore, we divided the data to roughly two equal parts, that is the years 2014-2016 and the years 2013 and prior. Figure 8 (b) shows the differences between the more recent (2014-2016) and the early years (2013 and prior). Figure 8 (b) shows the word comparison cloud (the words on top in green are more common in 2014-2016 while the words on the bottom in orange are more common prior to 2014). Both figures are plotted with R Package ‘wordcloud’. In the word comparison cloud, the size of each word is determined by its deviation from the overall occurrence average. For example, words like “classification” and “text” disappear completely meaning that they are roughly equally common in both periods. We can observe that sentiment analysis of “Chinese” language was more common in the early years while recent research activities highlight “Arabic” and “Indian” languages. Early years focused on “customers” “online” “product” reviews. While later years seem to focus on reviews made in “social” “media” and particularly in “Twitter” and “Facebook” (top left in Figure 8b). We can also observe that early years had focus on “web” and “online” while later years focus more on “mobile”. There has also been a shift in who produces the reviews. Early years referred the producers as “customer” or “consumer” while later years highlight the word “users”. We can see the word “big” coming from the big data concept. We note here that we have removed common words like results, data, sentiment, analysis, mining, opinion etc. to make the word clouds more meaningful.
3.5 Areas of Research - Topic modelling and human based classification of topics (RQ3)

In addition to word clouds, we performed text clustering with topic modelling to study the detailed areas of research in sentiment analysis. The word clouds provide a higher level view from top down as they focus on the most frequent words. Topic modelling works from bottom up as it tries to come up with a set of topics that could generate the given documents. Details of our method are given in Section 2.

Our topic modelling computation showed that 98 was the optimal number of topics when measured with log-likelihood. We further analyzed these 98 topics with qualitative coding to form higher level groups. In comparison to our previous work [Garousi and Mäntylä 2016], where we also utilized the same topic modelling approach, we noticed during the manual classification that topics in the area of sentiment analysis tended to be more incoherent in terms of the most probable words for each topic. We think this incoherence of the topics was a reflection of the structure of the papers, i.e., the papers had a) a dataset, b) a data/sentiment analysis method, c) the analysis were created for a particular language, and d) a research goal.

In comparison to our past work, we were less satisfied with the results of the topic modelling in here. For example, we had a topic with the top words of “ensemble”, “risk”, “forest”, “schema”, and “german”. Here “ensemble” and “forest” describe the data analysis method random forest that is an ensemble learning method. The word “risk” relates to the research goal that is to use sentiment analysis to study risk, for example the various risk in business [Güven et al. 2014] or medicine [LongDeng et al. n.d.]. Word “schema” is general and not very helpful in this case. Finally, the word “german” refers to the language that was used to perform the analysis in some papers, e.g., [Winkler et al. 2015]. To solve this issue of incoherence during manual qualitative coding, see Section 2.5, we coded such topics under multiple classes. Next, we present our classification from the coarsest grained groups to the fine-grained groups.

We formed three groups in the highest level, namely Data, Data Analysis, and Goal. Figure 9 shows our top level classification tree while Figures 10 to 14 show the details of each branch.
3.5.1 Data

The Data area collects all topics related to the source of information that is utilized. The Data class contained 10 topics directly under it where the most probable words about data were like: "feed", "music", "photo", "image", "video", "microblogging", "email", "chat", mobile, smartphone, newspaper, and chat. We were able to distinguish just one sub class for Data named Different Languages.

- Different Languages contained topics that described the language in which the data was expressed, e.g., spanish, italian, russian, arabic, german, etc. Interestingly, English was not found to be a high probable word. The reason is likely that the English language is implicitly granted in research articles.
3.5.2 Data Analysis

The Data Analysis class had two sub-classes that contained the topics related to sentiment analysis Tools and Techniques for analyzing the data.

Tools class contained nine tool-related topics for sentiment analysis, where the most probable words were like: translator, crawler, software, sentistrength, scanner, streamworkflow.

Techniques was further divided into three sub-classes.

