Research On Sentiment Analysis: The First Decade

Oskar Ahlgren

Abstract—The first publications on sentiment analysis and opinion mining were published roughly a decade ago. Now it is time to look back on the achievements so far. This paper presents statistics on the evolution of sentiment analysis. What kind of topics have been discussed? How has their popularity changed over time? Who have been the leading researchers? Answers to these questions are provided by statistical analysis on keywords and by applying Latent Dirichlet Allocation to the titles and abstracts of the publications. The aim of this paper is to provide background information on the big picture of semantic analysis and its development over time.

1 Introduction

Sentiment analysis or opinion mining is the process of identifying and detecting subjective information using natural language processing, text analysis, and computational linguistics. In short, the aim of sentiment analysis is to extract information on the attitude of the writer or speaker towards a specific topic or the total polarity of a document.

The first papers that used sentiment analysis among their keywords were published about a decade ago, but the field can trace its roots back to the middle of the 19th century. One of the pioneering resources for sentiment analysis is the General Inquirer [27]. Although it was launched already in the 1960s, it is still being maintained. Sentiment identification is a very complex problem, and thus much effort has been put into analyzing and trying to understand its different aspects, see for instance [3], [7], [14], [16], [17], [32], [38]. Common sources of opinionated texts have been movie and product reviews [24], [39], [42], blogs [13], [40], [41] and Twitter posts [11], [18], [29]. As news stories have traditionally been considered neutral and free from sentiments, little focus has been on them. However, the interest in this domain is growing, as automated trading algorithms account for an ever-increasing part of the trade. Refer to the works by Malo et al. [21] and Takala et al. [28] for more details. A fast and simple method for determining the sentiment of a text is using a pre-defined collection of sentiment-bearing words and simply aggregating the sentiments found [10], [19], [26], [30], [31], [37]. More advanced methods do not treat all words equally but assign more weight to important words depending on their position in the sentence. For instance, Malo et al. [21] have developed advanced methods for analyzing sentiments in the financial domain. Unfortunately, most domains are very specific, which means that one collection of words that is efficient for one domain most likely will not perform as well in another domain. Efforts have been made to solve this shortcoming for instance by Li and Zong [20] with their multi-domain sentiment classification approach. Another branch of sentiment analysis has been using a more linguistic approach, and they have been focusing on extracting the opinion holders and the quotes in texts [1], [2], [6], [9]. As natural language processing techniques keep improving and computational power keeps getting cheaper, even more efforts are likely to be put into sophisticated automatic text processing methods. Therefore, it is time to summarize the first decade of sentiment analysis, and thus the main objective of this paper is to investigate:

> *What specific research topics have there been in sentiment analysis literature?*

This will be achieved by analyzing the usage of keywords and by applying Latent Dirichlet Allocation (LDA) to the abstracts of the publications. The keywords are provided by the authors and the journal editors and they are intended to describe the publication at hand as accurately as possible. The purpose of the LDA analysis is to examine what words are used in the titles and abstracts and to investigate how the publications can be clustered into topics based on the words they contain. The results of the simpler keyword analysis and the LDA can then be compared. Topic models have been used in other fields [12], [15], [23] but to the best of our knowledge not in sentiment analysis. Therefore, the second objective of this paper is:

> *Show how topic models can be used to add value to traditional literature reviews.*

The layout of the rest of this paper is as follows. Sections 2 and 3 will describe the data collection process and the software used, respectively. In Section 4 the results are presented. Section 5 will discuss the results while Section 6 concludes the paper with acknowledgments.

2 Data Collection

In order to investigate the development of the sentiment analysis field, all scientific publications matching the keywords *Sentiment analysis*, *Opinion Mining*, *Sentiment Classification*, or *Polarity classification*, published before 2015, were downloaded from the Scopus database. The result was then filtered, and all non-English publications were discarded

• *Department of Information and Service Economy, Aalto University School of Economics, P.O. Box 21210, FI-00076 Aalto, Finland*
E-mail: oskar.ahlgren@aalto.fi

as well as all publications other than journal articles and conference publications. The remaining entries were cleaned from irrelevant data and checked for consistency, e.g. misspelled author names or with initials, keywords written in full-length or with acronyms, etc. After applying the filters and the data cleaning process, 2592 scientific publications remained; of these, roughly one third were journal articles, and the rest were conference papers. The clean data were finally analyzed in R and VOSviewer [34] in a Windows 7 environment. The Latent Dirichlet Allocation was done with the Mallet package in R.

3 METHODS

A research profiling study should answer questions like: *What?*, *Who?*, and *Where?*. To answer these questions, the data will be analyzed in two different ways: with a statistical keyword analysis[1] and with Latent Dirichlet Allocation. While the keyword analysis looks at the keywords only, LDA analyzes the words used in the title and the abstract of the publications as well. The purpose of the LDA analysis is to see what conclusions regarding the topics could be made based on the words used in the title and abstract. The following subsections will briefly discuss these methods.

3.1 Keyword Analysis and VOSviewer

A pure keyword analysis is a rather straight forward approach. All keywords are indexed and processed, and when the results are displayed in graphs and tables, conclusions can be drawn. These conclusions are typically trends and top lists. However, based on the keywords, co-occurrence maps can also be created. These maps display the relationship between co-authors and keywords used in a single paper, and they can be created using VOSviewer.

VOSviewer is a program developed by Van Eck and Waltman [34] for constructing and viewing bibliometrics maps. The program creates co-occurrence maps based on how often two keywords are mentioned together. The more similar (both being used in the same publication) they are, the closer they are in the map. The maps are created based on the co-occurrence matrix in three steps. In the first step, a similarity matrix is calculated based on the co-occurrence matrix. Then in the second step, a map is created by using the VOS mapping technique on the similarity matrix. In the final step, the map is optimized. For more details, please refer to the original work by Van Eck and Waltman [36].

Step 1: Similarity Matrix

VOSviewer uses a similarity measure called association strength [33], [35], and using this measure the similarity s_{ij} between two items i and j is then calculated as

$$s_{ij} = \frac{c_{ij}}{w_i w_j}, \tag{1}$$

where c_{ij} denotes the number of co-occurrences of items i and j, respectively, and w_i and w_j denote either the total number of occurrences of items i and j or the total number of co-occurrences of these items.

