The Information Value of Unstructured Analyst Opinions

–

Studies on the Determinants of Information Value and its Relationship to Capital Markets

Dissertation zur Erlangung des wirtschaftswissenschaftlichen Doktorgrades der Wirtschaftswissenschaftlichen Fakultät der Georg-August-Universität Göttingen

Vorgelegt von: Matthias Eickhoff, M.Sc.

Göttingen, 2017

Betreuungsausschuss

Erstbetreuer	Prof. Dr. Jan Muntermann
Zweitbetreuer	Prof. Dr. Matthias Schumann
Drittbetreuer	Prof. Dr. Lutz M. Kolbe

Table of Contents

List of Figures

List of Tables

Abbreviations

AN ..... Analyst Call Participant
API..... Application Programming Interface
BV ..... Book Value
CAPM..... Capital Asset Pricing Model
CEO ..... Chief Executive Officer
CFO ..... Chief Financial Officer
CORP..... Corporate Presentation Participants
CTM ..... Correlated Topic Model
$DJIA_{30}$..... The Dow Jones Industrial Average
DSR ..... Design Science Research
EPS ..... Earnings per Share
HDP ..... The Hierarchical Dirichlet Process
I/B/E/S ..... Institutional Brokers' Estimate System
ICT..... Information and Communication Technology
IS(R) ..... Information Systems (Research)
LASSO ..... Least absolute Shrinkage and Selection Operator
LDA ..... Latent Dirichlet Allocation
LMD ..... The Loughran & McDonald 10K Sentiment Dictionary
LSA ..... Latent Semantic Analysis
LSI ..... Latent Semantic Indexing
MIS ..... Management Information Systems
MRT ..... Media Richness Theory
OLS ..... Ordinary Least Squares
PCA ..... Principal Component Analysis
pLSA ..... Probabilistic Latent Semantic Analysis
PRES ..... Presentation Section of Conference Call
QA ..... Question and Answer Section of Conference Call
R&D ..... Research and Development
ROA..... Return on Assets
ROE ..... Return on Equity
RQ ..... Research Question
SE ..... Standard Error
SOM ..... Self-Organizing Map
SVD ..... Singular Value Decomposition
TDM ..... Term Document Matrix
TF-IDF..... Term Frequency inverse Document Frequency
TRAA ..... Thomson Reuters Advanced Analytics
URL ..... Uniform Resource Locator
WoC..... Wisdom of Crowds
XML ..... Extensible Markup Language

Symbols

Di ..... Document Nr. i
Ti ..... Topic Nr. i
AR(X) ..... Abnormal Return of X
CAR(X) ..... Cumulative abnormal Return of X
E(X) ..... Expectation of X
Epsilon (ϵ) ..... Error term
LDA - w_n ..... A Word
LDA - z_n ..... A Topic
LDA - Θ ..... Per Document Topic Distribution
LDA - α ..... Parameter of Dirichlet Prior
LDA - ξ ..... Parameter of Dirichlet Prior
LDA - αC ..... Call Topic Average
LDA - β- ..... Pre-Call Report Topic Average
LDA - $\beta +$ ..... Post-Call Report Topic Average
Mu (μ) ..... Mean
Sigma (σ) ..... Mean
Var(X) ..... Variance of X
$w_{i,j}$ ..... Word in category i, Nr. j
z-score (z) ..... A centered time series

A. Foundations

In the first part of this cumulative thesis, the foundations relevant to the individual research contributions presented in part B are developed. This includes the motivation of the presented research, the development of research questions, and an overview of the relevant literature, theories, methods, as well as the data used throughout this thesis. See section 3 for a detailed overview of the structure of this thesis.

1 Motivation

The role of information and communication technology (ICT) in the financial industry has recently undergone a fundamental evolution. Traditionally, computer systems served as a method of data storage intended to support firms' ongoing operations. As digital technologies evolve, ICT's functions shift towards a more active role in the firm. Two examples of this shift in the financial industry are given by the sector's recent interest in block chain-based distributed databases (Cohen et al., 2016) and the adoption of digital business strategies by incumbent financial industry firms, which intended to react to new technology-based competitors (Sia et al., 2016).

These changes impact many aspects of the business processes of financial market participants. This thesis focuses on the increased need for decision support that this changing role of ICT entails. Again, this need for decision support concerns many facets of a bank's business. One of these facets is given by the increased data volume resulting from the adoption of digital technologies. While this increase in available data creates opportunities for industries (Chen et al., 2012) and research (Rai, 2016), it also creates new challenges for decision makers. This is particularly true when copious quantities of unstructured, often textual, data become decision relevant. A growing body of research shows the relevance of this content type regarding investment decisions (Bollen et al., 2011; Li, 2010b; Nofer and Hinz, 2015).

Stock analysts play a significant role in information dissemination in financial markets. The analysis of their recommendations (Brown and Rozeff, 1978) and market forecasts in general (Cowles 1933) have a long tradition in financial research. More recently, the information content of the unstructured content of analyst reports (Asquith et al., 2005; Frankel et al., 2006; Huang et al., 2015), as well as analyst conference calls accompanying earnings announcements (Mayew et al., 2013), have been studied since these forms of analyst opinion became widely available. Thus, this research stream depends on content analysis methodology to extract metrics from this unstructured content. The focus in this thesis is given by quantitative content analysis using computational methods, in contrast to manual content analysis techniques such as hermeneutics or grounded theory.

Quantitative content analysis predates the development of the first computational systems (Speed, 1893), and it has always been a subject of interest when the development of business intelligence systems was concerned. Indeed, Luhn (1958), to whom originating the term business intelligence is commonly ascribed, refers to the auto abstracting and auto-encoding of documents as two of the key tasks of business intelligence systems.

The ever-increasing volumes of available data and the resulting need for approaches to its analysis have led to the development of many sub tasks within the content analysis domain, solutions to which may be used in decision support systems aimed at coping with this increased demand for data analysis. In this thesis, topic modeling (Blei, 2012) and sentiment analysis (Liu, 2012) are of special interest because these methods attempt to extract what is being said and how it is said. Thus, it is useful to quantify analyst opinion in unstructured data. This thesis builds on the intersection of these financial and methodological research streams. It investigates the drivers of this relevance based on information systems theory and contributes in three principal ways.

This is done first by providing an overview of the business models of FinTech companies, which helps to explain the changing environment financial markets operate within. Second, this research surveys the available research methodology in information systems and other technical disciplines and compares this state of the art with the methodology currently being applied in financial research. Opportunities for the application of content analysis techniques are identified in the financial domain and explored in the presented studies. Third, information systems theory is applied to the financial domain to search for explanations for the usefulness of unstructured content in the context of investment decisions.

These contributions are cumulative. The main area of this investigation is given by the analysis of unstructured analyst opinion and its information value, supported by an analysis of the changing entrepreneurial landscape and methodological contributions that establish the necessary overview of content analysis techniques relevant to the conducted research. The next section provides a more detailed overview of these research areas and develops the individual research questions, which are addressed in the individual contributions of this cumulative thesis.

2 Research Questions

The research questions of this thesis are separated into three principal areas of research. This section provides a brief introduction to the aims of each of these three areas and consequently derives the research questions for these areas.

The first area of interest concerns the entrepreneurial surroundings within which analysts work. Recently, the financial industry has undergone substantial transformation due to the ubiquity of digital technology, which has impacted not only the financial sector but all areas of entrepreneurial activity (Bharadwaj et al., 2013; El Sawy and Pereira, 2013). How this transformation is impacting the financial industry and the business models of this sector is crucial to understanding the impact of analyst opinion.

This digitization of the financial industry has fundamentally changed the creation and reception of analysts' information output. The increase in the means of automation of analysts' workflows and the analysis of their work are of interest when considering the changes within this area of activity. This is why Eickhoff et al. (2017, **paper I.1**) investigate this transformation of the financial industry by developing a taxonomy of digital business models for FinTech startups. This taxonomy enables readers to assess the degree to which ubiquitous digital technologies have changed the competitive landscape within the financial industry by highlighting the threat that such startups pose to incumbent firms. In this context, the first research area of this thesis is proposed.

Research Area I: This research area is concerned with establishing an overview of the entrepreneurial environment, in which the results of the other two research areas are to be understood. It introduces the entrepreneurial environment and the changing technological landscape, which have made it necessary to find innovative ways to analyze data and to remain competitive with increasingly diverse competitors and new analytical demands. The development of a business model taxonomy will show whether analytical business models have gained traction in this industry. This research area answers the following individual research questions:

> **Research Question I.1:** What are the dimensions and characteristics of typical business models of FinTech companies?
>
> **Research Question I.2:** How can these business models be grouped into different FinTech niche markets?

After an overview of the entrepreneurial landscape has been provided by the answers to the research questions posed in research area I, the next logical step towards the analysis of unstructured data within the financial domain is to determine which tools are available to support this task. First, text mining methods are needed to extract information from textual data in a format suitable for further analysis. Second, because the result of text mining methodology is typically given by a large quantity of numeric

data, the information extracted by these methods needs to be analyzed in a manner that provides decision-relevant metrics to decision makers, thus creating a result that is useful. Therefore, research area II addresses these methodological questions.

Research Area II: By what methodology can the information within unstructured content related to capital markets be transformed to be analyzed by traditional statistical methods, and how are these methods currently used in managerial research disciplines to provide meaningful answers to research questions that are relevant in these disciplines? This research section answers the following individual research questions, focusing on the roles of topic modeling and sentiment analysis:

> **Research Question II.1:** What is the state of the art of topic mining methodology used to process unstructured content in the methodological literature, and how are these methods being applied in the managerial sciences to provide meaningful information relevant to researchers and decision makers?
>
> **Research Question II.2:** How can dictionary-based and machine learning-based sentiment analysis be combined to mitigate some of their individual shortcomings, such as the need for labeled training data?

Thus, the output of research area two is constituted by deciding how to apply text mining methodology for the purposes of this thesis. When conducting text mining-based research, this typically constitutes the first half of the analysis. The second part of the analysis is to put the results of the textual analysis to use and produce information, based on which questions can be answered or which are useful to decision makers. Consequently, the third research area addresses the analysis of this information, deriving decision-relevant metrics on its basis and answering questions about the information flow between unstructured business communications and the financial markets.

Research Area III: Based on the results of area II, how can unstructured analyst opinion be analyzed in an informative manner? This research area is concerned both with the processing of unstructured analyst opinion and other information sources related to capital markets, as well as the impact these different media types have on individual companies. This area of research also examines how information systems and business administration theory can be applied to these problems to provide explanations as to why such effects exist. This research area answers the following individual research questions:

> **Research Question III.1:** What structure is there to the relationship between the opinions of social media users and stock analysts, and can wisdom of crowds theory be used to identify the situations in which the crowd or

stock analysts are more likely to provide timely information, reflecting changes in a firm's circumstance?

Research Question III.2: What constitutes a decision-relevant metric in the context of business communications regarding a firm's earnings announcement, and how can the metrics of analyst opinion determined by sentiment analysis and topic modeling be used to provide such decision relevant information?

Research Question III.3: To what extent do the topics contained in analyst reports that are released prior to an earnings call influence the topics contained therein, and does the call influence the content of reports released thereafter?

Research Question III.4: To what extent can the media richness of unstructured analyst opinion, as described by media richness theory, help to explain its effect on post earnings call firm stock returns when compared to information sources of lower richness?

Thus, the three areas of research explored in the context of this thesis are given by first providing an overview of the context within which this research is conducted, followed by establishing the necessary methodological foundations regarding the analysis of unstructured data, and finally by using the results of these methods to investigate questions of theoretical or practical importance. The next section provides an overview of the thesis' structure based on these research areas.

3 Structure of the Thesis

This section provides an overview of the different papers contributing to this cumulative thesis. Figure 1 provides an overview of this thesis. The top third of the figure outlines the foundation section of the thesis (A), in which the research background of the presented contributions is outlined and upon which their research paradigms are elaborated. The center part shows the individual research contributions (B) and their division into three research areas, for which research questions were developed in section 2.

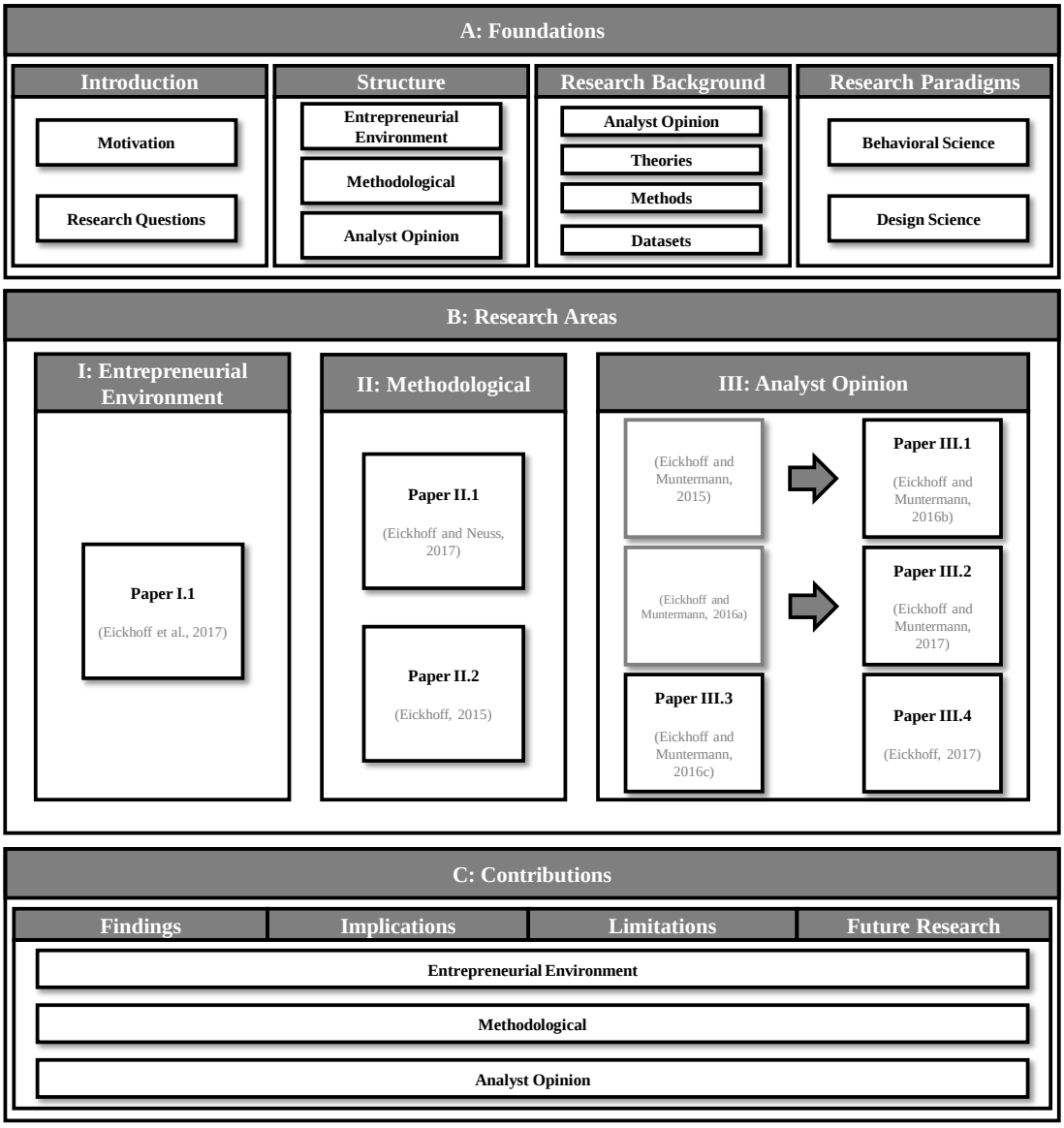


Figure 1: Structure of the thesis. Papers included in this dissertation are grouped by their respective research streams. Stream I contains a study concerned with the impact of digitization on the financial industry. Stream II contains methodological contributions regarding the methods used in stream III. Stream III contains studies using different content analysis methodologies to study the impact and information value of analyst opinion.

The bottom third of the figure outlines the structure of the result summary (C), in which the results of the individual contributions are summarized and their implications, limitations, and resulting opportunities for future research based on the presented results are discussed for each research area of the thesis.

3.1 Part A: Foundations

As shown in Figure 1, the foundational part of this thesis continues to outline the research background of the presented results after this overview of the thesis' structure. This is done by first providing a brief overview of the theories, which are relevant to the results presented here. Crowd wisdom theory, media richness theory, and a structured decision making process and its relationship to the phenomenon of information overload are discussed. Afterwards, the most important methods used throughout the different research contributions are introduced. This section focuses on the text mining methods used to process analyst opinion and social media data throughout the thesis, event studies, as well as literature reviews and taxonomy development. In turn, the datasets used throughout this thesis are presented and their individual characteristics are elaborated upon. Finally, behavioral science and design science are introduced because these two research paradigms constitute the theoretical underpinnings of the research design of this thesis.

3.2 Part B: Research Areas

As noted, the contributions included in this thesis concern three different areas of research, each of which is needed to develop a well-rounded impression of the market reaction to analyst opinion. Table 1 provides an overview of the individual research contributions, their research paradigms, and main contributions. The remainder of this section discusses how each paper is situated within its respective research area.

Research Area I – Managerial Environment: The first research area concerns the entrepreneurial environment within which this thesis is situated. Eickhoff et al. (2017, **paper I.1**) develops a taxonomy for FinTech business models and provides a thematic introduction to the changing landscape of business models in the financial industry.

Research Area II – Methodological: The second area of interest is given by the methodological foundations needed for the analysis of textual analyst opinion. In Eickhoff (2015, **paper II.1**), a framework for sentiment analysis using a hybrid method incorporating word lists and machine learning-based sentiment classification is developed. In Eickhoff and Neuss (2017, **paper II.2**), a literature review of the topic modeling methodology and its use in information systems research and other managerial disciplines is conducted.

Research Area III – Analyst Opinion: Finally, the third area of interest and the main theme of this thesis is presented by the analysis of analyst opinion throughout various media types and its effect on the stock market, as well as the comparison between analysts' opinion and other media types. Eickhoff and Muntermann (2016b, **paper III.1**) uses crowd wisdom theory to analyze the relationship between social media content and analyst opinion. This paper is an extension of the results presented in Eickhoff and Muntermann (2015). Eickhoff and Muntermann (2017, **paper III.2**) provide an approach to the analysis of topics contained in analyst reports and earnings conference calls, supporting decision makers in financial markets by reducing the complexity of these unstructured data sources. This paper is an extension of the results and the approach of Eickhoff and Muntermann (2016a). Eickhoff and Muntermann (2016c, **paper III.3**) investigate the topic relationship between earnings conference calls between stock analysts and company representatives, and analyst reports. Finally, Eickhoff (2017, **paper III.4**) uses media richness theory to provide explanations for differences in the information value between structured and unstructured sources of analyst opinion.

Paper Citation	**Outlet** Research Type	**Publication Status**	**Main Contribution**
I.1 (Eickhoff et al., 2017)	Redacted Taxonomy Development	Under Review	Development of a business model taxonomy for FinTech startups.
II.1 (Eickhoff and Neuss, 2017) Published	ECIS 2017 Literature Review	Published	Identification of topic modeling approaches and their use in management literature, emphasizing the differences between IS and other disciplines.
II.2 (Eickhoff, 2015)	DESRIST 2015 Design Science	Published	Development of a framework for hybrid sentiment analysis.
III.1 (Eickhoff and Muntermann, 2016b)	Information & Management Behavioral Positivist	Published	Quantifying the drivers of crowd wisdom based on social media data, and investigating the relationship between social media users' and analysts' sentiment on this basis.
III.2 (Eickhoff and Muntermann, 2017)	Redacted Behavioral Positivist	Under Review	Using topic models to relate earnings conference calls to stock returns and supporting the process by using the Simons decision process model.
III.3 (Eickhoff and Muntermann, 2016c)	PACIS 2016 Behavioral Positivist	Published	Investigation of topic spillovers between analyst reports and earnings conference calls.
III.4 (Eickhoff, 2017)	HICSS 2017 Behavioral Positivist	Published	Using media richness theory to explain the differences in media usefulness for abnormal return predictions.

Table 1: Papers included in this thesis. For each paper, a brief description of its main contribution and a handle to facilitate identification are provided along with its citation. The different research designs are elaborated upon in section 5.

3.3 Part C: Contributions

As shown in Figure 1, this section summarizes the results of the individual research papers presented in part B of this thesis. This is done by aggregating their results at the level of each research area.

The contribution part of this thesis begins with a result summary section. Here, the results are discussed regarding the research questions of this thesis that are developed in section 2. In turn, the implications, limitations, and opportunities for future research based on the presented results are discussed. For each of these subjects, the summary follows the three-area structure of this thesis and the order of papers within the research areas.

4 Research Background

This section gives a brief introduction to the research background of the presented contributions based on extant literature and provides an overview of the different theories relevant throughout this thesis. Afterwards, the most important methods used in the contributions of this thesis are discussed before providing an overview of the datasets on which these methods are used throughout this thesis.

4.1 The Information Value of Analyst Opinion

The information value of analyst estimates regarding the future development of the stock market is a long-standing research topic. Starting with the early work of Cowles (1933), who evaluate estimates on future stock returns, this research has been critical regarding the information value of such estimates.

This criticism is unsurprising and was later supported by the development of the efficient market hypothesis (Dimson and Mussavian, 1998; Malkiel and Fama, 1970), which suggests that such estimates cannot consistently outperform the market unless they are based on private information because all public information is to be quickly incorporated in the current stock price.

However, not all analyst research attempts to predict the future valuation of a firm. Instead, Bradshaw (2009) proposes the information processing model outlined in Figure 2. As shown, analysts rely on the forecasting of future financial reports to arrive at firm valuations to be able to react in case their estimate of the future firm value changes. It is important to note that this is not based on confidential information but only on public knowledge. Thus, the only possible value of such analyst research is given by either being quicker than other market participants regarding the speed of this process or having developed more accurate models for future firm performance, which make an individual analyst's estimate more accurate than others. Naturally, the market for analyst opinion is competitive, and no single analyst or firm can be expected to be the fastest or provide the most accurate recommendation consistently. Consequently, research assessing the information value of analyst opinion can focus on determining the circumstances under which analyst recommendations are valuable.

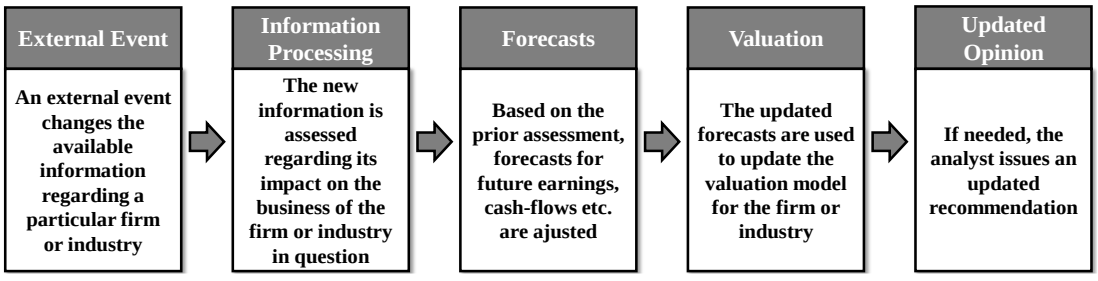


Figure 2: Analyst information processing following Bradshaw (2009).

There are two ways in which this can be done. First, individual analysts can be compared to one another to determine which analyst or firm provides the best estimates. Second, stock analysts can be compared to other sources of information to assess if analyst opinion creates complementary value beyond freely available information or to present the stock price at a given point in time. In both cases, the value of analyst opinion may also depend on the nature of the external event, which may change future firm performance. The research presented in the third research area of this thesis focuses on the second case and not on the assessment of individual analysts.

The early study of financial analysts' opinion focused on the accuracy of analysts' forecasts in comparison to other forecasts, such as management forecasts (Brown et al., 1985), the superiority of analysts' forecasts to purely time-series based forecasts (Brown and Rozeff, 1978), or their suitability as a replacement of older metrics for expected earnings (O'Brien, 1988). According to Brown et al. (1985), early research using analyst forecasts focused on five main areas overall (Brown et al., 1985, p. 1):

1. Properties of earnings forecasts by security analysts
2. Capital markets and security analyst earnings forecasts
3. Properties of earnings forecasts by management
4. Capital markets and management earnings forecasts
5. Benchmark comparisons of security analysts, management, and mechanical model earnings forecasts

This area of research has adopted many new facets. For instance, earnings conference calls between firms and analysts have gained traction as a related field of interest. In this context, management discrimination regarding the possibility to ask questions in a call has been studied (Mayew, 2008). Additionally, intra-call returns have been studied (Matsumoto et al., 2011). These and earlier studies focused on the question of whether such calls mattered (Frankel et al., 1999; Tasker, 1997). On this basis, the study of analysts' opinions has expanded beyond the analysis of structured analyst recommendations, such as the data available through the Institutional Broker Estimate System (*I/B/E/S*).

The research presented in this cumulative thesis is built upon this background of research regarding the role of analyst opinion on capital markets. This research focuses on the relationship between analyst opinion and the role of information systems in the context of their analysis. Thus, it combines theories and methods used in financial or accounting research, with those used in information systems research. Consequently, the aims and means of the presented research differ from those of research situated entirely in either domain individually, which is why the next sections give an overview of the theories and methods used throughout the individual contributions of this thesis.

4.2 Theoretical Background

This section provides brief introductions to the theories that are most important to the individual research contributions of this thesis. For each theory, an overview of its main constructs is provided.

4.2.1 Wisdom of Crowds

The term Wisdom of Crowds describes the phenomenon in which groups often outperform experts, even if the individual estimations of the group members are inferior to the expert assessment. The study of this effect has a long history in the sciences. An early study regarding this phenomenon was conducted by Galton (1907), who found that median group estimates can outperform expert opinion. Surowiecki (2005) provides a high-level overview of this area of interest and proposes that the quality of a crowds' assessment depends on the constructs shown in Figure 3, which shows the five constructs constituting the main themes of his view of the drivers of Crowd Wisdom.

As shown, these relate to both the composition of the group and the characteristics of its individual members. Within this theoretical framework, diversity is perhaps the most crucial factor influencing a group performance because the overarching idea of diversity underlies the other crowd characteristics outlined in the figure.

Increases in data availability and contexts, in which the effect can be observed, have enabled a resurgence of research in this area. Lorenz et al. (2011) study the negative impact of social influence on group decisions, Nofer and Hinz (2014) assess the performance of stock prediction communities, and Chen et al. (2014) examine the value of stock predictions transmitted through social media. In this thesis, Eickhoff and Muntermann (2015) and the extended version by Eickhoff and Muntermann (2016b, **paper III.1**) investigate ways to make these constructs measurable and compare social media users and stock analysts' opinion evolution based on per-situation metrics.

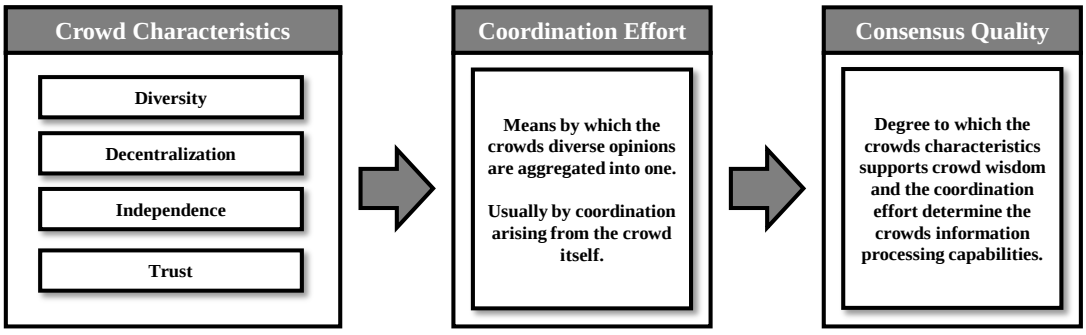


Figure 3: Wisdom of Crowd Theory (Surowiecki, 2005). Constructs influencing the quality of the average crowd opinion according to WoC theory (left). The role of coordination in arriving at a group consensus (center) and how the quality of this consensus depends on the drivers of crowd wisdom (right).

4.2.2 Decision Making and Information Overload

As discussed, the information value of stock analysts' recommendations depends on their information processing capability. Likewise, the value readers derive from them constitutes another information processing task, which consists of using the available analyst research, along with other sources of information, to arrive at an investment decision.

However, how do investors arrive at their investment decisions? Simon (1977) describes a general model for decision processes, which can help to structure this question into distinct phases, making it easier to understand the process. In the context of this thesis, Figure 4 provides an overview of how this process integrates with investment decision making. The upper part of the figure (1) shows some of the information sources available regarding listed companies in the example of data sources used throughout this thesis, which are elaborated in section 4.4. The central part of the figure (2) shows the decision process itself, which consists of surveying the available information and arriving at a problem statement, creating several potential solutions to the problem, and finally choosing from this pool of potential solutions and acting upon this alternative. Supporting decision makers in overcoming this problem has always been one of the main tasks of information systems. However, as noted by Simon (1976), information systems also contribute to this problem themselves.

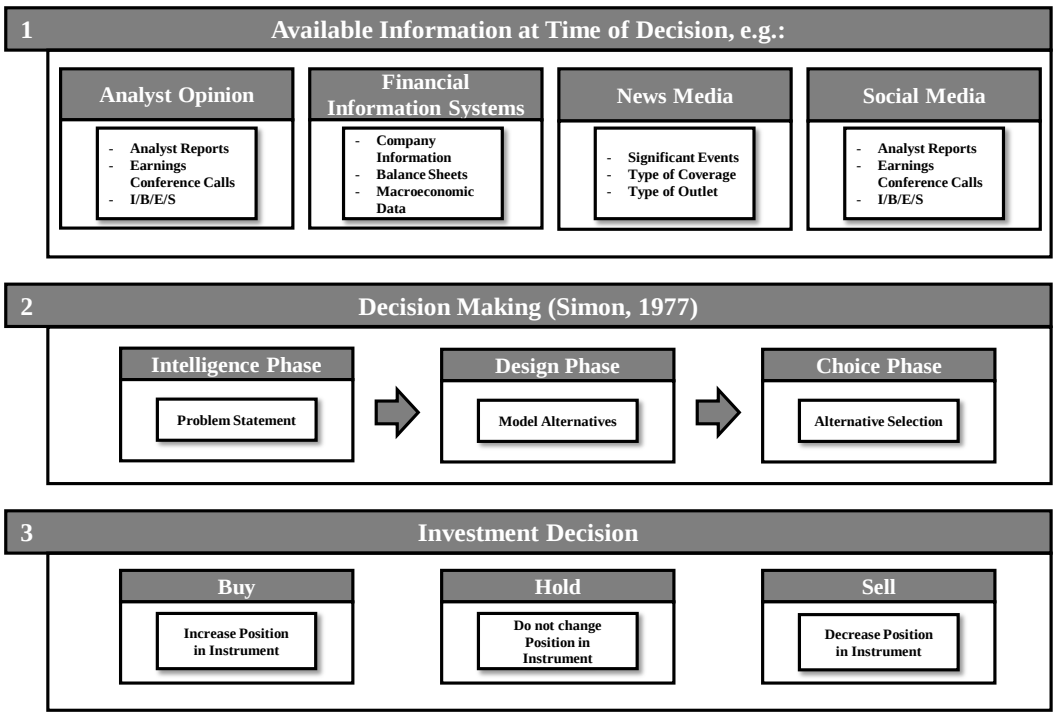


Figure 4: *Investment Decision Making. Investors assess the available information at any given point in time, as shown in the top part of figure (1). This process can be structured by the three phases of decision making proposed by Simon (1977), as shown in the middle part of figure (2). On this basis, an investment decision can be made (3).*

These diverse sources of information can overwhelm the information processing capabilities of decision makers, especially when operating under time constraints, in which case information overload can occur (Pennington and Tuttle, 2007). Due to the ever-increasing volume of digitally available information, the risk of information overload becomes more relevant as time progresses.

Making investment decisions is made more difficult by the need to assess the quality of the information made available by analysts and other sources of information. In the case of analyst opinion, prior research suggests that stock analysts exhibit several inefficiencies, which may influence the quality of their analyses in any situation. For example, analysts tend to *"stick to the herd"* by being careful to voice dissenting opinions (Twedt and Rees, 2012). One reason for this behavior is the concern that an incorrect opinion may have a negative impact on the future careers of analysts if the majority of their peers made a correct assessment in the same situation (Clement and Tse, 2005; Hong et al., 2000). Another reason is given by misguided incentive structures, which aim at increasing a firm's brokerage or investment banking revenue instead of rewarding analysts for the accuracy of their predictions (Groysberg et al., 2011). Thus, the research presented in research area III focuses on the properties of different information sources in the context of investment decisions and how understanding these properties can help decision makers arrive at informed judgements.

4.2.3 Media Richness Theory

Media richness theory, or sometimes the information richness theory, as proposed by Daft and Lengel (1983), analyzes the properties of different media types to determine what media type is suited best for the transmission of a particular type of information or a specific circumstance of the intended transmission (Daft and Lengel, 1983; Daft and Lengel, 1986; Daft and Macintosh, 1981). In its context, richness refers to the overall complexity of the medium regarding its information transmission capabilities. It argues that information transmission is most effective when the complexity of the transmitted information and the complexity of the medium used to transmit it are aligned. As shown in Figure 5, media richness theory (MRT) uses four constructs to explain this richness:

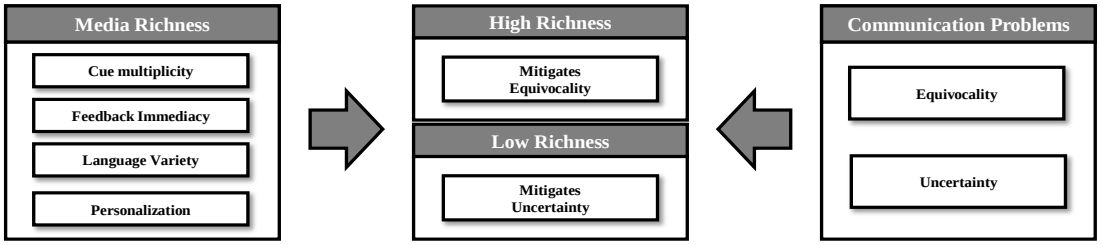


Figure 5: Media richness theory. Overview of the constructs of media richness theory based on Daft and Lengel (1983).

1. **Language variety (symbol variety)** refers to the number of different symbol types that a medium can transmit. Symbol types can, as the name suggests, be given by different human languages, but the idea of language variety exceeds this. High language variety also refers to the ability of the medium to transmit a wide spectrum of concepts and ideas. For instance, Daft and Lengel (1983) considers music to be a medium with high language variety.
2. **Cue multiplicity (channel variety)** considers the number of simultaneous channels a medium uses to transmit information. For example, face-to-face communication involves many different channels such as facial expressions, the spoken text itself, and the posture of all people taking part in the conversation.
3. **Personalization** concerns the extent to which a medium allows for messages to be customized for specific recipients. For example, a text written for children can be designed to be easier to understand than a technical document.
4. **Feedback immediacy** is defined by how interactive a medium is. For example, the ability to ask questions by the recipient of the communication or the ability to correct wrong perceptions constitute high feedback immediacy.

A medium is evaluated on the basis of these criteria and is consequently ranked on a low-richness to high-richness spectrum. Obviously, this is not a categorical assignment but rather a judgement call on a continuous scale of media richness. MRT considers two main problems that can inhibit effective communication (Daft and Lengel, 1986):

1. **Equivocality** refers to a situation of information oversupply, in which a decision maker has access to conflicting sources of information, which make it difficult to discern what information is relevant.
2. **Uncertainty** refers to a situation in which the decision maker has not been supplied with enough information to reach a decision.

The relationship between these two problems and the media richness property is given by the mitigation of either problem based on the richness of a given media type. Within the scope of MRT, high richness media mitigates uncertainty, while low richness media mitigates equivocality.

4.3 Methods

In the following sections, the main methods used in this thesis are presented. These sections are intended to be a brief introduction to each of these methods, with an emphasis on the practical applications of the methods and their individual strengths and limitations regarding the research goals of this thesis.

