Data Annotation for Support Ticket Data: A Literature Review

Simon Fuchs
Technical University Munich (TUM)
s.t.fuchs@tum.de

Janik Schnellbach
Technical University Munich (TUM)
janik.schnell-bach@tum.de

Lukas Schmidt
Technical University Munich (TUM)
luk.schmidt@tum.de

Dr. Holger Wittges
Technical University Munich (TUM)
holger.wittges@tum.de

Abstract

Supervised Machine Learning is still the most prevalent Machine Learning approach used across the field of Natural Language Processing. As it needs labels to work properly, labeling text data sets is a discerning step in supervised Machine Learning projects. Many industry projects involving supervised Machine Learning never reach a productive phase due to the absence of sufficient labeled data. Against this background, we conducted a Literature Review investigating state of the art approaches to label text data sets for later Natural Language Processing projects. We concentrated on solutions that could be applicable to annotate a support ticket data set. We found that there are three major approaches: Crowdsourcing, Learning Algorithms and Weak Supervision. We concluded that in annotation projects there seems to be a trade-off between label quality and cost/effort. We discuss our findings and share our thoughts on the special challenges of annotating a support ticket data set.

Keywords: Data annotation, Support Ticket Systems, Machine Learning, Labeling

1. Introduction

1.1 Motivation

Responding to questions and problems of customers as also providing technical support in set-up, operation, and maintenance and has a high significance for Information Technology Service companies. To provide this help, most of these organizations use a support ticket system (STS) in which clients can describe their requests and support staff members can send back possible solutions (Gupta et al., 2018). Distributing the incoming tickets to the experts is a time consuming process, which takes rarely available domain knowledge. Nevertheless, companies often outsource or delegate these low-level support tasks to third companies or less-skilled workforce, which increases the error rate within the service desk and therefore lowers customer satisfaction (Chagnon, 2017).

With the rising interest in Machine Learning (ML) in recent years, interest revived in how ML can be used to automate support service desks (Simon Fuchs et al., 2022). The hopes are that ML-automated service desks are more cost-effective and less error-prone than service desks run by the current less-skilled workforce (Qamili et al., 2018).

The key factor for a ML model's performance is the training data (Mitchell, 2007), i.e. number of observations with certain features (Goodfellow et al., 2016). Depending on the data set and the training type, the learning process is distinguished between supervised and unsupervised (Mitchell, 2007).

Whereas unsupervised learning identifies underlying structures to cluster the data (Goodfellow et al., 2016), supervised learning uses data labeling, i.e. to annotate a tag on the still untreated data (Monarch, 2021), to guide the learning process. The label can be a continuous variable (regression) nor a discrete category (classification) (Bishop & Nasrabadi, 2006). Data preprocessing, which includes the creation of labels, can take 80% (Shah & Kumar, 2019) or even 90% (Whang et al., 2023) of the work implementing a ML algorithm.

Additional, there are concepts that are called reinforcement learning in which a human agent works as active trainer and that only need small or no initial training data sets (Li, 2017). However, reinforcement learning will only touched lightly in this paper.

Certainly, when looking into the standard literature in the field of ML, such as Goodfellow et al. (2016), the authors often neglect the underlying data and limit themselves to mathematical foundations and the implementation of models. When looking at the universities, there are numerous lectures on ML, but hardly any courses focusing on data preparation (Monarch, 2021).

Depending on the source, between 78-87% of all ML projects started in industry fail or at least never reach a productive state (Gartner, 2018; Gates, 2019; venturebeat, 2019). According to Weber et al. (2022), data management is one of the four key domains in an enterprise, where ML projects tend to fail. Within data management training data collection and curating are key tasks, which largely includes data labeling. In fact, poor data management in general and missing or noisy labels in specific, are one of the 5 main reasons, why ML projects fail in praxis (Gates, 2019; Today, 2020; Trends, 2022).

URI: https://hdl.handle.net/10125/106573
978-0-9981331-7-1

1.2 Research Purpose & Research Question

These often neglected dimensions of data preprocessing in mind we analyze the process of data preparation, in detail the procedure of data annotation, in the automation of STSs using ML. Therefore, we undertake a narrative Literature Review (Paré et al., 2015) and follow the principle of Webster and Watson (2002) to identify what has been written about the status quo of annotating in support ticket data sets. We define the following research question: *What is the state-of-the-art in data annotation methods, which could be applied to a support ticket data set?*

This question aims to provide an overview of different labeling methods relevant for support ticket data sets and a collection of their advantages and disadvantages. We aim at providing a clearly arranged overview about state-of-the-art technology and approaches in text data annotation. Such that ML researchers can use this overview as starting point for their own text data annotation projects.

This literature review is also part of a larger STS automation project, in which we aim at implementing an ML classifier in the STS at our chair. The auxiliary goal of this literature review is therefore to create a knowledge base, how to label our own support ticket data set. Further intermediate results were already published in (Andraichuk, 2021; Simon Fuchs et al., 2022; Simon Fuchs et al., 2022; Wollendorfer, 2021).

2. Literature Review Design

The main purpose of this Literature Review is to examine the state-of-the-art labeling methods relevant for support ticket data sets, which itself is a subdivision of the ML research field of natural language processing (NLP). The first fast searches for this topic revealed that there is not much literature yet, specifically in the field of labeling support ticket data. For this reason, we extended our view on data annotation in NLP-fields with medium sized text sets. This means we excluded data annotation projects that (1) did not annotate text, (2) annotated large documents (multiple pages), or (3) annotated only text snippets of below one sentence.

The literature review presented was conducted between December 2022 and May 2023. We included the databases IEEE Xplore, Springer Link, Scopus and Web of Science. We used the keywords “label*ing”, “training data”, “text”, “annotation”, “machine learning” and “classification” and connected them using the Boolean Operators AND and OR for a first universal search on labeling and annotation methods (1-4). Based on that we search more specific on the methods Crowdsourcing (5-6), Learning Algorithms (7-8), and Weak Supervision (9-10) with the queries shown in Table 1. Unfortunately, we were not able to find one common search string for all databases due to the specifications of each database.

