Decision Support Systems 158 (2022) 113782

Contents lists available at ScienceDirect

Decision Support Systems

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

Explainability and fairness of RegTech for regulatory enforcement: Automated monitoring of consumer complaints

Michael Siering *

Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany

ARTICLE INFO

Keywords:
Regulatory technology (RegTech)
Regulatory enforcement
Consumer complaints
Explainable AI
Fair AI
Predictive analytics

ABSTRACT

The application of regulatory technology (RegTech) for monitoring comprehensive data sources has gained increased importance. Nevertheless, previous research neglects that the output of RegTech applications has to be explainable and non-discriminatory. Within this study, we propose design principles and features for a RegTech approach which provides automated assessments of financial consumer complaints. We follow three main design principles. First, we build upon information diagnosticity theory to address the need for explainable classifications empowering regulators to justify their actions. Second, we consider a bag-of-words representation and ensemble learning to ensure high classification accuracy. Third, we take into account author characteristics to avoid discriminating classifications. We evaluate our approach in the financial services industry and show its value for identifying consumer complaints resulting in monetary compensations. The proposed design principles and features are highly relevant for regulators, corporations as well as consumers.

1. Introduction

In recent years, regulatory technology (RegTech) for monitoring, reporting and compliance has become more and more important [1]. Specifically in a financial context, an increased amount of regulation has paved the way for technologies supporting regulatory authorities conducting supervisory tasks [2,3]. Apart from high predictive performance, RegTech applications have to fulfill two important additional requirements which have been under-researched so far: applications based on machine learning techniques have to offer explanations why a specific classification decision was made to foster transparency and encourage trust towards system output [4]. Additionally, related systems should be fair, i.e. discrimination of members of specific groups or populations should be avoided [5].

We address these important requirements in the field of financial consumer complaints. In the United States, the Consumer Financial Protection Bureau (CFPB) supervises financial service providers, enforces laws and has the task to protect consumers [6]. One important trigger for regulatory actions are consumer complaints. Consumers can submit their experiences made with financial service providers. These complaints are investigated and forwarded to the targeted companies. Typically, a supervisory authority's resources are limited. Thus, an approach which helps to direct regulatory efforts by identifying consumer complaints which are likely to be successful, i.e. which result in monetary compensations, is valuable. Nevertheless, previous research on consumer complaints has neglected the problem to forecast the complaint outcome.

With this study, we close these research gaps. Therefore, we follow a design science research approach [7,8] and propose several design principles and features which address the design requirements of explainability, classification performance and fairness. We evaluate the proposed design principles and features using multiple datasets: First, as a basis for the explanation of classification decisions, we take into account theoretically-derived linguistic features and validate our features using a dataset of 1500 financial product reviews. Second, for efficiently training our classifiers, we consider a balanced sample of 6000 financial consumer complaints. Third, for examining the practical relevance of our results, we evaluate the classifiers with a holdout sample of 75,000 consumer complaints with a natural class distribution.

We contribute to research by focusing on explainability and fairness of RegTech applications. Specifically, we propose new design principles for predicting the outcome of consumer complaints, including theory-based feature engineering to increase explainability and specialized classifiers to foster fairness. We thereby extend previous research, which has not proposed predictive models within this context yet. The proposed classifiers can be applied by different stakeholders. Our study has

* Corresponding author.
E-mail address: siering@wiwi.uni-frankfurt.de.

https://doi.org/10.1016/j.dss.2022.113782
Received 9 August 2021; Received in revised form 25 March 2022; Accepted 26 March 2022
Available online 2 April 2022

several practical implications for regulatory authorities, financial service providers and financial consumers.

The remainder of this paper is structured as follows. Section 2 presents the background of this study, encompassing RegTech applications, explainability and fairness of artificial intelligence and prior research on consumer complaints. Section 3 presents our research approach, including the derivation of design requirements, design principles and design features. Within section 4, we present the evaluation of the proposed design principles and the resulting classifiers. Section 5 discusses the results, whereas section 6 concludes.

2. Research background

2.1. Regulatory technologies

Regulatory Technology (RegTech) encompasses the usage of information technology for regulatory monitoring, reporting and compliance [1]. In the financial context, a highly regulated environment has paved the way for an increased usage of RegTech [2,9]: An increased amount of regulation necessitates enhanced resources for regulatory monitoring and compliance. This also increases the possibility for automatization. On the one hand, market participants can use RegTech to comply with regulations (i.e. for regulatory reporting). On the other hand, regulators can apply RegTech to increase their monitoring capabilities (i.e. analyzing large amounts of data). Overall, increased automation leads to cost savings and near-real time surveillance capabilities of regulators [10] and corporations [11].

In previous research, different approaches have been proposed to support regulators and market participants. In the field of fraud detection, various studies address the detection of bank fraud, insurance fraud, and other related financial fraud [12]. For instance, financial statements, insurance claims and credit card transaction data have been analyzed to identify fraudulent events [13]. Additionally, several studies address the identification of suspicious situations in financial markets [14,15]. In another stream of research, automated approaches have been proposed for summarizing legal documents [16,17].

Prior research mainly focuses on the analysis of information published by companies, transaction data or legal documents. Nevertheless, explainability and fairness in the more and more important field of RegTech applications has been under-researched so far [18] and a highly-relevant source of information has not been taken into account yet: information published by consumers regarding their product or service experiences made. Such information can help to identify situations that require further regulatory examination. An automated approach analyzing complaints published by consumers of financial products or services can help regulators to align their supervisory efforts. With this study, we close this research gap.

2.2. Explainable and fair AI

Explainability of artificial intelligence has gained increased importance due to the emphasized role of machine learning in today's more and more computerized environment [19]. Explainability refers to the notion that stakeholders of machine learning models are enabled to understand how and why machine learning models come to a specific decision [20]. With increased explainability, model acceptance is fostered and the model development process is supported [21].

In prior research, different approaches have been proposed to foster explainability. Whereas one strategy is to use model induction techniques that infer explainable models after model building (e.g. post-hoc interpretability), another strategy is to directly alter the model building process so that results are easier to interpret (e.g. intrinsic interpretability) or to use modeling techniques which per se result in more explainable models, such as decision trees [22,23]. Machine learning techniques are associated with differing levels of transparency, whereas an overview is provided by [23] as well as [24].

Prior approaches on explainability enable users and developers to identify *which* features contribute *how* much to a classification decision [25]. Nevertheless, theory-based feature engineering is under-researched so far, although an understanding of *why* a specific feature has an influence on a classification decision is imperative for model acceptance, specifically in a regulatory context. With this study, we contribute to this research stream by deriving features from information diagnosticity theory to explain why they influence complaint success. Thereafter, other well-established technologies for post-hoc explanation could be applied.

Next to model explainability, fairness of classification decisions has gained increased importance as well. Specifically, it is necessary that a model's output does not discriminate certain groups, for instance by race, gender or age – which particularly holds true in the regulatory context [26]. For instance, nonrepresentative training data might influence the model's output and cause discriminatory decisions [27]. Furthermore, certain user groups might be discriminated by the usage of specific features. Consequently, representative training data should be used and features which are uncorrelated with protected groups should be taken into account [28]. In the field of text mining, specific protected user groups might have a differing writing style, which might therefore lead to a potentially biased model output of text-based classifiers. Within our study, we propose an approach based on a specialized model for a protected group in order to overcome this issue potentially causing unfair situations.

2.3. Consumer complaints

Consumer complaints encompass feedback submitted by consumers to companies or to the responsible regulatory authorities. Consumer complaints are supplementary to electronic word-of-mouth, where consumers publish their opinions on online review platforms not specifically targeting the company, but other online consumers [29]. Previous studies investigate complaint behavior, how companies deal with consumer complaints and how consumers react on companies' responses.

Prior research has found that consumer complaints can be triggered by problems with products or services offered, but also by difficulties with the distribution channel, such as the salesperson involved [30]. If consumers are offered the possibility to submit their complaints, this increases consumer satisfaction [31]. Nevertheless, companies have been shown to refrain from inciting customers to complain – despite from its positive consequences [32].

Another stream of research deals with the question how companies address consumer complaints. Here, defensive strategies such as apologizing for potential misbehavior have been shown to be successful [33], whereas customer satisfaction should be a primary goal of company representatives [34]. Nevertheless, some companies rather try to postpone their reactions on consumer complaints and only react when regulatory authorities are involved [35].

