A Case Study of AI performance enhancement in Incident Management processes with low data-quality

Author: Dominic Stomp
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT

Artificial Intelligence has created new possibilities to increase operational efficiency in many fields, including the field of IT Service Management. Specifically in the resolution process, which focuses on the resolution of issue tickets. These tickets are essentially a bundle of correspondence between the IT agent and the problem holder. These tickets are categorized manually to generate data for continuous improvement and for further escalation down the escalation paths of the support team. AI is already able to classify these tickets based on the initial issue description, replacing the need for manual classification, and therefore increasing operational efficiency. However, some businesses suffer from low quality data sets, which, proven in prior research, will negatively affect the classification capabilities of AI. It was also proven that addressing the issue of low data quality in a data set through extensive data analysis will improve the performance of the corresponding AI models. However, for small to medium-sized businesses without the in-house capabilities for such methods, easier methodologies had to be explored. In this case study, such a low-quality data set was provided by an IT service provider in the nautical tourism sector and enhanced in multiple ways to measure their effect on AI classification performance. For evaluation, AI models had to be created, tested, and trained on the native and enhanced versions of the data set. The results of this research show that minor gains in performance can be achieved through systematic changes in the data set, like a better separation of categories based on their semantic meaning. The larger gains were achieved by removing the lowest quality entries from the data set. This was done, alternatively to extensive data analysis, by having an expert from the support team look for the common indicators of low-quality entries.

Graduation Committee members:
Matthias de Visser
Vincent Göttel

Keywords

Artificial Intelligence, Incident Management, Ticket classification, accessible alternatives, data-quality, case study

1. INTRODUCTION

In modern IT businesses, the effective management of service provision has become imperative to maintain operational efficiency. The related field of IT Service Management (ITSM) is a subset of Service Science that specifically focuses on all service-related aspects of IT business. In the ITSM domain the most common best practices framework to guide IT service processes is called Information Technology Infrastructure Library (ITIL). "ITSM, as defined in the ITIL, is both a glossary, to ensure a uniform vocabulary, and a set of conceptual processes intended to outline IT best practices." (Gulap et al, 2009). ITSM encompasses a wide variety of concepts, but this report focuses on the initial incident and resolution thereof called Incident Management (IM), which covers the practices from a problem being raised by a user to resolving that initial problem. In practice, the problem holder contacts the service department by e-mail or phone, and then the ticket system application creates a unique ticket regarding the specific issue to avoid confusion when working on multiple issues simultaneously. Following the ITIL framework, IT agents in the so-called "1st line support" (Service Desk) address and categorize every ticket, to subsequently escalate the resolution of the ticket to the correct agent in the 2nd line support for either resolution of incidental issues or larger systematic issues. Notice that multiple sequential lines of support are conceptualized in the ITIL framework covering multiple roles encompassing all ITSM processes, resembling a decision tree. The framework is designed to handle large volumes of tickets while maintaining efficiency and generating data to improve or facilitate processes within.

The increasing volume of tickets and the repetitive categorization and description practices in the ITSM field, combined with the current societal interest in artificial intelligence (AI), have spiked research into the theoretical benefits and practical implementation. The first commercial AI-assisted ticketing solutions have already entered the market, but the novel research domain of AI-ITSM is yet to be fully explored. The existing AI models are tailored toward the more traditional IT businesses with larger service desks incentivized to maintain good data quality to manage and improve internal processes. However, IT service providers that do not predominately focus on Incident Management in their operations have additional inherent challenges in the creation of effective models. In the AI-ITSM research domain, Reinhard et al. (2023) and Baresi et al. (2020) have introduced intensive automated methodologies to enhance the precision and sensitivity of AI-based ticket classification and resolution. However, smaller IT businesses might not have the resources necessary to apply these extensive methodologies. This paper is a case study of the implementation of AI in the ITSM practices of a niche IT service provider in the nautical tourism sector, which has an inherently high issue diversity and low-quality ticket data and lacks the resources for extensive automated ticket quality improvement efforts.

Existing research in the AI-ITSM domain highlighted the necessity and effectiveness of improving data quality in the AI model's training dataset. "It is generally known that assessing data quality is important for information systems research as low data quality results in expensive data quality costs" (Batini et al., 2007). In ITIL practices, every ticket generates data that is fundamental to the continuous improvement of the service provided. This data varies from descriptions of the initial issue that caused the ticket and how it was ultimately resolved to numerical values for how long it took to respond and resolve a ticket. However, the resulting dataset is often insufficient in both accuracy and completeness to be used as the training dataset for an effective AI model because "…due to various reasons (e.g., time pressure or convenience) and the complexity of support services, support agents tend to insufficiently describe issues and summarize resolutions, which in consequence limits the capabilities of the AI-driven systems" (Reinhard et al., 2023, page 280). Low data quality in the training dataset will impact the ability of an AI model to classify tickets based on the textual data input correctly. This was proven by Heinrich et al. (2019) who tested the ticket quality dimension of completeness on the performance of recommender systems. Therefore, inexpensive data quality improvement methods will be tested in this case study and evaluated to highlight the challenges and accessible solutions for implementing AI in IT service practices.