In the end, we received 139 hits, of these we consider 47 as relevant. Eliminating the 9 duplicates and adding 4 papers found by Forward and Backward Search (Webster & Watson, 2002) results in 42 relevant pieces. For a paper to be regarded as relevant, it must deal with the topic of annotating medium texts (as explicated above) with the purpose of afterwards training an ML model with these data and meet the three criteria stated above.

In recent years, there were also several approaches using large language models (LLMs) for data annotation, for example Zhou et al. (2022). Unfortunately, these experiments seemed not to fit the requirements of our own use case such that we neglected LLMs in our literature search.

Table 1 Overview of search keywords and search results

#	Database	Search String	Hits	Relevant
1	IEEE Xplore	("Document Title":"labeling") AND ("All Metadata":"training data") AND ("All Metadata": “text”)	19	3
2	Springer Link	(title:"label*ing") AND "text data" AND ("machine learning" OR "classification")	13	5
3	Scopus	TITLE(data AND annotation) AND TITLE-ABSKEY(label AND text)	14	3
4	Web of Science	(TI=(training data)) AND (ALL=(Machine Learning) OR ALL=(Classification)) AND ALL=(Text data) AND ALL=("Label*ing")	11	7
5	IEEE Xplore	(("Document Title":crowdsourc*) AND (("All Metadata":machine learning) OR ("All Metadata":classification)) AND "All Metadata":text data)	18	1
6	Web of Science	TI=(crowdsourc*) AND (ALL=(Machine Learning) OR ALL=(Classification)) AND ALL=(Text data) AND ALL=(Label*)	19	5
7	Scopus	(TITLE(active AND learning AND "label*ing") AND TITLE-ABS-KEY(machine AND learning) OR TITLE-ABS-KEY(classification) AND TITLEABS-KEY(text AND data))	8	4
8	Web of Science	TI=(learning "label*ing") AND (AB=(Machine Learning) OR AB=(Supervised Learning)) AND ALL=(Text data) AND ALL=(Label*)	15	7
9	Springer Link	(title:"weak supervision*") AND "labeling" AND "text data" AND ("machine learning" OR "classification")	7	4
10	Web of Science	TI=(weak supervision*) AND (AB=(Machine Learning) OR AB=(Supervised Learning)) AND ALL=(Text data) AND ALL=(Labeling	15	8
		Sum	47	
		Duplicates	9	
		Forward / Backward Search	4	
		Total	42	

Table 2 Concept matrix

	Method			Purpose		Details			
	Crowdsourcing	Learning Algorithm	Weak supervision	Labeling	subsequent Machine Learning project	Own Data Set	Only text-based	Transformers used	Application
Frequency of occurrence	**38%**	**64%**	**26%**	**81%**	**19%**	**50%**	**69%**	**21%**	
Alabduljabbar and Al-Dossari (2017)	X			X					Informatics
Altınel et al. (2017)		X			X		X		Informatics
Chang et al. (2015)	X	X		X		X	X		Informatics
Chang et al. (2017)	X			X					Informatics
Chen et al. (2022)		X	X	X		X	X		Informatics
Costa et al. (2011)	X	X		X			X		Informatics
Cusick et al. (2021)		X	X		X	X	X		Medicine
Dumitrache et al. (2017)	X			X			X		Medicine
Enkhsaikhan et al. (2021)		X		X		X	X		Geology
Eyuboglu et al. (2021)		X	X	X		X		X	Medicine
Haralabopoulos et al. (2021)	X			X		X	X		Informatics
Hassanzadeh and Keyvanpour (2013)		X			X				Informatics
Humbert-Droz et al. (2022)	X			X		X	X		Medicine
Jacobs et al. (2022)		X		X			X	X	Informatics
Jin et al. (2020)	X			X					Informatics
Kee et al. (2018)		X		X					Informatics
Kim et al. (2021)		X			X	X	X		Informatics
Lison et al. (2021)			X	X			X	X	Informatics
Miller et al. (2020)		X		X			X		Politics
Nimo-Járquez et al. (2019)		X			X	X	X		Informatics
Passonneau et al. (2008)	X			X		X			Art
Piroonsup and Sinthupinyo (2018)		X			X				Informatics
Poursabzi-Sangdeh et al. (2016)		X		X			X		Informatics
Ratner et al. (2017)			X	X					Informatics
Raykar et al. (2010)	X			X					Informatics
Rokicki et al. (2015)	X			X		X			Informatics
Rothwell et al. (2015)	X			X		X	X		Informatics
Salma et al. (2021)		X		X			X	X	Informatics
Sharma et al. (2019)	X			X		X	X		Psychology
Shen et al. (2022)		X	X		X	X	X	X	Medicine
Singh et al. (2021)		X		X			X	X	Medicine
Takanobu et al. (2018)		X	X	X		X	X		Informatics
Tchoua et al. (2019)	X	X		X		X	X	X	Engineering
Tekumalla and Banda (2021)			X	X		X	X	X	Medicine
Trivedi et al. (2019)		X		X		X	X		Medicine
Varma and Ré (2018)		X		X					Informatics
Wang et al. (2017)		X		X		X			Informatics
Wang et al. (2019)		X	X		X	X	X		Medicine
Yang et al. (2020)	X	X	X	X			X	X	Informatics
Yitagesu et al. (2021)		X		X		X	X		Informatics
Zhu et al. (2010)		X		X			X		Informatics
Zhu et al. (2009)	X		X	X			X		Informatics

Within our research, we focused on methods, processes, sequences and flows of data annotation projects, while omitting concrete tools. We also concentrated on data annotation that is done by workers of a company or project with business knowledge. This is due to the fact that we wrote this research in preparation for our own ticket data annotation project that requires skilled annotators because of the complex workings within our service desk.