Finally, different studies analyze how consumers react on companies' responses. If corporations respond, consumers are more satisfied and their purchase likelihood increases [36]. Additionally, it is very important how the company reacts – and whether a remedy is provided in response to a complaint [37]. Here, refunds have been shown to be more effective than sole apologies not accompanied by a measurable action [38]. Thus, consumer satisfaction can be increased by an efficient complaint handling process [39].

Whereas previous research shows how corporations deal with consumer complaints and outlines the value of an efficient complaint handling, previous research neglects the processing of consumer complaints by regulatory authorities. Apart from increasing customer satisfaction, regulators also face the more critical problem of assigning regulatory resources efficiently and to select the most promising complaints for manual investigation. In prior research, several studies classify user generated content as either positive or negative [40].

Nevertheless, prior studies do not focus on predicting the result of consumer complaints – i.e. the question of whether a complaint is substantial and results in a compensation paid to the consumer. With this study, we address this research gap.

2.4. Consumer complaints in the financial services industry

In the United States, the Consumer Financial Protection Bureau (CFPB) has been founded to supervise companies, enforce consumer financial laws and to propose rules to protect financial consumers [6]. One important function of the CFPB is to collect, investigate and respond to consumer complaints. Here, the CFPB provides a platform for consumers to submit complaints. The CFPB then forwards the submitted messages to the affected companies and also investigates the accusations on its own. Until 2020, the CFPB has received nearly two million complaints. The CFPB provides a database where consumer complaints are publicly accessible so that consumers can inform themselves on financial service providers, the submitted complaints as well as the actions taken.

Previous studies have analyzed whether CFPB regulation has an impact on service provision and has only found weak evidence [41]. Additionally, prior research leverages the consumer complaint database and provides several insights on the complaints submitted [42], such as the topics discussed [43]. Specifically, consumers use the complaint function in order to express anger, frustration, sadness and fear about a company's misbehavior [44].

In contrast, prior research neglects the question of whether the outcome of a consumer complaint – i.e. the fact whether a compensation is paid to the consumer – can be predicted by a RegTech approach. Due to the sheer amount of consumer complaints submitted to the CFPB, such an approach is very helpful to assign regulatory resources more efficiently and to particularly investigate those cases which are deemed to be promising.

3. Designing a RegTech approach for predicting the outcome of consumer complaints

3.1. Research approach

Within this study, we follow the design science research paradigm [7,8]. Our goal is to propose design principles guiding the process to build and evaluate a RegTech approach for predicting the outcome of consumer complaints. Thereby, design principles represent prescriptive knowledge guiding design and which can thereby also be applied within other domains apart from the classification of consumer complaints [45]. For developing the proposed artifact, we follow the research process proposed by [46]. First, we gain awareness of the classification problem by incorporating design requirements (section 3.2). Second, we suggest different design principles to address these design requirements (section 3.3). Third, we develop specific design features which are implemented to address our research problem (section 3.4). Fourth, we evaluate the proposed design principles and features based on different evaluation hypotheses (section 4). Fifth, we discuss our results (section 5).

The design requirements, principles as well as features including their interdependencies are presented in Fig. 1. In the next sections, we specifically motivate and discuss these different elements guiding our artifact development.

3.2. Design requirements

One major requirement for RegTech applications focusing on critical decisions such as the identification of fraudulent behavior or the question of whether a consumer complaint is actually substantial is whether the classification results can be made transparent to the users. In classical machine learning models, classifications are rather regarded as "black boxes", whereas it is unclear how a classification decision is made [47]. Nevertheless, in a regulatory context, the regulator has to justify why a specific course of action is followed. More important, if regulatory sanctions are taken or if a trial is prepared, a regulator has to be able to explain why a specific case is selected for further investigation. Specifically in ethical situations, the willingness to accept classification decisions without such explanations is reduced [4] and non-explainability can hamper trust in RegTech applications [48]. Consequently, RegTech applications have to be as transparent as possible to be able to explain classification results. Consequently, we formulate design requirement 1 (DR1):

DR1: Explainability of Classification Results: A RegTech application has to provide explanations on why a specific classification decision was made.

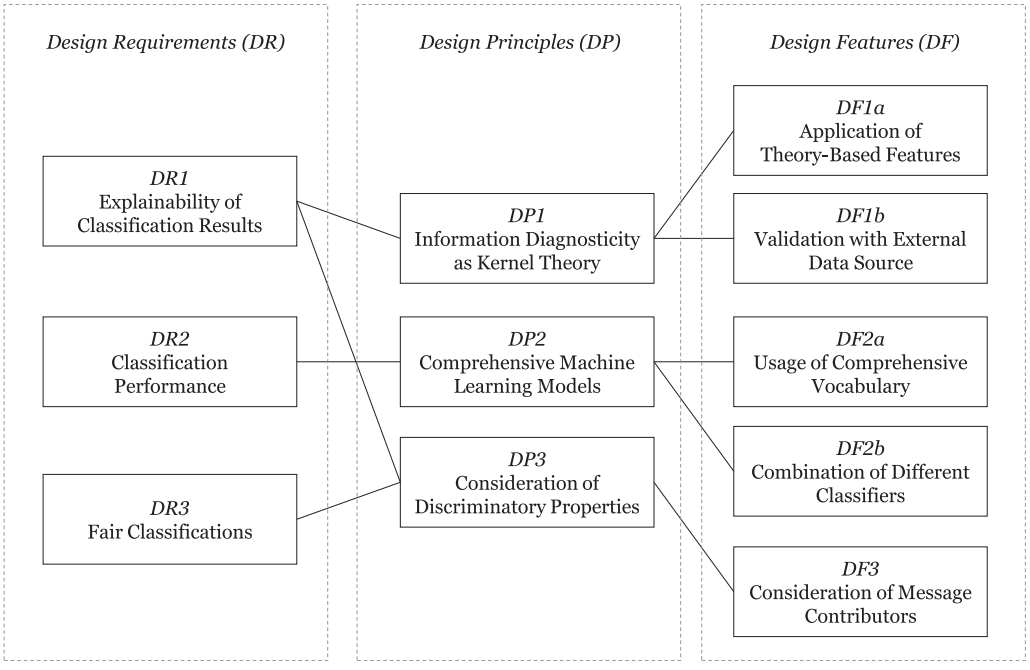

Fig. 1. Associations of Design Requirements, Design Principles and Design Features for an Explainable and Fair RegTech Approach.

Specifically in a regulatory context, classification performance is of high importance for applications based on artificial intelligence. When a system identifies a person as suspect for potentially committing a crime, the related system has be as precise as possible [5]. Furthermore, the proposed system should be as accurate as possible to increase trust and the willingness to consider the system's output [49]. As follows, we formulate design requirement 2 (DR2):

DR2: Classification Performance: A RegTech application has to provide classification decisions with a satisficing classification performance.

Finally, it is important that RegTech applications provide decision support which is fair, i.e. which is not prone to biases and which is not discriminating members of specific groups or populations [5]. Classifiers should not favor or penalize individuals who belong to a certain group of persons [50]. On the one hand, this is necessary to comply with current regulation – and on the other hand, this increases trust towards the systems itself [51]. Consequently, we formulate design requirement 3 (DR3):

DR3: Fair Classifications: A RegTech application hast to provide fair classifications.

3.3. Design principles

To address these design requirements, we propose different design principles. In prior research, different methods have been proposed to enhance the explainability of machine learning models (DR1). Whereas model induction techniques can be used to infer explainable models from black box models, the model building process can be altered as well so that results are easier to interpret, i.e. by using explainable features and to build easy-to-understand models [22].

Consumer complaints are submitted in textual form, so a text mining approach is appropriate for predicting complaint success. In this context, information diagnosticity theory motivates which aspects make texts helpful [52]. We assume that if a text is helpful for its readers, it is also more likely that a consumer complaint is successful, i.e. that it convinces the targeted company to pay a compensation. Thus, for addressing DR1, i.e. to ensure the explainability of the proposed models, we propose that features based on information diagnosticity theory are taken into account for classifier configuration during theory-based feature engineering. In this case, a classifier's output can be explained and justified by directly referring to the linguistic features and their relation to the diagnostic value of texts:

DP1: Information Diagnosticity as Kernel Theory Guiding Classifier Configuration

A focus on machine learning models whose decisions are easy to explain oftentimes results in models whose classification performance is potentially lower compared to traditional machine learning applications that maximize overall classification accuracy [50]. Therefore, there is a tradeoff between explainability and classification accuracy [53].

