Masters Seminar Report

State-of-the-art Evaluation for Incident Ticket Resolution with NLP & Process Mining

By

Saahithi Pradhan Paramalla

7015405

Department of Computer Science

Saarland University

German Research Center for Artificial Intelligence (DFKI)

Supervisor

Prof. Dr. Peter Loos

Content

Table of figures

Table Directory

1 Introduction

1.1 Motivation

Incident management is a crucial step in ensuring smooth functioning of businesses across various industries. From IT outages to customer complaints, incidents can disrupt operations and impact customer satisfaction. To effectively handles these incidents, companies often rely on incident tickets, which document the details of each issue but the enormous volume of tickets in large-size industries can be overwhelming, making it challenging to identify patterns and optimize the incident management. Often, incident resolutions can span from a few days to several months. Delays in incident solving may occur due to various reasons, such as the R&D team needing time to address the issue, lags from IT support, or global support taking time to consult with the customer. It's crucial to pinpoint where these delays occur within the incident management process. These insights allow for targeted improvements, aiming to reduce processing times at different stages of issue resolution.

This is where Natural Language Processing (NLP) and process mining come into play where NLP helps us classify the unstructured ticket text through which activities can be extracted which enable us to generate process mining model. Process mining enables us to visualize and understand the underlying workflows of incident management, identifying bottlenecks and inefficiencies. Large Language Models (LLMs) such as GPT-3.5 offer remarkable accuracy without requiring fine-tuning, thanks to the prompt engineering, in-context learning, and few-shot learning. By combining NLP and Process Mining, we can develop a comprehensive approach to incident management, categorizing tickets into different classes and generating a process model that provides valuable insights into various operational metrics, including response time, bottlenecks, incident resolution efficiency, and resource utilization. Ultimately, this improved process can lead to higher customer satisfaction, reducing churn rates and fostering long-term relationships with customers.

1.2 Problem Statement

The challenges encountered in incident management within businesses present an intriguing ground for research exploration. At the core of these challenges lies the complex interplay between operational inefficiencies, customer satisfaction, and the ability to adapt to dynamic business environments. Utilizing Large Language Models (LLMs) like GPT-3.5 (Brown et al., 2020) & BERT (Devlin et al., n.d.), not only provides state-of-the-art results but also eliminates the need for training because of their real-world knowledge. GPT-3.5 has shown groundbreaking results for a lot of computing tasks as shown in (Amin et al., 2023). Incident tickets usually contain a lot of typos, grammar errors, emojis and this data needs to be thoroughly cleaned to obtain good results for any task. Since training the models is not required for most of the tasks, there is no need to label large amounts of data.

The main research questions of this thesis are:

- How can Large Language Models (LLMs) be effectively integrated into the classification of incident management tickets without any training to enhance the accuracy and efficiency of incident resolution within an organization?
- To what extent can the combination of LLMs and Process Mining techniques contribute to the reduction of churn rate & optimization of the incident management process?

1.3 Structure of the report

In Section 2, we will discuss the background and the techniques related to Natural Language Processing (NLP) and Process Mining. Section 3 discusses about the methodology for documenting the literature search and Section 4 gives a detailed description of the recent research papers where we address the research gap based on the concept matrix. Section 5 deals with the discussion of the NLP techniques & their results and Section 6 concludes the report by addressing the possible outcomes of the research questions.

Keywords: Process Mining, Natural Language Processing (NLP), Large Language Models (LLMs), Incident Management classification

2 Background

2.1 Natural Language Processing (NLP)

Recent advancements in the field of NLP not only give promising results for any natural language task, but also reduce the need for training & longer computation times. Large Language Models (LLMs) like GPT-3.5 and BERT possess extensive pre-trained real-world knowledge and require very minimal to no fine-tuning to achieve state-of-the-art performance on any language generation and understanding tasks.

Large Language Models (LLMs):

Large language models (LLMs) represent a significant leap forward in artificial intelligence, capable of processing and generating human-like text with remarkable fluency. These models are trained on huge datasets of text and code, allowing them to perform a multitude of tasks without any training that were once thought to be the exclusive domain of human intelligence.

The applications of LLMs extend across various industries, fostering innovation and transforming workflows. In the business world, LLMs are utilized for tasks such as automating customer service interactions, conducting market research through sentiment analysis of social media data. Scientific research is also benefiting from LLMs, with applications in analyzing vast scientific literature, summarizing complex research papers, and even assisting in drug discovery by identifying potential target molecules.

Manual data labeling is a meticulous task that requires human expertise to categorize, tag, or annotate raw data. This often translates to time-consuming work, as large datasets necessitate labeling a vast number of individual entries. Furthermore, the accuracy and consistency of labels depend on the expertise of the human labelers. Inevitably, this expertise comes at a cost, with skilled labeling teams requiring a significant amount of time. The combination of time, expertise, and cost associated with manual data labeling can be a major hurdle for companies and researchers.

As depicted in Figure 1, task adaptation techniques like Fine-tuning, Prompt Engineering, Chain-of-Thought (CoT) prompting, and In-Context Learning (ICL) have

evolved recently which eliminate the need for fine-tuning LLMs and unlock their true potential. They also provide the feasibility to lessen the burden of manually labelling the data, which is one of the important & most required tasks for training or fine-tuning the model. The work done by (Mosbach et al., 2023) proves that fine-tuning and ICL techniques achieve comparable results, and their performance improves as the model size increases. The authors also conclude that fine-tuned models can generalize OOD better than the models that use ICL.