3. Findings of the Literature Review

3.1 Topic Analysis

Our findings are displayed in the concept matrix of Table 2. Most of our results come from the field of informatics, several from medicine and also some from politics, literature, art and geology.

We identified three major labeling methodologies: Crowdsourcing, Learning Algorithms and Weak Supervision. As all of them are built for different requirements, which vary in effort and label quality. We explicate our findings to the three methods below.

The main purposes discussed in the majority of the papers are labeling or annotating a data set, at which we expected to find many data annotation projects for a direct/subsequent ML project. However, 34 of 42 papers only described data labeling with no directly subsequent ML project, while 8 papers do.

We also recognized a trend of transformer models being increasingly used for data annotation. In 8 of the 15 paper published in the years 2020-2022 transformer models were used in some way.

3.2 Crowdsourcing

The first approach for labeling data is Crowdsourcing, where humans label the data points Especially for generating high-quality labels humans are used (Monarch, 2021). In particular Crowdsourcing means distributing tasks to a large, sometimes external, group, which makes solving extensive tasks more efficient than done by one specific person (Howe, 2008). With the increasing use of ML and therefore the increasing need for labeled data, Crowdsourcing gained a lot of momentum (Chang et al., 2017).

There are several approaches using platforms like Amazon Mechanical Turk for Crowdsourcing. In which poorly-skilled and poorly-payed workers annotate data following very basic flowcharts or manuals (Huynh et al., 2021). Unfortunately, bad accuracy rates for this approach are reported (Marshall et al., 2023). For this reason and because we aim at doing our own data annotation project (see above), we omitted crowdsourcing platforms form our further research focus.

To set up a successful Crowdsourcing project organization and structure are needed. Alabduljabbar and Al-Dossari (2017) point out the importance of passing explanations and guidelines to the workers for the labeling task, which is further developed by Chang et al. (2017). Also the device and interface for labeling is crucial, especially the design and interaction needs to be efficient and user-friendly. Monarch (2021) recommended to take into account human-computer interaction, minimizing eye movement and input validation.

Label Design: As described above the general concept of Crowdsourcing is multiple workers assign one or more labels to the given data samples. But conflicts arise, when distinct labels are assigned by different crowd workers. In this case Raykar et al. (2010) suggest a majority voting, but also note that this concept assumes equal proficiency of the workers. Chang et al. (2015) develop the approach further by counting in their before measured reliability. Therefore some papers use the Cohen's kappa coefficient (Cohen, 1960; Sharma et al., 2019; Wan et al., 2019) or Krippendorff (2018)'s alpha to identify the agreement between two annotators. The latter paper shows that these inter-annotators cannot be generalized but have to be determined empirically as they depend on the labeling use case, in particular the number of assignable labels, the number of crowd workers and the length of the data sample. Chang et al. (2015) use a simple binary label setup, however Passonneau et al. (2008) observe that the inter-annotator agreement rises when more than one label can be assigned. Dumitrache et al. (2017) point out that the disagreement of the workers can signal the ambiguity of a text sample, which is natural in language and must be addressed in Crowdsourcing and label design.

Labor distribution: Until now we considered a group of workers labeling parallel, partially overlapping and partially weighted. Another approach, analyzed by Rokicki et al. (2015), would be to divide the annotator pool in independent groups competing with each other. According to the authors this procedure increases the productivity through a higher team spirit and motivation and can reduce the costs, if only the winning team is paid. A similar team based procedure is proposed by Yang et al. (2020), where two groups are created. One tries to generate rules for labeling the data, whereas the second group tries to find observations to disprove those rules.

Quality Assurance: Next to efficiency also the quality of the labels has to be ensured. Jin et al. (2020) outline two dimensions. First, design aspects like workflows, gamification or the payment of crowd workers. Second, the statistical foundation like the aggregation of labels. Rothwell et al. (2015) find a susceptibility of Crowdsourcing, namely automatic/blind labeling. Therefore they suggest a captcha mechanism, in analogy to reCAPTCHA from Von Ahn et al. (2008), i.e. installing a quality check by periodically showing test samples with known solution to control the workers diligence.

Criticism: Still, criticism is raised. Hutchinson et al. (2021) point out the lack of responsibility for the data after the Crowdsourcing process is finished, as wrong labels can have negative influences also on future models. Therefore, Geiger et al. (2020) even say Crowdsourcing could be "a different kind of black boxing". Consequently, Hutchinson et al. recommend clear responsibilities and detailed documentation. Furthermore, Haralabopoulos et al. (2021) indicate the lack of data privacy when external crowd workers are hired and Sambasivan et al. (2021) show the often insufficient results as due to cost reasons less skilled workers or people without the required domain knowledge are engaged.

3.3 Learning Algorithms

As second approach for label acquisition, we identify a group of methods, which we summarize

under the term "learning algorithms". The characteristic of these techniques is that these base on ML or Deep Learning (DL) with learnable parameters. One can distinguish between semi-automated, still requiring human interaction, and fully automated systems.

3.3.1. Semi-automated

Semi-automated algorithms combine ML and human interaction, which is mostly known as Active Learning (AL). In this case only a minimal part of the data set is labeled by a human, whereby AL structures the process and tries to cover the complete characteristic of the data set through prioritization (Miller et al., 2020). The data set can then be used for semi-supervised learning. In comparison to manual labeling Hassanzadeh and Keyvanpour (2013) show that AL can reduce the costs by around 90%. Furthermore, Miller et al. (2020) demonstrate the advantages of AL in the case of imbalanced data. In contrast to random sampling, AL also considers the less frequent categories.

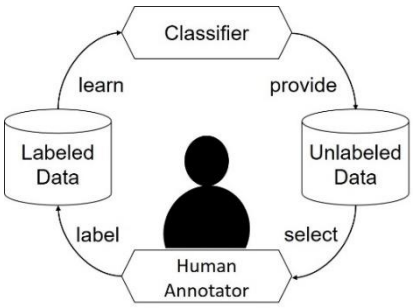

Figure 1 Active Learning workflow. Figure adapted from Settles (2009).

