Towards Extracting Customer Needs from Incident Tickets in IT Services

Lena Eckstein
Karlsruhe Service Research Institute
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: lena.eckstein@kit.edu

Niklas Kuehl
Karlsruhe Service Research Institute
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: kuehl@kit.edu

Gerhard Satzger
Karlsruhe Service Research Institute
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: gerhard.satzger@kit.edu

***Abstract*—In many service relationships, customer encounters are not systematically exploited in order to gain valuable insights. However, text mining and analytics methods would provide effective means to systematically screen customer responses and automatically extract relevant business information.**

In this work, we develop a machine learning method as an artifact for screening incident information in IT Services to detect customer needs. We implement and evaluate the method in a real-world context with an IT provider covering several thousands of incident tickets per year.

We show that it is feasible to map incoming tickets to a domain-specific selection of needs—and, hence, enable the providers' customer contacts to address unfilled needs with tailored service offerings. Thus, we contribute a methodology to service marketing and innovation managers to automatically and scalably monitor their customer base for additional sales opportunities.

I. Introduction

Customer Relationship Management (CRM) has turned into a key concern in various industries since many products and services have become commodities. Hence, an increasing focus is put on customer needs instead of distinct product or service features in order to innovate and to offer valuable services [1]. This is especially relevant where a high volume of transactions occur and, thus, huge amounts of data are acquired [1]. It is by observing customer behaviour, remembering past experience, learning from it and acting upon it that a relationship is built [2, p. 5].

Thus, one possibility of gaining information about customer needs in order to improve customer relationship is the analysis of service encounters. In IT services, incident, problem or complaint handling constitutes important service encounters in the customer relationship. Documentation of these service encounters frequently happens via so-called tickets. Over time, large amounts of data sets are created and stored by IT service providers. Some providers resort to manual ticket analysis in order to identify so far unmet customer needs — relying on knowledge and experience of technical support engineers. However, it becomes obvious that huge and fast growing data volumes as well as the need to externalize engineers' knowledge require more automated, scalable and data-driven solutions. This paper develops an approach based on machine learning methods and evaluates it in a feasibility study in industry. Input data are incident tickets of a particular product family of an IT service provider covering several thousands of tickets per year, mainly in B2B settings.

The central research question is: "How can structured and unstructured data of incident tickets be analyzed with a data-driven approach to identify customer needs?" The contribution of the paper is twofold: First, it proposes a method to train machine learning models in order to detect customer needs in (IT maintenance) incident ticket data. Second, it evaluates the approach on actual data and shows its feasibility by portraying different performances and their interpretation.

II. Related Work

A variety of research papers exist on the topics *customer need identification* and *incident ticket analytics* with machine learning methods. For *customer need identification* typical articles focus on needs related to products and use feedback in online customer centers as input data sources (e.g. [3]–[5]). There are two related examples of practice-oriented research on using advanced data analytics for customer need identification in services. Bae et al. [6] extract customer needs from complaints in a life insurance company while Kuehl et al. [7] determine whether Twitter messages express needs for e-mobility services. Our work differs in the following aspects: In contrast to Bae et al., we use text mining as an almost fully automated process. Compared to Kuehl et al., we focus on specifying particular needs. In addition, we also tap a different data source and choose a B2B domain.

With regard to the topic *incident tickets*, three main groups of research can be distinguished: IT system monitoring [8], [9], grouping of similar tickets [10], [11] and extraction of further useful information from tickets [12]–[14]. Among these, the work by Godbole and Roy [12] is most closely related to our approach: They try to judge customer satisfaction from incident tickets and, hence, use a comparable data source. While their setting is similar, we target a different level of insight: We do not intend to analyze whether underlying needs are satisfied but more specifically what these (implicitly) expressed needs actually are.

III. Foundations

In order to provide a common understanding of the terminology used in this paper, we define some prerequisites.

2378-1971/16 $31.00 © 2016 IEEE
DOI 10.1109/CBI.2016.30

Needs are "states of felt deprivation" [15, p. 34] created by a "discrepancy between actual and desired state of being" [16, p. 599]. In a business context, needs result from the value creation process and are problems for which a solution is desired [17, pp. 360f.].

Needs influence customer expectations that in turn affect perceived service quality, which is an important competitive factor in services [17]. In other words, high quality service means satisfying customer needs—whether stated explicitly or not—since quality is about "the characteristics of a product or service that bear on its ability to satisfy stated or implied needs" [18].

Quality attributes are, hence, an expression of underlying needs. Here, primary needs in services, i.e. rather general customer requirements [19], are matched to high-level service quality attributes. Based on an extensive literature review, we identified 25 relevant needs for services in general, for IT services and for B2B services. The relevancy of these needs for the combination of the three (B2B IT Services) has been validated by interviewing a business expert. This results in the following 14 customer needs (in alphabetical order):

- availability/responsiveness [20]
- capacity [21]
- competence (of the provider) [20]
- continuity [21]
- convenience [22]
- customer knowledge [20]
- efficiency [23]
- information [20]
- performance [23]
- personalization [24]
- reliability/dependability [20], [25]
- security/safety [20], [24]
- simplicity [22]
- training (of the customer) [22]

Traditionally, needs of a specific customer are identified via customer interviews or surveys, in one-to-one or group settings, as well as via regular threads of communication like correspondence, phone calls, or meetings. In addition, complaint and incident documentation can be a valuable source for collecting customer needs [19], [26].

