Journal of the Association for Information Systems (2021) **22**(1), 156-178
doi: 10.17705/1jais.00657

RESEARCH ARTICLE

ISSN 1536-9323

Design Principles for Robust Fraud Detection: The Case of Stock Market Manipulations

Michael Siering[1], Jan Muntermann[2], Miha Grčar[3]

[1]Goethe University Frankfurt, Germany, siering@wiwi.uni-frankfurt.de
[2]University of Goettingen, Germany, muntermann@wiwi.uni-goettingen.de
[3]Jožef Stefan Institute, Slovenia, miha.grcar@sowalabs.com

Abstract

We address the challenge of building an automated fraud detection system with robust classifiers that mitigate countermeasures from fraudsters in the field of information-based securities fraud. Our work involves developing design principles for robust fraud detection systems and presenting corresponding design features. We adopt an instrumentalist perspective that relies on theory-based linguistic features and ensemble learning concepts as justificatory knowledge for building robust classifiers. We perform a naive evaluation that assesses the classifiers' performance to identify suspicious stock recommendations, and a robustness evaluation with a simulation that demonstrates a response to fraudster countermeasures. The results indicate that the use of theory-based linguistic features and ensemble learning can significantly increase the robustness of classifiers and contribute to the effectiveness of robust fraud detection. We discuss implications for supervisory authorities, industry, and individual users.

Keywords: Fraud Detection, Market Manipulation, Design Principles, Text Mining, Data Mining, Instrumentalism, Ensemble Learning

Sandeep Purao was the accepting senior editor. This research article was submitted on February 22, 2016 and underwent four revisions.

1 Introduction

Fraud detection systems (FDS) have gained importance in both business and societal contexts. For instance, FDS have been used to identify suspicious employee communications (Holton, 2009), fraudulent corporate disclosures (Ravisankar et al. 2011), and unauthorized financial transactions (Chen, Chen, & Lin, 2006). A common problem in the field of fraud detection is that fraudsters constantly adapt their behavior to avoid being detected by contemporary systems (Bolton & Hand, 2002). For instance, consider a text categorization system that uses certain keywords to determine whether a document is suspicious. If the keywords become known, fraudsters will refrain from using them and adapt the content of their messages (Webb, Chitti, & Pu, 2005). However, the robustness of fraud-detection efforts against these types of countermeasures, especially in terms of identifying fraudulent texts, has rarely been addressed thus far. We respond to this theoretical and practical research gap by conducting a multiyear design science research (DSR) project with a multinational project consortium to address the problem of information-based market manipulation.

In this type of market manipulation, fraudsters frequently attempt to manipulate stock prices by disseminating highly positive but false information through fraudulent websites, spam messages, and advertising campaigns on legitimate websites (SEC, 2012b). Fraudsters often follow a "buy low and spam high" strategy: They begin by purchasing a certain stock, then they recommend the stock to internet users

to increase demand for it, thereby raising the stock's price, and, finally, the fraudsters sell their stocks at a profit (Frieder & Zittrain, 2006). These types of so-called "pump and dump" schemes have become a serious problem and a number of spam campaigns have led to significant financial losses (FBI, 2011). Investors duped by such schemes risk losing significant portions of their investments after the spam campaign concludes, when prices typically fall below their original levels (Aggarwal & Wu, 2006). Moreover, the firms that issued the affected stocks suffer significant reputational loss (Hanke & Hauser, 2008). The research consortium that addressed this problem consisted of nine partners, including universities, financial institutions, and IT service providers from the finance and market surveillance domains. In addition, a financial market surveillance authority contributed within an advisory board.

Previous studies have proposed various methods of detecting fraudulent websites or messages (Abbasi et al., 2010; Caruana & Li, 2012). Financial fraud detection is an important field (Ngai et al., 2011), and scholars have repeatedly addressed the problem of identifying securities fraud in general (Fast et al., 2007). Nevertheless, the problem of information-based fraud in its various forms, such as the dissemination of fraudulent stock recommendations, remains underexplored, especially in terms of providing robust classifications. Specifically, a robust classifier is one that will "resist change without adapting its initial stable configuration" (Wieland & Marcus Wallenburg, 2012, p. 890).

To address the problem, this study develops an IT artifact that can act as a robust classifier by providing an assessment of whether a given document is suspected of being fraudulent. The artifact is based on new design principles and exhibits new design features that make these classifications robust against potential fraudster countermeasures. To develop the artifact, we follow the problem-solving design science research (DSR) paradigm (Hevner, March, & Park, 2004; Newell & Simon, 1972) with a constructive and proactive approach (Iivari, 2007; Iivari, 2015). More specifically, we followed the process model of Kuechler & Vaishnavi (2008) to formulate specific design principles and design features (at the mesolevel) to address the identified problem and the specific design requirements in the field of information-based fraud.

From a methodological perspective, our research illustrates how classifiers constructed on the basis of relevant kernel theories can support problem solving. Our work therefore differs significantly from traditional data mining research, which strictly follows the logic of induction, generating new knowledge by applying data mining methods to detect patterns within the existing data. In contrast, we adopt an instrumentalist perspective, which provides the "freedom to play around with different theories and different traditions of scientific knowledge production in a way that rival philosophies of science neglect" (Kilduff, Mehra, & Dunn, 2011, p. 1011). Specifically, we employ theories drawn from marketing and financial economics as kernel theories that inform our artifact construction (Gregor & Hevner, 2013). We demonstrate that our research approach, design principles, and design features are advantageous for problem solving and generate practicable outcomes. We conduct an empirical evaluation of the artifact's validity in the context of stock market manipulations and assess its robustness by simulating a fraudster taking countermeasures against our solution. The remainder of this paper is structured as follows: Section 2 presents the research background, Section 3 focuses on the research methodology applied and our artifact design, Section 4 outlines our artifact evaluation, Section 5 discusses the results, and Section 6 concludes the paper.

2 Research Background

2.1 Fraud Detection in Finance

Data mining techniques have been applied to address diverse types of fraud, especially in the financial context. Ngai et al. (2011) provide an overview of this field and the major categories of financial fraud: *bank fraud*, *insurance fraud*, *securities and commodities fraud*, and *other finance-related fraud*.

Regarding *bank fraud*, the extant research has focused primarily on credit card fraud (Chen et al., 2006), although *insurance fraud* and *other finance-related fraud* have been explored in diverse contexts, such as automotive insurance fraud (Caudill, Ayuso, & Guillén, 2005) and financial statement fraud (Glancy & Yadav, 2011; Ravisankar et al., 2011). By contrast, few studies have examined the process of detecting manipulations of securities and commodities markets (Ngai et al., 2011). Regarding *securities fraud*, three types of stock market manipulation schemes have been described in the literature: *information-based*, *trade-based*, and *action-based* manipulations (Allen & Gale, 1992). These schemes seek to manipulate stock prices through the release and spread of false information (information-based manipulation), the buying or selling of a stock (trade-based manipulation), or the execution of certain management activities (action-based manipulation). Scholars have extensively studied *trade-based* manipulation (Felixson & Pelli, 1999). The restrictions imposed upon managers who trade their own firms' stock have led to *action-based* manipulation becoming rare (Öğüt et al., 2009). *Information-based* manipulation has gained increasing attention in recent years because the internet has facilitated the spread of fraudulent stock recommendations to large audiences. The manipulators typically attempt to profit by purchasing

a stock at a low price, recommending it to other investors, and then selling the stock at a higher price (Siering et al., 2017). Research has demonstrated that trading volumes increase if stocks are advertised through fraudulent recommendations (Böhme & Holz, 2006). Furthermore, several studies have revealed that these fraudulent recommendations can generate increases in stock prices during the manipulation period. However, when no further recommendation messages are published, the prices of the manipulated stocks decrease rapidly to below their original levels (Aggarwal & Wu, 2006; Böhme & Holz, 2006; Hanke & Hauser, 2008). Even though the United States Securities and Exchange Commission (SEC) has taken countermeasures against these forms of manipulation (i.e., by releasing warnings, suspending trading, and prosecuting manipulators), manipulation campaigns can still be effective (Siering, 2019).

In general, the detection of stock market manipulation remains underexplored (Ngai et al., 2011). While the general characteristics of such manipulation schemes and potential system designs have been taken into account (Gregory & Muntermann, 2014; Siering et al., 2017), the use of unstructured data sources such as financial news or investment newsletters does not appear to have been analyzed. This is a critical gap because this type of textual data is a frequent source of malicious and misleading information in the context of information-based manipulations. Furthermore, the potential countermeasures that fraudsters may use to circumvent fraud-detection mechanisms also remain underexplored.

2.2 Theoretical Perspectives on the Robustness of Fraud Detection

2.2.1 Related Work from Machine Learning

Fraud-detection systems must satisfy the general requirement of being able to achieve good classification performance. However, the development of robust fraud-detection classifiers is a challenging task: If fraudsters are aware that their activities may be detected, they might implement appropriate countermeasures to evade the fraud-detection systems. A robust classifier is one that will "resist change without adapting its initial stable configuration" (Wieland & Marcus Wallenburg, 2012, p. 890). This consideration significantly complicates the classification task for these systems, making their challenge "quite different from traditional classification problems, as intelligent, malicious, and adaptive adversaries can manipulate their samples to mislead a classifier or a learning algorithm" (Biggio et al., 2011, p. 350). Different approaches have been explored to increase the robustness of classifiers against the countermeasures of potential attackers. Several studies suggest adaptations of classifiers during feature processing (Kolcz & Teo, 2009), but the potential use of linguistic features as textual representations has rarely been investigated.

Linguistic features are derived from an original feature set, such as a "bag of words" from a document (Djeraba, 2002). Such features have been successfully applied for author identification (Zheng et al., 2006) and speaker recognition tasks (Campbell et al., 2007) but only to increase classification performance, not to increase classifier robustness. Furthermore, the selection of linguistic features has typically been ad hoc, rather than based on theoretical insights drawn from kernel theories serving as "justificatory knowledge" (Gregor & Jones, 2007) to improve classification robustness.

A different category of studies seeks to increase the robustness of classifications by training collections of different classifiers and implementing various rules such as majority voting or classification averages to combine classification results (Biggio, Fumera, & Roli, 2010; Perols, Chari, & Agrawal, 2009). Although this research stream provides guidance for the development and combination of multiple classifiers that use the same input data, no study has yet attempted to construct classifiers guided by relevant kernel theory to achieve better robustness against potential countermeasures.

