Communication of new energy forms: Ways to detect topics and stakeholders

Draft to appear in Proceedings of the *IEEE International Professional Communication Conference (ProComm) 2015*

Bianka Trevisan
RWTH Aachen University
b.trevisan@tk.rwth-aachen.de

Claas Digmayer
RWTH Aachen University
c.digmayer@tk.rwth-aachen.de

Eva Reimer
RWTH Aachen University
e.reimer@tk.rwth-aachen.de

Eva-Maria Jakobs
RWTH Aachen University
e.m.jakobs@tk.rwth-aachen.de

This paper addresses professional communication experts who develop communication concepts for companies and the introduction of new technologies, e.g., in the energy sector. It presents an approach exploiting social media applications (Facebook®, blogs) as an information resource for the communication of technologies by detecting information on relevant topics and stakeholders. For the topic and stakeholder detection, text mining-methods are applied. As an application example, new energy forms serve. The results show exemplarily that social media applications, such as blogs provide a wealth of information on relevant topics in public discourse, while social networking sites such as Facebook® are particularly useful for the detection of stakeholder profiles. The proposed approach can help communication experts to retrieve and gain information, for instance, as input for communication concepts and strategies.

Introduction

The paper is aimed at professional communication experts who develop communication strategies and concepts for companies and the introduction of new technologies; in this case, the topic is *new energy forms.* The objectives of the strategy can be to introduce a technology in principle or to promote related plants, infrastructure, products, and services. A main assumption of this paper is that technology-related communication strategies, e.g., if they are part of a PR campaign, must build on a profound understanding of the audience and, therefore, research is required. In this context, relevant questions are as such: Who are my audience members? What do I know about them? What do they already know and think about the given technology? What are ongoing discussions concerning the respective topic and related arguments (pros and cons regarding a technology)? Are there already stereotypes and misunderstandings anchored in the public?

However, research is both time-consuming and expensive, particularly if traditional approaches are used, such as interviews, focus groups, or media analysis. This paper describes new ways to search for relevant information by analyzing ongoing topic-related discussions in the Internet using new text technologies, such as text and Web mining. The

following discussion focuses on two often by the public used social media applications–blogs and social networking sites. The discussion is guided by two research questions:

RQ1: How to identify and understand public perspectives on a new technology in terms of often discussed topics?

RQ2: How to identify target groups by analyzing public discourse?

The discussion is based on two studies–a topic study and a stakeholder study–focusing on new energy forms as application example. In the topic study, German blog comments are analyzed identifying frequently discussed topics of smart grid. For the analysis, text mining-methods are applied (techniques from Computer Science refined by linguistic approaches). The stakeholder study examines Facebook® pages on deep geothermal energy, wind energy, and solar energy. It aims to identify stakeholders who use this place in the Web to talk about the aforementioned technologies (group profiles) and, moreover, detect which target groups dominate technology discussion.

Related Work

For this study, literature is presented in three areas. First, approaches in topic detection are described. Second, state of the art in stakeholder analysis is reported. Third, text mining-methods are introduced.

I. Topic detection

The research field of *topic detection and tracking (TDT)* is concerned with the development of technologies that make it possible (1) to gather news from the Internet, (2) to separate these news into individual text snippets, and (3) allocated these snippets to a specific topic [1]. There are a number of approaches in computational linguistics, which deal with topic detection on the Web such as [2] or the approaches in [3]. Moreover, some studies focus on the retrieval of topics in specific social media applications, for instance, [4] for Twitter® or [5] for news and blog corpora.

All these approaches have in common that they are based on a coarse-grained analysis at the document and text level, i.e., texts or sentences as a whole are assigned to a topic. Only a few approaches use a fine-grained analysis technique to assign text snippets on the word and phrase level to topics and, thereby, allow for in-depth analysis. Previous work presents an approach for TDT that is based on simple semantics and categorizes topic-specific terms from text into four groups of terms (proper names, locations, temporal expressions, and normal terms) that have the same meaning [6]. Moreover, [7] provides a more detailed approach with six different topic categories. A further approach is published by [8], where topic detection is performed without prior knowledge on topic categories. Here, the extraction and clustering of topic-specific terms is based on similarity measures. The major disadvantage of this approach is that a large amount of data is needed before any statement regarding the text content and topics can be made.