An incident in the context of IT services is defined as an "unplanned interruption to an IT service or reduction in the quality of an IT service" [27, section 4.2]. A widespread method for incident documentation is found in the IT Infrastructure Library (ITIL) which lists several best practice elements of a so-called "incident ticket" [27, section 4.2.7.2].

IV. Methodology

To extract customer needs from ticket data, we pursue the following approach (see figure 1): We generate a list of case-specific needs and then manually label a set of representative tickets as to which, if any, of the customer needs they express. We then prepare the data to be used as a training set in a classification model, apply text mining techniques, and evaluate the classification result against a random guess benchmark.

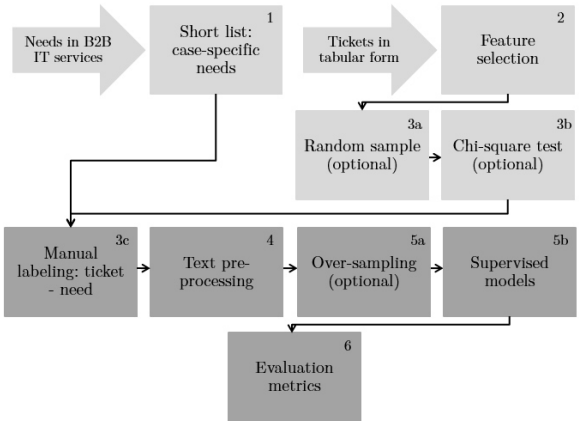

Fig. 1. Approach design

A. General Preparation

As stated above, a "long list" of 14 customer needs in B2B IT services has been derived based on a literature review. As a next step, we reduce the set to a shortlist of most relevant needs (1) to obtain a more manageable amount of distinct needs for manual labeling. Therefore, the list of 14 customer needs (in alphabetical order to prevent the impression of ranking by importance) is sent to an odd number of business experts, at least three, who independently select the most important needs in their opinion and rank them by importance. All needs which are chosen by majority are then used for the subsequent labeling and classification. In order to ensure overlap in the expert evaluations, we ask the experts to identify their top six needs, thus ensuring that more than one need is selected. Rankings are used as weights if more than six needs get chosen by majority vote. Experts selected for this task need to be familiar with the product group(s) from which the respective tickets arise but also have a thorough understanding of IT services and customer needs.

B. Feature Selection

Next we have to characterize tickets by selecting features (2) that later on can be used to point to particular needs identified above. A good starting point are the incident ticket elements listed in ITIL [27, section 4.2.7.2] since ITIL is a known framework providing best practices for IT service management [28, pp. 189f.].

Figure 2 shows the elements we consider relevant or discard, respectively. We assume that a short summarizing free-text is available as part of a ticket. This text shall include problem description and solution, i.e. briefly describe problem symptoms and solution steps. Both are important since either problem or solution alone might not be specific enough to determine an underlying need. For example, a "simple" solution like a software update may be an indicator for a lack

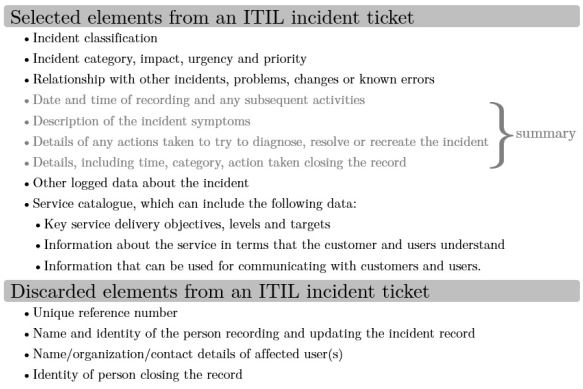

Fig. 2. Selected and discarded features, based on [27, section 4.2.7.2]

of competence on customer side. We decide to start with a summary instead of detailed incident problem and solution description for three reasons. First of all, we assume that a summary contains the most important information about the incident and discards any sidesteps that did not lead to the correct solution. Secondly, ticket labeling is speeded up since experts do not have to read long incident details. Finally, text mining is assumed to work better on short texts containing most relevant and recurring key words.