2.2.2 Related Work from Financial Economics and Marketing Research

In the following, we focus on related work from the field of financial economics and marketing research to explain the aspects that make stock recommendations effective. We incorporated this work into the development of our design features. In financial economics, it is assumed that information processing is the basis of investment decisions (Fama, 1970). Behavioral finance theory states that investment decisions can also be driven by irrational factors such as information presentation, including the sentiment expressed within a stock recommendation (de Bondt, 1998). Persuasive communication is also typically the focus of marketing research: Stock recommendations represent a form of advertising that is sent to internet users to influence their information processing and ultimately promote desired behavior—specifically, the purchase of a specific stock (Vakratsas & Ambler, 1999).

Marketing research has recognized the important role of advertisements' *information content* (Abernethy & Franke, 1996). Advertisements are often used by consumers to acquire product-related information, which is then incorporated into purchase decisions (Nelson, 1970). Moreover, if advertisements disregard customers' search for relevant product information, the advertisers' "non-informative advertising policy may

self-destruct" (Resnik & Stern, 1977, p. 53). In addition, in the financial context, the price-determination process for various instruments such as stocks is driven primarily by the information available to market participants (Fama, 1970). Therefore, the information content of advertisements is particularly important and should be considered by advertisers who promote financial products (Jones & Smythe, 2003).

Text *readability* encompasses the question of how easily a text can be read and the educational level required to understand its content (Bailin & Grafstein, 2001; Korfiatis, García-Bariocanal, & Sánchez-Alonso, 2012). It has been shown that readability is a prerequisite for advertising efficacy (Abruzzini, 1967). Thus, advertisers seek to increase the productive attention devoted to their advertisements by ensuring that they are easy to read (Clark, Kaminski, & Brown, 1990). The effect of text readability on investors' reactions has also been investigated in the financial context. In particular, the readability of corporate disclosures has been found to influence trading behavior, with investors demonstrating delayed reactions to corporate disclosures that are difficult to read (You & Zhang, 2009), and improved disclosure readability significantly affects small investor trading (Loughran & McDonald, 2010).

The important role of *sentiment* within advertisements and the effects of these emotions on consumers' moods and reactions have been the subject of various studies. Emotional advertising appears to increase consumers' attention to a product and bolster consumers' memories of product-related features (Chandy et al., 2001), and product-related emotional communications can intensify consumers' attitudes (Sonnier, McAlister, & Rutz, 2011). These arguments are supported in the financial context by behavioral finance theory. In particular, it is assumed that investors are influenced by the tone of discussions that involve certain financial instruments, and it has been shown that investors are influenced by sentiments expressed in newspapers, message boards, and even Twitter messages (Bollen & Huina, 2011; Das & Chen, 2007).

3 Research Methodology and Artifact Design

3.1 Design Science Research

We adopt the DSR paradigm, which is generally related to the development of IT artifacts (Hevner et al., 2004; March & Smith, 1995; Peffers et al., 2007). A key characteristic of this research paradigm is that DSR researchers search for satisficing (though not necessarily the best) problem solutions that meet the formulated problem requirements (Simon, 1996). Because DSR is focused on problem solving, problem analysis and appropriate domain knowledge are especially important for developing suitable problem solutions (Peffers et al., 2007). In this case, both gained insights and justificatory knowledge become integral parts of the developed problem solution (Simon, 1996).

The role of theory in DSR is twofold (Kuechler & Vaishnavi, 2008). First, so-called "kernel theories," which often originate from non-IS disciplines, may inform the search for a satisficing problem solution. We consider the work introduced in the previous section to be such kernel theory. Second, DSR seeks to make theoretical contributions by providing explicit prescriptions for "how to do something/solve a problem." Such prescriptive guidance is provided by design principles that represent "core principles and concepts to guide design" (Vaishnavi & Kuechler, 2015, p. 20), which can be applied for "use in the design and implementation of the IS product" (Hevner & Chatterjee, 2010, p. 49). We develop and present such design principles, which are mapped to design features at the instantiated level. Our design principles provide "a clear statement of truth that guides or constrains action" for the development of robust fraud-detection systems (Hevner & Chatterjee, 2010, p. 66) and can thus be considered to be essential design principles (Gregor, Müller, & Seidel, 2013). By offering a more effective solution to a well-known class of problem (fraud detection), our study belongs to improvement research: Here, new and better solutions are developed for known problems (Gregor & Hevner, 2013).

3.2 Research Process

Our DSR project follows the process model of Kuechler and Vaishnavi (2008), which provided guidance during our research process (see Figure 1). In the first step (awareness of the problem), the goal is to develop an understanding of the problem faced by stakeholders. After collecting, structuring, and condensing this information, the problem description and design requirements are formulated (see Section 3.3). These may be revised during the problem-solving process. The design requirements are addressed in the following step (suggestion), in which the initial ideas (tentative designs) for solving the problem are produced. New ideas may be brought forward deductively on the basis of a relevant kernel theory or abductively from other sources (e.g., similar cases; Kuechler & Vaishnavi, 2012) and are condensed in the form of design principles (see Section 3.4). However, while our approach to problem solving is inductive and data-driven, our logic of action is also characterized by truth-independent problem solving. Here, we consider theories to be "useful instruments in helping predict events and solve problems" (Kilduff et al., 2011, p. 302).

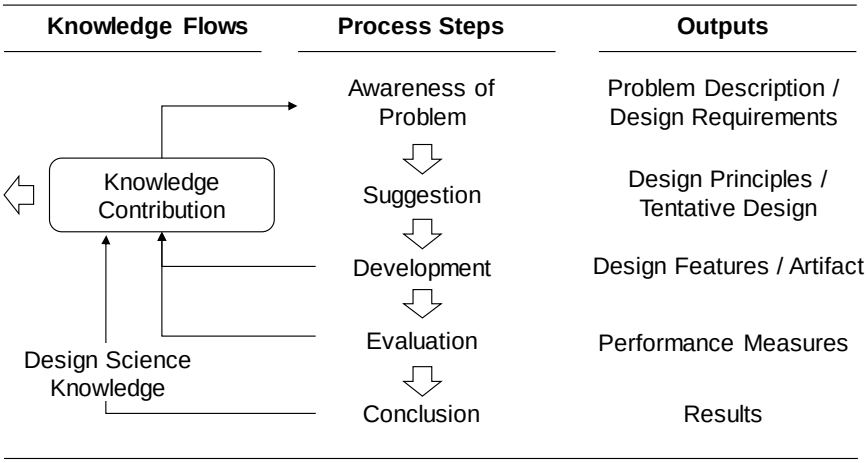

Figure 1. Employed Design Science Research Process Model Based on Kuechler & Vaishnavi (2008)

In the third step (development), the design principles are mapped to design features—the specific artifact capabilities that result from (for example) a chosen algorithm (Meth, Mueller, & Maedche, 2015). We present these design features in Section 3.5 in terms of an instantiated IT artifact—algorithm implementations that are evaluated in step four. Here, suitable measures are used to assess the performance of the IT artifact. The results may provide support for the previously coded design knowledge, as illustrated in Section 4 or, alternatively, may necessitate alterations during the previously taken steps. When the evaluation results provide support for the successful design of a satisficing problem solution, the codified design knowledge is finalized and presented in the context of future research in the final step (conclusion). The knowledge contribution is thereby made. In the following sections, we outline the steps taken to develop robust FDS.

3.3 Problem Description and Design Requirements

The phenomenon of information-based market manipulations (i.e., the spread of false information to affect stock prices) has existed for many years. As seen in the historical cases reported by the SEC (1959), information-based market manipulation used to be the exclusive preserve of privileged market participants such as broker-dealers, who capitalized on the fact that investors attentively listened to them. Today, the group of manipulators has grown and the way in which they use technology has changed significantly. Now, almost anyone can use the Internet to spread rumors throughout the world at nearly no cost. Thus, the problem of information-based market manipulation has become more urgent, while its detection and prevention have become more difficult (SEC, 2012b).

In our DSR project, this problem was explained by the participating domain experts. Our group of experts consisted of representatives of a market supervisory authority and an IT company that develops software for capital market surveillance. They reported that it is imperative to process the large and ever-growing universe of web documents to obtain knowledge of this type of market manipulation. Based on these insights, we derived design requirement DR1.

DR1: Process a large volume of unstructured data. To detect information-based securities fraud, FDS should support the processing of large collections of documents published on the internet.

Further interviews with domain experts showed that being able to easily access large collections of documents is not sufficient. Manually processing and assessing documents is not adequate because of the large number of documents available. Consequently, an automated assessment of documents is required. However, full automation in the field of market manipulation detection is not feasible. As a domain expert explained during an interview, it is ultimately up to the courts to decide whether to find a market participant guilty of market manipulation. Instead, FDS should direct its attention to cases in which documents are found to be suspicious and require further manual analysis. Against this background, design requirement DR2 was derived.

DR2: Provide automated identification of suspicious documents. The FDS should direct its attention to cases that merit further manual detailed exploration and provide an automated classification of documents (suspicious versus non-suspicious).

After the first steps within the research process (Section 3.2) were taken, we presented an initial tentative design (see sections below) to domain experts. While the initial reaction to design requirements DR1 and DR2 was positive, the domain experts sensed a problem with the suggested artifact that had not been clearly articulated.

Based on their experiences with other types of market manipulation, the experts intuited that market manipulators will adjust their behavior after becoming aware that corresponding FDS have been developed. Consequently, an FDS must provide reliable document classifications to address manipulators' adjustments of their writing style to prevent documents from being classified as suspicious. This feedback led us to derive a third and final design requirement, DR3.

DR3: Limit system vulnerability to fraudster countermeasures. The FDS should, without reconfiguration, provide reliable classifications of documents when the manipulator adjusts the writing style to mislead the system.

3.4 Design Principles of Robust Fraud Detection Systems

To address these design requirements, we developed several design principles that guided our artifact development. Following the requirements whereby a large volume of unstructured data must be processed (DR1) and document classifications should be conducted automatically (DR2), a related knowledge discovery process (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) must extract patterns from existing documents. The most important aspect of this process is the development of a proper problem understanding. Based on that problem understanding, an appropriate feature set can be derived for data mining purposes. Therefore, it is essential to understand which features are well-suited for identifying suspicious documents.