- **Machine Learning** had 8 topics where the most probable words were words such as: fuzzy, logic, bayesian, neural, deep, lda, crf, evolutionary and graph based methods, ensemble learning.

- **Natural Language Processing (NLP)** had 16 topics in which the most probable words were like: wordnet, composite, bigram, propagation, adjective, noun, verb, bias, morphology, ontology, n-gram, term weighting scheme.

- **Sentiment Analysis** had 8 topics, and the most probable words were like: segment, collaboration, word of mouth, cold, emoticon, tendency, symbol, thesaurus, senticnet, common sense.
Figure 11 Tool branch of Data Analysis branch of our classification tree
Figure 12 Techniques branch of Data Analysis branch of our classification tree

3.5.3 Goal

The goal of the paper tells what problem each paper tries to address. We identified two classes based on the types of goals: Application Domain oriented and Human and Behavior oriented.
Application domain oriented goals focused on the areas what we could call the “business” domain of sentiment analysis. It was divided further to six classes.

- **Travel** had four topics where the most probable words where words like: “city”, “tourism”, “tip”, “hotel”, “travel”, “restaurant”.
- **Entertainment** had five topics where the most probable words where words like: “program”, “broadcast”, “shop”, “buyer”, “book”, “box”, “office”, “app”, “game”, “player”
- **Finance and corporate** had six topics where the most probable words where words like: “finance”, “stock”, “risk”, “revenue”, “brand”, “firm”.
- **Medical** had three topics where the most probable words where words like: “disease”, “patient”, “health”, “drug”, “cancer”, “autism”, “stress”
- **Society** had three topics where the most probable words where words like: “security”, “crime”, “election”, “poll”, “policy”, “citizen”
- **Other** had six topics that each specified other application domains, such as citation analysis, advertisement, education, traffic, blogs. The most probable words were: “citation”, “paper”, “cite”, “advertisement”, “university”, “student”, “learner”, “style-blogger”, “traffic”.

Figure 13 Application Domain branch of Goal branch of our classification tree
Human and Behavior-oriented goal focused on the areas that could be used in several application domains. Yet, we think the classes here still are research goal oriented rather than data or data analysis methods.

- **Expertise and Influence** had five topics with the most probable and describing words like: “expert”, “leader”, “influential”, “reputation”, “recommendation”, “answer”
- **Interaction** had four topics with the most probable and describing words like: “conversation”, “engaging”; “discourse”, “crowd”, “rhetoric”, “audience”, “conference”, “reaction”, “debate”, “stance”.
- **Organization** had two topics with the most probable and describing words like: “virtual”, “team”, “hierarchy”, “self-organizing”
- **Truth** had six topics with the most probable and describing words like: “clue”, “irony”, “sarcasm”, “sense”, “disambiguate”, “implicit”; “feature-opinion”, “statement”, “judgement”, “inconsistent”, “truth”, “fake”, “spam”, “deception”
- **Language** had seven topics with the most probable and describing words like: “cross-domain”, “domain-specific”, “domain-independent”, “genre”, “slang”, “dialect”, “translation”, “multi-lingual”
- **Other** had seven topics that each specified other areas, such as culture, behavior, mood, cognition, trust, demographics, appraisal theory, geo location. The most probable words were: “behavior”, “mood”, “manipulation”, “trust”, “cognition”, “demographics”, “gender”, “geography”, “spatial”, “appraisal”.
3.6 Highest cited papers (RQ5)

Here we present the top cited papers for both original research and literature reviews in subsections 3.6.1 and 3.6.2.

3.6.1 Original research

This section provides short overviews of original research papers among the top five cited all time and the top cited paper from past five years individually (2011-2015). Additionally, we present a mapping on how the papers match the research topics presented as word clouds in Section 3.4 and to research topics in Sections 3.5.1-3.5.3. Table 5 lists the 5 top-cited of all time and top-3 cited for each year between 2011 and 2015.