Nevertheless, since classification accuracy is also particularly important for RegTech applications as formulated in design requirement 2, it is imperative to combine the advantages of easy-to-understand modeling and classical machine learning maximizing classification accuracy. Thus, we formulate as a second design principle (DP2) for explainable and fair RegTech applications that comprehensive machine learning models shall be applied which combine the strength of explainability and performance:

DP2: Application of comprehensive machine learning models

To address the requirement of fair classifications (DR3), we focus on an important aspect underlying most machine learning models: it is imperative to gain a proper domain and data understanding to identify and deal with potential discriminatory situations [54]. In case of consumer complaints submitted to the CFPB, it shall be avoided that specific consumer groups are discriminated by the developed classifiers. Specifically in text mining, certain user groups could be characterized by a specific writing style, which might then bias the predictions of text-based classifiers. We propose that RegTech applications analyzing consumer complaints shall therefore carefully consider such user information to avoid discrimination. Thus, we formulate our third design principle (DP3):

DP3: Consideration of potential discriminatory properties

3.4. Design features

In the following, we propose the design features which guide the implementation of our design principles. To address the design requirement of explainable classification results (DR1), we make use of information diagnosticity theory (DP1) and include features within our models which are directly motivated by this theoretical lens. Such features can be seen as easy to interpret and to understand [49]. Therefore, we assume that consumer complaints which have characteristics making them helpful and informative will also be more likely to result in compensations paid to consumers.

According to information diagnosticity theory, textual aspects influence a text's helpfulness, whereas the product type also has an influence on information diagnosticity. Typically, products can be divided into search goods and experience goods. Search goods can be assessed easily before purchase, whereas experience goods have to be purchased in order to evaluate their quality [55]. Financial services share some search good characteristics (like the number of free transactions in case of a bank account) as well as experience good characteristics (like the quality of investment advisory). Apart from search good qualities, financial services involving personal interaction therefore also have experience good characteristics, for example in the field of investment advice or financing, and are thus worthwhile to explore.

We focus on the following factors (see Table 1). First, we consider review depth, measured by the number of words. Here, an increased depth is assumed to be positively related to text's helpfulness [52]. Nevertheless, if a text is too comprehensive, this positive effect is reduced due to information overload [56]. Therefore, squared review depth is also considered. Furthermore, we take into account the sentiment, measured by the amount of positive and negative words contained in the text [57,58]. Here, we assume that neutral texts are more helpful than extremely positive or negative ones, since financial products or services can be seen as rather non-emotional. As financial products or services have experience good characteristics, personal experiences are important to come to an evaluation. In case of differing evaluations, reviews which are more neutral (and therefore less controversial) are assumed to be perceived as more helpful, specifically in case of products with experience good characteristics [59]. For instance, reviews stating that a bank provides "horrible service" or that a bank is the "worst bank I have ever banked with" will be seen as rather provoking compared to less emotional ones and will thus be less helpful for users with differing perceptions.

Additionally, we consider the temporal focus of the text since prior research has shown that temporal aspects are useful to draw conclusions

Table 1
Variable Operationalization.

	Variable	Definition
Linguistic Aspects	wordcount	Amount of Words of the consumer complaint
	positive_words	Ratio of positive words
	negative_words	Ratio of negative words
	past_words	Ratio of verbs related to the past
	active_words	Ratio of words indicating activity
	uncertain_words	Ratio of words indicating uncertainty
	perceiv_words	Ratio of words indicating perceptions
	know_words	Ratio of words indicating knowledge
Complaint characteristics	(sub-) product_type	(Sub-) Product type
	(sub-)issue	(Sub-) Issue of the complaint

on helpfulness in the field of online review platforms [60]. Here, we assume that a text is more helpful when consumers report on their experiences made in the past (indicated by a consequent writing style), instead of referring to future events. This specifically holds true for consumer opinions which should be backed by prior experiences made instead of mere future expectations.

Regarding the question of how consumers express their evaluation, we also focus on whether consumers follow an active writing style and whether they indicate that they know the situation - or whether they are rather uncertain and express their perceptions. Here, we assume that actively written texts are more helpful. Furthermore, we assume that texts written by consumers who are rather cautious when explaining their experiences (i.e. using words indicating uncertainty or perceptions) are more helpful than texts indicating that consumers specifically intend to know a particular situation and thus potentially overestimate themselves. This specifically applies to financial online reviews, as financial services share search and experience good characteristics. Here, personal experiences and perceptions are important, for instance when assessing the friendliness of an investment advisor. Specifically in case of conflicting perceptions of different individuals, texts which are written in a rather cautious manner will therefore be perceived as more helpful. For instance, a statement such as "perhaps, they have made a mistake" will be more helpful than "but they sure hate us as war veterans".

Therefore, we formulate design feature 1a and construct classifier A, which includes the proposed theory-based linguistic features. These features and the obtained relations to information diagnosticity can help to explain the classifier's output:

DF1a: Application of theory-based features

Previous research focuses on the diagnosticity of online reviews posted on online review platforms such as amazon or yelp. Here, customers can share their experiences made with the products purchased. Nevertheless, these platforms do not enable customers to share their experiences on financial services. As factors influencing review diagnosticity differ for search and experience goods and since financial services share some search and experience good characteristics, it has to be validated which factors are also applicable in the financial context.

As a consequence, we validate the features making online reviews helpful in the financial context using reviews on financial services. Therefore, we have acquired a dataset of online reviews posted on creditkarma.com and first investigate which linguistic features make those online reviews helpful. This data source is appropriate as online review platforms typically provide the opportunity to assess whether an online review is perceived as helpful or not. In the second step, we use these features for predicting the success of consumer complaints. Due to the rationale that the observed factors make online reviews helpful to the reader for making decisions, we assume that they will also be helpful for predicting complaint success and can serve as explanations for a model's output. This is also advantageous as consumer complaints are typically not accompanied by a possibility to evaluate their diagnosticity for decision making, and considering online reviews overcomes this shortcoming. Furthermore, these factors theoretically linked with diagnosticity foster understandability of the model. Consequently, we formulate design feature 1b:

DF1b: Validation with external data source

Although linguistic features capture an important part of the contents expressed within a text, a bag-of-words model using the comprehensive vocabulary of a text can perform better than models solely based on linguistic features. Therefore, we also build a bag-of-words model based on the consumer complaints within our sample [61]. For that purpose, the term-document matrix is constructed using the different words of the text. Based on the bag-of-words model, we build classifier B and formulate design feature 2a:

DF2a: Usage of comprehensive vocabulary

To further address the design principle to apply comprehensive machine learning models, we also consider ensemble learning to combine the outputs of different classifiers. Previous research has shown that a combination of different models can have a positive impact on classification performance, especially if the individual classifiers disagree [62]. As a consequence, we construct classifier C as an ensemble classifier taking into account the results of classifiers A and B, i.e. the predicted classes as well as the predicted probabilities. We consequently formulate design feature 2b:

DF2b: Combination of different classifiers

When making automated classifications, it is imperative that the proposed classifier is not discriminating certain groups or populations [5]. Since the proposed classifiers take into account texts of consumer complaints, it has to be made sure that the training data (i.e. the texts analyzed) do not lead to biased decisions [50].