Regarding "pump and dump" campaigns, we recognize that fraudsters will try to convince customers to buy a specific stock. Thus, we assume that the campaigns are formulated in a way that maximizes fraud effectiveness. Consequently, we infer that theories from financial economics and marketing seeking to explain information processing in financial markets and purchase decision-making behavior might be useful in the identification of relevant document characteristics. We therefore formulated our first design principle to focus on these kernel theories during the knowledge-discovery process.

DP1: Theory-guided knowledge discovery process: The FDS development process should be informed by kernel theories explaining fraud effectiveness.

Additionally, following DR1 and DR2, we inferred that the automated processing of stock recommendations and of classifying documents as either suspicious or non-suspicious is required. This finding is in line with earlier FDS from other domains, which has largely relied on automated solutions for data processing and automated classifications of cases via machine learning technologies (Ngai et al., 2011). Thus, following these design requirements as well as the literature stream outlined in the research background section, we formulated the second design principle, DP2.

DP2: Automation of document processing and classification: FDS should provide automated document processing and classification (suspicious vs. non-suspicious).

Finally, to fulfill the design requirement of limiting system vulnerability to fraudsters' countermeasures (DR3), we inferred that these countermeasures must be anticipated if the system is to be made more robust against them. This inference is particularly important because fraudsters have been shown to manipulate their deceptive content to mislead existing FDS (Biggio et al., 2011). This phenomenon has been observed in the field of spam detection, especially with regard to textual content (Goodman et al., 2007). Awareness of such potential countermeasures should thus help increase FDS robustness—specifically, the degree to which the classification process functions correctly in the presence of stressful environmental conditions (IEEE, 1990). Consequently, we formulated the third design principle.

DP3: Anticipation of fraudsters' countermeasures: FDS should provide reliable document classifications even when adapted documents prevent correct FDS classifications.

3.5 Artifact Design Features

Based on our design principles, we developed the design features that guide our artifact development to realize a robust FDS classifier. The resulting classifier can be integrated within an FDS as the core component to provide such classifications. The design features thus resemble the specific artifact characteristics that are necessary to satisfy the design principles (Meth et al., 2015). The specific mapping between design principles and features is shown in Figure 2. We present two design features that are related to document transformation (DF1a, DF1b), one design feature used for automated document classification (DF2), and two design features used to increase classifier robustness (DF3a, DF3b). In the case of document transformation, we first focus on the classic "bag-of-words" model and then emphasize the theoretically derived linguistic features. The theory-guided knowledge discovery process plays a central role in determining the design features, as the linguistic features are used for document transformation, classification, and increased classifier robustness. The specific design features and their relationships to the design principles are outlined in the following sections.

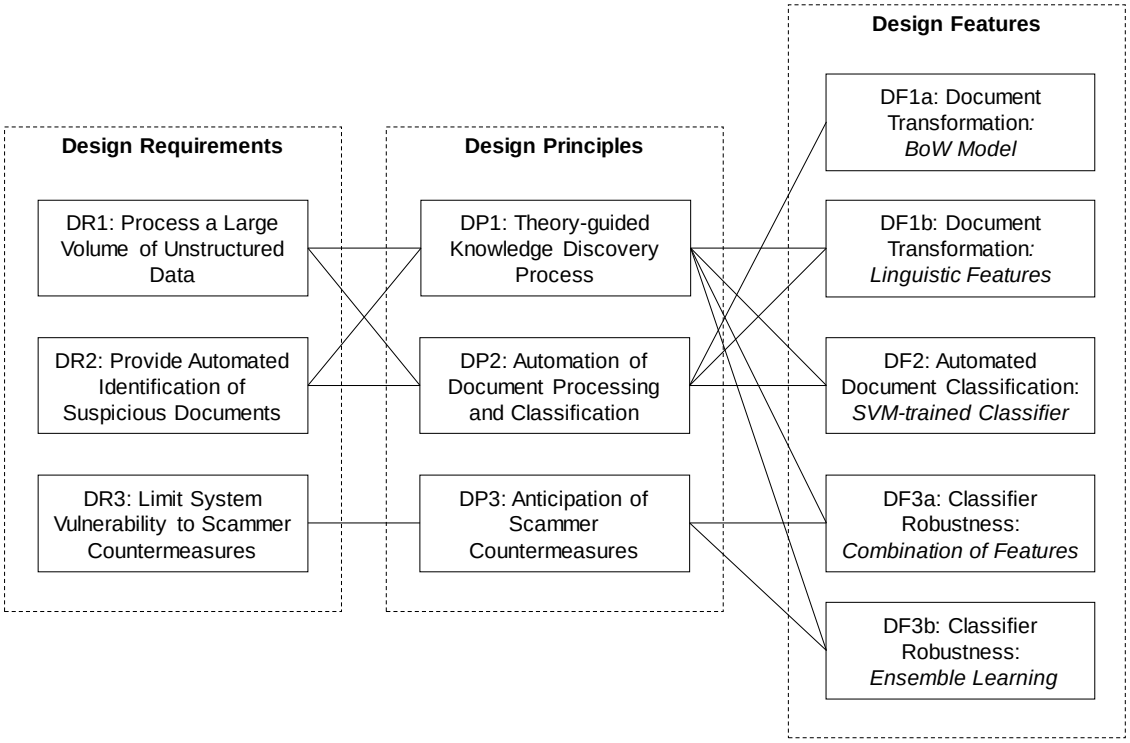

Figure 2. Mapping of Design Requirements, Principles, and Features

3.5.1 Design Feature DF1a: Document Transformation with a Bag-of-Words Model

We implemented document transformation via a bag-of-words model as a basic design feature (DF1a) to address the design principle of the automation of document processing and classification (DP2; Russell, Norvig, & Davis, 2010). Because classical machine learning techniques cannot assess plain text, we first used several pre-processing steps for the text of the examined recommendations (Apté, Damerau, & Weiss, 1994; Wei & Dong, 2001). We decomposed each document into its individual words, regarding each word as a feature (i.e., a bag-of-words model; Russell et al., 2010). To increase computational efficiency and classification performance, we reduced the number of features by removing stop words and applied minimum and maximum thresholds for the number of documents in which each feature should occur (Groth, Siering, & Gomber, 2014). We also applied a stemmer (Porter, 1980). To avoid overly optimistic classification results, we filtered out stock symbols, firm names, publisher names, and disclaimers that are contained only in suspicious stock recommendations. The remaining features were used to construct a document-feature matrix for the training and evaluation of the models. The term frequency-inverse document frequency (TF-IDF) measure was used to calculate the corresponding weights (Hotho, Nürnberger, & Paaß, 2005).

3.5.2 Design Feature DF1b: Document Transformation with Linguistic Features

We implemented another design feature (DF1b), document transformation with linguistic features to address the design principles of a theory-guided knowledge discovery process (DP1) and to enable automated document processing (DP2). In line with our instrumentalist research perspective, we sought to discover the theory "that has the highest likelihood of solving [our] particular problem" (Kilduff et al., 2011, p. 303). Theoretical foundations from financial economics and marketing serve as justificatory knowledge for our artifact design, guiding us to take into account information content, readability, and sentiment as linguistic features.

Information content. To increase the advertising effect of their stock recommendations, fraudsters need to provide a significant amount of relevant information about that stock. Thus, we determined that the document information content in the context of DF1b had the capacity to facilitate the identification of suspicious stock recommendations. We measured information content by relying on the "entropy measure" (Shannon, 1951). Entropy is a widely used measure of information content and can also be applied to measure the information content and redundancy of text samples (Shannon, 1951). In this study, we used an adaptation of Shannon entropy (Shannon, 1948),

which is provided by Equation (1) below. This metric is also extensively used in the field of machine learning (Han & Kamber, 2006):

$$Entropy = -\sum_{i=1}^{n} p_i \log(p_i) \quad (1)$$

In the above calculation of entropy, *n* denotes the words contained in a document, and p_i represents the probability that specific word *i* will occur. Here, high entropy values symbolize high information content (Martin & Rey, 2000; Teahan, 2000).

Readability. Fraudsters seek to increase the demand for a stock; therefore, because readability increases advertising efficacy and investors' reactions, suspicious stock recommendations should be easy to understand. Consequently, we used document readability as another linguistic feature to identify suspicious stock recommendations. We measured text readability by calculating the automated readability index (*ARI*), the Flesch Reading Ease Score (*Flesch*), and the Fog Index (*Fog*), which are provided by Equations (2), (3), and (4), respectively (Hu, Bose, Koh, & Liu, 2012; Loughran & McDonald, 2010; Smith & Senter, 1967). The *ARI*, as calculated by Equation (2) 2 below, has been used in the context of manipulation detection (Hu et al., 2012):

$$ARI = 0.5\frac{words}{sentences} + 4.71\frac{strokes}{words} - 21.43 \quad (2)$$

$$Fog = 0.4\left(\frac{words}{sentences} + 100\frac{complex\ words}{words}\right) \quad (3)$$

$$Flesch = 206.835 - 1.015\frac{words}{sentences} - 84.6\frac{syllables}{words} \quad (4)$$

In the above equations, *words*, *sentences*, *syllables*, and *strokes* represent the total number of words, sentences, syllables, and strokes in the text, respectively. *Complex words* indicates the total number of words consisting of three or more syllables. Both *ARI* and *Fog* are intended to represent the grade level required to understand a text; thus, lower scores for these metrics indicate that a document is easier to read. By contrast, low *Flesch* scores indicate documents that are difficult to read (Loughran & McDonald, 2010).

Sentiment. It can be assumed that suspicious stock recommendations will have a very positive tone because fraudsters seek to increase the demand for and the stock price of the targeted stock. By contrast, stock recommendations published by professional journalists are not aimed simply at convincing readers to purchase particular stocks but should instead aim to provide an unbiased analysis. Against this background, we propose document sentiment as an appropriate linguistic feature for identifying suspicious documents.

We examined the sentiments expressed in stock recommendations using an unsupervised, dictionary-based approach (Zhou & Chaovalit, 2008). We used the Harvard-IV-4 dictionary, which is commonly used in studies related to the current investigation (Hu et al., 2012; Tetlock, 2007; Tetlock, Saar-Tsechansky, & Macskassy, 2008). We counted the occurrences of positive and negative words using the categories defined by this dictionary, and also considered negations (Loughran & McDonald, 2011).