4.3.1 Text Mining Pre-Processing

As textual data are the basis for most of the presented analyses, this section gives a brief overview of the pre-processing needed to analyze such content. While the pre-processing needed for the different content analysis techniques used throughout this thesis differs, it is useful to introduce the terminology used in this task on the example of a straightforward text. Figure 6 provides an overview of a basic pre-processing approach. As shown, pre-processing a text can be structured in five phases. First, the document is read from a file or database.

Second, this text is split into separate pieces; in this case, the text is split directly into words. This step typically also removes non-word content, such as punctuation. Some of the literature also refers to these tokens as *features*. What constitutes a token and what is filtered out at this stage of pre-processing depends on the needs of a given analysis. For example, emoticons may be of interest despite not being words.

Third, *stopwords* are removed. These are words such as “I” or “in” that are not expected to comprise the information content of a document. Fourth, the remaining words are reduced to their stems, which increases the likelihood of the same words appearing across multiple documents or words being matched with word lists.

Finally, the remaining word stems are added to a term document matrix, which typically contains words from many documents. Often, this matrix is transformed to accommodate term weighting schemes.

A popular example is presented by *term frequency – inverse document frequency*, where the individual occurrences of words are replaced by the proportion that a term contributes to an individual text and is weighted against the overall occurrence frequency in the entire document corpus. This results in a measure for the abnormal portion of a word’s importance.

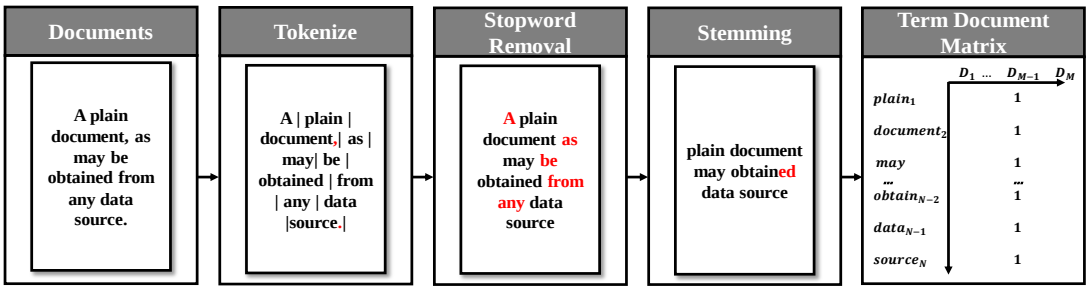


Figure 6: *Pre-processing of textual data. Plain text documents are pre-processed by creating separate tokens, removing uninformative words, reducing these words to word stems, and finally arranging the resulting features in a term-document matrix along with other documents.*

As noted, the needs for pre-processing are determined by the intended analysis. For example, topic models may benefit from maintaining word order per text instead of relying on a bag-of-words approach; a dictionary-based sentiment analysis may benefit from the use of more advanced methods than a simple stemmer, such as determining parts of speech or finding different words with the same meaning (word sense disambiguation). Indeed, the different presented papers use different pre-processing logics, but the principle of generating a term document matrix containing the tokens generated for each particular text holds true for most text mining approaches.

4.3.2 Sentiment Analysis

Sentiment analysis addresses extracting measures of authors' opinions from unstructured textual data. Two popular basic approaches to this problem currently exist. First, a dictionary-based sentiment analysis uses pre-determined word lists to determine the sentiment value of a text. Second, a machine learning-based sentiment analysis uses classification algorithms trained on pre-classified texts. Figure 7 provides an overview of this approach using a positive and negative word list, which corresponds to the approach taken in the papers included in contribution group III. The figure shows excerpts from a positive and a negative word list and how a positivity score is calculated for three example documents based on the intersection between the documents and the two word lists. The first document (D_1) is classified as positive because more positive words are found in the document than negative ones. In contrast, the second document, (D_2) is classified as negative because more negative words are contained therein. Finally, D_5 is considered neutral because the document contains an equal amount of words contained in the two categories.

Obviously, the success of this dictionary-based approach to sentiment analysis hinges upon the selection of a dictionary containing domain-appropriate words, which can help to represent authors' sentiment regarding the subject matter of interest. Consequently, many sentiment dictionaries for different domains have been developed. Within this thesis, two principle types of dictionaries are needed.

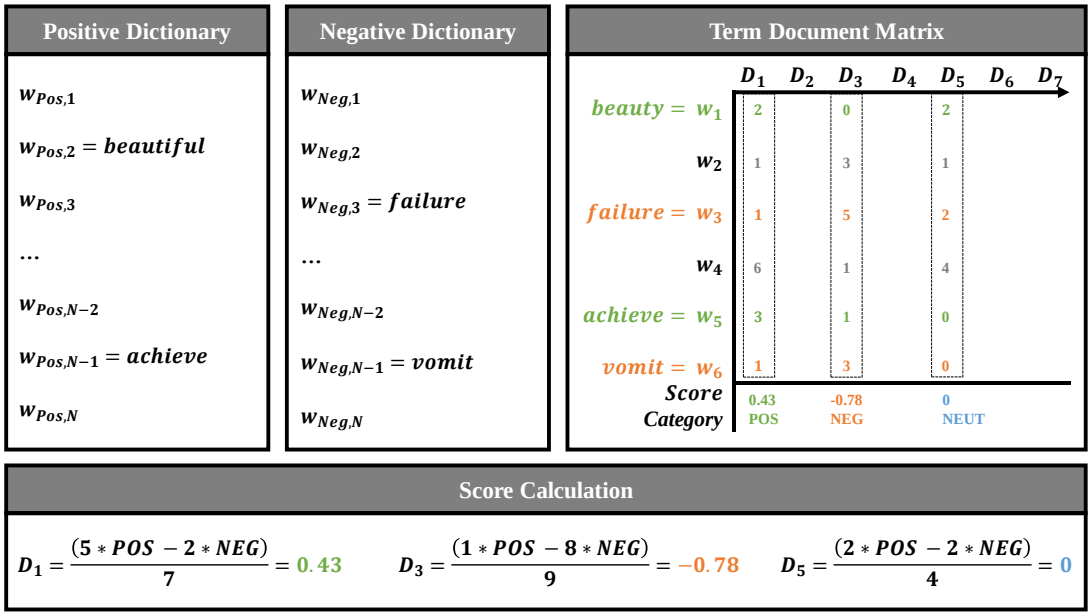


Figure 7: *Depiction of dictionary-based sentiment calculation. Positivity score calculation based on a positive and a negative sentiment dictionary. For each text, the word co-occurrences with each category are determined. On this basis, a measure scaled to text-length is computed by dividing with the sum of all category hits. One positive, one negative, and one neutral example is given.*

First, social media users' opinions need to be analyzed. Second, stock analysts' and other finance-oriented authors' opinions are of interest. For financial content, the 10-K dictionary (Loughran and McDonald, 2011), and a press release (Henry, 2008) dictionary are used. For social media users' opinion, the more general purpose Hu-Liu dictionary (Hu and Liu, 2004) and the valence categories from the General Inquirers dictionary — in its current version after the integration of *Harvard IV-4* and *Lasswell* categories — are used (Stone et al., 1966).[1] When a comparison between these two content types is desired, the latter type of dictionary is used because of the problems arising from the comparison between measures derived using different dictionaries, such as different proportions of positive and negative words. In contrast, machine learning-based sentiment analysis does not usually rely on such sentiment dictionaries but on a set of training documents. In the case of sentiment analysis, this set of training documents is usually assigned to categories by human coders. While there are a considerable number of classification algorithms suited to this task, this prior labeling of training documents is perhaps the most crucial step in successful machine learning-based sentiment analysis because any algorithm applied to this categorization task is limited by the quality of its training data.

[1] No comprehensive published work is available that describes the current unified version of the General Inquirer dictionary. The best resource for information regarding this dictionary is available online (GI-Team, 2002).

Overall, while sentiment analysis is usually performed at the document level, its results are most reliable when aggregated over a large number of documents to determine the average sentiment for a specific group of authors or a timeframe. This is because, regardless of the chosen approach to sentiment analysis, these models are only accurate to a certain degree, which makes judgements on individual documents unreliable.

Still, the choice of a suitable algorithm impacts both model accuracy and interpretability (Pang et al., 2002). In this thesis, machine learning-based sentiment analysis is only used in combination with dictionary-based techniques, which are used as a substitute for the manual coding of texts (Eickhoff, 2015, **paper II.1**). In general, modern machine learning-based techniques can perform at accuracy and recall rates similar to human judgement (Sharma and Dey, 2012), but this performance comes at the cost of the need for per-corpus training data.

4.3.3 Topic Modeling

Topic modeling is a technique intended to extract the core themes discussed in a given document and have been developed with the intent of easing the browsing of document collections regarding such underlying topics (Blei, 2012). Early solutions to this task include a latent semantic analysis (Croft and Harper, 1979; Deerwester et al., 1990; Landauer and Dumais, 1997) and non-negative matrix factorization (Lee and Seung, 2001).

The topic modeling technique used throughout this thesis, introduced by Blei et al. (2003), is called Latent Dirichlet Allocation (LDA) and differs from these earlier approaches. It not only clusters documents regarding the topics contained in them but also provides *topics*, which are intuitively interpretable by humans if introspection of the algorithmic results is desired. Another difference between LDA and other methods is given by the fact that for each document, more than one topic can be assigned. In fact, each document is represented as a mixture of underlying topics. This enables a more granular analysis of documents and is useful when using the topic assignments as regression variables in the studies in research area III of this thesis. Figure 8 shows how the model assigns words to topics and in turn assigns these topics to the documents in the document collection.

This two-stage approach to topic modeling enables the interpretation of both word to topic and topic to document assignments. Word to topic assignments can be used to interpret the meaning of the topics and are often used to assign labels to topics, while topic to document assignments can help to clarify which document in a collection contains content that is relevant to each of these labels.

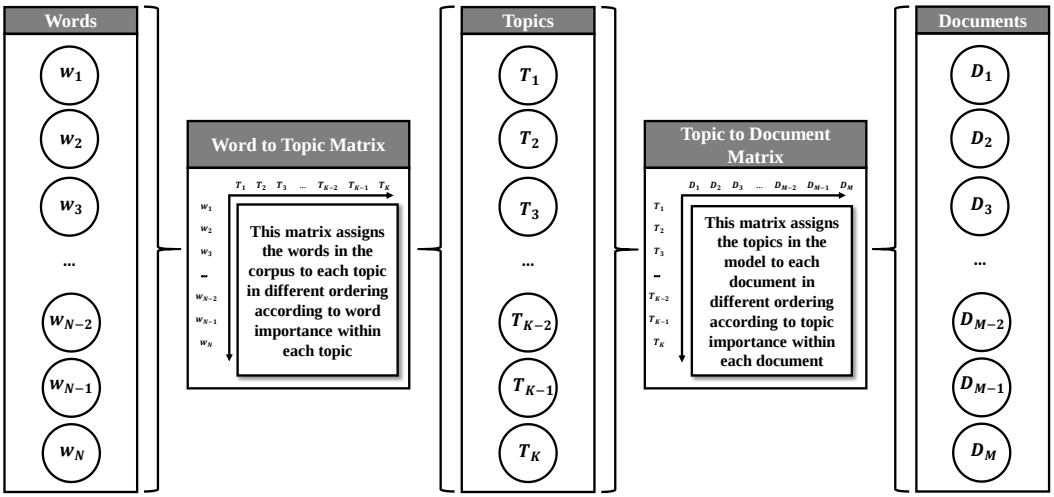


Figure 8: *Overview of LDA model components. The figure shows the relationships between words, topics and documents in a corpus of M documents, K topics in the model, and N words in the corpus. The word to topic matrix assigns a word probability for each word in the corpus within each topic. The topic to document matrix assigns the estimated topics to the documents.*

Figure 9 provides an example of how a topic model can be used to investigate the content of a document. As shown, when looking at a text, topics can overlap regarding their allocation to a document. Additionally, because each word is assigned to each topic with differing likelihood, topics can share important words. The higher the number of topics estimated by a topic model becomes, the more overlap between topics arises, while also increasing the model fit to the training data. Thus, the number of estimated topics is a trade-off between model fit and the interpretability of the estimated topics. Another way to use topic models is given by the possibility of using the resulting topics as explanatory variables in regression models. Topic to document assignments are a numerical representation of the information contained in a document. The advantage of topic models when compared to other ways to generate such numerical representations of documents, such as *Doc2Vec* (Le and Mikolov, 2014), is given by the interpretability of this model type.

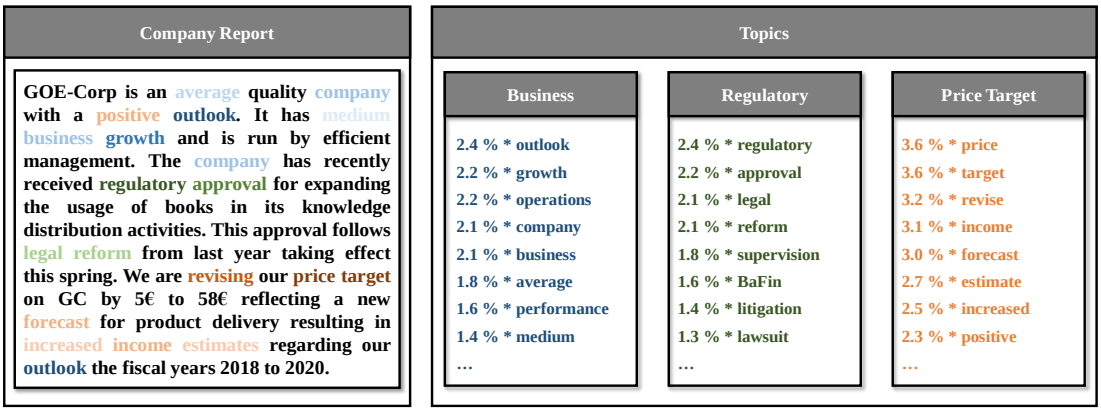


Figure 9: *Example of a company report containing different subjects. The text is compared to the word to topic assignments estimated using a topic model.*

Recently, topic models have become an increasingly popular tool in information systems research (Müller et al., 2016). Methodological recommendations regarding this technique have begun to appear in methodological IS research (Debortoli et al., 2016), and outlets have begun to encourage the submission of research using this methodology (Rai, 2016). In this thesis, Eickhoff and Neuss (2017, **paper II.2**) expands on the use of this method in the managerial sciences and information systems research in particular, and topic modeling is used extensively in the studies presented in research area **III**.

4.3.4 Event Study Analysis

Event studies are a common method in financial research aimed at investigating the effect of events on the future development of stock prices and implicitly on the value of the issuing firm. Unsurprisingly, this type of question was posed early on. Dolley (1933) investigated the effect of stock splits on firm value. This method continues to be a common tool for empirical research in accounting and finance (Corrado, 2011). Event studies are typically conducted by examining a larger sample of similar events and averaging the effect observed after the event. Events are often grouped into several event categories to highlight the differences among these types. The main methodological problem is given by the estimation of how the stock price would have behaved if no event had taken place at a given event date to determine the difference between this counterfactual estimate and the observed development of the stock price. Figure 10 provides an overview of this concept. While there are several types of event studies regarding the underlying market conditions and how the effect of an event is estimated, this thesis utilizes the market model approach (MacKinlay, 1997).

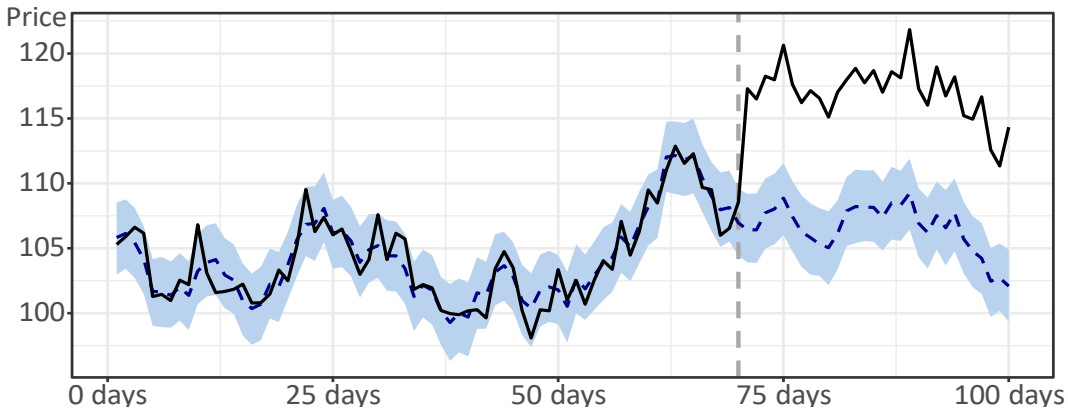


Figure 10: *Event study example. Actual stock price in black, estimated generative process shown with a dotted line. At $days = 70$, an event occurs and has a lasting effect on the future development of the series. The dotted estimate is based on the pre-event period. The purpose of the event study methodology is to quantify this change in the process generating the series by estimating the difference between the two lines following the event.*

As noted by Corrado (2011), this type of event study was popularized in accounting and finance research by Ball and Brown (1968) and Fama et al. (1969) due to the recent introduction of the capital asset pricing model (CAPM) by Sharpe (1964).

Another reason is given by increases in data availability to researchers after the introduction of early financial information systems. Event studies have been a common sight in financial research and have been adopted by other domains to explain phenomena other than stock prices (Skiera et al., 2017). In contrast to constant mean return models, market models include a reference index to relate the movement of the stock in the event of the general movement of the stock market, thus reducing the bias of the abnormal return estimate.

In this thesis, event studies are used to relate the content of analyst opinion in the form of analyst reports or analyst conference calls to the abnormal returns following earnings announcements.

4.3.5 Literature Review

As noted by Webster and Watson (2002), literature reviews are an essential basis for any research project because they enable authors to base a contribution on prior knowledge and help to identify gaps in the prior research. Additionally, literature reviews are of special importance to junior researchers attempting to understand the status quo within a research domain, as well as senior researchers attempting to remain current with an ever growing body of knowledge (Templier and Paré, 2015).

Consequently, an ongoing methodological discussion regarding the manner in which such reviews are conducted has developed in information systems research (e.g., Greenhalgh et al., 2005; Rowe, 2014; Sylvester et al., 2013). Templier and Paré (2015) differentiate between the nine types of literature reviews shown in Table 2.

Type	**Content**
Narrative	Simplest type: Find what has been written on a subject (unstructured)
Descriptive	Identify trends or structure in prior research (structured approach)
Scoping/Mapping	Exploratory review to determine if a structured review makes sense
Meta-analysis	Quantitative meta-analysis such as meta-regression
Qualitative systematic	Qualitative meta-analysis, discussion of results of multiple quantitative studies
Umbrella	A review that summarizes multiple reviews
Theoretical	Aims at theory development, often for new phenomena
Realist	Investigate complex non-deterministic phenomena
Critical	Assess extant literature regarding its fulfillment of review criteria

Table 2: Types of literature reviews in IS research. Nine types of literature reviews exist according to Paré et al. (2015).

These types can help to clarify the structure and intent of specific reviews, and mixed approaches are to be understood as archetypes because they are not mutually exclusive.

In the context of this thesis, Eickhoff and Neuss (2017, **paper II.1**) explicitly use this structured approach to literature analysis. The methodological section of this paper details the review type and the motivation for choosing it over the other types. Of course, the other presented contributions also base their results on the context of prior research and consequently use literature review methodology implicitly.

4.3.6 Taxonomy Development

A taxonomy is a system of classification for objects or concepts. For example, a taxonomy of business models can be used to determine which archetype of business model a particular firm follows (Baden-Fuller and Morgan, 2010; Porac and Thomas, 1990) or developing a classification for different types of management information systems (Earl, 2001). Taxonomies may also be referred to as typologies or frameworks for classification (Nickerson et al., 2013, p. 337). By developing taxonomies, researchers organize knowledge in a formalized manner (Wand et al., 1995, p. 291). This systemization of knowledge is a long-running theme in information systems research (Hirschheim et al., 1995). In a larger context, taxonomies relate to ontological classifications, which have also attracted increasing attention in computer science and information systems research (Guarino, 1998). Taxonomies can be viewed as one artifact type that is relevant in the development of ontologies, among other artifact types such as thesauri or controlled vocabularies (Gruninger et al., 2008), but taxonomies are considered a separate research approach (Dogac et al., 2002).

Fiedler et al. (1996, pp. 11-12) note that systems of classification were relevant to research *"since Aristotelian applications"*. Regarding the use of taxonomies in information systems research, Glass and Vessey (1995) highlight the role of taxonomies in structuring fields of knowledge, thereby allowing researchers to hypothesize about relationships. Likewise, according to McKnight and Chervany (2001), taxonomies bring order to concepts and make it easier for researchers to highlight relationships between them. Thus, taxonomies enable theory building (Doty and Glick, 1994) if the developed taxonomy is robust (Bapna et al., 2004). Additionally, taxonomies can make divergence in prior research visible (Sabherwal and King, 1995). Regarding the philosophical foundations of taxonomy development, Iivari (2007) notes that taxonomies are descriptive or prescriptive forms of conceptual knowledge in the epistemology of design science.

The taxonomy developed in Eickhoff et al. (2017, **paper I.1**) is based on the methodology proposed in Nickerson et al. (2013). The taxonomy development process as described by Nickerson et al. (2013) begins by determining a meta characteristic, which determines the focus of the further development process. In the case of the presented

contribution, this focus is given by searching for the determinants of FinTech business models. Consequently, the satisfaction of several ending conditions, such the ability of the developed taxonomy to classify all observations, is determined. An iterative development process then occurs, which alternates between empirical-to-conceptual and conceptual-to-empirical development cycles until the ending conditions are satisfied. This process is described in more detail in the methodology section of Eickhoff et al. (2017, **paper I.1**) and results in a taxonomy developed using a transparent development process, which enables other researchers to verify the results or extend the presented taxonomy.

4.4 Datasets

This section introduces the datasets used throughout the thesis and has two objectives. First, it introduces the reader to the used data and explains the pertinent terminology and expands on the information about the data sources available in each presented contribution. Second, it discusses the properties of the different datasets, which have an impact on research or the conclusions that can be drawn from the presented research.

4.4.1 Social Media

The social media data used in Eickhoff and Muntermann (2016b, **paper III.1**) was obtained from the SDLs SM2 database (SDL, 2017). Primarily, SM2 is a social media monitoring tool. However, it can also be used to perform searches on historical data and allows data exports using a semi-structured format based on the extensible markup language (XML). The database contains information from many social media platforms as well as other related content platforms, such as blogs, microblogs, and social news sites. In addition to the full content of each post, data exports also contain several metadata fields, including information about authors, such as gender or age, if this information was made available by the platform on which the content was hosted. Table 3 provides an overview of this metadata and elaborates on its availability.

As shown, some fields such as author age are only available if the platform of a posting provides this information readily and the author of the post has opted to provide it. The papers using this information as a basis for analysis are therefore limited to examining aggregate information about these fields for specific time periods instead of operating at the level of individual social media posts.

Field	Availability	Description
Media type	Yes	E.g., blog, microblog
Platform	Yes	Name of platform, e.g., Twitter
Author name	No	Name, often pseudonym, of author
Gender of author	No	Male or female
Age of author	No	Age in years
Location of blogger	Partially	City or state level in the US or UK, mostly country level elsewhere
Full Content	Yes	Full unstructured textual post
Location hosted	Yes	City or state level in US or UK, mostly country level elsewhere
Time discovered	Yes	When the post was discovered by the content provider
Time published	Partially	When the post was authored
Blog URL	Yes	Link to platform
Permalink	Yes	Link to individual post

Table 3: Description of data fields present in SDL SM2 XML exports. Limited to the fields used as the basis of variables of models in Eickhoff and Muntermann (2016b, paper III.1) and basic information about the content.

4.4.2 News Media

The news media data used in Eickhoff and Muntermann (2016b, **paper III.1**) were accessed using The Guardian's open data platform, from which all textual content published in The Guardian's online or print version is available, including rich metadata for each content item (The Guardian, 2017). This data source is used to identify the type of events that occurred within a given period using the categories of news published regarding a company.

4.4.3 Analyst Opinion

The analysis of analyst opinion is the foundation of research area III and its research questions. Thus, the sources of analyst opinion used for this purpose are of special importance when interpreting the presented results. This section introduces the three types of analyst opinion data used in this thesis, which are given by analyst reports, earnings call transcripts, and analyst estimate data obtained from the *I/B/E/S* system.

4.4.3.1 Earnings Calls

Earnings calls are telephone conferences that are typically held on the day of a firm's earnings announcement. These calls normally consist of two sections. First, the company presents their results in a monologue. Afterwards, a question and answer section follows in which participating analysts can ask questions about the firm's business. The information value of such calls has been studied extensively.

Early work on this subject focused on the question of what facts a call conducted voluntarily by a firm can tell investors about the quality of financial reporting and why managers opt to hold calls (Frankel et al., 1999; Tasker, 1997; Tasker, 1998). Another important aspect is the question of what role these calls play in the dissemination of information on capital markets, and call participants may be privileged in this respect (Sunder, 2002).

Since these beginnings, the analysis of tone measures such as dictionary-based sentiment measures has become a focus of this stream of literature in analyzing its incremental value (Price et al., 2012), the differences in investors' ability to interpret this tone (Blau et al., 2015), the differences in the tone of analysts and managers (Brockman et al., 2015), as well as what portion of this tonal measure is specific to individual managers (Davis et al., 2015). In this thesis, this research stream is contributed to by using topic modeling as an alternative to and a combination of the same with tonal measures.

4.4.3.2 Analyst Reports

The analyst reports used throughout this thesis refer to sell-side financial analyst reports, which are written for a broad audience and are typically either written by analysts employed by large banks intending to inform clients or are written directly for

sales. In contrast, buy-side analysts conduct their research for use by their employer. Such reports have been subject to continuous research efforts investigating their information value. Kloptchenko et al. (2004) use self-organizing maps to analyze the content of analyst reports. Asquith et al. (2005) study the market reaction to analyst reports to assess their information value. In contrast to the studies presented here, this analysis is based on extensive meta-data for each report while also incorporating some metrics for the textual portion of the reports. Similarly, Twedt and Rees (2012) study the relationship between analyst tone and market reaction. Huang et al. (2014) also study this reaction to analyst tone. Franco et al. (2015) focus on the effects of report readability. These studies are only a small sample of the active research stream surrounding analyst reports, which is mainly published in outlets focused on accounting and finance research. In this thesis, this prior research is extended by using new methods for the analysis of their unstructured content. This is done using media richness and crowd wisdom theory to explain their information value and to relate it to other content sources.

4.4.3.3 Broker Estimates

The Institutional Broker Estimate System (*I/B/E/S*) was developed by the brokerage firm Lynch, Jones and Rian Inc. in the early 1970s to systemically aggregate analysts' forecasts. Here, this estimate data are used as an augmentation to the data extracted from unstructured sources of analyst opinion and as a benchmark thereof. Initially, the primary focus of this database was to provide aggregate earnings forecasts, but the scope of the database has expanded since then. The current version of the database, accessed via Thomson Reuters for the purposes of this thesis, contains a 20-year history of analyst estimates (Reuters, 2015). This modern version of *I/B/E/S* contains both summary estimates, i.e., the average of all analyst estimates submitted to the system, as well as statistical properties of these averages, such as the standard deviation of the average. Additionally, the underlying individual estimates are available to some extent, although these are often anonymized. The main strengths of the database relevant to this thesis are the extensive historical data available in the system today, as well as the considerable number of covered firms. The most important limitation of the database relevant here is the possibility of reporting lags, for which Brown et al. (1985, p. 25) identify three different sources:

1. Lags can occur when analysts revise their forecasts and inform their clients but do not immediately submit this revision to *I/B/E/S*
2. There can be a mismatch between the reporting period of *I/B/E/S* (initially end of month) and the reporting period of analysts
3. Lags between submission to the system and availability to its users

For this thesis, only the first two lag types are relevant because historical data are analyzed, which negates the problem of lags due to data processing within the system.

However, the first two lag types permanently change the date of a forecast revision if the problem arose anywhere in the historical data. However, the bias introduced by this is presumably small when looking at data for recent years, which constitute the observation periods of the different papers of this thesis (all papers are based on post year 2000 data).

4.4.4 Startup Profiles

Crunchbase (2016) is a database containing information about a wide range of companies with a focus on startups. In contrast to "traditional" financial information systems, the database is focused on providing information about the funding structure of startups and the individuals and companies providing this funding, as well as the funders of the startups themselves. This focus on non-listed startup companies means that for a typical company listed on Crunchbase, much less information is publicly available about a firm than is typically available for the larger listed companies, about which information is gathered from other data sources in the presented contributions. However, the information provided by Crunchbase is unique regarding the considerable number of startups covered in the database.

This focus on unlisted startup companies enables researchers to use the database to investigate changes in business models due to changes in the socio-technological landscape of this rapidly evolving type of company instead of limiting the analysis on listed companies, which are inherently more inert due to their size and corporate structure. Thus, these data can be used to look for patterns in startup companies' business models, which may later be adopted by industries at large. In this thesis, this is done for the case of FinTech companies in Eickhoff et al. (2017, **paper I.1**), which uses the company category tags provided by the database to identify FinTechs contained in the database and continues to develop a taxonomy of their business models based on this subset of the database.

5 Research Paradigms

In this section, the different research paradigms and designs used throughout this thesis are introduced, and their different assumptions regarding the goals and means of research are discussed.

5.1 Behavioral Science

What constitutes scientific research and what does not is perhaps an unanswerable question. However, for the purposes of this thesis, a definition is both necessary to be able to develop the presented research project and is useful in the interpretation of the presented results. Thus, the following definition is followed because it summarizes many important principles and is in line with the type of research conducted here:

Scientific research is [a] systematic, controlled, empirical, amoral, public, and critical investigation of natural phenomena. It is guided by theory and hypothesis about the presumed relations among such phenomena (Kerlinger and Lee, 2000, p. 14).

Kerlinger and Lee (2000) emphasize three aspects of this definition. First, research should be systematic and ordered, which means that it should follow clearly defined guiding principles or, ideally, a process. Second, research is empirical in the sense that scientists should not report their beliefs as a result but instead subject this belief to a test using a defined method, which may invalidate it. Third, the results of research are not filtered by the moral conceptions of the researcher.

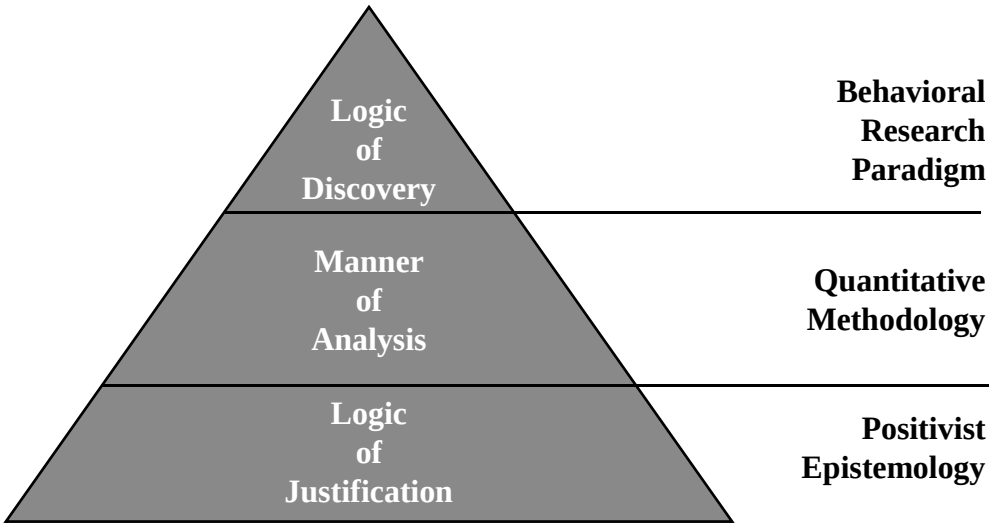


Figure 11: Epistemological foundation of the presented thesis. Hierarchy of behavioral research paradigm, quantitative methods used to investigate the phenomena of interest, and epistemological justification of the interpretation of the presented results. This follows the terminology discussed by Teddlie and Tashakkori (2009, p. 57). While the logic of discovery focuses on the formulation of theories, the logic of justification refers to the justification of why a certain approach generates trustworthy results (Johnson and Onwuegbuzie, 2004).

However, because this thesis addresses the study of human behavior and the behavior of groups such as social media users or market participants, this definition appears too focused on natural phenomena. Therefore, a behavioral research paradigm is followed that focuses on the interpretation and explanation of human behavior.

As noted by Devereux (1967, p. 10), *"it is customary but unsatisfactory to differentiate between the life sciences and the physical sciences primarily in terms of whether the object studied is an organism or inanimate matter."* Instead, Devereux (1967, p. 10) notes that a more prudent distinction between the behavioral and physical sciences can be based upon "intervening variables," which act as a filter between what is considered the "cause" and "effect" within the scope of a given analysis.

This thesis follows this idea of behavior as a mediator between causal relationships, and, in the strictest sense, all data used in the presented studies are a result of human behavior. It is therefore behavioral in nature as opposed to strictly naturalistic data, such as the weight of a carbon atom.

Within information systems research, in the context of systems design, this positivist framework is sometimes referred to as functionalist. According to Hirschheim et al. (1995, p. 69), within a functionalist mindset, *"the role of information systems is to provide timely information, which is relevant to decision makers for organizational problem solving and control."* As this definition suggests, this approach to information systems design considers a data model to be a representation of an observable reality, which in turn can be used to inform decision makers about this reality to enable them to act based on this information.

Thus, the contributions of this thesis are generated using a behaviorist research approach and quantitative methods and are based upon a positivist epistemological mindset. Figure 11 shows this hierarchy of research paradigm, method, and epistemological foundation of the conducted research.

5.2 Design Science

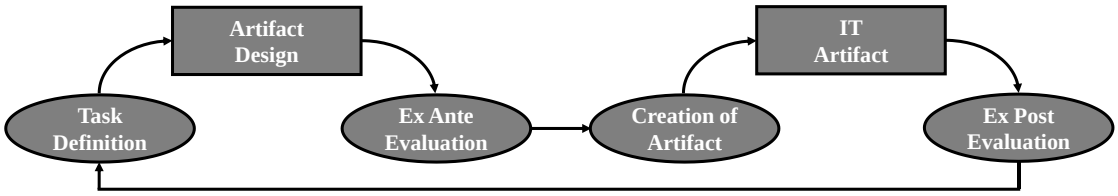


Figure 12: An idealized Design Science Research Process. The process highlights the sequential creation of an artifact design and the artifact itself, and the iteration between the finished artifact and the search for a problem. This process closely follows Baskerville et al. (2009).

Outside of information systems research, Design Science originated as a method in architectural design that intended to provide a structured design process template, which aims at anticipating the needs of structures' future inhabitants (Fuller, 1957). Design Science Research, as a formalized research method in information systems research (Hevner et al., 2004; March and Smith, 1995), is a problem-solving paradigm for artifact design. It is often focused on providing useful results and is less focused on discovering underlying patterns of *truth* (Winter, 2008). Of course, this distinction is by no means mutually exclusive. This formalization within the IS domain goes back to Simon (1996). Baskerville et al. (2009) differ from prior sequential definitions of artifact design by highlighting the iterative nature of the design process. An overview of their version of the design science process is provided in Figure 12.

As shown, the process begins with the formalization of the task at hand, resulting in an artifact design that is evaluated before its implementation. In the case of empirical model development, this ex-ante evaluation can be presented by pre-testing a design on a smaller sample. Finally, the completed artifact is also submitted to an evaluation. Again, in the case of empirical models, this can be done by comparing the performance of multiple designs or an evaluation of the out-of-training sample performance of a given model. This also relates to the iterative nature of this process; any finalized artifact, e.g., the implementation of an empirical model, may point to the need for changes in artifact design, such as choosing another algorithm for model estimation or a different set of model-hyperparameters.

The relationship between Design Science Research (DSR) and this thesis is given by its use in Eickhoff (2015, **paper II.2**), where the approach is used to create a framework for sentiment analysis using a combination of dictionary and machine learning-based methods. Additionally, the taxonomy development process of **paper I.1**, conducting the literature review for **paper II.1**, and the development of the empirical models of the papers in research area II can be regarded as design science problems, although, in those cases, this is not made explicit in their research designs.

B. Studies: Individual Research Contributions

In this part of the cumulative thesis, the individual research contributions are presented. The presentation follows the three research areas developed in part A of this thesis. Figure 13 reiterates this grouping to provide an overview of the following papers.