When focusing on texts, it can be assumed that specific user groups have differing experiences contributing online contents. More specifically, it can be assumed that older persons have a lower proficiency with information systems and are on average less capable of contributing online consumer complaints. It can also be assumed that their contributions are less suitable to be automatically assessed by the classifiers proposed. Thus, the message contributor has to be considered and a more profound manual investigation is necessary. Furthermore, this should also result in a specialized classifier trained for this user group. We therefore formulate design feature 3:

DF3: Consideration of message contributors

4. Evaluation

4.1. Evaluation design

For evaluating our design principles and features as well as the resulting classifiers, we follow a structured data mining process as proposed by [54], consisting of the steps *data selection*, *data preprocessing and transformation*, *data mining* and *evaluation.*

During *data selection, we* first select appropriate datasets: (1) to evaluate the features based on information diagnosticity theory, we acquire a random sample of 1500 online reviews on financial products published on creditkarma.com. On this online platform, consumers can report their experiences on financial products or services. (2) to evaluate the proposed classifiers A to C (see Table 2), we acquire a random balanced sample of 6000 consumer complaints from the CFPB consumer complaint database (3000 complaints resulting in a compensation paid, 3000 not resulting in a compensation paid). We only select those consumer complaints whose texts are published so that they can be processed by our classifiers. (3) to evaluate the proposed classifiers using the natural class distribution, we also acquire a random hold out sample consisting of 75,000 consumer complaints (3000 resulting in monetary compensations, 72,000 not resulting in monetary compensations).

We *preprocess* and *transform* the data in several steps [61]. We tokenize the texts to obtain the different terms. We use a stop word filter and a stemmer to reduce the feature set. To further grasp the textual aspects, we also generate n-grams [63]. Additionally, using the word

Table 2
Summary of Classifier Configuration.

Classifier	Configuration	Description
A	Features based on Information Diagnosticity Theory	Classifier taking into account topic as well as product type and factors influencing information diagnosticity according to our analysis of financial product reviews.
B	Features based on Bag-of-Words Model	Classifier based on bag-of-words model considering the texts' words as features.
C	Ensemble of Classifier B and C	Ensemble classifier based on bagging taking into account the predicted classes as well as the probabilities for each class according to classifier A and B.

lists of the General Inquirer [64], we determine the ratios of specific word categories as presented in Table 1. For constructing the bag-of-words model which is necessary for classifier B, we use the tf-idf measure to determine the feature weights [61].

For *data mining*, we apply several machine learning techniques [65]: we evaluate classification performance using naïve bayes, neural networks, decision tree, random forest, as well as support vector machine (SVM). Finally, to evaluate the classifiers based on the balanced sample, we apply ten-fold cross validation [66], and calculate the performance measures accuracy, precision, recall and F1 [61,65].

We validate and *evaluate* the proposed design principles and the resulting classifiers based on the three datasets (see Fig. 2). Therefore, we propose a set of *evaluation hypotheses* which are described in section 4.2. We then validate the theory-based features (see section 4.3) and verify the evaluation hypotheses for the proposed classifiers (see section 4.3, 4.4 and 4.5). Finally, it is also of high importance to provide a *practical evaluation* of our RegTech approach [67]. Therefore, we evaluate the best-performing classifier based on the random hold-out sample with a natural class distribution (see section 4.6).

4.2. Evaluation hypotheses

In the following, we develop our evaluation hypotheses for the proposed design principles and features. Specifically, we first evaluate whether information diagnosticity theory as kernel theory guiding explainable classifier configuration is actually useful for predicting the success of consumer complaints (DP1) by hypothesizing on the predictive value of classifier A. Thereafter, we hypothesize on the additional contribution of comprehensive classifier configurations (DP2) by specifically focusing on the performance of classifier B and C in comparison with classifier A. Finally, we examine whether author characteristics influence classification results and might therefore influence the fairness of the classification approach (DP3).

According to information diagnosticity theory, aspects such as a text's depth or its valence influence its helpfulness [52]. In case of consumer complaints, companies have to decide on how to deal with a specific complaint. If the complaint seems clear, is actually well-justified and diagnostic, a company will be more likely to pay a compensation. We assume that features resembling information diagnosticity are thus valuable for predicting the success of a consumer complaint (DP1). Furthermore, as such features are directly motivated by information diagnosticity theory, they can also be used to explain a classifier's decision. Thus, we formulate evaluation hypothesis H1:

H1: A classifier considering features explained by information diagnosticity theory has predictive value.

Furthermore, we hypothesize on the value of comprehensive machine learning models (DP2). Here, previous research has shown that classifiers based on a bag-of-words approach can outperform classifiers based on linguistic features alone [49]. In addition, ensemble learning can also add predictive value [68]. When the output of different classifiers is combined, local disadvantages can be overcome and the classification results can be improved [69]. Consequently, we hypothesize that comprehensive machine learning models outperform a model solely based on explainable features:

H2: A comprehensive model outperforms a model based on features explained by information diagnosticity theory alone.

Finally, we also hypothesize on the impact of potential discriminatory properties influencing the fairness of classification results (DP3). As the proposed classifiers specifically consider textual aspects of the complaints, the writing style can be assumed to have an influence on classification performance. The complaints submitted to the CFPB can be entered using a web interface. Here, "digital natives", which are most probably relatively young persons, are already used to the way of communicating on online platforms, for instance by means of online reviews. In contrast, older adults typically show a lower technical addiction, which might also influence their writing style, the resulting complaint success and classification performance. Based on this reasoning, we hypothesize that author characteristics influence classification performance. For providing a fair assessment of consumer complaints, regulators then have to put more effort into manual analyses or introduce specific classifiers for analyzing complaints which bear a risk of leading to discriminated classifications:

H3: Author characteristics have an influence on classification performance.

4.3. Information diagnosticity as kernel theory guiding classifier configuration

At first, we evaluate the features motivated by information diagnosticity theory. Therefore, we analyze the sample of online reviews related to financial products and services published at creditkarma. Based on this analysis, we ensure that the features can be used for explaining the classification results. As a second step. we construct classifier A, where we only take into account those features which significantly influence review helpfulness. This underlines the explainability of classification results.

For the first step, we acquire 1500 online reviews on financial products and services which have received at least five votes on review

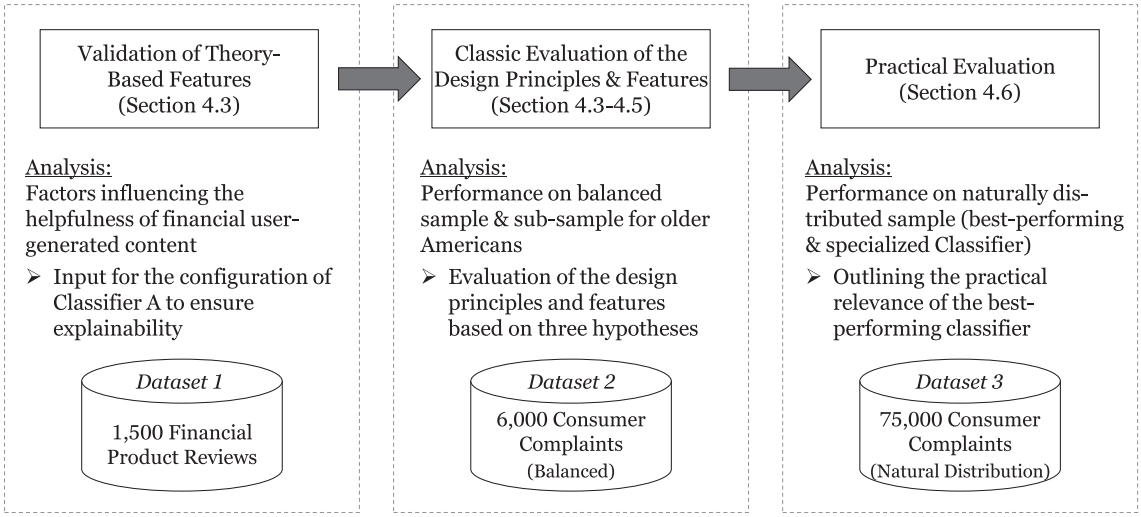

Fig. 2. Evaluation Process and Datasets Used.

helpfulness, so review helpfulness can be measured based on a sufficient amount of votes. For each review, we determine the linguistic features as described above. Table 3 provides the results of our tobit regression analysis explaining the percentage of review helpfulness with the different linguistic features. Thereby, we find significant influences of most of the variables: We confirm that review depth has a significant positive influence on review helpfulness, whereas we also find a negative effect if reviews are too comprehensive causing information overload. Furthermore, we observe a significant negative influence of positive sentiment on review helpfulness, whereas negative sentiment has no influence. Thus, neutral (or at least not too positive) online reviews can be regarded as more helpful.