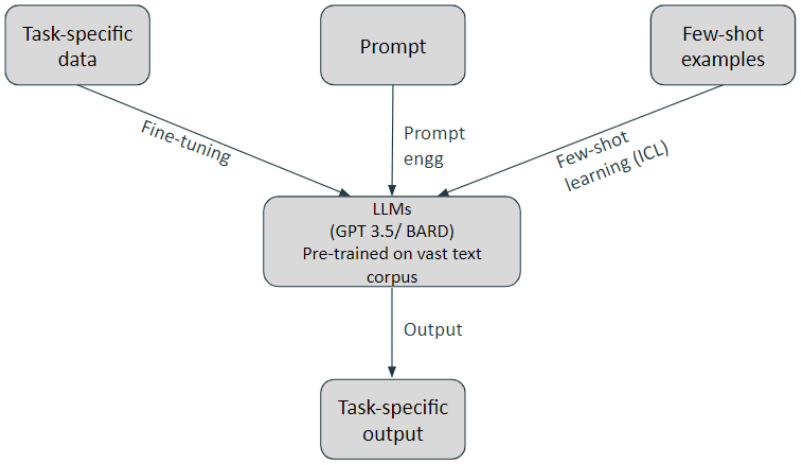

Figure 1: Illustration of adaptation techniques with LLMs

Fine-tuning:

Fine-tuning is a supervised learning setup in NLP where the model is trained on a particular task and its parameters are updated. Fine-tuning leverages the pre-trained knowledge of an LLM and tailors it to a specific task. By exposing the LLM to a targeted dataset of text and code relevant to your domain, you can significantly improve its performance on specific tasks. Fine-tuning unlocks the true potential of LLMs, transforming them from general natural language processors into specialists at tasks like creative story writing, translating languages with great efficiency, or answering user queries in an informative way, all within the context of your specific domain.

By exposing the LLM to a targeted dataset relevant to your field, like legal documents for summarization or e-commerce details for a customer service chatbot, you can significantly improve its performance for those specific tasks.

Prompt Engineering:

Prompt Engineering has evolved as a game-changer in recent NLP tasks acting as bridge between human context and machine output. Just like crafting the perfect book guides a student's learning, well-designed prompts leverage LLM's towards desired outcomes. This involves crafting clear instructions or prompts that provide context and guide the model's response in terms of style, tone, content accuracy, and even mitigating biases. Recent NLP tasks have seen prompt engineering shine in various applications like story or content writing, language generation and classification.

Suppose we want to perform multi-class classification on movie reviews using GPT-3.5, we can design the input prompt like this:

Input prompt: Imagine you're a film critic with a sharp eye for emotional tone. Read this review: 'The storyline was very exciting and pretty eye catching. But the choice of actors could have been better'. Would you categorize this review as a thumbs up (positive), a thumbs down (negative), or a shrug (neutral) for the movie?

In-context Learning (ICL) & Few-shot Learning:

In-context learning (ICL) leverages the principles of prompt engineering & empowers LLMs to learn about the new task "on the fly" instead of training or fine-tuning. As we have seen in prompt engineering, ICL leverages this by providing examples and explanations in the input prompt so that the model can learn well about the task. This eliminates the need for extensive data labeling and fine-tuning, also provides focused learning to the model making ICL incredibly efficient.

In-context Learning (ICL) and few-shot learning are similar in a way that they both provide examples to the language model in the prompt. ICL concentrates on providing the model with additional context of the scenario along with the examples whereas few-shot learning provides examples without any context so that the model can learn and perform the task.

Let's imagine a scenario where the language model needs to write a report. With In-Context Learning (ICL), you can guide the language model to generate efficient reports on any domain you want:

Input prompt with ICL: Imagine you're writing a report about renewable energy sources. Describe the benefits of solar power and wind energy compared to traditional fossil fuels. Include examples of countries or regions that have successfully implemented these renewable energy sources and discuss their impact on the environment and economy.

There are different types of few-shot learning techniques, and these techniques basically differ by the number of examples provided to the model. For example, if the model must perform a classification task as shown in Table 1 – in zero-shot setting, the model will not be provided with any training examples, and it must perform the task based on its pre-trained knowledge. In the few-shot setting, the model will be given any number of training examples (1-shot, 5-shot and so on) from the target classes to improve accuracy over zero-shot setting.

Zero-shot setting	**1-shot setting**
Classify the following movie review: The movie was thrilling to watch. Sentiment:	Classify the following movie review: The storyline of the movie was not good. Sentiment: Negative Classify the following movie review: The movie was thrilling to watch. Sentiment:

Table 1: Examples of few-shot learning

Chain-of-Thought (CoT) prompting:

Unlike traditional prompting that simply seeks an answer, CoT prompts dives deeper into the reasoning process of Large Language Models (LLMs). CoT prompts require the LLM to reveal its reasoning steps alongside the final answer. This output can be achieved by breaking the task into several smaller steps within the prompt itself.

For instance, a CoT prompt for sentiment analysis might ask the LLM to first identify key phrases that express positive or negative emotions, then explain how those phrases contribute to the overall sentiment of the text. Chain-of-Thought (CoT) prompting can be done in two different settings, namely zero-shot CoT, and few-shot CoT as shown in Table 2.

Zero-shot CoT pushes the boundaries even further where the model reasons through a task and reveals its thought process without any specific training examples. Few-shot CoT provides a middle ground by providing the LLM with just a few relevant examples to nudge it in the right direction. This can improve the model's performance on unseen tasks compared to zero-shot CoT.