Next, we adapted several document-level sentiment metrics, as presented in Equations (5), (6), and (7) below (Hu et al., 2012; Tetlock et al., 2008; Zhang & Skiena, 2010):

$$Polarity = \frac{pos - neg}{pos + neg} \quad (5)$$

$$Positivity = \frac{pos}{n} \quad (6)$$

$$Negativity = \frac{neg}{n} \quad (7)$$

These metrics consider *pos*, which represents the number of positive words, and *neg*, which represents the number of negative words, both calculated as described above. In addition, *n* is defined as the total number of words. If a document contains neither positive nor negative words, the value of the above metrics is defined as zero. A positive *polarity* value indicates the predominance of positive words in a document; similarly, a negative value indicates the predominance of negative words. We also calculated the proportion of positive and negative words (relative to total words) in each document (*positivity* and *negativity*, respectively).

3.5.3 Design Feature DF2: Automated Document Classification with an SVM-based Classifier

As a further design feature that addresses the design principle of the theory-guided knowledge discovery process (DP1) and the design principles of the automation of document processing and transformation (DP2), we applied automated document classification using SVM-based classifiers that identify suspicious stock recommendations (DF2). We thus followed a supervised learning setup whereby we used suspicious and non-suspicious stock

recommendations to train and evaluate several classifiers that should then be able to classify new recommendations. For this training, we used a Support Vector Machine (SVM) because it has been proven useful for analyzing both structured and unstructured data (Joachims, 1998; Kim, 2003; Tay & Cao, 2001). Based on this design feature, we built two fundamental classifiers.

Classifier A is based on a bag-of-words model and thus builds upon design feature DF1a. For this classifier, the text is pre-processed, and the words are used as features to represent the text. This approach reflects a classical text categorization task. *Classifier B* utilizes linguistic features to represent the stock recommendations and thus builds upon design feature DF1b. For this classifier, the different measures for information content, readability, and sentiment are used as input variables to determine whether a document is suspected to be a fraudulent stock recommendation.

3.5.4 Design Feature DF3a: Classifier Robustness with Combined Feature Sets

To implement design principle DP3—to consider the fraudster's countermeasures and to develop a classifier that is robust to these countermeasures—we implemented design feature DF3a by increasing classifier robustness with combined feature sets. We assumed that combining the bag-of-words model and linguistic features would improve classifier robustness, as avoiding being detected by classifiers that rely on two feature sets can be assumed to be more difficult than taking countermeasures against one feature set. Consequently, we addressed the theory-guided knowledge discovery process by focusing on the related feature set (DP1). Thus, we trained *Classifier C*, which builds upon both feature sets. This classifier incorporates the linguistic features for information content, readability, and sentiment and the features of the bag-of-words model.

3.5.5 Design Feature DF3b: Classifier Robustness Based on Ensemble Learning

In addition to directly combining the feature sets in a single classifier, ensemble learning can also increase the robustness of a fraud-detection approach (Dietterich, 1997). Therefore, we also implemented design feature DF3b by training and combining several classifiers to increase the robustness of the resulting classifier by anticipating the fraudsters' countermeasures (DP3). Given our focus on building robust classifiers, we constructed two additional classifiers based on an ensemble learning approach. To do this, we combined the outputs of Classifiers A and B and thus also considered the feature set resulting from the theory-guided knowledge discovery process (DP1). As a simple approach, Classifier D combines the outputs of Classifiers A and B as follows:

$$\mathrm{D}(\mathbf{x}) = \begin{cases} \text{suspicious}, & \mathrm{A}(\mathbf{x}) > 0 \vee \mathrm{B}(\mathbf{x}) > 0 \\ \text{non-suspicious}, & \text{otherwise} \end{cases} \quad (8)$$

Thus, document x is classified as suspicious by Classifier D if either Classifier A or Classifier B evaluates it as being suspicious (> 0); this technique represents the basic multiple-classifier approach proposed by Jorgensen, Zhou, and Inge (2008).

Finally, we constructed *Classifier E*, which addresses the concern that a fraudster may adopt countermeasures that involve adjusting the message content. Classifier E combines the outputs of Classifiers A and B in a more complex manner. Because of the nature of the SVM classification, the vector space underlying a classifier is separated into two half-spaces by a hyperplane. Consequently, a document can lie on either the "suspicious" or "non-suspicious" side of the hyperplane. A hyperplane can be formally described as $\mathbf{w} \cdot \mathbf{x_0} + b = 0$, where $\mathbf{x_0}$ is a point lying on the hyperplane, $\mathbf{w}$ is the weight vector (normal to the hyperplane), and b denotes the hyperplane bias (offset from the origin of the vector space). The parameters $\mathbf{w}$ and b are both determined by the SVM training algorithm in an attempt to separate the positive training examples (i.e., suspicious documents) from the negative ones (i.e., non-suspicious documents) by the widest possible margin with respect to the SVM optimization function.

Let us examine Classifier A more closely to explain the concept of document manipulation. In the case of Classifier A, each document is represented as a high-dimensional vector of TF-IDF weights, with each weight corresponding to one feature in the document. Given document vector $\mathbf{x}$, Classifier A performs the following assessment to determine whether the document is suspected of being fraudulent:

$$\mathrm{A}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

In this formulation, n is the size of the vocabulary (i.e., the number of different features in the document collection), x_i is the TF-IDF weight of the i-th feature (it is 0 if that particular feature is not present in the document), and w_i is the SVM weight that corresponds to the i-th feature. If $\mathrm{A}(\mathbf{x})$ is positive, the document lies on the positive side of the hyperplane and is considered to be suspicious; if it is negative, the document lies on the negative side of the hyperplane and is considered non-suspicious.

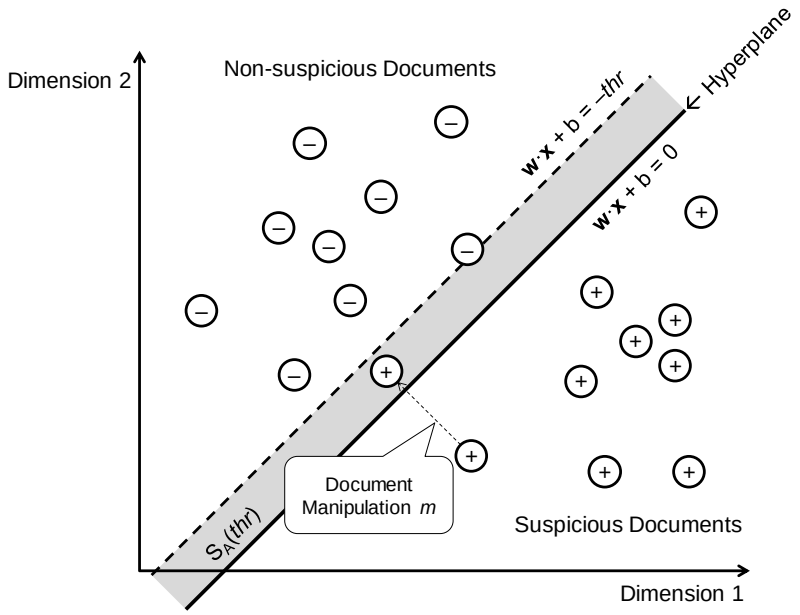

Figure 3. Pushing Documents from One Side (Suspicious) of the Hyperplane to the Other (Non-Suspicious)

To present a suspicious document as non-suspicious, the fraudster needs to replace words that indicate fraud with words that indicate trustworthiness (according to Classifier A). Technically, this requires replacing feature i with feature j so that $w_i x_i > w_j x_j$, which decreases the overall value of $\mathrm{A}(\mathbf{x})$. By performing such swaps, the fraudster "pushes" the document from the positive side of the hyperplane toward the negative side. Pushing a suspicious document far into non-suspicious territory is not feasible, as doing so requires a high degree of manipulation. An altered document is thus most likely to lie relatively close to the hyperplane on the non-suspicious (i.e., negative) side. In fact, such a document is expected to be found in subspace S_A, which is parameterized by $thr \geq 0$ and defined by $S_A(thr) = \{\mathbf{x}; -thr < \mathrm{x}\cdot\mathrm{w} + b \leq 0\} = \{\mathbf{x}; -thr < \mathrm{A}(\mathbf{x}) \leq 0\}$. The basis for a robust classifier follows the intuition that, even for a relatively small value of the threshold *thr*, the manipulated documents are pushed from the suspicious space into $S_A(thr)$. Figure 3 illustrates the described approach in a two-dimensional space.

This reasoning implies that the documents that fall into $S_A(thr)$ may have been altered. Therefore, for Classifier E, if a document falls outside of $S_A(thr)$, the output of Classifier A is accepted. However, if the document falls into $S_A(thr)$, Classifier B is instead employed to categorize the document. Changing single words in a document (i.e., the bag-of-words document representation) is straightforward, whereas changing linguistic features requires more effort and is not typically desirable for fraudsters because they want their recommendations to retain their advertising effects; thus, Classifier B is considered to be a superior approach for assessing potentially altered documents. This new design feature of robust classifiers, which is represented by Classifier E_{thr}, is defined as follows:

$$\mathrm{E}_{thr}(\mathbf{x}) = \begin{cases} \text{suspicious}, & \mathrm{A}(\mathbf{x}) > 0 \vee \\ & (\mathbf{x} \in \mathrm{S_A}(thr) \wedge \mathrm{B}(\mathbf{x}) > 0) \\ \text{non-suspicious}, & \text{otherwise} \end{cases} \quad (9)$$

This definition can be restated as follows:

$$\mathrm{E}_{thr}(\mathbf{x}) = \begin{cases} \text{suspicious}, & \mathrm{A}(\mathbf{x}) > 0 \vee \\ & (|\mathrm{A}(\mathbf{x})| < thr \wedge \mathrm{B}(\mathbf{x}) > 0) \\ \text{non-suspicious}, & \text{otherwise} \end{cases} \quad (10)$$

The following conclusions hold for extreme conditions, when the boundary of S_A lies on the hyperplane and S_A thus effectively does not exist (E_0), and when S_A occupies the entire negative half-space (E_∞):

$$\mathrm{E}_\infty(\mathbf{x}) = \mathrm{D}(\mathbf{x}) = \begin{cases} \text{suspicious}, & \mathrm{A}(\mathbf{x}) > 0 \vee \mathrm{B}(\mathbf{x}) > 0 \\ \text{non-suspicious}, & \text{otherwise} \end{cases} \quad (11)$$

$$\mathrm{E}_0(\mathbf{x}) = \begin{cases} \text{suspicious}, & \mathrm{A}(\mathbf{x}) > 0 \\ \text{non-suspicious}, & \text{otherwise} \end{cases} \quad (12)$$

4 Evaluation

The important considerations for conducting evaluations of IT artifacts in DSR include choice of evaluation criteria (the "what") and evaluation method (the "how"; Prat, Comyn-Wattiau, & Akoka, 2015). Our selections are guided by the design requirements. As evaluation criteria, we selected "validity," which suggests "that the artifact works correctly, i.e., correctly achieves its goal" (Prat et al., 2015, p. 265). This is, referring to our design requirements, closely related to robustness, "the ability of the artifact to handle invalid inputs or stressful environmental conditions" (Prat et al., 2015, p. 266). To gain insights into how well the classifier performed, we first examined how the classifier identified suspicious documents in normal circumstances (see section 4.3). Next, to understand robustness, we evaluated how the classifier performed in the presence of countermeasures (see Section 4.4). The evaluation followed 10-fold cross-validation and a simulation-based evaluation setup that modeled the fraudsters' behavior on the basis of inputs by the domain experts.