1. Even though the name implies otherwise, also bibliological data, such as the names of the authors and their affiliations, are included in the keyword analysis.

Step 2: VOS Mapping Technique

The VOS mapping technique creates a two-dimensional map in which the items are placed in such a way that the distance between two items i and j reflects their similarity s_{ij}. The higher the similarity between two items, the closer they are in the map. The objective function that is to be minimized is given by

$$V(\mathbf{x}_1, ..., \mathbf{x}_n) = \sum_{i<j} s_{ij} ||\mathbf{x}_i - \mathbf{x}_j||^2, \tag{2}$$

where the vector $\mathbf{x}_i = (x_{i1}, x_{i2})$ is the location of item i in the map, and $||\bullet||$ is the Euclidean norm. The objective function is minimized subject to the constraint

$$\frac{2}{n(n-1)} \sum_{i<j} ||\mathbf{x}_i - \mathbf{x}_j|| = 1. \tag{3}$$

Step 3: Translation, Rotation, and Reflection

The optimization in Step 2 does not yield a single globally optimal solution, as any translation, rotation, or reflection of the solution is also globally optimal. The following three transformations are applied to the solution:

- **Translation** transforms the map to be centered at the origin.
- **Rotation** rotates the map so that the variance on the horizontal dimension is maximized.
- **Reflection** is applied in the appropriate axis, if the median of $x_{11}, ..., x_{n1}$ or $x_{12}, ..., x_{n2}$ is larger than 0.

3.2 Latent Dirichlet Allocation

Statistical models, such as principal component analysis, factor analysis, clustering algorithms and latent semantic indexing are able to identify topics being discussed in sets of documents. The problem with these models is that they can only associate a document with one topic. Since documents tend to discuss several different topics, this is a serious limitation. The LDA model is capable of capturing the multi-topic characteristics of documents [5], and it is the simplest topic model that is suited for analyzing text documents [8]. LDA assumes that the data are structured and have patterns, even if these cannot be observed easily. In this case, the hidden patterns will form the basis that allow the documents to be classified into the different topics. In LDA, documents are assumed to be just a set of words, i.e. the ordering of the words is unimportant. The same also applies to the order of the documents in the collection. Say, for instance, that an LDA model can classify documents as either *Newspaper_related* or *Website_related*. If the document contains words like *print*, *columns*, and *Gutenberg*, it has a higher probability of being classified as *Newspaper_related*. In the same manner documents classified as *Newspaper_related* are more likely to generate these words. Words like *the* and *we* can be ignored altogether, as they should have equal probability of being classified as *Newspaper_related* or *Website_related*.

The idea that documents consist of a random mixture of topics, and that topics are characterized by the specific distribution of words, is clearly a probabilistic process,

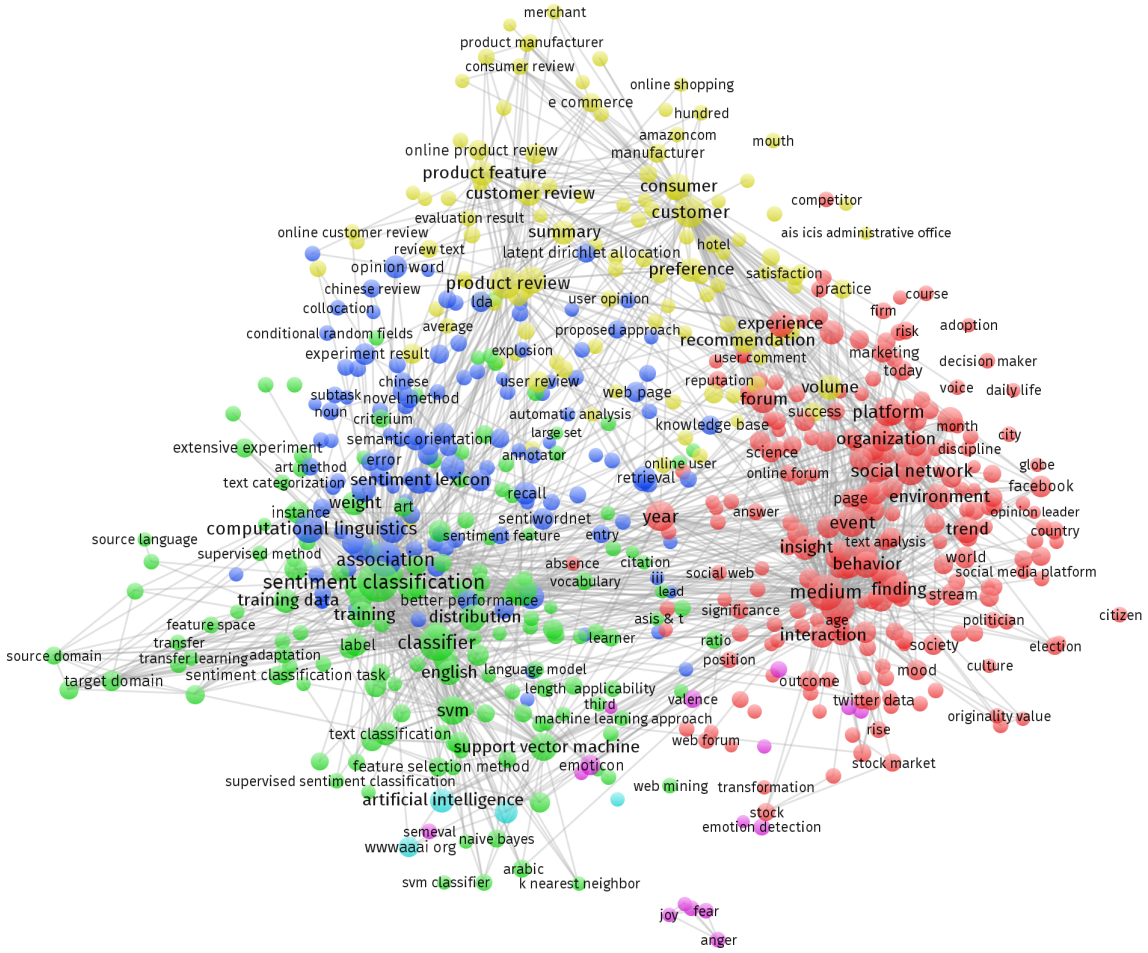


Fig. 1: Co-occurrence map of most used keywords

and for the calculation of the probability density of topic proportion two parameters are needed: one for the document distribution for a given topic, and one for the topic distribution for a specific document. These parameters are denoted α and θ, respectively. We then get