1.1 Research Questions

As described by Reinhard et al. (2023), there are various reasons for low-quality datasets limiting the capabilities of AI-driven systems. Additionally, a relatively high degree of variance in possible issues and resolutions in the IM landscape further limits the potency of AI models. Given the potential operational benefits of successful AI implementation and the impact of low data quality on AI prediction accuracy, it is important to explore the possibilities of dataset improvement methods and their effect on the resulting AI model. This case study will provide evidence of the effectiveness of inexpensive and accessible empowerment methods specifically for datasets that suffer from low data quality. With the primary objective of creating AI models with improved performance compared to the native version. The scope of this case study regarding AI implementation will contain a low level of complexity, the predictive categorization of issue descriptions based on the initial textual data.

1.2 Research Questions

To achieve the previously mentioned research objectives within the stated scope of the research, the following research questions must be answered:

(i) What dataset enhancement methodologies are known in prior research on AI-Incident Management that can improve the AI classification of tickets?

(ii) How can these methodologies be translated for small and medium-sized businesses with limited outsourcing resources to improve ticket AI classification?

1.3 Academic Relevance

At the time of writing, the existing research on Artificial Intelligence in the ITSM domain is limited, especially in the context of low-quality datasets. Given the overlap with other research fields, sufficient research could be found on the concept of data quality, as presented by Cai & Zhu (2015). And more in scope with this research, Heinrich et al. (2019) have explored the impact of low-quality data on the performance of AI models in the ITSM field. Fundamentally, this research aims to determine and enhance data quality when confronted with unideal datasets, and Baresi et al. (2020) and Reinhardt et al. (2023) have explored extensive statistical analysis methodologies in line with that aim. However, more basic methodologies to enhance data quality in a dataset and the effectiveness thereof in the ITSM field are yet to be explored.

1.4 Practical Relevance

In this case study, the overarching goal is to utilize AI to reduce the manual workload within the Incident Management process. IT businesses are generally highly scalable, except for the IT service requests that follow, which require human interaction. Frequently Asked Questions lists and AI chatbots are potential solutions to reduce the workload within a business's IT service management department, but these are not applicable to all business cases. The business in this case study requires highly technical, and therefore expensive, experts to interpret and resolve all incoming issue tickets. The specific task within the Incident Management process in which AI will be implemented to alleviate workload, is that of ticket categorization. Within the ITIL best practice framework, each ticket must be carefully interpreted and categorized to generate ticket data necessary for proper IT service management. The workload that this manual interpretation brings, given the large influx of tickets on a daily basis, can now be attempted to be reduced with the possibility of AI. However, an effective AI model that categorizes each ticket correctly based on unstructured issue descriptions can only be achieved if the data it was founded on is accurate. In this case study, that is not the case. If systematic improvements were made to ensure a higher quality of incoming tickets, it would still take years before a sufficient database is gained and ready for AI implementation. Therefore, this research gains relevance by exploring the methodology of improving the variable of data quality necessary to create an effective AI model that can alleviate workload in the categorization of tickets. Additionally, the resource constraint element of this research is an important element due to limited budgets for outsourcing AI implementation efforts. This constraint can have various reasons. In this case study a portion of the service delivery is project-based and bypasses the ITIL practices, resulting in a more limited budget for systematic improvements like AI implementation.

2. LITERATURE REVIEW

The aim of this case study is to highlight the basic challenges of AI implementation in the service delivery processes within the ITIL framework and to test enhancement methods presented in the AI-ITSM research domain. To do that, the current knowledge gained from research in the various theoretical fields must be explored. This includes the concepts presented in the ITSM and AI research domains, the definition and dimensions of data quality, the AI empowerment methodologies through data quality improvement, and the evaluation methods for AI model performance measurement.

2.1 IT Service Management

"ITSM is defined as an approach to IT operations that is characterized by its emphasis on IT services, customers, service level agreements, and an IT function's handling of its daily activities through processes" (Conger et al., 2008). Essentially, ITSM does not focus on the IT function itself but on managing the service provision. This is reflected in the most commonly used ITSM best practice framework, Information Technology Infrastructure Library (ITIL). ITIL is a conceptual framework and collection of best practices and processes that streamline the service provision processes holistically and facilitate enhancement through data collection. ITIL is generally combined and dependent on ticketing systems designed specifically for these processes. These ticketing systems allow the service agents to work on multiple issues simultaneously and include the different roles and parallel processes presented in the respective ITSM framework. The applications also allow the collection of data to be used by service managers to enhance the processes. The collected data includes metrics related to the tickets and the performance of the service agents, for example, the categorization of issues and nominal values of ticket resolution times and first response times. The ticketing system generally measures the nominal values. However, the categorization or detailed descriptions of ticket issues and resolutions are interpreted and manually generated by the service agents.

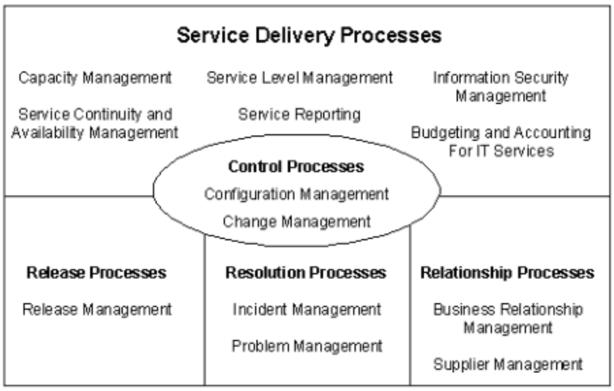


Figure 1: Overview of ITSM processes (ISO/IEC 20000-1, 2005)

See Figure 1 for a conceptual overview of the ITSM processes. The scope of this case study is limited to AI implementation in the resolution process, specifically Incident Management (IM).