II. Stakeholder analysis

Stakeholder analysis serves for the detection of individuals or groups (1) who affect an action (e.g. a project or event) or (2) which are affected by an action [9]. Moreover, the analysis aims at the identification of stakeholders' attitudes towards an action. Usually, the information is used to set which stakeholders have to be addressed in during particular

action phases, such as project plans, policies, or programs. Depending on the degree of involvement, different types of stakeholders are distinguished.

The detection of stakeholders is examined in different disciplines and research areas. Usually, the database consists of text data such as emails [10] or blog comments [11]. The method portfolio ranges from manual stylometric analysis as performed in traditional forensic linguistics to automated stylometric analysis in computer science and computational linguistics, which uses (statistical) text mining-methods and techniques [12]. Applying these methods, [13] determines style- and content-based features in English blog texts providing evidence on the gender and age of an author. Other approaches in stakeholder analysis originate primarily from the management area, but do not use text data for stakeholder identification, e.g., [14]. Rather, the self-reports of the stakeholders are in most studies used as database, as it is also common in traditional empirical social research.

III. Text Mining

Text mining refers to a methodology for the semi-automatic text analysis. It is based on computer science methods, and it aims to identify relevant topics in text data, e.g., relevant topics in technology discussions [15]. Thereby, text mining has the advantage that large amounts of data (> 1 million words) can be studied with relatively little effort. Moreover, it allows eliciting authentic and unaltered data that reflect users' topics of interest.

In market and opinion research, text mining-methods have been applied for many years [16]. They are primarily used for the text analysis of user-generated content from social media applications, for instance, in the context of *social media monitoring* [17]. Moreover, text mining-methods are adopted for topic analysis [18], site analysis |19], or stakeholder analysis [15] in similar research projects.

Summary

The presented approaches originate predominantly from computational linguistics and computer science; their implementation requires a high degree of programming skills. Nevertheless, the approaches provide also for non-computer scientists and cases useful hints for topic and stakeholder detection, such as the approaches described by [4] and [11].

In the following, we propose an innovative approach that offers fine-grained text mining-methods and -techniques for the exploitation of user-generated content from social media applications as information resource. The approach consists of two separate studies: a topic study and a stakeholder study. It allows professional communication experts to retrieve, collect, and identify relevant information in technology discussion efficiently from social media applications and prepare these for communicative purposes, such as communication concepts.

Topic study

In the topic study, comments from a German topic-specific blog are analyzed (*text data*). *Blogs* and similar social media applications (e.g. forums, communities) serve as social market places where people meet and chat about topics they are interested in [20]. People post content with the intention of documenting their life, providing commentary and opinions, expressing deeply felt emotions, articulating ideas through writing, forming and maintaining community forums [21]. Having this wealth of information, relevant topics in public discourse

can be detected. The methodological process of data collection, preparation, and analysis in topic detection is explained below.

I. Data collection

For the topic study, it is searched for topic-specific pages whose subject are *smart grids*. Thereby, a key search criterion was to find a sufficiently large amount of user-generated content in order to make valid statements about comments' content. Finally, the German news site www.heise.de was chosen, which contains several blogs and forums. In www.heise.de, several topic-relevant news articles were identified by keyword search from which one article was selected; the selection criterion was the amount of comments per article. Comments related to the article (text data) are collected from the source code automatically (time period: 2011/08/15 to 2011/08/20). Finally, the corpus contains 212 German topic-specific comments with approximately 53,000 tokens dealing with the topic *smart grid*.

II. Data preparation

For each elicited comment (text data), a .txt-file is created. In the first step of data processing, the .txt-files are Part-of-Speech (PoS) tagged. *PoS tagging* is the initial and fundamental step in computational linguistics' data preparation; each additional processing step is based on the added information in this step. Thereby, to each unit of the text (*token* or space-separated unit, e.g. word, punctuation mark, emoticon) a morpho-syntactic category is assigned (e.g. NN=normal noun, NE=named entity) according to the STTS-tagset [22]. In total, 54 categories were distinguished.