The article by Dave, Lawrence and Pennock [Dave et al. 2003] is overall the most cited original research paper in our data set. In this paper, an approach to opinion mining is presented, where opinions of products are mined from the Web and analyzed using Natural Language Processing (NLP) techniques. Opinions are automatically divided into positive and negative
sentiments, while feature opinions and context are considered as well. The paper used several data analysis techniques that are also available in our classification in Section 3.5.2, such as Bayesian and n-gram. Our classification in Section 3.5.3 had three groups, namely Entertainment, Travel and Medical that can be considered as products and partially match this paper. Our text clustering or the following manual classification did not produce a class for “product reviews” or a class for “web” as both terms are very general. Obviously, due to this generality both “product reviews” and “web” were clearly present in the word clouds in Section 3.4 in Figure 8. The paper was published in 2003 and since then the topic of sentiment analysis has expanded and become more specialized.

| Year | Title                                                                 | Citation Count | Authors                                                                 |
|------|----------------------------------------------------------------------|----------------|-------------------------------------------------------------------------|
| 2003 | Mining the peanut gallery: Opinion extraction and semantic classification of product reviews | 576 | Dave et al. 2003 |
|      | Recognizing contextual polarity in phrase-level sentiment analysis | 561 | Wilson et al. 2005 |
|      | Lexicon-based methods for sentiment analysis                           | 370 | Taboada et al. 2011 |
|      | Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales | 303 | Pang and Lee 2005 |
| 2008 | A holistic lexicon-based approach to opinion mining                   | 266 | Ding et al. 2008 |
| 2011 | Lexicon-based methods for sentiment analysis                           | 370 | Taboada et al. 2011 |
| 2012 | An ELM-based model for affective analogical reasoning                  | 15 | Cambria et al. 2015 |
| 2014 | Convolutional neural networks for sentence classification              | 40 | Kim 2014 |
| 2013 | Estimating the helpfulness and economic impact of product reviews     | 191 | Ghose and Ipeirotis 2011 |
|      | Opinion word expansion and target extraction through double propagation | 163 | Qiu et al. 2011 |

Wilson, Wiebe and Hoffman [Wilson et al. 2005] present phrase level sentiment analysis approach using machine learning algorithm, which mainly judges whether an expression is polar or neutral and the polarity of the expression. Word “polarity” is also seen in wordcloud Figure 8b in the section of words more common before 2014. The lexicon used together with machine learning algorithm has subjectivity values for words, which take into account the context dependent nature of analyzed sentiments. This approach can be used both with larger documents and single phrases, and in the experiments of the authors the approach outperformed the human baseline. From the identified data sources in Section 3.5.1, this paper uses news articles from MPQA corpora [Anon n.d.] as data. Machine learning is a match as a data analysis technique (Section 3.5.2), while the Human and Behavior-oriented goals of this study are mostly related to topics under class language.

Taboada et al. [Taboada et al. 2011] enhance the existing analysis tool “Semantic Orientation CALculator (SO-CAL)” by supplementing rules such as intensification and negation to a pre-existing lexicon based method. This new approach is then validated, evaluated and compared to several existing sentiment analysis tools and approaches, in an experiment where various data sets are analyzed with all the approaches in order to draw conclusions. The authors used several data sets to evaluate the tool: product reviews, social media comments, news articles and headlines. In Section 3.5.1, we did not identify topics for product reviews or social media comments as data source, however one could consider microblogs and chat to include social media comments. SO-CAL uses Natural Language Processing techniques identified in Section 3.5.2. We see the goal of this paper as
validation and evaluation of the SO-CAL tool, and with multiple data sets used, from Section 3.5.3 there are several matches to topics with main words such as lexicon under Language class and Human and Behavior oriented goals.

Pang and Lee [Pang and Lee 2005] introduce a problem rating-inference problem where sentiment is classified in multi-point scale rather than positive, neutral and negative sentiments. They proceed to solve it using three different approaches utilizing support vector machines. Movie reviews are used as data sources to evaluate the approaches, which do not correspond to the data sources identified in Section 3.5.1, but we had Entertainment under goals in Section 3.5.3, which is a good match for the paper. The paper used support vector machine based algorithms, which corresponds well to the machine learning class identified in Section 3.5.2, however support vector machines (SVM) also do appear as their own topic under machine learning.