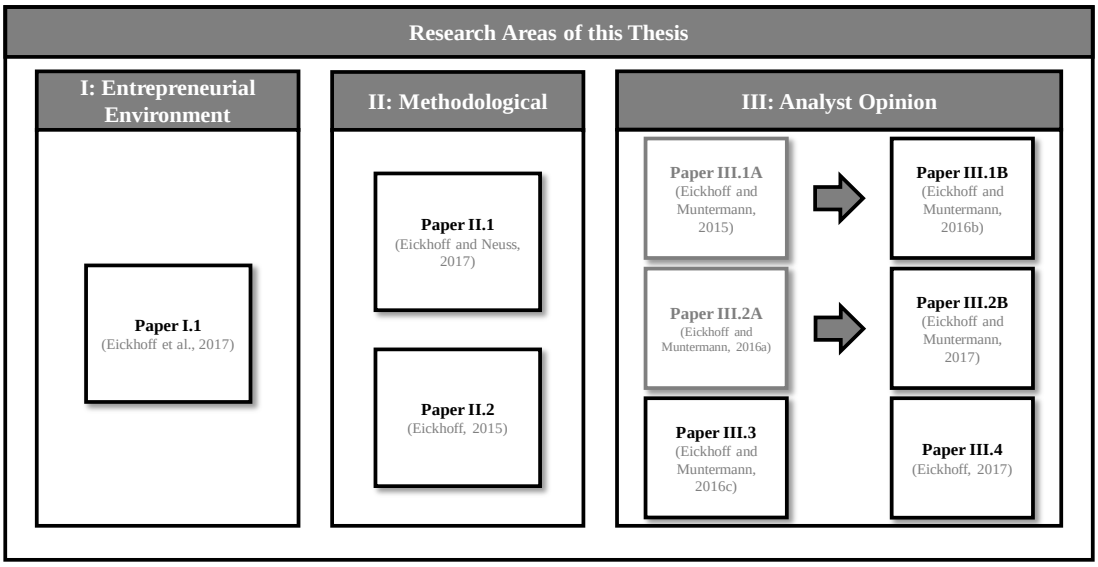


Figure 13: Grouping of research papers in research areas.

I. Research Area: Entrepreneurial Environment

This research area provides an examination of the changing entrepreneurial landscape in the financial industry. The following research paper develops a business model taxonomy for FinTech startups in the financial domain. This highlights the pressure this new competition exerts on incumbent firms in the financial industry. It addresses the following research questions:

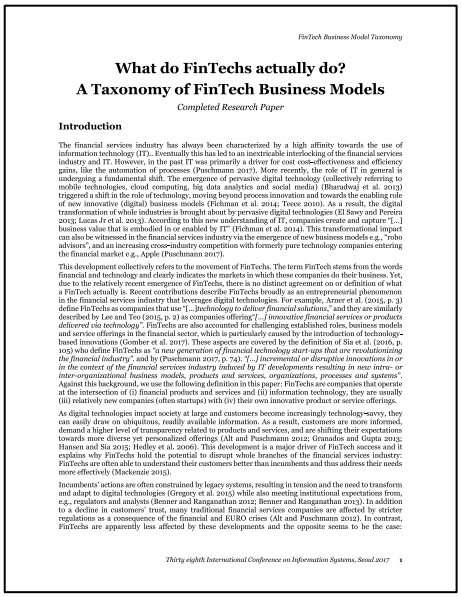
Research Question I.1: What are the dimensions and characteristics of typical business models of FinTech companies?

Research Question I.2: How can these business models be grouped into different FinTech niche markets?

I.1. FinTech Business Model Taxonomy

(not included in this document due to copyright)

What do FinTechs actually do?
A Taxonomy of FinTech Business Models



FinTech Business Model Taxonomy

What do FinTechs actually do?
A Taxonomy of FinTech Business Models

Completed Research Paper

Introduction

The financial services industry has always been characterized by a high affinity towards the use of information technology (IT). Eventually this has led to an inextricable interlocking of the financial services industry and IT. However, in the past IT was primarily a driver for cost cost-effectiveness and efficiency gains, like the automation of processes (Puschmann 2017). More recently, the role of IT in general is undergoing a fundamental shift. The emergence of pervasive digital technology (collectively referring to mobile technologies, cloud computing, big data analytics and social media) (Bharadwaj et al. 2013) triggered a shift in the role of technology, moving beyond process innovation and towards the enabling role of new innovative (digital) business models (Fichman et al. 2014; Yoo et al. 2010). As a result, the digital transformation of whole industries is brought about by pervasive digital technologies (El Sawy and Pereira 2013; Lucas Jr et al. 2013). According to this new understanding of IT, companies create and capture "[…] business value that is embodied in or enabled by IT" (Fichman et al. 2014). This transformational impact can also be witnessed in the financial services industry via the emergence of new business models e.g., "robo advisors", and an increasing cross-industry competition with formerly pure technology companies entering the financial market e.g., Apple (Puschmann 2017).

This development collectively refers to the movement of FinTechs. The term FinTech stems from the words financial and technology and clearly indicates the markets in which these companies do their business. Yet, due to the relatively recent emergence of FinTechs, there is no distinct agreement on or definition of what a FinTech actually is. Recent contributions describe FinTechs broadly as an entrepreneurial phenomenon in the financial services industry that leverages digital technologies. For example, Arner et al. (2015, p. 3) define FinTechs as companies that use "[…]*technology to deliver financial solutions,*" and they are similarly described by Lee and Teo (2015, p. 2) as companies offering "*[…] innovative financial services or products delivered via technology*". FinTechs are also accounted for challenging established roles, business models and service offerings in the financial sector, which is particularly caused by the introduction of technology-based innovations (Gomber et al. 2017). These aspects are covered by the definition of Sia et al. (2016, p. 105) who define FinTechs as "*a new generation of financial technology start-ups that are revolutionizing the financial industry*", and by (Puschmann 2017, p. 74). "*[…] incremental or disruptive innovations in or in the context of the financial services industry induced by IT developments resulting in new intra- or inter-organizational business models, products and services, organizations, processes and systems*". Against this background, we use the following definition in this paper: FinTechs are companies that operate at the intersection of (i) financial products and services and (ii) information technology, they are usually (iii) relatively new companies (often startups) with (iv) their own innovative product or service offerings.

As digital technologies impact society at large and customers become increasingly technology-savvy, they can easily draw on ubiquitous, readily available information. As a result, customers are more informed, demand a higher level of transparency related to products and services, and are shifting their expectations towards more diverse yet personalized offerings (Alt and Puschmann 2012; Granados and Gupta 2013; Hunwen and Sia 2015; Hedley et al. 2006). This development is a major driver of FinTech success and it explains why FinTechs hold the potential to disrupt whole branches of the financial services industry: FinTechs are often able to understand their customers better than incumbents and thus address their needs more effectively (Mackenzie 2015).

Incumbents' actions are often constrained by legacy systems, resulting in tension and the need to transform and adapt to digital technologies (Gregory et al. 2015) while also meeting institutional expectations from, e.g., regulators and analysts (Benner and Ranganathan 2012; Benner and Ranganathan 2013). In addition to a decline in customers' trust, many traditional financial services companies are affected by stricter regulations as a consequence of the financial and EURO crises (Alt and Puschmann 2012). In contrast, FinTechs are apparently less affected by these developments and the opposite seems to be the case:

Thirty eighth International Conference on Information Systems, Seoul 2017 1

Abstract: Redacted in this version

Citation: Redacted in this version

Keywords: FinTech; Business Models; Taxonomy; Digital Transformation; Financial Technology; Startups

II. Research Area: Methodological

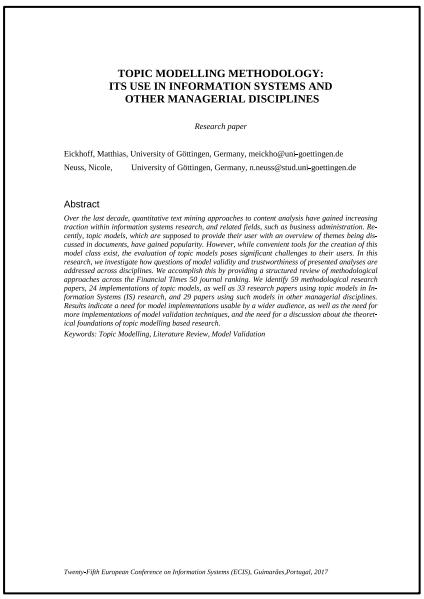
This research area provides a closer look at the text mining methodology used in the third research area. Two approaches are discussed in this area. First, a literature review of the available topic modelling methodology and its use in information systems research and other entrepreneurial disciplines is conducted. This highlights how this methodology is applied in these domains and what best practices and pitfalls can be learned from these examples. Second, a hybrid framework for sentiment analysis, using dictionary and machine learning based methods is presented. The discussion of the strengths and weaknesses of the individual approaches are an important basis for the selection of the approaches taken in research area III. This area addresses the following research questions:

Research Question II.1: What is the state of the art of topic mining methodology used to process unstructured content in the methodological literature, and how are these methods being applied in the managerial sciences to provide meaningful information relevant to researchers and decision makers?

Research Question II.2: How can dictionary-based and machine learning-based sentiment analysis be combined to mitigate some of their individual shortcomings, such as the need for labeled training data?

II.1. Topic Modelling Methodology Review

Topic Modelling Methodology: Its Use in Information Systems and other Managerial Disciplines



TOPIC MODELLING METHODOLOGY:
ITS USE IN INFORMATION SYSTEMS AND
OTHER MANAGERIAL DISCIPLINES

Research paper

Eickhoff, Matthias, University of Göttingen, Germany, meickho@uni-goettingen.de
Neuss, Nicole, University of Göttingen, Germany, n.neuss@stud.uni-goettingen.de

Abstract

Over the last decade, quantitative text mining approaches to content analysis have gained increasing traction within information systems research, and related fields, such as business administration. Recently, topic models, which are supposed to provide their user with an overview of themes being discussed in documents, have gained popularity. However, while convenient tools for the creation of this model class exist, the evaluation of topic models poses significant challenges to their users. In this research, we investigate how questions of model validity and trustworthiness of presented analyses are addressed across disciplines. We accomplish this by providing a structured review of methodological approaches across the Financial Times 50 journal ranking. We identify 59 methodological research papers, 24 implementations of topic models, as well as 33 research papers using topic models in Information Systems (IS) research, and 29 papers using such models in other managerial disciplines. Results indicate a need for model implementations usable by a wider audience, as well as the need for more implementations of model validation techniques, and the need for a discussion about the theoretical foundations of topic modelling based research.

Keywords: Topic Modelling, Literature Review, Model Validation

Twenty-Fifth European Conference on Information Systems (ECIS), Guimarães,Portugal, 2017

Abstract: Over the last decade, quantitative text mining approaches to content analysis have gained increasing traction within information systems research, and related fields, such as business administration. Recently, topic models, which are supposed to provide their user with an overview of themes being discussed in documents, have gained popularity. However, while convenient tools for the creation of this model class exist, the evaluation of topic models poses significant challenges to their users. In this research, we investigate how questions of model validity and trustworthiness of presented analyses are addressed across disciplines. We accomplish this by providing a structured review of methodological approaches across the Financial Times 50 journal ranking. We identify 59 methodological research papers, 24 implementations of topic models, as well as 33 research papers using topic models in Information Systems (IS) research, and 29 papers using such models in other managerial disciplines. Results indicate a need for model implementations usable by a wider audience, as well as the need for more implementations of model validation techniques, and the need for a discussion about the theoretical foundations of topic modelling based research.

Citation: Eickhoff, M., and Neuss, N. (2017). "Topic Modelling Methodology: Its Use in Information Systems and Other Managerial Disciplines," In: *Proceedings of the 25th European Conference on Information Systems (ECIS)*.

Keywords: Topic Modelling, Literature Review, Model Validation

1 Introduction

The rise of social media platforms and the availability of news online have created textual "big data", which has outgrown the feasibility of in-depth qualitative analysis. Quantitative methods to the analysis of textual data, such as sentiment analysis (Liu, 2012), have consequently become an established tool in the methodological spectrum of information systems research. Recent developments, such as efforts towards a "web of data", will only increase the need for an automated analysis of textual content (W3C, 2013). Among the approaches to analyzing large document collections, topic models, such as Latent Dirichlet Allocation (Blei et al., 2003), have recently gained traction in applied (non-methodological) research. Debortoli et al. (2016) provide a tutorial for using topic modelling as a tool in information systems research and provide readers with an example analysis showcasing the use of this model class. The recent focus on topic modelling as a quantitative research method has enabled researchers to address questions that previously would have been considered out of reach. As noted by Rai (2016), evaluation strategies for topic modelling include the reference to expert opinion, as well as quantitative approaches, such as the comparison of models estimated using varied parameters. However, modelling the contents of document collections is a challenging task and remains an area of active research in natural language processing and computer science literature. The "unreasonable effectiveness" (Halevy et al., 2009) of current models representing large document collections continues to be a challenge regarding the question on how to use these models in social-sciences and information systems research. In research concerned with testing hypothesis on the basis of theory, it is of critical importance to be able to convince readers that the models actually represent large document collections accurately, in order to establish the trustworthiness of conclusions based upon the models (Lincoln and Guba, 1985). In this paper, we investigate how researchers across different disciplines deal with this problem by conducting a structured review of literature in the top outlets of business related literature, on the basis of which we categorize different strategies to address this challenge. The paper is structured as follows: In section 2, the concept of topic modelling is introduced before a brief introduction to the relation between topic modelling methodology and (meta) theoretical considerations is given, based on which we discuss some of the results of the review. In section 3, the research design of this study is developed and presented. Section 4 presents and discusses the results of the review, while section 5 summarized this research.

2 Topic Models

The aim of topic modelling is to determine structures in underlying document collections. Initially, topic models were developed as an information retrieval tool, intended to make browsing large document collections easier (Salton et al., 1975). For example, topic models can be used to browse collections of scientific journals according to the subject of articles, without relying on metadata (Blei and Lafferty, 2009a). The first widely used model in this class was Latent Semantic Indexing (LSA), which as this review shows is still a popular option (Croft and Harper, 1979; Dumais et al., 1988). LSA extracts the underlying topics from a term-document matrix by applying singular value decomposition (SVD), which results in mathematically orthogonal topics. While this assumption of orthogonality contradicts human intuition about topics, topic models are essentially a data compression technique and this approach leads to the maximization of topic variance on a compressed representation of the document collection like how principal component analysis (PCA) does when used to reduce the number of features in a regression problem, which many researchers may be more familiar with. This assumption of topics' mutual exclusiveness is softened by probabilistic LSA (pLSA), which models topics as word distributions (Hofmann, 1999), leading to a notion of topics more in line with human intuition. After all, we would not assume most topics to be completely distinct from one another. This model type is extended upon by Latent Dirichlet Allocation (LDA), which differs from pLSA by imposing Dirichlet distributed priors to its word to topic and topic to document distributions (Blei et al., 2003).

Again, this is more in line with the human notion of topics, as it leads to sparse topic assignments to documents due to the sparse nature of the Dirichlet distribution. What this means is that not each document is a mixture of all topics in a model, but few

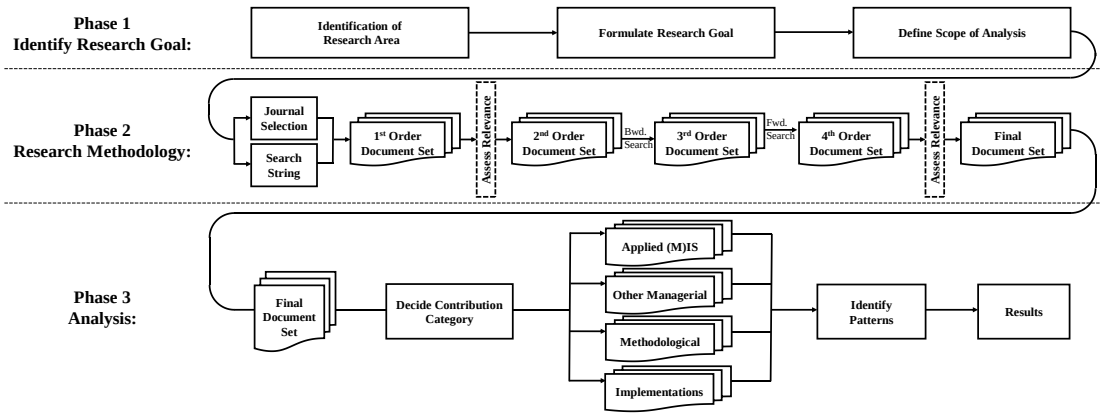


Figure 14: Research design segmented in three phases. Following Ngai et al. (2011). The first phase identifies the scope and goal of the presented research. Phase 2 describes the methodology of the conducted literature review. Phase 3 gives an overview of the analysis conducted based on this literature review.

topics are disproportionally more important for a document than others. A more detailed description of the intuition underlying this model type can be found in Blei (2012). This increased resemblance of human intuition does not necessarily mean that the more modern approaches outperform their precursors. For example, Bergamaschi and Po (2014) implement a plot-based movie recommender system and find that LSA outperforms LDA in their example. Since these methods became available, they have also been applied in (M)IS research. In recent years, this use has also arrived in top-tier outlets within the discipline (Kulkarni et al., 2014; Larsen and Bing, 2016; Sidorova et al., 2008). Superficially, it seems that M(IS) is not making use of the newer model types and their advantages regarding their closeness to human intuition, but determining whether this is the case requires a closer look and is one of the objectives of the following review.

2.1 Meta theoretical Foundations of Topic Modelling Research

When analyzing approaches to automated text analysis, such as topic mining, taking a step back to look at the (meta) theoretical foundation of such approaches can help in the analysis of their use. Ignatow (2015) provides an overview of the theoretical foundations of digital text analysis and argues that the meta theoretical foundations of such methods have not been sufficiently established and that applied social research using them often lacks adequate theories of language supporting the use of the method. According to Ignatow (2015), this lack of theoretical foundation stems from the unique positioning of automated text analysis between the natural and social sciences and weakens its relative positioning in comparison to exegetical methods and other inductive qualitative approaches. In principle, there are three possible meta theoretical foundations of text mining research resulting in three types of research designs. First, realist designs use models of text in a positivist framework to develop testable theories. See Elder-Vass (2014) for an extensive discussion of different variants of this approach. Second, constructivist designs use models of text to augment exegetical methods for qualitative text analysis, such as Grounded Theory (Lai and To, 2015). Third, mixed methods research designs (Venkatesh et al., 2016). Such studies often have comparably rigorous meta theoretical underpinnings because they are not conducted within the "safety" of either positivist or constructivist reference frames. While the IS, like many disciplines, traditionally focused on research designs build upon positivist mindsets, recently both qualitative (Bagozzi, 2011; Gregory, 1993; Mingers, 1995) and mixed methods research designs have become a common sight in the discipline (Ågerfalk, 2013; Venkatesh et al., 2013).

3 Research Design

As described, the goal of this analysis is to provide insight into the available methods for topic mining, and how these methods are applied both in (M)IS and other managerial disciplines. Ngai et al. (2011) present a similar analysis for the applications of data mining techniques within the domain of financial fraud detection and structure their review into three distinct research design phases, which represent a suitable research design for the case at hand. Thus, a comparable three-stage design is chosen for this study. In the first stage, the research goal is defined and the analysis is scoped. In the second phase, the research-methodology is outlined. In the third phase, we describe how the study is conducted on this basis. Figure 14 shows this process.

1.1 Phase 1: Identify a Research Goal

By determining the area of research, formulating the goal of the study, and defining the scope of the research, the studies relation to the wider research landscape is determined. In this research, the area of research is given by the search for available methodology for the training and evaluation of topic models, as well as their application in (M)IS and other managerial sciences. The goal of this study is to identify methodological opportunities for future studies and to examine how prior research has used the available methodology. To strengthen the focus of the analysis, this goal is formalized to the following three research questions:

RQ1 (Methodological pervasiveness): How widespread is the usage of topic models in the management literature and for what purposes are these models used therein?

RQ2 (Validation methodology): How do researchers address the problem of establishing trust into the results of their analysis when using topic models to analyze large document collections?

RQ3 (Interdisciplinary differences): How does the usage of topic models differ between M(IS) and other managerial disciplines?

The scope of the review is presented by an initial search within all journals included in the Financial Times 50 (FT50) ranking, which represents major outlets across numerous management-related fields. Further outlets are accepted into the study if they are deemed relevant regarding the aims of the study and are discovered by the structured literature review described in the next section. In order to formalize this relevance criterion, the following relevance definition is used throughout this research: Research is considered relevant for the scope of this study, if it falls into one of the four categories outlined in Table 4.

Category Name	Description
Methodological Foundations	A methodological contribution towards topic modelling, either a new topic modelling approach, a task specific document pre-processing logic, or an evaluation method for topic models, which is sufficiently different from other approaches included in the review.
Implementation (DSR)	An implementation of any of the above, which is made available to the public in a usable state, which means the software should be available and working.
Applied (M)IS (Empirical)	Applied research papers using topic modelling, or methodological considerations regarding topic modelling, within the IS community.
Applied Non-IS (Empirical)	Applied research papers using topic modelling, or methodological considerations regarding topic modelling, within studies from management related fields.

Table 4: Relevance criteria for literature. As discovered during the literature search. Structured into four relevance categories.

3.1 Phase 2: Research Methodology

Our goal in this phase is to arrive at a formalized abstraction of the conducted research process. This serves two purposes. First, the resulting design helps when conducting the study by splitting the research process into individual work units. Second, it helps readers to assess the quality and rigor of a study by providing a clear indication on how the study was conducted. In the case at hand, the first task during this phase is given by the identification of a suitable approach to the identification of literature using the relevance criterion stated in phase 1, resulting in the question what ways of literature exploration have been identified by methodological literature regarding literature reviews. Due to the continuous growth of the IS discipline, and the need for junior researchers to gain an overview of extant research, as well as the increasing difficulty to remain knowledgeable for senior researchers (Templier and Paré, 2015), a growing body of work regarding the methodology of literature reviews has evolved. Webster and Watson (2002) may be considered the starting point of this methodological discussion within IS. Since, Greenhalgh et al. (2005), Sylvester et al. (2013), Rowe (2014), and Boell and Cecez-Kecmanovic (2015) are only a small sample of this diverse toolset of methodological approaches towards literature based research.

Webster and Watson (2002) are perhaps the most notable example of guidelines to performing a structured literature search in the IS literature. They propose to divide the search for literature into three steps. Figure 14 (phase 2) provides an overview of this approach. First, a set of outlets is identified in which to search for relevant articles. Second, the references of these articles are examined to identify prior work (backward search). Third, the results of the two prior search phases are used to perform a search for articles citing them (forward search). As noted above, the relevant outlets for the first phase have already been identified as the journals included in the FT50. To search for relevant articles in the online databases listing the journals, a search string needs

to be determined, which covers a broad spectrum of work related to topic modelling. The following string is used and was determined by iterating between including more search terms and removing those, which produce non-relevant results:

"Topic Mining" OR "Topic Model*" OR "Topic Distribution" OR "Hierarchical Dirichlet Process" OR "Multinomial Asymmetric Hierarchical Analysis" OR "Latent Dirichlet Allocation" OR "Latent Semantic Indexing" OR "Latent Semantic Analysis" OR Mallet OR Gensim.

As shown, the string contains several relevant variants of "Topic*", where the star denotes the appropriate *any* search wildcard for each database. Furthermore, different topic modelling techniques are included, as well as MALLET (McCallum, 2002) and Gensim (Řehůřek and Sojka, 2010), which represent two popular implementations of topic models. These two are included because, as opposed to most other topic modelling software, they do not include topic modelling in their name. As the result of the literature search showed, most papers can be identified using either "Topic Mining" or "Topic Model*". The search string is used to search for titles, abstracts, keywords, as well as the full text of papers. Initially, a longer search string was used, which also included abbreviations where applicable, such as LDA in addition to Latent Dirichlet Allocation, however the results of searches including the abbreviated terms do not provide more relevant results and instead clutter the search results with other meanings for the abbreviations, which are not related to topic modelling. Regarding the reviews' scope in time, no assumptions were made during the initial search, but results indicate that no relevant content exists before 1978. Of course, arguably, text pre-processing literature precedes this year but this literature is not specific to topic modelling as a research method. The database search using this search string resulted in 108 results (1st order document set, Figure 14 phase 2). These documents were consequently assessed using the criteria outlined in Table 4, resulting in 23 2nd order documents. On this basis, the backward search resulted in 86 additional papers, increasing the 3rd order document set to 109 candidates. The forward search added another 44 papers, resulting in 153 documents. At this state, due to the large number of documents in the analysis, we conducted another relevance check and 8 documents were removed. The remaining 145 documents were assigned to the four relevance categories outlined in Table 4, resulting in 33 "Applied IS" papers, 29 "Applied Non-IS" papers, along with 24 implementations and 59 methodological contributions. Figure 14 (phase 3) provides an overview of this categorization into methodological research, implementations, and applied research papers stemming from M(IS) or other managerial disciplines. The analytic part of this research is based on this final set of documents.

3.2 Phase 3: Analysis

Methodological work: First, the methodological works are reviewed, to arrive at an overview of the available methodology, which can be used by applied studies. To this

end, the main methodological contribution of each paper is identified by examining each paper in the sample and summarizing its main contribution. Based on the sum of these contributions, the typology shown in Figure 15 is developed, which considers six archetypes of contribution. It should be noted that this is not a formal typology or taxonomy, in which the characteristics of each paper would be mutually exclusive from one another (Nickerson et al., 2013). Of course, a paper can contribute in more than one way regarding these categories. For methodological papers, a *model type* contribution is given by the introduction of a new topic model, which may be done by using an entirely new approach (Blei et al., 2003), augmenting existing approaches (Blei and Lafferty, 2006), or changing what is being modelled (Chang and Chien, 2009). *Computational or mathematical* works included in the sample concern algorithms or data structures of special importance to topic modelling.

Purely statistical work, such as numeric optimization techniques, are not included in this review because, typically, they are not inherently interesting for applied work (although, of course, they are of prime importance to enable it). *Introductory* papers provide a methodological overview, or give recommendations for the use of a specific model type. Often, these papers are domain specific. *Comparative* papers assess multiple model types, to determine which model is most suitable for a specific domain or document type. *Utility* papers help researchers by providing guidance apart from modelling itself. For example, Mei et al. (2007) develop a method for automated topic-label generation. *Model validation* is concerned with assessing topic model quality and often develops or compares metrics for this task. Finally, a research *domain* is identified for each paper. The list of discovered methodological papers is presented in Appendix A.

Implementations: If methodological papers make their method publicly available or applied papers mention the used implementation, this is added as a separate reference category. Each implementation is checked regarding its public availability. An implementation is considered public if it is usable (license not considered) and a download is available. Also, since many implementations are software libraries and not standalone programs, the programming language used for each implementation is noted. The list of discovered implementations is presented in Table 5.

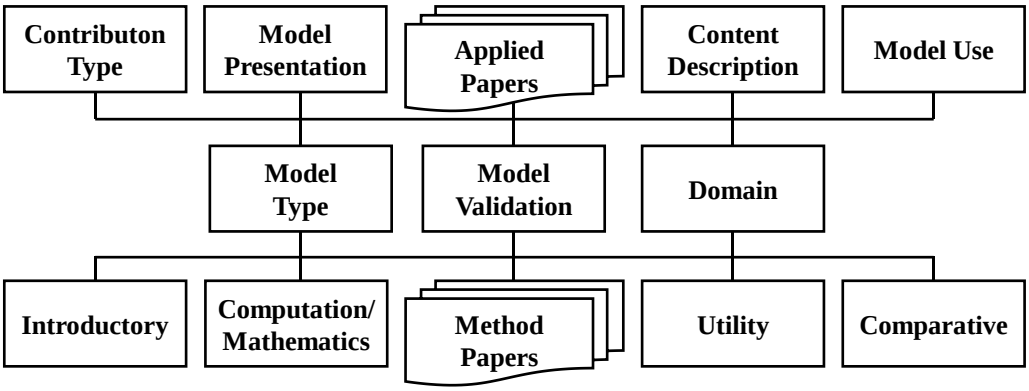


Figure 15: Literature assessment categories. For methodological and applied contributions. The set of applied categories is identical for IS and other managerial literature. Methodological contributions are assessed using a separate but partially overlapping set of categories.

Applied work: Likewise, applied papers identified as relevant to the review are assessed. This is done by a two-pass procedure. First, like the treatment of methodological papers, each applied paper is assessed regarding its *main contribution* (not in figure)**.** These main contributions are consequently divided into *contribution types*. Second, the papers are assessed regarding their use of topic models. Again, the *model type* used in a paper is determined. *Model use* describes to what end the topic model is used in the paper. For example, if it is used to inspect topics or if the topics are used in a regression model. Consequently, *model validation* (how the model is validated) and *model presentation* (how the model is presented to readers) are derived by examining each paper in more detail. Finally, a *domain* is coded for each applied paper. As shown, the model type, model validation, and domain criteria are shared between the two paper categories, while the other criteria are distinct for each paper type. The list of discovered applied papers is presented in Appendices B and C.

4 Results and Discussion

The answer to RQ1 is presented by Figure 16, which shows the annual paper counts across the different review categories and the distribution of all papers in appendices A-C over their respective disciplines. As shown, while in earlier years most discovered contributions are methodological, more recently this relation has inversed and topic models are being used in applied studies more frequently. It is important to keep in mind that recent methodological work is less likely to be discovered by the backward- and forward search mechanism applied in this review. Nonetheless, this still shows the growing relevance of topic modelling methodology for both (M)IS and other managerial disciplines. **Methods (Appendix A):** As shown in the result tables and Figure 16, topic modelling methodology has become a vibrant research subject. Starting with LSA (Dumais et al., 1988) and LDA (Blei et al., 2003), which represent the two most

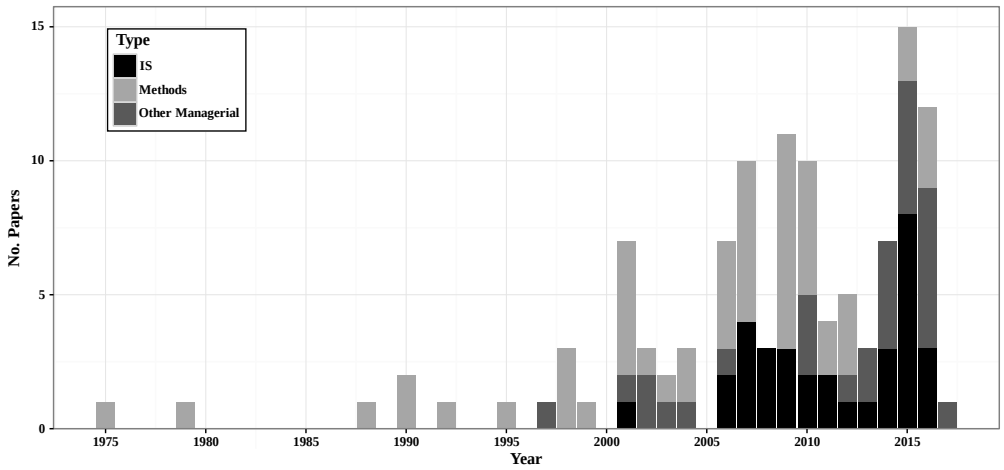


Figure 16: Annual distribution of contributions. Grouped by applied (M)IS and other applied managerial contributions, as well as methodological contributions. As shown, in recent years applied contributions have begun to outpace methodological works.

common model types used in applied papers, 25 model types are identified in this review, many of which focus on extensions of the two archetypes. Notably, despite beginning the literature review with the FT50 journal selection as a starting point, many methodological contributions stemming from the IS domain were identified. Thus, IS seems to have been established as a reference discipline for researchers looking for methodological guidance in the use of topic models. As can be expected, computer science (CS) and statistics present the two other most important methodological disciplines for topic modelling. **Implementations (Table 5):** As noted, implementations of the methodological contributions included in the review are reported separately if they are available, to highlight methods readily available for use. Unfortunately, while most methodological contributions explain the statistical approach to their work, publicly available implementations remain the exception. BleiLab (2016) provide an example of enabling others to benefit from methodological work and release implementations and working examples when possible.

When comparing, the citation counts of papers with released implementations to those without, it is obvious that this approach pays off. Accessibility remains a major problem (Ramage et al., 2009). All publicly available implementations identified in this review are either programming libraries or command line applications. In the interest of making topic models as usable as other statistical techniques, implementations with graphical user interfaces are needed.

Applied research papers: Table 6 provides a condensed overview of the results regarding applied research contributions, and contrasts M(IS) with other managerial disciplines. Appendices B and C report the full results for these categories in more detail, such as brief descriptions for each contribution. As shown, papers applying topic modelling are dominated by the two ‘basic’ model types LDA and LSA. When comparing

IS to other domains, LSA is favored over LDA, while this relation is reversed elsewhere.

Also, IS research included twice as many discussion and review articles when compared to other managerial disciplines, with 30% of IS articles being reviews and 27% being discussion pieces. Also, while 69% of non-IS articles use topic models for content analysis, only 12% of IS articles do so. In IS research, 36% percent of papers actively validate a model, while 59% of non-IS articles do so. Thus, regarding RQ2, we find that while many applied research papers make use of the topic model validation techniques proposed by methodological contributions, many researchers who use the topic model as a part of regression models abstain from a dedicated validation of the topic model.

Citation	Title	Public	Comment	URL
(Miller and Fellbaum, 1998)	Wordnet	Public	Digital dictionary for word relations	https://wordnet.princeton.edu
(McCallum, 2002)	MALLET	Public	Java, several topic models	http://mallet.cs.umass.edu
(Blei et al., 2003)	LDA-C	Public	C, Blei et al. (2003)	https://github.com/blei-lab/lda-c
(Blei and Lafferty, 2007)	Correlated topic model	Public	C, correlated topic model (CTM)	https://github.com/blei-lab/ctm-c
(Blei and Lafferty, 2009b)	Turbotopics	Public	Python, multiword phrases in topics	https://github.com/blei-lab/turbotopics
(Chong et al., 2009)	Supervised LDA for classification	Public	C++	https://github.com/blei-lab/class-slda
(Řehůřek and Sojka, 2010)	Gensim	Public	Python, several model types, flexible	https://radimrehurek.com/gensim
(Gerrish and Blei, 2010)	Dynamic and Influence Topic Model	Public	Command line implementation	https://github.com/blei-lab/dtm
(Hoffman et al., 2010)	Online var. Bayes for LDA	Public	Python	https://github.com/blei-lab/onlinelda-vb
(Wang and Blei, 2010)	Hierarchical Dirichlet Process	Public	C++	https://github.com/blei-lab/hdp
(Crossno et al., 2011)	TopicView	No	-	-
(Grün and Hornik, 2011)	Topicmodels (R-Package)	Public	R implementation of LDA	https://cran.r-project.org/web/packages/topicmodels/index.html
(Ramage and Rosen, 2011)	Stanford topic modeling toolbox	Public	Not maintained anymore	http://nlp.stanford.edu/software/tmt/tmt-0.4
(Wang and Blei, 2011)	Collaborative modeling	Public	C++	https://github.com/blei-lab/ctr
(Wang, 2011)	Online Hierarchical Dirichlet Process	Public	Python	https://github.com/blei-lab/online-hdp
(Zhai et al., 2012)	Mr. LDA	Public	-	https://github.com/lintool/Mr.LDA
(Roberts et al., 2014)	R, stm: structural topic models	Public	-	https://cran.r-project.org/web/packages/stm/index.html
(Sievert and Shirley, 2014)	LDAvis	Public	R Package for Visualization	https://github.com/cpsievert/LDAvis
(Chaney, 2014)	Online Topic Model Visualization	Public	Python, browsing topics	https://github.com/blei-lab/tmv
(Gopalan et al., 2014)	COLLABTM	Public	Nonnegative Collaborative Modeling	https://github.com/blei-lab/collabtm
(Blei, 2014)	Hierarchical latent Dirichlet allocation	Public	C,Hier. LDA,fixed depth tree and a stick breaking prior on the depth weights	https://github.com/blei-lab/hlda
(Günther et al., 2015)	LSAfun	No	R Package for LSA	https://cran.r-project.org/web/packages/LSAfun/index.html
(Charlin et al., 2015)	Dynamic Poisson factorization (dPF)	Public	Command line implementation	https://github.com/blei-lab/DynamicPoissonFactorization
(Ranganath et al., 2015)	Deep Exponential Family	Public	Command line implementation	https://github.com/blei-lab/deep-exponential-families
(BleiLab, 2016)	Blei Group Implementations	Public	David Blei Github repository, many implementations (see this table).	https://github.com/blei-lab

Table 5: Implementations identified by the literature review. If available, the citation of the methodological research paper is provided. If no such paper could be identified, a web reference is provided pointing to the implementation itself.