2.2 The importance and challenges of ticket classification

In the ITIL best practices framework one of the first steps of the resolution process of a ticket is ticket classification. This manual categorical classification and other metrics generated by the ticketing system provide the necessary data for the continuous improvement cycle of the IT service delivery. The quality of this data is not only important within the field of ITSM, but Heinrich et al. (2019) also found a negative impact of low data quality on the consequent AI models. When looking to enhance data quality it is also important to distinguish between the types of data. In the scope of this research, the initial textual data submitted by the service requester and subsequent classification entries affixed by the assigned IT agent to each ticket. In the case of ticket classification, incorrect entries made by the IT agent will affect the performance of consequent AI models. Thus, efforts must be made to recognize these errors and remove or adjust them in the AI training set.

There are various challenges in the proper classification of tickets, also depending on the level of detail asked for by management. Regarding AI ticket classification, the initial text data is inherently unstructured data, necessitating proper preprocessing. Furthermore, the textual data can vary highly in the level of detail given by the service requester, and spelling errors are common. Regarding ticket classification by the IT agent in the context of the case study, unclear issue descriptions and a lack of detail in the initial text data lead to classification errors. Additionally, the service requester most often does not know what the root cause of the problem is and merely describes the issue from their perspective. The IT agent could later adjust the ticket classification to be accurate, but for reasons like a heavy workload, this does not always happen. As an example, the service requester initially only describes that their laptop is not working properly. This information alone does not suffice to get a clear picture of the issue and classify the ticket, as it could be caused by a variety of reasons like hardware issues, internet issues, software issues, and authentication issues to name a few. Therefore, the IT agent must contact the service requester to gather further details and adjust the classification later. This example highlights a challenge in both manual ticket classification and AI ticket classification, which is further compounded if the AI model is based on a training set that includes the classification errors made by the IT agents.

2.3 Artificial Intelligence Categories

Three commonly used classification algorithms for text data are the Bayesian classifier, decision tree classifier, and artificial neural networks (ANN). Depending on the context and the intended goal, these classifiers, individually or combined, must be tested against the training set to determine the best fit.

2.3.1 The Bayesian classifier

The Bayesian classifier is used to classify text data based on the occurrence and weights of features (words) in the text. For example, the word "bark" can point to the outer layer of a tree or the noise a dog makes. The actual meaning of a text including the word "bark" is then calculated by the occurrence of other features and their assigned weights. So, the occurrence of features like "noise" or "tree" in the same text will be used as indicators to classify based on the true meaning of the text.

2.3.2 The Decision Tree classifier

The decision tree classifier resembles a flowchart with the decision points being the classification points. These classification points based on the features of the text can both be categorical or numerical values and lead to classification based on the presence, absence, or value threshold of certain features. An example of how a decision tree classifier would work in this case study: If the text data mentions that the internet connection is slow but not lost, and all individuals on the ship are affected, and the numerical value of internet speed would fall below 5 MB/s, then the issue would be classified as "bad signal area". The challenge of this classification method is that the text data needs to be accurate and complete, as the classification is conditional and requires the presence of certain features and values to complete the decision tree pathway.

2.3.3 Artificial Neural Networks

Artificial Neural Networks (ANN) is the most versatile and least transparent classifier of the three. ANN derives its name from a biological neural network due to the resemblance of its conceptual layout. Multiple input nodes deliver information to layers of hidden nodes that each have their respective weights to then finally end in the output layer to classify the data. The connections between nodes called synapses, have their weight adjusted to correctly classify input data depending on the training set or subsequent deep learning. The complexity increases with the number of hidden node layers that are required to correctly classify the input data; this gives the ANN flexibility in its use as it can be used on seemingly unrelated data. However, the complexity of how ANN can produce an accurate model of classification also causes it to be untransparent.

2.4 Data quality

The primary variable causing challenges in the creation of an effective AI model within the scope of this research is data quality. But what determines the quality of data, and how does it affect artificial intelligence? As previously mentioned, AI models are generated through statistical analysis of large datasets to create complex models that can predict outcomes based on the input. The AI model is therefore generated through and dependent on the learning dataset. The quality of which being highly important in the effectiveness of the resulting AI model. According to Cai and Zhu (2015) "… the use and analysis of big data must be based on accurate and high-quality data, which is a necessary condition for generating value from big data". Multiple definitions have been proposed in various research fields, but simply put the quality of data is "fitness for use" (Wang & Strong, 1996). This means that the quality of data is determined by its usability and its reliability. However, more dimensions and subdimensions have been proposed by Cai and Zhu. See Figure 2 for the total conceptual overview of the data quality dimensions by Cai and Zhu.