Zero-shot CoT	**Few-shot CoT**
Q: B has 3 apples. He buys 2 bags of apples. Each bag has 3 apples. How many apples does B have now? A: Let's think step by step	Q: B has 3 apples. He buys 2 bags of apples. Each bag has 3 apples. How many apples does B have now? A: B had 3 apples. 2 bags of 3 apples each is 6 apples. 3+6=9 apples. The answer is 9. Q: C has 14 books. Half of the books are English books and other half are German books. How many German books does C have? A:

Table 2: Examples of CoT prompting (ref: (Kojima et al., 2022)*)*

2.2 Process Mining

In recent years, many software companies have been adapting to cloud-based services and subscription-based models which provides more standard for the software vendors in the industry. These subscription-based models provide timely updates, security and flexibility to the customers but also increase the churn rate because of the ongoing costs over time. This may lead to some of the customers not

renewing the subscriptions and this leads to less revenue for the company. This has a significant business impact and thus, companies strive to increase their customer satisfaction and their service to prevent the churn of the customers. Every time the customer raises an incident ticket, the efficiency of the incident solving is measured by 'Resolution time'. Resolution time is defined as the time from when the customer opened the ticket to the time when the ticket was closed. Huge resolution time impacts customer satisfaction and this can lead to higher churn rates.

The Net Promoter Score (NPS) is a market research metric and a key performance indicator (KPI) for analyzing customer loyalty and satisfaction. The NPS survey tries to determine if the customer would recommend the company's product to new customers. The customers are requested to rate the product ranging from 0 to 10 where the customers are categorized into three groups namely, promoters, passives, and detractors.

Figure 2 depicts how NPS score can compromise on customer satisfaction and have an overall effect on the company growth, Promoters are those loyal customers who would likely recommend the product to others whereas passives are those who are satisfied with the product but not likely to recommend it. Detractors are the unhappy customers who would create a negative impact on the product and company. To calculate the final NPS score which ranges from -100 to +100, companies simply subtract the percentage of detractors from the percentage of promoters. Higher resolution time indicates a higher percentage of detractors or passives.

A positive NPS score indicates that the company has more promoters than detractors and vice versa. But the NPS score is calculated is only calculated once or twice in a year and there exists a high possibility that many customers do not fill the NPS survey. Hence companies have no overall view of how customer satisfaction can lead to churn and have no visibility of early indicators for churn because of the performance of incident management and customer satisfaction. This is where the incident management process model can help software companies to obtain timely overview of how the resolution times are maintained for each issue.

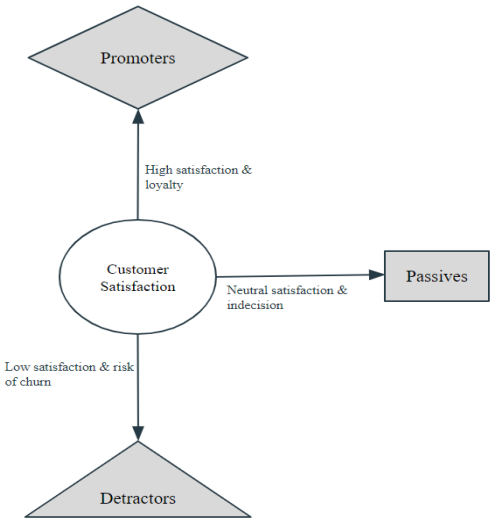

Figure 2: Effect of NPS on churn

Process mining has become a game-changer, offering software companies a powerful tool to analyze and optimize their development processes dynamically. Unlike traditional methods that rely on manual analysis, process mining leverages real-time data. It extracts valuable insights from event logs – digital footprints left behind by development tools and repositories. This data allows process mining to reconstruct the actual flow of work, uncovering deviations from the planned ideal.

Incident management functions as a centralized system for tracking and resolving issues. By creating and logging tickets for each problem, developers gain a clear overview. Each ticket contains several numbers of updates throughout its life cycle. These updates convey deep insights about lags in the incident management, time taken to resolve sub issues within an issue and communication with customers regarding the issue. It also allows for prioritization of critical issues, improved communication within teams, and faster resolution thanks to detailed information within tickets. Extracting a process model from incident management tickets not only builds a roadmap for effectively solving the issues but also ensures customer satisfaction and reduces churn rate providing long-term loyalty.

Firstly, it prioritizes critical issues and facilitates swift resolution, reducing downtime for users. Secondly, it eliminates confusion and wasted time by clearly defin-

ing roles and communication channels, leading to improved efficiency. Furthermore, the model encourages thorough investigation and documentation, building a knowledge base of root causes that prevents similar issues from recurring. Transparency is also enhanced, as everyone involved has a clear view of ongoing problems and their progress. The process model acts as a safety net, ensuring smooth workflows even during unexpected glitches, ultimately leading to faster development cycles, higher quality software, and a competitive edge.

3 Methodology

To understand the research gap existing with respect to the motivation of the thesis, we define different concepts and search terms which help us to obtain the statistics of the research established over the course of years. To better understand the research gap, (Webster & Watson, 2002) discusses development of two different matrices, namely author-centric matrix, and concept-centric matrix.

In this report, we develop a concept-centric matrix which summarizes the various concepts which motivate the problem statement of this thesis and helps us find the research gap in the defined concepts. The concepts defined in the matrix determine the main framework of the literature survey and serve as a foundation for understanding the gap in existing literature.

In addition to the concept matrix, in this section we report the search statistics based on the defined search terms. Based on the search results, two plots are constructed as shown in Figure 3 & Figure 4. The specified search terms include *support tickets classification, incident tickets classification, extracting process model using NLP, process mining and LLMs, few-shot text classification, classification using GPT models*. These search results were acquired using "EBSCOHost" and "Web of Science" portal and are as follows for 'all fields' and, with a focus of entries within the timeline range from 01-01-2019 to 01-04-2024 to ensure quality and to include latest research advancements.