Regarding the temporal focus of the reviews, we observe that texts are more helpful if they refer to experiences made within the past. Additionally, we also find that an active writing style has a significant positive influence on review helpfulness. Finally, texts are more helpful if they are written more cautiously (expressing uncertainty or perceptions). In contrast, there is a significant negative effect if consumers overemphasize their knowledge. Furthermore, we control for the overall evaluation of the product or service (the amount of stars) and find a significant positive effect. The amount of votes received has no significant influence.

At the second step, we now focus on consumer complaints and observe that successful consumer complaints are longer, less positive and more related to the past. Furthermore, their authors less often emphasize their knowledge and more often express perceptions, which is in line with our results from the analysis of online reviews. In combination with the complaint characteristics, we thus include the factors significantly influencing the diagnosticity of financial online reviews when constructing classifier A for predicting whether a consumer complaint results in a financial compensation or not. For that purpose, we build our predictive models taking into account different machine learning techniques. The results using our balanced sample of consumer complaints are presented in Table 4.

Considering the results of classifier A, we observe that for each machine learning technique, an accuracy of more than 80.00% is achieved. Related to the performance of both classes, we observe that the F1-scores are higher for those cases where a monetary compensation is paid. Focusing on precision and recall, we find that the precision when identifying cases with no monetary compensation is higher, whereas recall is better for cases with monetary compensation. Overall, a classifier based on naïve bayes performs best, whereas we obtain an accuracy of 84.03%. The F1-scores are also promising, classifier A results in a F1-scores of 83.24% (no monetary compensation) and 84.76% (monetary compensation).

Related to research hypothesis 1, we can confirm that classifiers based on explainable features are valuable to predict the outcome of consumer complaints. A naïve classifier assigning each case to the same class would result in an accuracy of 50.00%. This is significantly outperformed by classifier A (accuracy: 84.03%), which is also confirmed by McNemar's test (presented in Table 6).

Table 3
Drivers of Financial Review Helpfulness based on Tobit Regression (***/**/*: $p < 1\%/5\%/10\%$, $n = 1500$).

	Variable	Estimate	*P*-Value	
Linguistic Aspects	wordcount	0.0004	<0.01	***
	wordcount * wordcount	0.0000	0.01	**
	positive_words	−0.3157	<0.01	***
	negative_words	−0.0812	0.64	
	past_words	0.1179	<0.01	***
	active_words	0.1958	0.08	*
	uncertain_words	0.7590	<0.01	***
	perceiv_words	0.6033	0.06	*
	know_words	−0.6574	<0.01	***
Control Variables	stars	0.0144	<0.01	***
	no_total	0.0001	0.17	
	Intercept	0.7172	<0.01	***
	Pseudo R^2	0.2357		

4.4. Application of comprehensive machine learning models

As a next step, we evaluate classifier B and C, both representing comprehensive machine learning models. Our results using the balanced sample based on 10-fold cross-validation are presented in Table 5.

Focusing on classifier B, which is based on a bag-of-words model taking into account the different terms of the consumer complaints, we find promising results. Except from the classifier based on a decision tree, the accuracy is again above 80.00%. Here, classifier B based on support vector machine performs best. We obtain an accuracy of 85.12% and favorable F1-scores (84.65% for class "no monetary compensation", 85.55% for class "monetary compensation"). Compared to classifier A, this is an increase in accuracy of more than one percentage point.

Classifier C is an ensemble based on bagging considering the predicted classes as well as confidences of the best performing classifiers A and B. For classifier C, we also find an increase in classification performance compared to the other classifiers. Here, the configuration based on neural network performs best, resulting in an accuracy of 85.92% and F1-scores of 85.18% and 86.58% respectively.

To evaluate hypothesis 2, we investigate whether classifiers based on comprehensive machine learning techniques (i.e. classifier B and C) outperform a classifier solely based on features resulting from information diagnosticity theory (i.e. classifier A). Next to comparing the performance figures as outlined above, we also conduct McNemar's test on classifier performance (see Table 6). Here, we find a significant difference both for classifier B and C, whereas the significance level is higher for classifier C ($p < 0.01$) when compared to classifier B ($p = 0.02$). Consequently, research hypothesis H2 is supported as well.

4.5. Consideration of potential discriminatory properties

To investigate whether the results are influenced by the message contributor, we consider if the consumer complaint was contributed by a user who is 62 years of age or older (labeled by the CFPB as "older American") or not. We first investigate whether both groups differ in writing style, whereas the results are presented in Table 7. Related to the linguistic features, we find significant differences: for instance, older Americans publish longer and less positive consumer complaints. Thus, a different writing style is observed, which therefore might also influence classification performance.

Next, we report the results of the best performing classifier differentiated by the two subgroups "older American" and "regular user" (Table 8). Related to accuracy, we consistently observe that classifier A, B, and C perform better for regular users when compared to older Americans. In case of classifier C, the difference in accuracy is almost 6 percentage points. Furthermore, the precision for the class "no monetary compensation" is substantially reduced (i.e. almost 9 percentage points for classifier C), whereas precision for classifier C for the class "monetary compensation" decreases by about 2 percentage points. Specifically, the decrease in classification for the class "no monetary compensation" causes unfair situations: a regulator will refrain from focusing on cases which are predicted to not result in monetary compensations. Thus, complaints contributed by older users will more often be misclassified and therefore disregarded. In other words, if a regulator relies on the classifier's predictions, the cases of older Americans are more often not dealt with correctly. Consequently, evaluation hypothesis H3 is confirmed.

To overcome this weakness, we now specifically train a classifier for a balanced sub sample of complaints submitted by older Americans only. The corresponding results are presented in Table 9. As shown by the results, the observed accuracy is slightly worse than the accuracy observed for classifiers A, B, and C. Nevertheless, we find an increase in

Table 4
Evaluation of Classifier A.

				No Monetary Compensation			Monetary Compensation		
Classifier	Configuration	Algorithm	Accuracy	Prec.	Recall	F1	Prec.	Recall	F1
A	Features based on Information Diagnosticity Theory	Naïve Bayes	84.03	87.59	79.30	83.24	81.09	88.77	84.76
		Neural Network	82.50	82.61	82.33	82.47	82.39	82.67	82.53
		Random Forest	81.48	91.01	69.87	79.05	75.55	93.10	83.41
		Decision Tree	82.05	92.19	70.03	79.60	75.84	94.07	83.98
		SVM	83.92	87.23	79.47	83.17	81.14	88.37	84.60

Metrics are based on stratified 10-fold cross-validation.

Table 5
Evaluation of Classifier B and C.

				No Monetary Compensation			Monetary Compensation		
Classifier	Configuration	Algorithm	Accuracy	Prec.	Recall	F1	Prec.	Recall	F1
B	Features based on Bag-of-Words Model	Naïve Bayes	82.08	86.53	76.00	80.92	78.60	88.17	83.11
		Neural Network	82.83	85.00	79.73	82.28	80.92	85.93	83.35
		Random Forest	80.32	86.54	71.80	78.48	75.90	88.83	81.86
		Decision Tree	68.20	90.44	40.70	56.14	61.74	95.70	75.06
		SVM	85.12	87.37	82.10	84.65	83.12	88.13	85.55
C	Ensemble of Classifier A and B	Naïve Bayes	84.63	85.17	83.87	84.52	84.11	85.40	84.75
		Neural Network	85.92	89.86	80.97	85.18	82.68	90.87	86.58
		Random Forest	85.70	89.40	81.00	84.99	82.63	90.40	86.34
		Decision Tree	85.65	88.91	81.47	85.03	82.90	89.83	86.23
		SVM	85.80	89.78	80.80	85.05	82.55	90.80	86.48

Metrics are based on stratified 10-fold cross-validation.

Table 6
McNemar's Test Results on Classifier Performance (***/**/*: $p < 1\%/5\%/10\%$).

Classifier	Configuration	vs. All "No Monetary Compensation"	vs. All "Monetary Compensation"	vs. Classifier A
A	Features based on Information Diagnosticity Theory	<0.01***	<0.01***	–
B	Features based on Bag-of-Words Model	<0.01***	<0.01***	0.02**
C	Ensemble of Classifier A and B	<0.01***	<0.01***	<0.01***

Table 7
Descriptive Statistics: Linguistic Features.