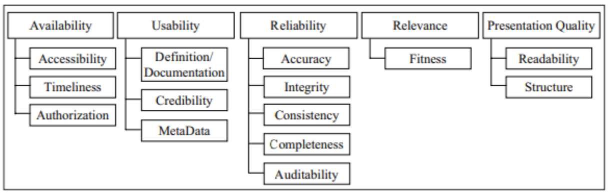


Figure 2: Two-layer big data quality standard by Cai & Zhu (2015)

In the context of this research, not all dimensions proposed by Cai and Zhu are equally relevant due to the use of ticketing systems ensuring that multiple dimensions and subdimensions are standardized or accounted for. The main dimensions and the respective subdimensions of data quality relevant to this case study are reliability and relevance. This means that to improve the performance of AI models, the tickets in the training set must be improved or selected based on their performance in the stated dimensions.

2.5 AI Empowerment Methodologies

With the main variable of data quality and its relevant dimensions presented, systematic approaches toward improving that variable can now be explored. Baresi et al. and Reinhardt et al. have developed automated AI enhancement methodologies and proven their effectiveness, but these methodologies consist of extensive data science efforts and are, therefore, less accessible in the presence of resource constraints. However, these methodologies and their conceptual approaches towards data quality improvement will act as a guideline for a more basic approach in this research.

Reinhardt et al. (2023) proposed a ticket analytics pipeline that improves AI-ITSM models, specifically Incident Management processes, by selecting data entries based on their quality. The proposed pipeline consists of five data analytical steps (DR1-5) that allow filtering out low data quality, and ultimately improves the accuracy, recall, precision, and F1 scores of resulting AI models for multiple common Machine Learning-based classifiers. Conceptually, in DR1 the characteristics of data quality are laid out depending on whether the textual data regards issue description or resolution description. In the scope of this research, the same should be done for the issue description using different methods. In DR2 the data is preprocessed, this is especially important since the textual data for issue descriptions are initial emails. Therefore, links, email signatures, attachments and other special characters and blank spaces must be removed. In DR3, topic clustering is used to ensure that only important topics are included in the dataset and redundant tickets are removed. Conceptually, this means that a ticket must contribute to AI model in the relevant context. If the ticket is not part of the Incident Management process, then it contaminates the training set. This step can be mimicked using more basic methodologies using expert knowledge to filter out non-relevant topics. In DR4, the tickets in the training set are scored in multiple dimensions of data quality as previously presented. And in DR5, the tickets are given a total, normalized nominal score that represents their usefulness based on data quality. DR4 and 5 combined allow for filtering based on data quality, and this ultimately enhances the data set as it removes low data-quality entries. As previously mentioned, in this research an attempt will be made to effectively enhance the training set in a similar way, using more basic and therefore accessible methods.

Baresi et al. (2020) presented a similar methodology in which ticket data quality is measured in extensive statistical analysis named ACQUA. ACQUA is a 15-step methodology aimed at predicting a ticket's data quality using the initial textual data only, differing from the ticket analytics pipeline methodology, which emphasizes determining data quality by a data entry's effect on the final AI model. The effectiveness of deductive analysis of the initial issue description and other metrics, such as text length, in determining data quality must be considered when creating a basic training set enhancement methodology. For instance, tickets with relatively short text length after preprocessing will likely be of low quality and thus filtered out of the training data set. Furthermore, with the input data as the main determinant of data quality, proper preprocessing to structure innately unstructured data is thus of great importance. Preprocessing is a concept that includes various activities aimed at structuring input data for use in statistical analysis, in this case, "The preprocessing consists in the following six activities: i) filtering to remove missing data, ii) text transformation to remove special characters and punctuation, iii) domain transformation to eliminate from the ticket text partial or blank parts, iv) encoding to transform values and labels onto pre-defined numbers, v) tokenization to obtain the list of words, and vi) stemming to normalized words to a root form." (Baresi et al., 2020).

2.6 The gap in the existing research in AI-Incident Management

The previously highlighted research by Baresi et al. (2020) and Reinhardt et al. (2023) are indicative of the existing research in the field of AI implementation in Incident Management. They rely on extensive data analysis to reach the desired outcome. Alternative options to improve low-quality data sets remain largely unexplored, specifically options that are accessible for small to medium-sized businesses. Heinrich et al. (2019) have proven the positive correlation between data quality and the performance capabilities of AI trained on that data, and others have researched how to determine and improve that data quality as a means to improve AI performance. However, the methods explored for doing this are extensive and, therefore, would require outsourcing that process. Which is not directly accessible to companies with limited budgets for such developments. Finding easier alternatives for improving the data quality of data sets than previously researched is the gap in knowledge that this research addresses.

3. METHODOLOGY

To answer the research questions, 'What dataset enhancement methodologies are known in prior research on AI-Incident Management that can improve the AI classification of tickets?' and 'How can these methodologies be translated for small and medium-sized businesses with limited outsourcing resources to improve ticket AI classification?' the following quantitative research had to be conducted. To ultimately evaluate the data quality enhancement methodologies, a predictive AI model had to be created first from the native dataset. Then different enhancement methodologies could be drafted that a company can do themselves and their effectiveness tested and evaluated against a native unenhanced version.

3.1 Data collection

The data set for this research was collected from the ticketing system backlog of an IT service provider in the nautical tourism sector. The data set consisted of more than ten thousand data entries, with each data entry being a single ticket, including the initial text data that caused the ticket and their manual classifications. Representatives from the company have confirmed that the data was of low quality, causing difficulties in performance reporting and service improvement. This is partly due to their client base, which consists of various personnel working on the ships with varying nationalities combined with high employee turnover, causing the initial ticket descriptions to be incomplete, inaccurate, and misspelled without any prospect of improvement. The second stated reason for low data quality is that Incident Management is not seen as one of the core functions of their business, limiting the availability of resources. With a lack of personnel and time, the Service desk is unable to correctly classify all tickets and reduce contamination of the data set. The previously stated factors lead to a data set with low data quality, wrongly classified tickets, and contamination of the data set with generated tickets that should not be part of the data set. This also meant that previous attempts at AI implementation into the IM process had failed as the predictive capabilities fell short and consequently did not reduce workload as intended.