The search results in the EBSCO portal provided research articles from a huge number of domains which were not relevant for this thesis methodology. To keep the search results bounded by the thesis's domain, the search theme in EBSCO portal is restricted to Natural Language Processing (NLP), Artificial Intelligence

(AI), Artificial Neural Networks, Machine Learning (ML), Data Processing & Analytics, Sentimental Classification.

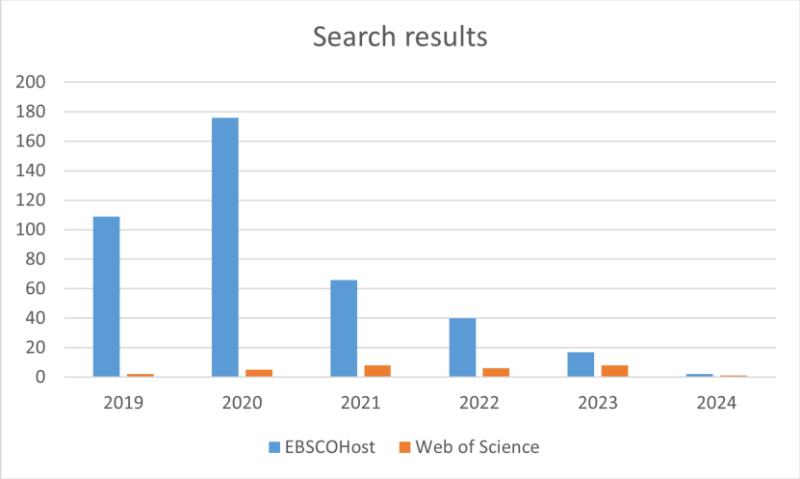

Figure 3: Search results - includes all search terms

The results as shown in the Figure 3 report the search results for the query ‘support tickets classification’ OR ‘incident tickets classification’ AND ‘few-shot text classification’ OR ‘classification using GPT models’ AND ‘extracting process models using NLP’ in the EBSCO host portal. The plot depicts that there is research work done to classify incidents or support tickets in various domains, but the number of articles submitted in recent years is much less. The results also report the search results for the same query as stated above but on the ‘Web of Science’ portal. The results suggest that there has been a significant amount of research including the extraction of process models from text but as the years progressed the number of research papers published have decreased significantly especially on ‘Web of Science’ portal.

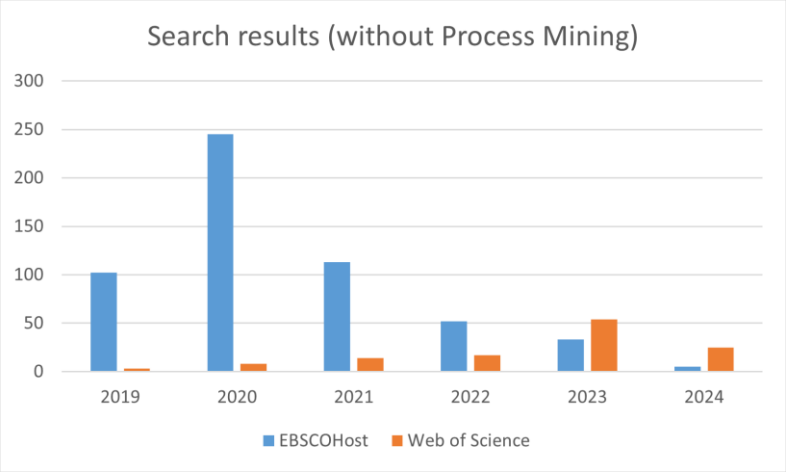

Figure 4: Search results - excludes Process Mining search term

The search results, excluding the 'extracting process models using NLP' search term in the 'EBSCOHost' and ‘Web of Science’ portal, are shown in Figure 4. As depicted, the classification of incident management tickets has gradually decreased over the years, with Transformer models emerging as the predominant approach in recent years. Similarly, the figure illustrates the search results from the 'Web of Science' portal, revealing a substantial volume of research conducted on incident ticket classification using Transformer models in 2023. However, including the search term for process mining in the query yielded considerably fewer results in the same year.

4 Literature Survey

In this section, we will be discussing previous work as presented in Table 3 according to this thesis domain. To accurately distinguish between the significance of Natural Language Processing (NLP) and Process Mining, we provide two distinct subsections below. The first subsection delves into research on classification in NLP, particularly focusing on IT tickets. In the second subsection, we explore how Process Mining aids in achieving solutions for enhanced incident management.

The integration of Natural Language Processing (NLP) with Process Mining represents an untapped potential, with a noticeable gap existing between these domains. The aim is to narrow this gap by harnessing the capabilities of Transformer models and incorporating their outputs into the generation of process models. Transformer models offer promising solution due to their ability to leverage extensive real-

world knowledge and their capacity to perform effectively with real world datasets without the need of extensive training. There is a conspicuous absence of research utilizing Transformer models to enhance Process Mining techniques, underscoring the novelty and significance of the planned approach.