C. Labeling

Next, a sample of x tickets has to be randomly chosen from the overall set of tickets (3a). This set will then get labeled and build the base for training a classification algorithm and testing its quality. Representativeness of this sample for the population (a particular product group, a country, etc.) may be validated. A common Chi-square goodness of fit test (3b) evaluates if observed data follow a certain probability distribution or, in other words, whether a random sample is representative for the population it was chosen from [29]. $p - values \geq 0.15$ imply acceptance of the null hypothesis [29, p. 249], i.e. representativeness of sample can be assumed. Random sampling and Chi-square testing may be skipped if the overall set of tickets is rather small or sufficient resources (time and skilled personnel) for manual labeling are available.

Then, tickets are manually labeled by business experts with the one most appropriate, implicitly expressed customer need (3c). For this, ticket data are presented to an odd number of experts, at least three, independently. For each ticket, either one of the selectable needs is chosen, or alternatively it can be labelled with "other (specify!)" or "no need". In addition, experts flag the ticket if it is very difficult to label with one main need only – as an indication of labeling quality. The main need for each ticket is ultimately determined by majority vote. If a ticket is labeled with three different needs by the experts, then it can be either refereed by a "trusted" expert (our preferred choice) or discarded from the analysis.

D. Text Mining

As a text mining process (4) we choose a bag-of-words approach with tokenization, stemming and stop word removal. This approach has not only been successfully employed in related research as presented previously [3], [4], [10], [11], [13], but is also advisable from two additional angles. First, incident tickets are expected to have a rather *limited vocabulary* of recurring technical terms, which makes it particularly suitable for a bag-of-words approach [2, pp. 783f.]. Second, the *implementation effort* for running a feasibility study has to be taken into account: Bag-of-words and simple natural language processing are fairly easily applied. More complex NLP approaches would require creation of comprehensive lexica with synonyms, homonyms and misspelling correction for a specific domain, in our case incident ticket language and technical terms.

For stemming we apply the Porter stemmer [30, p. 7] and then tokenize each resulting word. For speeding up stop word removal, standard text mining is enhanced by named entity recognition, which identifies e.g. email addresses, names (mainly of technicians on customer and provider side), URLs, IP addresses, percentages, date- and timestamps, which are to be removed [30, p. 22]. If necessary, further stop words are removed manually. For example, a semi-structured mask/blueprint can be used when entering free-text to an incident ticket. These mask words must be discarded since they do not contain ticket specific information. The remaining tokens are then used to create a term-document matrix with boolean values.

E. Classification

We can well expect cases where records are not approximately equally distributed over different customer need classes and so-called "class imbalance" is found. Classification algorithms usually assume balanced distribution of classes to work properly [31]. To remedy this and create a balanced training data set, either a majority class needs to be under-sampled (i.e. records are removed) or a minority class over-sampled (i.e. records are replicated). The Synthetic Minority Over-sampling Technique (SMOTE) is an extension of the latter usually delivering better results [32], [33]: Instead of over-sampling with replacement, synthetic examples are created for a minority class based on the k-nearest neighbors approach. However, SMOTE has shown to be generally unsuitable for high-dimensional data [34]. Since we assume a small data set arising from the need for manual labeling, under-sampling would reduce the data set further, which is not preferable here. Thus, we use over-sampling to balance the training data set (5a).

Decision trees, Support Vector Machines (SVM), k-nearest neighbours (kNN) and naïve Bayes are promising algorithms for training classification models (5b). They have successfully been used for similar tasks in related work (decision trees: [4], [6], [10], [14]; SVM: [7], [10]–[12]; kNN: [10]; naïve Bayes: [7], [10], [12]) and are mentioned among the top ten

algorithms in data mining [35]. In addition, they are suitable for text classification [36, p. 213].

F. Evaluation Metrics for Classification

For the limited amount of labeled tickets, we split data into 80% training and 20% test data and apply ten-fold stratified cross-validation [39]. For the evaluation of each generated model (6), confusion matrices (see figure 3) are created for test sets. Based on these, overall accuracy or error rate and class-specific precision and recall are calculated for each customer need.

$$accuracy = \frac{TP + TN}{P + N} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$F\text{-}score = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{4}$$

Then, overall (average) metrics for multi-class classification are calculated [40]. We choose macro-averaging in case of a class-imbalance problem, i.e. when the number of records per customer need in the sample is significantly different. Macro metrics treat all classes equally and are, hence, insensitive to large classes. Else micro-averaging can be used [40]. We use the F-score as an additional evaluation metric since it considers both precision and recall and allows for emphasizing one by adjusting β. A harmonic $F\text{-}score$ weight between precision and recall is $\beta = 1$. Depending on the application purpose, other values can be chosen, e.g. $\beta = 0.5$ (more weight on precision) or $\beta = 1.5$ (weight on recall). Precision measures the share of tickets from a predicted need that actually belongs to this need. Recall regards the share of tickets captured correctly for a given need. From a managerial perspective, the prioritization of precision vs. recall (and, thus, the choice of β in the F-score) depends on the business impact of suffering the error. A focus on high precision reduces potential effort for re-classifying tickets after applying a trained model—as a predicted need for a ticket is an actual need with a high probability. If this effort—needed to sort out false positives—is to be kept low, focus (and weight) can be set on precision. High recall, on the other hand, means that tickets with a specific need are in fact recognized as such. This would be important if decision-makers are particularly interested in not missing out on relevant needs - at the expense of ending up with more "false positives" that erroneously indicate a need. In this case, recall may be weighted higher than precision.