Model Type	Word Collocation	Naïve Bayes	Hierarchical	LDA	LSA	LSA&LDA	CTM	SOM	SVD	Clusters	None
IS	3%	3%	0%	18%	45%	0%	3%	0%	3%	3%	21%
Other	0%	3%	3%	41%	24%	3%	7%	7%	0%	0%	10%

Paper Type	DSS	Information Retrieval	Review	Statistical	Text Similarity	Tool	Tutorial	Validation	Discussion	Content Analysis
IS	3%	12%	30%	0%	3%	3%	3%	6%	27%	12%
Other	0%	0%	14%	3%	0%	0%	10%	0%	3%	69%

Validation	Yes	No
IS	36%	63%
Other	59%	41%

Table 6: Overview of applied research contributions.Papers in information systems compared with other managerial domains. Note that for the purposes of this summary table, less granular categories are reported than in the detailed tables.

One reason for the omission of a dedicated validation may be presented by the argument that if a topic model produces topics which are useful as variables in the context of statistical analysis, this itself validates the model for the purposes of these studies. However, the presentation of the topic model in such cases should be especially careful, to establish trustworthiness of the presented analysis. However, the lack of implementations of such model types in software intended for the use by social scientists remains a major hurdle for such work. 53% of all applied articles stem from the IS domain, followed by accounting research with 10%, while general management and marketing are tied at 8% each. 39% of all applied papers use their topic model as a tool for content analysis. Models are mostly used as a variable augmenting existing regression models, or to gain a general sense of topics included in text collections. This type of article is much more common in other managerial disciplines, but IS research using the methodology to this end still exists. The second most common applied paper type is presented by review articles, which review research domains (Moqri et al., 2015; Sidorova and Isik, 2010) or journals (Cohen Priva and Austerweil, 2015; Wang et al., 2015). Overall, the review indicates a diverse research landscape using topic modelling as a content analysis tool, as well as many research papers using it to review entire disciplines or outlets. Regarding their use of topic models some of the discovered contributions distinguish themselves and can serve as examples providing interesting ways to describe the usage of a model in an applied paper or integrating topic models in an analysis in another interesting way, which sets them apart from papers that ‘end’ after a topic model has been estimated.

First, Bao and Datta (2014), who investigate risk types in corporate risk disclosures, highlight topic models’ capability to simultaneously discover and quantify categories in a document collection, coupled with an extensive model evaluation, which enables readers to assess the reliability of the presented approach. Their evaluation includes both quantitative measures for model fit, comparisons to alternative topic models, as

well as presentations of the chosen approach using graphs and word-clouds. Second, Paul and Girju (2009), who compare research domains using topic models, show how topic similarity between different models can be used to compare different document collections, and support their arguments using a mix of reporting the words included in their estimated topics and graphs showing the evolution of topic similarity over time. As these examples show, how a model is displayed in a contribution is crucial to establishing trust in presented results. Going one step beyond the idea of presentation, Ramage et al. (2009) argue that readers should be able to explore models for themselves. Mützel (2015), a sociologist, discusses the lack of student method training in topic modelling, and data processing in general, as another challenge hindering the integration of the method in non-technical domains but notes that non-technical fields can draw on a vast experience regarding the study of meaning, which can support the automated analysis of large data sets, raising the question of the **theoretical foundations of topic modelling**: Using topic models for the purposes of advancing theory has been one of the uses of this model type early on (Landauer and Dumais, 1997) but remains the exception when surveying the applied literature. As discussed, accessibility may present one major cause for this. Another is given by the lack of theoretical foundations of topic modelling, which makes it more challenging to establish trust in results based upon its use. While few studies explicitly state their (meta) theoretical foundations, for studies classified as *content analysis* a positivist underpinning aiming at the empirical validation of established theory is often implicitly clear. On the other hand, constructivist foundations or mixed methods approaches to the analysis of topic models remain largely unexplored. However, similarities and differences between topic modelling and human coding have been discussed (Quinn et al., 2010). Also, many studies use topic labels coded from the top words of topics as a tool to present their results. Yet, this is usually done for presentation only and not using qualitative methodology, which may be suitable for this purpose. Since qualitative researchers have developed rigorous coding techniques, this methodology can support quantitative topic modelling creating opportunities for collaboration. Thus, the combination of qualitative methodology and topic modelling remains an interesting opportunity for future research. Evans and Aceves (2016) survey text mining methodology and provide recommendations on how it can be used as a tool for theory generation in the social theory. Wagner-Pacifici et al. (2015) discuss similar issues with a focus on using big data to access knowledge about social phenomena. Ignatow (2015) remains the only article discovered in the review in which the theoretical foundations of topic modelling are discussed. However, these articles do not discuss topic models in particular. As shown, (M)IS has established itself as a reference discipline for other managerial fields regarding topic modelling methodology. The exploration of the theoretical foun-

dations of the use and interpretation of topic models, as well as their capabilities regarding the generation and testing of social-, economic- and systems theory present an opportunity to strengthen this referential role of IS.

5 Conclusion

In this review, we surveyed the topic modelling literature regarding the methodological possibilities and uses thereof in applied research papers. To this end, we formalize our research design and conduct a structured literature review, resulting in a sample of topic modelling methodology and applied research papers in IS and other managerial disciplines. Also, we provide an overview of available implementations of topic modelling approaches. Our results indicate that while, in recent years, topic modelling has become a tool used across many disciplines and is especially prevalent in IS research, most researchers use "vanilla" LSA or LDA, instead of more specialized modelling approaches. A likely reason for this focus on two approaches is given by the lack of publicly available implementations for many methods. However, some researchers have had great success with making their implementations available (BleiLab, 2016). More such "open-access methodology" is needed to advance the use of topic modelling methodology in IS and other domains, especially regarding model validation, which many toolkits for topic modelling do not yet address as a priority and the resulting lack of methodological accessibility remains a problem (Ramage et al., 2009).

Looking at the non-IS research landscape, there is a need for modelling tools which are suited to the needs of researchers who do not use command line interfaces or software libraries, as there are no graphical user interfaces for most available implementations. A key factor in the quality of topic modelling based research is given by the presentation of the model in a paper. As discussed, looking beyond the boundaries of individual disciplines can help to identify successful solutions to this task. Also, the (meta) theoretical foundations of topic modelling remain to be established to make it easier to integrate the methodology in studies aimed at validating or expanding theory in the social and managerial sciences. One promising avenue for the creation of this theoretical foundation is presented by mixed methods, aiming at combining the advantages of modelling large document collections with qualitative approaches to content analysis. In conclusion, topic modelling has become a useful tool for many researchers, but specialized models and the development of suitable implementations for applied researchers remain largely unsolved problems offering perspectives for future research.

Appendix A: Methodological Research Contributions

Citation	Main Contribution	Type	Domain
(Salton et al., 1975)	Document storage in vector space	Computational	CS
(Croft and Harper, 1979)	Document search without prior content information	Model	IR
(Dumais et al., 1988)	LSA Model	Model	CS
(Deerwester et al., 1990)	Document retrieval using higher order term relations	Model	IS
(Spence and Owens, 1990)	Words that statistically co-occur often have a contextual association	Model	Psychology
(Cutting et al., 1992)	Clustering as an information retrieval tool	Model	IR
(Raftery, 1995)	Bayesian model selection	Validation	Sociology
(Landauer et al., 1998)	LSI: Explanation and interpretation	Validation	Interdisciplinary
(Dumais et al., 1998)	Comparison of approaches to text categorization	Comparative	IS
(Papadimitriou et al., 1998)	LSI: Evaluation of method	Validation	CS
(Hofmann, 1999)	Probabilistic-LSI: Modelling approach	Model	CS
(Lee and Seung, 2001)	Algorithmic comparison regarding non-negative matrix factorization	Computational	IS
(Park et al., 2001)	Model including prior document knowledge	Computational	IR
(Heylighen, 2001)	Comparison of word sense disambiguation approaches	Validation	IR
(Turney, 2001)	IR using pointwise mutual information (PMI-IR)	Comparative	CS
(Hofmann, 2001)	Unsupervised Learning by Probabilistic Latent Semantic Analysis	Model	ML
(Visa et al., 2002)	Document comparison by prototype matching	Model	IS
(Blei et al., 2003)	Modelling document topics using latent topics (LDA)	Model	CS
(Griffiths and Steyvers, 2004)	MCMC approach to LDA inference	Model	Interdisciplinary
(Dumais, 2004)	Overview of LSI/LSA	Model	IS
(Wei et al., 2006)	Two hierarchical agglomerative clustering (HAC) techniques	Comparative	IS
(Blei and Lafferty, 2006)	Model similar to LDA but topics change over time	Model	CS
(Teh et al., 2006)	Mixture model similar to LDA for unknown number of topics	Model	Statistics
(Teh et al., 2006)	Hierarchical Dirichlet Processes	Model	Statistics
(Wallach, 2006)	Combining n-grams and topics for document description.	Model	CS
(Blei and Lafferty, 2007)	A correlated topic model (CTM), inter-topic relations	Model	Statistics
(Mei et al., 2007)	Automated label generation for multinomial topic models	Utility	CS
(Foltz, 2007)	Book chapter: Discourse coherence and LSA	Introductory	Interdisciplinary
(Landauer, 2007)	Book chapter: Interpretation of LSA as theory of meaning.	Introductory	Interdisciplinary
(Steyvers and Griffiths, 2007)	Book chapter: Introduction to probabilistic topic models (LDA)	Introductory	Interdisciplinary
(Graesser et al., 2007)	Book chapter: Case study: Using LSA as part of a tutoring system	Theoretical	Interdisciplinary
(AlSumait et al., 2008)	Adaptive Topic Models for Mining Text Streams	Model	IS
(Wallach et al., 2009b)	Empirical evaluation methods for topic modelling.	Validation	CS
(Lin and He, 2009)	Joint Sentiment and Topic model (JST).	Model	CS
(Chang and Chien, 2009)	Sentence based Latent Dirichlet Allocation (SLDA)	Model	CS
(Wang et al., 2009)	Using topic models for multi-document summarization	Model	Comp. Ling.
(Asuncion et al., 2009)	Algorithmic comparison regarding inference in topic models	Comparative	ML
(Wallach et al., 2009a)	Comparison of structured priors for LDA	Comparative	IS
(Blei and Lafferty, 2009a)	Book chapter: Introduction to topic models	Introductory	Interdisciplinary
(Liu et al., 2009)	Joint author community and topic modelling	Model	CS
(Blei and Lafferty, 2009b)	Visualizing topics with multi-word expressions	Validation	CS
(Du et al., 2010)	Topic modelling method incorporating document segmentation	Model	ML
(Lee et al., 2010)	Comparison of topic modelling methods	Comparative	IS
(Newman et al., 2010b)	Automated evaluation of topic coherence	Validation	Comp. Ling.
(Ramage et al., 2010)	“Labeled LDA” for tweet and user characteristics	Model	IS
(Newman et al., 2010c)	Automated evaluation of topic coherence	Validation	Comp. Ling.
(Newman et al., 2010a)	Visualizing search results and document collections using topic maps	Utility	IS
(Grimmer and King, 2011)	Unsupervised clustering and evaluation thereof	Model, Evaluation	Interdisciplinary
(Newman et al., 2011)	Improving topic coherence with regularized topic models	Validation	IS
(Lu et al., 2011)	Topic modelling and multi-aspect sentiment analysis	Model	IS
(Nguyen et al., 2012)	Hierarchical nonparametric model using speaker identity	Model	Comp. Ling.
(Evangelopoulos et al., 2012)	Methodological recommendations for LSA studies	Introductory	IS
(Blei, 2012)	Overview article regarding probabilistic topic models	Introductory	CS
(Ramirez et al., 2012)	Automated topic model validation	Validation	CS
(Ignatow, 2015)	Discussion of theoretical foundations of textual analysis	Theoretical	Sociology
(Nikolenko et al., 2015)	Interval semi-supervised topic model (ISLDA) and coherence metric	Metric	IS
(George et al., 2016)	Model use cases in management research	Theoretical	Management
(Evans and Aceves, 2016)	Discussion of theory development based on text mining	Theoretical	Sociology
(Loughran and McDonald, 2016)	Overview of textual research in finance	Theoretical	Finance

Table 7: Methodological contributions. As identified by the structured literature review. As shown, a multitude of different models has been developed over the years. Also, quantitative validation strategies for their output have become an active area of methodological research. More recently, a discussion about the theoretical foundations and the use of topic models for theory testing and advancement has begun. Domains: Information Systems (IS), Computer Science (CS), Information Retrieval (IR), Machine Learning (ML), Computational Linguistics (Comp. Ling.).

Appendix B: Applied Research Papers (Other Managerial Disciplines)

Citation	Model Type	Content	Domain	Validation	Model Use	Presentation	Description
(Landauer and Dumais, 1997)	LSA	Content Analysis	Psychology	Human benchmark	Abstraction	Statistical	LSA for analyzing Plato's problem
(Back et al., 2001)	SOM	Content Analysis	Accounting	Benchmark	Annual reports vs. quant. Data	Plots, Labels	Use of SOM for annual reports
(Kintsch and Bowles, 2002)	LSA	Content Analysis	Language	-	Metaphor comprehension	Similarities	What makes metaphors difficult to understand?
(Landauer, 2002)	LSA	Tutorial	Psychology	-	Model meaning	Example models	Introduction to LSA as a representation of learning
(Wolfe and Goldman, 2003)	LSA	Tutorial	Behavior	Guidelines	Discuss model use	LSA similarity scores	Methodological guidance for LSA use in psychology
(Kloptchenko et al., 2004)	SOM	Content Analysis	Accounting	Qualitative clustering	Explain market variation	SOM example shown	Financial reports, information regarding fut. performance
(Boukus and Rosenberg, 2006)	LSA	Content Analysis	Accounting	-	Explain market variation	Labels	LSA of FOM minutes correlated w. economic conditions
(Li, 2010a)	Naïve Bayes	Content Analysis	Accounting	Cross-validation	Classify corp. Filings	Statistical	Using bayes classification for thematic and sentiment
(Quinn et al., 2010)	LDA	Content Analysis	Pol. Science	K-choice, extensive	Generate topics from political texts	Evolution	Topic modelling with political texts
(Grimmer, 2010)	Own (Hier.)	Content Analysis	Pol. Science	Over time variation	Per-author agenda	Evolution, result clustering	Measuring expressed agendas in pol. texts, new model
(Cicon et al., 2012)	LSA	Content Analysis	Finance	-	Cluster by topics	Theme clustering	Thematic analysis of corporate governance codes
(Grimmer and Stewart, 2013)	LSA,LDA	Content Analysis	Pol.Science	Validity measures	Discussion of use cases	-	Different models, assumptions, capabilities, problems
(Mohr and Bogdanov, 2013)	LDA	Tutorial	Language	-	Example model	Labels	Nontechnical introduction to topic models (LDA)
(Bao and Datta, 2014)	LDA	Content Analysis	Management	Perplexity, pred. validat.	As variable	Labels, word clouds	Identification of risk categories
(Campbell et al., 2014)	LDA	Content Analysis	Accounting	-	As variable	-	Information content of 10-K risk factor section
(Tirunillai and Tellis, 2014)	LDA	Content Analysis	Marketing	Dimension validation	Interpretation of topics/factors.	-	Consumer satisfaction dimensions (social media)
(Huber et al., 2014)	-	Review	Marketing	-	Topic evolution	Importance over time	Topics in JMR
(Huang et al., 2015)	LDA	Content Analysis	Accounting	Topic change	As variable	Labels	Analyst report topic modelling
(Kaplan and Vakili, 2015)		Content Analysis	Management	-	Ideas in patents	-	Topic modelling of patents
(Giorgi and Weber, 2015)	LDA	Content Analysis	Management	Word intrusion	Extract topics from analysts' reports	Labels	Analysts' framing repertoires and analyst evaluation.
(Cohen Priva and Austerweil, 2015)	LDA	Review	Cognition	-	Topic evolution	Top words, importance over time	Journal topic article: "Cognition"
(Wang et al., 2015)	LDA	Review	Marketing	-	Topic evolution	Labels, importance over time	50 Years "Journal of Consumer Research"
(Trusov et al., 2016)	CTM (no TM)	Content Analysis	Marketing	Accuracy	Profile clustering	Statistics for dimensions	Profiling in customer-base analysis, behavioral Targeting
(Castelló et al., 2016)	LSA	Content Analysis	Management	-	Analysis of tweet topics	First and second order topic labels	Stakeholders' sustainable development agendas
(Bellstam et al., 2016)	LDA	Content Analysis	Finance	k-choice by experiment	Topics and sentiment	Word clouds	Text-based measure of innovation using analyst reports
(Bendle and Wang, 2016)	LDA	Discussion	Management	-	Discussion of use cases	-	Discussion of LDA use cases in business
(Guerreiro et al., 2016)	CTM	Review	Ethics	Likelihood and perplexity	Key themes of research area	Discussion of each topic of interest	Review of cause-related marketing literature
(Jacobs et al., 2016)	-	Statistical	Marketing	Success rate	As variable	Statistical	Model-Based Purchase Predictions for Large Assortments
(Guo et al., 2017)	LDA	Content Analysis	Tourism	Benchmark	Compare to review ratings	Evaluation plots	Tourist satisfaction analysis

Table 8: *Applied papers in other managerial disciplines (non-IS). Results indicate a strong focus on the use of topics models as a tool for content analysis, which often involves using the topic to document assignments as variables in regression models. The temporal distribution of the discovered contributions within this category indicate a rapid increase in the use of topic modelling methods. While there are several tutorials and methodological advice papers within these fields, there is still room for future research regarding a broader spectrum of model use and topic model validation. While most studies within these fields use very extensive validation techniques for other statistical methods, topic model validation has not yet been adopted to the same degree. Likewise, most studies either use LSA or LDA, while there may still be many use cases for derivatives of these methods.*

Appendix C: Applied Research Papers (Information Systems)

Citation	Model Type	Content	Domain	Validation	Model Use	Presentation	Description
(Husbands et al., 2001)	SVD	IR	IS	Precision measure	Model is main contribution	Statistical	Using SVD for document retrieval
(Wei and Croft, 2006)	LDA	IR	IS	Average precision	Find similar documents	Model not shown	Using LDA for ad-hoc information retrieval
(Mihalcea et al., 2006)	LSA	Text Similarity	IS	Precision, recall, F-Score	Find similar documents	-	Corpus- and knowledge-based measures of similarity
(Wei et al., 2007)	Clustering	IR	IS	Performance metric	Model is main contribution	Statistical	Topic based query expansion for IR
(Arazy and Woo, 2007)	Collocation	IR	IS	F-score	Model is main contribution	Word collocation	Information retrieval using collocation indexing
(Sidorova et al., 2007)	LSA	Review	IS	-	Interpretation of topics/factors	Interpretation and description	Using LSA to identify research streams
(Graesser et al., 2007)	LSA	Tutorial	IS	?	Part of virtual tutor	?	Explanatory case study
(Titov and McDonald, 2008)	MG-LDA	Content Analysis	IS	LDA benchmark, metric	Model is main contribution	Labels	Extract aspects from product reviews (Multi-Grain LDA)
(Hall et al., 2008)	LSA	Review	IS	-	Interpretation of topics/factors	Interpretation and description	Using LDA to identify historical research trends.
(Sidorova et al., 2008)	LSA	Review	IS	-	Interpretation of topics/factors	Interpretation and description	Using LSA to identify research streams
(Ramage et al., 2009)	-	Discussion	IS	-	Exploration of output	Model should be explorable	Accessibility, trust in topic models (social sciences)
(Paul and Girju, 2009)	Naïve Bayes	Review	IS	-	Interpretation of topics/factors	Labels, evolution, inter-model	Topic comparison between research domains
(Chang et al., 2009)	LDA	Validation	IS	Benchmark (other metric)	Model output evaluation	Word and topic intrusion.	Quantitative metrics for semantic topic coherence
(Turney and Pantel, 2010)	-	Discussion	IS	-	Review article	-	Different text representations using vector space models
(Sidorova and Isik, 2010)	LSA	Review	IS	-	Exploration of output	Topic labels, importance	Review using LSA: Business process literature
(Aral et al., 2011)	LDA	Content Analysis	IS	Model comparison	As variable	Labels	Impact of stock recommendations on stock returns
(O'Connor et al., 2011)	-	Discussion	IS	Review article	Review article	-	Model complexity and model assumptions
(Chen et al., 2012)	-	Discussion	IS	-	Special issue about BI research	-	Overview of big data landscape, including topic models
(Jin et al., 2013)	LDA	DSS	IS	-	As variable	-	Forex trend modelling system
(Koukal et al., 2014b)	LSA	Discussion	IS	-	Literature review	-	LSA for literature reviews and prototype tool
(Kulkarni et al., 2014)	LSA	Review	IS	-	Literature review	Importance over time	Operations management research
(Koukal et al., 2014a)	LSA	Validation	IS	Purpose of article	Literature review	Benchmarks	Validation of Koukal et al. (2014b)
(Ahmad and Laroche, 2015)	LSA	Content Analysis	IS	-	Measure emotions	Statistical	Review helpfulness and emotions shown in review
(DiMaggio, 2015)	-	Discussion	IS	-	-	-	Different research perspectives in CS and social sciences
(Mützel, 2015)	-	Discussion	IS	Discussion article	Discussion article	-	Topic modelling in sociology, challenges, opportunities
(Wagner-Pacifici et al., 2015)	-	Discussion	IS	Review article	Literature review	-	Discussion: Big data in the social and cultural sciences
(Moqri et al., 2015)	LSA	Review	IS	No Full text	Literature review	No Full text	Identifying Research Trends in IS
(Chen and Zhao, 2015)	CTM	Review	IS	-	Literature review	Plots	Correlated topic model: Information systems
(Aryal et al., 2015)	LSA	Review	IS	-	Literature review	Period-comparison of key terms	Healthcare research
(Kundu et al., 2015)	LSA	Review	IS	-	Literature review	Importance over time	Supply chain management
(Müller et al., 2016)	LSA	Content Analysis	IS	Varying topic count	Interpretation of topics/factors	Term- and document loadings	Develop a typology of BPM professionals
(Rai, 2016)	LDA	Discussion	IS	-	Discussion article	-	Call for use of LDA for theory generation
(Larsen and Bing, 2016)	LSA	Tool	IS	Recall, Precision, F-Score	Construct identity	Constructs, graphs for evaluation	Addressing construct identity in literature reviews

Table 9: Applied research papers in Information Systems (IS). As shown, IS researchers have, so far, mainly used topic models for reviewing purposes in several contexts. Like researchers in other managerial disciplines, they focus on LSA and LDA for their studies. In comparison, more IS papers discuss the use of topic modelling methodology, while using the model as part of another analysis is less common. As was observed in other domains, the usage of topic models has recently spiked within the discipline.

II.2. Hybrid Sentiment Analysis Framework

(not included in this document due to copyright)

Enabling reproducible Sentiment Analysis: A hybrid domain-portable Framework for Sentiment Classification

Enabling reproducible Sentiment Analysis: A hybrid domain-portable Framework for Sentiment Classification

Matthias Eickhoff

Georg-August University, Göttingen, Germany
matthias.eickhoff@wiwi.uni-goettingen.de

Abstract. In this paper a hybrid framework for Sentiment Analysis is presented. In the first part, dictionary based and machine learning based Sentiment Classification are introduced and the two approaches are contrasted. In the second part of the paper, the HSentiR framework, which combines the two approaches, is introduced. Consequently, the framework is evaluated regarding scoring accuracy and practical concerns.

Keywords: Sentiment Analysis • Reproducible Research

1 Introduction & Research Problem

Content and sentiment analysis as fields of study have intrigued researchers for a long time. As early as the nineteenth century, the quality of newspaper-articles was studied on a statistical basis [1]. However, due to the exponential increase in readily available digital texts that also has resulted from the rise of social media, sentiment analysis has become one of the most active data mining topics. Popular techniques include the use of (1) machine learning approaches, such as of Support Vector Machines (SVM) and naïve Bayes classification, or (2) scoring by comparing the words in a text with a dictionary of sentiment words of known polarity [2]. These two approaches to sentiment analysis have specific advantages and limitations. While dictionary based scoring methods offer a higher level of domain portability than machine learning based ones, their ability to detect sentiment in a document remains limited to the used dictionary and, for optimal scoring accuracy, a domain-specific dictionary is desirable nonetheless [3]. On the other hand, machine learning based classifiers are typically not domain-portable at all because they are based on different statistical measures of similarity and consequently perform much worse, if the documents at hand are not comparable to initial training data [4]. Due to these drawbacks of the individual approaches, researchers have combined them in hybrid models, which strive to combine advantages of both methods. It has been shown that the combination of two or more methods can improve scoring

Abstract: In this paper a hybrid framework for Sentiment Analysis is presented. In the first part, dictionary based and machine learning based Sentiment Classification are introduced and the two approaches are contrasted. In the second part of the paper, the HSentiR framework, which combines the two approaches, is introduced. Consequently, the framework is evaluated regarding scoring accuracy and practical concerns.

Citation: Eickhoff, M. (2015). "A Hybrid Domain-Portable Framework for Sentiment Classification," In: *Proceedings of the 10th International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, B. Donnellan, M. Helfert, J. Kenneally, D. VanderMeer, M. Rothenberger and R. Winter (eds.): Springer International Publishing, pp. 215-219.

Keywords: Sentiment Analysis; Reproducible Research

III. Research Area: Analyst Opinion

In this research area the information value of analyst opinion is studied using several theoretical and methodological lenses.

Eickhoff and Muntermann (2016b, **paper III.1**) looks at the relation between the information processing of stock analysts and social media users. The paper examines under which circumstances social media users may be able to outperform the timeliness of analysts' assessments. Eickhoff and Muntermann (2017, **paper III.2**) is concerned with the risk of information overload arising from the wealth of information to investors, who have to make investment decisions. It uses topic modelling and sentiment analysis to reduce the decision complexity in this situation. Eickhoff and Muntermann (2016c, **paper III.3**) compares the content of two different sources of analyst opinion, analyst reports and earnings conference calls, and elaborates on the topic transfer between these two media types. Finally, Eickhoff (2017, **paper III.4**) applies media richness theory to the investment decision situation investigated in paper III.2, in order to provide explanations for the usefulness of unstructured analyst opinion.

> **Research Question III.1:** What structure is there to the relationship between the opinions of social media users and stock analysts, and can wisdom of crowds theory be used to identify the situations in which the crowd or stock analysts are more likely to provide timely information, reflecting changes in a firm's circumstance?
>
> **Research Question III.2:** What constitutes a decision-relevant metric in the context of business communications regarding a firm's earnings announcement, and how can the metrics of analyst opinion determined by sentiment analysis and topic modeling be used to provide such decision relevant information?
>
> **Research Question III.3:** To what extent do the topics contained in analyst reports that are released prior to an earnings call influence the topics contained therein, and does the call influence the content of reports released thereafter?
>
> **Research Question III.4:** To what extent can the media richness of unstructured analyst opinion, as described by media richness theory, help to explain its effect on post earnings call firm stock returns when compared to information sources of lower richness?

III.1. Stock Analysts vs. the Crowd

(not included in this document due to copyright)

Stock Analysts vs. the Crowd: Mutual Prediction and the Drivers of Crowd Wisdom

Information & Management 53 (2016) 835–845

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

Stock analysts vs. the crowd: Mutual prediction and the drivers of crowd wisdom

Matthias Eickhoff*, Jan Muntermann

University of Göttingen, Chair of Electronic Finance and Digital Markets, Germany

ARTICLE INFO

Article history:
Received 20 September 2015
Received in revised form 22 February 2016
Accepted 22 March 2016
Available online 28 May 2016

Keywords:
Wisdom of crowds
WoC
Sentiment analysis
Stock analysts
Social media

ABSTRACT

We examine the drivers of crowd wisdom in the financial domain by relating analyst report and social media sentiment via Granger causality (GC) testing based on the wisdom of crowds (WoC) theory. The significance of a large number of the tested time series indicates that analyst reports and social media content are suitable for mutual prediction. We elaborate on the conditions under which crowd cognitive diversity matters, and we derive related measures. The results suggest that the WoC theory can partially explain the GC between the two media types and that both professional analysts and the crowd can outperform one another under favorable circumstances.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The impact of financial analyst reports has been subject to increasing scientific scrutiny. In particular, the herding behavior of financial analysts, that is, how some analysts seem to affect the opinion of others [44], has been analyzed extensively [12,25,43]. The reaction of the capital market has also been analyzed. Traditional analyses have dealt with buy/hold/sell recommendations, from the Institutional Brokers' Estimate System (I/B/E/S). More recently, researchers have been applying text-mining methods to analyst reports to automatically extract more information than previously available. The analysis of the entire report is desirable, as it may yield further information beyond the constrained categorization of the stock [44]. In contrast to prior studies, this study focuses on how the opinions of professional stock analysts, that is, individuals who are paid to provide regularly updated opinions about certain companies, relate to those of social media users. The study also analyzes whether the two related content types can be used to predict each other's sentiment. Although studies have shown that social media content can be used to forecast stock returns similar to analyst recommendations, this does not necessarily imply that the prediction powers of the two mediums are inherently related. Such a relation is of interest because professional analysts have access to privileged information, as illustrated in Section 2.

However, analysts also face a number of constraints that can influence their recommendations. These constraints include incentives to generate in-house brokerage volume [27] and the tendency to stick to established recommendations [11,43]. The average social media user does not face these constraints.

Therefore, such a relation could allow the incorporation of social media content in models traditionally using traditional stock analyst recommendations for predictive purposes. This incorporation could serve as a control variable for the analysts' biases. Therefore, the aim of this research is to investigate the presence of such a relationship between content types in either direction. Social media and analyst report data regarding the 30 component companies of the Dow Jones Industrial Average (DJA) index are collected from the year 2013. First, we perform sentiment analysis on analyst reports and social media data to establish time periods (quarterly and annual) and company combination. Granger causality (GC) between social media and analyst sentiment. We then aim to determine reasons for different cases of GC direction. Afterthis section, the theoretical background of the analysis is introduced. Theoretical insight on the information value of analyst reports and the concept of wisdom of crowds (WoC) is compared in Section 2, focusing on inefficiencies in analyst opinions likely to be mitigated by social media content. Based on these foundations, hypotheses are developed and tested. In Section 3, methodological foundations are introduced and comprise sample generation, GC, sentiment analysis, and variable operationalization related to WoC constructs. We then present our analysis and empirical results in Section 4. Using binary response models, we provide evidence regarding which types of companies and public interest foster GC

* Corresponding author.
E-mail address: Matthias.Eickhoff@wiwi.uni-goettingen.de (M. Eickhoff).

http://dx.doi.org/10.1016/j.im.2016.03.009
0378-7206/© 2016 Elsevier B.V. All rights reserved.

Abstract: We examine the drivers of crowd wisdom in the financial domain by relating analyst report and social media sentiment via Granger causality (GC) testing based on the wisdom of crowds (WoC) theory. The significance of a large number of the tested time series indicates that analyst reports and social media content are suitable for mutual prediction. We elaborate on the conditions under which crowd cognitive diversity matters, and we derive related measures. The results suggest that the WoC theory can partially explain the GC between the two media types and that both professional analysts and the crowd can outperform one another under favorable circumstances.

Citation: Eickhoff, M., and Muntermann, J. (2016b). "Stock Analysts Vs. The Crowd: Mutual Prediction and the Drivers of Crowd Wisdom," *Information & Management* 53 (7), pp. 835-845.

Keywords: Wisdom of crowds; WoC; Sentiment analysis; Stock analysts; Social media

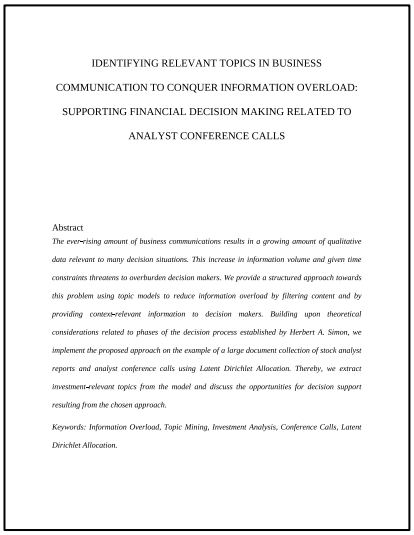
Prior Version: Eickhoff, M., and Muntermann, J. (2015). "Stock Analysts Vs. The Crowd: A Study on Mutual Prediction, ", In: Proceedings of the 19th Pacific Asia Conference on Information Systems, Singapore: AISeL.

Best Paper: The prior version received the best Completed Research Paper Award at the 19th Pacific Asia Conference on Information Systems.

III.2. Identifying relevant Topics in Business Communication

(not included in this document due to copyright)

Identifying relevant Topics in Business Communication to conquer information overload: Supporting Financial Decision Making related to Analyst Conference Calls



IDENTIFYING RELEVANT TOPICS IN BUSINESS COMMUNICATION TO CONQUER INFORMATION OVERLOAD: SUPPORTING FINANCIAL DECISION MAKING RELATED TO ANALYST CONFERENCE CALLS

Abstract

The ever-rising amount of business communications results in a growing amount of qualitative data relevant to many decision situations. This increase in information volume and given time constraints threatens to overburden decision makers. We provide a structured approach towards this problem using topic models to reduce information overload by filtering content and by providing context-relevant information to decision makers. Building upon theoretical considerations related to phases of the decision process established by Herbert A. Simon, we implement the proposed approach on the example of a large document collection of stock analyst reports and analyst conference calls using Latent Dirichlet Allocation. Thereby, we extract investment-relevant topics from the model and discuss the opportunities for decision support resulting from the chosen approach.

Keywords: Information Overload, Topic Mining, Investment Analysis, Conference Calls, Latent Dirichlet Allocation.

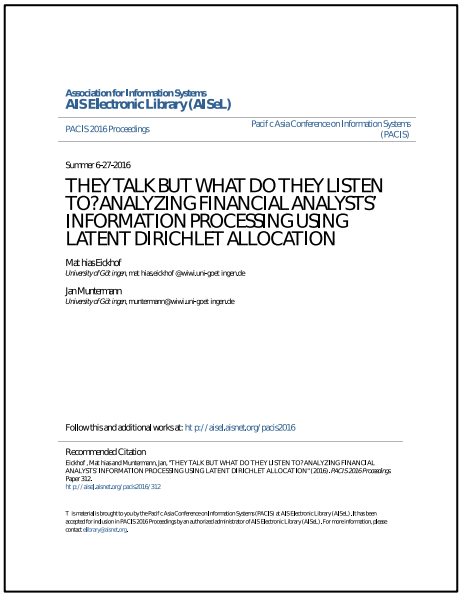
Abstract: Redacted in this version

Citation: Redacted in this version

Keywords: Information Overload; Topic Mining; Investment Analysis; Conference Calls; Latent Dirichlet Allocation

III.3. Topic Transfer between Earnings Calls and Analyst Reports

They talk but what do they listen to? Analyzing Financial Analysts' Information Processing using Latent Dirichlet Allocation



Abstract: In this study, we examine stock analyst information processing behavior on the example of information transfer between analyst conference calls and analyst reports. From a theoretical perspective, the study contributes to an understanding of analysts' recommendation biases resulting from their information processing. It provides new insights on how information is actually used by analysts, while practical implications for both sides of conference calls and other market participants are examined. Results indicate that analysts are exposed to new information during conference call events, which they consequently incorporate in their reporting.

Citation: Eickhoff, M., and Muntermann, J. (2016c). "They Talk but What Do They Listen To? Analyzing Financial Analysts Information Processing Using Latent Dirichlet Allocation," In: *Proceedings of the 20th Pacific Asia Conference on Information Systems*, Chiayi, Taiwan: AISeL.