Article	Concepts					
	Transformers	Process Mining	Classification algorithms	Training	Fine-tuning	In-context learning/ Prompt Engineering
(Zhou et al., 2014)			✓	✓		
(Zaidi et al., 2021)			✓	✓		
(Kallis et al., 2019)				✓		
(Wahba et al., 2020)			✓	✓		
(Revina et al., 2020)			✓	✓		
(Liu et al., 2023)	✓		✓	✓	✓	
(Kojima et al., 2022)	✓					
(Louka et al., n.d.)	✓			✓	✓	✓
(Maksai et al., 2014)			✓	✓		
(Mosbach et al., 2023)	✓				✓	✓
(Terragni & Hassani,		✓				

2018)						
(Jlailaty et al., 2016)		✓	✓	✓		
(Qian et al., 2019)		✓	✓	✓		
(Amin et al., 2023)	✓			✓		✓
(Kourani et al., 2024)	✓	✓		✓		✓

Table 3: Concept matrix

4.1 NLP for classification

Incident ticket classification is a special case of multi-class text classification because having a wide range of classes makes it more challenging to distinguish accurately. As discussed above, because of the high number of classes, many classes overlap with the other classes and make the task more challenging. This requires fine-grained decision making and we can't rely on the true semantics of the word. As depicted in Figure 5, true meaning of the incident texts changes in the incident texts and are based on the operational dynamics and protocols of the software company. For instance, textual terms like 'priority' and 'severity' have specialized meanings within the incident management frameworks and are often associated with the seriousness of the issue. The meanings of these texts can also differ among different departments within the company reflecting diverse perspectives in software incidents.

Previous research papers provide important insights into improving the classification accuracy of unstructured data through fine-tuning, in-context, and few-shot learning. One of the early techniques that concentrate on classifying bug reports using text mining and data mining in supervised learning (Zhou et al., 2014) where they classify bugs into top three probabilities (high, medium, low) using Multinominal Naïve Bayes Classifier technique and classify whether a bug report is corrective one or not. Using Bayesian Net Classifier, they make predictions on the bug reports. Research work by (Zaidi et al., 2021) provides an automated solution sug-

gesting mechanism for IT support tickets based on two different approaches namely, “Similarity Search Model” and “End-to-End Model”. The first approach depends on the text similarity in the ticket texts converting texts into vectors and the second approach depends on the encoder-decoder mechanism to predict the recommended solutions for resolving the support tickets. This work also aims to recommend solutions for the support tickets from the past resolutions.

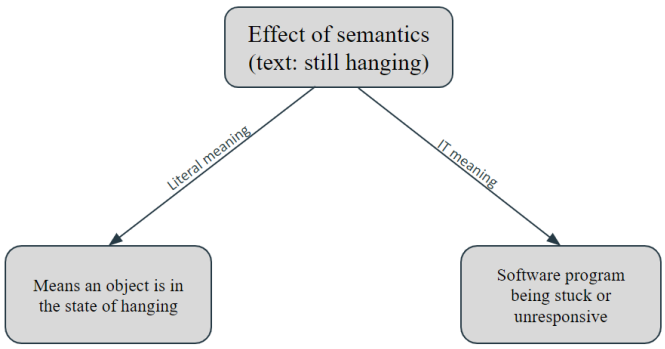

Figure 5: Example showing the change of text semantics

In another study conducted by (Kallis et al., 2019) developed a Machine Learning mechanism using fastText, a tool developed by Facebook (Joulin et al., 2016) to tag the tickets and the framework is called as ‘Ticket Tagger’. This innovative framework was designed to automatically tag IT tickets within GitHub repositories. By integrating a GitHub app into any repository, the authors enabled the seamless tagging of incoming IT tickets based on both their titles and descriptions. However, while these previous works primarily focused on ticket titles and descriptions, it's worth noting that tickets often undergo numerous updates (comments) throughout their lifecycle. The classification of these comments into distinct categories holds significant potential for applying process mining techniques—a key research objective of this thesis.

In the work done by (Maksai et al., 2014), the authors propose a two-step approach which uses hierarchical clustering, and the aim of this paper is to keep the effort for the manual labeling very minimal. The incident tickets are categorized into a multi-

level structure and then classified based on their importance & frequency. They use three real IT datasets where each class is very unbalanced, and the results achieved are reported using precision, recall and F1-score. They observe that the method proposed by the authors works perfectly not just with larger classes but also with smaller classes with the highest F1-score in most of the classes.

Using linguistic features of the incident texts can also make the classification simpler and better as shown in (Revina et al., 2020). In this paper, the authors focus on text representation techniques like linguistic features, and TF-IDF rather than extensive feature engineering required for incident classification. To perform the ticket classification task, the authors compare machine learning approaches and rule-based approaches. To describe the linguistic features of the IT tickets, three levels of text understanding techniques such as objective-level (answering questions of who, when and where), subjective (by semantic analysis) and meta-knowledge (out of the domain info) are implemented.

The results suggest that basic classification algorithms like KNN, Naïve Bayes and Hubness-Fuzzy KNN can also achieve high accuracy when they are combined with efficient text representation techniques which can captures the semantics and meaning of the tickets. However, ML-based approaches outperform the rule-based approach developed in this paper.

Another work by (Wahba et al., 2020), investigates the effectiveness of static word embeddings for classifying IT support tickets. Static word embeddings are simpler and need less computational power compared to other language models. This work compares the performance of static word embeddings with more complex techniques like Pre-trained Language Models (PLMs) for IT support ticket classification. PLMs are powerful language models that can adapt their meaning representation based on context. The models use static word embeddings to represent the textual contexts of the IT tickets and are trained on datasets with labeled IT support tickets with 32 different labels. The results conclude that static word embeddings are not very useful in the classification of IT tickets since they contain a high number of in-domain words which are usually considered as Out-of-Vocabulary (OOV) words in pre-trained embeddings. Using models with contextual embeddings like BERT can provide better performance for this use case.