3.2 The AI model, classifiers, and evaluation

In this research, the Orange data mining tool was used to process the data, create the AI model, and evaluate the different data quality enhancements. The tool may not allow for direct integration into any ticketing application, but it is helpful to clearly conceptualize the different steps of this research and to compare the findings in isolation. Orange data mining offers the tools necessary for this research, such as preprocessing, bag-of-words, the previously mentioned classifiers, and the ability to connect all variations of the model to performance measurements like a confusion matrix. The easy-to-use preprocessing feature is especially important since the text data in IM will primarily consist of emails, which are inherently unstructured and include many confusing elements like blank spaces and email signatures that need to be removed.

3.3 The data-quality enhancement methodologies

As the theory on data-quality enhancement presented, there are two approaches when filtering out data entries that must be replicated to an extent without outsourcing the process. Firstly, Reinhardt et al. (2023) scored each entry individually based on the dimensions of data quality and compared the native model's performance to enhanced versions that had low-scoring entries filtered out. The second approach presented by Baresi et al. (2020) emphasized prediction and thus scored the text data at an earlier stage to achieve the best possible prediction performance. These two approaches indicate that efforts must be made to select based on the initial text data as a form of preselection, and subsequently on the enhanced data set after establishing a working model to further improve its performance. Both approaches rely on extensive data analysis which cannot be replicated given the constraints. Alternatively, in-depth expert knowledge from the subject of this case study was used to approach the results that data analysis would have given. This alternative approach of pinpointing low data-quality entries, in combination with other enhancement efforts presented in this research, would allow businesses to enhance AI performance without outsourcing the process if the resulting AI model performance is sufficient. In short, expert knowledge was used to preselect the data set and filter out subjectively low-scoring entries. In the next step, an AI model could be drafted using different classifiers and preprocessing techniques. And finally, additional efforts were made to optimize the enhanced model even further. The results were then evaluated by comparing the performance of the same conceptual model with the native version of the data set, with minor preprocessing efforts to achieve a minimally viable model, against the performance of the model with the enhanced version of the dataset.

4. FINDINGS

In this section, both the AI model creation process as well as the model's performance results will be presented to determine whether the basic data quality enhancing steps affect the prediction accuracy. This includes generating a minimum viable AI model based on the unaltered data set, the subsequential steps taken to enhance the data set within the same model layout, the following adjustments made on the model layout itself to further enhance performance, and the test results of these steps individually.

4.1 The base model

First, a minimally viable model must be made in the Orange data mining tool with the native data set to allow for further enhancement of the data set and settings, and to compare the enhanced version against this native data set and its subsequent model. See Figure 3 for the base model layout. Moving from left to right, the Corpus widget labeled as "Native" represents the unaltered data set with a Corpus Viewer attached to view the individual data entries. Following that are the Preprocess Text and Bag of Words widgets. Combined, they preprocess the data and transform them into text features consisting of multiple combinations of words that indicate the semantic meaning of the data entries. The Word Cloud widget was used to visually present the most common words; this allowed unnecessary words inherent to the email format, like "dear" and "Hello," to be manually excluded in the Preprocess Text step.

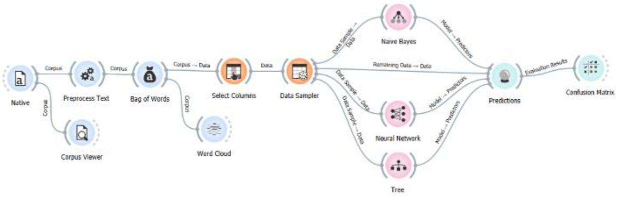
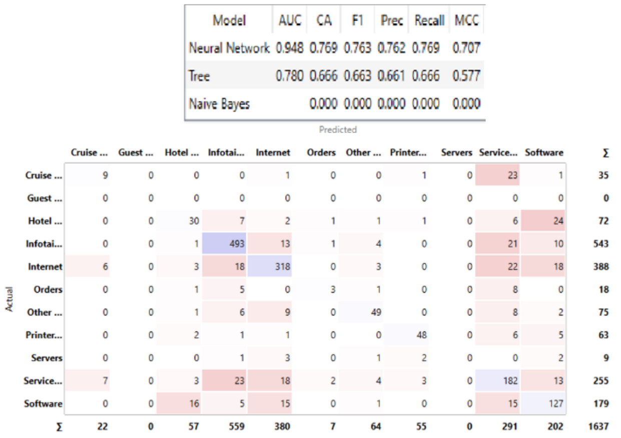


Figure 3: Base model layout using the native, unenhanced, data set

Following that is the Select Columns widget that allows the selection of the target variable, which is the column of issue categories in this case. The Data Sampler widget splits the input data into a training data set and a test data set, as can be seen by the labels of the connections to the classifiers and the Prediction widget. The last two widgets, Predictions and the Confusion Matrix, are for evaluating the model's performance. The common widget is Test&Score, but this includes cross-validation against the training set to measure overall performance. This is in contrast to the Prediction widget used in this research, which only measures the performance of the actual predictions, which is more applicable given the low quality of the training data. See Figure 4 for the performance metrics of the base model. The main metric of importance is Classification Accuracy (CA). The performance metrics also showed that the most viable classifier for this research was the Artificial Neural Network classifier, and therefore it was used in all further models.