In the following, we outline our evaluation hypotheses concentrating on the question of which classifiers are most suitable to address the design requirements. Thereafter, we outline the acquisition of the corpus of documents used to train and evaluate the classifiers. Finally, we outline our evaluation approach and the corresponding results.

4.1 Hypotheses

The ability to manage a large volume of unstructured data (DR1) and to support the automated identification of suspicious documents (DR2) are basic characteristics of all classifiers. Thus, we concentrate on the question of which implementation of our design features performs best in the provision of robust classifications (DR3) when formulating our evaluation hypotheses.

Fraudsters seek to evade classifiers by avoiding terms that identify suspicious contents and/or replacing such terms with words that are typically contained in non-suspicious messages (Biggio et al., 2010; Jorgensen et al., 2008). In the following, we define this behavior as an "attack" on the functioning of the classifier. If the classifiers subjected to these countermeasures are not retrained, the classification performance of the attacked classifiers will decrease significantly (Webb et al., 2005). However, in a scenario that involves stock recommendations intended to convince readers to buy the advertised stock, we assume that fraudsters seek to maintain their advertising efficiency. Thus, fraudsters have a vested interest in retaining the message features that influence advertising efficiency. Based on this reasoning, we formulate the following hypothesis for classifiers that provide automated document classifications (DF2), following DF1b and taking into account linguistic features relating to advertising efficiency (in contrast to classifiers that solely follow DF1a and thus rely solely on a bag-of-words model).

H1: When under attack, a classifier based on linguistic features outperforms a classifier based solely on a bag-of-words model.

In addition to taking linguistic features into account, classification performance can be increased by combining feature sets (DF3a) or by applying ensemble learning (DF3b). In the case of a combination of feature sets, it can be assumed that it is more difficult to manipulate classifiers that consider both bag-of-words and linguistic features than classifiers that consider bag-of-words or linguistic features alone.

In the case of ensemble learning, the individual decisions of different classifiers are combined to classify new examples (Dietterich, 1997). Ensembles can be more accurate if individual classifiers disagree (Dietterich, 1997; Hansen & Salamon, 1990) because "multiple learner systems try to exploit the local different behavior of the base learners to enhance the accuracy and the reliability of the overall inductive learning systems" (Valentini & Masulli, 2002, p. 4). Given the background of these general advantages in the case of different classification tasks, we hypothesize that these characteristics will continue to be advantageous if such classifiers, based on DF3a or DF3b, are attacked:

H2a: When under attack, a classifier that combines linguistic features and the bag-of-words model will outperform other classifier configurations based solely on linguistic features or a bag-of-words model.

H2b: When under attack, a classifier based on ensemble learning incorporating linguistic features and the bag-of-words model will outperform other classifier configurations based solely on linguistic features or a bag-of-words model.

4.2 Dataset Acquisition and Descriptive Statistics

4.2.1 Dataset Acquisition

Training and evaluating classifiers require documents that represent both document classes: documents suspected to be fraudulent stock recommendations and documents that contain reliable recommendations. The identification of appropriate documents was carefully conducted in cooperation with our domain experts; it also incorporated feedback from financial institutions and the financial supervisory authority. The SEC has published several criteria that provide the basis for identifying documents that represent stock

recommendations as suspicious and/or fraudulent.[1] We searched for stock recommendations that fulfilled these criteria and included small-cap stocks traded primarily in markets with little regulation, labeling these recommendations as *suspicious*. To acquire newsletters promoting stocks that matched these criteria, we used the newsletter.hotstocked.com archive. This internet service does not publish its own stock recommendations but aggregates diverse stock recommendations that are published either on the web or in investment newsletters.

The identification of reliable stock recommendations was also carefully conducted. Stock recommendations published on the internet that do not fulfill the SEC criteria for suspiciousness are not guaranteed to be reliable (they can be manipulative without triggering the conditions required for a suspicious or fraudulent designation (Aggarwal & Wu, 2006)). Thus, we only considered documents that were published in more reliable sources, specifically financial newspapers. We used analyst reports that contain stock recommendations published by Dow Jones Newswires. Based on feedback from our domain experts, we selected Dow Jones Newswires as an appropriate source for reliable documents because it is a major financial news provider that is well-regarded by financial professionals (Tetlock, 2007) and because its documents are created by many different authors. Thus, we downloaded the analyst reports published by Dow Jones Newswires and designated these reports as *non-suspicious* stock recommendations.

Following the above procedures, we acquired a total of 14,556 suspicious and 3,342 non-suspicious stock recommendations published between December 15, 2010 and February 10, 2012. We removed stock symbols, firm names, and publisher names from the documents to ensure the generalizability of the results. In the Discussion section below, we elaborate on the finding that our classification results remain robust when taking a second dataset into account.

We considered only the first suspicious recommendation that was published with regard to a specific stock to remove identical recommendations and to avoid overfitting. This restriction reduced the final number of suspicious stock recommendations used in this study to 896. In addition, a review of the non-suspicious documents obtained from Dow Jones Newswires reveals that some of these documents consisted only of tables that span a large number of stocks but do not include any analyses. Thus, we discarded these documents from the analysis, and a total of 2,088 documents were used to train our fraud-detection classifiers. Our results remain robust regardless of whether the complete or the reduced datasets for suspicious and non-suspicious recommendations were assessed.

All of the classifiers were trained with a biased cost function because of the unbalanced dataset (Witten, Frank, Hall, & Pal, 2016). Therefore, the error on suspicious examples was multiplied by the total number of non-suspicious examples divided by the total number of suspicious examples (2,088/896 = 2.33) during the training. We also trained the classifiers with a non-biased cost function; the recall of the suspicious documents was most heavily affected by this (it decreased), significantly affecting the overall classification performance, as shown by the *F*-measure.

4.2.2 Descriptive Statistics

For both suspicious and non-suspicious stock recommendations, we determined *information content*, *readability*, and *sentiment*, as described above. We tested whether the theoretically derived linguistic features were suited for differentiating between the two document classes and were consequently useful in fraud detection by performing Wilcoxon rank-sum tests to assess the equality of medians (see Table 1). With respect to the *information content* of the examined recommendations, we found that *Entropy* was significantly higher for suspicious stock recommendations than for non-suspicious stock recommendations. This result indicates that suspicious stock recommendations contain more information than non-suspicious stock recommendations. With respect to *readability*, each readability measure indicates that suspicious stock recommendations are easier to understand than non-suspicious stock recommendations. Therefore, the null hypothesis of equal medians could be rejected for *ARI*, *Flesch*, and *Fog* at a 1% confidence level. Moreover, Table 1 shows that suspicious stock recommendations indicate sentiments that are more positive than the sentiments of non-suspicious stock recommendations. In addition, compared with non-suspicious recommendations, suspicious stock recommendations contain a higher fraction of positive sentiment-bearing words (positivity) and a lower fraction of negative sentiment-bearing words (negativity). Thus, all the examined linguistic features discriminate between the two document categories, and the observed differences are consistent with the kernel theories that justified our feature selection. Given these results, we consider the linguistic feature set to be useful in the fraud detection context.

[1] In this context, the SEC warns investors against trading stocks that are recommended if it is unclear whether the recommender holds a position in the recommended stock, whether compensation was paid to the recommender (if the recommendation was an advertisement), or whether the recommended stock is a small, thinly traded company (SEC, 2012a).

Table 1. Descriptive Statistics for Linguistic Features and Results of Wilcoxon Rank-Sum Tests for the Equality of Medians (*/**/*:p<1%/5%/10%)**

Variable	Linguistic feature	Suspicious stock recommendations		Non-suspicious stock recommendations		*p*-value
		Mean	*Median*	*Mean*	*Median*	
Information content	Entropy	7.1826	7.2858	6.8579	6.8833	< 0.01***
Readability	ARI	13.951	13.755	15.739	15.561	< 0.01***
	Flesch	45.111	44.276	39.240	39.475	< 0.01***
	Fog	15.947	15.949	17.072	16.971	< 0.01***
Sentiment	Polarity	0.4322	0.4390	0.1218	0.1261	< 0.01***
	Positivity	0.0861	0.0864	0.0688	0.0686	< 0.01***
	Negativity	0.0344	0.0333	0.0538	0.0525	< 0.01***

4.3 Naive Evaluation

We evaluated the performance of the different classifiers and the general validity of the proposed problem solution utilizing *k*-fold stratified cross-validation (k = 10), which avoids overly optimistic results (Mitchell, 1997). We created a contingency table that contains the number of correctly and incorrectly classified examples. These results were classified as true positives (*TP*), true negatives (*TN*), false positives (*FP*), or false negatives (*FN*). On this basis, the performance metrics of accuracy, precision, recall, and F_1 (Hotho et al., 2005; Kotsiantis, 2007; van Rijsbergen, 1979) were calculated through micro-averaging (Chau & Chen, 2008). We calculated precision, recall, and F_1 for the "suspicious" and "non-suspicious" classes. The evaluation metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{16}$$

The results of the 10-fold cross-validation are presented in Table 2. This table presents the results for Classifier A, which accounts only for the bag-of-words model; for Classifier B, which accounts only for the linguistic features of information content, readability, and sentiment; for Classifier C, which utilizes both feature sets; and for Classifiers D and E, which are based on ensemble learning. For this classic evaluation, we selected *thr* = 0.5 for Classifier E, but other values between 0 and 1 produced similar results, as illustrated in the following section. If only the basic text-based features are taken into account (Classifier A), an accuracy of 99.67% is achieved. In addition, the precision, recall, and F_1 scores are above 98% for all of the classes of results. These are excellent scores, although previous text mining studies have reported comparable results for related document classification tasks (Joachims, 1998; Webb et al., 2005).