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}, \tag{4}$$

where $\Gamma(x)$ is the Gamma function. To find the hidden patterns, the joint distribution must be calculated at three different levels: word, document, and corpus level. These probabilities will assess the likelihood of words being able to describe a given topic or the likelihood of a document belonging to a specific topic. Using the previous result, the joint distribution at the word level is calculated for every word in all documents in the corpus, and it is given by

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_b, \beta). \tag{5}$$

In the above equation, z denotes the topics associated with specific words, and β is a parameter for the topic distribution of a given word. By integrating over θ and summing over z, we obtain the marginal probabilities for a single document

$$p(w|\alpha,\beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta. \tag{6}$$

Once these probabilities are calculated for all documents, the corpus level probability is given by the product of all the marginal document probabilities

$$p(C|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d, \tag{7}$$

For more details on LDA, please refer to the original works of Blei et al. [4].

4 Results

Keyword analysis is a quick and reliable way to analyze the field, as the keywords are given by the authors and publicists themselves. LDA, on the other hand, gives deeper insights because it sees beyond the keywords and looks at the words used to describe the conducted research.

	Journal Publications		Conference Proceedings	
Rank	Keyword	#	Keyword	#
1.	Sentiment Analysis	888	Sentiment Analysis	1643
2.	Opinion Mining	500	Opinion Mining	833
3.	Data Mining	429	Data Mining	789
4.	Sentiment Classification	263	Sentiment Classification	435
5.	Text Mining	152	Natural Language Processing Systems	370
6.	Social Networking (Online)	145	Social Networking (Online)	369
7.	Natural Language Processing Systems	130	Computational Linguistics	285
8.	Social Media	114	Semantics	218
9.	Natural Language Processing	113	Social Media	197
10.	Semantics	102	Classification (Of Information)	192
11.	Artificial Intelligence	100	Natural Language Processing	176
12.	Classification (Of Information)	99	Twitter	170
13.	Text Processing	91	Text Mining	162
14.	Twitter	85	Text Processing	157
15.	Internet	78	Learning Systems	147
16.	Feature Extraction	69	Artificial Intelligence	143
17.	Learning Systems	68	Information Retrieval	143
18.	Algorithms	65	Knowledge Management	119
19.	Information Retrieval	65	World Wide Web	113
20.	Machine Learning	65	Feature Extraction	103

TABLE 1: Publication statistics: Top 20 Keywords used in Journal Publications and in Conference Proceedings

4.1 Keyword Analysis

It is natural to start by analyzing the scope of the research literature by examining which keywords have been most commonly used. Figure 1 shows the co-occurrence map of most-used keywords.[2]

For obvious reasons, *Sentiment Classification* is at the heart of the map in the green cluster together with closely related keywords like *SVM* and *Classifier*. The blue cluster contains keywords associated with *Semantics* and *Lexicons*. Therefore, the green and the blue clusters are partly integrated. Product reviews is one of the most researched fields in sentiment research, and it is represented by its own yellow cluster in the top left corner of the figure. The red cluster is tied to one other frequent domain of sentiment research: media. Here are common keywords like *Social Network*, *Platform* and *Interaction*. Despite being used in 2531 of 2592 publications, *Sentiment Analysis* is not present in the graph. At first this might seem strange, but the simple explanation is that its explanatory value is non-existent due its high frequency. In total, well over 9000 unique keywords were used in the publications.

In Table 1, the keywords used in journal publications and in conference proceedings are compared. The vast majority of the used keywords are present in both top lists and also at similar positions. Turns out that the keywords that are only found in one of the lists were left just outside the top list. One interesting observation is that *Internet* are more popular than *World Wide Web* for journal publications, even if the words are close to perfect synonyms.

Another interesting question is *When?*. The actual publishing date of individual publications are not interesting, but the general trend is. A decade ago, there was only a handful of publications, and in 2014, well over 600 of them were published, see Figure 2.

2. In Scopus, there are two sets of keywords: author keywords and publisher keywords. As some publications only have either set, these were combined and any overlapping removed.

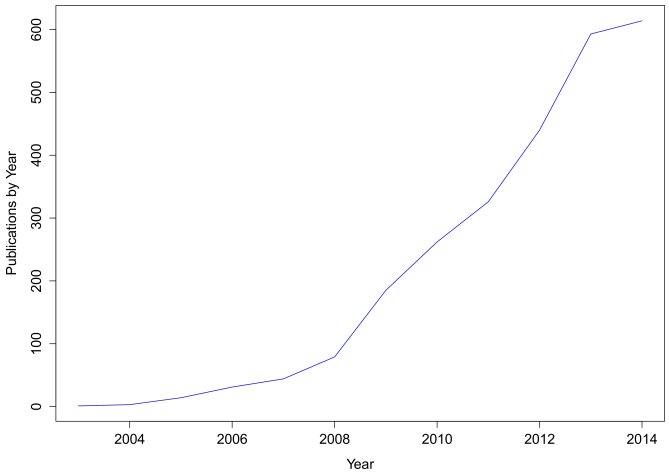


Fig. 2: Publications by year

Based on Figure 3, there is a clear rising trend for all the plotted keywords. There are, however, differences between them. For instance, *Sentiment Classification* and *Opinion Mining* did not have more uses in 2014 than in the previous year, and *Social Networking* was hardly used at all before 2010, while in 2014, it was used almost 200 times. As *Sentiment Analysis*, *Opinion Mining*, and *Sentiment Classification* were used as search criteria, they are naturally at the top of the table. Still, *Data Mining* split the trio and is the second most-used keyword. Data mining is the process of finding useful information in databases, and thus the usefulness of these techniques to this field is clear. One common denominator for the Top 15 keywords is that most of them are all related to various techniques. This is further evidence of the multitude of domains researched, as few topic rise above the methods and techniques used. It is also worth noting that while some keywords are perfect synonyms, in this survey they are still considered to be different. In this domain, *Sentiment Classification* and *Classification* could almost be considered synonyms, as sentiments are the likely target of the classification process.