Ding, Liu and Yu [Ding et al. 2008] present a lexicon based approach for feature based analysis of reviews. The algorithm used for analysis is presented in pseudo-code. Various rules that enhance the algorithm, such as negation rule which can reverse the sentiment (e.g., word “not” in front of word “good” changes the meaning of word good from positive to negative) are introduced in detail. Other rules concern conjunctions, synonyms and antonyms, sentences with word “but” and some rules are related to conditionals that are used for context dependent opinion words. This approach is then implemented with programming language C++ and evaluated together with two other methods by analyzing customer review data. The implemented approach by the authors, “Opinion Observer”, performs better on average when analyzing reviews of various products than the two other methods called “FBS” and “OPINE”. Product reviews are used as a data set, for which we have no clear match in the Section 3.5.1. However, some of the product reviews are for mobile phones. The approach authors presented and evaluated corresponds to topic, which has the stemmed word “featurebas” as its most probable word, it is under class Natural Language Processing introduced in Section 3.5.2. Besides evaluating and validating the developed approach, the paper corresponds Human and Behavior goals with topics under classes Language and Truth.

Among the papers between 2011 and 2015 we find for the year 2015 that the most cited paper was by Cambria, Gastaldo, Bisio and Zunino [Cambria et al. 2015] who present an extreme learning machine (ELM) approach, which aims to for deeper understanding of language semantics. Instead of just classifying sentiments only on positive-negative axis, the approach can give strength of different emotions such as grief, anger, admiration and joy in the analyzed data. The authors experiment and discuss validity of their approach, by evaluating different parts of the analysis framework with two different benchmarking data sets, and lastly by integrating the approach to an existing opinion mining engine to analyze opinions from a United Kingdom’s national health service feedback system and comparing the results to other available approaches. The results indicate faster learning speed compared to existing machine learning algorithms, as well as better generalizability of predictors over prior approaches. For evaluation, a data set of feedback to the United Kingdom’s national health service is used. The research data corresponds poorly to the data class we identified in section 3.5.1, but is a match to the topic with most probable word health under class Medical identified in Section 3.5.3 for goals. The employed data analysis techniques match the class Machine learning in Section 3.5.2. Besides evaluating and validating the presented approach, the paper corresponds Human and Behavior goals with topics under classes Language and Truth.

In 2014, Tang et al. [Tang et al. 2014] introduced three new neural networks using word embedding for analyzing tweets. These neural networks are evaluated by using them with three different lexicons and comparing them to various existing approaches. The developed neural networks outperform the earlier ones by a wide margin. The authors argue that improvements are due to the better distinguishing of words with similar context. The data used in the paper is a partial match to the topic with most probable word “microblog”, but does not correspond to the rest of the words in that topic. Data is analyzed using neural networks and the approach has its own topic under Machine learning class identified in Section 3.5.2. Besides evaluating the developed approaches, one partial match to topic under goals introduced in Section 3.5.3 can be identified: because tweets are analyzed, topic with most probable words “hashtag” and “twitterspecif” under
class Finance and Corporate matches well. Twitter can also be seen on both wordclouds Figure 8a and 8b, figure 8b showing that Twitter is recently used more regularly.

In 2013, Warriner, Kuperman and Brysbaert [Warriner et al. 2013] gathered words from existing sources and provide emotional grades for the word. The words are graded by Amazon Mechanical Turk users and the results are gathered by the authors. The authors also consider possible uses for the created lemmas, report gender differences on grading certain categories of words, e.g., taboo words or weapons. Goal of the paper is to provide lemmas which can be used for emotion and sentiment analysis, thus is does not correspond well with identified goals in Section 3.5.3, however topic with most probable words “signal”, “demograph” and “gender” matches the paper partially as many gender differences relating to lemmas are reported in the paper. Also this paper provides a lexicon that is match for Section 3.5.2 Data analysis and natural language processing.