Furthermore, Classifier B achieves a classification accuracy of 83.61%; thus, 83.61% of all cases are classified correctly through this approach. Misclassification costs (i.e., the consequences of classifying suspicious recommendations as non-suspicious and vice versa) are particularly important in fraud detection (Phua et al., 2010). Thus, the classification results for both classes should also be taken into account. In the case of Classifier B, significantly lower precision appears to be achieved for the suspicious class than for the non-suspicious class. However, the difference in recall between these two classes is less substantial: 86.84% of the suspicious recommendations are classified as suspicious, whereas 82.31% of the non-suspicious recommendations are classified as non-suspicious.

Classifier C, which incorporates the bag-of-words model and linguistic features, produces results that are comparable to, but slightly lower than, the results of Classifier A. Regarding the classifiers based on ensemble learning, Classifier D demonstrates an overall classification performance that appears to be between those of Classifiers A and B. Finally, Classifier $E_{0.5}$ produces an overall classification performance that is comparable to the performance of Classifiers A and C. Thus, Classifiers A, C, and $E_{0.5}$ achieve very good results in the identification of suspicious recommendations and produce slightly better overall performance than Classifiers B and D.

Table 2. SVM Classification Results (All values Are Given as Percentages)

		Class *suspicious*			Class *non-suspicious*		
	Accuracy	**Precision**	**Recall**	F_1	**Precision**	**Recall**	F_1
Classifier A	99.67	98.99	99.89	99.44	99.95	99.57	99.76
Classifier B	83.61	67.72	86.84	76.10	93.54	82.31	87.57
Classifier C	98.79	97.33	98.61	97.97	99.43	98.86	99.14
Classifier D	87.50	70.54	100.00	82.73	100.00	82.17	90.21
Classifier $E_{0.5}$	98.69	95.83	100.00	97.87	100.00	98.14	99.06

Table 1. The 20 Most Important Features for Classifier C by SVM Weight

Rank	(Linguistic) feature	Weight	Rank	(Linguistic) feature	Weight
1	Alert	1.5032	11	fitch	0.5315
2	Sp	1.3050	12	upgrade	0.5219
3	*Polarity*	0.8393	13	gbp	0.5059
4	Analyst	0.7964	14	bank	0.4732
5	*Entropy*	0.6138	15	moodys	0.4674
6	Said	0.6123	16	eur	0.4530
7	Technology	0.6097	17	chart	0.4204
8	pick	0.5844	18	read	0.4177
9	Mid	0.5642	19	list	0.3927
10	ratings	0.5400	20	*Flesch*	0.3572

4.4 Robustness Evaluation

To evaluate the robustness of the proposed classifiers, we first analyzed the relative importance of the linguistic features. Thereafter, we simulated an attack on the classifiers to evaluate how these performance figures change if the input documents are manipulated according to a document manipulation strategy described in the Appendix.

During the training process, SVM assigns certain weights to the features that it assesses. We used these assigned weights to evaluate the importance of individual features (i.e., individual words or linguistic features). In particular, weights with higher absolute values exert greater influence on the classification decision (Guyon et al., 2002). Table 3 reports the 20 most important features for Classifier C, sorted by weight. This table shows that linguistic features are of great importance. For Classifier C, polarity (i.e., sentiment) has the highest rank among the linguistic features, whereas entropy (i.e., information content) is ranked #5. Furthermore, Flesch (i.e., readability) is ranked #20 (out of the 9,990 features that are relevant in the model).

Furthermore, a number of features of the bag-of-words model are also among the 20 most important features for Classifier C. For example, many suspicious stock recommendations *alert* (#1) investors about stock *picks* (#8). From a fraudster's point of view, these words should be avoided in future stock recommendations to prevent detection by the classifiers. However, a fraudster would also need to alter the linguistic features of a message. As a consequence, we expect Classifier C to be more robust than Classifier A against manipulations because important linguistic features pose a dilemma for fraudsters—as marketing theory points out, t message manipulation of linguistic features to avoid identification by Classifier C would decrease the advertising effect of the fraudster's recommendations.

To further explore the robustness of the classifiers, we performed a simulation of a worst-case attack. We assumed that the fraudster has obtained or could fully replicate the feature weights of Classifier A and is thereby fully aware of the most relevant words that should be avoided; this assumption is much more strict than related attack simulation approaches that do not assume this type of insider knowledge (Jorgensen et al., 2008; Webb et al., 2005). Second, we assumed that the fraudster did not want to reduce the advertising effect of the document. As a result, the linguistic features (and thus also Model B) were expected to be relatively stable. For each suspicious document, given

the degree of manipulation *m*, the fraudster replaced the m% most important features that drive suspiciousness with suitable synonyms that are considered to be less suspicious (the detailed algorithm for document manipulation is presented in the Appendix).

To evaluate the robustness of the different classifiers, we evaluated their classification performance by increasing the manipulation degree *m* (i.e., the percentage of words that are replaced by suitable synonyms). This assessment is graphically depicted in Figure 4. In accordance with H1, we see that, when under attack (i.e., if *m* is increased), the classifier based on linguistic features only (Classifier B) outperforms the classifier based on the bag-of-words model with respect to accuracy (Classifier A). Although the accuracy of Classifier B is below that of Classifier A at $m = 0$, Classifier B outperformed Classifier A for values of *m* that are equal to or greater than 0.3. The same result was observed for the F_1 measure in the case of $m \geq 0.4$, which combines the precision and recall factors. Thus, the results of this simulated attack support H1.

Furthermore, Classifier C appears to be more robust than a classifier that uses only the bag-of-words model (Classifier A) or linguistic features (Classifier B), which supports H2a. With respect to the performance of the developed classifier based on ensemble learning approaches, it can be concluded that Classifier D and the various Classifier E configurations (i.e., for several different *thr* values) exhibit by far the best robustness to the attacks, as demonstrated by the various performance measures. However, Classifier E outperformed Classifier D in most scenarios and performed reasonably well at $m = 0$.

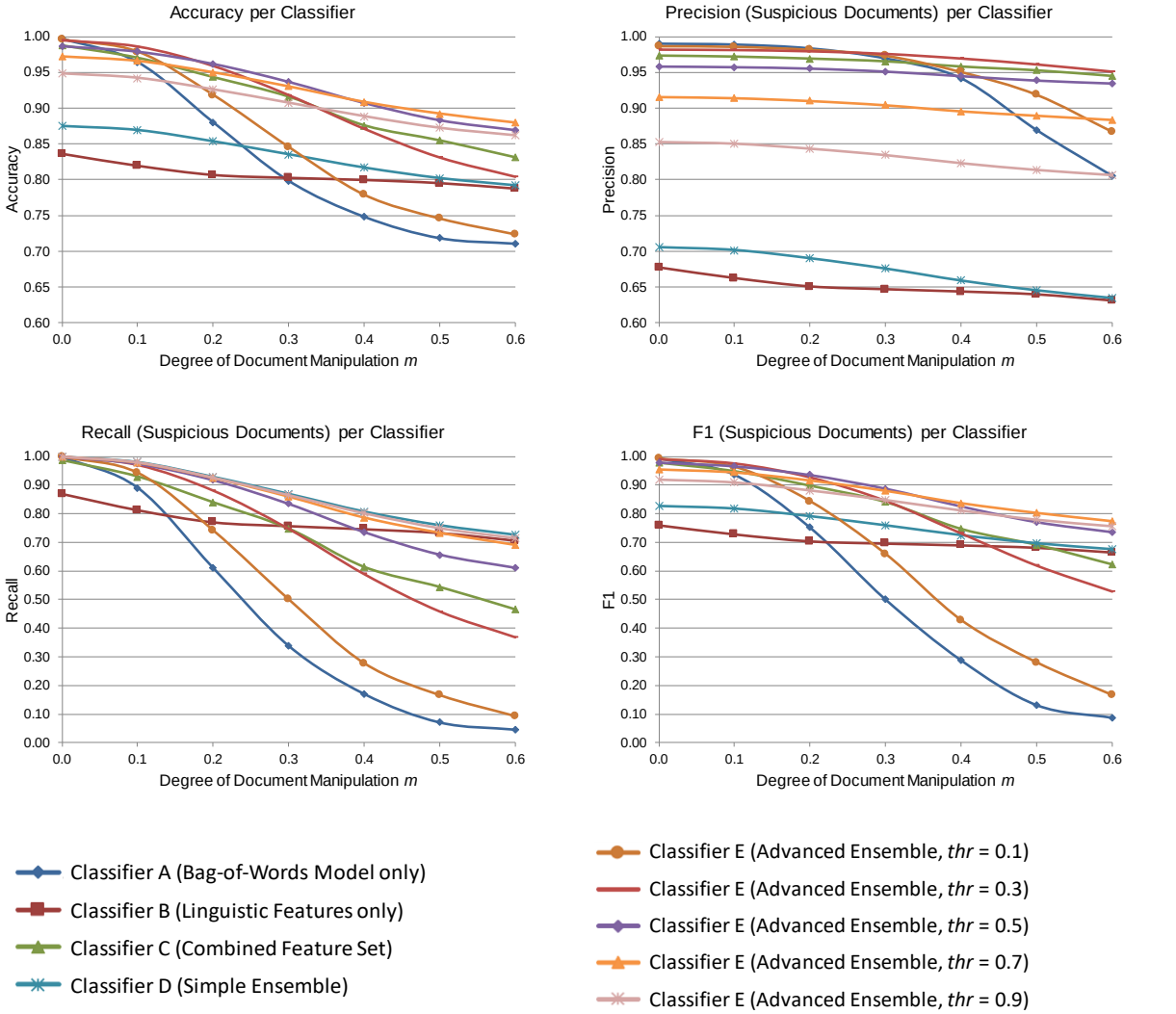

Figure 4. The Robustness of Classifiers Against Countermeasures

In short, given a performance metric and a document manipulation level, Classifier $E_{0.5}$ is either the outright best-performing classifier or performs similarly to the best-performing classifier (for instance, with regard to accuracy, the absolute difference between Classifier $E_{0.5}$ and the best-performing classifier was always equal to or less than 1%).