Comparing a multi-class model to the outcome of a random guess will be the baseline of the evaluation: The *expected* amount of true positives has to be calculated—either by assuming an already known distribution of tickets over n need classes or with a probability of $1/n$ for each class. Combined with the actual number of tickets per need ("positive samples") and the expected number of predictions for this need ("positive labels"), class-specific and average metrics over all classes can be determined for random guessing.

		Predicted class		
		A	B	*Total*
Class	A	true positives (TP)	false negatives (FN)	*positive samples (P)*
	B	false positives (FP)	true negatives (TN)	*negative samples (N)*
	Total	*positive labels (P′)*	*negative labels (N′)*	*all samples/labels*

Fig. 3. Confusion matrix, based on [32, p. 366]

TABLE I
INCIDENT TICKET FEATURES AND MAPPING TO ITIL INCIDENT TICKET

Feature	**ITIL**
severity during ticket life cycle	incident impact, urgency and priority
product group	incident classification
total amount of days until closure	date and time of recording and closing
solution code and fix number	category, action taken closing the record, known errors
indicators regarding critical situations with higher management involvement	incident impact, urgency and priority
language preferences	service catalogue
problem and solution summary	date and time of recording and any subsequent activities; description of the incident symptoms; details of any actions taken to try to diagnose, resolve or recreate the incident; details, including time, category, action taken closing the record

V. RESULTS

In the following section, we apply the presented methodology in a feasibility study and present and discuss the results.

A. Feasibility Study Setting

The feasibility study is set in the context of a large, internationally operating provider of hardware, software and IT services. The incident tickets arise from software maintenance requests of a specific product group and customers are mainly B2B private sector companies. For reasons of confidentiality, no further details about the product group of tickets examined are disclosed here.

The following features are defined as interesting by three business experts for identifying customer needs in this setting: information on severity during ticket life cycle, product group, total amount of days until closure, solution code and fix number, indicators regarding critical situations with higher management involvement and language preferences, and problem and solution summary in textual format. Table I relates these features to the information contained in an incident ticket according to ITIL as described previously.

TABLE II
NUMBER OF TICKETS ASSIGNED TO EACH CUSTOMER NEED

Customer need	Labeled tickets
availability/ responsiveness	57
competence	62
continuity	15
customer knowledge	1
efficiency	1
reliability/ dependability	6
no need	21
other	3
(blank)	34
Total	200

B. Preparatory Steps

First, we send our list of 14 needs to three experts of the feasibility study's environment in order to retrieve essential needs.

As a result, the following most important needs are chosen (in alphabetical order):

- availability/responsiveness *(mentioned by all)*
- competence (of provider) *(mentioned by all)*
- continuity (continuous service) *(mentioned by two)*
- customer knowledge (provider knows customer) *(mentioned by two)*
- efficiency *(mentioned by all)*
- reliability/dependability *(mentioned by two)*

The needs "information", "performance" and "security/safety" are only mentioned by one expert each and, thus, not included. From the complete list, the following are not selected by any expert: "capacity", "convenience", "personalization", "simplicity" and "training".

Subsequently, a random sample of 200 incident tickets is selected. In order to ensure the representativeness of the sample, we perform a Chi-square goodness of fit test. The result is $T \approx 8.4$ and $p \approx 0.4$ which meets our requirement $p \geq 0.15$. Therefore, the random sample is assumed to be representative for the original list of tickets. Finally, the 200 randomly sampled tickets are labeled with the six corresponding customer needs by an expert[1].

As shown in table II, 34 tickets are not labeled. Thereof, 26 tickets are marked as "very difficult to label". For the remaining eight tickets neither a need is selected nor is a remark found in "very difficult to label". The tickets labeled as "other" or left blank comprise tickets that consist of multiple related tickets and, hence, cannot be traced to one distinct topic by the expert. We do not include tickets marked as "other" or left blank for further analysis. Moreover, we discard classes with only one example, i.e. "customer knowledge" and "efficiency". For these, the same record would have to be used for training and test which obviously would not produce reliable results.

[1]Only one expert is available for this task. This is due to the amount of work required to manually label tickets and time constraints for this work. The expert is software engineer and has professional knowledge in technical support for the product group examined here.

This leaves a total of 161 tickets of which 13% (21 tickets out of 161) are labeled as containing no need. The two predominant labels are "competence" (39% of remaining tickets) and "availability/responsiveness" (35%) while only 9% of records belong to "continuity" and 4% to "reliability/dependability". In order to reduce the effects of this class-imbalance problem, we over-sample the minority classes "no need", "continuity" and "reliability/dependability" in the training set. The 161 tickets are set as the final data set[2] and used as the basis for calculating evaluation metrics in the remainder of the paper.