Figure 1 shows the pool-based labeling process, which provides the annotator with a "data pool" to choose from, whereas the stream-based process would deliver the next sample to label in a continuous chain (Zhu et al., 2010). Which sample is considered next to be annotated by the human is decided by the sampling or querying strategy. We will analyze different strategies from Settles (2009) and Miller et al. (2020).

Uncertainty-based (UB) is the most common used strategy, it needs as a starting aid randomly drawn samples, which will be labeled (Settles, 2009). A classifier predicts the unlabeled data samples and the ones with the highest uncertainty are selected again to be labeled by a human (Lewis, 1995). Then this process is iteratively repeated such that the classifier improves (Monarch, 2021). Jacobs et al. (2022) showed how this procedure can reduce the effort in labeling a text data set. A variation of this approach is the margin sampling (Tong & Koller, 2001) considering the margin between the two most probable labels. A low margin signifies a high uncertainty. Costa et al. (2011) therefore use a Support Vector Machine (SVM).

Query-by-committee (QBC) uses multiple classifiers in parallel voting independently for a label. The sample with the highest disunity between the classifiers is chosen to be classified by a human next (Gilad-Bachrach et al., 2005). Kee et al. (2018) show how this method surpasses the UB algorithm by taking into account the variety in the data with different specialized classifiers. QBC is combinable with a data distribution analysis.

Data-distribution-based (DDB) methods are looking at the data's variance or density instead of using a classifier (Settles, 2009). Wang et al. (2017) show a high generalizability using density clustering.

Model-performance-based (MPB) approaches are based on the two methods "expected model change" and "expected error reduction" by Settles (2009) and aim at selecting the sample whose label would boost the performance or reduce the error of the total model the most (Miller et al., 2020). However, the authors point out the rising computational costs for the calculations with increasing data size.

The question remains when to stop the labeling and to start with e.g. semi-supervised learning. Zhu et al. (2010) propose the criteria maximum uncertainty, overall uncertainty, selected accuracy, and minimum expected error. Piroonsup and Sinthupinyo (2018) suggest to reach a balanced distribution of labeled and unlabeled data in a data clustering before the annotation process can be terminated.

3.3.2 Fully-automated

Next to the presented semi-automated methods of learning algorithms the data samples can also be labeled fully automated, without any human interaction. Hereby, the borders to unsupervised learning are not clear-cut. In our research we found several approaches which we present hereafter. Enkhsaikhan et al. (2021) suggest a DL algorithm performing iterative Named Entity Recognition (NER), i.e. identifying and classifying words and their semantic meaning. By using an optimized domain dictionary, the data can then be automatically labeled or complex samples can be sorted out to be annotated by a human. Similarly Altınel et al. (2017) predict the labels by computing significance values for words and Kim et al. (2021) therefore use the frequency of entities in a text.

Yitagesu et al. (2021) find syntactic similarities between text samples with sentence parsing trees using a categorical variational autoencoder. Here, the input is embedded using different language models. Their results are comparable to manually labeled data sets, even for widely varying sentence structures. Singh et al. (2020) require a small pre-labeled sub-dataset that is used to train transformer models that afterwards then generate artificial labels for the rest data set. They also present some data augmentation strategies for improving their transformers' performance.

Salma et al. (2021) use a DL algorithm based on HuggingFace's XLNet[1]. After intense preprocessing, the text data is embedded into a vector representation and the cosine similarity scores between these vectors are calculated. After that, they use the infomap algorithm to extract specific communities in the text data. From these communities the authors derive rules to annotate the data points with distinct class and label assignments.

All fully-automated approaches analyzed during this literature review use some kind of ML approaches to generate data labels. Almost none of the reviewed papers spent thoughts on the question, if ML-labeled data might impart systematic errors to later ML models trained with it (for more, see section 5. Discussion).

3.4 Weak Supervision

The third group of labeling methods we found is Weak Supervision (WS). The WS approach aims to offer an automated alternative to costly manual annotation. Instead of accurate but expensive labels, WS generates cheap but noisy labels (Ratner et al., 2017). Noisy, in this case, is equivalent to a lower quality of labels, as described in subsection 2.2.1. Therefore, it is often used in use cases with large unlabeled data sets. Instead of manual data annotation by humans, WS uses so called Labeling Functions (LFs) that programmatically (and therefore rule-based) decide data labels In general, we found the following types of LFs:

Pattern-based LFs often base upon keyword matching/search or feature annotation. Here, the LFs focus on easy/obvious patterns (Lison et al., 2021). Often, meta parameters like document tags, ticket categories or user stamps are used for the LFs. Lison et al. (2021) call this LFs "gazetteers".

Distant Supervision LFs use external knowledge of corresponding labeled data sets to derive the labels. Here, the labeled source data set is used to make inferences about labels for an unlabeled data set (Lison et al., 2021). This can also be part of semi-supervises ML approaches (Revina et al., 2020).

Weak Classifier LFs make use of ML models, which already successfully performed on a similar/ related data set/task (Ratner et al., 2017). The idea is often to only use a subset of the text data with "obvious" data points to get a starting set of LFs that are afterwards fine-tuned by different methods (Ratner et al., 2017).

Many WS approaches are discussed in the field of the medical ML. For example, Cusick et al. (2021) use pattern-based LFs to structure unstructured clinical records. In the paper, they refer to two specific particularly common text patterns in their clinical records: target lexicons and modifier lexicons. Here, the term target lexicons refers to condition-specific keywords in the text, e.g. disease or disorder. The term modifier lexicons refers to drastic semantic changes within the text, e.g. negations. Another paper identified in this literature review and also in a medical context comes from Eyuboglu et al. (2021). They use WS to annotate unstructured radiology reports. The authors use LFs, that work based on the anatomical regions of the reports that are available as keyword, e.g. lung, liver, heart, etc. The paper describes a subsequent classification model with (pre)probabilities that the authors use to analyze the reports in more detail and to suggest treatment recommendations. Shen et al. (2022) use pattern-based LFs to annotate electronic health records with view on the patients lifestyle (e.g. sports, nutrition, etc.). The authors emphasize their finding that LFs, which primarily rely on keywords, are susceptible to biased data and lacks of data diversity.