These findings support H2b, which states that, when under attack, a classifier based on ensemble learning that accounts for both linguistic features and the bag-of-words model outperforms models based on either linguistic features or the bag-of-words model alone. However, the sensitivity of this classifier must be established by selecting an appropriate *thr* value. Classifier $E_{0.5}$ is more robust to manipulations than Classifier C, which shows that the application of ensemble learning is more appropriate than the combination of feature sets at the classifier level.

5 Discussion

Our results show that the proposed design principles and features can be used to address the design requirements for robust fraud detection. We found that prior theories from marketing and financial economics provide a foundation (justificatory knowledge) for identifying suspicious stock recommendations. Notably, we found that such recommendations are easier to read, incorporate more positive sentiments, and provide greater information content (which supports advertising success). For the forecasting models, we confirmed the usefulness of theory-based linguistic features (see H1). A classifier based on just the linguistic features provides good results. The robustness evaluation confirmed the usefulness of theory-based linguistic features (see H2a, H2b) and demonstrates that an ensemble learning approach that uses linguistic features and bag-of-words models is appropriate for generating a robust fraud-detection classifier.

We acknowledge that our approach has limitations. First, our approach addressed two different types of documents that can be regarded as examples of suspicious and non-suspicious stock recommendations (relying on criteria published by the SEC to identify suspicious stock recommendations and on analyst reports published by Dow Jones Newswires to identify non-suspicious recommendations). An alternative approach would be to assess recommendations by domain experts. This approach was criticized by our domain experts because one cannot be certain whether a recommendation that is labeled as suspicious actually aims to manipulate stock prices because they would not know the specific intentions of the publisher (supported by Bolton & Hand, 2002). As also argued by the involved market supervisory authority, any such assessment for training a classifier must follow documented criteria that can be disclosed.

Our predictions are based on stock recommendations for which publishers self-disclosed that they were paid to advertise the stocks in question. Thus, the study does not assess recommendations without this disclaimer. However, the inclusion of this statement is obligatory (Hu, McInish, & Zeng, 2009), and the SEC cannot prohibit the publication of fraudulent stock recommendations that include this statement as doing so would obstruct "freedom of speech" (SEC, 2012a). Thus, we cannot claim that our study incorporates all possible types of suspicious stock recommendations, although it does include a significant subset of them. By excluding the disclaimers during training, we ensured that the classifiers could detect the remaining suspicious stock recommendations that did not contain disclaimers.

To rule out the possibility that the results of this study were driven by fundamental differences in the document sources used for training (e.g., a news agency such as Dow Jones Newswires might have guidelines for the composition of related documents) that were different from suspicious documents (which are published by various promoters), we reran our experiments using another source of non-suspicious documents (recommendations published in the Yahoo! Finance category "Investing Ideas & Strategies,"). In this setting, the classification results remained robust. This allowed us to further establish robustness and overcome a major limitation of fraud-detection systems (manipulators adapt to them after their characteristics have been published) (Bolton & Hand, 2002).

6 Conclusion

In this study, we present a fraud-detection approach for identifying suspicious stock recommendations. To improve the robustness of this approach, we propose new design principles, design features, and different classifiers that utilize both a bag-of-words model and linguistic features derived from domain kernel theories.

We contribute theoretically and methodologically to the literature in several ways. Most importantly, we propose design principles and specific design features for robust fraud-detection systems that address the problem class of information-based market manipulations, and we demonstrate robust evaluations based on attack simulations. Our approach (that includes bag-of-words models and theory-motivated linguistic features in combination with ensemble learning) significantly increases the robustness of fraud-detection. Through our work, we demonstrate that the shift from foundationalism to instrumentalism in contemporary data mining research can contribute to problem solving. In this case, foundationalism seeks to progress toward truth by following inductive logic, whereas instrumentalism attempts to engage in

problem solving, provides the flexibility to build an approach on the basis of relevant theories, and utilizes different reasoning principles, including both induction and deduction (Kilduff et al., 2011). To the best of our knowledge, our study is the first to investigate the problem of information-based fraud detection by analyzing and classifying stock recommendations.

The practical contributions of this study are threefold. First, the proposed fraud-detection classifiers can be included in a fraud detection system (FDS) to enhance the "information-based market manipulation detection capabilities" of firms and market surveillance authorities. In particular, existing detection schemes can be improved to clearly and correctly identify stock recommendations serving in pump-and-dump schemes. Additionally, the proposed fraud-detection classifiers could also be used to complement established FDS covering other manipulation scenarios (Gregory & Muntermann, 2014). Second, our findings may be relevant to security software developers who are addressing this problem domain, at least with respect to stock scam emails (Symantec, 2011). Our classifiers could be included in browser toolbars, which already generate warnings for phishing websites. Finally, the design principles and design features for improving classifier robustness and its evaluation could be applied to other fields or languages apart from English to investigate the robustness of text-based classifiers—for example, for opinion spam (Liu, 2012) in the social commerce context.

References

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., & Nunamaker, J. F. (2010). Detecting fake websites: The contribution of statistical learning theory. *MIS Quarterly*, *34*(3), 435-461.

Abernethy, A. M., & Franke, G. R. (1996). The information content of advertising: A meta-analysis. *Journal of Advertising*, *25*(2), 1-17.

Abruzzini, P. (1967). Measuring language difficulty in advertising copy. *Journal of Marketing*, *31*(2), 22-26.

Aggarwal, R. K., & Wu, G. (2006). Stock market manipulations. *The Journal of Business*, *79*(4), 1915-1953.

Allen, F., & Gale, D. (1992). Stock-price manipulation. *The Review of Financial Studies*, *5*(3), 503-529.

Apté, C., Damerau, F., & Weiss, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, *12*(3), 233-251.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: a critique. *Language & Communication*, *21*(3), 285-301.

Biggio, B., Corona, I., Fumera, G., Giacinto, G., & Roli, F. (2011). Bagging classifiers for fighting poisoning attacks in adversarial classification tasks. *Lecture Notes in Computer Science*, *6713*, 350-359.

Biggio, B., Fumera, G., & Roli, F. (2010). Multiple classifier systems under attack. *Lecture Notes in Computer Science*, *5997*, 74-83.

Böhme, R., & Holz, T. (2006). The effect of stock spam on financial markets. *5th Workshop on the Economics of Information Security*.

Bollen, J., & Huina, M. (2011). Twitter mood as a stock market predictor. *Computers and Operations Research*, *44*(10), 91-94.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, *17*(3), 235-255.

Campbell, W. M., Campbell, J. P., Gleason, T. P., Reynolds, D. A., & Wade, S. (2007). Speaker verification using support vector machines and high-level features. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2085-2094.

Caruana, G., & Li, M. (2012). A survey of emerging approaches to spam filtering. *ACM Computing Surveys*, *44*(2), 1-27.

Caudill, S. B., Ayuso, M., & Guillén, M. (2005). Fraud detection using a multinomial logit model with missing information. *The Journal of Risk and Insurance*, *72*(4), 539-550.

Chandy, R. K., Tellis, G. J., MacInnis, D. J., & Thaivanich, P. (2001). What to say when: Advertising appeals in evolving markets. *Journal of Marketing Research*, *38*(4), 399-414.

Chau, M., & Chen, H. (2008). A machine learning approach to web page filtering using content and structure analysis. *Decision Support Systems*, *44*(2), 482-494.

Chen, R., Chen, T., & Lin, C. J. (2006). A new binary support vector system for increasing detection rate of credit card fraud. *International Journal of Pattern Recognition and Artificial Intelligence*, *20*(2), 227-239.

Clark, G. L., Kaminski, P. F., & Brown, G. (1990). The readability of advertisements and articles in trade journals. *Industrial Marketing Management*, *19*(3), 251-260.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375-1388.

de Bondt, W. F. M. (1998). A portrait of the individual investor. *European Economic Review*, *42*(3-5), 831-844.

Dietterich, T. G. (1997). Machine-learning research: Four current directions. *AI Magazine*, *18*(4), 97-136.

Djeraba, C. (2002). Content-based multimedia indexing and retrieval. *IEEE MultiMedia*, *9*(2), 18-22.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383-417.

Fast, A., Friedland, L., Maier, M., Taylor, B., Jensen, D., Goldberg, H. G., & Komoroske, J. (2007). Relational data pre-processing techniques for improved securities fraud detection. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 941-949.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37-54.

FBI. (2011). *Financial crimes report to the public*. http://www.fbi.gov/stats-services/publications/financial-crimes-report-2010-2011

Felixson, K., & Pelli, A. (1999). Day end returns: Stock price manipulation. *Journal of Multinational Financial Management*, *9*(2), 95-127.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press.

Frieder, L., & Zittrain, J. (2006). *Spam works: Evidence from stock touts and corresponding market activity*. Berkman Center Research.

Glancy, F. H., & Yadav, S. B. (2011). A computational model for financial reporting fraud detection: On quantitative methods for detection of financial fraud. *Decision Support Systems*, *50*(3), 595-601.

Goodman, J., Cormack, G. V., & Heckerman, D. (2007). Spam and the ongoing battle for the inbox. *Communications of the ACM*, *50*(2), 24-33.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337-355.

Gregor, S., & Jones, D. (2007). The Anatomy of a design theory. *Journal of the Association for Information Systems*, *8*(5), 312-335.

Gregor, S., Müller, O., & Seidel, S. (2013). Reflection, abstraction, and theorizing in design and development research. *Proceedings of the 21st European Conference on Information Systems*.

Gregory, R. W., & Muntermann, J. (2014). Research note—Heuristic theorizing: Proactively generating design theories. *Information Systems Research*, *25*(3), 639-653.

Groth, S. S., Siering, M., & Gomber, P. (2014). How to enable automated trading engines to cope with news-related liquidity shocks? Extracting signals from unstructured data. *Decision Support Systems*, *62*, 32-42.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389-422.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann.