In the second step of data processing, the PoS tagged data was imported into the tool EXMARaLDA. The tool is primarily used as an analysis tool for discourse analysis. In this study, the tool was applied for multi-level annotation of text data, i.e., the assignment of different information to the same text snippet. Afterwards, all topic-specific terms were searched, identified, and classified in the text on the lexical-semantic level. The underlying classification or *annotation schema* includes eight categories as depicted in Table 2. Six categories or *tags* (PER, ORG, LOC, VEH, FAC, GPE) are taken from the scheme of [7], two additional categories (EL, PROD) are added due to topic-specification [see also 14]. The categorization is based on a topic-specific lexicon that lists relevant terms of a topic area (e.g. *smart grid*) and their linguistic variants (e.g. *Smart Grid*, *smart-grid*). The listed words originate from texts, such as topic-related brochures, articles, or books and have been identified and listed manually.

TABLE 2. ANNOTATION SCHEMA [23].

Type	Tag	Sample categories
Topic-specific term	EL*	Noun, adjective or verb without further specification
Facility	FAC	Bridges, buildings,

Type	Tag	Sample categories
		airport
Geo-political entity	GPE	Countries, states, provinces, counties
Location	LOC	Physical extents, mountains, lakes, seas
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
People	PER	Individuals, fictional characters, small groups
Product	PROD*	Wares and services
Vehicles	VEH	Planes, trains and automobiles

In the third step, the categorized terms are complemented with information about their affiliation to a lexical field. A *lexical field* brings together all semantically related terms in one category, e.g., *biomass* is a term in the lexical field *resource* [24],[25]. Figure 1 shows an excerpt of the processed data in EXMARaLDA.

	173	174	175	176	177	178
X [txt]	Ein	Einfamilienhaushalt	hat	eine	monatliche	Stromrechnung
X [pos]	ART	NN	VAFIN	ART	ADJA	NN
X [lemma]	ein	Einfamilienhaushalt	haben	ein	monatlich	Stromrechnung
Lexical-semantic level		EL				EL
Lexical field		Haushalt				Kosten

FIGURE 1. EXCERPT OF THE PROCESSED DATA.

The complete sentence of the example in Figure 1 is this: *Ein Einfamilienhaus hat eine monatliche Stromrechnung von überschlagen um die 100 – 200 Euro. / A single family house has a monthly electricity bill of around 100-200 euro.* In the sentence, the topic-specific terms (tag: EL) *Einfamilienhaus* (*single family house*) and *Stromrechnung* (*electricity bill*) occur. In this case, the term *Einfamilienhaus* (*single family house*) is part of the lexical field *Haushalt* (*household*) and the term *Stromrechnung* (*electricity bill*) is part of the lexical field *Kosten (costs)*.

III. Data analysis

In the step of data analysis, the occurrence of tags (on the lexical-semantic level), topic-specific terms, and lexical fields are determined. The results are displayed in absolute frequency.

IV. Results

First, the weighting of assigned tags is examined. The results show that general topic-specific terms (EL, n=1781) are mentioned most often. References to specific products (PROD, n=126) or organizations (ORG, n=54) as well as to geo-political entities (GPE, n=94) and places (LOC, n=20) are rare. These results suggest that the discussion of smart grids in the present text data is carried out on a more general level. The second highest ranked category is people (PER, n=175) that is predominantly represented by blogger names, such as *Prigogyne* or *HeWhosePathIsChosen*. These names meet the aforementioned purpose of personnel concealment and do not have any topic relevance; consequently, they can be neglected. Moreover, mentions of facilities (FAC, n=5) and vehicles (VEH, n=4) are also negligible due to low occurrences. The result overview is given in Table 3.

TABLE 3. TAG RANKING.

Tag	**n**
EL	1781
PER	175
PROD	126
GPE	94
ORG	54
LOC	20
FAC	5
VEH	4

Second, it is examined which topic-specific terms or “hot topics” are mentioned most frequently. The Top-10 of the results includes only terms of the category EL (see Table 3). The results show that by far the most common term is *electricity* (n=128). Hereafter, the terms *block-type thermal power plant* (n=70), *price* (n=65), *energy* (n=64), and *system* (n=54) follow. The entire results are summarized in Table 4.

TABLE 4. TOP-10 RANKED TOPIC-SPECIFIC TERMS.