Pre-trained language models (LLMs) and transformers like BERT (Bidirectional Encoder Representations from Transformers) have revolutionized classification tasks by offering a powerful framework for understanding and processing textual data. These sophisticated models are pre-trained on vast amounts of text from diverse sources, allowing them to learn intricate patterns and semantic nuances inherent in language. By leveraging this pre-training, BERT excels at capturing contextual information and semantic relationships within sentences, enabling more accurate understanding of text compared to traditional methods. Furthermore, BERT's bidirectional nature enables it to consider the entire context of a word or phrase, resulting in richer representations that enhance classification performance. Additionally, fine-tuning BERT on specific classification tasks requires minimal additional training data, making it a highly efficient and versatile tool for various applications.

In (Liu et al., 2023), the authors fine-tune BERT on Microsoft's incident management data to classify the tickets by considering not just the title and description of the incident ticket but also the comments (modified by engineers & system) which makes the incident management system more accurate for the support team of any industry. They use ten fine-grained labels to classify around 76,000 incident management tickets taken from Microsoft Kusto Database.

This work also implements an active learning system which learns about the new incoming tickets and creates new labels accordingly. They experiment with baseline-models like Naïve Bayes with Bag of Words (BoW) and Naïve Bayes with TF-IDF, Logistic Regression with BoW, Logistic Regression with TF-IDF and LS model – text classifier on Microsoft Language Studio. The authors prepare three datasets to show the classification performance difference by texts modified by machines (D-Machine), texts modified by human engineers (D-Human) and a mixture of both (D-Mixture). To design a labeled dataset for evaluation, the authors considered subject matter experts for better understanding and accuracy. They only extract the [CLS] token from the last layer of the fine-tuned BERT and feed it to the Multi-layer Perception (MLP) to perform multi-class classification (named it as TicketBERT). They obtain best performance on classifying D-human tickets on TicketBERT but on D-machine and D-mixture Ticket-BERT tends to not perform

very well because these datasets contain machine generated texts which have different semantics in the meaning of the texts when compared to human generated texts. Naïve-Bayes models performed worst on the D-Human dataset than Logistic Regression models which concludes that discriminative models are better than generative models in labeling tasks which contains unstructured human text.

LLMs not only work well with prompt engineering techniques and in-context learning (Cahyawijaya et al., 2024) but are also known to be zero-shot learners where they do not require any examples to understand the meaning of the given task (Kojima et al., 2022).

In (Kojima et al., 2022), the authors use Chain-of-Thought (CoT) prompting along with zero-shot learning to evaluate the pre-trained LLMs on various real-world tasks like arithmetic and symbolic reasoning. They achieved state-of-the-art results just by adding "Let's think step by step" to the prompt. The Chain-of-Thought process is performed with two different prompts – reasoning extraction and answer extraction. In the reasoning extraction step, the reasoning of the language model for the prompt is obtained and the same reasoning text is forwarded to the model along with the prompt in the answer extraction step. They conducted experiments on six Arithmetic tasks, two common sense tasks, and four symbolic reasoning tasks. The authors experimented with four different techniques namely few-shot, few-shot with CoT, zero-shot, and zero-shot CoT on 17 different language models and obtained state-of-the-art results for the zero-shot CoT technique for 90% of the tasks. This paper tells us how large language models are "decent zero-shot reasoners" when given proper instructions to leverage their reasoning capabilities.

In the most recent techniques, Large Language Models (LLMs) like GPT-3.5 and GPT-4 have outperformed fine-tuned models and non-generative models with few examples. In (Loukas et al., n.d.), they show how generative models can leverage the accuracy of various tasks in few-shot setting or in-context learning. They conduct few-shot learning on generative models like GPT-3.5 and GPT-4 in the financial domain. To perform this task, they use Banking77 (Casanueva et al., 2020) dataset, a financial dataset for intent classification contains 77 labels where most of the labels overlap with each other which makes this research work close to our thesis objective and realistic. In the in-context learning process, the authors provide

the GPT models with a task description and the classes in the first section of the prompt, they provide example of labeled texts for each class in the second section and then they add the text which is to be classified. In this work, the authors conclude that using conservative GPT models with in-context learning can be easy and accurate when fast responses/ solutions are required compared to Masked Language Models (MLMs).

4.2 Process Mining & NLP

In Incident Management system, process mining plays an important role in organizing the bug reports, reporting the time taken to resolve and ensures that the process of the solution delivery to the customer is smooth. By analyzing the event log generated from the combination of activities, timestamps, and case IDs, process mining techniques offer invaluable insights into the underlying workflows and procedures within incident management processes. This facilitates the identification of bottlenecks, optimization opportunities, and areas for improvement in the bug resolution process.

Moreover, process mining aids in enhancing the overall efficiency and effectiveness of incident management operations by providing actionable insights derived from the analysis of event logs. By visualizing the sequence of activities and their interdependencies, process mining enables stakeholders to gain a comprehensive understanding of the incident management process, thereby fostering informed decision-making and continuous process improvement initiatives. Ultimately, the integration of process mining methodologies within Incident Management systems serves to streamline operations, enhance customer satisfaction, and drive organizational success in delivering timely and effective solutions to reported issues (Terragni & Hassani, 2018). With this groundwork established, further exploration of related papers promises to unveil additional insights and advancements in this domain.

The activities (labels) created using NLP techniques serve as a backbone for the generation of the event log. For every respective activity, a timestamp, and a case ID (ticket ID) is already allocated in the dataset given. Using these three attributes an event log generated a process mining model. The quality of this event log is a crucial step for the easy application of process mining techniques.