When it comes to the number of citations, one paper stands out from the rest. The article, *Opinion Mining and*

Rank	Year	Author	Title	#
1.	2008	Pang B. and Lee L.	Opinion Mining And Sentiment Analysis	1583
2.	2004	Hu M. and Liu B.	Mining And Summarizing Customer Reviews	962
3.	2003	Dave K., Lawrence S. and Pennock D.M.	Mining The Peanut Gallery: Opinion Extraction And Semantic Classification Of Product Reviews	510
4.	2005	Wilson T., Wiebe J. and Hoffmann P.	Recognizing Contextual Polarity In Phrase-Level Sentiment Analysis	445
5.	2007	Blitzer J., Dredze M. and Pereira F.	Biographies, Bollywood, Boom-Boxes And Blenders: Domain Adaptation For Sentiment Classification	275
6.	2005	Pang B. and Lee L.	Seeing Stars: Exploiting Class Relationships For Sentiment Categorization With Respect To Rating Scales	240
7.	2008	Ding X., Liu B. and Yu P.S.	A Holistic Lexicon-Based Approach To Opinion Mining	212
8.	2008	Abbasi A., Chen H. and Salem A.	Sentiment Analysis In Multiple Languages: Feature Selection For Opinion Classification In Web Forums	211
9.	2006	Kennedy A. and Inkpen D.	Sentiment Classification Of Movie Reviews Using Contextual Valence Shifters	197
10.	2007	Mei Q., Ling X., Wondra M., Su H. and Zhai C.	Topic Sentiment Mixture: Modeling Facets And Opinions In Weblogs	168

TABLE 2: Publication statistics: Top 10 Cited Publications

Rank	Year	Author	Title	#	Per Year
1.	2013	Moraes R., Valiati J.F., Gaviao Neto W.P.	Document-level sentiment classification: An empirical comparison between SVM and ANN	39	19.5
2.	2009	Lin C., He Y.	Joint sentiment/topic model for sentiment analysis	121	20.2
3.	2009	Prabowo R., Thelwall M.	Sentiment analysis: A combined approach	123	20.5
4.	2011	Jo Y., Oh A.	Aspect and sentiment unification model for online review analysis	84	21.0
5.	2014	Cambria E., Olsher D., Rajagopal D.	SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis	23	23.0
6.	2011	Jiang L., Yu M., Zhou M., Liu X., Zhao T.	Target-dependent Twitter sentiment classification	95	23.8
7.	2011	Ghose A., Ipeirotis P.G.	Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics	121	30.3
8.	2012	Thelwall M., Buckley K., Paltoglou G.	Sentiment strength detection for the social web	108	36.0
9.	2013	Feldman R.	Techniques and applications for sentiment analysis: The main applications and challenges of one of the hottest research areas in computer science	74	37.0
10.	2013	Cambria E., Schuller B., Xia Y., Havasi C.	New avenues in opinion mining and sentiment analysis	91	45.5

TABLE 3: Publication statistics: Top 10 Recent Publications with the highest citation average

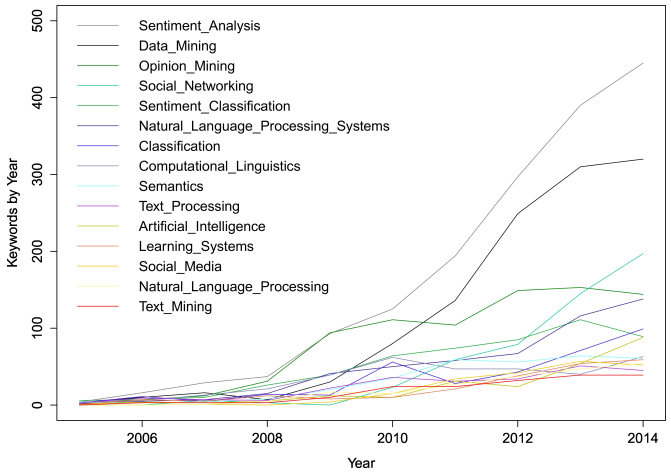


Fig. 3: Publication statistics: Top 10 Most used Keywords

Sentiment Analysis, by Pang and Lee [25] is often considered to be the Bible of this field. The high regard is clearly also reflected in the sheer number of citations of this work. It has been cited 1583 times, which means that more than half of all authors have cited this work. In this aspect, it outperforms all other papers even on the Top 10 list in Table 2 multiple times. The same authors have a second top position at number 6.

If Table 2 is compared with the list of the most published authors in Table 4, there is surprisingly little overlap. The only authors to make both lists are B. Liu and H. Chen. In total, over 5200 authors have contributed to this field. However, Table 2 is not completely fair to more recent publications, as it takes a fair amount of time to gather citations. To highlight influential recent publications Table 3 displays the mean annual citations instead. In the last few years E. Cambria and M. Thelwall have published several well received publications. Interesting observations can also be drawn from Figure 4, which displays co-operation between authors. A line between two authors in the figure is an indication of co-authored articles. The actual distance between them in the figure has no implications; it is just a matter of presenting the map in a clear fashion. Distinctive clusters are separated by color and a few names pop out: H. Wang (purple), Y. Liu (red), B. Liu (brown), Y. Zhang (olive), and E. Cambria (pink). Not surprisingly, the top co-operating authors are also the most published ones.