Term	**n**
electricity	128
block-type thermal	70
price	65
energy	64
system	54
expensive	50
consumer	46
save	43
operate	37
oil	37

Third, the occurrences of lexical fields are counted. The results show that among the top-10-ranked, the lexical field *COSTS* (n=464) is taking the far largest share. It is followed by the lexical field *ECONOMY* (n=388), which is–similar to the lexical field *COSTS*–primarily dedicated

to economic and monetary issues. Hence, both highest ranked lexical fields deal with topics, which concern the aspect *cost-benefit-ratio*. The following lexical fields focus on topics, such as energy (*ELECTRICITY*, n=290; *HEATING & WARMTH*, n=204), natural resources (*RESOURCE,* n=272), energy usage (*CONSUMPTION,* n=236; *DEVICE & MACHINE*, n=210), and energy production (*POWER PLANT*, n=220; *GENERATION*, n=168; *PROVIDER*, n=164). The results are depicted in Table 5.

TABLE 5. TOP-10 RANKED LEXICAL FIELDS.

Lexical field	Term example	n
COSTS	price, tariff, expensive	464
ECONOMY	market, capital, investment	388
ELECTRICITY	green / off-peak electricity	290
RESOURCE	biomass, coal, oil	272
CONSUMPTION	energy wasters	236
POWER PLANT	coal-fired power plant	220
DEVICE & MACHINE	fridge, dryer, cell phone	210
HEATING & WARMTH	waste heat, pellet heating	204
GENERATION	power / heat production	168
PROVIDER	electricity supplier	164

Stakeholder study

In the stakeholder study, a new approach is chosen: the analysis of Facebook® data for stakeholder detection (*metadata*). Facebook® is a common social networking site that is daily visited by 28 million users in Germany [26]. In Facebook®, comment authors' leave information about personal details, such as their age, sex, or marital status that allow for the reconstruction of user or stakeholder profiles. The process of data collection, preparation, and analysis differs from the methodology of the topic study since metadata require specific methodological procedures.

I. Data collection

To build up a data corpus for the stakeholder study, it was searched for topic-specific German Facebook® pages dealing with one of the three kinds of renewable energies: deep geothermal energy, wind energy, and solar energy. The key criterion of the data collection was that the topic of the page should relate exclusively to one of the mentioned renewable energies [15].

When a topic-specific Facebook® page is identified that meets the key criterion, the entire self-reported metadata of the site's members or subscribers are fully collected. The metadata is collected from the source code, automatically (time period: 2009-2014). Finally, the corpus contains Facebook® data from 795 users and 15 different Facebook® pages (e.g. news sites). A predominantly number of Facebook® pages (n=13) represents the topic deep geothermal energy. The elicited corpora are summarized in Table 1.

TABLE 1. CORPUS OF THE STAKEHOLDER STUDY (WWW.FACEBOOK.DE, COMMENTS FROM 2009 TO 2014).

	Deep geothermal energy	**Solar energy**	**Wind energy**
Σ t	2009-2013	2013-2014	2009 – 2013
Pages	13	1	1
Users	634	122	39

II. Data preparation

The collected metadata are stored in plain text (.xml). The metadata (file) provides self-reported information of the user, such as information about user's age provided by the birthdate (e.g. 4.1.1983) or user's sex of involved stakeholders (male vs. female). The metadata are quoted as type-token ratio. Figure 2 shows an excerpt of a metadata .xml-file.

```
<file>FB_Linksammlung_Erneuerbare_Energien_2013-05-23_04-07-23_350_article</file>
<project>FuEne</project>
<textsorte>Article</textsorte>
<media>Website</media>
<method>FBParser</method>
<source>
    <name>Facebook</name>
    <url>../FuEne/Facebook/Geothermie/Linksammlung_Erneuerbare_Energien.html</url>
</source>   <date>2013-05-23 04-07-23</date>
<timestamp>1369318043</timestamp>
<created>2013-11-12 04-58-28</created>
<topic>Energy</topic>
<title>(2) Linksammlung Erneuerbare Energien</title>
<author>
```

FIGURE 2. EXCERPT OF A METADATA FILE.

For the subsequent metadata analysis, each .xml-file is imported in a .csv-file.

III. Data analysis

The metadata are analyzed quantitatively. The aim of the quantitative analysis was to determine the distribution of sex, age, marital status, and education status, based on the self-reported information from the users' profiles.