In 2012, Thelwall, Buckley and Paltoglou [Thelwall et al. 2012] evaluate a tool named “SentiStrength 2”, which is a lexicon based classifier used together with various machine learning algorithms. This tool is evaluated with a variety of data sets that feature shorts texts, e.g., YouTube comments and tweets. The tool performs well analyzing short texts found on social web and performs fairly context independent, this is a partial match to topic “microblog” under class Data in Section 3.5.1. SentiStrength itself is included in topic identified in Section 3.5.2 under class Tools. The top-cited paper for 2011 [Taboada et al. 2011] was among the top-five cited all time and was already summarized in fifth paragraph in this section.

Many of the most cited original research papers are studies introducing new approaches for sentiment analysis, thus they are mainly concerned with validating and evaluating the approach. Hence the goals identified in Section 3.5.3 are only used for validation purposes. For example, movie reviews are used to validate the performance of a new approach to sentiment analysis rather than starting with the goal of building a sentiment analysis tool that works for movie reviews themselves. Therefore, based on the reviews of the top cited original works we think these papers are solution oriented, i.e., a solution looking for a problem, rather than problem oriented, i.e., a problem looking for a solution.

3.6.2 Literature reviews

Here we provide brief overviews of the literature review papers among the top five cited of all time and the top cited paper from past five years individually (2011-2015). Additionally, we performed mapping on how the papers match the topics introduced in Sections 3.5.1- 3.5.3. In addition, Table 6 lists the 5 top-cited of all time and top-3 cited for each year between 2011 and 2015.

| Year | Paper Title | Citation Count |
|------|-------------|----------------|
| 2015 | A survey on opinion mining and sentiment analysis: Tasks, approaches and applications | 10 |
| 2014 | Sentiment analysis and applications: A survey | 66 |
| 2013 | New avenues in opinion mining and sentiment analysis | 157 |
| 2012 | A survey of opinion mining and sentiment analysis | 2169 |

Table 6 Top-cited literature reviews by years, paper title, year, (citation count) and citation count
The work by Pang and Lee [Pang and Lee 2008] is the most cited literature review on sentiment analysis we found. The review focuses on the fundamentals and basic applications of sentiment analysis. The introduced applications of sentiment analysis roughly correspond to the class application domain identified in section 3.5.3. Themes relating to voting are touched in sections 2.3 and 2.4 in the paper by Pang and Lee, which is a specific topic under class Society. The paper does not delve into technical details of different algorithms, focusing more on a broader overview of the field. However, several natural language processing and machine learning approaches are discussed by Pang and Lee [Pang and Lee 2008] in their sections 4 and 5, where they match the identified classes natural language processing (NLP) and machine learning in Section 3.5.2. The paper also contains a brief overview of broader implications of studies using sentiment analysis and additionally concerns on privacy and manipulation relating to sentiment analysis. The financial themes discussed match topic identified in Sections 3.5.3, as topics under classes Expertise and Finance corporate. The concerns for manipulation match topic with most probable words “mood” and “manipul” under class Other from Section 3.5.3. Opinion spam detection is also discussed in section 5.2.4.1 of Pang and Lee [Pang and Lee 2008] paper, which matches the topic with most probable words “spam” and “fake” under class Truth in Section 3.5.3. Finally, the work also covers some free external resources for researchers, such as lexical resources and data sets, which correspond to the general section Data in Section 3.5.1 and class Language in Section 3.5.3.