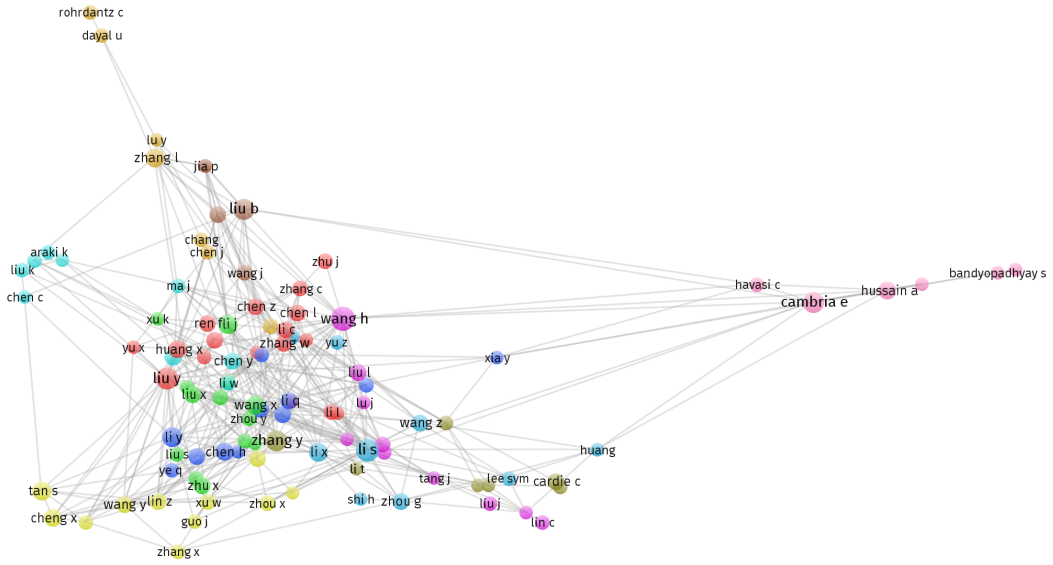


Fig. 4: Co-occurrence map of most productive authors

Rank	Author	#	Rank	Author	#
1.	Wang, H.	37	17.	Chen, Y.	13
2.	Liu, Y.	27	18.	Li, X.	13
3.	Liu, B.	25	19.	Zhou, G.	13
4.	Li, S.	21	20.	Zhu, X.	13
5.	Li, Y.	19	21.	Li, Q.	12
6.	Zhang, Y.	19	22.	Liu, X.	12
7.	Li, S.	18	23.	Wang, T.	12
8.	Cambria, E.	17	24.	Wang, Y.	12
9.	Hussain, A.	16	25.	Xu, H.	12
10.	Li, J.	16	26.	Zhang, J.	12
11.	Chen, H.	15	27.	Cambria, E.	11
12.	Cheng, X.	15	28.	Ren, F.	11
13.	Huang, X.	15	29.	Thelwall, M.	11
14.	Tan, S.	15	30.	Zeng, D.	11
15.	Wang, X.	15	31.	Balahur, A.	11
16.	Cardie, C.	14	32.	Wang, H.	11

TABLE 4: Publication statistics: Top 32 Published Authors

Rank	Country	#	%
1	China	920	21.1
2	United States	846	19.4
3	India	278	6.4
4	United Kingdom	173	4
5	Spain	171	3.9
6	Japan	162	3.7
7	Italy	157	3.6
8	Germany	156	3.6
9	Taiwan	107	2.5
10	South Korea	105	2.4
11	Hong Kong	103	2.4
12	Singapore	98	2.3
13	Canada	84	1.9
14	France	80	1.8
15	Greece	66	1.5
16	Australia	60	1.4
17	Malaysia	52	1.2
18	Netherlands	50	1.1
19	Brazil	47	1.1
20	Switzerland	34	0.8
	TOP-20	3749	86.2
	In Total	4350	100.0

TABLE 5: Publication statistics: Top 20 Countries

In Table 5, the home countries of the affiliations are summarized. Unfortunately, Scopus does not associate the authors with their affiliations, it only list them both. It is therefore not possible to know who belongs to which institution, if there is more than one affiliation. As a result, the table is a bit skewed, but the trend is still clear. Many small countries in South East Asia, (Hong Kong[3], Singapore, and Taiwan) have a fair amount of publications compared to their sizes. Also Greece has highly successful researchers in this field. Regardless, this result correlates well with Table 6, where the most publishing institutions are listed. The results show that there is a lot of focus on sentiment analysis in Chinese and American institutions. This Table relies on what affiliation the authors have stated: some of them have given a specific department as their affiliation, while others have chosen to give only their university, school, or research institute. By going with the biggest entity (university or research institute), there might be some changes in the ranking.

4.2 Latent Dirichlet Allocation

The titles and abstracts were used exactly as they were written, and this resulted in both *Word* and *Words* being top ranked for the topic of *Sentiment Analysis*. If the input would have been stemmed[4], the differentiating words would naturally have been different. Without stemming, it is possible that the singular and plural forms will be strongly associated with different topics. If this is desirable or not will depend on the application, but the user needs to be aware of the matter. Table 7 presents the Top lists for the words used in the abstracts and in the titles. Many of these words are commonly associated with sentiment analysis, such as *Sentiment*, *Analysis*, and *Opinion*, but many are also popular

3. Hong Kong is not an independent country, it is an autonomous territory belonging to the People's Republic of China. Still it is often reported as the location of contributing authors.

4. Stemming is the process of reducing all words to their base form. The base form of both *reads* and *reading* is *read*.

Rank	Research Institution	Publications
1.	Microsoft Research Asia, Beijing, China	21
2.	Department of Computer Science, Cornell University, Ithaca, NY, United States	17
3.	Department of Computer Science, University of Illinois at Chicago, Chicago, IL, United States	17
4.	School of Computer Engineering, Nanyang Technological University, Singapore	17
5.	Department of Information Systems, City University of Hong Kong, Hong Kong	16
6.	National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China	16
7.	L3S Research Center, Hannover, Germany	15
8.	School of Computer Science and Technology, Shandong University, Jinan, China	15
9.	Natural Language Processing Lab, Soochow University, Suzhou, China	14
10.	Knowledge Media Institute, Open University, Milton Keynes, United Kingdom	13

TABLE 6: Publication statistics: Top 10 Institutions

words that describe research in general, such as *Data* and *Approach*.

Rank	Word	Term Freq	Rank	Word	Doc Freq
1	Sentiment	5300	1	Sentiment	1574
2	Analysis	2912	2	Analysis	1326
3	Opinion	2249	3	Based	1165
4	Based	2135	4	Results	927
5	Classification	1825	5	Using	858
6	Data	1663	6	Opinion	835
7	Reviews	1489	7	Data	823
8	Information	1393	8	Classification	785
9	Social	1322	9	Information	785
10	Approach	1319	10	Approach	762

TABLE 7: Publication statistics: This table show the total number of times a word is used in abstracts and titles (Term Frequency) and in how many documents it is present in total (Document Frequency)

It is a delicate task to choose the number of topics LDA should identify. If the number is too large, many of the topics will be very similar, and if the number is too small, there is a risk that key topics will remain undetected. For each topic, the 50 most relevant words were identified, and based on these words, the subject of the topic was determined. Based on the data at hand, ten topics were found to be a reasonable choice, and the ten identified topics were

Opinion Mining, Sentiment Classification, Sentiment Analysis, Sentiment Models, Methods, Data Analysis, Reviews, Emotions, Social Media, and Finance.