Model	AUC	CA	F1	Prec	Recall	MCC
Neural Network	0.948	0.769	0.763	0.762	0.769	0.707
Tree	0.780	0.666	0.663	0.661	0.666	0.577
Naive Bayes		0.000	0.000	0.000	0.000	0.000

Predicted

Actual	Cruise ...	Guest ...	Hotel ...	Infotai...	Internet	Orders	Other ...	Printer...	Servers	Service...	Software	Σ
Cruise ...	9	0	0	0	1	0	0	1	0	23	1	35
Guest ...	0	0	0	0	0	0	0	0	0	0	0	0
Hotel ...	0	0	30	7	2	1	1	1	0	6	24	72
Infotai...	0	0	1	493	13	1	4	0	0	21	10	543
Internet	6	0	3	18	318	0	3	0	0	22	18	388
Orders	0	0	1	5	0	3	1	0	0	8	0	18
Other ...	0	0	1	6	9	0	49	0	0	8	2	75
Printer...	0	0	2	1	1	0	0	48	0	6	5	63
Servers	0	0	0	1	3	0	1	2	0	0	2	9
Service...	7	0	3	23	18	2	4	3	0	182	13	255
Software	0	0	16	5	15	0	1	0	0	15	127	179
Σ	22	0	57	559	380	7	64	55	0	291	202	1637

Figure 4: Base model performance, Predictions scores & Neural Network Confusion Matrix

In this research, "Bag of Words" was used as the default feature creation method. The other common method, called "Document embedding," was also tested. Document embedding turns the word combinations, called features, into vectors and places them in a vector space, thus keeping features with similar semantic meanings closer to each other. This method appeared to be more fitting since the data set contained many misspellings and synonyms that described the same meaning. However, in the tests, all performance metrics were lower compared to the models using the more generic Bag of Words. The most likely explanation is that all tickets are created for issues within a certain business scope and, therefore, already have similar semantic meanings.

4.2 The enhanced models

With the base model layout and settings established, the data quality of the data sets could now be improved, and the effects of these enhancements could be measured against the base model with the native data set.

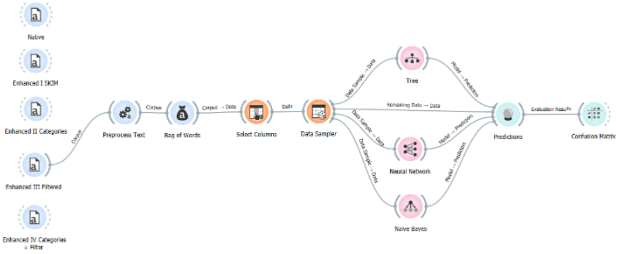
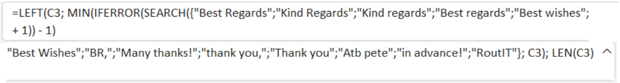


Figure 5: Example of the layout and how the models were trained/tested per data set.

Based on the results of the base model and the underlying theory, a few data set enhancement methods could be outlined and separated to measure their individual and combined effects on the model's performance. See Figure 5 for the overview of the models with the native and enhanced data sets. In Enhanced I, the data set had the email signatures removed from the text data, aiming to reduce obsolete information (features) from the text data. Removing the email signatures is not a standard feature in the Orange Data Mining tool, so a workaround was used by applying the formula shown in Figure 6 on the text containing columns in the data set.



```
=LEFT(C3; MIN(IFERROR(SEARCH({"Best Regards";"Kind Regards";"Kind regards";"Best regards";"Best wishes";
+ 1)) - 1)
"Best Wishes";"BR,";"Many thanks!";"thank you,";"Thank you";"Atb pete";"in advance!";"RoutIT"}; C3); LEN(C3)
```

Figure 6: Email signature skim formula used in Enhanced II

The formula essentially reproduces all text information “left” of the specifically stated closing sentences, such as “Kind regards” and “Best wishes,” if present.

In Enhanced II, the categories of the data set were altered by combining low-frequency and predictability categories with their semantic neighbors based on the base model's results in the confusion matrix. For example, the rare category “Servers” was combined with the more frequent category “Internet,” as the symptoms of these issue categories are very similar. The distinction between server defects and other core component defects was now made on the lower sub-category level.

In Enhanced III, the low data quality entries were removed based on the theoretical dimensions of data quality. This was done by experts in the respective business who could pinpoint the common indicators of low-quality tickets. The entries that were filtered out were mostly the remnants of merged tickets, which are created when two tickets concern the same issue and are merged, and business communication addressed to the Servicedesk resulting in the creation of a ticket unrelated to service issues. But also the automatic notifications of internal monitoring systems addressed to the support team.

Lastly, the data set Enhanced IV was a combination of previous enhanced data sets that had a positive effect on their model’s predictive performance. So, the method of category combination from Enhanced II and the method of removing low data quality entries from Enhanced III were combined in Enhanced IV.