Hanke, M., & Hauser, F. (2008). On the effects of stock spam e-mails. *Journal of Financial Markets*, *11*(1), 57-83.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*, 993-1001.

Hevner, A., & Chatterjee, S. (2010). *Design research in information systems: Theory and practice*: Springer.

Hevner, A.R., March, S. T., & Park, J. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105.

Holton, C. (2009). Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem: IT decisions in organizations. *Decision Support Systems*, *46*(4), 853-864.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). a brief survey of text mining. *GLDV Journal for Computational Linguistics*, *20*(1), 19-62.

Hu, B., McInish, T., & Zeng, L. (2009). The CAN-SPAM Act of 2003 and stock spam emails. *Financial Services Review*, *18*, 87-104.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, *52*(3), 674-684.

IEEE. (1990). *IEEE standard glossary of software engineering terminology (IEEE Std 610.12-1990)*. IEEE.

Iivari, J. (2007). A paradigmatic analysis of information systems as a design science. *Scandinavian Journal of Information Systems*, *19*(2), 39-64.

Iivari, J. (2015). Distinguishing and contrasting two strategies for design science research. *European Journal of Information Systems*, *24*(1), 107-115.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning*, 137-142.

Jones, M. A., & Smythe, T. (2003). The information content of mutual fund print advertising. *The Journal of Consumer Affairs*, *37*(1), 22-41.

Jorgensen, Z., Zhou, Y., & Inge, M. (2008). A multiple instance learning strategy for combating good word attacks on spam filters. *Journal of Machine Learning Research*, *8*, 1115-1146.

Kilduff, M., Mehra, A., & Dunn, M. B. (2011). From blue sky research to problem solving: A philosophy of science theory of new knowledge production. *The Academy of Management Review*, *36*(2), 297-317.

Kim, K.-j. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1-2), 307-319.

Kolcz, A., & Teo, C. H. (2009). Feature weighting for improved classifier robustness. *Proceedings of the Sixth Conference on Email and Anti-Spam.*

Korfiatis, N., García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, *11*(3), 205-217.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*(3), 249-268.

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: Anatomy of a research project. *European Journal of Information Systems*, *17*, 489-504.

Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information Systems*, *13*(6), 395-423.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1-167.

Loughran, T., & McDonald, B. (2010). *Measuring Readability in Financial Text* (Working paper). University of Notre Dame.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*, 251-266.

Martin, M. A., & Rey, J.-M. (2000). On the role of Shannon's entropy as a measure of heterogeneity. *Geoderma*, *98*(1-2), 1-3.

Meth, H., Mueller, B., & Maedche, A. (2015). Designing a requirement mining system. *Journal of the Association for Information Systems*, *16*(9), 799-837.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, *78*(2), 311-329.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 9). Prentice-Hall.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & 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), 559-569.

Öğüt, H., Mete Doğanay, M., & Aktaş, R. (2009). Detecting stock-price manipulation in an emerging market: The case of Turkey. *Expert Systems with Applications*, *36*(9), 11944-11949.

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45-77.

Perols, J., Chari, K., & Agrawal, M. (2009). Information market-based decision fusion. *Management Science*, *55*(5), 827-842.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *Proceedings of the International Conference on Intelligent Computation Technology and Automation*, 50-53.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 211-218.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems*, *32*(3), 229-267.

Ravisankar, P., Ravi, V., Raghava Rao, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*(2), 491-500.

Resnik, A., & Stern, B. L. (1977). An analysis of information content in television advertising. *Journal of Marketing*, *41*(1), 50-53.

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

SEC. (1959). *A 25 year summary of the activities of the securities and exchange commission*. SEC.

SEC. (2012a). Internet fraud: Tips for checking out newsletters. http://www.sec.gov/investor/pubs/cyberfraud/newsletter.htm

SEC. (2012b). Investor alert: Social media and investing—Avoiding fraud. http://www.sec.gov/investor/alerts/socialmediaandfraud.pdf

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423.

Shannon, C.E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*(1), 50-64.

Siering, M. (2019). The economics of stock touting during internet-based pump and dump campaigns. *Information Systems Journal*, *29*(2), 456-483.

Siering, M., Clapham, B., Engel, O., & Gomber, P. (2017). A taxonomy of financial market manipulations: Establishing trust and market integrity in the financialized economy through automated fraud detection. *Journal of Information Technology*, *32*(3), 251-269.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). MIT Press.

Smith, E.A., & Senter, R.J. (1967). *Automated readability index*. Aerospace Medical Research Laboratories.

Sonnier, G. P., McAlister, L., & Rutz, O. J. (2011). A dynamic model of the effect of online communications on firm sales. *Marketing Science*, *30*(4), 702-716.

Symantec. (2011). Global debt crises news drives pump-and-dump stock scams. http://www.symantec.com/connect/blogs/global-debt-crises-news-drives-pump-and-dump-stock-scams

Tay, F. E. H., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega*, *29*(4), 309-317.

Teahan, W.J. (2000). Text classification and segmentation using minimum cross-entropy. *Proceedings of the International Conference on Content-Based Multimedia Information Access*, 943-961.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*(3), 1437-1467.

Vaishnavi, V. K., & Kuechler, W. (2015). *Design science research methods and patterns: Innovating information and communication technology* (2nd ed.). CRC Press.

Vakratsas, D., & Ambler, T. (1999). How advertising works: What do we really know? *Journal of Marketing Research*, *63*(1), 26-43.

Valentini, G., & Masulli, F. (2002). Ensembles of learning machines. *Lecture Notes in Computer Science*, *2486*, 3-20.

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Butterworths.

Webb, S., Chitti, S., & Pu, C. (2005). An experimental evaluation of spam filter performance and robustness against attack. *Proceedings of the First International Conference on Collaborative Computing*.

Wei, C.-P., & Dong, Y.-X. (2001). A mining-based category evolution approach to managing online document categories. *Proceedings of the 34th Hawaii International Conference on System Sciences*.

Wieland, A., & Marcus Wallenburg, C. (2012). Dealing with supply chain risks: Linking risk management practices and strategies to performance. *International Journal of Physical Distribution & Logistics Management*, *42*(10), 887-905.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.

You, H., & Zhang, X. (2009). Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting Studies*, *14*, 559-586.

Zhang, W., & Skiena, S. (2010). Trading strategies to exploit blog and news sentiment. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378-393.

Zhou, L., & Chaovalit, P. (2008). Ontology-supported polarity mining. *Journal of the American Society for Information Science and Technology*, *59*(1), 98-110

Appendix: Algorithm for Document Manipulation

1. Extract all of the unique features from the document.
2. Use the SVM decision function to rank the features according to their contribution to classifying the document into the suspicious class. Because classifier A is based on a linear kernel, this decision function takes the following form (Guyon et al., 2002):

$$\mathrm{d}(\mathbf{x}) = \mathbf{x}\cdot\mathbf{w} + b = x_1 w_1 + x_2 w_2 + \cdots + x_n w_n + b \tag{17}$$

   In this equation, **x** is the TF-IDF vector, **w** is the SVM weight vector, and b is the hyperplane bias. As shown above, each of the components (the summands) contributes to the final value of $\mathrm{d}(\mathbf{x})$. The components $x_i w_i$ correspond to the features in the bag-of-words vocabulary. If the suspicious documents are labeled "1" and the non-suspicious documents are labeled "-1" in the training set, then a positive value for a particular component $x_i w_i$ would indicate that it contributes to classifying the document into the suspicious class, and the absolute value of this component would represent the degree to which the feature contributes to the final outcome. The fraudster therefore ranks the features in descending order (i.e., largest to smallest) according to their $x_i w_i$ values, such that the features that provide the greatest contributions to classifying the document into the suspicious class are at the top of this ranked list.
3. In the document, the fraudster locates the words corresponding to the topmost $(100 \times m)$% features from the list. The fraudster considers only features with positive $x_i w_i$ values (even if this means that fewer than $(100 \times m)$% features are considered). The fraudster replaces each of these words with a suitable synonym. The fraudster's lexical knowledge is modeled with two lexical resources: WordNet (Fellbaum, 1998) and SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010). The fraudster modifies a word in the following manner:
   a. The fraudster looks up the word's lemma in WordNet and retrieves all of its synonyms
   b. The fraudster uses SentiWordNet to determine the amount of positivity p_i and the amount of negativity n_i to assign to each synonym s_i. If the word to be replaced bears a positive sentiment ($p_i > n_i$), then only the words with $p_i > 0$ would be regarded as suitable replacements. Similarly, if the word to be replaced bears a negative sentiment ($n_i > p_i$), only the words with $n_i > 0$ are regarded as suitable replacements. The intuition behind this supposition is that the fraudster wishes to preserve the marketing effect of the document (see assumption 2).
   c. The fraudster looks at the weight w_i for each of the synonyms. If a synonym does not exist in the bag-of-words vocabulary, its weight w_i equals 0. The synonyms are ranked in ascending order (i.e., smallest to largest) according to their weights. This classification procedure means that the synonyms with the most negative weights (i.e., the synonyms that contribute the most to classifying the document in the non-suspicious class) will be at the top of the list. The fraudster uses the topmost synonym from the list to replace the original word in the document.

About the Authors

Michael Siering is a postdoctoral research associate at Goethe University Frankfurt and works as a project manager in the financial services industry. His research focuses on decision support systems in electronic markets, with a focus on the analysis of user-generated content. His work has been published in outlets such as *Journal of Management Information Systems*, *Journal of the Association for Information Systems*, *Information Systems Journal*, *Journal of Information Technology*, and *Decision Support Systems*.

Jan Muntermann is a full professor of electronic finance and digital markets at the University of Goettingen, Germany. His research interests include (big) data analytics and managerial decision support, digital business strategy development and execution, and the conceptual and methodological foundations of theory development in information systems research. He has published in journals such *as Information Systems Research*, *Decision Support Systems*, and the *European Journal of Information Systems*.

Miha Grčar was a researcher in social media and news analytics at the Department of Knowledge Technologies at Jožef Stefan Institute, Slovenia, for over 10 years. His area of expertise is a mix of data mining, text mining, stream mining, machine learning, recommender systems, information retrieval, network analysis, and language technologies. He is also a skilled software developer and a co-founder of Sowa Labs GmbH, a company recently acquired by Boerse Stuttgart Group. Miha now works as the CTO and a managing director of Sowa Labs.