Even among these topics there is some partial overlap. Clearly, *Sentiment Classification* and *Sentiment Analysis* are just different aspects of the same topic. *Sentiment Models* and *Methods* are also closely related. In fact, the first six topics are highly similar. If the number of detected topics would have been higher, there would likely have been even more almost identical topics. On the other hand, all ten of the identified topics are reasonably different.

The word clouds in Figure 5 show the most important words for each topic. In this figure, the size of the word corresponds to its relative importance for the topic. It is worth noting that some words can be important for more than one topic. For instance, *Learning* is a top-ranked word for both topics of *Emotions* and *Methods*[5]. At this point, it is important to stress that topic classification is not random, but how a publication is classified might be. In the most extreme cases, the classification might depend on whether a single specific word is used or not, as synonyms are not taken into account in LDA. Next follows a short description of each topic and the words that best describe it[6].

Opinion Mining gave the name for the entire genre. The top publications in this topic aim to detect opinions mainly in political texts and debate transcriptions and often also those of the opinion holder. In politics, the sentiment is not necessarily positive or negative. Therefore, the classification can be into supporting or opposing, respectively, *pro* or *con* as well. It is safe to assume that many would associate a *Blog* with the topic *Social Media* instead of *Opinion Mining*. However, this just shows that intuition can lead one astray.

Identifying words: *Opinion, Web, Mining, Topic, Blog*

Sentiment Classification tends to use well-known text classification algorithms, such as naïve Bayes, maximum entropy, and support vector machines, to classify texts as positive or negative. Much effort has been put into analyzing Chinese texts, which is not surprising taking into account that China is one of the leading publishers in this field.

Identifying words: *Sentiment, Classification, Analysis, Based, Learning*

Sentiment Analysis focuses on tools and other lexical resources for analyzing texts. One of the most well-known sentiment tools, SentiFul by Neviarouskaya et al. [22], belongs to this topic.

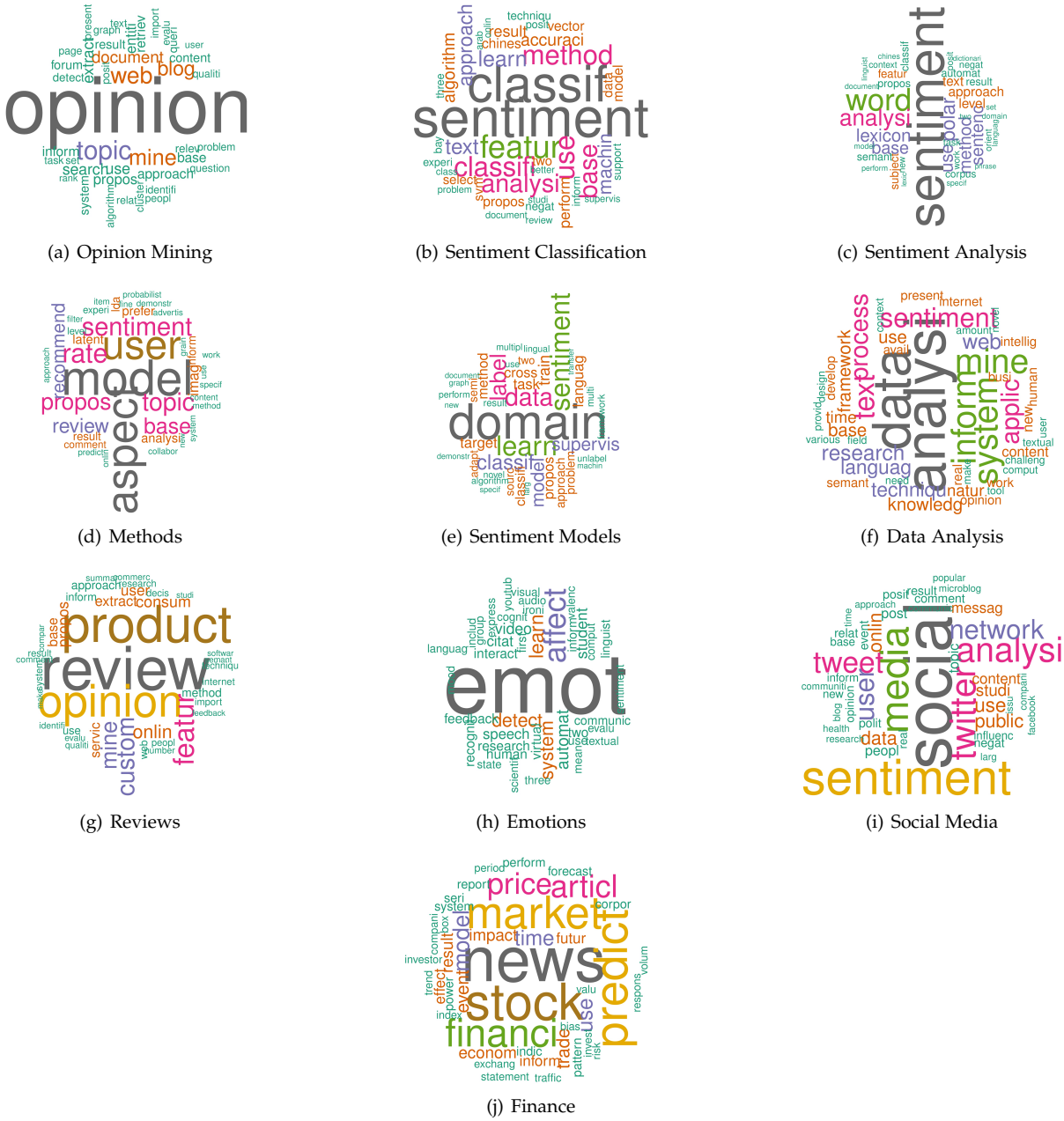
Identifying words: *Sentiment, Analysis, Words, Polarity, Based*

Sentiment Models try to model sentiments and opinions, mainly in reviews and rankings. High-ranking publications tend to add value to the analysis by taking user preferences or behavior into account.