Keywords: Topic Mining; Text Mining; Financial Analysts; Information Processing

1 Introduction

The rapid increase of available textual information has been discussed in research for a long time (Baker and McCallum, 1998). Financial analysts, who we define as professional analysts being paid to provide their analysis, as a group are especially exposed to this increase in amount and velocity of information, as they are, by job definition, expected to provide an analysis that encompasses the available information. Analysts always been facing an overwhelming amount of information as the stock markets have always provided amounts of information that escape an individual's scrutiny. Thus, the group may have developed coping mechanisms that are also suitable for adoption in other domains. However, the information processing of analysts has been criticized regarding a multitude of biases in the past, some of which we will discuss in the following section. Due to this union of capability and (at least partially) understood inefficiency, financial analysts provide an interesting research field regarding information processing behavior. In order to make this behavior visible, we employ Latent Dirichlet Allocation (LDA), a modern topic mining algorithm, to two kinds of textual data produced by financial analysts and their communication with company representatives. Using this topic-modelling approach, we study how topics transfer between the two content domains by tracking the topic-similarity around conference call events. This research is structured as follows. In following section, we provide theoretical background regarding financial analysts and their known biases described in literature. In the third section, we describe the two types of content used in our analysis and how we prepare the texts for the following analysis. Following the description of our pre-processing, we provide a short introduction to the applied methodology, which begins by explaining the LDA topic model and introduces cosine similarity as a measure for topic-distance in the context of our analysis. In the fifth section of the paper, we provide the results of our analysis for both a single conference call and the entire sample for varying subsample selections. Finally, we discuss the implications of this research for future research in this domain, as well as implications for concerned practitioners, and the transferability of the chosen approach to other domains. We conclude the paper with a discussion of limitations of the chosen approach and a result summary.

2 Theoretical Background

Professional financial analysts are known to exhibit a number of inefficiencies regarding their information processing habits. Prior research suggests that this is caused by a multitude of factors, such as herding behavior, i.e. the tendency to stick to the consensus estimate (Twedt and Rees, 2012), due to career concerns (Clement and Tse, 2005),

fear of being singled out in the case of wrong predictions (Hong et al., 2000) or inadequate incentive structures, focusing on increasing brokerage or investment banking revenue instead of rewarding correct predictions (Groysberg et al., 2011). While this prior research focuses on identifying inefficiencies in analysts' information processing, we analyze how analysts actually arrive at their conclusion. Specifically, we focus on the topics they talk about in conference calls and write about in their reports on the basis of which they finally justify their conclusion. Professional analysts are faced by a multitude of possible sources of unstructured information, such as newswire services, corporate disclosures or social media. Obviously, these sources of information are available to anyone truly interested in analyzing a firm's performance. However, analysts may also have access to privileged information. Since decreasing information asymmetry between market participants is a prerequisite for efficient capital markets, fair disclosure regulation aims at reducing the occurrence of privileged communication between firms and analysts. Still, the effect of such regulation may have unintended adverse consequences (Irani and Karamanou, 2003; Sunder, 2002). One possible channel for such privileged information is presented by analyst telephone conferences in which analysts are able to directly interact with high level, often C-level, representatives of the concerned company. As this degree of direct interaction with high-level employees is unusual, analyst calls are supposed to reveal new information about a company's current and future prospects. Even if the content of the calls is disclosed to the public in a timely manner, the call itself provides analysts with the opportunity to ask company representatives the questions that are important for their specific information needs, thus providing the analyst with information that may not be of value to other market participants. We apply topic-mining techniques to both types of content and examine the topic similarity between analyst conference calls and reports written prior to and after these telephone conferences. This allows us to compare the topic structure of the calls to reports in two directions. First, we examine how related reports that were released prior to the call events are to the topic structure of the calls. This allows us to assess if the topics discussed in the calls are 'new', or if calls mainly discuss topics already addressed in previous reports. This also relates to the question of analyst herding behavior. If analysts do not bring up new topics in the calls, this may be a sign of 'sticking to the herd'. Based on these considerations, we pose the following research questions**:**

(1) To what extend do financial analysts include analyst reports released prior to conference calls in their topic-selection during the call?

If conference calls are sources of valuable new information, the topics discussed therein should be picked up after the call, i.e. the reports following the call should show increased similarity to the call when compared to the ones released prior to the call. Thus, we pose our second research question:

(2) To what extend do financial analysts include conference call topics when writing post-call reports?

The answers to these questions are of both theoretical and practical importance. First, if analysts do not incorporate novel information revealed in analyst calls in their assessments, analyst calls are of little informational value to the analyst and merely a media outlet for companies. Second, if analysts do make use of novel information revealed in calls, it enables market participants to study their content and predict analyst opinions. While the topics discussed during conference calls are influenced by both corporate and analyst participants, the firm holding the call could try to influence the development of the call by adjusting their presentation, which is usually a comparatively long monologue at the beginning of the call. The same holds true for analysts themselves. By gaining insights into what topics have affected analyst behavior in the past, analysts may be able to recognize manipulation attempts by corporate representatives or improve their own information processing.

3 Data and Pre-Processing

For the purposes of this study, we make use of two sources of unstructured data, which we subject to modern text mining methods. For both data sources, we have collected comprehensive data for the multinational technology and consulting corporation IBM between 2000 and 2015. In this period, 59 calls and 4735 reports were collected. Of these 59 calls 48 will be used for the analysis as some calls do not satisfy the conditions required for later analysis, i.e. that they are surrounded by enough analyst reports. Both datasets were downloaded from Thomson Reuters Advanced Analytics (TRAA).

Analyst Earnings Call Transcripts: These transcripts contain the communication between corporate representatives and selected analysts. In the case of IBM, its vice president for investor relations, as well as its CFO typically represent the company, while a small group of analysts (~10) joins them. A typical call consists of an initial presentation by the company, followed by a Q&A section, in which questions by the analysts are answered by the corporate representatives.

Analyst Reports: These reports are available as PDF files from which we extract the textual content needed for our analysis. Furthermore, we identify the date on which reports were added to the TRAA database and the company responsible for their release. Both types of documents are pre-processed by removing ‘stopwords’, i.e. common English words which are not expected to help us establish the differences between texts due to their omnipresence and non-textual content, such as tables, figures, or reoccurring numbering (which would otherwise become a ‘topic’), before submitting these documents to the topic-mining described in the next section. Also, the disclaimer contained in each call transcript is removed for the same reasons. Consequently, both

types of documents are combined into a single textual corpus, stemmed, and converted to a bag-of-words vector space representation, i.e. a term-document-matrix (TDM). We filter the TDM for both sparse and very frequent terms, neither of which are expected to be helpful in establishing statistical differences between the texts, before dropping all documents, which contain no words satisfying these restrictions. We do not convert the TDM to a TF-IDF matrix because of theoretical considerations concerning the chosen topic mining algorithm (Blei et al., 2003).

4 Method

Our analysis will be conducted in two steps: First, we address RQ1 by utilizing Latent Dirichlet Allocation (LDA), a topic mining algorithm, on both types of documents and compute similarity scores between the resulting vectors of topic to document probability. If analysts incorporate novel information released in conference calls into their reports, similarity scores between the two types of content should increase if a call took place prior to the release of a report. RQ2 is addressed by interpreting our results regarding their theoretical implications and possible implications for both financial analysts and corporate representatives in analyst calls. Latent Dirichlet Allocation (Blei et al., 2003) generates a statistical model representing the latent topic distribution of a given document collection. The model describes the documents as a mixture of topics (Θ), which in turn are a distribution of the words contained in the documents (N). For both mixtures, a vector of assignment probabilities is calculated. Each word receives a (conditional) per-topic assignment probability and each topic is in turn assigned to each document with another conditional probability. Or formally, following the definition of the generative process by (Blei et al., 2003): A corpus **D** is a vector of documents **w**, each of which is in turn comprised of **N** individual words w_n. Thus, for each document vector **w** in corpus **D** the topic distribution over documents and per-topic word distributions are computed as shown in Table 10.

1.	Choose **N** ~ Poisson(ξ), i.e. a word distribution.
2.	Choose **Θ** ~ Dirichlet(α), i.e. a topic distribution.
3.	For each of the **N** words w_n:
	I: Choose topic z_n ~ Multinomial(**Θ**).
	II: Choose a word w_n from $p(w_n\|z_n, \beta)$, i.e. the multinomial conditional probability of the word conditioned on the topic z_n.

*Table 10: Description of the Latent Dirichlet Allocation. Topic-Mining Algorithm following (Blei et al., 2003). **N** refers to the number of words contained in a document, **Θ** denotes the topic distribution.*

This process results in two matrices. Matrix **A** contains the word to topic probabilities, while matrix **B** contains the topic to document probabilities. Matrix **A** can be used to identify the overarching theme of a particular topic by ordering the matrix by descending word probabilities for each topic. Matrix **B** can be used to compare different documents using their (dis)similarity regarding these topics. We use the R 'topicmodels' package for our analysis (Grün and Hornik, 2011). When training the topic model, the main parameter that needs to be chosen is *k*, the number of topics included in the model. The choice of *k* is a trade-off between choosing a small value, which trains a model with very few topics that are quite distinct from one another and a large *k*, which results in a model with many topics, however, these topics may be more similar to each other. Another factor in this choice is the number of topics that are expected to be naturally included in the analyzed content. As discussed, the LDA model provides a matrix of topic to document assignment probabilities (B). This can be used to compare documents in a number of ways. A common approach is the computation of similarity scores between the respective topic probability vectors of two documents. A higher similarity implies a more similar topic structure and consequently similar documents that are more alike. Different measures for these resemblances have been proposed and such similarity scores are not inherently alike human intuition about document similarity (Lee et al., 2005). Here, we will stick to cosine similarity, while the evaluation of different measures in this context presents an opportunity for future research. Cosine-Similarity measures the difference between vectors by calculating the cosine of the angle between them (Han et al., 2011). The resulting value is bounded in [0,1] since probability vectors only contain elements between 0 and 1 (typically single topics have small probabilities). The cosine-similarity between two vectors **a** and **b** is defined as:

$$\text{Cosine-Similarity(a,b)} = \frac{crossprod(\boldsymbol{a},\boldsymbol{b})}{\sqrt{crossprod(\boldsymbol{a}) * crossprod(\boldsymbol{b})}}$$

A larger cosine-similarity is related to 'more similar' documents and consequently provides a more intuitive scale for similarity, as opposed to a larger angle between the vectors, which is related to less similar documents. If a topic has a high (or low) probability in document A and document B, there might be a relationship between these assignment probabilities. However, the inverted case is unlikely to occur. When comparing the documents in our analysis, we utilize the topic probability vector for the documents, i.e. compare $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)$ to $\boldsymbol{\beta} = (\beta_1, \dots, \beta_N)$ using the cosine measure. However, since we are not concerned with the relationship between a single conference call and a single analyst report, but rather the relation between the two content types in general, a second step of aggregation is necessary before the analysis can take place. In order to assess if there is a change in the topic structure of analyst reports following conference calls (C), we select a sample of reports prior to and following

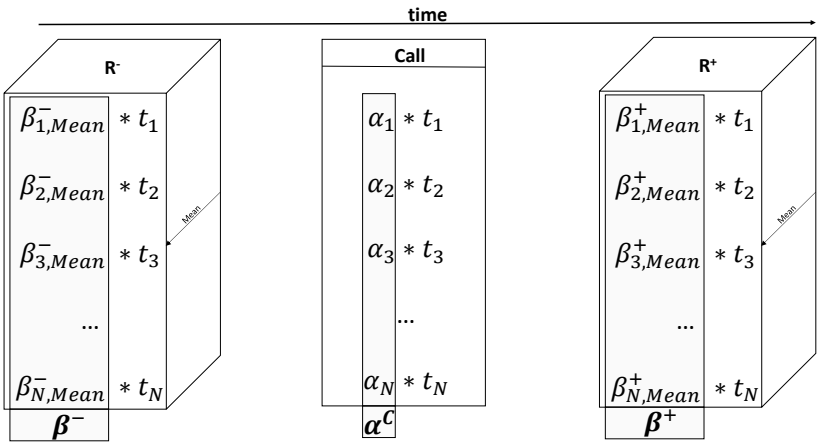


Figure 17. Illustration of sample selection surrounding a conference call. $\boldsymbol{R}^{+/-}$ refer to the pre-call and post-call report samples. The topic-to-document probabilities of the call are denoted as $\boldsymbol{\alpha}^C$, the average topic-to-document probabilities of the pre- and post-call samples are denoted as $\boldsymbol{\beta}^-$ and $\boldsymbol{\beta}^+$ respectively. Each t_n refers to a topic.

each conference call. For both periods we compute the average of the samples' respective topic to document probabilities and denote the average topic probabilities prior to the call as $\boldsymbol{\beta}^-$, while referring to the post-call average as $\boldsymbol{\beta}^+$. If the call has an impact on the topic structure of the reports written after it has taken place, the similarity between the calls topic structure and the reports should increase. If there is no relation between the two content types at all, no stable pattern should emerge. This leaves us with the task of determining the appropriate sizes of the pre- and post-call report samples. There are, roughly, one hundred reports written between the occurrence of two calls (the number varies) and consequently a pre- and post-call sample of 50 reports each would capture the mid-point between two calls. However, there is the concern that if too many reports are selected in the sample, a present relation between the content types might be missed because, for example, only the reports written in the week prior to and after the call may relate to the call. However, if the sample is chosen too small, one cannot be certain that this doesn't miss reports of especially well-informed analysts (pre-call similarity, had a topic early on), or reports written very thoroughly after the call (simply take longer to create). Therefore, we conduct the analysis for a large amount of sample sizes and compare the results regarding the mean of pre- and post-call topic similarities to all calls on average. The results presented in the next section exhibit interesting patterns based on this parameter.

5 Results

To give a meaningful impression of the results for the averages described at the end of the last section, we begin the presentation of our results by reporting one example of the data that constitute the average results. Table 11 provides the pre-call ($Cos(\alpha^c, \beta^-)$) and post-call ($Cos(\alpha^c, \beta^+)$) averages for a report sample size of 10 reports in each

direction from the call. As Table 11 shows the pre-call similarity tends to be lower than the post-call similarity for this sample size. The mean pre-call similarity across calls is 7.395% and the mean post-call similarity is 16.737%. The mean-difference is statistically significant on a 99% confidence level. This creates the question how this difference depends on the size of the pre- and post-call report samples. To answer this question, we calculate the mean pre- and post-call similarity for sample sizes between 2 and 300 reports before and after the call.

As shown in Figure 18, the post-call similarity peaks immediately after the call and continues to be larger than the pre-call similarity throughout the chosen report sample sizes. The fact that the difference is significant on a 90% confidence level up until over 100 reports after the call is interesting in itself, as a report sample of this size may very well include the next call. Still, there is no notable peak in similarity for either pre- or post-call similarity, which indicates that topics from one call will typically not be taken up in the next one.

Call ID	$Cos(\alpha^c, \beta^-)$	$Cos(\alpha^c, \beta^+)$	Change	Call ID	$Cos(\alpha^c, \beta^-)$	$Cos(\alpha^c, \beta^+)$	Change
1	0.079	0.326	+	25	0.003	0.322	+
2	0.074	0.040	-	26	0.045	0.592	+
3	0.040	0.064	+	27	0.022	0.302	+
4	0.047	0.023	-	28	0.020	0.015	-
5	0.037	0.695	+	29	0.020	0.271	+
6	0.026	0.089	+	30	0.338	0.023	-
7	0.246	0.041	-	31	0.016	0.623	+
8	0.045	0.023	-	32	0.100	0.235	+
9	0.332	0.025	-	33	0.027	0.285	+
10	0.042	0.262	+	34	0.033	0.011	-
11	0.115	0.078	-	35	0.294	0.007	-
12	0.043	0.071	+	36	0.054	0.238	+
13	0.025	0.018	-	37	0.057	0.344	+
14	0.138	0.106	-	38	0.044	0.472	+
15	0.046	0.034	-	39	0.025	0.444	+
16	0.045	0.188	+	40	0.075	0.209	+
17	0.017	0.047	+	41	0.125	0.014	-
18	0.033	0.049	+	42	0.017	0.057	+
19	0.033	0.298	+	43	0.073	0.021	-
20	0.007	0.016	+	44	0.020	0.019	-
21	0.022	0.293	+	45	0.507	0.095	-
22	0.008	0.030	+	46	0.025	0.211	+
23	0.004	0.018	+	47	0.031	0.028	-
24	0.005	0.007	+	48	0.070	0.353	+

Table 11. Mean Cosine-Similarities. Between each conference call and the mean topic to document probabilities ($\boldsymbol{\beta}^-$ and $\boldsymbol{\beta}^+$) for 10 pre- and post-call reports. Positive difference in grey, negative in white. The call IDs are ordered in time, i.e. call 1 is the first call in the sample and call 48 the last.

More importantly, the peak in post-call similarity is in line with the assumption that analysts are provided with valuable new information during conference calls, which leads to a topic change in post-call reports.

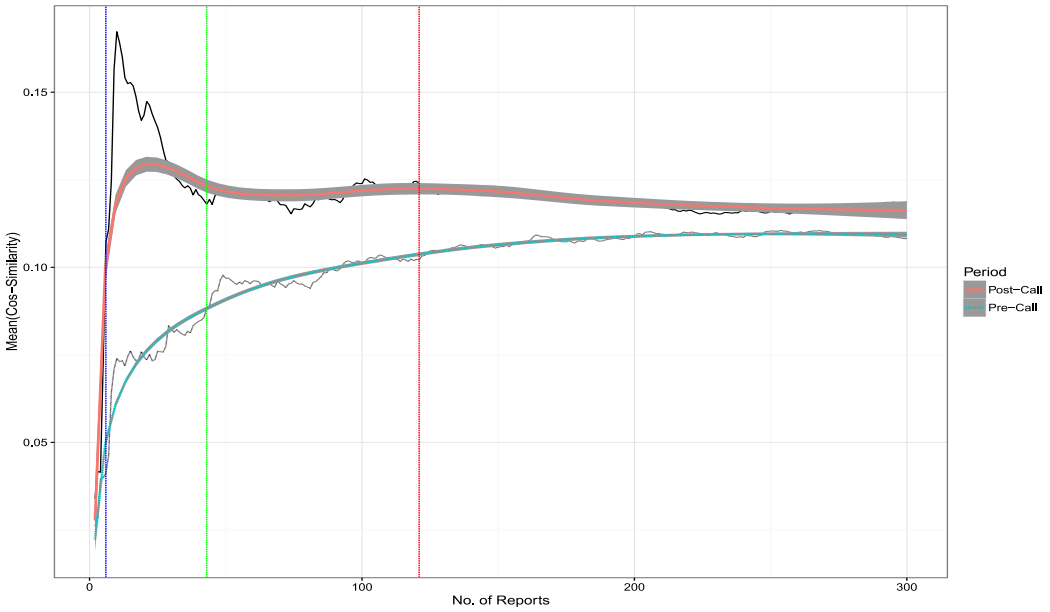


Figure 18. *Topic transfer between media. Average (across all calls) of means (across $\boldsymbol{\beta}^{-}$ and $\boldsymbol{\beta}^{+}$) cosine similarities for report samples of varying size n= [2,…,300]. The peak of Pre-Call is located around n=[2,…,20]. The blue vertical line indicates the beginning of statistically significant differences. The green line indicates the end of 95% confidence. The red line the end of 90% confidence.*

5.1 Limitations

There are a number of limitations, which should be kept in mind when considering the results of this research. First, while the sample analyzed is quite large regarding the number of analyst reports and of reasonable size (n=48) regarding the calls included in the analysis, we examine the case of a single firm in this study. While there is a methodological reason for this (a cross-company trained topic model would be unlikely to work well for our analysis), validation of these results in future research using different samples is desirable.

5.2 Future Research

As topic mining is relatively unexplored in IS research compared to other text mining approaches, it does not surprise that more questions remain unanswered regarding the usability of the topics for data mining in general and the implications of topic mining in the domain studied in this research in particular. We identify three possible types of extension of the presented research. First, a similar analysis could be applied to other non-financial content domains. Second, single topics can be explored regarding their

individual domain-transfer behaviors. Third, more than two content types could be analyzed simultaneously:

Transfer to other content domains: The approach discussed in this research may be useful in exploring information processing in other application domains. For example, marketing research regarding the adoption of topics from campaigns in social media posts or the analysis of political debates by the examination of topic proliferation in different content types might benefit from a similar analysis. The main requirement for this type of analysis is the existence of historical data that may be used for training the LDA model and that is suitable to be categorized into separate categories.

Introspection of individual topics: Of course, the presented results pose the question which kinds of topics are most likely to be transferred from one content type to the other. LDA does provide the necessary output for this kind of analysis. For example, an inter-author coding of the top words (most probable) in the topics with the highest pre- to post-call volatility could reveal which types of topics are responsible for the observed effects. Both directions of possible topic-transfer are potentially interesting for this kind of analysis. The question which kinds of topics are more likely to be picked up in a call when they are contained in a pre-call report can contribute to further understanding of analyst herding behavior, while the question which topics are most likely to transfer from calls to future reports could inform analysts which call topics are especially valuable information. This may also help to understand in which situations analysts do not herd, i.e. are especially 'daring'. Finally, for both types of topics more likely to transfer from one domain to the other, the question of 'topic value' may be explored by relating these topics to abnormal stock returns during or after the call events. Another interesting question regarding the behavior of individual topics is given by the question if topics introduced by the analysts in the Q&A part of the call, which were not present in the presentation part of call, are more likely to have a long lasting impact on the topic structure of reports or an immediate market reaction. **Expansion of current approach to more content types:** While the presented analysis reveals interesting patterns of topic-transfer between the content types, it would be daring to assume a causality between the two types of content in a general sense. Financial analysts are, as their name suggests, interpreters of information and not originators of events. Thus, it would be interesting to introduce more content types to this analysis and investigate in which other content types new topics arise before they are picked up by financial analysts or the corporate representatives present during conference calls. There are a number of promising candidates for this expansion of the current analysis, such as news media, social media, corporate filings or regulatory announcements.

6 Conclusion

In this research we apply LDA, a topic mining algorithm, to analyst reports and conference calls, in order to investigate financial analysts' information processing behavior. To this end, we collect a consecutive sample of reports and calls about IBM covering the period from 2000 to 2015. Keeping in mind the limitations discussed in the previous section, we examine the topic to document assignment probabilities resulting from this model and determine average topic similarities between report samples prior to and after each conference call in the sample, while varying the scope of the pre- and post-call samples. Results indicate that, although individual topics exhibit different behavior, on average, analyst reports written in a short period after conference call events show a significant topic-uptake from conference call events. This finding is in line with the consideration that analyst conference calls are a valuable source of new information for stock analysts. On the other hand, there is no similar spike in to-call similarity regarding analyst reports released prior to call events, which may be seen as support for the "herding" tendency of analyst opinion, i.e. the tendency to stick to the consensus estimate until new information has reduced the risk of changing ones' opinion. Future extensions of this study may include the extension of the approach to more companies, which may allow to examine industry differences in topic-transfer between analyst reports and conference calls, content domains, the introspection of the effects of individual call topics, as well as the integration of other measures, such as conference call sentiment and its relation to call topics.

III.4. Media Richness and the Information Value of Analyst Opinion

What do They mean? Using Media Richness as an Indicator for the Information Value of Stock Analyst Opinion regarding post-earnings Firm Performance

What do they mean?
Using Media Richness as an Indicator for the Information Value of Stock Analyst Opinion regarding post-earnings Firm Performance

Matthias Eickhoff
University of Göttingen
Matthias.Eickhoff@wiwi.uni-goettingen.de

Abstract

In this research the impact of media-richness on the investor reaction to earnings announcements is investigated. To this end, unstructured (high-richness) sources of analyst opinion are subjected to text-mining and combined with structured (low-richness) sources of analyst opinion, as well as other commonly used structured data relevant to company performance. Results indicate that equivocality is a major problem faced by investors, while uncertainty as understood by media-richness theory appears to be less dominant.

1. Introduction

The quarterly earnings announcements of publically traded companies pose significant challenges for both academia and investors alike. Research regarding this quarterly ritual is extensive and often concerns the impact of the newly revealed information on the expectations of investors. Meanwhile, investors face a more ad-hoc decision problem. They need to decide whether to buy, to hold, or to sell a share on the ground of so far announced information. In todays networked society, many information sources are available to investors. Among those, the variant opinions of stock analysts are a key source of analysis regarding past and future company performance. However, these opinions are available over a multitude of channels including both financial information systems, which provide structured recommendations, telephone earnings calls, which are publically available in most jurisdictions, and analyst reports, which interested investors can purchase. This poses a selection problem for investors: Which of these sources are (is) most important regarding the investment decision? This research addresses that decision problem using media richness theory as a background on which different media types can be evaluated regarding a number of criteria. Furthermore, media richness theory provides pointers on what kind of media type (rich or less rich) is more efficient, given the presence of uncertainty and equivocality in any given decision problem. On this basis, the available media types are evaluated and tested regarding their explanatory power as to historical investment decisions, which are measured as the abnormal stock return following earnings announcements. Section 2 provides a brief introduction to recent finance and accounting research concerning earnings releases and the related sources of analyst opinion. Following, it discusses the relation of media richness theory to the investment decision problem. Section 3 introduces the data used in the following analysis. Section 4 gives a methodological overview and shows how sentiment analysis, topic-mining and the estimation of abnormal returns are performed as a basis for this research. Consequently, section 5 shows how the analysis is performed and presents its results. Section 6 discusses the implication of the presented results before this research is concluded in section 7.

2. Theory

2.1. Analyst opinion

Earnings releases in general and earnings call transcripts in particular have been studied regarding their information content by continuous streams of accounting and finance research. Previous work focusses on structured information regarding the earnings release itself and non-textual measures regarding the earnings call. The information content of earnings announcements have been found to be stable over several decades regarding both abnormal trading volume and return volatility [1-3]. More recently, research focusses on textual measures derived from the transcripts of earnings' calls. Commonly, studies use wordlists based measures to capture call tone [4]. Also, a lack of spontaneity in management responses to analyst questions in earnings calls has been investigated [4]. Furthermore, investment reaction to analyst tone is usually stronger than to management tone [1]. In line with these findings, positive tone increases the effect of positive earnings surprise [5]. Moreover earnings forecasts published by analysts who participated in earnings' calls immediately after

Abstract: In this research the impact of media-richness on the investor reaction to earnings announcements is investigated. To this end, unstructured (high-richness) sources of analyst opinion are subjected to text-mining and combined with structured (low-richness) sources of analyst opinion, as well as other commonly used structured data relevant to company performance. Results indicate that equivocality is a major problem faced by investors, while uncertainty as understood by media-richness theory appears to be less dominant.

Citation: Eickhoff, M. (2017). "What Do They Mean? Using Media Richness as an Indicator for the Information Value of Stock Analyst Opinion Regarding Post-Earnings Firm Performance," In: *Proceedings of the 50th Hawaii International Conference on System Sciences*, Hawaii, USA: AISeL.

Keywords: Earnings Release; Firm Performance; Media Richness Theory; Stock Analysts; Topic Mining

1 Introduction

The quarterly earnings announcements of publically traded companies pose significant challenges for both academia and investors alike. Research regarding this quarterly ritual is extensive and often concerns the impact of the newly revealed information on the expectations of investors. Meanwhile, investors face a more ad-hoc decision problem. They need to decide whether to buy, to hold, or to sell a share on the ground of so far announced information. In todays networked society, many information sources are available to investors. Among those, the variant opinions of stock analysts are a key source of analysis regarding past and future company performance. However, these opinions are available over a multitude of channels including both financial information systems, which provide structured recommendations, telephone earnings calls, which are publicly available in most jurisdictions, and analyst reports, which interested investors can purchase. This poses a selection problem for investors: Which of these sources are (is) most important regarding the investment decision? This research addresses that decision problem using media richness theory as a background on which different media types can be evaluated regarding a number of criteria. Furthermore, media richness theory provides pointers on what kind of media type (rich or less rich) is more efficient, given the presence of uncertainty and equivocality in any given decision problem. On this basis, the available media types are evaluated and tested regarding their explanatory power as to historical investment decisions, which are measured as the abnormal stock return following earnings announcements. Section 2 provides a brief introduction to recent finance and accounting research concerning earnings releases and the related sources of analyst opinion Following, it discusses the relation of media richness theory to the investment decision problem. Section 3 introduces the data used in the following analysis. Section 4 gives a methodological overview and shows how sentiment analysis, topic-mining and the estimation of abnormal returns are performed as a basis for this research. Consequently, section 5 shows how the analysis is performed and presents its results. Section 6 discusses the implication of the presented results before this research is concluded in section 7.

2 Theory

2.1 Analyst opinion

Earnings releases in general and earnings call transcripts in particular have been studied regarding their information content by continuous streams of accounting and finance research. Previous work focusses on structured information regarding the earnings release itself and non-textual measures regarding the earnings call.

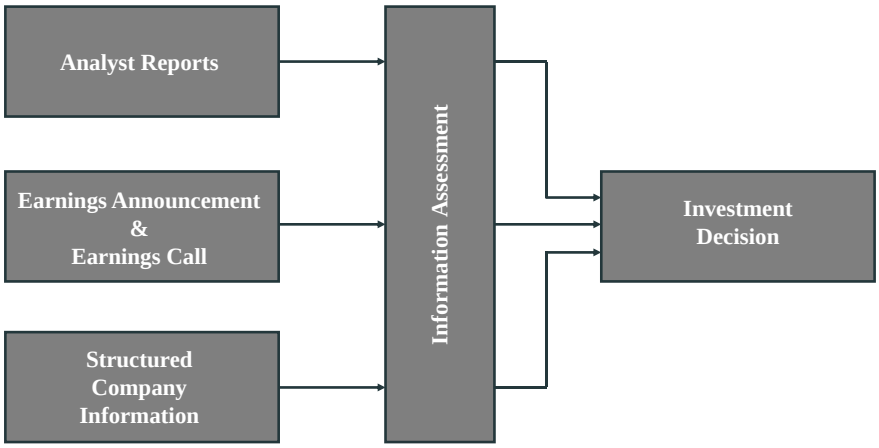


Figure 19: Investment Decision.

The information content of earnings announcements have been found to be stable over several decades regarding both abnormal trading volume and return volatility (Brockman et al., 2015; Davis et al., 2015; Landsman and Maydew, 2002). More recently, research focusses on textual measures derived from the transcripts of earnings' calls. Commonly, studies use wordlists based measures to capture call tone (Lee, 2015). Also, a lack of spontaneity in management responses to analyst questions in earnings calls has been investigated (Lee, 2015). Furthermore, investment reaction to analyst tone is usually stronger than to management tone (Brockman et al., 2015). In line with these findings, positive tone increases the effect of positive earnings surprise (Blau et al., 2015). Moreover, earnings forecasts published by analysts who participated in earnings' calls immediately after them are more accurate than those published by analysts who did not participate (Mayew et al., 2013). It has been shown that allowing analysts to participate in conference calls gives them access to private information, even though the calls are public, and management can use such measures to discriminate against unfavorable analysts in earnings calls (Mayew, 2008). Still, earnings calls have been found to be a valuable source of information for investors. Likewise, the reports and non-earnings' calls related forecasts released by stock analysts have been subject to continuous interest by accounting and finance researchers. As is the case in the area of earnings' announcements, earlier research focused on structured recommendations released by analysts. One source for these structured analyst opinions is given by the institutional broker estimate system (*I/B/E/S*), in which analysts publish and continuously update estimates on numerous financial indicators related to a company's performance. On the basis of such data, research has shown that mean forecasts (the average opinion of all analysts submitting an estimate) overemphasizes the common information all analysts share over the private information that makes the estimates interesting in the first place (Kim et al., 2001). In addition analysts with historically more accurate forecasts are more likely to make bold predictions (as opposed

to sticking to the consensus estimate) than those with poorer forecast accuracy in their past (Hope, 2003). Like the purely earnings related analyst opinion research, this research stream has also begun to analyze textual analyst opinions, which are presented in analyst reports and examined the choice of peer companies used by sell-side equity analysts (De Franco et al., 2015). On the same note it's showing that report readability correlates with analyst capability (Franco et al., 2015). It has also been shown that report tone can provide excess information beyond structured forecasts (Huang et al., 2014).

As this overview of research on analyst opinion shows, research regarding the textual sources of analyst opinion has begun, but so far has mainly focused on the augmentation of traditional models regarding the accuracy and impact of analyst opinion. As shown by this prior research, this textual content can improve upon the structured recommendations given by analysts through *I/B/E/S* and similar systems. This research contributes towards this growing corpus of knowledge by investigating two of the reasons why this is so using established IS theory. Furthermore, topic-mining is used to extract information about the impact of specific topics discussed in both analyst reports and earnings calls. In this study, the investor reaction to the release of quarterly earnings will be considered in conjunction with the release of analyst reports and estimates for earnings surprise. Figure 19 shows this basis for investment decisions following earnings announcements.

2.2 Media Richness Theory

Media Richness Theory as proposed by Daft and Lengel (1983) analyzes the effectiveness of different media types regarding the transportation of information between different individuals or organizations. It argues that in order to convey information effectively the transport medium needs to match the complexity of the transmitted information regarding four core criteria: (1) Language variety, (2) multiplicity of cues (channel variety), (3) personalization (source), and (4) feedback immediacy. **Language variety** does not necessarily refer to the use of different natural languages but the mediums' capability to transmit a wide spectrum of concepts and ideas. For example, Daft and Lengel list music and art as media with a high language variety in their seminal work on the subject, as opposed to mathematics as an example of a low variety language. **Cue multiplicity** alludes to the variety in channels through which a medium transmits information. For example, face-to-face communication offers more channels (facial expressions, audio, visual) than a phone call (audio). **Personalization** or the source of communication refers to the soft factor of being able to interact with another person instead of a machine or written communication. Finally, **feedback immediacy** extends this notion by allowing to correct faulty perceptions by the recipient of transmitted information. If a medium ranks high across these categories, it is considered

rich. Based upon this richness categorization the theory argues that rich media types perform superior to less rich media types in equivocal tasks, while less rich media types can support information transmission in the presence of uncertainty (Daft and Lengel, 1986). Uncertainty categorizes situations in which a decision maker has not been supplied with enough information to reach a well based decision. Equivocality describes a situation in which the decision maker is faced with numerous and possibly conflicting sources of information, making it difficult to reach a firm decision (Daft and Macintosh, 1981; Dennis and Kinney, 1998). Earnings announcements present both uncertain and equivocal problems to decision makers (investors). On the one hand, the investor wants to assess the future performance of a company based on information, which in the best case presents the current state of the company. As all predictions are, this assessment is highly uncertain. On the other hand, the investor is faced with numerous opinions pertaining to the subject at once. The management of a company will often interpret a given situation differently than stock analysts, media, or the investor. Consequently, the information presented to the investor is highly equivocal. Assuming both of these assumptions hold true, media richness theory suggests that there is good reason to listen to both low and high richness media types regarding earnings announcements. Low richness media types may help to reduce uncertainty, while high richness media types may mitigate equivocality. This leads to the question if the hypothesized effects are measurable in media types related to earnings announcements. To address this question, the following research Questions are proposed:

RQ1: Do low richness media types transmit investment-relevant information regarding earnings announcements, i.e. does low media complexity help to investors to reduce uncertainty after earnings announcements?

RQ2: Do high richness media types transmit investment-relevant information regarding earnings announcements, i.e. do increased language variety, cue multiplicity, personalization and feedback immediacy in their union increase the information content of materials related to earnings announcement by reducing equivocality? These isolated considerations naturally lead to a third question considering the combination of both high and low richness media types:

RQ3: What is the incremental value of high and low richness media types when their antipode is already being considered?

3 Structured, unstructured Data and Media Richness Theory

In this section the data types used in the following analysis are presented. In the context of media richness theory, structured and unstructured data can be considered as extreme representations of low and high richness media types. Typically, structured data is highly formalized and consequently scores low in the discussed categories, assessing media richness. In contrast, unstructured data typically consists of higher richness media types, such as earnings call transcripts and analyst reports in our case. Thus, sources of structured and unstructured data can serve as proxies for low and high richness media types. All data used in this study is obtained from Thompson Reuters' Datastream and Advanced Analytics (TRAA) platforms.

3.1 Low Richness (Structured Data)

Three kinds of low richness data sources of interest to investors are investigated in this study. First, the stock price of companies in the sample is collected. Secondly, several balance sheet related variables, commonly used in relation to earnings announcements are added to augment the stock price. Finally, several analyst consensus estimates are collected from *I/B/E/S*, which reflect the mean estimates of all analysts who had submitted their opinion about each variable on the call date. Table 12 provides a detailed description of each collected variable.