The removal of tickets based on the completeness dimension of data quality also became apparent, as the error “Missing instances” above the Prediction widget was not present in both the Enhanced III and Enhanced IV models.

Table 1: Model performance per data set using the Neural Network classifier

	Description
Enhanced I	Email signatures skimmed in the text data
Enhanced II	Low-frequency categories combined based on semantic meaning
Enhanced III	Low data quality tickets filtered based on expert knowledge
Enhanced IIII	Enhanced II and III combined

	Neural Networks			
Data set	**CA**	**F1**	**Precision**	**Recall**
Native	0,769	0,763	0,762	0,769
Enhanced I	0,751	0,746	0,751	0,751
Enhanced II	0,782	0,779	0,781	0,782
Enhanced III	0,810	0,801	0,800	0,810
Enhanced IIII	0,817	0,810	0,808	0,817

Table 1 shows the individual models' predictive performance. Taking the CA (Classification Accuracy) as the leading variable, Enhanced I resulted in a decrease in performance, while Enhanced II, Enhanced III, and the combined Enhanced IV resulted in increased performance.

In Enhanced I, the intent was to clean up the data by removing the email signatures and, therefore, improve prediction performance by removing redundant information, increasing the contrast between tickets. However, with all other things kept equal, the direct result of the signature removal is a decrease in CA. That means that the email signature contains information that is helpful in the prediction of the target variable of ticket category. This is probably due to a correlation between requester and category; for example, engineers are more likely to raise tickets regarding core components.

In Enhanced II, combining categories based on their semantic meaning resulted in a relatively small increase in prediction performance. Conceptually, this was expected to have a positive effect on performance, but given that the selected categories that were combined into their closest semantic neighbors had low ticket counts, the overall benefit was minimal.

The largest performance increase was gained in Enhanced III. Filtering tickets out of the dataset that score low on data quality, interpreted by an expert on the content matter in this case, improved the prediction performance of the model, as indicated by the theory. In the presented related research, extensive data analysis was done to score each entry based on the dimensions of data quality. However, the use of content-specific expert knowledge to filter out low-data-quality tickets as an alternative to that approach gained positive results.

Ultimately, combining the positive enhancement methodologies in Enhanced IV resulted in an almost five percent increase in prediction performance. While attempting to improve a model's CA, it has to be noted that the data set itself contains human errors in the target variable, which affects the test set and, subsequently, the CA score.

Using the same layout and methods, the models were also trained to predict different target variables as a test to see the extent of AI automation possibilities in this case study. The target variables tested were the site location where the problem occurred, the Agent to which the ticket was assigned, and the highly specific Sub-Category that follows after the initially targeted Category. However, the trained models performed far below the benchmark and were considered not viable. The text data did not always contain enough information to predict where the issue occurred or to determine the specific underlying problem. Predicting the Agent variable was also unsuccessful, as there very often was no correlation between who handled the ticket and the content of that ticket. This could theoretically be resolved by designing this variable in advance so that it does not contain names but rather departments with their own respective fields of expertise.

5. CONCLUSION & DISCUSSION

5.1 Theoretical Contribution

The findings of this research presented in the previous section show that using an existing low-quality data set, AI prediction performance increases can be gained through relatively simple methods aimed at improving the overall quality of the data. The positive effect that increasing the data quality in the data set had on the AI's predictive performance supports the prior research by Heinrich et al. on the correlation of these two variables. The various tests showed that making systematic changes to the data set, like combining categories based on their semantic meaning, yielded minimal performance increases. Attempting to remove redundant data, in this case, the email signatures could even yield a negative result. These two things result from the Neural Network classifier determining the weight of the features when classifying the target variable and the supposedly redundant data still contributing slightly to the classification. It turns out that the largest performance increases are gained by addressing the low data quality problem at the per-data entry level, and that expert knowledge on the subject matter can serve as an alternative to extensive data analysis in determining which entries need to be removed or altered in the data set. These findings answer the second research question of "How can these methodologies be translated for small and medium-sized businesses with limited outsourcing resources to improve ticket AI classification?" and address the current gap in the literature regarding data set enhancement for AI implementation purposes for small to medium-sized businesses specifically.

The first research question "What dataset enhancement methodologies are known in prior research on AI-Incident Management that can improve the AI classification of tickets?" was answered in the theory section of this report. Unfortunately, a direct comparison cannot be made between the alternative methodology presented in this research and the methodologies of Reinhardt et al. and Baresi et al. due to a difference in classifiers. In the more related research by Reinhardt et al., the effect of their data quality enhancement methodology was tested using multiple AI classifiers, but not the Artificial Neural Network classifier that was used in this research. Multiple classifiers were tested to be able to make that comparison, but the other classifiers did not work sufficiently with the used data set and its contents. But, to make an indirect comparison, Reinhardt et al. were able to achieve greater performance increases with their methodology on some classifiers.

5.2 Practical contribution

Although the prediction accuracy increase may appear minor, the size of the ticket influx can warrant this additional effort during the AI implementation phase. This means that with the increased accessibility of AI creation tools, a Service Manager can decide to take the steps explored in this research and create an AI model specific to their business for ticket classification purposes despite working with a suboptimal data set. Outsourcing the AI creation and data set enhancement process to external experts might yield higher prediction performances, and Service Managers will have to weigh the ramifications of probable lower prediction performance against the cost of outsourcing. When the classification done by AI is used for data analysis of service performance and/or automated escalation down the support levels, then there will be ramifications for prediction errors. Depending on the nature of the business and the error itself, these ramifications can range from minor inconveniences to very costly mistakes.