	Regular User		Older American		
Variable	Mean	Median	Mean	Median	*P*-Value
wordcount	220.26	153.00	234.63	182.00	<0.01***
positive_words	0.1022	0.0980	0.0939	0.0930	<0.01***
negative_words	0.0566	0.0532	0.0579	0.0541	0.28
past_words	0.4577	0.4717	0.4919	0.5000	<0.01***
active_words	0.1654	0.1628	0.1556	0.1540	<0.01***
uncertain_words	0.0274	0.0238	0.0225	0.0202	<0.01***
perceiv_words	0.0115	0.0095	0.0129	0.0110	<0.01***
know_words	0.0463	0.0426	0.0375	0.0355	<0.01***

Wilcoxon Rank-Sum Tests for Equality of Medians (***/**/*: $p < 1\%/5\%/10\%$).

precision for the class "no monetary compensation" by up to 5% and thus an increased level of fairness.

4.6. Practical evaluation

The analyses presented above are based on a balanced sample taking into account the same amount of consumer complaints resulting in monetary compensations and not resulting in monetary compensations. This ensures proper classifier training due to oversampling. Nevertheless, it is also important to evaluate the classification performance based on the real class distribution [15], since only about 4% of all consumer complaints actually result in monetary compensations. For that purpose, we evaluate the best performing classifier (i.e. the best "regular" classifier as well as the fair classifier for older Americans) using an additional hold-out sample with a real class distribution. The evaluation results are presented in Table 10,

For the regular classifier, we obtain an accuracy of 80.51%, which is below the original accuracy of classifier C (85.92%). Furthermore, when considering the performance regarding both classes, we observe that cases resulting in no monetary compensation are identified with very high precision, whereas the precision for cases resulting in monetary compensation is much lower. For recall, we find contrary results, but the F1-score is considerably higher for consumer complaints which do not result in monetary compensations. For the fair classifier, we find comparable results, whereas overall accuracy is worse, but precision for "no monetary compensation" is comparable.

At a first sight, this shows that the proposed classifiers perform not as good as expected when compared to the results obtained using the balanced sample. Nevertheless, specifically the lift chart presented in Fig. 3 outlines the high practical relevance of both classifiers, also when applied using the unbalanced hold-out sample.

For generating the lift chart, we first classify each case in the hold-out sample. Then, we sort the cases by the classifier's prediction confidence. The lift chart shows how many complaints resulting in monetary compensations are selected when a specific amount of complaints having the highest confidences are taken into account. The lift chart shows that if only 1.00% of all complaints are selected, about 13.00% of all complaints resulting in monetary compensations are included. If 10.00% of the cases are taken into account, already more than 65.00% of all cases resulting in monetary compensations are covered. For the fair classifier trained on older Americans, we find comparable, but slightly worse results. This outlines the practical relevance of the proposed classifiers: a regulator can use the classifiers to reduce the amount of cases to be

Table 8
Classifier Evaluation, by Author (Best Performing Classifier).

				No Monetary Compensation			Monetary Compensation		
Classifier	Configuration	Author	Accuracy	Prec.	Recall	F1	Prec.	Recall	F1
A	Features based on Information Diagnosticity Theory	Regular User	84.48	88.14	80.34	84.06	81.30	88.77	84.87
		Older American	78.00	75.83	59.87	66.91	78.89	88.72	83.52
B	Features based on Bag-of-Words Model	Regular User	85.48	87.82	83.01	85.35	83.30	88.04	85.60
		Older American	80.20	77.95	65.13	70.97	81.21	89.11	84.98
C	Ensemble of Classifier A and B	Regular User	86.32	90.23	82.02	85.93	82.94	90.78	86.68
		Older American	80.44	81.58	61.18	69.92	80.00	91.83	85.51

Metrics are based on stratified 10-fold cross-validation.

Table 9
Classifier Evaluation, Classifier Specifically Trained for older Americans.

			No Monetary Compensation			Monetary Compensation		
Classifier	Configuration	Accuracy	Prec.	Recall	F1	Prec.	Recall	F1
A_{OLD}	Features based on Information Diagnosticity Theory	77.05	78.87	73.90	76.30	75.45	80.20	77.75
B_{OLD}	Features based on Bag-of-Words Model	79.70	83.07	74.60	78.61	76.95	84.80	80.68
C_{OLD}	Ensemble of Classifier A and B	79.35	80.99	76.60	78.73	77.87	82.00	79.88

Metrics are based on stratified 10-fold cross-validation.

Table 10
Best Performing Classifiers Evaluated with Unbalanced Hold-out Sample.

		No Monetary Compensation			Monetary Compensation		
	Accuracy	Prec.	Recall	F1	Prec.	Recall	F1
Regular Classifier	80.51	99.54	80.07	88.75	15.99	91.03	27.20
Fair Classifier for Older Americans	73.80	99.53	73.51	84.56	7.28	85.71	13.42

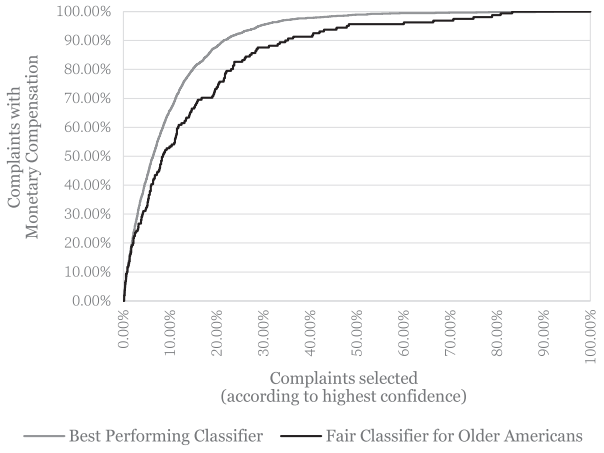

Fig. 3. Lift Chart (Best Performing Classifier, Unbalanced Holdout Sample).

analyzed manually, whereas the complaints selected first are very likely to result in monetary compensations.

5. Discussion

5.1. Design principles for predicting the success of consumer complaints

Based on the design requirements that a RegTech approach for regulatory enforcement has to provide explainable classifications, ensure proper classification performance and should not discriminate certain user groups, we propose and evaluate several design principles and features. Based on our evaluation, we find that features building upon information diagnosticity theory as kernel theory should be used, comprehensive machine learning models should be applied and that discriminatory properties should be taken into account to be able to mitigate them.

As shown by our analysis of theory-based features, we find that factors driving the helpfulness of online reviews in the financial field are also valuable to predict the outcome of consumer complaints. Due to the theoretical linkage of these features to the diagnosticity of texts, a good explanation can be provided why a model results in a specific recommendation. Here, drivers of diagnosticity can be highlighted and explanations can be given to the model's users. Our evaluation of the features based on financial online reviews provides additional justification for these theoretical explanations.

Overall, the application of comprehensive machine learning models can increase the performance and can thus improve the overall applicability of the proposed classifiers. Classifier B, which is based on a bag-of-words model, as well as classifier C, which is an ensemble classifier taking into account the classifications of classifier A and B both show promising results. The best-performing model based on ensemble learning combines the advantages of high explainability and high classification performance.

Focusing on the consumer complaint's author, we outline that applying the proposed classifiers can also lead to potentially unfair situations. Whereas the proposed classifiers perform well in most of the cases, classification performance decreases if consumer complaints are contributed by older Americans. One explanation can be that this group of persons less frequently uses online platforms and is thus less experienced in contributing appropriate texts. To ensure that this group is treated fairly, regulatory authorities should be specifically cautious when dealing with the related complaints and apply a specialized classifier for older Americans. Thus, an increased understanding of the model can lead to a reduction of negative consequences and to increased fairness.

Finally, for proper classifier training, a balanced sample is taken into account. As shown by our additional evaluation that focuses on a hold-out sample with a real-world class distribution, the proposed classifiers also have important practical value. Although the performance figures

for identifying consumer complaints resulting in monetary compensations decrease, the lift chart shows that the proposed classifiers are very helpful when regulators have to identify a specific amount of consumer complaints for manual investigation. Thus, the proposed classifiers are of high practical relevance in the context of limited regulatory resources.

5.2. Implications for research

Our study has several implications for research. Most important, we propose design principles and features for a RegTech approach predicting the success of consumer complaints. Within this context, we provide a knowledge contribution in form of a known solution transferred to a new problem i.e. the classification of financial consumer complaints by means of a RegTech approach (exaptation research, [70]).