One of the recent state-of-the-art evaluations done by (Schüler & Alpers, 2024) on various automatic process model generation approaches talks about the process models generated using machine learning approaches. In this approach, process models are extracted using rule-based approaches, pattern matching techniques and using Natural Language Processing (NLP). The authors also talk about the issues with the huge amount of unstructured data produced at enterprises and companies. The quality of the process models generated depends entirely on the quality of the textual data and when using natural language approaches, the incident tickets often contain quite complex sentences. To improve the accuracy of the process models, the input textual data must be heavily pre-processed, and addition of additional information could be required often. This work also proposes different ways to improve the quality of generated models by using repair algorithms based on adding additional data or based on the model properties.

There is some quality research work done in the process of extracting process models from email logs (Jlailaty et al., 2016) and using multi-grained text classification (Qian et al., 2019). Email is a rich source of information about business processes, but it's often unstructured. This paper addresses the challenge of automatically extracting process models from email logs. Unsupervised machine learning techniques are used in (Jlailaty et al., 2016) to label email logs with activities for process model generation and this can be useful for activity generation for new incoming emails. This paper structures the process model extraction in three stages, namely Process Topic Discovery, Process Instances Discovery, and Process Activities Discovery.

Traditional Process Model Extraction (PME) relies on manually defined features to extract process relationships. In (Qian et al., 2019), the authors reconstruct the Process Model Extraction (PME) task as multi-grained text classification problem. They develop two different datasets from different domains and then train and evaluate their model on Process Model Extraction (PME). They break down PME into three tasks mainly – sentence classification (sentence-level), sentence semantics recognition (sentence-level) and semantic role labelling (word-level). The authors utilized a hierarchical neural network to automatically capture these aspects from text data at different granularities (word, sentence, document). For training

the neural network, they define a coarse-to-fine learning mechanism where the neural network is trained in stages, starting with coarse-grained tasks (identifying procedural text), and progressing to fine-grained tasks (classifying process steps and relationships). This allows the network to leverage high-level knowledge for more specific tasks. The proposed approach achieved better performance compared to existing state-of-the-art methods for PME.

The framework developed in (Kourani et al., 2024) automates the generation of process models from the textual descriptions using the contextual understanding capabilities of LLMs. They implement prompt engineering techniques, error handling and code generation to extract processes from textual descriptions and add an interactive feedback loop that can refine the generated process models. They use role prompting techniques where the LLM acts as a process model expert in the field of process mining and allow the LLM to fill in any gaps in the process description. They also use few-shot learning and injective learning techniques to incorporate knowledge that the model is unaware of. To make the LLM understand how to avoid errors that can occur during the process model generation, they use Negative Prompting technique where they use few-shot demonstrations including common mistakes that should be avoided.

To evaluate the performance of the LLM, the authors raise two research questions – how well does their framework perform when integrated with state-of-the-art LLMs and its performance comparison with other process modeling systems. The results convey that GPT-4 performed better than Gemini and GPT-4 error handling capabilities ensured quality and efficient process model generation. They also concluded that Gemini failed to incorporate user feedback in the refinement of process models.

5 Discussion

In the literature survey, we have discussed about research papers that have contributed to incident/support ticket management that help the IT service desk employees to better understand, organize and solve enormous number of IT issues generated in any enterprise. Few papers delve into extracting a process model from the inci-

dent management tickets but most of the work does not consider the life cycle of the tickets except (Liu et al., 2023).

The traditional classification algorithms may provide state-of-the-art results for this task, but they need large amounts of labeled data for training. Labeling incident tickets not only requires a huge amount of time but also requires vast knowledge regarding the incident tickets, products, and their in-domain vocabulary. Recent advancements in NLP as mentioned in [(Loukas et al., n.d.), (Amin et al., 2023)] eliminate the need for training and provide solutions using latest NLP techniques which are accurate and easier to implement.

In Section 4.1, we also discussed how the semantics of the text matter in the performance of the language models. In the context of incident tickets, semantics of the text undergo transformation compared to their literal interpretation. The meaning of the texts is changed with technical annotations specific to software industry. LLMs are known widely for capturing semantic meanings and contextual cues embedded within incident text. LLMs consider the surrounding context of words which allows extraction of the meaning based on a larger context. They can analyse the patterns and interpret the content within the incident tickets and help us classify the incident tickets efficiently.

As discussed in (Liu et al., 2023), Transformer models like BERT require no need for training and just fine-tuning the model for incident classification task using a smaller dataset yields great accuracy. Advancing beyond BERT, LLMs like GPT-3, GPT-3.5 and GPT-4 give promising results for any language task using prompt engineering or in-context learning. Building upon this, we want to eliminate the use of training language models and utilize their context understanding capabilities for performing the multi-class classification task. Techniques like prompt engineering and few-shot/ in-context learning settings can be used for the GPT models and fine-tuning approach for BERT to achieve the desired output.

6 Conclusion

In this report, we have addressed the research gap between recent NLP techniques and Process Mining to solve the classification of the incident management tickets. We have found that using recent NLP techniques not only excludes the need for

training language models but also eliminates the need for extensive manual labelling of the incident management data.

It is evident that incident ticket classification presents unique challenges due to the wide range of classes and overlapping nature of many classes. Previous research has explored various techniques, such as fine-tuning, clustering and training the language models, to improve classification accuracy. Studies have demonstrated the effectiveness of techniques like Multinomial Naïve Bayes Classifier, auto-suggest mechanisms, and Machine Learning mechanisms like fastText for ticket tagging. Moreover, the advent of advanced models like BERT has revolutionized classification tasks by capturing contextual information and semantic nuances, enhancing classification performance. However, it is crucial to note that while static word embeddings may not be effective for IT ticket classification, models with contextual embeddings like BERT show promising results.