IV. Results

In the stakeholder study, four metadata types were analyzed: distribution of sex, age, marital status, and education status. The results overview is given in Table 6.

TABLE 6. STAKEHOLDER PROFILES (RESULTS DISPLAYED IN ROUNDED PERCENTAGE).

		Deep geothermal energy	**Solar energy**	**Wind energy**
Sex	m	69	94	83
	f	31	6	17
Age	-20	3	22	n too small
	20-40	60	25	
	40-60	31	47	
	60+	5	6	
Marital status	single	20	37	
	liaised	24	30	
	married	54	33	
Education status		high school graduation +		

Sex: The results of the stakeholder study show that in the given sample mostly members, which identified themselves as men, are subscribers of Facebook® pages dealing with new energy forms, such as deep geothermal energy, solar energy, and wind energy. In detail, the distribution among the sexes is in this sample as follows: deep geothermal energy (n_m=69%, n_f=31%), solar energy (n_m=94%, n_f=6%), and wind energy (n_m=83%, n_f=17%).

Age: The results regarding the age structure show that most people in the present sample, which are members of Facebook® pages related to deep geothermal energy, report their age in the range from 20 to 40 years (n=60%); therein, a relatively small proportion make people aged under 20 years (n=3%) and over 60 years (n=5%). Focusing on the topic of solar energy, it is evident that the members of the Facebook® page report their age mostly in the age group 40-60 years (47%). Moreover, a relatively high proportion of members reported themselves in the age group of 20 years and minor (n=22%) and in the age group between 20 and 40 years (25%). Lastly, only 6% of the members displayed their age as 60 years and older. Meaningful results on the age structure of Facebook® pages members related to wind energy cannot be derived, due to a low number of members.

Marital status: The results addressing the marital status show that most members of Facebook® pages related to deep geothermal energy report their status as married (n=54%), a much smaller portion as single (n=20%), or liaised (24%). In the case of solar energy, the distribution of the marital status is evenly spread: singles (n=37%), liaised (n=30%), and married (n=33%). Meaningful results regarding the marital status of Facebook® pages members related to wind energy cannot be displayed, due to a low number of members.

Education: On average, the majority of stakeholders from the examined sample have reached–according to their indications–at least the educational status of high school graduation.

Discussion

The present study investigates how to exploit social media applications as an information resource. Getting insights on relevant topics (RQ1) and stakeholder involvement (RQ2) in public discussion on new energy forms is essential for successful communication concepts and strategies, as prepared by professional communication experts and pointed out by [27] and [28]. Therefore, the following section contains a discussion of results and of our methodological approach with regard to their usefulness and applicability.

I. Social media applications as information resource

RQ1: The results show that the proposed approach is useful for the identification of relevant topics and the reconstruction of stakeholders' perspectives on new energy forms. However, cultural conditions and the domain in which or for which a text (*here*: blog comment) is produced can influence the topic discussion, as pointed out by [29]. The influence of cultural conditions addresses different attitudes and approaches to literacy, text, and authorship. The influence of the domain implies special norms, conventions, value systems, and expectations that have to be considered in text production and, furthermore, in text analysis. This applies in the broadest sense for the behavior within the Internet and the different social media applications. The influence of cultural conditions and the domain can determine the selection of the topics discussed.

In addition, to get more insights about relevant topics in public discussion on *new energy forms*, more text data from different social media applications need to be considered. For the present approach, this could be, for instance, text data from blogs as well as from news sites, Facebook®, forums, etc. In doing so, different stakeholder perspectives are merged and interconnected. Finally, a more holistic result can be gained.

RQ2: The results lead to the assumption that the discussed topic determines the composition of the stakeholder group. As previous studies showed, particularly young users (82% of 14-29-year-old Internet users) use Facebook® [30]. With regard to the results of the present stakeholder study, it is evident that the reported age structure of the users in the topic area *new energy forms* differs of the general reported age structure in Facebook®. The target groups in our study are mostly aged 30 years and older (87,7%). Moreover, it is expected that the profiles and, in particular, the age structure varies on Facebook® pages that focus on complete opposite topics, e.g., on topics such as *family* or *sports*. Professional communication experts that make use of Facebook® as an information resource for stakeholder analysis should be aware of this fact.