We specifically address the design requirements of explainability, classification performance as well as fairness. Within this context, we outline the relevance of features based on information diagnosticity theory, the application of comprehensive machine learning models as well as the consideration of a complaint's author. Our design principles can therefore serve as building blocks for information systems analyzing consumer complaints or other user generated content. We extend prior research in the field of RegTech by particularly focusing on explainability and fairness of the machine learning models used. Here, we outline how classifier output can be made more understandable so that users' decisions based on the model can be justified. We therefore also provide a new solution for a known problem – i.e. ensuring explainable classifications (improvement research, [70]).

Additionally, by focusing on the consumer complaints' authors, we emphasize how the fairness of the approach can be improved. Nevertheless, our results also show that increased fairness is accompanied by decreased overall classifier performance. We also demonstrate that information diagnosticity theory is applicable in the financial field, whereas textual aspects making user-generated content helpful are also useful for predicting the success of consumer complaints published. We specifically show the relevance of a text's depth, the sentiment expressed, its temporal focus and its cautious writing style. We therefore also extend the research on review diagnosticity, by specifically focusing on financial services which share search and experience good characteristics. Thus, we broaden the research on the search and experience good dichotomy.

Finally, previous research analyzing consumer complaints has focused on contents discussed within and corporate responses towards consumer complaints, but has neglected the prediction of complaint success. With this study, we close this research gap.

5.3. Implications for practice

Our study has important practical implications. First, the proposed RegTech approach can be applied by regulatory authorities to help identifying consumer complaints which should be further investigated. Here, regulators can assign their resources more efficiently for monitoring financial market participants. Moreover, regulatory decisions based on the proposed model are more transparent as linguistic features based on information diagnosticity theory are used.

Second, the proposed approach can also be applied by companies in the financial services industry to monitor consumer feedback. Critical complaints which lead to compensations paid to consumers can be identified, whereas handling of such complaints can be prioritized so that customer satisfaction increases.

Third, the proposed RegTech approach might also be used to show consumers whether their complaints have the potential to be successful or not. If a complaint is deemed to be unsuccessful, consumers can be advised based on the linguistic features which aspects of the text have to be improved to make it more diagnostic and thus to increase the probability of success.

Finally, the approach is also valuable for consumer groups which would normally not be treated equally by the machine learning model. In our case, older Americans have the chance that their complaints are treated fairly by persons using the proposed RegTech approach, since they are aware of a potential influence of the user group submitting a complaint.

5.4. Limitations

We are aware that our study has several limitations. First of all, we analyze consumer complaints available in a public database provided by the CFPB. To be included within the study (and to be suitable to be analyzed by a text mining approach), consumer complaints have to be provided in full text. Consequently, one might question whether the results of this study also hold true for those complaints which are not publicly disclosed in full text. Nevertheless, as the database is very comprehensive and contains hundreds of thousands of consumer complaints, we argue that the complaints analyzed can be seen as rather representative. If access to consumer complaints not disclosed in full text was possible, future research could also apply and evaluate the proposed design principles for such consumer complaints.

Additionally, our RegTech approach can be used to predict whether a complaint actually leads to a monetary compensation paid to the consumer or not. However, in those cases where no compensation is paid, other corporate reactions are possible. These might include no response at all, denial, or providing an excuse but no payment. We are aware that our approach cannot differentiate between these possible courses of action. Nevertheless, from a regulatory perspective, our approach answers the most important question: whether a complaint leads to a monetary compensation paid to a consumer or not. Additionally, the response to complaints might also differ in other domains. Whereas monetary compensations are common in case of financial consumer complaints, physical compensations might be widespread in other contexts. This might therefore require a specific training of the proposed approach. Nevertheless, the training procedures of the different classifiers, as well as the application of theory-based features are independent from the application domain and can therefore also be applied in other contexts.

Finally, to investigate the fairness of the proposed RegTech approach, we focus on the author submitting a complaint. From the dataset analyzed, we extract the information whether a complaint was submitted by an Older American or by a "regular" user. Consequently, the differentiation of author groups is rather broad. Nevertheless, it is still valuable to provide indications how the author group influences classification performance – and, more important – how a potential discriminating influence can be mitigated.

5.5. Future research

Our study also provides several avenues for future research. Our design principles are evaluated using consumer complaints from the financial field. Future research might investigate the applicability of our approach in other domains. As already outlined in the limitations section, we predict the most important reaction to a consumer complaint, i.e. whether a monetary compensation is paid to the consumer or not. However, other reactions on consumer complaints are possible, which might therefore also be examined in future studies.

Related to document processing, more pronounced techniques as well as additional features can be taken into account to evaluate whether the performance of the obtained machine learning models can further be improved. Specifically, work from the rich research stream on opinion mining could be considered. Finally, our study uses information publicly available from the CFPB database. Other data sources might be analyzed which provide more fine-grained author information. Apart from that, if a dataset including a regulatory authority's reaction is available in future, such information can be analyzed as well.

6. Conclusion

Our study shows that the design principles of taking into account information diagnosticity theory as a kernel theory for classifier configuration, applying comprehensive machine learning models, and considering potentially discriminating properties are valuable for the development of explainable, well-performing and fair RegTech applications. We extend previous research by addressing explainability and fairness in the field of RegTech and by predicting the outcome of financial consumer complaints.

As shown by our practical evaluation, the proposed approach is specifically valuable for regulatory authorities, who can concentrate on those consumer complaints with a high probability of success for manual analysis and thus allocate resources more efficiently. Furthermore, financial service providers and consumers can also profit from the proposed approach. Apart from this application in the financial domain, future research might focus on the applicability of the proposed design principles and features in additional fields.

Author information

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*.