C. Text Mining and Classification Evaluation

After performing preparatory work, we implement the actual approach and calculate evaluation metrics. For model creation two different tools are combined: IBM SPSS Modeler Premium 16.0 and RapidMiner Studio 6.5.002.

1) Text Pre-Processing: Solution and incident summary are in textual format and are assumed to contain relevant information about customer needs. Hence, the text is preprocessed as described previously.

2) Results for Five Need Classes: Table III portrays evaluation metrics for this scenario[3]. Numbers in bold indicate results superior to both variants of random guessing. Assuming a known distribution of tickets over need classes means probabilities of $\frac{57}{161}$ for "availability", $\frac{62}{161}$ for "competence", $\frac{21}{161}$ for "no need", $\frac{15}{161}$ for "continuity" and $\frac{6}{161}$ for "reliability". When the distribution of tickets per class is assumed to be unknown, probability is equal, i.e. $\frac{1}{5}$, for each class.

With C4.5 and kNN, "continuity" and "reliability" show precision and recall of 0, i.e. no true positives are found for these two customer needs. For SVM and naïve Bayes, additionally, no true positives exist for "no need" while recall and precision for "availability" and "competence" are superior to random guess. Overall, these results are below expectations since some needs have no correctly classified tickets at all.

Hence, we test an additional setup: Instead of distinguishing between all identified need classes, we combine the three minor classes "no need", "continuity" and "reliability/dependability" together to one class "minor needs".

3) Results for Three Need Classes: Table IV portrays evaluation metrics and expected outcomes of random guessing. As before, bold numbers illustrate superiority to random guessing. With each algorithm, we find true positives for all three classes. Recall and precision for "minor needs" are in general lower than for the two major needs except with kNN.

Ordered by macro-averaged F-score (with $\beta = 1$), SVM performs best, followed by kNN, C4.5 and naïve Bayesian classifier. The ranking is equal for macro-averaged precision

[2]It could be critically remarked that this amount only contributes to 81% of the original ticket sample. However, it is assumed that tickets that turned out to have no label should have been removed in advance (during ticket preparation) since they are a combination of multiple tickets. The two excluded tickets for "customer knowledge" and "efficiency" would have a negligible effect on evaluation metrics.

[3]Micro-averaging leads to $precision = recall = F\text{-}score$ as found here if all records are assigned to a class and none is labeled with "null".

TABLE III
FIVE CLASS CLASSIFICATION RESULTS, ROUNDED TO TWO DECIMAL PLACES

Algorithm	C4.5	SVM	kNN	naïve Bayes	random guess, known class distribution	random guess, unknown class distribution
overall accuracy	**0.32**	**0.55**	0.29	**0.48**	0.30	0.20
average per-class accuracy	0.20	**0.30**	**0.22**	**0.26**	0.20	0.20
"availability" recall	0.27	**0.91**	0.27	**0.82**	0.35	0.20
"availability" precision	0.30	**0.56**	**0.50**	**0.45**	0.35	0.35
"competence" recall	**0.50**	**0.58**	0.33	**0.50**	0.39	0.20
"competence" precision	**0.50**	**0.54**	**0.40**	**0.60**	0.39	0.39
"no need" recall	**0.25**	0.00	**0.50**	0.00	0.13	0.20
"no need" precision	**0.17**	0.00	**0.18**	0.00	0.13	0.13
"continuity" recall	0.00	0.00	0.00	0.00	0.09	0.20
"continuity" precision	0.00	0.00	0.00	0.00	0.09	0.09
"reliability" recall	0.00	0.00	0.00	0.00	0.04	0.20
"reliability" precision	0.00	0.00	0.00	0.00	0.04	0.04
micro-averaged recall. precision. f1-score	**0.32**	**0.55**	0.29	**0.48**	0.30	0.20
macro-averaged recall	0.20	**0.30**	**0.22**	**0.26**	0.20	0.20
macro-averaged precision	0.19	**0.22**	**0.22**	**0.21**	0.20	0.20
macro-averaged f1-score	0.20	**0.25**	**0.22**	**0.23**	0.20	0.20

TABLE IV
THREE CLASS CLASSIFICATION RESULTS, ROUNDED TO TWO DECIMAL PLACES

Algorithm	C4.5	SVM	kNN	naïve Bayes	random guess, known class distribution	random guess, unknown class distribution
overall accuracy	**0.39**	**0.45**	**0.35**	**0.35**	0.33	0.33
average per-class accuracy	**0.35**	**0.42**	**0.38**	0.33	0.33	0.33
"availability" recall	0.18	**0.55**	0.27	**0.45**	0.35	0.33
"availability" precision	0.33	**0.46**	**0.50**	**0.36**	0.35	0.35
"competence" recall	**0.75**	**0.58**	0.25	**0.42**	0.39	0.33
"competence" precision	**0.43**	**0.44**	**0.43**	**0.42**	0.39	0.39
"minor needs" recall	0.13	0.13	**0.63**	0.13	0.26	0.33
"minor needs" precision	0.25	**0.50**	**0.28**	0.20	0.26	0.26
micro-averaged recall. precision. f1-score	**0.39**	**0.45**	**0.35**	**0.35**	0.33	0.33
macro-averaged recall	**0.35**	**0.42**	**0.38**	0.33	0.33	0.33
macro-averaged precision	**0.34**	**0.47**	**0.40**	0.32	0.33	0.33
macro-averaged f1-score	**0.34**	**0.44**	**0.39**	0.33	0.33	0.33