Identifying words: *Model, Aspect, User, Based, Sentiment*

5. The complete list of words and their relative importance can be found in Appendix A.

6. To provide a bigger picture, the words are stemmed.



(a) Opinion Mining (b) Sentiment Classification (c) Sentiment Analysis

(d) Methods (e) Sentiment Models (f) Data Analysis

(g) Reviews (h) Emotions (i) Social Media

(j) Finance

Fig. 5: Word clouds of the 10 identified topics

Methods concerns itself with any method used in sentiment mining. The most common ones are classifying and learning methods.

Identifying words: *Domain, Sentiment, Data, Learning, Classification*

Data Analysis covers matters ranging from approaches to sentiment analysis to application architecture and interfaces.

Identifying words: *Analysis, Data, Information, Mining, Text*

Reviews are a popular source for opinions and sentiments. Apart from extracting sentiments, also product features are mined, and texts are summarized.

Identifying words: *Reviews, Product, Opinion, Mining, Opinions*

Emotions, in this topic, are greatly associated with human learning situations. Issues researched cover matters from visual learning to how emotions influence students' learning progress.

Identifying words: *Emotion, Emotional, Affective, Learning, Affect*

Social Media is not only a well-researched topic, it is one of the core topics of sentiment analysis. In this topic, the focus is on detecting sentiments in various social media.

Identifying words: *Social, Sentiment, Media, Twitter, Analysis*

Finance is not one of the identified clusters in Figure 1, so in this case, LDA provided some additional insights beyond the keywords. Scientists have been researching matters from risk evaluation to identifying game changers, simulating trades, and modelling movements of commodities.

Identifying words: *News, Stock, Financial, Market, Articles*

LDA calculates a measure describing how related a document is to a given topic. All documents are related to all topics to a certain extent, and the sum of all of the relatedness indexes always equals 1. This means that some publications might be closely associated with several topics, while others are not related particularly strongly to any topic. The strongest relation a publication had in this survey was 0.98, and the average of all documents to their most relevant topic was 0.49. Therefore, a cutoff of 0.25 was used to determine whether a document belongs to a specific topic or not. This resulted in the 2592 analyzed documents being associated with a total of 3657 topics, while 32 documents had no significant relation to any topic.

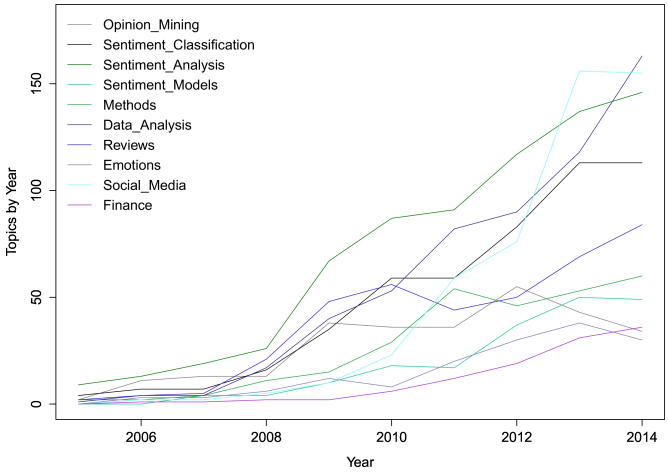


Fig. 6: Evolution of the 10 identified topics in LDA

The evolution of the identified topics is shown in Figure 6. As with the keywords in Figure 3, most topics show a substantial increase over time. The exceptions are *Opinion Mining* and *Emotions*, but these topics have rather narrow scopes. The most researched topics according to LDA are *Sentiment Classification*, *Sentiment Analysis*, *Data Analysis*, and *Social Media*, all of which are core topics also found in the co-occurrence map of the keywords. Even if the results of the keyword analysis and LDA cannot be compared directly—as each keyword should be considered a distinct and unique topic—the maps are still able to provide additional knowledge about larger quasi-topics. One important difference is that the map topics are clusters and not clearly separated, while the topics in LDA, on the other hand, are distinctively separated. Taking these differences into account, both methods generate very similar results. It is hardly surprising that the topics with the wider research spectrums have more publications and also show a stronger growth.

5 Discussion and Conclusions

In this paper, we have conducted a bibliometrics research study of sentiment analysis using the Scopus database. First, based on a few central keywords, basic statistics of the field were reported. Despite sentiment analysis being a fairly new subject, a few observations can still be made. The most obvious one is that the field has experienced exponential growth. This is true for most keywords, but the keyword analysis showed that some, like *Social Networking*, experience a higher growth rate than others, and some new keywords, like *Twitter*, have gained popularity recently. Co-occurrence maps are helpful in finding similarities and patterns between keywords and authors. The map of frequent keywords confirms that much focus has been on reviews and social media, but its real advantage lies in its usefulness in identifying new research opportunities and possibilities. Thus, it can help researchers find areas where they can contribute the most to the scientific community.

Second, to gain insights beyond the keywords assigned by authors and editors, Latent Dirichlet Allocation was applied on the titles and abstracts of the publications. LDA clustered the publications into topics similar to those found by the keyword co-occurrence maps, which confirms the focal points of the sentiment research.

One genre of social media seems to be almost entirely unused for sentiment analysis: discussion forums. This is quite surprising, as practically any conceivable topic is discussed online. Furthermore, people often disagree, and thus, sentiments are bound to be found in them. Another source of sentiments that is surprisingly underutilized is news, probably partly due to the issues discussed in the introduction. Even if news are assumed to be objective, the reality is that they seldom are. By comparing how different magazines or websites report on the same events or by comparing countries or regions, we are bound to find significant differences. So far, most of the focus on sentiments in news has been on either finding option holders and extracting quotes [2] or on financial implications of the news [21].

6 Acknowledgment

I appreciate Bikesh Upreti for his invaluable help with the data acquisition and the LDA model. I would also like to thank Anton Frantsev for his input during the finalization of this work.

References

[1] Balahur, A., Steinberger, R., van der Goot, E., Pouliquen, B., and Kabadjov, M. (2009): Opinion Mining on Newspaper Quotations. Proceedings of the workshop Intelligent Analysis and Processing of Web News Content (IAPWNC), held at the 2009 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 523-526.

[2] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2010): Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), pp. 2216-2220.

[3] Bar-Hillel, Y. and Carnap, R. (1953): Semantic Information The British Journal for the Philosophy of Science Vol. 4, pp. 147-157

[4] Blei, D., Ng, A., and Jordan, M. (2003): Lafferty, John, ed. Latent Dirichlet allocation. Journal of Machine Learning Research 3 (4-5): pp. 993-1022.