Variable	Description
Total Assets	Total Assets as reported on call date
Pretax ROA [%]	Pretax return on assets in percent
BV / Outstanding Share	Book value per outstanding share
Price to Book	$\frac{\text{Stock Price}}{\text{Total Assets} - \text{Intangibles}}$
Insider Ownership [%]	Percent of shares owned by shareholders >10% ownership or officers
ROE Surprise Mean	Return on equity surprise (*I/B/E/S*) mean
ROE # Estimates	Number of Estimates for ROE
EPS Surprise Mean	Earnings per share surprise (*I/B/E/S*) mean
EPS # Estimates	Number of Estimates for EPS
Market Cap	No. Outstanding shares x price
Consolidated Market Cap	No. Outstanding shares (all issues) times price
Reports LMD Uncertainty	Reports % of text match with LMD Uncertainty
Call QA AN LMD ModalStrong	Call Q&A analyst questions % of text match with LMD ModalStrong
Call QA AN LMD Negative	Call Q&A analyst questions % of text match with LMD Negative

Table 12: Variable descriptions.

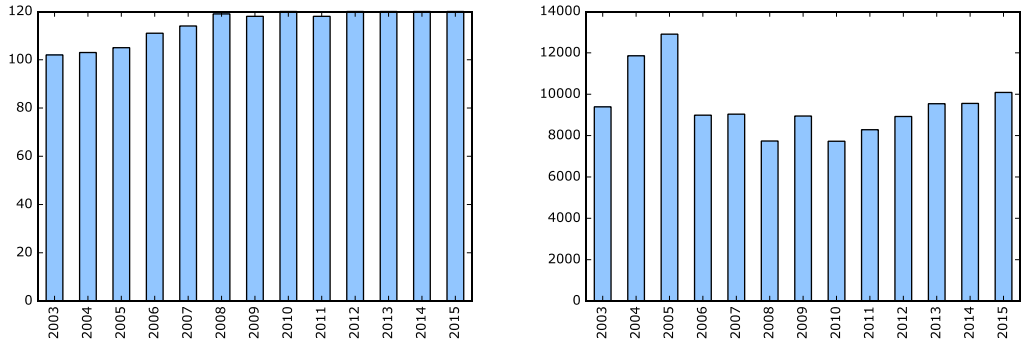


Figure 20: Histograms of call (left) and report (right) counts.

Prior research regarding the effects of analyst opinion on investor behavior shows that investors do listen to this low-richness information source. For example, trading strategies based on the consensus estimate have been analyzed (Barber et al., 2001), as have the effects of boldness on forecast accuracy (Clement and Tse, 2005).

3.2 High Richness (Unstructured Data)

Two sources of unstructured analyst opinion data are used in the following analysis. Both were collected from Thompson Reuters Advanced Analytics (TRAA). Both analyst reports and earnings conference calls have been studied regarding to their effect on investor behavior. Analyst reports have been studied regarding the market reaction to their release (Asquith et al., 2005), the effect of their readability on abnormal trading volumes (Franco et al., 2015), and effect of report ambiguity on investor reaction (Winchel, 2015). Likewise, investors' reactions to earnings conference calls have been studied in regard to the link between effects of call tone and investor sophistication (Blau et al., 2015), as well as the effects of call tone on abnormal returns (Price et al., 2012). As these prior studies have repeatedly shown, these sources of unstructured but high-richness media often lead to significant investor reactions. Thus, both low- and high-richness media sources have been shown to be feasible predictors for investor behavior. The following data are used in this analysis. At first, analyst reports about the companies included in the Dow Jones Industrial Average ($DJIA_{30}$) between 2003 and 2016 were collected. These analyst reports typically contain a review of the current financial situation of the company and an estimate of its future development, in a mixture between freely written text as well as tables and figures.

This paper focuses on analyzing the freely written portion of the reports and consequently extracting the textual content of each document for further analysis. Second, the transcripts of all earnings calls for the same period were collected. Earnings calls typically consist of two separate segments.

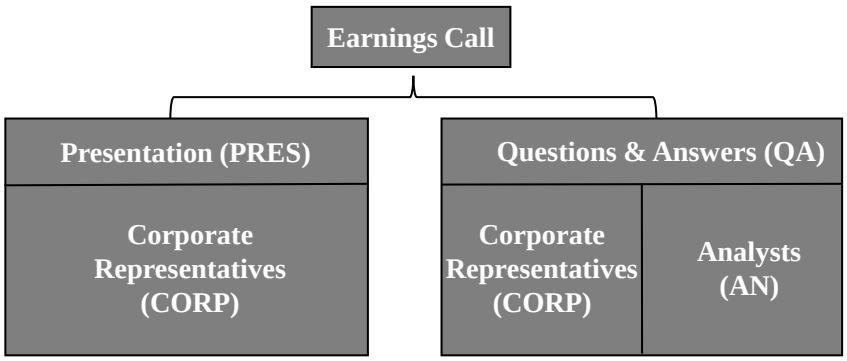


Figure 21: Earnings Call Structure.

First, the management of the company holding the call presents the earnings announcement in form of a monologue of the CEO or CFO. Second, analysts ask questions about the announcement or other topics of interest for possible future company performance. Figure 21 gives an overview of this structure. Three types of contribution to the call, i.e. presentation by the corporation, questions by analysts, and answers by the corporation, are extracted to calculate separate measures for each. If a call does not exhibit all three segments it was dropped from the sample. Figure 20 shows the annual counts for each media type. In the case of the earnings calls 120 per year is the natural limit for a 30 company ($DJIA_{30}$) sample (one call per quarterly earnings per year). As shown, data are available continuously for the analyzed period with a notable reduction in the report count during the global financial crisis (~2008).

4 Method

In this section, an overview of the methodologies and tools utilized in this paper will be provided. First, the text mining methods used to extract variables from textual content are elaborated. Second, the abnormal return model is developed. Finally, the commonly used methods to analyze the resulting data are presented.

4.1 Sentiment Analysis

Sentiment Analysis involves extracting the emotional contents of documents or document collections with the intension of providing an overview on the opinions and feelings of their authors. Both supervised and unsupervised learning has been applied to this field of content analysis (Liu, 2012). While both approaches have been used with success, clearly certain trade-offs exist using either. On the one hand, supervised learning can offer excellent accuracy but domain-portability poses a challenge (Aue and Gamon, 2005). On the other hand, dictionary based (unsupervised) sentiment scoring is inherently limited by the dictionary of sentiment laden words texts are compared to, resulting in the need for domain-appropriate sentiment dictionaries (Muhammad et al.,

2013). As this analysis intends to asses both analyst reports and earnings call transcripts and finance specific sentiment dictionaries are available, the dictionary based approach is chosen. Two dictionaries to score both types of texts were chosen, a general purpose dictionary developed by Hu and Liu with Positive and Negative categories (Hu and Liu, 2004) and a finance specific dictionary developed by Loughran and McDonald (LMD) with Positive, Negative, Litigious, Modal-Weak, Modal-Strong and Uncertainty categories (Loughran and McDonald, 2011). In the case of analyst reports, a 30-day report sample prior to each earnings call event is averaged regarding these categories. For the calls themselves, the three discussed segments are scored individually for each call.

4.2 Topic Mining

In contrast to sentiment analysis, topic mining aims to extract what is being said in a document and not how authors feel about a specific topic. Latent Dirichlet Allocation (Blei et al., 2003) is a topic mining algorithm and generates a pre-determined number of topics in a document collection **D** consisting of a vector of documents **w**, each of which consists of **N** individual words w_n. A topic z_n is represented as a mixture of the words contained in the document collection. In turn, each topic is assigned to each document with a certain probability.

For the purpose of our analysis the topic-assignments to each earnings call in the sample and the average of the topic assignments to all analyst reports 30 days prior to the call date (equivalent to the sentiment measures) are computed using the MALLET topic mining package (McCallum, 2002). For the calls themselves, topics are computed for the entire call and not for the three individual segments. The main reason behind this is, that the Q&A portion of the call often contains very short text parts, such as short questions which are asked by analysts often receive short answers and consequently are not as suited for being analyzed by using LDA on each individual text. The topic model is trained on all textual data at once, i.e. both analyst reports and the conference call transcripts are used as training data.

1.	Choose **N**~ Poisson(ξ)
2.	Choose $\Theta \sim Dir\text{ichlet}(\alpha)$
3.	For each of the **N** words w_n:
	I: Choose topic $z_n \sim$ Multinomial(Θ)
	II: Choose a word w_n from $p(w_n \mid z_n, \beta)$, i.e. the multinomial conditional probability of the word conditioned on the topic z_n.

Table 13: Latent Dirichlet Allocation (Blei et al., 2003).

The model is set to identify 100 topics. The number of topics is the crucial model parameter in our case. If the parameter is chosen too low, the resulting topics lack granularity if it is chosen too high the resulting topics are increasingly indistinguishable for humans. As noted, a topic consists of the likelihood that each word in the corpus is part of the topic. Consequently, topics can be very similar in regard to their most likely words but very different in their overall composition, consequently making them hard to distinguish. The number of 100 topics is determined by experiments keeping this trade-off in mind and seems reasonable as this relatively high number of topics helps to avoid company specific topics, i.e. topics that simply classify a text regarding the company it belongs to. To validate the number of chosen topics, a HDP (hierarchical dirichlet process) model (Teh et al., 2006) was trained using Gensim (Řehůřek and Sojka, 2010). This model, depending on the cutoff-likelihood (relative importance of one topic compared to others), points to 80-95 topics as the 'optimal' amount. As the following analysis uses a variable selection approach, choosing a slightly higher number seems reasonable as superfluous topics will not be included in the resulting models. Another option is to train the topic model dependent on companies or industries with lower individual topic numbers. The main reason this was not done here is that it would result in topics that do not allow for inter-company or inter-industry comparisons of topic impacts.

4.3 Abnormal Returns

The analysis of the investment decisions following an earnings announcement aims to capture the reaction of the average investor to said announcement. It is important to keep in mind that only changes in opinion can be measured on the stock market. In general, three possible reactions to an earnings announcement are conceivable. First, an investor can be positively surprised by the announcement and consequently buy the share. Second, the opposite reaction follows a negative surprise. However, if an investor does not change her opinion and consequently does not alter her position regarding a stock, no trade occurs and consequently no change in price can be observed on the market. Nonetheless, both negative and positive changes in opinion should lead to a change in the investors' position and consequently would be observable on the market. Thus, the measurement is strictly limited to the unexpected portion of the information contained in an earnings announcement. In order to monitor the aggregate change in investor opinions, after an earnings announcement, the abnormal return of a company's share after the event can serve as a proxy. This approach to measuring investor opinion relies on the efficient market hypothesis, which stipulates that any new information should be represented in the stock price immediately (Fama et al., 1969; Malkiel and Fama, 1970).

Type	ID	Label (Coding)	Top Words
R	8	Earnings	report securities andor affiliates financial eps companies subject
R	43	Credit Cards	volume growth debit payments payment credit card revenue
R	51	Agreement	agreement announced technology company systems development health
R	63	Chem. Products	sales materials company chemicals performance protection products segment
R	69	Investment	report information investment price research securities limited financial
R	72	Pharma Products	cancer disease phase products infections trial life science
R/C	75	Pharma Research	sales patients data phase product products drug study
R/C	86	Rating	research report securities investment companies stock industry months
R	94	Sports	footwear apparel brand growth china futures product athletic
C	35	Listing	listed listing sales investext data deleted services corporate
C	37	Risk	rating report markets firm risk global securities investment
C	66	Prudence	prudential group equity llc rating york report analyst
C	88	Financial Data	source exhibit data research yoy survey figure index
C	99	Markets	world markets report company sector securities investment research

Table 14: Topics relevant in regression models.Type = R denotes report topics, C call topics.

In order to monitor the abnormal component of returns following an event, an expectation of the normal return for the same period in the absence of the event needs to be formulated and the abnormal return is defined as the difference between observed and expected returns following the event: $AR_{i,t} = R_{i,t} - E(R_{i,t}|X_t)$. $E(R_{i,t}|X_t)$ refers to the expected return given X_t, the development of a reference group (S&P500) of shares during the period. The estimation of these normal returns is performed by using the market model approach (MacKinlay, 1997), assuming a time-constant relation between the reference group and the stock in question, using the following OLS model: $R_{i,t} = \alpha_i + \beta_i R_M + \epsilon_{i,t}$ with $E(\epsilon_{i,t}) = 0$ and $Var(\epsilon_{i,t}) = \sigma^2_{\epsilon,i}$ using -200 days up to -1 day prior to the earnings announcement as training data for the return model. This yields the abnormal return for a given day, starting with the day of the earnings announcement itself. Indeed, both AR_{t0} to AR_{t10} and cumulative return measures (CARs) were calculated for the sample. AR_0, the abnormal returns on the day of the earnings call, was most suitable for the analysis and consequently is used as the independent variable for the resulting models.

4.4 Topic Selection

As noted, the topic model trained on earnings call transcripts and analyst reports is used to compute the topic composition for each call transcript in the sample and the average topic composition of all analyst reports released 30 days prior to the call. As each of these two topic groups consists of 100 topics, this alone results in numerous topic-variables per call event. As it is infeasible to use all 200 topic-variables in the following regression models and more importantly it is unknown which topics computed by the model, are of interest to investors, a selection needs to be performed. A two-step approach to this selection problem is chosen. First, a topic set suitable for regression is identified using the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), which selects a subset of variables out of the pool of topics.

The sentiment based variables are also supplied to the Lasso. The selected sentiment variables are those included in the regression models in Table 15. Second, because automated variable selection by itself is problematic, the resulting topic set is manually inspected as a sanity check. The coding is performed by looking at the top 20 words for each topic (10 displayed for space reasons). The resulting topic set is displayed in Table 14. As shown, a label is assigned to each topic by coding the top words of each selected topic. The type column indicates whether the topic was selected to be relevant in calls (C), analyst reports (R), or both. This has two purposes: Firstly, only topics, which are interpretable, should be used for further analysis. Secondly, it helps to make the following regression models easier to read, by replacing the topic numbers with the resulting codes. As shown, a mixture of topics indicating discussions of both financial topics and company or industry specific topics was selected. This points to the possibility of estimating the topic models on industry specific samples to generate more granularity. However, this would require a dataset covering a larger index than the DJIA because within this index only a couple of companies per industry are represented.

	M1: Structured Data		**M2: Unstructured Data**		**M3: Both (M1 + M2)**	
	Coefficient	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
Total Assets	-0.027	0.104			-0.093	0.091
Pretax ROA [%]	-0.101	0.105			0.012	0.099
BV / Outstanding Share	0.046	0.187			0.048	0.19
Price to Book	0.511***	0.152			0.461***	0.157
Insider Ownership [%]	0.065	0.074			0.022	0.065
ROE Surprise Mean	-0.430**	0.178			-0.393**	0.163
ROE # Estimates	0.009	0.117			0.072	0.109
EPS Surprise Mean	-0.106	0.168			-0.177	0.179
EPS # Estimates	0.013	0.088			-0.085	0.082
Consolidated Market Cap	-1.945*	0.994			-2.573**	0.999
Market Cap	1.898*	0.998			2.602**	1.006
Reports Topic Earnings			0.111*	0.061	0.1	0.065
Reports Topic Credit Cards			-0.131**	0.061	-0.141**	0.061
Reports Topic Agreement			-0.148**	0.061	-0.141**	0.063
Reports Topic Chem. Products			-0.133**	0.06	-0.148**	0.066
Reports Topic Investment			0.140**	0.06	0.137**	0.062
Reports Topic Pharma Products			-0.069	0.061	0.007	0.085
Reports Topic Pharma Research			-0.125**	0.061	-0.115*	0.062
Reports Topic Rating			0.082	0.193	0.221	0.197
Reports Topic Sports			0.157**	0.062	0.197***	0.065
Call Topic Listing			0.077	0.064	0.112	0.073
Call Topic Risk			0.068	0.079	0.05	0.079
Call Topic Prudence			-0.188***	0.062	-0.190***	0.063
Call Topic Pharma Research			0.123	0.08	0.136*	0.078
Call Topic Rating			0.074	0.194	-0.242	0.216
Call Topic Financial Data			0.119**	0.06	0.133**	0.062
Call Topic Markets			0.129**	0.059	0.126**	0.06
Reports LMD Uncertainty			0.102*	0.061	0.166**	0.067
Call QA AN LMD ModalStrong			0.126**	0.061	0.126**	0.061
Call QA AN LMD Negative			-0.198***	0.067	-0.188***	0.068
Constant	0	-0.068	0	0.058	0	0.056
Observations	206		206		206	
R2	0.089		0.381		0.443	
Adjusted R2	0.038		0.318		0.348	
Residual Std. Error	0.981	df = 194	0.826	df = 186	0.808	df = 175
F Statistic	1.730*	df = 11; 194	6.030***	df = 19; 186	4.642***	df = 30; 175

Table 15: Regression model ($Y=AR_0$) summaries.

5 Analysis and Results

The topics selected using the described approach are combined with the variables obtained from structured data sources and for all variables z-scores are computed ($z = (x - \mu)/\sigma$) in order to normalize the scale of the variables. It is important to keep this step in mind when interpreting the following regression models since the coefficients refer to the deviation from the mean of each variable, i.e. a negative coefficient refers to a value smaller than the average value of the variable and vice-versa. All calls on which the full set of variables could not be obtained were dropped from the sample. This mainly concerns two types of missing variables. First, calls for which no more than 5 reports were available 30 days prior to the earnings release were dropped in order to obtain better averages. Second, the *I/B/E/S* estimates were not available for earlier years. In principle, some calls could be reintegrated after variable selection is performed but the smaller sample size is kept as this seemed like the cleaner approach. In order to analyze the data with regard to the research questions three models need to be estimated. The first model (M1) only contains variables from the structured data sources and addresses RQ1 (Do low richness media types transmit investment relevant information?). The second model (M2) only contains variables estimated from unstructured data and addresses RQ2 (Do high richness media types transmit investment relevant information?). The third model (M3) contains both types of variables addresses RQ3 (What incremental value can be gained by combining both low and high richness media types?). Regarding the interpretation of these models and their connection to the research questions three aspects are of particular interest. Firstly, the coefficient sizes within the individual models and their stability (especially regarding their sign) across the models. Secondly, which coefficients are significant across the different models? And thirdly, and most interestingly, how to the models compare to one another on the model level. Table 15 shows the three resulting models.

Regarding **M1** media richness theory predicts that low-richness media types should help information transmission by mitigating issues of uncertainty (**RQ1**). As M1 contains such low-richness information, it should be able to explain the investment decisions following earnings announcements if uncertainty is a relevant problem for this decision. As shown, the adjusted R^2 of this model is comparatively small, indicating that the variables obtained from structured data are not able to explain much of the variance within the abnormal returns on the call day. Keeping this in mind, both price to book ratio and unconsolidated market capitalization show significant positive effects, while return on equity surprise mean and consolidated market capitalization exhibit negative effects. Overall, while the examined structured data offers some insight regarding the investment decisions following the earnings call, the comparatively

small adjusted R^2 (3.8%) points at a small effect of uncertainty as described by media richness theory for this decision type.

Regarding **M2** media richness theory predicts that high-richness media types should help information transmission by mitigating issues of equivocality (**RQ2**). Thus, M2 should be able to explain the investment decisions following earnings announcements if equivocality is a relevant problem for decision makers. As shown, the adjusted R^2 (31.8%) of this model is comparatively large, indicating that the variables extracted from unstructured content explain a larger portion of the variance within abnormal stock returns on the earnings call day. Thus, these unstructured data sources seem to transmit more investment relevant information when compared to the structured data sources contained in M1. Within the framework of media richness theory this points to equivocality regarding the interpretation of earnings announcements as a major problem investors need to mitigate using high-richness media sources. Within the topics of analyst reports released in a period of 30 days before the earnings calls several topics are identified that allude toward a positive effect of the earnings announcement. In particular, the discussion of earnings, investments and sports (likely and industry specific topic) show positive coefficients. Likewise, the discussion of financial data and market activity in the call itself shows positive effects. Finally, strong modal words from the LMD word list used by analysts in their Q&A questions relate to positive post-call returns.

On the other hand, reports containing credit cards, corporate agreements, chemical products, pharma research and prudence exhibit negative effects. Of these, especially the prudence topic is interesting because it isn't as industry specific as the other negative topics. Finally, negative words used by analysts in their Q&A questions show a negative relation to returns. Overall, this model explains a much larger portion of return variance than the one incorporating traditional structured data, thus indicating that equivocality seems to be a bigger problem than uncertainty regarding the investment decisions after earnings calls. Finally, **M3** combines both low- and high-richness media types in one model and investigates the complementary value of this combination beyond the value of the individual models (**RQ3**). As shown, this model slightly improves upon the adjusted R^2 from 31.8% to 34.8%, which means the improvement is slightly smaller than the adjusted R^2 of M1. This finding is in line with the intention of the measure, which penalizes models including more covariates. With regard to RQ3, this relatively small improvement over M2 may be interpreted as a small incremental value of combining low- and high-richness media types. However, keeping the limitations of the chosen text mining approaches in mind, which do not extract all information contained in the unstructured data, it is doubtful, if this incremental value would be present if an informed investor actually reads the analyst reports and listens

to the earnings call. Both of the above also may contain the relevant information contained in the structured data sources. Thus, the incremental value of the combined model can be doubted. Still, it can serve as a stability check for M2. As shown, the addition of the variables contained in M1 does not alter the sign of the significant variables of M2. The only change in covariate significance is given by the pharma research topic, which is not significant in M2 but is in M3. In summary, the models indicate a low impact of uncertainty for post-earnings call investment decisions, while providing much clearer evidence for the impact of equivocality. Finally, the incremental value of the combination of low- and high-richness media types is doubtful if the low impact of uncertainty as understood by media richness theory and the information likely contained in the unstructured content but not captured by text-mining approaches are kept in mind.

5.1 Implications and Limitations

The implications of the presented results can be spilt up into practical, theoretical and policy implications. From a **practical** perspective, investors should focus both on analyst reports and earnings calls as their primary sources of information regarding the impact of earnings announcements, if no automation is required or text mining is feasible. Still, when this is not the case the available structured data can support investment decisions. Even when these reports explained much less of the variance within abnormal returns in this study. Also, as also shown by other studies (Blau et al., 2015; Brockman et al., 2015; Davis et al., 2015; Price et al., 2012; Twedt and Rees, 2012), the presented analysis reinforces the crucial effect of call and report tone towards investor reaction. The **theoretical** contribution of this study is twofold. First, its results contribute towards the growing body of research regarding the impact of unstructured, high-richness, media types in accounting and finance research by exploring the feasibility of topic-mining within the problem domain of earnings announcements. Secondly, it contributes towards the discussion of media richness theory and its applications by exploring the effects of uncertainty and equivocality in regard to the same problem domain. On a whole results indicate that equivocality presents a major challenge to investors. Finally, results indicate that **policy** makers should keep a watchful eye on the private information gained by analysts because of their privileged access to top level corporate representatives as the high correlation to abnormal post-earnings returns may point to the presence of such private information (or analysts simply to a good job). The presented results should be interpreted only while keeping the following limitations in mind. As shown, some of the topics that are identified by the chosen approach are industry specific. Thus, an extension of the chosen approach to a larger sample with industry specific topics is desirable. Also, the identified topics are time-constant and a topic model can only identify topics, which are previously contained in

a corpus. Thus, entirely new topics will be missed and would be interesting to add in the future. Furthermore, the observation count in the final regression models is limited by data availability regarding structured analyst consensus estimates. Like the inclusion of industry specific topics, this could possibly be mitigated, by using a larger index as the basis of the analysis (data collection would pose a significant hurdle to such an extension).

5.2 Future Research

The discussed limitations and presented results outline several avenues for future research. First, future research can examine the industry specific effects of analyst opinion on investor reaction to earnings announcement and their relation to low and rich media types. Secondly, many other media types are of interest to investors. For example, social media has been used to explain stock returns (Bollen et al., 2011). Also, the audio versions of earnings calls present an opportunity to work with higher-richness media than with their transcripts. Furthermore, as noted in the limitations, the presented topic model does treat all calls and report equally over time. However, of course the topics discussed in earnings calls and analyst reports change over time. For example, during the global economic crisis call topics were likely much more negative than the model based on all data can represent and vice-versa. This point of topic sentiment hints at another possible extension. The per-topic sentiment of analyst opinion represents a final possible addition to the presented approach. Finally, other topic modelling techniques, dynamic topic models (Blei and Lafferty, 2006), exist and a comparison of their suitability for post earnings call return analysis provides an opportunity for future research.

6 Conclusion

This study examined the impact of earnings-call related low- (structured) and high-richness (unstructured) content on abnormal stock returns on the earnings-call day before the background of media richness theory. Results indicate that uncertainty poses a smaller challenge for investors than equivocality, i.e. that the decision problem of investors is dominated not by a lack of clarity or information availability but a lack of consensus among the available information sources. Furthermore, the complementary value of combining low- and high-richness media types is examined but results regarding this combination are unclear. While combined models show minor improvements over one only including high-richness media, it is doubtful if this improvement would be present if investors manually examine high-richness media sources. Additionally, results reinforce the evidence for the relation between analyst tone and investor reaction to analyst opinion and explore topic-mining as another information extraction technique that can aid in understanding the content of analyst communication.

C. Contributions

The presented contributions aim at investigating three research areas to provide a multi-faceted view of the origins, means of analysis, and use cases for unstructured opinion data in the financial industry.

The first research area focuses on the entrepreneurial environment of the financial industry in the context of FinTech companies and the innovative business models that these companies have developed.

The second research area focuses on methodological questions and provides an overview of different content analysis techniques, which are needed to cope with growing amounts of unstructured data relevant to financial markets, as well as how researchers have addressed this challenge in information systems research and other managerial disciplines.

The third and final research area constitutes the focus of this thesis. Here, unstructured content produced by stock analysts and distributed over different media types, such as analyst reports or analyst conference calls after earnings announcements, are analyzed by using the methodologies discussed in the second research area.

In this chapter, the contributions of the individual research papers constituting this thesis are briefly reiterated and related to the overarching research questions of each research area. Consequently, the practical, theoretical, and policy implications of these findings are discussed. Finally, the limitations of the presented results are discussed before outlining future research avenues based on the presented thesis.

.

1 Summary of Results

This section summarizes the findings regarding the research questions outlined in section 2. This is done by restating each research question and providing a summary of the results. See the results and discussion sections of the individual contributions for more extended result discussions based on the more granular research questions of the individual papers.

1.1 Research Area I: Entrepreneurial Environment

This research area is aimed at establishing an overview of the entrepreneurial environment, in which the results of the other two research areas are to be understood. It introduced the entrepreneurial environment and the changing technological landscape, which have made it necessary to find innovative ways to analyze data and remain competitive given increasingly diverse competitors and new analytical demands. Its first research question was as follows:

> **Research Question I.1:** What are the dimensions and characteristics of typical business models of FinTech companies?

The answer to this question is given by the FinTech business model taxonomy developed in Eickhoff et al. (2017, **paper I.1**). As discussed in more detail in that contribution, a taxonomy, as defined by Nickerson et al. (2013), consists of a unique combination of dimensions, referring to the categories in which an object is considered, and characteristics, meaning the possible values within these dimensions. The dimensions identified in this iterative taxonomy development process are given by a company's dominant technology component, its value proposition, the delivery channel for its products, its main source of revenue, and its product or service offering. In turn, for each of these dimensions, a set of possible characteristics were identified. For example, automation, collaboration, and customization constitute some of the characteristics of the value proposition dimension in the developed taxonomy. See the results section of Eickhoff et al. (2017, **paper I.1**) for the full taxonomy. Definitions for each dimension and characteristic are also provided in the paper.

Regarding the focus of this thesis, the developed taxonomy shows how analytical business models making use of decision support systems or Blockchain technology gain increasing relevance in the financial domain. On this basis, the second research question of this research area was as follows:

> **Research Question I.2:** How can these business models be grouped into different FinTech niche markets?

This question is addressed in the evaluation section of Eickhoff et al. (2017, **paper I.1**), where hierarchical clustering is used to cluster FinTechs in the sample to validate the developed taxonomy by using this clustering to assess the comprehensiveness of the taxonomy within the sample. This is done to determine whether the taxonomy enables us to explain each resulting company cluster, and the taxonomy satisfies this criterion. We consequently use the developed taxonomy to classify each company from each cluster regarding its business model using the developed dimensions and characteristics. Based on this application of the developed taxonomy, we look at the dominant characteristics within each company cluster and derive FinTech business model archetypes from the resulting distinct combinations of characteristics, such as payment service providers, alternative trading venues, and lending communities.

The results show that cryptocurrency-based and machine learning- or analytics-based businesses are challenging incumbent market participants in the financial domain.

1.2 Research Area II: Methodological

This research area concerns the methodology with which the information within unstructured content related to capital markets can be transformed to be analyzed by traditional statistical methods and how these methods are currently being used in managerial research disciplines to provide meaningful answers to research questions relevant in these disciplines.

Its first research question was given by determining the current state regarding topic modeling literature in information systems and other managerial research disciplines:

> **Research Question II.1:** What is the state of the art of topic mining methodology used to process unstructured content in the methodological literature, and how are these methods being applied in the managerial sciences to provide meaningful information relevant to researchers and decision makers?

The results of Eickhoff and Neuss (2017, **paper II.1**) concern three different areas of research. First, methodological contributions regarding topic modeling are considered. Second, available implementations of these topic modeling approaches are discussed. Third, applied research papers using the methods are investigated regarding their approach towards this methodology. Finally, the theoretical foundations of research using topic modeling methodology and its use in theory testing are considered.

First, topic modeling methodology has become a vibrant research subject. Starting with LSA (Dumais et al., 1988) and LDA (Blei et al., 2003), which represent the two most common model types used in applied papers, 25 model types are identified in the review, many of which focus on extensions of the two archetypes. Notably, many methodological contributions stemming from the IS domain are identified. Thus, IS

appears to have been established as a reference discipline for researchers looking for methodological guidance in the use of topic models. As can be expected, computer science and statistics present the two other most important methodological disciplines for topic modeling.

Second, regarding implementations, while most methodological contributions explain the statistical approach to their work, publicly available implementations remain the exception. BleiLab (2016) provides an example of enabling others to benefit from methodological work and released implementations and working examples if possible. When comparing the citation counts of papers with released implementations to those without, it is obvious that this approach is useful in this metric. However, methodological accessibility remains a major problem (Ramage et al., 2009).

Third, papers applying topic models are dominated by the two "basic" model types of LDA (27%) and LSA (35%), indicating many opportunities to augment the prior work by using more case-specific model types. Fifty-three percent of all applied articles stem from the IS domain, followed by accounting research at 10%, while general management and marketing are tied at 8% each. Thirty-nine percent of all applied papers use their topic model as a tool for content analysis. Within this category, models are mostly used as a variable augmenting existing regression models or to gain a general sense of topics included in text collections. This type of article is much more common in other managerial disciplines, but IS research using the methodology to this end still exists. The second most common applied paper type is presented by review articles, which review entire research domains (Moqri et al., 2015; Sidorova and Isik, 2010) or focus on individual journals (Cohen Priva and Austerweil, 2015; Wang et al., 2015).

Finally, using topic models for the purposes of advancing theory has been one of the uses of this model type early on (Landauer and Dumais, 1997) but remains the exception when surveying the applied literature. As discussed, accessibility may present one major cause for this. Another is given by the lack of theoretical foundations of topic model usage, which makes it more challenging to establish trust in results based upon its use. While few studies explicitly state their (meta) theoretical foundations, for studies classified as *content analysis,* a positivist underpinning aiming at the empirical validation of established theory is often implicitly clear. On the other hand, constructivist foundations or mixed method approaches to the analysis of topic models remain largely unexplored. However, similarities and differences between topic modeling and human coding have been discussed (Quinn et al., 2010). Since qualitative researchers have developed rigorous coding techniques, this methodology can support quantitative topic modeling, thereby creating opportunities for collaboration (Teddlie and Tashakkori, 2009).

This research area's next research question was given by investigating how different sentiment analysis techniques can be combined and what advantages such a combination potentially has:

> **Research Question II.2:** How can dictionary-based and machine learning-based sentiment analysis be combined to mitigate some of their individual shortcomings, such as the need for labeled training data?

The answer to this question is provided by the sentiment analysis framework and software artifact presented in Eickhoff (2015, **paper II.2**). This approach uses dictionary-based sentiment analysis to bootstrap training data for supervised learning techniques. While as outlined in the contribution, this approach does work and combines some of the benefits of supervised learning techniques, such as being able to capture domain-specific word valence with the domain portability of general purpose sentiment dictionaries, it should be remembered that the accuracy of the resulting model is unlikely to exceed that of a purely dictionary-based solution by much, as is discussed in more detail in the paper.

1.3 Research Area III: Analyst Opinion

This research area was concerned with both the processing of unstructured analyst opinion and other unstructured information sources related to capital markets and the impact that these different media types have on individual companies. It also examined how information systems and business administration theory can be applied to these problems and provide explanations as to why such effects exist. This research area's first research question was given by the relationship of social media and analyst opinion to crowd wisdom theory:

> **Research Question III.1:** What structure is there to the relationship between the opinions of social media users and stock analysts, and can wisdom of crowds theory be used to identify the situations in which the crowd or stock analysts are more likely to provide timely information, reflecting changes in a firm's circumstance?

This question was investigated in Eickhoff and Muntermann (2016b, **paper III.1**). The results indicate that social media platform diversity, such as the number of different platforms used by social media users in a sample, increases the likelihood of Granger cause between the two types of content in either direction. However, the same could not be shown in this research for age diversity, and an increase in the average age of social media users decreases the likelihood of social media user sentiment being able to explain analyst opinion. While this appears counterintuitive, as one may expect older users to be more knowledgeable on average, this does not contradict the crowd

wisdom theory because an increase in average age does not in itself constitute an increase in cognitive diversity. In line with crowd wisdom theory, we find that an increase in social media user authority measures results in a decrease of explanatory power for social media users, the theoretical explanation being that a group more assertive in its tone may result in conflicting opinions not being voiced.

Interestingly, an increase in social media user certainty also decreases the likelihood of analysts predicting social media sentiment. This, however, does not relate to crowd wisdom theory because the authority measure only concerns social media users. Investigating this for security analysts as another form of a crowd would itself be interesting. Overall, our results regarding the crowd wisdom measures for the theoretical constructs described by Surowiecki (2005) are promising. The results indicate that there is indeed a measurable connection between the makeup of the crowd and its ability to explain analyst opinion. Although it is difficult to compare the results between different content domains, this result corresponds with previous crowd wisdom research, such as the efforts to explain the content quality of Wikipedia (Arazy et al., 2006) or its comparable quality to classic encyclopedias (Giles, 2005).

Interestingly, this suggests that crowd wisdom can arise without a system that is specifically designed to allow the crowd to aggregate their opinion. Prior research highlights the importance of group coordination for content quality (Kittur and Kraut, 2008). In our case, this aggregation only occurs after the fact using sentiment analysis. Examining how the support of the crowd's coordination may improve their information processing capabilities is an interesting question for future research.

The second half of this research question concerns the situational component of the relevance of crowd wisdom. This is investigated regarding industry dummies, company characteristics, and the types of news released within a given timeframe. Overall, the industry dummies confirm prior research, indicating that stock analysts' recommendations do indeed carry inherent value (Womack, 1996). At the same time, no significant support for any industry in the sample adds explanatory power to the crowds' opinion. Regarding company-specific effects, we observe a decrease in the capabilities of the crowd to predict the analyst sentiment for companies with a higher number of subsidiaries. This might be because large multinational companies are too complex to be summarized by a single measure of crowd opinion. The crowd may have a positive opinion of one division of a firm while expressing a negative opinion about another. Regarding the effects of the news, we find that the type of news released in the observation period has a significant effect on the crowds' performance. For example, business news decreases the likelihood of analysts' ability to predict social media users. This supports the known tendency that analysts have to adhere to prior assessments (Trueman, 1994).

In summary, the results of Eickhoff and Muntermann (2016b, **paper III.1**) suggest that the crowd wisdom theory indeed provides useful constructs that can be operationalized to explain when the crowd can arrive at opinions prior to the availability of expert assessment. The second research question within this research area concerned how topic modeling can be used to support decision makers in overcoming the problem of information overload in the context of earnings releases:

> **Research Question III.2:** What constitutes a decision-relevant metric in the context of business communications regarding a firm's earnings announcement, and how can metrics of analyst opinion determined by sentiment analysis and topic modeling be used to provide such decision-relevant information?