and recall. Compared to previous setups, only small class-imbalance is found here. Thus, we include ranking by micro-averaged F-score. SVM is the best performing algorithm in this list, too, followed by C4.5, kNN and naïve Bayesian classifier.

Finally, we compare the generated models for the three class setup to the expected outcome of a random guess. Here, a known distribution of tickets over need classes means probabilities of $\frac{57}{161}$ for "availability", $\frac{62}{161}$ for "competence" and $\frac{42}{161}$ for "minor needs". Unknown class distribution suggests a probability of $\frac{1}{3}$ for each class.

In both cases, SVM classification provides higher recall, precision and F-score values except for "minor needs" recall. KNN is the only model where "minor needs" recall is better than a random guess. In addition, precision for all classes is higher while recall for the major needs is lower. Naïve Bayesian classifier presents better recall and precision than random guessing but only for "availability" and "competence". C4.5, in general, does not perform better than random guessing except for "competence" recall and precision.

D. Interpretation

Overall, the three class setup provides models that result in better metrics than a random guess. The selection of an applicable model depends on managerial inclination. Two methods to deal with "minor needs" can be envisioned. On the one hand, corresponding tickets could be discarded from further processing or just regarded as "less important". Then, high precision and recall values for "availability" and "competence" are desirable and the presented SVM model would be selected. On the other hand, "minor needs" tickets could receive "special attention" in further processing. In this case, SVM would be chosen for highest precision and kNN for highest recall. When the effort for re-classifying (i.e. manually removing tickets that do not belong to "minor needs") should be low, precision is weighted more, else higher recall is preferred.

Knowing customer needs is a prerequisite to offer adequate solutions to customer problems. However, identifying customer needs in interviews or focus groups can be

a cumbersome and time-consuming process [41]. With our approach, needs are retrieved from incident tickets, i.e. a service encounter documentation. By this, we find main needs for a specific customer on the one hand, and, on the other hand, are able to group incidents and, thus, customers with similar needs. Subsequently, two directions of action are proposed. Operationally, gained knowledge may drive more effective behaviour in customer interactions: the provider may be able to react faster to customer requests when "availability" is a main customer need, or route the incident to the right expert to meet a "competence" need. From a business development point of view, appropriate solutions and offerings can be provided to a customer. For example, for a customers whose main concern is "availability", providing on-site technical engineers might be an attractive offer. Another customer, who is relying on provider competence because of his own lack of know-how, might benefit from an education program to improve incident resolution in the future. All of these measures aim at improving customer relationship by providing additional value and encouraging long-term business.

VI. Conclusion

We developed an analytical approach based on classification and text mining algorithms to identify customer needs from incident tickets. We successfully validated our approach in a feasibility study with a large IT service provider. Results show that the proposed approach is capable to elicit information from service encounter data in an automated and scalable fashion–after an initial training.

Naturally, the work has certain limitations and we envision future extensions to further improve results and applicability.

First, the amount of records used was rather small, and only one expert was available for labelling—as in practice it is hard to carve out time for qualified business experts. Hence, individual bias cannot be ruled out which could be reduced by more labeling experts and using majority votes. By labelling a larger amount of tickets with underlying customer needs, classification models could be further improved. This especially holds for classes with few samples.

Second, word semantics, e.g. synonyms, phrases and part-of-speech are not considered so far—as texts are represented as term-document matrix with stemmed tokens and removed stop words. Building domain-specific lexica could extend linguistic processing of words in addition to a mainly statistical bag-of-words approach.

Third, with regard to classification methods, we used only a small number of available methods applying specific parameters (e.g. tree depth for C4.5). A broader set of methods and parameter choices may further drive performance — in particular if incorporated into learning ensemble models that have proven to be effective in other contexts [42].

Finally, also multi-label classifications may be considered, where $1...n$ needs are assigned to each ticket. By this, multiple needs underlying a ticket would be respected instead of focusing on the "main" need which may sometimes be hard to select.

Despite these limiting factors, the paper contributes research insights to need identification in services and to methods for customer relationship management. Business implications are obvious: Not only can available data be used to be screened for additional (need) information, but the application of data and text mining approaches will allow to do this in an automated and scalable manner — once the one-time setup effort to calibrate the model has been invested. Tapping incident data from service encounters may yield several advantages: in the short-term providers may benefit from operational improvements dealing with a particular incident ticket, while mid-term tailored offerings to the customer may enhance business development, and dependency on scarce and futile expert knowledge is reduced. Long-term this will support customer relationship management and, in particular, back customer intimacy strategies where knowledge about the customer contributes to competitive advantage [43].