A Service Manager can also decide to take measures to ensure a higher-quality data set beforehand. Systematic changes such as implementing a ticket format that demands the problem holder to fill in certain fields, including drop-down lists to avoid synonyms and misspellings where possible, will ensure a higher quality of each entry. Especially in the reliability dimension of data quality, the data's completeness, consistency, and accuracy can be ensured to a certain extent in this way. These changes will have to be made preemptively with AI implementation in mind and run for a certain period to result in a sufficiently sized data set with higher data quality. In practice, for small and medium-sized businesses with a lesser influx of tickets, gathering this sufficiently sized training set can take years. This means that this research, which enhances the existing data set, can serve as a solution if the intent is to implement AI in the Incident Management processes sooner, under the condition that systematic changes are still made to ensure higher data quality of new tickets following that AI implementation. If these changes are not made, the model that was not trained to classify the low-quality tickets will be unable to classify these tickets correctly.

5.3 Future research

As stated in the practical contribution, the data set enhancement methodologies explored in this research allow small to medium-sized businesses to enhance the initial data set and start their AI implementation process with a feasible AI model. However, the following systematic changes that should be made to ensure higher data quality in the influx of tickets still need to be researched. The effects of combining categories based on their semantic meaning for example indicate that such systematic changes must be made with AI implementation in mind, since the distinction in certain categories made practical sense but with the symptoms of the issue being so similar and the escalation path the same, it was better to combine them for AI implementation purposes alone. This example highlights that a business's design of the processes within the Incident Management framework matters for the effective implementation of AI, and this is yet largely unexplored.

Additionally, the scope of this research, given the case study subject, was focused on emails as the only format of ticket data. Specifically, the text data of the initial email of the problem holder. But in practice many companies use both contact by email and phone as the available methods of ticket submittal for their clients. This means that similar research must be done for small to medium-sized businesses on effective AI implementation with that specific data format.

6. REFERENCES

Conger, S., Winniford, M., & Erickson-Harris, L. (2008, 14-17 August). Service management in operations. Paper presented at the Fourteenth Americas Conference on Information Systems, Toronto, ON, Canada.

Wang, R. Y., & Strong, D. M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12(4), pp 5–33.

Zschech, P.: Beyond descriptive taxonomies in data analytics: a systematic evaluation approach for data-driven method pipelines. Inf. Syst. E-Bus Manage. 1–35 (2022). https://doi. org/10.1007/s10257-022-00577-0

Batini, C., Barone, D., Mastrella, M., Maurino, A., Ruffini, C.: A framework and a methodology for data quality assessment and monitoring. ICIQ, pp. 333–346 (2007)

Revina, A., Buza, K., Meister, V.G.: IT ticket classification: the simpler, the better. IEEE Access 8, 193380–193395 (2020). https://doi.org/10.1109/access.2020.3032840 19.

Koehler, J., et al.: Towards Intelligent Process Support for Customer Service Desks: Extracting Descriptions from Noisy and Multi-lingual Texts, S 36–52

Marcuzzo, M., Zangari, A., Schiavinato, M., Giudice, L., Gasparetto, A., Albarelli, A.: A multi-level approach for hierarchical Ticket Classification. In: Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), pp. 201–214 (2022)

Zicari, P., Folino, G., Guarascio, M., Pontieri, L: Discovering accurate deep learning based predictive models for automatic customer support ticket classification. In: Proceedings of the 36th Annual ACM Symposium on Applied Computing. ACM, New York, NY, USA (2021)

Agarwal, S., Aggarwal, V., Akula, A.R., Dasgupta, G.B., Sridhara, G.: Automatic problem extraction and analysis from unstructured text in IT tickets. IBM J. Res. Dev. 61(1):4:41–4:52 (2017). doi:https://doi.org/10.1147/jrd.2016.2629318 2

Galup, S. D., Dattero, R., Quan, J. J., & Conger, S. (2009). An overview of IT service management. Communications of the ACM, 52(5), 1

International Organization for Standardization. (2005). ISO/IEC 20000-1:2005 Information technology - Service management - Part 1: Service management system requirements. ISO.

Baresi, L., Quattrocchi, G., Tamburri, D. A., & Van Den Heuvel, W. (2020). Automated Quality Assessment of Incident Tickets for Smart Service Continuity. In *Lecture notes in computer science* (pp. 492–499). https://doi.org/10.1007/978-3-030-65310-1_35

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(0), 2. https://doi.org/10.5334/dsj-2015-002

Reinhard, P., Li, M. M., Dickhaut, E., Peters, C., & Leimeister, J. M. (2023). Empowering Recommender Systems in ITSM: A Pipeline Reference Model for AI-Based Textual Data Quality Enrichment. In *Lecture notes in computer science* (pp. 279–293). https://doi.org/10.1007/978-3-031-32808-4_18

Heinrich, B., Hopf, M., Lohninger, D., Schiller, A., & Szubartowicz, M. (2019). Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems. *EM*, *31*(2), 389–409. https://doi.org/10.1007/s12525-019-00366-7