Regarding the given stakeholder study, it should be noted that only the information shared or reported by the user (*self-reported information*) can be collected. For instance, in the present data only a binary encoding of sex or gender is detected. This is due to the fact that in the German version of Facebook® only two genders are suggested (male vs. female) contrary, for instance, to the 59 gender options in the American version of Facebook® [31]. Hence, it cannot be determined whether the user by selecting male or female indicates his biological sex *or* gender.

The uncertainty about the correctness and authenticity of shared information applies also to the rest of the profile or demographic data. Nevertheless, it must be emphasized that the provided information amount could not be found in any other social media application, which strictly indicates for the use of Facebook® as information resource. As long as no alternative social networking sites find it's way onto the Web that holds as much demographic information as Facebook®, Facebook® remains for the collection of stakeholder profiles the benchmark.

II. Methodological approach

RQ1: Regarding methodological constraints, simple counts of topic-specific terms deliver only first insight into relevant topics. Rather, topic-specific terms should be summarized into lexical fields as proposed in the given approach. For instance, the term *electricity* is a high ranked topic-specific term (see Table 4: *electricity*, n=128). Nevertheless, the lexical field *ELECTRICITY* (n=290) containing the term *electricity* only reaches the third rank in Table 5. Here, the lexical field *COSTS* gets the first place in the ranking.

Additionally, a deeper understanding of stakeholder-relevant topics can be gained by in depth-analysis, focusing on the identification of term polarities. Therefore, sentiment analysis techniques can be applied that allow for the differentiation of positive or negative connotated terms and topics. Hence, considering stakeholders' attitude towards these relevant topics, communication experts can tailor communication strategies and concepts conform to the public requirements and opinions.

RQ2: Approaches and methods for the indication of sex (or gender), age, or education status by linguistic indicators are on the rise, as indicated before. When these approaches achieve a sufficient degree of maturity, also other social media applications with missing self-reported metadata (e.g. text data in blogs) can be used for stakeholder detection. In this case, information from text data and metadata can be gathered in combination and correlated with each other. As a result, an enormous methodological gain would be achieved.

Outlook

The described approach provides for a number of research questions and fields connecting points, e.g., for purposes of risk communication [32] or acceptance research. An extension of the presented approach through the integration of methods and techniques from opinion detection would be desirable. A combination of stakeholder analysis, topic detection, and opinion detection would enable professionals to get deeper insights into the perception and evaluation of new technologies and, consequently, allow for a more adequate communication. The combination of the three approaches is promising but requires further research.

Acknowledgements

Funded by Bundesministerium für Bildung und Forschung, Forschungscampus Elektrische Netze der Zukunft (FKZ 03FS0488). Moreover, thanks go to Tim Hemig for technical support.

References

[1] W.G. Stock, “Themenentdeckung und -verfolgung und ihr Einsatz bei Informationsdiensten für Nachrichten,” *Information – Wissenschaft und Praxis*, vol. 58, no. 1, pp. 41–46, 2007.

[2] S. Chakrabarti, *Mining the Web: Discovering knowledge from hypertext data*. San Francisco: Elsevier, 2003.

[3] J. Allan, *Topic Detection and tracking: Event-based information organization.* Norwell: Kluwer Academic Publishers, 2002.

[4] M. Cataldi et al. “Emerging topic detection on Twitter based on temporal and social terms evaluation,” in *Proceedings of the 10th ACM International Workshop on Multimedia Data Mining (MDMKDD '10),* 2010, 10 pages.

[5] L.-W. Ku et al. “Opinion extraction, summarization and tracking in news and blog corpora,” in *Proceedings of the Spring Symposia on Computational Approaches to Analyzing Weblogs (AAAI-CAAW’06),* 2006, 8 pages.

[6] J. Makkonen et al., “Simple semantics in topic detection and tracking,” *Information Retrieval*, vol. 7, no. 3–4, pp. 347–368, 2004.

[7] D. Jurafsky and J.H. Martin, *Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing.* New Jersey: Prentice Hall, 2009.

[8] C. Wartena and R. Brussee “Topic detection by clustering keywords” in *Proceedings of the 19th International Workshop on Database and Expert Systems Application (DEXA '08)*, 2008, 54–58.