Already yielding improvements for the problem area of incident tickets in IT service management, the general approach of eliciting information from customer encounters may bear a far richer potential: on the one hand, other sources of available service encounter data, like customer satisfaction surveys, meeting minutes, technician reports or interaction data itself may be exploitable. On the other hand, service encounter data may be also used to extract other relevant information, like customer experience or customer intimacy ratings [44]. We believe that the application of data and text mining techniques will be an effective means to support marketing and innovation managers, specifically, for building lasting service relationships and individualized offers. Thus, this will enhance servitization strategies of enterprises [45] that provide critical differentiation in competitive markets.

References

[1] C. Rygielski, J.-C. Wang, and D. C. Yen, "Data mining techniques for customer relationship management," *Technology in Society*, vol. 24, no. 4, pp. 483–502, 2002.

[2] G. Linoff and M. J. A. Berry, *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management*, 3rd ed. Indianapolis, IN: Wiley, 2011.

[3] T. Y. Lee, "Needs-based analysis of online customer reviews," in *Proceedings of the Ninth International Conference on Electronic Commerce*, D. Sarppo, M. Gini, R. J. Kauffman, C. Dellarocas, and F. Dignum, Eds. New York, NY: ACM, 2007, pp. 311–317.

[4] Y. Park and S. Lee, "How to design and utilize online customer center to support new product concept generation," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10 638–10 647, 2011.

[5] J. Jin, P. Ji, Y. Liu, and S. C. Johnson Lim, "Translating online customer opinions into engineering characteristics in QFD: A probabilistic language analysis approach," *Engineering Applications of Artificial Intelligence*, vol. 41, pp. 115–127, 2015.

[6] S. M. Bae, S. H. Ha, and S. C. Park, "A web-based system for analyzing the voices of call center customers in the service industry," *Expert Systems with Applications*, vol. 28, no. 1, pp. 29–41, 2005.

[7] N. Kuehl, J. Scheurenbrand, and G. Satzger, ""Needs from Tweets": Towards deriving customer needs from micro blog data," in *Multikonferenz Wirtschaftsinformatik (MKWI) 2016*, V. Nissen, D. Stelzer, S. Straburger, and D. Fischer, Eds. Univ.-Verl. Ilmenau, 2016, vol. 2, pp. 1229–1232.

[8] L. Tang, T. Li, L. Shwartz, F. Pinel, and G. Y. Grabarnik, "An integrated framework for optimizing automatic monitoring systems in large IT infrastructures," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, R. L. Grossman,

R. Uthurusamy, I. Dhillon, and Y. Koren, Eds. New York, NY: ACM, 2013, pp. 1249–1257.

[9] V. Nair, A. Raul, S. Khanduja, V. Bahirwani, S. Sellamanickam, S. Keerthi, S. Herbert, and S. Dhulipalla, "Learning a hierarchical monitoring system for detecting and diagnosing service issues," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, L. Cao, C. Zhang, T. Joachims, G. Webb, D. D. Margineantu, and G. Williams, Eds. New York, NY: ACM, 2015, pp. 2029–2038.

[10] G. A. Di Lucca, M. Di Penta, and S. Gradara, "An approach to classify software maintenance requests," in *Proceedings of the International Conference on Software Maintenance*. Piscataway, NJ: IEEE, 2002, pp. 93–102.

[11] S. Agarwal, R. Sindhgatta, and B. Sengupta, "Smartdispatch: Enabling efficient ticket dispatch in an IT service environment," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Q. Yang, D. Agarwal, and J. Pei, Eds. New York, NY: ACM, 2012, pp. 1393–1401.

[12] S. Godbole and S. Roy, "Text classification, business intelligence, and interactivity: automating c-sat analysis for services industry," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Y. Li, B. Liu, and S. Sarawagi, Eds. New York, NY: ACM, 2008, pp. 911–919.

[13] A. Chandramouli, G. Subramanian, and D. Bal, "Unsupervised extraction of part names from service logs," in *Proceedings of the World Congress on Engineering and Computer Science*, ser. Lecture Notes in Engineering and Computer Science, S. I. Ao, C. Douglas, W. S. Grundfest, and J. Burgstone, Eds., vol. 2. San Francisco: IAENG, 2013, pp. 826–828.

[14] G. Satzger and P. Hottum, "Management der Interaktionsqualität in industriellen Dienstleistungsnetzwerken: Ein "Service Analytics"-Ansatz für die Störungsbearbeitung," in *zfbf Sonderheft 69/2015 Steuerung von Industrial Service Networks*, L. Grünert, P. Horváth, and M. Seiter, Eds. Düsseldorf: Handelsblatt Fachmedien, 2015, pp. 150–173.