Liu and Zhang [Liu and Zhang 2012] show a thorough and extensive literature review and provide fundamental definitions. These definitions touch the topic with most probable word “segment” under class “Sentiment Analysis” identified in Section 3.5.2. Sentiment classification is divided into document level and phrase level classification, while literature covering both subjects is overviewed. Section 2 in their paper discusses various machine learning approaches which are identified in Section 3.5.2 of our paper. Naive Bayes is the most discussed approach which is also a topic under machine learning. Themes related to class natural language processing in Section 3.5.2 are found in Liu and Zhang [Liu and Zhang 2012] in sections 2.1 “Terms and their frequency” and 5.4 “Propagating and bootstrapping for lexicon expansion” of their paper. Their literature review also covers opinion spam and how to detect it, which corresponds to a topic under class Truth in Section 3.5.3. Besides covering the recent developments with 120 citations, the work includes open research questions related to the field.

Cambria, Schüller, Xia and Havasi [Cambria et al. 2013] give broad introductions on different techniques concerning sentiment analysis and their recent developments. Video and audio are predicted to be future data sources for sentiment analysis and further discussed under heading “Multimodal Sentiment Analysis”, which matches the class Data in section 3.5.1 and topics with most probable words such as speech and video. Aspect based sentiment analysis is discussed under heading “From Coarse-to Fine-Grained Analysis”, this matches a topic with most probable words “segment” and “aspectbas” under class “Sentiment Analysis”. Overall the paper is only 7 pages long and does not go in detail to one specific topic, thus serving better as introductory material.

Feldman [Feldman 2013] introduces basic techniques and some key applications of sentiment analysis. Sentiment analysis is divided into document and sentence level analysis, while lexicon acquisition and aspect-based (sometimes called feature based) sentiment analysis is also covered. Both word “lexicon” and “aspectbased” can be found on the word cloud on Figure 8b, where its in the top half meaning it has been used more recently. Particularly good match for the paper is the class Language introduced in Section 3.5.3, as themes such as analysis of a whole text versus a sentence and comparative opinion analysis are discussed. Sentence level analysis is also related to irony, which is a topic under class Truth in Section 3.5.3. Many of the topics under classes natural language processing (NLP) and machine learning from Section 3.5.2 are mentioned, but the paper
does not go into detail explaining them. Applications and sentiment lexicon acquisition are also discussed, which are related to classes language and finance corporate. One of the main contributions of the article is the collection of research problems the authors see most relevant.

Akehurst [Akehurst 2009] examines the relation of blogs and tourism in scientific literature, which matches several topics under class Travel from Section 3.5.3. Range of papers focusing on extracting relevant blog content, but also touching word of mouth, on the context of tourism are overviewed. Word of mouth is also identified as a topic under class sentiment analysis in Section 3.5.2.

In 2015, Ravi and Ravi [Ravi and Ravi 2015] gave a broad introduction to the field of sentiment analysis. The article has a section on subjectivity classification, sentiment classification, review usefulness measurement, lexicon creation, opinion word extraction and product aspect extraction, opinion spam detection and lastly various applications. Authors identify support vector machines as the most used technique in literature they have reviewed, which is a machine learning approach. Cross-domain and cross-lingual techniques are discussed in sections 3.2.3 and 3.2.4 of their paper, which have corresponding topics under class language in section 3.5.3 in the present article. Opinion spam is discussed in section 3.4 of their paper and matches topic under class Truth. Sub sections “Market and FOREX Rate Prediction” and “Box Office Prediction” match topics under class finance corporate in Section 3.5.3.

In 2014, Medhat, Hassan and Korashy [Medhat et al. 2014] main contribution is the classification and counting of articles using specific approaches related to sentiment analysis. The review goes in depth to the technical solutions used in the field. Two topics under class natural language processing (NLP), latent semantic indexing (LSI) and dictionary based approaches, are overviewed. Topics under class machine learning highly match the paper as well, with specific topics such as Bayesian networks, decision tree classifiers, corpus based approaches and transfer learning mentioned in more detail. One of the main contributions of this study is the gathering and classification of 54 original research papers, while mainly analyzing the techniques used on these reviewed papers.

The top-cited papers for 2012 and 2013 were both among the top-five cited all time and were already summarized [Cambria et al. 2013; Liu and Zhang 2012].