Chen et al. (2022) propose a combination of a distant supervision approach with clustering. The authors assume in their combined approach that WS-annotated data always follows a similarity structure and therefore can be labeled by an iterative distant supervision KNN-clustering. Varma and Ré (2018) propose a further instance of Weak Classifier LFs. Here, they first embed the text data, using bag of words. After that, they iteratively derive heuristics for LFs using different classifiers on the embedded vectors. The authors also suggest an initial quality measurement of the data set to prevent the LFs from becoming too noisy or granular.

Regarding the label quality produced by WS approaches, we found that the calculation of quality/accuracy is not as straightforward as for Crowdsourcing or Learning algorithms. Unlike in Crowdsourcing, where the data is hand-labeled, the accuracy of LFs is only measurable to a limited extent (Ratner et al., 2017). Lison et al. (2021) therefore recommend assessing the impact and interaction of LFs by training an auxiliary generative model. Since LFs are difficult to apply from scratch, over time multiple libraries were developed to help researchers to create WS algorithms (Ratner et al., 2017). One famous such is Ratner et al. (2017)'s "Snorkel" , which was a pioneer work for large-scale use of WS. Snorkel provides an interface and uses a simplified syntax that allow easy writing and easy use of LFs. Another such toolkit is Lison et al. (2021)'s "skweak". Additional to its simplification of writing and managing LFs, it is also able to aggregate the results of the WS and to contrast the interaction of the created LFs.

3.5 Combining Weak Supervision and Crowdsourcing

Some papers we found combine WS and Crowdsourcing. Zhu et al. (2009) in a first step requested experts to annotate a subset of the data set. After that, they asked the experts to determine which attributes or parts of the document/text, e.g., sentences, keywords, tags, etc. led to their annotation decision. The authors then used these answers for developing handcrafted LFs, with which they labeled

1 https://huggingface.co/docs/transformers/modeldoc/xlnet

the rest of the data set. Wang et al. (2019) combine expert knowledge and historical knowledge bases from their clinical use case, to create LFs following a distant supervision approach.

4. Discussion

4.1 Machine Learning for Data annotation

The primary goal of this Literature Review was to provide an overview over the field of labeling methods relevant for support ticket data sets. Apart from the label annotation methods using human workforce (Crowdsourcing), we found several different methods using semi-supervised or even fully automated ML to generate labels. However, using un-/semi-supervised ML to guide a later ML algorithm raises the question if not monitored systematic errors caused by the ML model annotating the training data set are then later bequeathed to the supervised ML models trained on this automatically annotated data set. Unfortunately, this topic is only minorly discussed in the literature analyzed in this review. We therefore regard this issue as an open research gap. We argue that there should be more research investigating if automated data annotation approaches are reliable enough to afterwards train other ML models based on their automatically annotated data sets. Also we want to raise the question, why ML staff do not directly use un- or semi-supervised ML approaches for their projects instead of using it for data annotation and afterwards using a supervised approach. Based on the research studied in this review, we argue that every data annotation project should still involve a human component, to at least randomly check the automatically generated labels and by that bringing in business/use case understanding.

4.2 Effort and Label Quality

Optimal training with ML, needs optimal label quality. From a business perspective, we want to minimize the associated effort. However, the papers we found show there are just two possibilities, namely good or low-cost labels. In general, the best results in regard of label quality are achieved by skilled human annotators (Monarch, 2021). In regard of human annotators in general, we must differentiate. Skilled human annotators tend to produce high label quality (Mandal et al., 2018), as for example the health sector shows (Wang et al., 2020). In contrast, less skilled workers, often provided by microtasking services like Amazon mechanical Turk, do often raise concern respective their annotation quality (Hossain & Kauranen, 2015). When we compare the accuracy or confidence values provided in the analyzed papers, we find that skilled human annotators still outperform the ML approaches. Yet, the highly automated processes form a rapid and very cost-effective alternative to human annotators.

Annotation project leaders therefore find themselves in the dilemma of deciding between cost/effort and label quality. As label quality is directly connected to ML classifier performance (Aggarwal, 2014), this dilemma is equivalent to the question of how much the later performance of an ML artifact is allowed to cost. The problem of too elaborate annotation projects or better said the unclearness of the cost necessary to create a productive ML artifact is still one of the major failing points across the industry (Weber et al., 2022).

More direct quantitative comparisons between human annotators and automated approaches therefore would be highly helpful, interesting and relevant for ML managers having to decide which way to go to annotate training data sets. Unfortunately, most automated approaches studied in this review did not compare their approaches directly and quantitatively to a crowdsourcing project leaving the question open, if their approaches are a fair trade between label quality and effort. This we regard as a research gap.

4.3 Annotating support ticket data sets

Annotating a support ticket data set holds its own special challenges. Other than annotation projects that involve more general knowledge, like labeling every day images or labeling chat messages, the most common support ticket classification tasks are level-classification, sentiment prediction, or request escalation prediction (Simon Fuchs et al., 2022). All these classification tasks involve high degrees of business understanding (Simon Fuchs et al., 2022; Simon Fuchs et al., 2022) which probably amplifies the gap in label quality between skilled and unskilled annotators. In addition, we suspect that the responsibilities between different service desk levels and divisions are often less obvious than the differences in more common ML classification tasks. Additionally, support ticket data sets are often smaller than the large data sets of more mainstream ML data sets (Thiée, 2021; Yamaoka et al., 2019). These factors all increase the impact of label quality on a support ticket ML classification project. Based on this study, we suspect that, especially for smaller companies, best label quality will be achieved by manually annotating the support ticket data set by their support agents. However, the combination of automated approaches like semi-automated Learning Functions and Weak Supervision with human annotation and its business understanding looks like a promising candidate for a good trade-off of label quality and labeling effort/cost. Her, more research is needed. As we already wrote in section 2, unfortunately there is not much research published yet how to best label a support ticket data set. We think this is a promising topic and call for more research in this particular field.