This research question was addressed in Eickhoff and Muntermann (2017, **paper III.2**). Regarding the definition of "relevance" in the context of business communications, we presented a topic-model based analysis and optimized the topic selection regarding abnormal returns as an investment-relevant criterion. Of course, stock returns are only one possible solution to this question. In other non-investment contexts, this criterion would have to be changed. For example, in a marketing context, product sales or ad impressions could serve as an appropriate alternative. The important characteristic of the criterion regarding the chosen selection procedure is "quantifiability," meaning the feasibility to use the criterion as the object of a statistical optimization procedure.

To address the question of how to train topic models with the aim of filtering non-relevant topics in business communications, our analysis has shown that there are several considerations when training topic models to describe the topics contained in quarterly earnings announcements and analyst opinion related to these announcements. On the one hand, it appears prudent to train topic models specific to a certain industry or company. On the other hand, this introduces the danger of missing topics, which have not impacted a specific company or industry in the past. Therefore, we opted to train our topic model without such specificity but rather an increased number of topics to capture a wide range of possible subjects. Concerning the question of how to identify a relevant subset of topics suitable as information regarding the decision problem, we present an approach to reduce the number of topics to be considered by using the Lasso (Tibshirani, 1996) as a variable selection and parameter shrinkage approach.

This leads to the reflection of our results in the broader context of information overload. Any model-based approach to complexity reduction sacrifices some information in the interest of simplification. Consequently, while this topic model-based approach can certainly help to reduce the information overload during the decision-making process, the corresponding loss of information must be considered carefully within the

confines of the domain-specific decision problem. Within the financial domain, the operationalization of topic relevance is comparatively easy. The goal of investments (profit) is clearly measurable and relatively well-understood. Even slight additions to this goal, such as the incorporation of moral considerations into the investment decision, prohibit the chosen approach to the problem. Still, if the goal is as clear as in the analyzed case, the presented approach can provide an important contribution given by the reduction of the decision complexity that the model achieves.

The third research question in this area concerned how topics transfer between stock analyst reports and earnings conference calls:

> **Research Question III.3:** To what extent do the topics contained in analyst reports released prior to an earnings call influence the topics contained therein, and does the call influence the content of reports released thereafter?

This research question was addressed in Eickhoff and Muntermann (2016c, **paper III.3**). This is done by comparing the topic cosine similarity of reports released prior to and after an earnings conference call with this call itself.

The results show that the mean pre-call similarity across calls in the sample is ~7.4%, and the mean post-call similarity is ~16.7%. This mean difference is statistically significant at a 99% confidence level. This leads to the question of whether this difference depends on the size of the pre- and post-call report samples. The post-call similarity peaks immediately after the call and continues to be larger than the pre-call similarity throughout the chosen report sample sizes.

The fact that the difference is significant at a 90% confidence level up until over 100 reports after the call is interesting because a report sample of this size may well include the next call. Still, there is no notable peak in similarity for either the pre- or post-call similarity, which indicates that topics from one call will typically not be raised in the next call. More importantly, the peak in post-call similarity is in line with the assumption that analysts are provided with valuable additional information during conference calls, which leads to a topic change in post-call reports.

Thus, the results indicate that earnings conference calls play an important role in disseminating information regarding earnings releases, and the topics contained in them are often discussed in the analyst reports released in the following days.

The fourth research question in this research area was given by the relationship of media richness theory and analyst opinion:

> **Research Question III.4:** To what extent can the media richness of unstructured analyst opinion, as described by media richness theory, help to explain its effect on post earnings call firm stock returns when compared to information sources of lower richness?

Media richness theory predicts that low-richness media types should help information transmission by mitigating issues of uncertainty. As shown in Eickhoff (2017, **paper III.4**), this effect appears to be comparatively small in the observed examples, pointing to a small effect of uncertainty about the information relevant to earnings announcements. Additionally, media richness theory predicts that high-richness media types should help information transmission by mitigating issues of equivocality.

Thus, high-richness media should be able to explain the investment decisions following earnings announcements if equivocality is a relevant problem for decision makers. As shown in

Eickhoff (2017, **paper III.4**), the explanatory power of such a model is comparatively large, indicating that the variables extracted from unstructured content explain a larger portion of the variance within abnormal stock returns on the earnings call day. Thus, these unstructured data sources appear to transmit more investment relevant information when compared to structured data sources within the confines of that analysis. Within the framework of media richness theory, this points to equivocality regarding the interpretation of earnings announcements as a major problem that investors need to mitigate using high-richness media sources. In summary, Eickhoff (2017, **paper III.4**) indicates a low impact of uncertainty for post-earnings call investment decisions while providing much clearer evidence for the impact of equivocality. Finally, the incremental value of the combination of low- and high-richness media types is doubtful if the comparatively low impact of uncertainty, as understood by media richness theory, and the information likely contained in the unstructured content but not captured by text-mining approaches are borne in mind.

2 Implications

2.1 Research Area I: Entrepreneurial Environment

The implications of the results in this research area can be grouped into implications for research, practical implications, and policy implications.

Implications for Research: The first implication of this research area is its contribution toward developing a consensus on the question of what constitutes FinTech. As discussed, due to the rapidly changing landscape in the financial industry in general, FinTech companies in particular, and the relative youth of the "FinTech phenomenon", the lines between incumbent firms and FinTech companies are not distinct. Furthermore, the lines are blurred between traditional tech firms selling their products to banks and the new phenomenon of FinTech, in which firms challenge the established financial industry by providing either what was traditionally considered a financial service or entirely new related services.

To this end, the presented taxonomy of business models can be interpreted in terms of what it does not include. Considered in conjunction with existing definitions of FinTech firms, this enables researchers to focus on the new phenomenon. Additionally, the presented taxonomy provides an overview of the studied phenomenon. Thus, the dimensions and characteristics of FinTech business models included in the taxonomy presented in Eickhoff et al. (2017, **paper I.1**) help to identify different types of FinTech business models by abstraction beyond the business model of individual firms. In conjunction with the presented clustering of firm attributes, this allows for the identification of firms that are especially unlike each other, each of which represents a different facet of the FinTech landscape. Furthermore, the presented dimensions and characteristics provide a basis for further theory development and theory testing related to the FinTech phenomenon (Varshney et al., 2015).

Practical implications: Taxonomies allow for the abstraction needed to identify unoccupied business models, as reflected by combinations of characteristics currently not offered by competing firms. Additionally, incumbents can use the taxonomy to gain an overview of which traditional business models are threatened by new competition and which new business models are being developed. Thus, the presented taxonomy allows practitioners to gain an overview of the current status quo in this rapidly changing environment.

Policy implications: The financial services and banking sector is a highly regulated industry, in which incumbents must comply with regulations regarding problems such as fraud prevention, identity theft, organized crime, and sanctions against nation states. Regulators have established processes to address these and many other concerns with

incumbent firms. FinTechs have not been subject to the same level of scrutiny if they themselves have not been classified as banks or providers of financial services. However, the nature of FinTech business models implies that these firms face many of the same risks as traditional banks. For example, they handle similarly sensitive customer information and may be targeted by illegal activity such as fraud or process financial transactions across national borders, which entails the risk of the malicious use of such services. Thus, it is imperative for regulators to gain an overview of what business models are being created in this new sector of the financial industry. Parts of this new industry segment may create a need for new forms of regulation or an extension of the applicability of existing rules.

2.2 Research Area II: Methodological

The implications of the results presented in Eickhoff and Neuss (2017, **paper II.1**) mainly regard the manner in which applied research using topic modeling techniques is conducted. As shown therein, applied studies often do not present the topic model in a way that makes the result appear trustworthy or accessible to the reader. Related to this concern, few studies use any manual model validation approach, such as coding topics to labels for introspection, or use model fit metrics.

As noted, no implementation of state of the art topic models exist, which feature a graphical user interface, thus rendering use of the methodology difficult for researchers who do not have a technical background. This is obviously prohibitive for entire research domains. The main implication of Eickhoff (2015, **paper II.2**) is given by highlighting the need for methodological transparency in content analysis-based research. This transparency can be achieved by using open source implementations of any applied content analysis techniques, enabling other researchers to replicate prior studies and build upon their results. Furthermore, the presented framework provides guidelines for conducting applied research using such methods.

2.3 Research Area III: Analyst Opinion

As discussed, the results of Eickhoff and Muntermann (2016b, **paper III.1**) add to the growing body of work by suggesting that crowd wisdom as a phenomenon can be used to explain the (sometimes surprising) quality of the content created by large groups. Our contribution to this theoretical body of research is twofold. First, the comparison to stock analysts allows for the benchmarking of crowd wisdom against an expert group. The results of the analysis suggest that in some situations, the crowd can add information in a timelier manner than experts. Second, our analysis supports crowd wisdom theory as proposed by Surowiecki (2005).

We divide the **practical implications** of this research into implications for the financial sector and those for social media users and platforms. Within the financial sector, our results inform the customers of analysts about the conditions under which analyst research is especially valuable but also when it may be wise to resort to social media monitoring tools to gauge the crowd's opinion. The results also reveal how to aggregate the opinions of social media users. Similarly, stock analysts are informed of the circumstances under which it may be wise to listen to social media users' opinions as an additional source of information but also when they are unlikely to provide valuable information. In addition, the operationalization of crowd wisdom-related constructs can help companies refine their social media monitoring tools to better reflect the diverse opinion of the crowd.

Aside from companies, this research should also be relevant to social media content aggregators who need to know the type of data on social media users that interest their customers. Finally, special purpose social networks such as stock recommendation communities or social lending communities are fundamentally based on the concept of crowd wisdom, and their success depends on understanding the conditions under which it arises. The members of such communities expect these platforms to deliver insights gained from this crowd wisdom. Thus, crowd wisdom theory-based results provide such communities with guidelines regarding the makeup of wise crowds.

We distinguish between the theoretical and practical implications of Eickhoff and Muntermann (2017, **paper III.2**). The **practical implications** of the presented research are given by the identification of decision-relevant topics and the resulting reduction of decision complexity, which is desirable because it reduces the risk of information overload. Another possible application of the chosen approach is given by the possible (partial) automation of the decision process. Because no step of the chosen approach requires manual intervention, this could technically be feasible. However, it may not be practically desirable because of the risk of misclassification resulting from a fully automated solution. Additionally, such automation would require extensive back-testing and should incorporate other non-topic-based approaches to stock return estimation beyond the analysis of **paper III.2**.

The **theoretical contribution** of the presented research is given by its addition to the growing stream of literature regarding the risk of information overload and the presented mitigation approach to such problems. We structured the decision problem related to the phenomenon of information overload according to the phases of the decision-making process, providing a theoretical foundation for determining the approachable steps of our analysis (Simon, 1977). As shown, topic mining and identifying the most relevant topics in business communications represent a suitable strategy to cope with information overload. However, the examined context does represent a typical

situation of information overload, meaning a situation in which too large volumes of information need to be processed by a decision maker within a constrained amount of time.

The implications of Eickhoff and Muntermann (2016c, **paper III.3**) are twofold. First, the implication of this study for future research is given by its method of studying the transfer of topics discussed in documents stemming from different media types. Studies of this nature are interesting regarding many other media types than the two sources of analyst opinion studied in this contribution; see section 4.3 (future research) for a more detailed discussion of this possibility. From a practical point of view, the implication of the presented results is given by stressing the importance of analyst conference calls as a means of information disclosure relevant to price discovery on capital markets.

The implications of the results presented in Eickhoff (2017, **paper III.4**) can be spilt up into practical, theoretical and policy implications. From a **practical** perspective, investors should focus both on analyst reports and earnings calls as their primary sources of information regarding the impact of earnings announcements if no automation is required or if text mining is feasible. Keeping in mind the results of Eickhoff and Muntermann (2016c, **paper III.3**), the role of conference calls regarding information dissemination is difficult to overstate. Additionally, as also shown by other studies (Blau et al., 2015; Brockman et al., 2015; Davis et al., 2015; Price et al., 2012; Twedt and Rees, 2012), the presented analysis reinforces the crucial effect of the call and report tone towards investor reactions to this information.

The **theoretical** contribution of this study is twofold. First, its results contribute towards the growing body of research regarding the impact of unstructured, high-richness media types in accounting and finance research by exploring the feasibility of the topic mining of earnings announcements. Second, it contributes to the discussion of media richness theory and its applications by exploring the effects of uncertainty and equivocality regarding the same problem domain. In total, the results indicate that equivocality presents a major challenge for investors. Finally, the results indicate that **policy** makers should keep a watchful eye on the "*private"* information gained by analysts because of their privileged access to top level corporate representatives because the high correlation to abnormal post-earnings returns makes the ability to contribute to the call content inherently valuable.

3 Limitations

3.1 Research Area I: Entrepreneurial Environment

The taxonomy presented in Eickhoff et al. (2017, **paper I.1**) should be interpreted while keeping in mind several assumptions and decisions made during its development, which influence the result of the development process. **First**, regarding the development of dimensions, the selection of dimensions based on the business model literature is inherently selective and thereby limits itself to dimensions that exist a priori in the extant literature. Due to the dynamic nature of business and the FinTech movement in particular, we cannot exclude the possibility that, for a given company, multiple possible characteristics exist that contradict the definition of Nickerson et al. (2013) of mutually exclusive characteristics. Additionally, the development of one taxonomy for all types of FinTech companies limits the granularity of our results in the presented research. Regarding the scope of our study, our taxonomy development was limited to companies contained in our sample. While this sample is quite large, not every company has an inherent need to be listed in such a database. This may be especially true for non-US or non-EU firms because the sample is skewed towards these regions.

3.2 Research Area II: Methodological

The literature review presented in Eickhoff and Neuss (2017, **paper II.1**) is inherently limited by the initial focus on journals contained in the Financial Times 50 journal selection. While this focus is in line with the goals of the analysis, especially regarding applied contributions, there might be another group of applied contributions using topic modeling methodology that is not cited within the explored domain. Additionally, the focus on topic mining limits the scope of the review because the use of other content analysis approaches, such as document summarization, may be able to provide pointers for the research using topic models. Furthermore, due to the backward-forward search approach used to find literature in the review, methodological contributions essentially have less of a chance of being discovered in the review. The framework for hybrid sentiment analysis presented in Eickhoff (2015, **paper II.2**) is limited by the domain portability of the sentiment dictionary used to create training data. Additionally, the resulting model is inherently limited by the accuracy of the dictionary within a given domain.

3.3 Research Area III: Analyst Opinion

Methodological and theoretical limitations are present in Eickhoff and Muntermann (2016b, **paper III.1**), which warrant discussion. There is no reliable way to determine

which portion of social media users are professional or "hobby" analysts, which makes the distinction between the two groups blurry. The available data do not enable insight into social media users' social hierarchy, which may provide a more suitable measure of independence and authority as a precursor to crowd wisdom. Finally, the aggregation of social media users' opinions to the singular sentiment measure sacrifices the diversity of the groups' opinions. Other measures may well be more suitable to capture this diversity within the crowd. Beyond these methodological considerations, it is also important to remember that the presented results may be domain specific. The reproduction of a similar analysis for a non-economic, or at least not as explicitly financial, domains could provide an interesting comparison. Additionally, the presented study focused on 3 out of 5 constructs that the crowd wisdom theory proposes. Future studies should focus on trust and coordination as drivers of crowd wisdom, and the study only derives metrics for one of the two groups.

The main limitation of Eickhoff and Muntermann (2017, **paper III.2**) is given by the static nature of its model. This concerns both the chosen topic modeling approach and the market model-based abnormal return metric, which assumes a time-constant linear relationship between the reference index and the firm in question during the event period. The topic modeling limitation could be mitigated using implementations of topic models, which allow for the trained classification to be updated in the face of new training data, enabling an evolution of the model over time (Blei and Lafferty, 2006; Řehůřek and Sojka, 2010). While the chosen return model appears prudent for the short event period used in the model, other techniques should be explored if longer event periods are of interest. Additionally, a larger company sample would likely improve upon the presented results.

Regarding Eickhoff and Muntermann (2016c, **paper III.3**), the main limitation of this short paper is given by the scope of its analysis. We examine the case of a single firm in this study. While there is a methodological reason for this (a cross-company trained topic model would be unlikely to work well for our analysis), validation of these results in future research using different samples is desirable.

As discussed in Eickhoff (2017, **paper III.4**), some of the topics that are identified by the chosen approach are industry specific. Thus, an extension of the chosen approach to a larger sample with industry specific topics is desirable. Additionally, the identified topics are time-constant, and a topic model can only identify topics that are previously contained in a corpus. Thus, entirely new topics will be missed and would be interesting to add in the future. Furthermore, the observation count in the final regression models is limited by data availability regarding structured analyst consensus estimates. Like the inclusion of industry-specific topics, this may be mitigated by using a larger firm sample as the basis of the analysis.

4 Future Research

4.1 Research Area I: Entrepreneurial Environment

In this research area, a FinTech business model taxonomy was developed in Eickhoff et al. (2017, **paper I.1**). As discussed, the aspiration to generalize the business models of all types of FinTech companies limits the granularity of both the dimensions and characteristics developed for this taxonomy. Thus, future research focusing on more specialized taxonomies may provide further insights. Additionally, more extensive use of the developed taxonomy is possible. For example, clustering can be performed on the characteristics assigned to each company as opposed to the approach chosen here: namely, to cluster the tags not used during taxonomy development as a confirmatory effort. This could help identify which combinations of characteristics are common and show patterns across diverse types of business models, identifying which roles are already being filled by companies and which are not. The dynamic development of the FinTech industry creates a need for future investigation. New companies may follow entirely different business models than those included in our dataset. Thus, future research may focus on exploring whether the presented taxonomy still holds.

4.2 Research Area II: Methodological

The main implication for future research of this research area is given by the identification of research gaps found in Eickhoff and Neuss (2017, **paper II.1**). (M)IS has established itself as a reference discipline for other managerial fields regarding topic modeling methodology. The exploration of the theoretical foundations of the use and interpretation of topic models, as well as their capabilities regarding the generation and testing of social, economic and systems theory, present an opportunity to strengthen this referential role of IS.

This is especially true when looking at the possibilities of topic modeling methodology in mixed method research designs, which combine the usefulness of topic modeling for analyzing large document collections and the intricacy of qualitative research designs, which enable researchers to analyze content in much more detail but are prohibitively time consuming when applied to large document collections.

Based on the framework for sentiment analysis presented in Eickhoff (2015, **paper II.2**), a possible extension of the presented framework is given by integrating a stage for bootstrapping sentiment dictionaries. A semi-supervised approach to dictionary creation could be based on existing lexical databases such as *SentiWordNet* (Baccianella et al., 2010).

4.3 Research Area III: Analyst Opinion

There are two distinct ways to extend on the results of Eickhoff and Muntermann (2016b, **paper III.1**). First, the two remaining constructs of crowd wisdom theory that are not operationalized in this study due to a lack of data about them could be integrated in the existing analysis to provide a more comprehensive picture of crowd wisdom in the studied domain.

Second, while analysts have been treated as the opposite of the social media crowd in this study, they themselves resemble a crowd, and the criteria under which this crowd of analysts is especially effective could be studied. Indeed, many of the same data points used to operationalize the constructs of crowd wisdom theory regarding social media users are also available for the analyst side. For example, the names and employers of analysts are included in analyst reports, and the location from which a report is published is also known or discoverable. Additionally, the number of different analysts covering a firm at a given point in time is determinable, especially if *I/B/E/S* data are used as a proxy, since the database makes the number of covering analysts readily available. Thus, the setting of Eickhoff and Muntermann (2016b, **paper III.1**) could be extended to compare two different types of crowds. Additionally, the study could be extended upon by adopting topic modeling as another technique for analyzing the two content types to study the topics for which each group is favored.

On the basis of the results presented in Eickhoff and Muntermann (2017, **paper III.2**), the presented approach to information complexity reduction could be combined with other approaches, such as rule-based filtering. Furthermore, information overload is a relevant problem beyond the financial domain. Consequently, other domains could be studied using similar approaches. The main task in such a domain transfer of the presented approach is finding a decision-relevant metric, such as stock returns in the financial case.

Three areas of possible future research are identified in Eickhoff and Muntermann (2016c, **paper III.3**). First, the presented approach can be transferred to other content domains. One domain that could benefit from such an approach is marketing research, where the transmission of topics introduced in a marketing campaign through user-generated content could be of interest. Because the main requirement for the presented approach is the availability of historical training data and because the marketing research already uses analyzed social media content, this appears to be a small hurdle in this domain.

Second, the level of analysis could be shifted towards a more granular inspection of the behavior of individual topics. Because this research only focused on document-

level topic similarity, this extension would both be methodologically feasible and interesting in the context of analyst opinion because there may be some topics that do not follow the conference-call-to-analyst-report transfer pattern that is shown in this research. Such a granular analysis could also contribute on a theoretical level because it could provide more insights on analyst herding behavior based on the topic. Another interesting question regarding the behavior of individual topics is given by the question of whether topics introduced by the analysts in the Q&A part of the call, which were not present in the presentation part of call, are more likely to have a long-lasting impact on the topic structure of reports or an immediate market reaction.

Third, the presented approach can be extended to more than two content types at once. One reason to be interested in such an extension is the fact that the presented approach cannot prove that a topic transfer occurred but only that one content type adopted the same topic after the other. This does not preclude that both content types adopt topics from a third or multiple other content domains and only differ in their information processing speed. Thus, an extension beyond two content types is desirable because financial analysts are, as their name suggests, interpreters of information and are not the originators of the information themselves.

Finally, the results presented in Eickhoff (2017, **paper III.4**) present several avenues of future research regarding the effects of media richness on information transmission in the financial domain. This study could be extended using a larger company sample, enabling the analysis of industry-specific effects regarding earnings announcements and their relationship to low- and high-richness media types. Additionally, other non-analyst-authored media types could be of interest to such an extension. Furthermore, because the model in this contribution is time constant, which contradicts the reality of ever-changing topics in companies' ongoing operations, a more adaptive topic model could be used to reflect the changes in the situation of firms over time. Indeed, adaptive topic models are available (Blei and Lafferty, 2006), as is discussed in Eickhoff and Neuss (2017, **paper II.2**). This adaptive modeling would also make the analysis more resistant to market shocks, such as economic crises. Another extension could be made by studying the effects of the opinions of individual analysts or analyst-employers because person- or firm-level data are available in the database used.

In summary, regarding all studies presented in this third research area of the thesis, comparisons between the effectiveness of other content analysis techniques are an interesting avenue for future research. For example, vector representations of textual content have recently risen in popularity (Le and Mikolov, 2014) and offer another interesting approach to the task of creating compressed representations of document content. Methodological comparisons would also reduce the risk of method bias in the presented results (Podsakoff et al., 2003).

References

Ågerfalk, P. J. (2013). "Embracing Diversity through Mixed Methods Research," *European Journal of Information Systems* 22 (3), pp. 251-256.

Ahmad, S. N., and Laroche, M. (2015). "How Do Expressed Emotions Affect the Helpfulness of a Product Review? Evidence from Reviews Using Latent Semantic Analysis," *International Journal of Electronic Commerce* 20 (1), pp. 76-111.

AlSumait, L., Barbará, D., and Domeniconi, C. (2008). "On-Line Lda: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking," in: *IEEE International Conference on Data Mining*. pp. 3-12.

Aral, S., Ipeirotis, P. G., and Taylor, S. J. (2011). "Content and Context: Identifying the Impact of Qualitative Information on Consumer Choice," In: *Proceedings of the International Conference on Information Systems*, Shanghai.

Arazy, O., Morgan, W., and Patterson, R. (2006). "Wisdom of the Crowds: Decentralized Knowledge Construction in Wikipedia," *16th Annual Workshop on Information Technologies & Systems (WITS) Paper*.

Arazy, O., and Woo, C. (2007). "Enhancing Information Retrieval through Statistical Natural Language Processing: A Study of Collocation Indexing," *MIS Quarterly* 31 (3), pp. 525-546.

Aryal, A., Gallivan, M., and Tao, Y. Y. (2015). "Using Latent Semantic Analysis to Identify Themes in Is Healthcare Research," In: *Proceedings of the Americas Conference on Information Systems*, Puerto Rico: AISeL.

Asquith, P., Mikhail, M. B., and Au, A. S. (2005). "Information Content of Equity Analyst Reports," *Journal of Financial Economics* 75 (2), pp. 245-282.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). "On Smoothing and Inference for Topic Models," In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*: AUAI Press, pp. 27-34.

Aue, A., and Gamon, M. (2005). "Customizing Sentiment Classifiers to New Domains : A Case Study," *Proceedings of Recent Advances in Natural Language Processing RANLP* 49, pp. 207-218.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). "Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," In: *Proceedings of the LREC*, pp. 2200-2204.

Back, B., Toivonen, J., Vanharanta, H., and Visa, A. (2001). "Comparing Numerical Data and Text Information from Annual Reports Using Self-Organizing Maps," *International Journal of Accounting Information Systems* 2 (4), pp. 249-269.

Baden-Fuller, C., and Morgan, M. S. (2010). "Business Models as Models," *Long range planning* 43 (2), pp. 156-171.

Bagozzi, R. P. (2011). "Measurement and Meaning in Information Systems and Organizational Research: Methodological and Philosophical Foundations," *MIS Quarterly* 35 (2), pp. 261-292.

Baker, L. D., and McCallum, A. K. (1998). "Distributional Clustering of Words for Text Classification," In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, New York, USA: ACM, pp. 96-103.

Ball, R., and Brown, P. (1968). "An Empirical Evaluation of Accounting Income Numbers," *Journal of Accounting Research* 6 (2), pp. 159-178.

Bao, Y., and Datta, A. (2014). "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures," *Management Science* 60 (6), pp. 1371-1391.

Bapna, R., Goes, P., Gupta, A., and Jin, Y. (2004). "User Heterogeneity and Its Impact on Electronic Auction Market Design: An Empirical Exploration," *MIS Quarterly*, pp. 21-43.

Barber, B., Lehavy, R., McNichols, M., and Trueman, B. (2001). "Can Investors Profit from the Prophets? Security Analyst Recommendations and Stock Returns," *The Journal of Finance* 56 (2), pp. 531-563.

Baskerville, R., Pries-Heje, J., and Venable, J. (2009). "Soft Design Science Methodology," In: *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, Malvern, Pa, USA: ACM, p. 9.

Bellstam, G., Bhagat, S., and Cookson, J. A. (2016). "A Text-Based Analysis of Corporate Innovation," *SSRN* 2803232.

Bendle, N. T., and Wang, X. S. (2016). "Uncovering the Message from the Mess of Big Data," *Business Horizons* 59 (1), pp. 115-124.

Bergamaschi, S., and Po, L. (2014). "Comparing Lda and Lsa Topic Models for Content-Based Movie Recommendation Systems," In: *Proceedings of the International Conference on Web Information Systems and Technologies*: Springer, pp. 247-263.

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., and Venkatraman, N. (2013). "Digital Business Strategy: Toward a Next Generation of Insights," *MIS Quarterly* 37 (2), pp. 471-482.

Blau, B. M., DeLisle, J. R., and Price, S. M. (2015). "Do Sophisticated Investors Interpret Earnings Conference Call Tone Differently Than Investors at Large? Evidence from Short Sales," *Journal of Corporate Finance* 31, pp. 203-219.

Blei, D. M. (2012). "Probabilistic Topic Models," *Communications of the ACM* 55 (4), pp. 77-84.

Blei, D. M. (2014). "Hierarchical Lda with a Fixed Depth Tree and a Stick Breaking Prior on the Depth Weights,".

Blei, D. M., and Lafferty, J. D. (2006). "Dynamic Topic Models," In: *Proceedings of the International Conference on Machine Learning*, Pittsburgh PA: ACM, pp. 113-120.

Blei, D. M., and Lafferty, J. D. (2007). "A Correlated Topic Model of Science," *The Annals of Applied Statistics*, pp. 17-35.

Blei, D. M., and Lafferty, J. D. (2009a). "Topic Models," *Text mining: Classification, Clustering, and Applications* 10 (71), p. 34.

Blei, D. M., and Lafferty, J. D. (2009b). "Visualizing Topics with Multi-Word Expressions," *arXiv* 0907.1013.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet Allocation," *The Journal of Machine Learning Research* 3, pp. 993-1022.

BleiLab. (2016). "Blei Lab Github Repository." Retrieved 12/2/2016, from https://github.com/blei-lab

Boell, S. K., and Cecez-Kecmanovic, D. (2015). "On Being 'Systematic' in Literature Reviews in Is," *Journal of Information Technology* 30 (2), pp. 161-173.

Bollen, J., Mao, H., and Zeng, X. (2011). "Twitter Mood Predicts the Stock Market," *Journal of Computational Science* 2 (1), pp. 1-8.

Boukus, E., and Rosenberg, J. V. (2006). "The Information Content of Fomc Minutes," *SSRN* 922312.

Bradshaw, M. T. (2009). “Analyst Information Processing, Financial Regulation, and Academic Research,“ *The Accounting Review* 84, pp. 1073-1083.

Brockman, P., Li, X., and Price, S. M. (2015). “Differences in Conference Call Tones: Managers Vs. Analysts,“ *Financial Analysts Journal* 71 (4), pp. 24-42.

Brown, L. D., and Rozeff, M. S. (1978). “The Superiority of Analyst Forecasts as Measures of Expectations: Evidence from Earnings,“ *The Journal of Finance* 33 (1), pp. 1-16.

Brown, P., Foster, G., and Noreen, E. (1985). *Security Analyst Multi-Year Earnings Forecasts and the Capital Market*. American Accounting Association.

Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H.-m., and Steele, L. B. (2014). “The Information Content of Mandatory Risk Factor Disclosures in Corporate Filings,“ *Review of Accounting Studies* 19 (1), pp. 396-455.

Castelló, I., Etter, M., and Nielsen, F. Å. (2016). “Strategies of Legitimacy through Social Media: The Networked Strategy,“ *Journal of Management Studies* 53 (3), pp. 402-432.

Chaney, A. J. B. (2014). “Online Topic Model Visualization,“.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). “Reading Tea Leaves: How Humans Interpret Topic Models,“ In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 288-296.

Chang, Y.-L., and Chien, J.-T. (2009). “Latent Dirichlet Learning for Document Summarization,“ In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*: IEEE, pp. 1689-1692.

Charlin, L., Ranganath, R., McInerney, J., and Blei, D. M. (2015). “Dynamic Poisson Factorization,“ In: *Proceedings of the ACM Conference on Recommender Systems*: ACM, pp. 155-162.

Chen, H., Chiang, R. H. L., and Storey, V. C. (2012). “Business Intelligence and Analytics: From Big Data to Big Impact,“ *MIS Quarterly* 36 (4), pp. 1165-1188.

Chen, H., De, P., Hu, Y., and Hwang, B.-H. (2014). “Wisdom of Crowds: The Value of Stock Opinions Transmitted through Social Media,“ *Review of Financial Studies* 27 (5), pp. 1367-1403.

Chen, H., and Zhao, J. L. (2015). “Istopic: Understanding Information Systems Research through Topic Models,“ In: *Proceedings of the International Conference on Information Systems*, Fort Worth, USA: AISeL.

Chong, W., Blei, D., and Li, F.-F. (2009). “Simultaneous Image Classification and Annotation,“ In: *Proceedings of the Computer Vision and Pattern Recognition, 2009. CVPR 2009*: IEEE, pp. 1903-1910.

Cicon, J. E., Ferris, S. P., Kammel, A. J., and Noronha, G. (2012). “European Corporate Governance: A Thematic Analysis of National Codes of Governance,“ *European Financial Management* 18 (4), pp. 620-648.

Clement, M. B., and Tse, S. Y. (2005). “Financial Analyst Characteristics and Herding Behavior in Forecasting,“ *The Journal of Finance* 60 (1), pp. 307-341.

Cohen, L., Tyler, R., Contreiras, D., and Buxton, P. (2016). “Blockchain's Three Capital Markets Innovations Explained,“ *International financial law review* 35 (26), pp. 9-9.

Cohen Priva, U., and Austerweil, J. L. (2015). “Analyzing the History of Cognition Using Topic Models,“ *Cognition* 135, pp. 4-9.

Corrado, C. J. (2011). “Event Studies: A Methodology Review,“ *Accounting & Finance* 51 (1), pp. 207-234.

Cowles , A. r. (1933). “Can Stock Market Forecasters Forecast?,“ *Econometrica: Journal of the Econometric Society*, pp. 309-324.

Croft, W. B., and Harper, D. J. (1979). "Using Probabilistic Models of Document Retrieval without Relevance Information," *Journal of Documentation* 35 (4), pp. 285-295.

Crossno, P. J., Wilson, A. T., Shead, T. M., and Dunlavy, D. M. (2011). "Topicview: Visually Comparing Topic Models of Text Collections," In: *Proceedings of the International Conference on Tools with Artificial Intelligence*: IEEE, pp. 936-943.

Crunchbase. (2016). "Crunchbase." Retrieved 03/13/2017, from http://www.crunchbase.com

Cutting, D. K., Karger, D. R., Pedersen, J. O., and Scatter, T. J. W. (1992). "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections," In: *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 318-329.

Daft, R. L., and Lengel, R. H. (1983). "Information Richness. A New Approach to Managerial Behavior and Organization Design,".

Daft, R. L., and Lengel, R. H. (1986). "Organizational Information Requirements, Media Richness and Structural Design," *Management science* 32 (5), pp. 554-571.

Daft, R. L., and Macintosh, N. B. (1981). "A Tentative Exploration into the Amount and Equivocality of Information Processing in Organizational Work Units," *Administrative science quarterly*, pp. 207-224.

Davis, A. K., Ge, W., Matsumoto, D., and Zhang, J. L. (2015). "The Effect of Manager-Specific Optimism on the Tone of Earnings Conference Calls," *Review of Accounting Studies* 20 (2), pp. 639-673.

De Franco, G., Hope, O.-K., and Larocque, S. (2015). "Analysts Choice of Peer Companies," *Review of Accounting Studies* 20 (1), pp. 82-109.

Debortoli, S., Müller, O., Junglas, I., and Brocke, J. (2016). "Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial," *Communications of the Association for Information Systems* 39 (7).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science* 41 (6), pp. 391-407.

Dennis, A. R., and Kinney, S. T. (1998). "Testing Media Richness Theory in the New Media: The Effects of Cues, Feedback, and Task Equivocality," *Information Systems Research* 9 (3), pp. 256-274.

Devereux, G. (1967). *From Anxiety to Method in the Behavioral Sciences*. The Hague, Paris: Mouton & Co.

DiMaggio, P. (2015). "Adapting Computational Text Analysis to Social Science (and Vice Versa)," *Big Data & Society* 2 (2), pp. 1-5.

Dimson, E., and Mussavian, M. (1998). "A Brief History of Market Efficiency," *European Financial Management* 4 (1), pp. 91-103.

Dogac, A., Laleci, G., Kabak, Y., and Cingil, I. (2002). "Exploiting Web Service Semantics: Taxonomies Vs. Ontologies," *IEEE Data Engineering Bulletin* 25 (4), pp. 10-16.

Dolley, J. C. (1933). "Characteristics and Procedure of Common Stock Split-Ups," *Harvard Business Review* 11 (3), pp. 316-326.

Doty, D. H., and Glick, W. H. (1994). "Typologies as a Unique Form of Theory Building: Toward Improved Understanding and Modeling," *Academy of Management Review* 19 (2), pp. 230-251.

Du, L., Buntine, W., and Jin, H. (2010). "A Segmented Topic Model Based on the Two-Parameter Poisson-Dirichlet Process," *Machine Learning* 81 (1), pp. 5-19.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). "Inductive Learning Algorithms and Representations for Text Categorization," In: *Proceedings of the International conference on Information and knowledge management*: ACM, pp. 148-155.

Dumais, S. T. (2004). "Latent Semantic Analysis," *Annual Review of Information Science and Technology* 38 (1), pp. 188-230.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). "Using Latent Semantic Analysis to Improve Access to Textual Information," In: *Proceedings of the ACM Conference on Human factors in computing systems*: SIGCHI pp. 281-285.

Earl, M. (2001). "Knowledge Management Strategies: Toward a Taxonomy," *Journal of management information systems* 18 (1), pp. 215-233.

Eickhoff, M. (2015). "A Hybrid Domain-Portable Framework for Sentiment Classification," In: *Proceedings of the 10th International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, B. Donnellan, M. Helfert, J. Kenneally, D. VanderMeer, M. Rothenberger and R. Winter (eds.): Springer International Publishing, pp. 215-219.