In 2011, Efron [Efron 2011] focused on microblogs, “consisting of brief textual entries written on an ongoing basis”. Services such as Facebook, Tumblr and FourSquare are considered to contain microblogs by Efron, but this paper is mainly concerned with data acquired from Twitter. Basics of data mining from microblogs and sentiment analysis are covered with existing scientific literature, examples of Efron’s own research are used to demonstrate data analysis. Microblogs is a topic under class data introduced in Section 3.5.1 and Twitter-related terms are in a topic under class finance and corporate from Section 3.5.3. Some studies reviewed are done on data containing political opinions, a topic under class Society from Section 3.5.3. Efron considers the biggest challenge of information retrieval research to be identifying the hidden problems of microblog users, as different search engines already offer answers to their implicit questions in microblogs themselves. The word “microblog” can also be found on the word clouds in Figure 8a and 8b, based on Figure 8b the word is used more after year 2013. Related words to “microblog” can be found on the upper part of Figure 8b, such “social”, “media” and “facebook”.

Overall comparison to existing literature reviews shows a good match with our classification and the ones created in the past. However, our classification is created by analyzing much larger pool of papers (over 5,000) than what has been used in prior studies. Due to our large data set we used computer-assisted analysis of the papers. The good match with prior classifications shows that using topic modelling clusters as a starting point for classification creation is a promising approach. On the other hand, large data allows us to be more certain than prior works that we have not missed significant research areas in the field. Additionally, our classification included some new areas not identified in prior highly cited literature reviews. For example, topics related to medicine, society, interaction and expertise were not covered in detail prior works. Overviews of different available languages for sentiment analysis were mostly absent in the past reviews. Data sources such as music and video were covered in only very limited capacity and brief mentions.
4 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 related to sentiment analysis.

Our choice of employing the Scopus database could be considered a limitation of this study. The most widely employed academic databases are Google Scholar, Scopus, and the Web of Science [Harzing and Alakangas 2016]. There has been much discussion about which database offers the least error-prone results and the highest coverage for knowledge discovery and researchers’ assessment, e.g., [Harzing and Alakangas 2016; Wildgaard 2015; Mongeon and Paul-Hus 2016; S. Adriaanse and Rensleigh 2013] . As reported by Harzing and Alakangas [Harzing and Alakangas 2016], Google Scholar offers the most comprehensive coverage of the literature, but Scopus follows closely and offers quality inclusion criteria and data exporting features that Scholar does not offer (furthermore, data scraping is forbidden by Scholar’s terms of services). Adriaanse and Rensleigh [S. Adriaanse and Rensleigh 2013] found that Scopus is the most reliable of most widely used academic databases, with least inconsistencies on author names, volume and issue number. Mongeon and Paul-Hus [Mongeon and Paul-Hus 2016] compared Scopus and the Web of Science across the fields of natural sciences and engineering, biomedical research, the social sciences, and arts and humanity. 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, while not perfect, has been optimal for the task at hand.

5 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 5,163 papers. 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, discovered 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 science of sentiment analysis and opinion mining has deep roots in the studies on public opinion analysis at the start of 20th century. First papers that matched our search strings were post-World War II studies that investigated the public opinion, for example towards communism, in countries recovering from the devastations of the war. Still, the topic was in hibernation until the mid-2000’s when it finally emerged as an important research topic due to the need and availability of online product reviews. In 2005, only 16 papers about this topic were published while in 2015 the number was nearly 1,300. This gives us a nearly 100-fold increase in a decade making sentiment analysis undoubtedly one of the fastest growing research areas of the previous years.

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 5,000, 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 29% 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, expertise, 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 produce 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.

So, what holds in the future of sentiment analysis? We assume 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 computing methods such as speech, gaze, and neuromarker analysis. However, we are doubtful whether sentiment analysis can achieve a similar 100-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 100,000 papers on sentiment analysis published in the year 2025.