[5] Blei, D., and Lafferty, J. (2009): Topic models. Text mining: classification, clustering, and applications 10, (2009), 71.

[6] Pouliquen, B., Steinberger, R., and Best, C. (2007): Automatic Detection of Quotations in Multilingual News. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP2007), pp. 487-492.

[7] Carnap, R. (1950): Empiricism, Semantics, Ontology, Revue Internationale de Philosophie 4: pp. 20-40.

[8] Chaney, A., and Blei, D. (2012): Visualizing Topic Models. ICWSM.

[9] Dave, K., Lawrence, S., and Pennock, D. M. (2003): Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in Proceedings of WWW, pp. 519-528.

[10] Garas, A., Garcia, D., Skowron, M., and Schweitzer, F. (2012): Emotional persistence in online chatting communities. Scientific Reports, 2, article 402.

[11] Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. (2005): The predictive power of online chatter. In R. L. Grossman, R. Bayardo, K. Bennett and J. Vaidya (Eds.), Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD'05, pp. 78-87

[12] Hall, D., Jurafsky, D., and Manning, C.D. (2008): Studying the History of Ideas Using Topic Models. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 363-371.

[13] Harb, A., Plantie, M., Dray, G., Roche, M., Trousset, F., and Poncelet, P. (2008): Web opinion mining: how to extract opinions from blogs?. In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology CSTST'08. ACM, New York, NY, USA, pp. 211-217

[14] Hatzivassiloglou, V., and Wiebe, J. (2000): Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the 18th conference on Computational linguistics-Volume 1, Association for Computational Linguistics, pp. 299-305

[15] He, Q., Chen, B., and Giles, C.L. (2009): Detecting Topic Evolution in Scientific Literature: How Can Citations Help, Cikm, pp. 957-966.

[16] Hintikka, J. (1970): On semantic information. In information and inference Synthese, Library, Reidel, Dordrecht, The Netherlands.

[17] Hu, M., and Liu, B. (2004): Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 168-177

[18] Jansen, B.J., Zhang, M., Sobel, K., and Chowdury, A. (2009): Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 60(11), pp. 2169-2188.

[19] Kucuktunc, O., Cambazoglu, B.B., Weber, I., and Ferhatosmanoglu, H. (2012): A large-scale sentiment analysis for Yahoo! Answers, Proceedings of the 5th ACM International Conference on Web Search and Data Mining.

[20] Li, S., and Zong, C. (2008): Multi-domain sentiment classification. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Association for Computational Linguistics, pp. 257-260.

[21] Malo, P., Sinha, A., Takala, P., Ahlgren, O., and Lappalainen, I. (2013): Learning the Roles of Directional Expressions and Domain Concepts in Financial News Analysis. In: Proceedings of IEEE International Conference on Data Mining Workshops (SENTIRE-2013): IEEE Press.

[22] Neviarouskaya, A., Prendinger, H. and Ishizuka, M. SentiFul: A Lexicon for Sentiment Analysis. In Proceedings of the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. (2009) pp. 1 - 6.

[23] Newman, D., Noh, Y., Talley, E., Karimi, S., and Baldwin, T. (2010): Evaluating Topic Models for Digital Libraries Categories and Subject Descriptors. JCDL, pp. 215-224.

[24] Pang, B. and Lee, L. (2005): Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. Proceedings of the Association for Computational Linguistics (ACL), pp. 115-124

[25] Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2), pp. 1-135.

[26] Pfitzner, R., Garas, A., and Schweitzer, F. (2012): Emotional divergence influences information spreading in Twitter, ICWSM-12.

[27] Stone, P. and Hunt, E. (1963): A computer approach to content analysis: studies using the General Inquirer system. In Proceedings of the May 21-23, 1963, spring joint computer conference (AFIPS '63 (Spring)): ACM, pp. 241-256.

[28] Takala, P., Malo, P., Sinha, A., and Ahlgren, O. (2014): Gold-standard for topic-specific sentiments in economic texts. To appear in: Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC-2014), 26-31 May, Reykjavik, Iceland.

[29] Thelwall, M. and Prabowo, R. (2007): Identifying and characterising public science-related concerns from RSS feeds. Journal of the American Society for Information Science and Technology, 58(3), pp. 379-390.

[30] Thelwall, M., Buckley, K., and Paltoglou, G. (2011): Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 62(2), pp. 406-418.

[31] Thelwall, M., Buckley, K., and Paltoglou, G. (2012): Sentiment strength detection for the social Web, Journal of the American Society for Information Science and Technology, 63(1), pp. 163-173.

[32] Turney, P. (2002): Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL02).

[33] Van Eck, N. J., Waltman, L., Van den Berg, J., and Kaymak, U. (2006): Visualizing the computational intelligence field. IEEE Computational Intelligence Magazine, 1(4), pp. 6-10.

[34] Van Eck, N., and Waltman, L. (2007): VOS: A new method for visualizing similarities between objects. In H.-J. Lenz & R. Decker (Eds.), Advances in data analysis: Proceedings of the 30th annual conference of the German Classification Society (pp. 299-306). Heidelberg: Springer.

[35] Van Eck, N. J., and Waltman, L. (2007b). Bibliometric mapping of the computational intelligence field. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 15(5), pp. 625-645.

[36] Van Eck, N. J., and Waltman, L. (2010): Software survey: VOSviewer, a computer program for bibliometric mapping, Scientometrics 84, pp. 523-538

[37] Weber, I, Ukkonen, A., and Gionis, A. (2012): Answers, not links: extracting tips from yahoo! answers to address how-to web queries, Proceedings of the fifth ACM international conference on Web search and data mining (WSDM12).

[38] Wiebe, J. (1994): Tracking point of view in narrative. Computational Linguistics, 20.

[39] Wiebe, J. and Rilo, E. (2005): Creating subjective and objective sentence classifiers from un-annotated texts. In Proceedings of CICLing 2005, pp. 486-497.

[40] Yang, H., Si, L., and Callan, J. (2006): Knowledge transfer and opinion detection in the TREC 2006 blog track. In Proceedings of TREC 2006, vol. 120

[41] Yang, C., Hsin-Yih Lin, K., and Chen, H. (2007): Emotion classification using web blog corpora. In WI'07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 275-278

[42] Yu, H. and Hatzivassiloglou, V. (2003): Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, pp. 129-136