As addressed in Section 1.2, it is evident that LLMs provide better results when leading with unstructured texts and texts which contain out-of-domain meanings. Eliminating the need for training and reducing the costs of manual labelling, LLMs make the accomplishment of tasks easier and computationally time effective. Particularly talking about incident management tickets, manually labelling them is a challenging task because the context of the text's changes accordingly with respect to the issue.

Furthermore, the integration of incident ticket classification with Process Mining methodologies offers a comprehensive approach to incident management. Process Mining techniques provide invaluable insights into the underlying workflows and procedures within incident management processes, facilitating the identification of bottlenecks and optimization opportunities. By visualizing the sequence of activities and their interdependencies, process mining enables enterprises to gain a comprehensive understanding of the incident management process, thereby fostering better resolution times which provide better customer satisfaction and less churn rate.

In conclusion, the synthesis of NLP techniques and Process Mining methodologies presents a compelling avenue for enhancing incident management processes. By leveraging advanced models like GPT-3.5 and innovative techniques in-context

learning, prompt engineering, organizations can streamline operations, enhance customer satisfaction, and drive organizational success. Baseline algorithms like KNN, Naïve Bayes also provided efficient results when correct representation techniques were chosen. As we look towards the planned approach for this thesis, continued research in this domain promises to unveil additional insights and advancements, furthering the efficiency and effectiveness of incident management processes.

7 Literature

Amin, M. M., Mao, R., Cambria, E., & Schuller, B. W. (2023). *A Wide Evaluation of ChatGPT on Affective Computing Tasks*. http://arxiv.org/abs/2308.13911

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. http://arxiv.org/abs/2005.14165

Cahyawijaya, S., Lovenia, H., & Fung, P. (2024). *LLMs Are Few-Shot In-Context Low-Resource Language Learners*. http://arxiv.org/abs/2403.16512

Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). *Efficient Intent Detection with Dual Sentence Encoders*. http://arxiv.org/abs/2003.04807

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (n.d.). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://github.com/tensorflow/tensor2tensor

Jlailaty, D., Grigori, D., & Belhajjame, K. (2016). *A framework for mining process models from emails logs*. http://arxiv.org/abs/1609.06127

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). *FastText.zip: Compressing text classification models*. http://arxiv.org/abs/1612.03651

Kallis, R., Di Sorbo, A., Canfora, G., & Panichella, S. (2019). Ticket Tagger: Machine Learning Driven Issue Classification. *Proceedings - 2019 IEEE International Conference on Software Maintenance and Evolution, ICSME 2019*, 406–409. https://doi.org/10.1109/ICSME.2019.00070

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large Language Models are Zero-Shot Reasoners*. http://arxiv.org/abs/2205.11916

Kourani, H., Berti, A., Schuster, D., & van der Aalst, W. M. P. (2024). *Process Modeling With Large Language Models*. http://arxiv.org/abs/2403.07541

Liu, Z., Benge, C., & Jiang, S. (2023). *Ticket-BERT: Labeling Incident Management Tickets with Language Models*. http://arxiv.org/abs/2307.00108

Loukas, L., Stogiannidis, I., Malakasiotis, P., Vassos, S., & Ai, H. (n.d.). *Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance*. https://chat.openai.com/

Maksai, A., Bogojeska, J., & Wiesmann, D. (2014). Hierarchical Incident Ticket Classification with Minimal Supervision. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2015-January*(January), 923–928. https://doi.org/10.1109/ICDM.2014.81

Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., & Elazar, Y. (2023). *Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation*. http://arxiv.org/abs/2305.16938

Qian, C., Wen, L., Kumar, A., Lin, L., Lin, L., Zong, Z., Li, S., & Wang, J. (2019). *An Approach for Process Model Extraction By Multi-Grained Text Classification*. http://arxiv.org/abs/1906.02127

Revina, A., Buza, K., & Meister, V. G. (2020). IT Ticket Classification: The Simpler, the Better. *IEEE Access*, *8*, 193380–193395. https://doi.org/10.1109/ACCESS.2020.3032840

Schüler, S., & Alpers, S. (2024). State of the Art: Automatic Generation of Business Process Models. *Lecture Notes in Business Information Processing*, *492 LNBIP*, 161–173. https://doi.org/10.1007/978-3-031-50974-2_13

Terragni, A., & Hassani, M. (2018). Analyzing Customer Journey with Process Mining: From Discovery to Recommendations. *Proceedings - 2018 IEEE 6th International Conference on Future Internet of Things and Cloud, FiCloud 2018*, 224–229. https://doi.org/10.1109/FiCloud.2018.00040

Wahba, Y., Madhavji, N. H., Steinbacher, J., & Steinbacher, J. 2020. (2020). *Evaluating the Effectiveness of Static Word Embeddings on the Classification of IT Support Tickets*.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. In *Quarterly* (Vol. 26, Issue 2).

Zaidi, S. S. A., Fraz, M. M., Shahzad, M., & Khan, S. (2021). A multiapproach generalized framework for automated solution suggestion of support tickets. *International Journal of Intelligent Systems*, *37*(6).

Zhou, Y., Tong, Y., Gu, R., & Gall, H. (2014). Combining text mining and data mining for bug report classification. *Proceedings - 30th International Conference on Software Maintenance and Evolution, ICSME 2014*, 311–320. https://doi.org/10.1109/ICSME.2014.53