We offer closure and some future work topics to our opening quote of the idea that “The pen is mightier than the sword”. As laid out in the Introduction, leaders have always been interested in public opinion, for example to increase popularity or to stump the attempts of revolution. French revolution sparked interest in public opinion and books gauging the public opinion started to appear already in early 1800’s [Prentice 1802]. Public opinion quarterly started to appear in 1937 and only a few years earlier Nazi’s had taken over in Germany. More recent events highlight future applications that could be essential in protecting democracy and civil liberties. Perhaps, sentiment analysis combined with IP-network analysis can help in detecting fake opinion post generated by a troll army, aka web-brigade [Anon 2016c]. Citizens of more established democracies can also benefit from sentiment analysis if it can detect opinionated or completely fake news articles that allegedly affected the presidential election of the USA [Anon n.d.]. Similarly, companies are influencing public opinion by generating fake news and organizations to protect their interest against scientific facts like climate change [Washington 2013]. Our analysis had a branch labeled as the Truth which contains research about detecting fake and spam opinions, see Figure 14.

6 ACKNOWLEDGEMENTS

The first and third author have been partially supported by the Academy of Finland grant 298020.

7 REFERENCES

ČENĚK ADAMEC and Ivan Viden. 1947. Polls come to Czechoslovakia. Public Opinion Quarterly 11, 4 (1947), 548–552.
Gary Akehurst. 2009. User generated content: the use of blogs for tourism organisations and tourism consumers. Service Business 3, 1 (2009), 51–61.
Anon. DBLP: Records in DBLP. Retrieved December 2, 2016a from http://dblp.uni-trier.de/statistics/recordsindblp
Anon. Fake News, Trump and the Pressure on Facebook. Retrieved December 2, 2016b from https://www.bloomberg.com/news/articles/2016-11-25/fake-news-trump-and-the-pressure-on-facebook-quicktake-q-a
Anon. 2014. Greedy Indian Publisher Charges Authors and Readers, Requires Copyright Transfer. (March 2014). Retrieved October 27, 2016 from https://scholarlyoa.com/2014/03/18/greedy-indian-publisher/
Anon. MPQA Corpora. Retrieved October 27, 2016c from http://mpqa.cs.pitt.edu/corpora/
R. Core Team and others. 2013. R: A language and environment for statistical computing. (2013).
Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* 63, 1 (2012), 163–173.
John Thorley. 2004. *Athenian democracy*, Psychology Press.
Mikalai Tsytsarau and Themis Palpanas. 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery* 24, 3 (2012), 478–514.
Richard Van Noorden, Brendan Maher, Regina Nuzzo, and others. 2014. The top 100 papers. *Nature* 514, 7524 (2014), 550–553.
Paul MB Vitányi, Rudi Cilibrasi, and others. 2010. *Normalized Web Distance and Word Similarity*.
Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.
Haydn Washington. 2013. *Climate change denial: Heads in the sand*, Routledge.
Reinhard Werner. 2015. The focus on bibliometrics makes papers less useful. *Nature* 517, 7534 (2015), 245.
Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 625–631.
Lorna Wildgaard. 2015. A comparison of 17 author-level bibliometric indicators for researchers in Astronomy, Environmental Science, Philosophy and Public Health in Web of Science and Google Scholar. *Scientometrics* 104, 3 (2015), 873–906.
Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 347–354.
Stephan Winkler, Susanne Schaller, Viktoria Dorfer, Michael Affenzeller, Gerald Potz, and Michał Karpowicz. 2015. Data-based prediction of sentiments using heterogeneous model ensembles. *Soft Computing* 19, 12 (2015), 3401–3412.
Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 427–434.
José P. Zagal, Noriko Tomuro, and Andriy Shepitsen. 2011. Natural language processing in game studies research: An overview. *Simulation & Gaming* (2011), 1046878111422560.
Liangzhao Zeng, Boualem Benatallah, Anne HH Ngu, Marlon Dumas, Jayant Kalagnanam, and Henry Chang. 2004. Qos-aware middleware for web services composition. *IEEE Transactions on software engineering* 30, 5 (2004), 311–327.
Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. 2012. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1528–1531.