[9] J. W. Weiss, *Business Ethics: A stakeholder and issues management approach.* San Francisco: Berrett-Koehler Publishers, 2014.

[10] W. Deitrick et al., “Author gender prediction in an email stream using neural networks,” *Journal of Intelligent Learning Systems and Applications,* vol. 4, no. 3, pp. 169–175, 2012.

[11] S. Goswami et al. “Learning age and gender of blogger from stylistic variation,” in *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence (PReMI’06),* 2009, pp. 205–212.

[12] T. McEnery and M. Oakes, “Authorship identification and computational stylometry,” in *Handbook of Natural Language Processing,* R. Dale et al., Ed. New York: Decker, 2000, pp. 545–562.

[13] S. Argamon et al., “Mining the blogosphere: Age, gender and the varieties of self-expression,” *First Monday* vol. 9, no. 12, 2007.

[14] M. Welch and P. R. Jackson, "Rethinking internal communication: a stakeholder approach," *Corporate Communications: An International Journal*, vol. 12, no. 2, pp. 177–198, 2007.

[15] B. Trevisan et al., “Facebook as a source for human-centred engineering. Web Mining-based reconstruction of stakeholder perspectives on energy systems,” in *Advances in Human Factors, Software, and Systems Engineering,* B. Amaba and B. Dalgetty, Ed. Krakau: AHFE Conference, 2014, pp. 180–191.

[16] S.D. Rappaport, *Listen first!: Turning social media conversations into business advantage.* New Jersey: Wiley & Sons, 2011.

[17] J. Sterne, *Social media monitoring. Analyse und Optimierung Ihres Social Media Marketings auf Facebook, Twitter, YouTube und Co.* Heidelberg: mitp, 2011.

[18] B. Trevisan and E.-M. Jakobs “Talking about mobile communication systems. Verbal comments in the Web as a source for acceptance research in large-scale technologies,” in *Proceedings of the IEEE International Professional Communication Conference (IPCC’10), 2010,* pp. 93-100.

[19] B. Trevisan et al. “Web comment-based trend analysis on deep geothermal energy,” in *Proceedings of the IEEE International Professional Communication Conference (IPCC’10), 2013,* 8 pages.

[20] B. Pang and L. Lee, *Opinion mining and sentiment analysis.* Boston: Now Publishers, 2008.

[21] B. A. Nardi et al., “Why we blog,” *Communications of the ACM*, vol. 12, no. 47, pp. 41-46, 2004.

[22] A. Schiller et al., *Guidelines für das Tagging deutscher Textcorpora mit STTS*. University of Stuttgart, 1999.

[23] B. Trevisan, *Bewerten in Blogkommentaren*. RWTH Aachen University, 2014.

[24] H. Geckeler, *Strukturelle Semantik und Wortfeldtheorie*. München: Wilhem Fink Verlag, 1971.

[25] E. Coseriu, *Einführung in die strukturelle Betrachtung des Wortschatzes*. Tübingen: Gunter Narr, 1970.

[26] Das Statistik-Portal - Statistiken und Studien aus über 18.000 Quellen. [Online]. Available: http://www.statista.com.

[27] E. Reimer et al., “New ways to develop professional communication concepts,” in *Proceedings of the Professional Communication Conference (ProComm’15),* 2015.

[28] S. Wirtz-Brückner et al., “The potential of Facebook for communicating complex technologies using the example of deep geothermal energy,” in *Proceedings of the Professional Communication Conference (ProComm’15),* 2015.

[29] E.-M. Jakobs, „Textproduktion als domänen- und kulturspezifisches Handeln,“ in *Domänen- und kulturspezifisches Schreiben*, K. Adamzik et al., Ed. Frankfurt am Main: Lang, 1997, pp. 9–30.

[30] ARD/ZDF-Onlinestudien. [Online]. Available: http://www.ard-zdf-onlinestudie.de.

[31] Facebook's new gender option: Fill in the blank. [Online]. Available: http://www.usatoday.com/story/tech/2015/02/26/ facebook-gender-option-fill-in-the-blank/24059551/.

[32] C. Digmayer et al. “Medusa and Pandora meet the Web 2.0: How risk types influence the communication in social media”, in *Proceedings of the Professional Communication Conference (ProComm’15),* 2015.