5. Limitations of this Literature Review

We intentionally narrowed the scope of this Literature Review to the topic of ticket data annotation. As explicated above, this excluded literature in broader fields of data annotation, even on the cost of a smaller generalizability of our results aside the field of support tickets. This intentional reduction of the paper's scope is caused by our desire to have (1) literature research for our subsequent research on labeling a support ticket data set and (2)

Nevertheless, the practical application of such theoretical known solutions to the field of STSs are full research projects for themselves. We deliberately only searched scientific databases. This means we deliberately excluded google, patent literature or other non-scientific publications. This delimitation is motivated by the fact that industry solutions are often confidential or not detailed enough to understand their inner workings. In addition, we created this literature review to develop and test our own approaches with full control over it and without being stuck to industry tools.

6. Conclusion

Supervised Machine Learning (ML) heavily depends on the labels of a data set used for training. Labeling, also called annotating, a training data set is therefore a vital part of many ML projects. This gets so far that missing labeled training data is one of the major reason ML projects across the industries are cancelled. In particular, labeling a support ticket data set has its own specialties and challenges.

We therefore conducted this literature to identify state-of-the-art approaches to label a support ticket data set. A fast finding was that there is not much literature specific on support tickets, which is why we extended our search on annotating medium-sized text data in general.

We found that there are three major lines of annotating text data sets: By Crowdsourcing, where groups of human annotators label the data points; by Learning Algorithms, where mostly Deep Learning models with learnable parameters annotate the data points; and by Weak Supervision where unexact Labeling Functions are created to create cheap but noisy labels.

We discuss our findings with special view on the question if ML models labeling data for later ML training relay systematic errors; the question of the balance of label quality vs. project cost/effort; and the applicability of the found approaches on support ticket data. We highlight the specifications of the service desk use case and the relevance of business understanding in it. We also propose ideas how to effectively annotate a support ticket data set and identify research gaps in the present research.

7. References

Aggarwal, C. C. (2014). *Data Classification: Algorithms and Applications*. Taylor & Francis. https://books.google.de/books?id=gJhBBAAAQBAJ

Alabduljabbar, R., & Al-Dossari, H. (2017). *Towards a classification model for tasks in crowdsourcing* Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing, Cambridge, United Kingdom. https://doi.org/10.1145/3018896.3018916

Altınel, B., Ganiz, M. C., & Diri, B. (2017). Instance labeling in semi-supervised learning with meaning values of words. *Engineering Applications of Artificial Intelligence*, *62*, 152-163.

Andraichuk, V. (2021). *Extraction Tool for Customer Support Tickets from the SAP HANA Database - Design and Implementation* Technische Universität München]. Lehrstuhl Publikationen. https://www.researchgate.net/publication/364038597_Extraction_Tool_for_Customer_Support_Tickets_from_the_SAP_HANA_Database_-Design_and_Implementation_Extraction_Tool_for_Customer_Support_Tickets_from_the_SAP_HANA_Database_-Design_and_Implementation

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Chagnon, C. T., A.C.; Djamasbi, S. (2017). *Creating a decision support system for service classification and assignment through optimization* AMCIS 2017 - America's Conference on Information Systems: A Tradition of Innovation,

Chang, J. C., Amershi, S., & Kamar, E. (2017). Revolt: Collaborative crowdsourcing for labeling machine learning datasets. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,

Chang, S., Dai, P., Chen, J., & Chi, E. H. (2015). Got many labels? deriving topic labels from multiple sources for social media posts using crowdsourcing and ensemble learning. Proceedings of the 24th International Conference on World Wide Web,

Chen, L.-M., Xiu, B.-X., & Ding, Z.-Y. (2022). Multiple weak supervision for short text classification. *Applied Intelligence*, *52*. https://doi.org/10.1007/s10489-021-02958-3

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Costa, J., Silva, C., Antunes, M., & Ribeiro, B. (2011). On using crowdsourcing and active learning to improve classification performance. 2011 11th International Conference on Intelligent Systems Design and Applications,

Cusick, M., Adekkanattu, P., Campion, T. R., Jr., Sholle, E. T., Myers, A., Banerjee, S., Alexopoulos, G., Wang, Y., & Pathak, J. (2021). Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *J Psychiatr Res*, *136*, 95-102. https://doi.org/10.1016/j.jpsychires.2021.01.052

Dumitrache, A., Aroyo, L., & Welty, C. (2017). Crowdsourcing ground truth for medical relation extraction. *arXiv preprint arXiv:1701.02185*.

Enkhsaikhan, M., Liu, W., Holden, E.-J., & Duuring, P. (2021). Auto-labelling entities in low-resource text: a geological case study. *Knowledge and Information Systems*, *63*, 695-715.

Eyuboglu, S., Angus, G., Patel, B. N., Pareek, A., Davidzon, G., Long, J., Dunnmon, J., & Lungren, M. P. (2021). Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT. *Nature Communications*, *12*(1), 1880. https://doi.org/10.1038/s41467-021-22018-1

Fuchs, S., Drieschner, C., & Wittges, H. (2022). *Improving Support Ticket Systems Using Machine Learning: A Literature Review* Hawaii International Conference on System Sciences 2022,

Fuchs, S., Wiehl, N., Wittges, H., & Krcmar, H. (2022). *Towards an Automated SAP Service Desk -Design and Implementation of a Prototype Machine*

Learning Classifier for Support Tickets on a Small Data Set Conference: SAP Academic Community Conference 2022 DACH, Magdeburg, Germany.