[15] G. Armstrong and P. Kotler, *Marketing: An Introduction*, 11st ed. Boston, MA and Munich: Pearson, 2013.

[16] C. Homburg, S. Kuester, and H. Krohmer, *Marketing Management: A Contemporary Perspective*, 2nd ed. London: McGraw-Hill Higher Education, 2013.

[17] C. Grönroos, *Service Management and Marketing: Customer Management in Service Competition*, 3rd ed. Chichester and Weinheim: Wiley, 2007.

[18] American Society for Quality, "Quality glossary," 2015. [Online]. Available: http://asq.org/glossary/q.html

[19] A. Griffin and J. R. Hauser, "The voice of the customer," *Marketing Science*, vol. 12, no. 1, pp. 1–27, 1993.

[20] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "A conceptual model of service quality and its implications for future research," *Journal of Marketing*, vol. 49, no. 4, pp. 41–50, 1985.

[21] D. Cannon, D. Wheeldon, S. Lacy, and A. Hanna, *ITIL Service Strategy*, 2nd ed. London: TSO, 2011.

[22] M. T. Cunnigham and D. A. Roberts, "The role of customer service in industrial marketing," *European Journal of Marketing*, vol. 8, no. 1, pp. 15–28, 1974.

[23] J. Zolkiewski, B. Lewis, F. Yuan, and J. Yuan, "An assessment of customer service in business–to–business relationships," *Journal of Services Marketing*, vol. 21, no. 5, pp. 313–325, 2007.

[24] J. A. Fitzsimmons and M. J. Fitzsimmons, *Service Management: Operations, Strategy, Information Technology*, 7th ed. Boston, MA: McGraw-Hill, 2011.

[25] P. Kotler and K. L. Keller, *Marketing Management*, 14th ed. Harlow: Pearson Education, 2012.

[26] J. F. Early and O. J. Coletti, "The quality planning process," in *Juran's Quality Handbook*, J. M. Juran and A. B. Godfrey, Eds. New York, NY: McGraw Hill, 1999, pp. 3:1 – 3:50.

[27] R. A. Steinberg, C. Rudd, S. Lacy, and A. Hanna, *ITIL Service Operation*, 2nd ed. London: TSO, 2011.

[28] J. Cardoso, H. Fromm, S. Nickel, G. Satzger, R. Studer, and C. Weinhardt, *Fundamentals of service systems*. Springer International Publishing, 2015.

[29] S. M. Ross, *Simulation*, 5th ed. Amsterdam: Academic Press, 2013.

[30] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining," *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, vol. 20, no. 1, pp. 19–62, 2005.

[31] N. V. Chawla and G. Karakoulas, "Learning from labeled and unlabeled data: An empirical study across techniques and domains," *J. Artif. Int. Res.*, vol. 23, no. 1, pp. 331–366, 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622503.1622511

[32] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA: Elsevier, 2011.

[33] C. C. Aggarwal, *Data Mining: The Textbook*. Cham and Heidelberg and New York: Springer International Publishing, 2015.

[34] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC bioinformatics*, vol. 14, pp. 106–121, 2013.

[35] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[36] C. C. Aggarwal and C. X. Zhai, "A survey of text classification algorithms," in *Mining Text Data*, C. C. Aggarwal and C. X. Zhai, Eds. Boston, MA: Springer US, 2012, p. 213.

[37] T. Joachims, "A statistical learning model of text classification for support vector machines," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel, Eds. New York, NY: ACM, 2001, pp. 128–136.

[38] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[39] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA: Morgan Kaufmann Publishers Inc, 1995, pp. 1137–1143.

[40] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[41] M. Fisher, M. Houghton, and V. Jain, *Cambridge IGCSE® Business Studies Coursebook*, ser. Cambridge International Examinations Series. Cambridge University Press, 2014. [Online]. Available: https://books.google.de/books?id=PdtkAwAAQBAJ

[42] A. Vogt, E. R. Mattfeldt, G. Satzger, L. Lueders, M. Piper, O. Gehb, and W. L. Jones, "Analytical support for predicting cost in complex service delivery environments," *IBM Journal of Research and Development*, vol. 58, no. 4, pp. 7:1–7:10, July 2014.

[43] M. Treacy and F. Wiersema, "Customer intimacy and other value disciplines," *Harvard Business Review*, vol. 71, no. 1, pp. 84–93, 1993.

[44] F. Habryn, "Customer intimacy analytics: Leveraging operational data to assess customer knowledge and relationships and to measure their business impact," Zugl.: Karlsruhe, KIT, Diss., 2012, 2012. [Online]. Available: http://digbib.ubka.uni-karlsruhe.de/volltexte/1000028159

[45] A. Neely, O. Benedetinni, and I. Visnjic, "The servitization of manufacturing: Further evidence: Academic paper to be presented at the 18th European Operations Management Association Conference," 2011. [Online]. Available: http://www.researchgate.net/publication/265006912