Eickhoff, M. (2017). "What Do They Mean? Using Media Richness as an Indicator for the Information Value of Stock Analyst Opinion Regarding Post-Earnings Firm Performance," In: *Proceedings of the 50th Hawaii International Conference on System Sciences*, Hawaii, USA: AISeL.

Eickhoff, M., and Muntermann, J. (2015). "Stock Analysts Vs. The Crowd: A Study on Mutual Prediction," In: *Proceedings of the 19th Pacific Asia Conference on Information Systems*, Singapore: AISeL.

Eickhoff, M., and Muntermann, J. (2016a). "How to Conquer Information Overload? Supporting Financial Decisions by Identifying Relevant Conference Call Topics," In: *Proceedings of the 20th Pacific Asia Conference on Information Systems*, Chiayi, Taiwan: AISeL.

Eickhoff, M., and Muntermann, J. (2016b). "Stock Analysts Vs. The Crowd: Mutual Prediction and the Drivers of Crowd Wisdom," *Information & Management* 53 (7), pp. 835-845.

Eickhoff, M., and Muntermann, J. (2016c). "They Talk but What Do They Listen To? Analyzing Financial Analysts Information Processing Using Latent Dirichlet Allocation," In: *Proceedings of the 20th Pacific Asia Conference on Information Systems*, Chiayi, Taiwan: AISeL.

Eickhoff, M., and Muntermann, J. (2017). "Identifying Relevant Topics in Business Communication to Conquer Information Overload: Supporting Financial Decision Making Related to Analyst Conference Calls," *Under Review*.

Eickhoff, M., Muntermann, J., and Weinrich, T. (2017). "What Do Fintechs Actually Do? A Taxonomy of Fintech Business Models," *Under Review*.

Eickhoff, M., and Neuss, N. (2017). "Topic Modelling Methodology: Its Use in Information Systems and Other Managerial Disciplines," In: *Proceedings of the 25th European Conference on Information Systems (ECIS)*.

El Sawy, O. A., and Pereira, F. (2013). *Business Modelling in the Dynamic Digital Space: An Ecosystem Approach*, (1 ed.). Electronic: Springer.

Elder-Vass, D. (2014). "Debate: Seven Ways to Be a Realist About Language," *Journal for the Theory of Social Behaviour* 44 (3), pp. 249-267.

Evangelopoulos, N., Zhang, X., and Prybutok, V. R. (2012). "Latent Semantic Analysis: Five Methodological Recommendations," *European Journal of Information Systems* 21 (1), pp. 70-86.

Evans, J. A., and Aceves, P. (2016). "Machine Translation: Mining Text for Social Theory," *Annual Review of Sociology* 42, pp. 21-50.

Fama, E. F., Fisher, L., Jensen, M. C., and Roll, R. (1969). "The Adjustment of Stock Prices to New Information," *International economic review* 10 (1), pp. 1-21.

Fiedler, K. D., Grover, V., and Teng, J. T. C. (1996). "An Empirically Derived Taxonomy of Information Technology Structure and Its Relationship to Organizational Structure," *Journal of Management Information Systems* 13 (1), pp. 9-34.

Foltz, P. W. (2007). "Discourse Coherence and Lsa," in *Handbook of Latent Semantic Analysis*. pp. 167-184.

Franco, G., Hope, O.-K., Vyas, D., and Zhou, Y. (2015). "Analyst Report Readability," *Contemporary Accounting Research* 32 (1), pp. 76-104.

Frankel, R., Johnson, M., and Skinner, D. J. (1999). "An Empirical Examination of Conference Calls as a Voluntary Disclosure Medium," *Journal of Accounting Research* 37 (1), pp. 133-150.

Frankel, R., Kothari, S. P., and Weber, J. (2006). "Determinants of the Informativeness of Analyst Research," *Journal of Accounting and Economics* 41 (1-2), pp. 29-54.

Fuller, R. B. (1957). "A Comprehensive Anticipatory Design Science," *Royal Architectural Institute of Canada* 34 (?).

Galton, F. (1907). "Vox Populi (the Wisdom of Crowds)," *Nature* 75 (7), pp. 450-451.

George, G., Haas, M. R., and Pentland, A. (2016). "From the Editors: Big Data and Data Science Methods for Management Research," *Academy of Management Journal* 59 (5), pp. 1493-1507.

Gerrish, S., and Blei, D. M. (2010). "A Language-Based Approach to Measuring Scholarly Impact," In: *Proceedings of the International Conference on Machine Learning*, pp. 375-382.

GI-Team. (2002). "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries." Retrieved 13.3., 2014, from http://www.wjh.harvard.edu/~inquirer/homecat.htm

Giles, J. (2005). "Internet Encyclopaedias Go Head to Head," *Nature* 438 (7070), pp. 900-901.

Giorgi, S., and Weber, K. (2015). "Marks of Distinction: Framing and Audience Appreciation in the Context of Investment Advice," *Administrative Science Quarterly*, pp. 1-35.

Glass, R. L., and Vessey, I. (1995). "Contemporary Application-Domain Taxonomies," *IEEE Software* 12 (4), pp. 63-76.

Gopalan, P. K., Charlin, L., and Blei, D. (2014). "Content-Based Recommendations with Poisson Factorization," In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 3176-3184.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., and Hu, X. (2007). "Using Lsa in Autotutor: Learning through Mixed Initiative Dialogue in Natural Language," in *Handbook of Latent Semantic Analysis*. pp. 243-262.

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., and Peacock, R. (2005). "Storylines of Research in Diffusion of Innovation: A Meta-Narrative Approach to Systematic Review," *Social Science & Medicine* 61 (2), pp. 417-430.

Gregory, F. H. (1993). "Soft Systems Methodology to Information Systems: A Wittgensteinian Approach," *Information Systems Journal* 3 (3), pp. 149-168.

Griffiths, T. L., and Steyvers, M. (2004). "Finding Scientific Topics," *Proceedings of the National Academy of Sciences* 101 (suppl 1), pp. 5228-5235.

Grimmer, J. (2010). "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases," *Political Analysis* 18 (1), pp. 1-35.

Grimmer, J., and King, G. (2011). "General Purpose Computer-Assisted Clustering and Conceptualization," *Proceedings of the National Academy of Sciences* 108 (7), pp. 2643-2650.

Grimmer, J., and Stewart, B. M. (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21 (3), pp. 267-297.

Groysberg, B., Healy, P. M., and Maber, D. A. (2011). "What Drives Sell-Side Analyst Compensation at High-Status Investment Banks?," *Journal of Accounting Research* 49 (4), pp. 969-1000.

Grün, B., and Hornik, K. (2011). "Topicmodels : An R Package for Fitting Topic Models," *Journal of Statistical Software* 40 (13), pp. 1-30.

Gruninger, M., Bodenreider, O., Olken, F., Obrst, L., and Yim, P. (2008). "Ontology Summit 2007-Ontology, Taxonomy, Folksonomy: Understanding the Distinctions," *Applied Ontology* 3 (3), pp. 191-200.

Guarino, N. (1998). "Formal Ontology and Information Systems," In: *Proceedings of the Proceedings of FOIS*, pp. 81-97.

Guerreiro, J., Rita, P., and Trigueiros, D. (2016). "A Text Mining-Based Review of Cause-Related Marketing Literature," *Journal of Business Ethics*, pp. 111-128.

Günther, F., Dudschig, C., and Kaup, B. (2015). "Lsafun-an R Package for Computations Based on Latent Semantic Analysis," *Behavior Research Methods* 47 (4), pp. 930-944.

Guo, Y., Barnes, S. J., and Jia, Q. (2017). "Mining Meaning from Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation," *Tourism Management* 59, pp. 467-483.

Halevy, A., Norvig, P., and Pereira, F. (2009). "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems* 24 (2), pp. 8-12.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). "Studying the History of Ideas Using Topic Models," In: *Proceedings of the Conference on empirical methods in natural language processing*: Association for Computational Linguistics, pp. 363-371.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques, Third Edition (the Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.

Henry, E. (2008). "Are Investors Influenced by How Earnings Press Releases Are Written?," *Journal of Business Communication* 45 (4), pp. 363-407.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). "Design Science in Information Systems Research," *MIS Quarterly* 28 (1), pp. 75-105.

Heylighen, F. (2001). "Mining Associative Meanings from the Web: From Word Disambiguation to the Global Brain," In: *Proceedings of the Trends in Special Language and Language Technology*, R. Temmerman (ed.), Brussels: Standaard Publishers.

Hirschheim, R., Klein, H. K., and Lyytinen, K. (1995). *Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations*. Cambridge: Cambridge University Press.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). "Online Learning for Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems*, J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel and A. Culotta (eds.). Curran Associates, Inc., pp. 856-864.

Hofmann, T. (1999). "Probabilistic Latent Semantic Indexing," In: *Proceedings of the ACM SIGIR conference on Research and development in information retrieval*: ACM, pp. 50-57.

Hofmann, T. (2001). "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning* 42 (1), pp. 177-196.

Hong, H., Kubik, J. D., and Solomon, A. (2000). "Security Analysts' Career Concerns and Herding of Earnings Forecasts," *The RAND Journal of Economics* 31, pp. 121-121.

Hope, O.-K. (2003). "Accounting Policy Disclosures and Analysts' Forecasts," *Contemporary Accounting Research* 20 (2), pp. 295-321.

Hu, M., and Liu, B. (2004). "Mining and Summarizing Customer Reviews," *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04* 04, pp. 168-168.

Huang, A., Lehavy, R., Zang, A., and Zheng, R. (2015). "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach," *Ross School of Business Working Paper* 1229.

Huang, A. H., Zang, A. Y., and Zheng, R. (2014). "Evidence on the Information Content of Text in Analyst Reports," *The Accounting Review* 89 (6), pp. 2151-2180.

Huber, J., Kamakura, W., and Mela, C. F. (2014). "A Topical History of Jmr," *Journal of Marketing Research* 51 (1), pp. 84-91.

Husbands, P., Simon, H., and Ding, C. H. (2001). "On the Use of the Singular Value Decomposition for Text Retrieval," *Computational information retrieval* 5, pp. 145-156.

Ignatow, G. (2015). "Theoretical Foundations for Digital Text Analysis," *Journal for the Theory of Social Behaviour* 46 (1), pp. 104-120.

Iivari, J. (2007). "A Paradigmatic Analysis of Information Systems as a Design Science," *Scandinavian Journal of Information Systems* 19 (2), p. 5.

Irani, A. J., and Karamanou, I. (2003). "Regulation Fair Disclosure, Analyst Following, and Analyst Forecast Dispersion," *Accounting Horizons* 17 (1), pp. 15-29.

Jacobs, B. J. D., Donkers, B., and Fok, D. (2016). "Model-Based Purchase Predictions for Large Assortments Model-Based Purchase Predictions for Large Assortments," *Marketing Science* 35 (3), pp. 389-404.

Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., and Ramakrishnan, N. (2013). "Forex-Foreteller: Currency Trend Modeling Using News Articles," In: *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 1470-1473.

Johnson, R. B., and Onwuegbuzie, A. J. (2004). "Mixed Methods Research: A Research Paradigm Whose Time Has Come," *Educational Researcher* 33 (7), pp. 14-26.

Kaplan, S., and Vakili, K. (2015). "The Double-Edged Sword of Recombination in Breakthrough Innovation," *Strategic Management Journal* 36 (10), pp. 1435-1457.

Kerlinger, F. N., and Lee, H. B. (2000). *Foundations of Behavioral Research*, (4 ed.). Fort Worth, USA: Harcourt College Publishers.

Kim, O., Lim, S. C., and Shaw, K. W. (2001). "The Inefficiency of the Mean Analyst Forecast as a Summary Forecast of Earnings," *Journal of Accounting Research* 39 (2), pp. 329-335.

Kintsch, W., and Bowles, A. R. (2002). "Metaphor Comprehension: What Makes a Metaphor Difficult to Understand?," *Metaphor & Symbol* 17 (4), pp. 249-262.

Kittur, A., and Kraut, R. E. (2008). "Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination," In: *Proceedings of the Proceedings of the 2008 ACM conference on Computer supported cooperative work*: ACM, pp. 37-46.

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., and Visa, A. (2004). "Combining Data and Text Mining Techniques for Analysing Financial Reports," *Intelligent systems in accounting, finance and management* 12 (1), pp. 29-41.

Koukal, A., Gleue, C., and Breitner, M. (2014a). "Enhancing Literature Review Methods - Evaluation of a Literature Search Approach Based on Latent Semantic Indexing," In: *Proceedings of the International Conference on Information Systems*, Auckland.

Koukal, A., Gleue, C., and Breitner, M. (2014b). "Enhancing Literature Review Methods - Towards More Efficient Literature Research with Latent Semantic Indexing," In: *Proceedings of the European Conference on Information Systems*, Tel Aviv, Israel: AISeL.

Kulkarni, S. S., Apte, U. M., and Evangelopoulos, N. E. (2014). "The Use of Latent Semantic Analysis in Operations Management Research," *Decision Sciences* 45 (5), pp. 971-994.

Kundu, A., Jain, V., Kumar, S., and Chandra, C. (2015). "A Journey from Normative to Behavioral Operations in Supply Chain Management: A Review Using Latent Semantic Analysis," *Expert Systems with Applications* 42 (2), pp. 796-809.

Lai, L. S., and To, W. (2015). "Content Analysis of Social Media: A Grounded Theory Approach," *Journal of Electronic Commerce Research* 16 (2), p. 138.

Landauer, T. K. (2002). "On the Computational Basis of Learning and Cognition: Arguments from Lsa," *Psychology of Learning and Motivation* 41, pp. 43-84.

Landauer, T. K. (2007). "Lsa as a Theory of Meaning," in *Handbook of Latent Semantic Analysis*. pp. 3-34.

Landauer, T. K., and Dumais, S. T. (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," *Psychological Review* 104 (2), p. 211.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). "An Introduction to Latent Semantic Analysis," *Discourse Processes* 25 (2-3), pp. 259-284.

Landsman, W. R., and Maydew, E. L. (2002). "Has the Information Content of Quarterly Earnings Announcements Declined in the Past Three Decades?," *Journal of Accounting Research* 40 (3), pp. 797-808.

Larsen, K. R., and Bing, C. H. (2016). "A Tool for Addressing Construct Identity in Literature Reviews and Meta Analyses," *MIS Quarterly* 40 (3), pp. 529-551.

Le, Q. V., and Mikolov, T. (2014). "Distributed Representations of Sentences and Documents," In: *Proceedings of the International Conference on Machine Learning*, Beijing, China, pp. 1188-1196.

Lee, D. D., and Seung, H. S. (2001). "Algorithms for Non-Negative Matrix Factorization," In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 556-562.

Lee, J. (2015). “Can Investors Detect Managers' Lack of Spontaneity? Adherence to Predetermined Scripts During Earnings Conference Calls,“ *The Accounting Review* 91 (1), pp. 229-250.

Lee, M., Pincombe, B., and Welsh, M. (2005). “An Empirical Evaluation of Models of Text Document Similarity,“ In: *Proceedings of the*: Cognitive Science Society.

Lee, S., Baker, J., Song, J., and Wetherbe, J. C. (2010). “An Empirical Comparison of Four Text Mining Methods,“ In: *Proceedings of the Hawaii International Conference on System Sciences*: IEEE, pp. 1-10.

Li, F. (2010a). “The Information Content of Forward-Looking Statements in Corporate Filings—a Naïve Bayesian Machine Learning Approach,“ *Journal of Accounting Research* 48 (5), pp. 1049-1102.

Li, F. (2010b). “Textual Analysis of Corporate Disclosures: A Survey of the Literature,“ *Journal of Accounting Literature* 29, pp. 143-143.

Lin, C., and He, Y. (2009). “Joint Sentiment/Topic Model for Sentiment Analysis,“ In: *Proceedings of the ACM Conference on Information and knowledge management*: ACM, pp. 375-384.

Lincoln, Y. S., and Guba, E. G. (1985). *Naturalistic Inquiry*. Beverly Hills, California et al.: SAGE Publications.

Liu, B. (2012). “Sentiment Analysis and Opinion Mining,“ *Synthesis Lectures on Human Language Technologies* 5 (1), pp. 1-167.

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). “Topic-Link Lda: Joint Models of Topic and Author Community,“ In: *Proceedings of the Annual International Conference on Machine Learning*: ACM, pp. 665-672.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. (2011). “How Social Influence Can Undermine the Wisdom of Crowd Effect,“ *Proceedings of the National Academy of Sciences* 108 (22), pp. 9020-9025.

Loughran, T., and McDonald, B. (2011). “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks,“ *The Journal of Finance* 66 (1), pp. 35-65.

Loughran, T., and McDonald, B. (2016). “Textual Analysis in Accounting and Finance: A Survey,“ *Journal of Accounting Research* 54 (4), pp. 1187-1230.

Lu, B., Ott, M., Cardie, C., and Tsou, B. K. (2011). “Multi-Aspect Sentiment Analysis with Topic Models,“ In: *Proceedings of the International Conference on Data Mining Workshops*: IEEE pp. 81-88.

Luhn, H. P. (1958). “A Business Intelligence System,“ *IBM Journal of Research and Development* 2 (4), pp. 314-319.

MacKinlay, A. C. (1997). “Event Studies in Economics and Finance,“ *Journal of Economic Literature* 35 (1), pp. 13-39.

Malkiel, B. G., and Fama, E. F. (1970). “Efficient Capital Markets: A Review of Theory and Empirical Work,“ *The Journal of Finance* 25 (2), pp. 383-417.

March, S. T., and Smith, G. F. (1995). “Design and Natural Science Research on Information Technology,“ *Decision Support Systems* 15 (4), pp. 251-266.

Matsumoto, D., Pronk, M., and Roelofsen, E. (2011). “What Makes Conference Calls Useful? The Information Content of Managers' Presentations and Analysts' Discussion Sessions,“ *The Accounting Review* 86 (4), pp. 1383-1414.

Mayew, W. J. (2008). “Evidence of Management Discrimination among Analysts During Earnings Conference Calls,“ *Journal of Accounting Research* 46 (3), pp. 627-659.

Mayew, W. J., Sharp, N. Y., and Venkatachalam, M. (2013). “Using Earnings Conference Calls to Identify Analysts with Superior Private Information,“ *Review of Accounting Studies* 18 (2), pp. 386-413.

McCallum, A. K. (2002). “Mallet: A Machine Learning for Language Toolkit,“.

McKnight, D. H., and Chervany, N. L. (2001). “What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology,“ *International Journal of Electronic Commerce* 6 (2), pp. 35-59.

Mei, Q., Shen, X., and Zhai, C. (2007). “Automatic Labeling of Multinomial Topic Models,“ In: *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 490-499.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). “Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity,“ In: *Proceedings of the AAAI Conference*, pp. 775-780.

Miller, G., and Fellbaum, C. (1998). *Wordnet: An Electronic Lexical Database*. Cambridge: MIT Press.

Mingers, J. C. (1995). “Information and Meaning - Foundations for an Intersubjective Account,“ *Information Systems Journal* 5 (4), pp. 285-306.

Mohr, J. W., and Bogdanov, P. (2013). “Introduction - Topic Models: What They Are and Why They Matter,“ *Poetics* 41 (6), pp. 545-569.

Moqri, M., Bandyopadhyay, S., and Cheng, H. K. (2015). “Identifying Research Trends in Is,“ In: *Proceedings of the Americas Conference on Information Systems*, Puerto Rico.

Muhammad, A., Wiratunga, N., Lothian, R., and Glassey, R. (2013). “Domain-Based Lexicon Enhancement for Sentiment Analysis,“ In: *Proceedings of the SMA@ BCS-SGAI*, pp. 7-18.

Müller, O., Schmiedel, T., Gorbacheva, E., and vom Brocke, J. (2016). “Towards a Typology of Business Process Management Professionals: Identifying Patterns of Competences through Latent Semantic Analysis,“ *Enterprise Information Systems* 10 (1), pp. 50-80.

Mützel, S. (2015). “Facing Big Data: Making Sociology Relevant,“ *Big Data & Society* 2 (2).

Newman, D., Baldwin, T., Cavedon, L., Huang, E., Karimi, S., Martinez, D., Scholer, F., and Zobel, J. (2010a). “Visualizing Search Results and Document Collections Using Topic Maps,“ *Web Semantics: Science, Services and Agents on the World Wide Web* 8 (2), pp. 169-175.

Newman, D., Bonilla, E. V., and Buntine, W. (2011). “Improving Topic Coherence with Regularized Topic Models,“ In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 496-504.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010b). “Automatic Evaluation of Topic Coherence,“ In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*: Association for Computational Linguistics, pp. 100-108.

Newman, D., Noh, Y., Talley, E., Karimi, S., and Baldwin, T. (2010c). “Evaluating Topic Models for Digital Libraries,“ In: *Proceedings of the Annual joint Conference on Digital libraries*: ACM, pp. 215-224.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). “The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature,“ *Decision Support Systems* 50 (3), pp. 559-569.

Nguyen, V.-A., Boyd-Graber, J., and Resnik, P. (2012). “Sits: A Hierarchical Nonparametric Model Using Speaker Identity for Topic Segmentation in Multiparty Conversations,“ In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*: Association for Computational Linguistics, pp. 78-87.

Nickerson, R. C., Varshney, U., and Muntermann, J. (2013). "A Method for Taxonomy Development and Its Application in Information Systems," *European Journal of Information Systems* 22 (3), pp. 336-359.

Nikolenko, S. I., Koltcov, S., and Koltsova, O. (2015). "Topic Modelling for Qualitative Studies," *Journal of Information Science* 0165551515617393.

Nofer, M., and Hinz, O. (2014). "Are Crowds on the Internet Wiser Than Experts? The Case of a Stock Prediction Community," *Journal of Business Economics* 84 (3), pp. 303-338.

Nofer, M., and Hinz, O. (2015). "Using Twitter to Predict the Stock Market," *Business & Information Systems Engineering* 57 (4), pp. 229-242.

O'Brien, P. C. (1988). "Analysts' Forecasts as Earnings Expectations," *Journal of Accounting and Economics* 10 (1), pp. 53-83.

O'Connor, B., Bamman, D., and Smith, N. A. (2011). "Computational Text Analysis for Social Science: Model Assumptions and Complexity," in: *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques," In: *Proceedings of the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*: Association for Computational Linguistics, pp. 79-86.

Papadimitriou, C. H., Tamaki, H., Raghavan, P., and Vempala, S. (1998). "Latent Semantic Indexing: A Probabilistic Analysis," In: *Proceedings of the ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*: ACM, pp. 159-168.

Paré, G., Trudel, M.-C., Jaana, M., and Kitsiou, S. (2015). "Synthesizing Information Systems Knowledge: A Typology of Literature Reviews," *Information & Management* 52 (2), pp. 183-199.

Park, H., Jeon, M., and Rosen, J. B. (2001). "Lower Dimensional Representation of Text Data in Vector Space Based Information Retrieval," *Computational information retrieval*, pp. 3-23.

Paul, M. J., and Girju, R. (2009). "Topic Modeling of Research Fields: An Interdisciplinary Perspective," In: *Proceedings of the International Conference on recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 337-342.

Pennington, R., and Tuttle, B. (2007). "The Effects of Information Overload on Software Project Risk Assessment," *Decision Sciences* 38 (3), pp. 489-526.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology* 88 (5), p. 879.

Porac, J. F., and Thomas, H. (1990). "Taxonomic Mental Models in Competitor Definition," *Academy of management Review* 15 (2), pp. 224-240.

Price, S. M., Doran, J. S., Peterson, D. R., and Bliss, B. A. (2012). "Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone," *Journal of Banking & Finance* 36 (4), pp. 992-1011.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). "How to Analyze Political Attention with Minimal Assumptions and Costs," *American Journal of Political Science* 54 (1), pp. 209-228.

Raftery, A. E. (1995). "Bayesian Model Selection in Social Research," *Sociological methodology*, pp. 111-163.

Rai, A. (2016). “Editor's Comments: Synergies between Big Data and Theory,“ *MIS Quarterly* 40 (2), pp. iii-ix.

Ramage, D., Dumais, S. T., and Liebling, D. J. (2010). “Characterizing Microblogs with Topic Models,“ *International Conference on Web and Social Media*, pp. 130-137.

Ramage, D., and Rosen, E. (2011). “Stanford Topic Modeling Toolbox,“.

Ramage, D., Rosen, E., Chuang, J., D. Manning, C., and McFarland, D. A. (2009). “Topic Modeling for the Social Sciences,“ in: *NIPS Workshop on Applications for Topic Models: Text and Beyond*. Whistler, Canada.

Ramirez, E. H., Brena, R., Magatti, D., and Stella, F. (2012). “Topic Model Validation,“ *Neurocomputing* 76 (1), pp. 125-133.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. M. (2015). “Deep Exponential Families,“ In: *Proceedings of the AISTATS*.

Řehůřek, R., and Sojka, P. (2010). “Software Framework for Topic Modelling with Large Corpora,“ In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45-50.

Reuters, T. (2015). “Thomson Reuters I/B/E/S Estimates - Fact Sheet,“.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2014). “Stm: R Package for Structural Topic Models,“ *R package* 1, p. 12.

Rowe, F. (2014). “What Literature Review Is Not: Diversity, Boundaries and Recommendations,“ *European Journal of Information Systems* 23 (3), pp. 241-255.

Sabherwal, R., and King, W. R. (1995). “An Empirical Taxonomy of the Decision-Making Processes Concerning Strategic Applications of Information Systems,“ *Journal of Management Information Systems* 11 (4), pp. 177-214.

Salton, G., Wong, A., and Yang, C.-S. (1975). “A Vector Space Model for Automatic Indexing,“ *Communications of the ACM* 18 (11), pp. 613-620.

SDL. (2017). “Sdl Social Intelligence: Sm2 Social Media Monitoring.“ Retrieved March 16'th, 2017, from http://www.sdl.com/de/download/sm2-social-media-monitoring/82522/

Sharma, A., and Dey, S. (2012). “A Comparative Study of Feature Selection and Machine Learning Techniques for Sentiment Analysis,“ in: *Proceedings of the 2012 ACM Research in Applied Computation Symposium*. San Antonio, Texas: ACM, pp. 1-7.

Sharpe, W. F. (1964). “Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk,“ *The Journal of Finance* 19 (3), pp. 425-442.

Sia, S. K., Soh, C., and Weill, P. (2016). “How Dbs Bank Pursued a Digital Business Strategy,“ *MIS Quarterly Executive* 15 (2), pp. 105-121.

Sidorova, A., Evangelopoulos, N., and Ramakrishnan, T. (2007). “Diversity in Is Research: An Exploratory Study Using Latent Semantics,“ In: *Proceedings of the International Conference on Information Systems*, p. 10.

Sidorova, A., Evangelopoulos, N., Valacich, J. S., and Ramakrishnan, T. (2008). “Uncovering the Intellectual Core of the Information Systems Discipline,“ *MIS Quarterly* 32 (3 September), pp. 467-482.

Sidorova, A., and Isik, O. (2010). “Business Process Research: A Cross-Disciplinary Review,“ *Business Process Management Journal* 16 (4), pp. 566-597.

Sievert, C., and Shirley, K. E. (2014). “Ldavis: A Method for Visualizing and Interpreting Topics,“ In: *Proceedings of the Workshop on interactive language learning, visualization, and interfaces*, pp. 63-70.

Simon, H. A. (1976). *Administrative Behavior*, (3. ed.). New York: The Free Press.

Simon, H. A. (1977). *The New Science of Management Decision*. New York: Prentice Hall.

Simon, H. A. (1996). *The Sciences of the Artificial*, (3 ed.). Cambridge, MA: MIT press.

Skiera, B., Bayer, E., and Schöler, L. (2017). "What Should Be the Dependent Variable in Marketing-Related Event Studies?," *International Journal of Research in Marketing* In Press, Corrected Proof.

Speed, J. G. (1893). "Do Newspapers Now Give the News?," *Forum* 15, pp. 704-711.

Spence, D. P., and Owens, K. C. (1990). "Lexical Co-Occurrence and Association Strength," *Journal of Psycholinguistic Research* 19 (5), pp. 317-330.

Steyvers, M., and Griffiths, T. (2007). "Probabilistic Topic Models," in *Handbook of Latent Semantic Analysis*. pp. 424-440.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilive, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, Massachusetts: The M.I.T. Press.

Sunder, S. V. (2002). "Investor Access to Conference Call Disclosures: Impact of Regulation Fair Disclosure on Information Asymmetry," *SSRN Electronic Journal*.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Random House LLC.

Sylvester, A., Tate, M., and Johnstone, D. (2013). "Beyond Synthesis: Re-Presenting Heterogeneous Research Literature," *Behaviour & Information Technology* 32 (12), pp. 1199-1215.

Tasker, S. C. (1997). "Voluntary Disclosure as a Response to Low Accounting Quality: Evidence from Quarterly Conference Call Usage," in: *Department of Management*. Massachusetts Institute of Technology.

Tasker, S. C. (1998). "Bridging the Information Gap: Quarterly Conference Calls as a Medium for Voluntary Disclosure," *Review of Accounting Studies* 3 (1-2), pp. 137-167.

Teddlie, C., and Tashakkori, A. (2009). *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Sage Publications Inc.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association* 101 (476), pp. 1566-1581.

Templier, M., and Paré, G. (2015). "A Framework for Guiding and Evaluating Literature Reviews," *Communications of the Association for Information Systems* 37 (1), p. 6.

The Guardian. (2017). "The Guardian Open Platform." Retrieved March 16th, 2017, from http://open-platform.theguardian.com/

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267-288.

Tirunillai, S., and Tellis, G. (2014). "Mining Marketing Meaning from Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," *Journal of Marketing Research* 51 (August), pp. 463-479.

Titov, I., and McDonald, R. (2008). "Modeling Online Reviews with Multi-Grain Topic Models," In: *Proceedings of the International conference on World Wide Web*: ACM, pp. 111-120.

Trueman, B. (1994). "Analyst Forecasts and Herding Behavior," *Review of Financial Studies* 7, pp. 97-124.

Trusov, M., Ma, L., and Jamal, Z. (2016). "Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting," *Marketing Science* 35 (3), pp. 405-426.

Turney, P. (2001). "Mining the Web for Synonyms: Pmi-Ir Versus Lsa on Toefl," In: *Proceedings of the European Conference on Machine Learning*, pp. 491–502.

Turney, P. D., and Pantel, P. (2010). "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research* 37 (1), pp. 141-188.

Twedt, B., and Rees, L. (2012). "Reading between the Lines: An Empirical Examination of Qualitative Attributes of Financial Analysts' Reports," *Journal of Accounting and Public Policy* 31, pp. 1-21.

Varshney, U., Nickerson, R., and Muntermann, J. (2015). "Towards the Development of a Taxonomic Theory," In: *Proceedings of the 21st American Conference of Information Systems*, Puerto Rico: AISeL.

Venkatesh, V., Brown, S. A., and Bala, H. (2013). "Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems," *MIS quarterly* 37 (1), pp. 21-54.

Venkatesh, V., Brown, S. A., and Sullivan, Y. W. (2016). "Guidelines for Conducting Mixed-Methods Research: An Extension and Illustration," *Journal of the Association for Information Systems* 17 (7), p. 2.

Visa, A., Toivonen, J., Vanharanta, H., and Back, B. (2002). "Contents Matching Defined by Prototypes: Methodology Verification with Books of the Bible," *Journal of Management Information Systems* 18 (4), pp. 87-100.

W3C. (2013). "The W3c Data Activity: Building the Web of Data.", from https://www.w3.org/2013/data/

Wagner-Pacifici, R., Mohr, J. W., and Breiger, R. L. (2015). "Ontologies, Methodologies, and New Uses of Big Data in the Social and Cultural Sciences," *Big Data & Society* 2 (2), p. 2053951715613810.

Wallach, H. M. (2006). "Topic Modeling: Beyond Bag-of-Words," in: *International conference on Machine Learning*. Pittsburgh, Pennsylvania, USA: ACM, pp. 977-984.

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009a). "Rethinking Lda: Why Priors Matter," In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1973-1981.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009b). "Evaluation Methods for Topic Models," In: *Proceedings of the Annual International Conference on Machine Learning*: ACM, pp. 1105-1112.

Wand, Y., Monarchi, D. E., Parsons, J., and Woo, C. C. (1995). "Theoretical Foundations for Conceptual Modelling in Information Systems Development," *Decision Support Systems* 15 (4), pp. 285-304.

Wang, C., and Blei, D. (2010). "Hierarchical Dirichlet Process (with Split-Merge Operations),". Github.

Wang, C., and Blei, D. M. (2011). "Collaborative Topic Modeling for Recommending Scientific Articles," In: *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 448-456.

Wang, C. J. W. B., David M. (2011). "Online Variational Inference for the Hierarchical Dirichlet Process," In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 752-760.

Wang, D., Zhu, S., Li, T., and Gong, Y. (2009). "Multi-Document Summarization Using Sentence-Based Topic Models," In: *Proceedings of the ACL-IJCNLP*

Conference Short Papers: Association for Computational Linguistics, pp. 297-300.

Wang, X. S., Bendle, N. T., Mai, F., and Cotte, J. (2015). "The Journal of Consumer Research at 40: A Historical Analysis," *Journal of Consumer Research* 42 (1), pp. 5-18.

Webster, J., and Watson, R. T. (2002). "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly* 26 (2), pp. xiii-xxiii.

Wei, C.-P., Chiang, R., and Wu, C.-C. (2006). "Accommodating Individual Preferences in the Categorization of Documents: A Personalized Clustering Approach," *Journal of Management Information Systems* 23 (2), pp. 173-201.

Wei, C.-P., Hu, P. J.-H., Tai, C.-H., Huang, C.-N., and Yang, C.-S. (2007). "Managing Word Mismatch Problems in Information Retrieval: A Topic-Based Query Expansion Approach," *Journal of Management Information Systems* 24 (3), pp. 269-295.

Wei, X., and Croft, W. B. (2006). "Lda-Based Document Models for Ad-Hoc Retrieval," In: *Proceedings of the ACM SIGIR conference on Research and development in information retrieval*: ACM, pp. 178-185.

Winchel, J. (2015). "Investor Reaction to the Ambiguity and Mix of Positive and Negative Argumentation in Favorable Analyst Reports," *Contemporary Accounting Research* 32 (3), pp. 973-999.

Winter, R. (2008). "Design Science Research in Europe," *European Journal of Information Systems* 15 (5), pp. 470-475.

Wolfe, M. B. W., and Goldman, S. R. (2003). "Use of Latent Semantic Analysis for Predicting Psychological Phenomena: Two Issues and Proposed Solutions," *Behavior Research Methods, Instruments, & Computers* 35 (1), pp. 22-31.

Womack, K. L. (1996). "Do Brokerage Analysts' Recommendations Have Investment Value?," *The Journal of Finance* 51 (1), pp. 137-167.

Zhai, K., Boyd-Graber, J., Asadi, N., and Alkhouja, M. L. (2012). "Mr. Lda: A Flexible Large Scale Topic Modeling Package Using Variational Inference in Mapreduce," In: *Proceedings of the International conference on World Wide Web*: ACM, pp. 879-888.

Appendix

Appendix A:
Overview of author contribution ratios for papers included in this thesis

Paper Citation	Publication Status	Outlet Research Type	Author	Contribution
I.1 (Eickhoff et al., 2017)	Under Review	Redacted Taxonomy Development	**Eickhoff** Muntermann Weinrich	**33%** 33% 33%
II.1 (Eickhoff, 2015)	Published	DESRIST 2015 Design Science	**Eickhoff**	**100%**
II.2 (Eickhoff and Neuss, 2017)	Published	ECIS 2017 Literature Review	**Eickhoff** Neuss	**95%** 5%
III.1 (Eickhoff and Muntermann, 2016b)	Published	Information & Management Behavioral Positivist	**Eickhoff** Muntermann	**75%** 25%
III.2 (Eickhoff and Muntermann, 2017)	Under Review	Redacted Behavioral Positivist	**Eickhoff** Muntermann	**80%** 20%
III.3 (Eickhoff and Muntermann, 2016c)	Published	PACIS 2016 Behavioral Positivist	**Eickhoff** Muntermann	**80%** 20%
III.4 (Eickhoff, 2017)	Published	HICSS 2017 Behavioral Positivist	**Eickhoff**	**100%**