Gartner. (2018). *Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence* Gatner. Retrieved 21.03. from https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence

Gates, N. (2019). *Almost 80% of AI and ML Projects Have Stalled, Survey Says* [Interview]. Robotics Business Review. https://www.roboticsbusinessreview.com/ai/almost-80-of-ai-and-ml-projects-have-stalled-survey-says/

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,

Gilad-Bachrach, R., Navot, A., & Tishby, N. (2005). Query by committee made real. *Advances in neural information processing systems*, *18*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gupta, M., Asadullah, A., Padmanabhuni, S., & Serebrenik, A. (2018). Reducing user input requests to improve IT support ticket resolution process. *Empirical Software Engineering*, *23*(3), 1664-1703. https://doi.org/10.1007/s10664-017-9532-2

Haralabopoulos, G., Torres, M. T., Anagnostopoulos, I., & McAuley, D. (2021). Privacy-preserving text labelling through crowdsourcing. Artificial Intelligence Applications and Innovations. AIAI 2021 IFIP WG 12.5 International Workshops: 5G-PINE 2021, AI-BIO 2021, DAAI 2021, DARE 2021, EEAI 2021, and MHDW 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings,

Hassanzadeh, H., & Keyvanpour, M. (2013). A two-phase hybrid of semi-supervised and active learning approach for sequence labeling. *Intelligent Data Analysis*, *17*(2), 251-270.

Hossain, M., & Kauranen, I. (2015). Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal*, *8*(1), 2-22.

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

Humbert-Droz, M., Mukherjee, P., & Gevaert, O. (2022). Strategies to Address the Lack of Labeled Data for Supervised Machine Learning Training With Electronic Health Records: Case Study for the Extraction of Symptoms From Clinical Notes. *JMIR Medical Informatics*, *10*(3), e32903.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,

Huynh, J., Bigham, J., & Eskenazi, M. (2021). A survey of nlp-related crowdsourcing hits: what works and what does not. *arXiv preprint arXiv:2111.05241*.

Jacobs, P. F., Maillette de Buy Wenniger, G., Wiering, M., & Schomaker, L. (2022). Active learning for reducing labeling effort in text classification tasks. Artificial Intelligence and Machine Learning: 33rd Benelux Conference on Artificial Intelligence, BNAIC/Benelearn 2021, Esch-sur-Alzette, Luxembourg, November 10–12, 2021, Revised Selected Papers,

Jin, Y., Carman, M., Zhu, Y., & Xiang, Y. (2020). A technical survey on statistical modelling and design methods for crowdsourcing quality control. *Artificial Intelligence*, *287*, 103351.

Kee, S., Del Castillo, E., & Runger, G. (2018). Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences*, *454*, 401-418.

Kim, J.-J., On, B.-W., & Lee, I. (2021). High-quality train data generation for deep learning-based web page classification models. *IEEE Access*, *9*, 85240-85254.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Lewis, D. D. (1995). A sequential algorithm for training text classifiers: Corrigendum and additional data. Acm Sigir Forum,

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

Lison, P., Barnes, J., & Hubin, A. (2021). *skweak: Weak Supervision Made Easy for NLP*.

Mandal, A., Malhotra, N., Agarwal, S., Ray, A., & Sridhara, G. (2018). Automated dispatch of helpdesk email tickets: Pushing the limits with AI.

Marshall, C. C., Goguladinne, P. S., Maheshwari, M., Sathe, A., & Shipman, F. M. (2023). Who Broke Amazon Mechanical Turk? An Analysis of Crowdsourcing Data Quality over Time. Proceedings of the 15th ACM Web Science Conference 2023,

Miller, B., Linder, F., & Mebane, W. R. (2020). Active learning approaches for labeling text: review and assessment of the performance of active learning approaches. *Political Analysis*, *28*(4), 532-551.

Mitchell, T. M. (2007). *Machine learning* (Vol. 1). McGraw-hill New York.

Monarch, R. M. (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Nimo-Járquez, D., Narvaez-Rios, M., Rivas, M., Yáñez, A., Bárcena-González, G., Guerrero-Lebrero, M. P., Guerrero, E., & Galindo, P. L. (2019). AL 4 LA: Active Learning for Text Labeling Based on Paragraph Vectors. Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part I 15,

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183-199.

Passonneau, R. J., Yano, T., Lippincott, T., & Klavans, J. (2008). Relation between agreement measures on human labeling and machine learning performance: Results from an art history image indexing domain. *Computational Linguistics for Metadata Building*, *49*.

Piroonsup, N., & Sinthupinyo, S. (2018). Analysis of training data using clustering to improve semi-supervised self-training. *Knowledge-Based Systems*, *143*, 65-80.

Poursabzi-Sangdeh, F., Boyd-Graber, J., Findlater, L., & Seppi, K. (2016). Alto: Active learning with topic overviews for speeding label induction and document labeling. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),

Qamili, R., Shabani, S., & Schneider, J. (2018). *An Intelligent Framework for Issue Ticketing System Based on Machine Learning*. https://doi.org/10.1109/EDOCW.2018.00022

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. (2150-8097 (Print)).

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, *11*(4).

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

Rokicki, M., Zerr, S., & Siersdorfer, S. (2015). Groupsourcing: Team competition designs for crowdsourcing. Proceedings of the 24th international conference on world wide web,

Rothwell, S., Carter, S., Elshenawy, A., Dovgalecs, V., Saleem, S., Braga, D., & Kennewick, B. (2015). Data

collection and annotation for state-of-the-art NER using unmanaged crowds. Sixteenth Annual Conference of the International Speech Communication Association,

Salma, T., Saptawati, g. a. p., & Rusmawati, Y. (2021). *Text Classification Using XLNet with Infomap Automatic Labeling Process*. https://doi.org/10.1109/ICAICTA53211.2021.9640255

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,

Settles, B. (2009). Active learning literature survey.