References

[1] D.W. Arner, J. Barberis, R.P. Buckey, FinTech, RegTech, and the reconceptualization of financial regulation, Nw. J. Int'l L. & Bus. 37 (2016) 371.
[2] N.G. Packin, RegTech, compliance and technology judgment rule, Chi.-Kent L. Rev. 93 (2018) 193.
[3] W.L. Currie, D.P. Gozman, J.J.M. Seddon, Dialectic tensions in the financial markets: a longitudinal study of pre-and post-crisis regulatory technology, J. Inf. Technol. 33 (2018) 304–325.
[4] D. Doran, S. Schulz, T.R. Besold, What does explainable AI really mean? A new conceptualization of perspectives, in: Working Paper, 2017.
[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, in: Working Paper, 2019.
[6] CFPB, Consumer Response Annual Report, 2019.
[7] A. Hevner, S. March, J. Park, Design science in information systems research, MIS Q. 28 (2004) 75–105.
[8] K. Peffers, T. Tuunanen, M. Rothenberger, S. Chatterjee, A design science research methodology for information systems research, J. Manag. Inf. Syst. 24 (2007) 45–77.
[9] D. Gozman, W. Currie, The role of investment management systems in regulatory compliance: a post-financial crisis study of displacement mechanisms, J. Inf. Technol. 29 (2014) 44–58.
[10] I. Anagnostopoulos, Fintech and regtech: impact on regulators and banks, J. Econ. Bus. 100 (2018) 7–25.
[11] T. Butler, Towards a standards-based technology architecture for RegTech, J. Fin. Transformat. 45 (2017) 49–59.
[12] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature, Decis. Support. Syst. 50 (2011) 559–569, https://doi.org/10.1016/j.dss.2010.08.006.
[13] J. West, M. Bhattacharya, Intelligent financial fraud detection: a comprehensive review, Comp. Security 57 (2016) 47–66.
[14] M. Siering, B. Clapham, O. Engel, P. Gomber, A taxonomy of financial market manipulations: establishing trust and market integrity in the financialized economy through automated fraud detection, J. Inf. Technol. 32 (2017) 251–269, https://doi.org/10.1057/s41265-016-0029-z.
[15] J. Lausen, B. Clapham, M. Siering, P. Gomber, Who is the next "wolf of wall street"? Detection of financial intermediary misconduct, J. Assoc. Inf. Syst. 21 (2020).
[16] Atefeh Farzindar, Guy Lapalme, Legal text summarization by exploration of the thematic structure and argumentative roles, Text Summarization Branches Out (2004) 27–34.
[17] A. Kanapala, S. Pal, R. Pamula, Text summarization from legal documents: a survey, Artif. Intell. Rev. 51 (2019) 371–402.
[18] J.W. Williams, Regulatory technologies, risky subjects, and financial boundaries: governing 'fraud' in the financial markets, Acc. Organ. Soc. 38 (2013) 544–558, https://doi.org/10.1016/j.aos.2012.08.001.
[19] I. Sample, Computer says no: why making AIs fair, accountableand transparent is crucial, The Guardian, 2017.
[20] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. Moura, P. Eckersley, Explainable Machine Learning in Deployment, Proceedings of the Conference on Fairness, Accountability, and Transparency, 2020.
[21] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Commun. ACM 63 (2019) 68–77.
[22] D. Gunning, Explainable artificial intelligence (xai), in: Defense Advanced Research Projects Agency, DARPA/I2O, 2017.
[23] V. Belle, I. Papantonis, Principles and practice of explainable machine learning, Front. Big Data 39 (2021).
[24] C. Molnar, Interpretable Machine Learning, 2020.
[25] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proc. Natl. Acad. Sci. 116 (2019) 22071–22080.
[26] S. Hod, K. Chagal-Feferkorn, N. Elkin-Koren, A. Gal, Data science meets law, Commun. ACM 65 (2022) 35–39.
[27] B. Cheatham, K. Javanmardian, H. Samandari, Confronting the risks of artificial intelligence, McKinsey Quarterly (2019).
[28] P.T. Kim, Addressing algorithmic discrimination, Commun. ACM 65 (2021) 25–27.
[29] D. Halstead, Negative word of mouth: substitute for or supplement to consumer complaints? J. Consumer Satisfact. Dissatisfact. Complain. Behav. 15 (2002) 1.
[30] D.T. McAlister, R.C. Erffmeyer, A content analysis of outcomes and responsibilities for consumer complaints to third-party organizations, J. Bus. Res. 56 (2003) 341–351.
[31] P.U. Nyer, An investigation into whether complaining can cause increased consumer satisfaction, J. Consum. Mark. 17 (2000) 9–19.
[32] T. Hansen, R. Wilke, J. Zaichkowsky, Managing consumer complaints: differences and similarities among heterogeneous retailers, Int. J. Retail Distrib. Manag. 38 (2010) 6–23.
[33] Gary L. Clark, Peter F. Kaminski, David R. Rink, Consumer complaints: advice on how companies should respond based on an empirical study, J. Consum. Mark. 9 (1992) 5–14, https://doi.org/10.1108/07363769210035189.
[34] A.J. Resnik, R.R. Harmon, Consumer complaints and managerial response: a holistic approach, J. Mark. 47 (1983) 86–97.
[35] A.L. Ryngelblum, N.W.H. Vianna, C.A. Rimoli, The ways companies really answer consumer complaints, Mark. Intell. Plan. 31 (2013) 54–71.
[36] J. Strauss, D.J. Hill, Consumer complaints by e-mail: an exploratory investigation of corporate responses and customer reactions, J. Int. Mark. 15 (2001) 63–73.
[37] A.M. Susskind, A content analysis of consumer complaints, remedies, and repatronage intentions regarding dissatisfying service experiences, J. Hosp. Tour. Res. 29 (2005) 150–169.
[38] C. Goodwin, I. Ross, Consumer evaluations of responses to complaints: What' s fair and why, J. Serv. Mark. 7 (1990) 39–47.
[39] J.M. Hogarth, M. English, M. Sharma, Consumer complaints and third parties: determinants of consumer satisfaction with complaint resolution efforts, J. Consumer Satisfact. Dissatisfact. Complain. Behav. 14 (2001) 74.
[40] S. Khedkar, S. Shinde, Deep learning and ensemble approach for praise or complaint classification, Procedia Comp Sci 167 (2020) 449–458.
[41] A. Fuster, M.C. Plosser, J.I. Vickery, Does CFPB oversight crimp credit? FRB New York Staff Rep (2018).
[42] I. Ayres, J. Lingwall, S. Steinway, Skeletons in the database: an early analysis of the CFPB's consumer complaints, Fordham J. Corp. & Fin. L. 19 (2013) 343–391.
[43] K. Bastani, H. Namavari, J. Shaffer, Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints, Expert Syst. Appl. 127 (2019) 256–271.
[44] P. Foohey, Calling on the CFPB for help: telling stories and consumer protection, Law & Contemp. Probs. 80 (2017) 177–209.
[45] V.K. Vaishnavi, W. Kuechler, Design Science Research Methods and Patterns: Innovating Information and Communication Technology, secondnd edition, Crc Press, 2015.
[46] B. Kuechler, V. Vaishnavi, On theory development in design science research: anatomy of a research project, Eur. J. Inf. Syst. 17 (2008) 489–504.
[47] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable AI systems for the medical domain? Working Paper, 2017.
[48] A. Rai, Explainable AI: from black box to glass box, J. Acad. Mark. Sci. 48 (2020) 137–141.
[49] M. Siering, J. Muntermann, M. Grčar, Design principles for robust fraud detection: the case of stock market manipulations, J. Assoc. Inf. Syst. 22 (2021).
[50] J. Zou, L. Schiebinger, AI can be sexist and racist—it's time to make it fair, Nature 559 (2018).
[51] S. Feuerriegel, M. Dolata, G. Schwabe, Fair AI: challenges and opportunities, Business & Information Syst. Eng. 62 (4) (2020) 379–384.
[52] S.M. Mudambi, D. Schuff, What makes a helpful online review? A study of customer reviews on amazon.com, MIS Q. 34 (2010) 185–200.
[53] Z.C. Lipton, The mythos of model interpretability, Queue 16 (2018) 31–57.
[54] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Mag. 17 (1996) 37–54.
[55] R.B. Ekelund, F.G. Mixon, R.W. Ressler, Advertising and information: an empirical study of search, experience and credence goods, J. Econ. Stud. 22 (1995) 33–43.
[56] Q. Gan, Q. Cao, D. Jones, Helpfulness of Online User Reviews: More Is Less, AMCIS 2012 Proceedings, 2012.

[57] M. Siering, The economics of stock touting during internet-based pump and dump campaigns, Inf. Syst. J. 29 (2019) 456–483.
[58] C. Janze, M. Siering, "Status effect" in user-generated content: Evidence from online service reviews, in: Proceedings of the International Conference on Information Systems, 2015.
[59] M. Siering, J. Muntermann, B. Rajagopalan, Explaining and predicting online review helpfulness: the role of content and reviewer-related signals, Decis. Support. Syst. 108 (2018) 1–12, https://doi.org/10.1016/j.dss.2018.01.004.
[60] J. Otterbacher, Inferring gender of movie reviewers: Exploiting writing style, content and metadata, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010, pp. 369–378.
[61] A. Hotho, A. Nürnberger, G. Paaß, A brief survey of text mining, GLDV J. Comput. Ling. 20 (2005) 19–62.
[62] T.G. Dietterich, Ensemble methods in machine learning, Lect. Notes Comput. Sci 2000 (1857) 1–15.
[63] S.S. Groth, M. Siering, P. Gomber, How to enable automated trading engines to cope with news-related liquidity shocks? Extracting signals from unstructured data, Decis. Support. Syst. 62 (2014) 32–42.
[64] P.J. Stone, E.B. Hunt, A computer approach to content analysis: studies using the general inquirer system, in: Proceedings of the AFIPS Spring Joint Computer Conference, 1963, pp. 241–256.
[65] S.B. Kotsiantis, Supervised machine learning: a review of classification techniques, Informatica 31 (2007) 249–268.
[66] T. Mitchell, Machine Learning, McGraw-Hill, London, 1997.
[67] B. Clapham, M. Siering, P. Gomber, Popular news are relevant news! How investor attention affects algorithmic decision-making and decision support in financial markets, Information Systems Frontiers 23 (2021) 477–494.
[68] T.G. Dietterich, Machine-learning research: four current directions, AI Mag. 18 (1997) 97–136.
[69] G. Valentini, F. Masulli, Ensembles of learning machines, Lect. Notes Comput. Sci 2486 (2002) 3–20.
[70] S. Gregor, A.R. Hevner, Positioning and presenting design science research for maximum impact, MIS Q. 37 (2013) 337–355.