Shah, V., & Kumar, A. (2019). The ML data prep zoo: Towards semi-automatic data preparation for ML. Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning,

Sharma, V., Shpringer, B., Yang, S. M., Bolger, M., Adewole, S., Brown, D., & Gharavi, E. (2019). Data collection methods for building a free response training simulation. 2019 Systems and Information Engineering Design Symposium (SIEDS),

Shen, Z., Schutte, D., Yi, Y., Bompelli, A., Yu, F., Wang, Y., & Zhang, R. (2022). Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. *BMC Medical Informatics and Decision Making*, *22*(1), 88. https://doi.org/10.1186/s12911-022-01819-4

Singh, G., Sabet, Z., Shawe-Taylor, J., & Thomas, J. (2020). *Constructing Artificial Data for Fine-Tuning for Low-Resource Biomedical Text Tagging with Applications in PICO Annotation* Explainable AI in Healthcare and Medicine,

Singh, G., Sabet, Z., Shawe-Taylor, J., & Thomas, J. (2021). Constructing artificial data for fine-tuning for low-resource biomedical text tagging with applications in pico annotation. *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*, 131-145.

Takanobu, R., Huang, M., Zhao, Z., Li, F.-L., Chen, H., Zhu, X., & Nie, L. (2018). A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning. IJCAI,

Tchoua, R. B., Ajith, A., Hong, Z., Ward, L. T., Chard, K., Belikov, A., Audus, D. J., Patel, S., de Pablo, J. J., & Foster, I. T. (2019). Creating training data for scientific named entity recognition with minimal human effort. Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part I 19,

Tekumalla, R., & Banda, J. M. (2021). Using weak supervision to generate training datasets from social media data: a proof of concept to identify drug mentions. *Neural Computing and Applications*, 1-9.

Thiée, L.-W. (2021). *A systematic literature review of machine learning canvases*. Gesellschaft für Informatik, Bonn. https://doi.org/10.18420/informatik2021-101

Today, A. D. (2020). *Top 10 Reasons Why AI Projects Fail*. AI & Data Today. Retrieved 21.03. from https://www.aidatatoday.com/top-10-reasons-why-ai-projects-fail/

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*(Nov), 45-66.

Trends, M. (2022). *5 Reasons Why AI Projects Fail*. Analytics Insight. Retrieved 21.03. from https://www.analyticsinsight.net/5-reasons-why-ai-projects-fail/

Trivedi, H. M., Panahiazar, M., Liang, A., Lituiev, D., Chang, P., Sohn, J. H., Chen, Y.-Y., Franc, B. L., Joe, B., & Hadley, D. (2019). Large scale semi-automated labeling of routine free-text clinical records for deep learning. *Journal of digital imaging*, *32*, 30-37.

Varma, P., & Ré, C. (2018). Snuba: Automating Weak Supervision to Label Training Data. *Proceedings VLDB Endowment*, *12*(3), 223-236. https://doi.org/10.14778/3291264.3291268

venturebeat. (2019). *Why do 87% of data science projects never make it into production?* Retrieved 21.03. from https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. *Science*, *321*(5895), 1465-1468.

Wan, Z., Xia, X., Lo, D., & Murphy, G. C. (2019). How does machine learning change software development practices? *IEEE Transactions on Software Engineering*, *47*(9), 1857-1871.

Wang, C., Han, L., Stein, G., Day, S., Bien-Gund, C., Mathews, A., Ong, J. J., Zhao, P.-Z., Wei, S.-F., Walker, J., Chou, R., Lee, A., Chen, A., Bayus, B., & Tucker, J. D. (2020). Crowdsourcing in health and medical research: a systematic review. *Infectious Diseases of Poverty*, *9*(1), 8. https://doi.org/10.1186/s40249-020-0622-9

Wang, M., Min, F., Zhang, Z.-H., & Wu, Y.-X. (2017). Active learning through density clustering. *Expert Systems with Applications*, *85*, 305-317.

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., & Liu, H. (2019). A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making*, *19*(1), 1. https://doi.org/10.1186/s12911-018-0723-6

Weber, M., Engert, M., Schaffer, N., Weking, J., & Krcmar, H. (2022). Organizational Capabilities for AI Implementation—Coping with Inscrutability and Data Dependency in AI. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-022-10297-y

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii-xxiii. http://www.jstor.org/stable/4132319

Whang, S. E., Roh, Y., Song, H., & Lee, J.-G. (2023). Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 1-23.

Wollendorfer, C. (2021). *Process and Data Mining to Improve Customer Support* Technical University of Munich (TUM)].

Yamaoka, H., Yamamoto, K., Nagai, T., & Masuda, H. (2019). *Case Study of Implementing an IT Service Desk Ticketing System at Small Computer Center* Proceedings of the 2019 ACM SIGUCCS Annual Conference, New Orleans, LA, USA. https://doi.org/10.1145/3347709.3347820

Yang, J., Fan, J., Wei, Z., Li, G., Liu, T., & Du, X. (2020). A game-based framework for crowdsourced data labeling. *The VLDB Journal*, *29*, 1311-1336.

Yitagesu, S., Xing, Z., Zhang, X., Feng, Z., Li, X., & Han, L. (2021). Unsupervised labeling and extraction of phrase-based concepts in vulnerability descriptions. 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE),

Zhou, N., Aw, A., Liu, Z. H., heng Tan, C., Ting, Y., Chen, W. X., & sim zheng Ting, J. (2022). CXR Data Annotation and Classification with Pre-trained Language Models. Proceedings of the 29th International Conference on Computational Linguistics,

Zhu, J., Wang, H., Hovy, E., & Ma, M. (2010). Confidence-based stopping criteria for active learning for data annotation. *ACM Transactions on Speech and Language Processing (TSLP)*, *6*(3), 1-24.

Zhu, Y., Jing, L., & Yu, J. (2009, 2009//). New Labeling Strategy for Semi-supervised Document Categorization. Knowledge Science, Engineering and Management, Berlin, Heidelberg.